UNBIASED RATIO AND REGRESSION TYPE ESTIMATORS

N. S. SASTRY

Institute of Agricultural Research Statistics
(I.C.A.R.)
Library Avenue
NEW DELHI-12

1964

UNBIASED RATIO AND REGRESSION TYPE ESTIMATORS

N.S. SASTRY

Dissertation
submitted in fulfilment of
the requirements for the award of
Diploma in Agricultural and Animal Husbandry Statistics
of the
INSTITUTE OF AGRICULTURAL RESEARCH STATISTICS(I.C.A.R)
NEW DELHI-12

1964

ACKNOWLEDGEMENTS

I have great pleasure in expressing my deep sense of gratitude to Dr. B.V. Sukhatme, Professor of Statistics, Institute of Agricultural Research Statistics (I.C.A.R.) New Delhi, for his valuable guidance, keen interest and constant encouragement throughout the course of investigation and immense help in writing up the thesis by critically going through the manuscript.

I am grateful to Dr. V.G. Panse, Statistical Adviser, I.C.A.R., New Delhi-12, for providing necessary facilities to carry out the investigation.

New Delhi-12
July 31, 1964.

N. S. Sastry
(N.S. SASTRY)

C O N T E N T S

INTRODUCTION AND SUMMARY

Simple random sampling is by far the most commonly used method of sampling in surveys. It is simple, operationally convenient and gives equal chance of selection for all the units in the population. When, however, the units vary considerably in size, as is often the case, simple random sampling does not take into account the possible importance of larger units in the population. Under such circumstances, without foregoing the operational convenience of simple random sampling, it is desirable to use auxiliary information, such as size of unit, at the estimation stage, for obtaining more efficient estimators of the population value in the sense of giving estimators with smaller standard errors. Two examples of such estimation procedures are the 'Ratio and Regression' methods of estimation.

The two classical ratio estimators are the ratio of means estimator $\bar{y}_R = \frac{\bar{y}_n}{\bar{x}_n} \bar{X}$ or equivalently the weighted mean of the ratios estimator $\bar{r}_w \bar{X} = \frac{\sum_i^n x_i r_i}{\sum_i^n x_i} \bar{X}$ and the mean of ratios estimator $\bar{y}_r = \bar{r}_n \bar{X}$, where $\bar{y}_n$ and $\bar{x}_n$ are the sample means, $\bar{r}_n$, the mean of individual ratios $r = \frac{y}{x}$ in the sample and $\bar{X}$ is the known population mean of the auxiliary variable x. Both the estimators are known to be biased. The latter is not even consistent. An exact expression for the bias in $\bar{y}_r$ is available which does not depend upon the sample size n. Since the unweighted mean $\bar{r}_n$ may be seriously biased if r tends to be larger (or smaller) for large x than for small x, estimator $\bar{y}_r$ is likely to be more biased than the estimator $\bar{y}_R$ based on the weighted mean $\bar{r}_w$. No exact expression for the bias in $\bar{y}_R$ is available; however, for samples of moderate size, from populations in which the linear regression of y on x passes near the origin, and in which the coefficient of variation of x is not too large, the bias in $\bar{y}_R$ is negligible. But the problem how large the sample should be to make the bias negligible has not yet been solved satisfactorily for all types of populations.

The classical regression estimator is obtained by evaluating the least squares line of best fit $y = \bar{y}_n + b_n(x - \bar{x}_n)$ at the point $\bar{X}$, giving $\bar{y}_{lr} = \bar{y}_n + b_n(\bar{X} - \bar{x}_n)$ as a regression estimator of the population mean, where b_n is the sample regression coefficient of y on x. Except when the true regression line passes through the origin, the regression estimator is less biased, and more precise, than the ratio estimator $\bar{y}_R$. On the other hand, the ratio estimator is more easily calculated.

The available bias expressions and variance formulae for both the regression estimator $\bar{y}_{lr}$ and the ratio estimator $\bar{y}_R$ are only approximate; the approximations assuming the sample size n to be sufficiently large. For small samples nothing is known about the nature of their bias and precision. This situation has led the research workers in the field to explore ways and means of obtaining 'Ratio and Regression type' estimators, which are either completely free from bias or subject to a smaller bias than the customary ones.

In this connection Quenouille (1956) has suggested a technique of reducing the bias in the ratio estimator $\bar{y}_R$.

By splitting a sample of size 2n at random into two sub-samples of sizes n each, he has considered a weighted average of the three ratio estimators of the form $\bar{y}_R$, applied to the total sample and the two sub-samples, where the weights are chosen in such a way as to reduce the bias to the order $\frac{1}{n^2}$. Ravindra Singh (1962) has further investigated the technique and examined the optimum sizes for the two sub-samples with respect to the order of decrease in bias and the efficiency of the modified ratio estimator as compared to the ordinary one. Murty and Nanjamma (1959) have developed a technique of estimating the bias of a ratio estimator unbiasedly to any given degree of approximation and used this estimator of bias to correct the ratio estimator for its bias, thereby getting an 'almost unbiased ratio estimator'.

Unbiased ratio and regression type estimators have been evolved in recent years, following two different approaches. Of the ratio and regression type, the ratio type has received much attention. The first approach consists in getting new types of unbiased ratio and regression estimators under commonly adopted sampling schemes. The second approach is to modify the sampling scheme so as to make the usual ratio estimator (i.e., Ratio of the unbiased estimator of the population total of y to the unbiased estimator of the population total of x under the original scheme of sampling) unbiased.

Hartley and Ross (1954) have been the pioneers in the group of authors who tried to obtain unbiased estimators under the commonly adopted sampling schemes. In simple random sampling without replacement, they have given an elegant expression for the bias in $\bar{y}_r$, and an unbiased estimator of that bias, thereby arriving at an unbiased ratio estimator

$$\bar{y}'_r = \bar{r}_n \bar{X} + \frac{(N-1)n}{N(n-1)} (\bar{y}_n - \bar{r}_n \bar{x}_n).$$

Robson (1957) has derived the exact formula for its variance, and an unbiased estimator of the variance. In large samples, more simple estimators of the variance of $\bar{y}'_r$ and an extensive discussion of the relative efficiencies of estimators $\bar{y}_R$, $\bar{y}_r$ and $\bar{y}'_r$ have been given by Goodman and Hartley (1958).

Sukhatme (1962) has obtained a generalized form of $\bar{y}'_r$ for multistage designs. A double sampling version of $\bar{y}'_r$, in which the unknown population mean of x is replaced by the sample mean of x based on a larger preliminary sample, without disturbing the property of unbiasedness, has been given by Sukhatme (1962) together with a comparison of its large sample efficiency with the double sampling versions of $\bar{y}_R$ and $\bar{y}_r$. Joso Nieto Pascual (1961) considers, in a stratified population, a 'separate' unbiased ratio estimator which is a straight-forward generalization of $\bar{y}'_r$, and a 'combined' unbiased ratio estimator, the latter being based on a slightly different sampling scheme. The scheme for the combined estimator consists in drawing K independent stratified samples, each sample containing one unit selected at random from each of the strata. In large samples, he has obtained a comparison of the combined unbiased estimator with the usual combined biased estimator and also a comparison of the separate Hartley and Ross unbiased estimator with the usual separate biased estimator.

Mickey (1959) has putforward a general theory for constructing 'unbiased ratio and regression type' estimators in simple random sampling without replacement, using information on the population means of several auxiliary variates. For a sub class of his general class of estimators he has obtained non-negative unbiased estimators of the variance. No attempt has, however, been made to investigate the variance of the proposed class of unbiased estimators.

Williams (1961, 1963) has considered a hypothetical two stage sampling scheme, in which at the first stage, one of the possible ~~steps~~ splits of the whole population into s mutually exclusive and exhaustive groups of size $\frac{n}{k}$ each (i.e. population size $N = \frac{ns}{k}$) is selected at random, followed by the selection with equal probability without replacement of k of the groups. For a given split of the population and a random selection of the groups, conditionally he has obtained a general class of unbiased ratio and regression type estimators. In actual usage the groups are obtained by splitting a simple random sample without replacement of size n from the whole population, but not by splitting the population itself. The same principle is extended to obtained unbiased estimators in multistage designs, as also to obtain a 'combined' unbiased estimator in stratified populations. He has also discussed the unbiased estimation of estimator variance and the precision of the regression type estimators.

The underlying principle in the approach of a second group of authors in evolving unbiased ratio type estimators stems from the following considerations. If $\hat{Y}_s$ and $\hat{X}_s$ are unbiased estimators of the population totals Y and X, based on the s^{th} sample selected with any given sampling design, then the ratio estimator $\hat{R} = \frac{\hat{Y}_s}{\hat{X}_s}$ will be unbiased for the ratio $R = Y/X$, if the design is so changed that P_s, the probability of selecting the s^{th} sample is proportional to $\hat{X}_s P'_s$ where P'_s is the probability of selecting the s^{th} sample in the original sampling design, i.e. if $P_s = \frac{\hat{X}_s P'_s}{X}$. If, further, P'_s is same for all s, then P_s should be made proportional to $\hat{X}_s$ to make the ratio estimator unbiased.

Thus in the case of simple random sampling without replacement the ratio $\frac{\bar{y}_n}{\bar{x}_n}$ of the sample means would be unbiased for the population ratio if the original sampling design is modified so as to make the probability of selecting the s^{th} sample proportional to $(\bar{x}_n)_s$, or in other words proportional to the total size of the sample. Lahiri (1951), Midzuno (1952) and Sen (1952) have independently given sampling procedures for obtaining a sample with probability proportional to its total size. Based on these procedures of selection, Des Raj (1954) has given modified sampling schemes appropriate to unistage, stratified, multistage and multiphase designs, in the case of simple random sampling without replacement, which eliminate the bias of the usual ratio estimator.

Murty, Nanjamma and Sethi (1959) have given modifications of many of the selection procedures, commonly adopted in practice, which, while retaining the form of the usual biased ratio-estimator, make them unbiased. The method suggested by them consists essentially in selecting the first unit with probability proportional to the auxiliary variate, and the remaining units in the sample according to the original sampling scheme. Indeed the method is very elegant and provides easily calculable unbiased estimators of the sampling variances also, but its utility is limited in practice as it assumes the knowledge of all the x values in the population, in which case a better sampling design can be formulated. Recently Pathak (1964) has shown that if in these modified sampling schemes a sufficient statistic is available and if the ratio estimator does not depend upon the sufficient statistic, it can be uniformly improved by Rao-Blackwell theorem. This result has been used by him in deriving unbiased ratio estimators, better than the ratio estimators given by Murty, Nanjamma and Sethi.

The present investigation is a critical study of Mickey's unbiased ratio and regression type estimators. Section 1 deals with the unbiased estimation of the variance of Mickey's estimator in its general form. Section 2 is concerned with the investigation of the precision of Mickey's unbiased ratio type estimators, utilizing information on a single auxiliary variable; and a comparison of their efficiency with the usual biased ratio estimator $\bar{y}_R$ in large samples. In section 3 an attempt is made to obtain a large sample formula for the variance of the unbiased regression type estimators based on a single auxiliary variable; and to compare their efficiency with the usual biased regression estimator and the corresponding Mickey's ratio type estimators. Section 4 is a study of the unbiased ratio type estimators based on two auxiliary variables in respect of their precision, in large samples, and their relative efficiency compared with the Olkin's weighted and biased ratio estimator. Section 5 deals with the development of Mickey's principle to obtain unbiased ratio and regression type estimators in two-phase sampling. It also includes the unbiased estimation of their variances and a discussion of their large sample efficiency compared with the usual biased ratio and regression estimators in two-phase sampling. In Section 6 separate and combined unbiased ratio type estimators, based on Mickey's principle, are given for stratified simple random sampling without replacement together with the unbiased estimation of their variance. Finally, Section 7 gives some numerical results concerning the performance of the unbiased ratio type estimators with respect to the usual biased ratio estimators.

1. UNBIASED ESTIMATION OF ESTIMATOR VARIANCE

Mickey's 'unbiased ratio and regression estimators' are particular cases of general class of unbiased estimators, developed by him. A brief account of the sampling frame, sampling design, and the construction procedure of the general class of estimators is necessary to outline the results of this Section.

Mickey's unbiased estimators

Let the finite population of size N be represented by a set of (p+1) component vectors

$$(y_j, x_{1j}, x_{2j}, \ldots\ldots\ldots, x_{pj}), \quad j=1,2,\ldots\ldots\ldots,N,$$

where $x_1, x_2, \ldots, x_p$ are p auxiliary variables with known population means $\bar{X}_1, \bar{X}_2, \ldots\ldots, \bar{X}_p$. The problem is to estimate the unknown population mean $\bar{Y}$ of the variable y under study. For this purpose, a simple random sample of size n is selected (without replacement) from the population. Let $\bar{y}$, and $\bar{x}_i$, $i=1,2,\ldots\ldots,p$ denote the sample means.

Given this sample, for any choice of m of the sample elements, $m < n$, the remaining n-m elements constitute a simple random sample of size n-m from the finite population of N-m elements derived by excluding the selected m elements from the given population. Let z_m represent the m sample elements so chosen out of the given sample and $a_i(z_m)$, $i=1,2,\ldots,p$ denote some known real valued functions of the observations z_m. Further let $\bar{y}_m$, and $\bar{x}_{im}$, $i=1,2,\ldots.,p$, be the means of the observations z_m.

Now define

$$\bar{y}_{n-m} = \frac{n\bar{y} - m\bar{y}_m}{n - m}, \quad \bar{x}_{in-m} = \frac{n\bar{x}_i - m\bar{x}_{im}}{n - m}$$

and
$$\bar{Y}_{N-m} = \frac{N\bar{Y} - m\bar{y}_m}{N - m}, \quad \bar{X}_{iN-m} = \frac{N\bar{X}_i - m\bar{x}_{im}}{N - m}, \quad i=1,2,\ldots,p.$$

Further let U_m denote the statistic given by

$$U_m = \bar{y}_{n-m} - \sum_{i=1}^{p} a_i(z_m)(\bar{x}_{in-m} - \bar{X}_{iN-m}) \quad .. \quad (1.1)$$

Then, if $E(U_m/m)$ denotes the conditional expectation for a given set z_m, we have

$$E(U_m/m) = \bar{Y}_{N-m}$$

Consequently, if
$$T_m = \frac{(N-m)U_m + m\bar{y}_m}{N}, \quad .. \quad (1.2)$$

then
$$E(T_m/m) = \bar{Y}$$

Hence unconditionally also T_m provides an unbiased estimator of the population mean $\bar{Y}$.

The estimator T_m which can also be written in the following equivalent forms:

$$T_m = \bar{y} - \sum_{i=1}^{p} a_i(z_m)(\bar{x}_i - \bar{X}_i) - \frac{m(N-n)}{N(n-m)} \left[\bar{y}_m - \bar{y} - \sum_{i=1}^{p} a_i(z_m)(\bar{x}_{im} - \bar{x}_i) \right] \quad (1.3)$$

$$= \frac{(N-m)n}{N(n-m)} \left[\bar{y} - \sum_{i=1}^{p} a_i(z_m)(\bar{x}_i - \bar{X}_i) \right] - \frac{m(N-n)}{N(n-m)} \left[\bar{y}_m - \sum_{i=1}^{p} a_i(z_m)(\bar{x}_{im} - \bar{X}_i) \right] \quad (1.4)$$

$$= \sum_{i=1}^{p} a_i(z_m)\bar{X}_i + \frac{(N-m)n}{N(n-m)} \left[\bar{y} - \sum_{i=1}^{p} a_i(z_m)\bar{x}_i \right] - \frac{(N-n)m}{N(n-m)} \left[\bar{y}_m - \sum_{i=1}^{p} a_i(z_m)\bar{x}_{im} \right] \quad (1.5)$$

is hereafter called Mickey's unbiased estimator, in its general form; meaning thereby Mickey's unbiased estimator T_m with arbitrary coefficient functions $a_i(z_m)$, $i=1,2,..,p$.

Here z_m may be taken as the observations on the first m draws of the sample of size n or as any subsample of size m from the given sample. Thus a general class of unbiased estimators can be obtained by taking weighted averages of estimators of the form T_m, applied to all possible permutations of the sample elements. Of particular interest is the estimator T_m^* obtained from T_m by averaging over all possible permutations. This is so because the unordered sample plays the role of a sufficient statistic and by an application of Rao-Blackwell theorem it follows that the variance of T_m^* is never greater than that of T_m.

Mickey has obtained unbiased Ratio and Regression type estimators as particular cases of the estimator T_m in its general form, by a proper choice of the coefficient functions $a_i(z_m)$, $i=1,2,...,p$.

Unbiased Ratio type estimators

For example, when information on only one auxiliary variable is available, the choice $a(z_m) = \bar{y}_m / \bar{x}_m = R_m$, applied to the form (1.5) of T_m, provides an unbiased ratio type estimator T_{1m} given by

$$T_{1m} = R_m\bar{X} + \frac{(N-m)n}{N(n-m)} \left[\bar{y} - R_m\bar{x} \right]. \quad .. \quad (1.6)$$

Averaging over all possible permutations, we obtain the more efficient unbiased ratio type estimator T_{1m}^* given by

$$\overset{*}{T}_{1m} = \overset{*}{R}_m\bar{X} + \frac{(N-m)n}{N(n-m)} \left[\bar{y} - \overset{*}{R}_m\bar{x} \right], \quad .. \qquad (1.7)$$

where $\overset{*}{R}_m$ is the average of R_m over all permutations.

Unbiased Regression type estimators

$$\text{With the choice } a(z_m) = \frac{\sum_j^m (y_j-\bar{y}_m)x_j}{\sum_j^m (x_j-\bar{x}_m)^2} = b_m ,$$

the usual linear regression coefficient based on the observations z_m, expression (1.3) for T_m yields a regression type estimator T_{2m} given by

$$T_{2m} = \left[\bar{y} - b_m(\bar{x}-\bar{X}) \right] - \frac{m(N-n)}{(n-m)N} \left[\bar{y}_m-\bar{y}-b_m(\bar{x}_m-\bar{x}) \right] (1.8)$$

Averaging over all permutations we obtain

$$\overset{*}{T}_{2m} = \left[\bar{y}-\overset{*}{b}_m(\bar{x}-\bar{X}) \right] + \frac{m(N-n)}{(n-m)N} \cdot \frac{1}{\binom{n}{m}} \sum^{\binom{n}{m}} b_m(\bar{x}_m-\bar{x}) , \qquad (1.9)$$

where $\overset{*}{b}_m$ is the average of b_m over all permutations.

The present section deals with the unbiased estimation of the variance of Mickey's unbiased estimators T_m and $\overset{*}{T}_m$, in their general form.

Unbiased estimator of the variance of T_m

From the well known formula, connecting variance, with the conditional expectation and conditional variance, we have

$$V(T_m) = E\left\{ V(T_m/m) \right\} + V\left\{ E(T_m/m) \right\}$$

$$= \bar{E}\left\{ V\left[\frac{(N-m)U_m+m\bar{y}_m}{N} \Big/ m \right] \right\}$$

$$= \frac{(N-m)^2}{N^2} E \left[V(U_m/m) \right]$$

$$= \frac{(N-m)^2}{N^2} E \left[V(\bar{y}_{n-m} - \sum_{i=1}^{p} a_i(z_m)\bar{x}_{i\,n-m} / m) \right], \quad (1.10)$$

from the definition (1.1) of U_m.

Now we observe that

$$\bar{y}'_{n-m} = \bar{y}_{n-m} - \sum_{i=1}^{p} a_i(z_m)\bar{x}_{i\,n-m}$$

is the arithmetic mean of the observations

$$y'_j = y_j - \sum_{i=1}^{p} a_i(z_m)x_{ij},$$

made on the n-m sample elements which are obtained by excluding z_m from the given sample of size n. Further, since for a given choice of z_m, the remaining n-m sample elements constitute a random sample from the derived population of size N-m, it can be seen that a non-ve unbiased estimator of

$$V(\bar{y}'_{n-m} / m)$$

is provided by

$$\frac{N-n}{(N-m)(n-m)} s^2_{y',n-m}, \quad m=1,2,\ldots,n-2 \quad (1.11)$$

where $$s^2_{y',n-m} = \frac{1}{n-m-1} \sum_{j}^{n-m} (y'_j - \bar{y}'_{n-m})^2,$$

the summation being taken over the remaining n-m sample elements. Consequently, from (1.10) and (1.11), a non-ve unbiased estimator of the variance of T_m is given by

$$\text{Est.}V(T_m) = \frac{(N-n)(N-m)}{N^2(n-m)} \cdot \frac{1}{n-m-1} \sum_{j}^{n-m} \left[(y_j - \bar{y}_{n-m}) - \sum_{i=1}^{p} a_i(z_m)(x_{ij} - \bar{x}_{i\,n-m}) \right]^2 \quad (1.12)$$

This estimator holds good for all values of m less than n except for m = n-1, in which case there is only one observation of the type y'_j. Further it can be seen that the reliability of the Est.$V(T_m)$ is more for small values of m than for the choices of m near to the total sample size n, as $s^2_{y',n-m}$ is based on n-m-1 degrees of freedom.

Unbiased estimator of the variance of T_m^*

Since for a given sample of size n, $E(T_m / n) = T_m^*$, we have

$$V(T_m) = E(T_m - T_m^*)^2 + V(T_m^*).$$

Averaging over all the possible $\binom{n}{m}$ estimators of the form T_m will therefore give

$$\frac{1}{\binom{n}{m}} \sum^{\binom{n}{m}} V(T_m) = E \frac{1}{\binom{n}{m}} \sum^{\binom{n}{m}} (T_m - T_m^*)^2 + V(T_m^*). \qquad (1.13)$$

From this it follows that an unbiased estimator of the variance of T_m^* is given by

$$\text{Est. } V(T_m^*) = \frac{1}{\binom{n}{m}} \sum_j^{n-m} \left[\text{Est.}V(T_m) - (T_m - T_m^*)^2 \right], \qquad (1.14)$$

where Est.$V(T_m)$ is provided by (1.12).

Particular cases

For the ratio type estimators T_{1m} and T_{1m}^*, from (1.12) and (1.14) unbiased estimators of the variance are provided by

$$\text{Est. } V(T_{1m}) = \frac{(N-n)(N-m)}{N^2(n-m)} \cdot \frac{1}{(n-m-1)} \sum_j^{n-m} \left[(y_j - \bar{y}_{n-m}) - R_m(x_j - \bar{x}_{n-m}) \right]^2$$

(1.15)

$$\text{and Est. } V(T^*_{1m}) = \frac{1}{\binom{n}{m}} \sum^{\binom{n}{m}} \text{Est.} V(T_{1m})$$

$$- \left(\bar{X} - \frac{(N-m)n}{N(n-m)} \bar{x}\right)^2 \frac{1}{\binom{n}{m}} \sum^{\binom{n}{m}} (R_m - R^*_m)^2 \qquad (1.16)$$

For the regression type estimators T_{2m} and T^*_{2m}, unbiased estimators of the variance are given by

$$\text{Est. } V(T_{2m}) = \frac{(N-n)(N-m)}{N2(n-m)} \cdot \frac{1}{(n-m-1)} \sum_{j}^{n-m} \left[(y_j - \bar{y}_{n-m}) - b_m(x_j - \bar{x}_{n-m}) \right]^2 \qquad (1.17)$$

$$\text{and Est. } V(T^*_{2m}) = \frac{1}{\binom{n}{m}} \sum^{\binom{n}{m}} \left[\text{Est.} V(T_{2m}) - (T_{2m} - T^*_{2m})^2 \right] \qquad (1.18)$$

2. PRECISION OF RATIO TYPE ESTIMATORS WITH ONE AUXILIARY VARIABLE

Goodman and Hartley (1958) have investigated, in large samples, the relative efficiencies of the ratio type estimators $\bar{y}_R = \frac{\bar{y}}{\bar{x}} \bar{X}$, $\bar{y}_r = \bar{r} \bar{X}$,

and $\bar{y}_r' = \bar{r}\bar{X} + \frac{(N-1)n}{N(n-1)} (\bar{y} - \bar{r}\bar{x})$, $(\bar{r} = \frac{1}{n} \sum_j^n \frac{y_j}{x_j})$

of which the first two are biased estimators and the third an unbiased estimator of $\bar{Y}$. They have shown that in large samples the estimator $\bar{y}_r'$ is more efficient than $\bar{y}_R$, if and only if the slope of the regression line of y on x is closer to

$\bar{r}_p = \frac{1}{N}\sum_{j=1}^{N}\frac{y_j}{x_j}$ than to the population ratio $R = \bar{Y} / \bar{X}$.

In this section, we shall investigate the relative efficiencies of Mickey's unbiased ratio type estimators:

$$T_{1m} = R_m \bar{X} + \frac{(N-m)n}{N(n-m)} (\bar{y} - R_m \bar{x}) \tag{2.1}$$

and

$$T_{1m}^* = R_m^* \bar{X} + \frac{(N-m)n}{N(n-m)} (\bar{y} - R_m^* \bar{x}) \tag{2.2}$$

with respect to the conventional biased ratio estimator $\bar{y}_R$, for large samples.

We shall first obtain the variance of T_{1m}, in large samples, for m sufficiently large and compare it with the variance of the usual biased ratio estimator $\bar{y}_R$. After this an expression for the variance of T_{1m}^*, in large samples, will be derived. It will be seen that considerable simplification is effected in the large sample variance expression of T_{1m}^*

when either m is small as compared to n or when m is sufficiently large. From the practical point of view these two cases are most important as in these two cases only computation of the estimator T^*_{1m} and of the unbiased estimator of its variance, given in section 1, is most convenient. Finally, assuming the population to follow a bivariate normal distribution, we shall further investigate the variances of T_{1m} and T^*_{1m}, when m is large; and discuss the relative efficiency of T^*_{1m} with respect to $\bar{y}_R$.

Variance of T_{1m} for large m

To derive an expression for the variance of the estimator T_{1m} in large samples for m sufficiently large, the results of the following lemma will be useful.

Lemma:- In simple random samples (without replacement) of size n from a bivariate finite population of N pairs (x_i, y_i) $i=1,2,\ldots,N$,

$$\mathrm{Cov}(\bar{x}, s_x^2) = K_{n,N}\ \mu_{03},\quad \mathrm{Cov}(\bar{y}, s_x^2) = K_{n,N}\ \mu_{12},$$

$$\mathrm{Cov}(\bar{x}, s_{xy}) = K_{n,N}\ \mu_{12},\quad \mathrm{Cov}(\bar{y}, s_{xy}) = K_{n,N}\ \mu_{21},$$

where $\bar{x}$ and $\bar{y}$ are the sample means, $\bar{X}$ and $\bar{Y}$ are the population means,

$$s_x^2 = \frac{1}{n-1}\sum_{i}^{n}(x_i-\bar{x})^2,\quad s_{xy} = \frac{1}{n-1}\sum_{i}^{n}(x_i-\bar{x})(y_i-\bar{y}),$$

$$\mu_{03} = \frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{X})^3,\quad \mu_{12} = \frac{1}{N}\sum_{i=1}^{N}(y_i-\bar{Y})(x_i-\bar{X})^2,$$

$$\mu_{21} = \frac{1}{N}\sum_{i=1}^{N}(y_i-\bar{Y})^2(x_i-\bar{X}) \text{ and}$$

$$K_{n,N} = \left[\frac{1}{n} - \frac{n-3}{n(N-1)} - \frac{2(n-2)}{n(N-1)(N-2)}\right].$$

Proof:- Without loss of generality, in the evaluation of these covariances, one may assume the population means $\bar{X}$ and $\bar{Y}$ to be zeros.

Then

$$\mathrm{Cov}(\bar{y}, s_x^2) = \frac{1}{n(n-1)} E\left[\left(\sum_i^n y_i\right)\left\{\sum_i^n x_i^2 - \frac{\left(\sum_i^n x_i\right)^2}{n}\right\}\right]$$

$$= \frac{1}{n(n-1)} E\left[\frac{n-1}{n}\left(\sum_i^n y_i x_i^2 + \sum_{i\neq j}^n y_i x_j^2\right) - \frac{2}{n}\sum_{i\neq j}^n y_i x_i x_j - \frac{1}{n}\sum_{i\neq j\neq k}^n y_i x_j x_k\right]$$

$$= \frac{1}{n(n-1)}\left[\frac{n-1}{n}\left(c_1\sum_{i=1}^N y_i x_i^2 + c_2\sum_{i\neq j}^N y_i x_j^2\right) - \frac{2c_2}{n}\sum_{i\neq j}^N y_i x_i x_j - \frac{c_3}{n}\sum_{i\neq j\neq k}^N y_i x_j x_k\right]$$

where, $c_1 = \frac{n}{N}$, $c_2 = \frac{n(n-1)}{N(N-1)}$ and $c_3 = \frac{n(n-1)(n-2)}{N(N-1)(N-2)}$.

Thus

$$\mathrm{Cov}(\bar{y}, s_x^2) = \frac{1}{n(n-1)}\left[\frac{n-1}{n}(c_1 N\mu_{12} - c_2 N\mu_{12}) + \frac{2c_2}{n} N\mu_{12} - \frac{2c_3}{n} N\mu_{12}\right]$$

$$= K_{n,N}\,\mu_{12}, \text{ where } K_{n,N} = \left[\frac{1}{n} - \frac{n-3}{n(N-1)} - \frac{2(n-2)}{n(N-1)(N-2)}\right].$$

Putting x = y in this result we have $\mathrm{Cov}(\bar{x}, s_x^2) = K_{n,N}\,\mu_{03}$.

Now

$$\mathrm{Cov}(\bar{x}, s_{xy}) = \frac{1}{n(n-1)} E\left[\left(\sum_i^n x_i\right)\left\{\sum_i^n x_i y_i - \frac{\left(\sum_i^n x_i\right)\left(\sum_i^n y_i\right)}{n}\right\}\right]$$

$$= \frac{1}{n(n-1)} E \Big[\frac{n-1}{n} \big(\sum_{i}^{n} x_i^2 y_i + \sum_{i \neq j}^{n} x_i x_j y_j \big) - \frac{1}{n} \sum_{i \neq j}^{n} x_i^2 y_j$$

$$- \frac{1}{n} \sum_{i \neq j}^{n} x_i x_j y_j - \frac{1}{n} \sum_{i \neq j \neq k}^{n} x_i x_j y_k \Big]$$

$$= \frac{1}{n(n-1)} \Big[\frac{n-1}{n} (c_1 N \mu_{12} - c_2 N \mu_{12}) + \frac{2c_2}{n} N \mu_{12} - \frac{2c_3}{n} N \mu_{12} \Big]$$

$$= K_{n,N} \mu_{12}.$$

Similarly $\mathrm{Cov}(\bar{y}, s_{xy}) = K_{n,N} \mu_{21}$. Q.E.D.

We now proceed to derive $V(T_{1m})$.

Since T_{1m} results from T_m, in its general form, by putting $p = 1$ and $a(z_m) = R_m$, we have from (1.10) Section 1,

$$V(T_{1m}) = \frac{(N-m)^2}{N^2} E \Big[V(\bar{y}_{n-m} - R_m \bar{x}_{n-m} / m) \Big]$$

$$= \frac{(N-m)(N-n)}{N^2(n-m)} E \Big[S^2_{N-m,y} + R_m^2 \; S^2_{N-m,x} - 2R_m \, S_{N-m,xy} \Big], \qquad (2.3)$$

where $S^2_{N-m,y}$, $S^2_{N-m,x}$ and $S_{N-m,xy}$ are the mean sums of squares and mean sum of products in the derived population of size $N-m$.

Write $\bar{y}_m = \bar{Y} + e_1$, $\bar{x}_m = \bar{X} + e_2$, $S^2_{N-m,x} = S^2_x + e_3$ and $S_{N-m,xy} = S_{xy} + e_4$, where S^2_y, S^2_x and S_{xy} are the mean sums of squares and mean sum of products in the population of size N, and

$$E(e_1) = E(e_2) = E(e_3) = E(e_4) = 0$$

$$E(e_1^2) = \frac{N-m}{Nm} S^2_y, \; E(e_2^2) = \frac{N-m}{Nm} S^2_x, \text{ and } E(e_1 e_2) = \frac{N-m}{Nm} S_{xy}.$$

Also $E(e_1e_3) = Cov(\bar{y}_m, s^2_{N-m,x})$

$$= -\frac{N-m}{m} Cov(\bar{Y}_{N-m}, s^2_{N-m,x})$$

$$= -\frac{N-m}{m} K_{N-m,N}\, \mu_{12}, \text{ from the lemma}$$

$$= -\frac{N}{(N-1)(N-2)}\, \mu_{12} .$$

Similarly

$$E(e_2e_3) = -\frac{N}{(N-1)(N-2)}\, \mu_{03} ,$$

$$E(e_1e_4) = -\frac{N}{(N-1)(N-2)}\, \mu_{21} ,$$

$$\text{and } E(e_2e_4) = -\frac{N}{(N-1)(N-2)}\, \mu_{12} . \qquad (2.4)$$

Now in the formula (2.3), to evaluate the term $E(R_m^2\, s^2_{N-m,x})$ we assume that m is sufficiently large, and write

$$E(R_m^2\, s^2_{N-m,x}) = R^2 s_x^2\, E\Big[(1+\frac{e_1}{\bar{Y}})^2(1+\frac{e_2}{\bar{X}})^{-2}(1+\frac{e_3}{s_x^2})\Big]$$

$$= R^2 s_x^2\, E\Big[(1+\frac{2e_1}{\bar{Y}}+\frac{e_1^2}{\bar{Y}^2})(1-\frac{2e_2}{\bar{X}}+\frac{3e_2^2}{\bar{X}^2})(1+\frac{e_3}{s_x^2})\Big]$$

$$= R^2 s_x^2\, E\Big[1+\frac{2e_1}{\bar{Y}}-\frac{2e_2}{\bar{X}}+\frac{e_3}{s_x^2}+\frac{2e_1e_3}{\bar{Y}s_x^2}-\frac{4e_1e_2}{\bar{Y}\bar{X}}$$

$$-\frac{2e_2e_3}{\bar{X}s_x^2}+\frac{e_1^2}{\bar{Y}^2}+\frac{3e_2^2}{\bar{X}^2}\Big], \quad .. \qquad (2.5)$$

neglecting expectations of cubic and higher powers in e's.

Substituting the expected values from (2.4) in (2.5), we obtain, to the order of approximation 1/m,

$$E(R_m^2\, s_{N-m,x}^2) = R^2 S_x^2 \left[1 + \frac{N-m}{Nm}\,(C_y^2 + 3C_x^2 - 4C_{xy}) + \frac{2N}{(N-1)(N-2)}\left(\frac{\mu_{03}}{\bar{X}S_x^2} - \frac{\mu_{12}}{\bar{Y}S_x^2}\right) \right] \quad (2.6)$$

where $C_y^2 = S_y^2/\bar{Y}^2$, $C_x^2 = S_x^2/\bar{X}^2$ and $C_{xy} = S_{xy}/\bar{Y}\bar{X}$.

Proceeding on similar lines, to the same order of approximation, it can be seen that

$$E(R_m s_{N-m,xy}) = R\, S_{xy} \left[1 + \frac{N-m}{Nm}(C_x^2 - C_{xy}) + \frac{N}{(N-1)(N-2)}\left(\frac{\mu_{12}}{\bar{X}S_{xy}} - \frac{\mu_{21}}{\bar{Y}S_{xy}}\right) \right] \quad (2.7)$$

Substituting the results (2.6) and (2.7) in (2.3) and simplifying, we obtain

$$V(T_{1m}) = \bar{Y}^2 \frac{(N-m)(N-n)}{N^2(n-m)} \left[(C_y^2 + C_x^2 - 2C_{xy})\left(1 + \frac{N-m}{Nm} C_x^2\right) + \frac{2(N-m)}{Nm}(C_x^2 - C_{xy})^2 \right]$$

$$- 2\, \frac{(N-m)(N-n)}{N(N-1)(N-2)(n-m)} \left[R^2\left(\frac{\mu_{12}}{\bar{Y}} - \frac{\mu_{03}}{\bar{X}}\right) - R\left(\frac{\mu_{21}}{\bar{Y}} - \frac{\mu_{12}}{\bar{X}}\right) \right]. \quad (2.8)$$

If the finite population correction factor is negligible, then $V(T_{1m})$ simplifies to

$$V(T_{1m}) = \frac{\bar{Y}^2}{(n-m)} \left[(C_y^2 + C_x^2 - 2C_{xy})\left(1 + \frac{C_x^2}{m}\right) + \frac{2}{m}\,(C_x^2 - C_{xy})^2 \right]. \quad (2.9)$$

<u>Efficiency of T_{1m} for large m</u>

In large samples from a large population, the variance of the usual biased ratio estimator $\bar{y}_R$ is given by

$$V(\bar{y}_R) = \frac{\bar{Y}^2}{n}(C_y^2 + C_x^2 - 2C_{xy}). \tag{2.10}$$

A comparison of (2.9) and (2.10) clearly shows that, when m is large and the finite population correction factor is negligible, the unbiased ratio estimator T_{1m} is less efficient than the usual biased ratio estimator $\bar{y}_R$.

Also from (2.10), we can write

$$V(T_{1m}) = \frac{A}{n-m} + \frac{B}{m(n-m)}$$

where A and B are constants independent of m; which indicates that $V(T_{1m})$ increases as m increases from n/2 to n-1. Thus in large populations, with a sufficiently large sample, efficiency of T_{1m} goes down as the choice of m is made closer and closer to the total sample size n.

It may be noted that T_{1m} is mainly dependent on the unbiased estimator of the mean of the derived population and that this unbiased estimator is based on the derived sample of size n-m. Now as the choice of m is made nearer and nearer to the total sample size n, the size of the derived sample decreases and consequently the precision of the unbiased estimator of the derived population is likely to decrease. This may be one of the reasons for the decrease in the efficiency of T_{1m} as m approaches the total sample size n, in large populations.

Variance of T_{1m}^* in large samples

Putting $k = \frac{(N-m)n}{N(n-m)}$, from the definition (2.2) of T_{1m}^*, we have

$$V(T^*_{1m}) = \bar{X}^2 V(R^*_m)+k^2 V(\bar{y})+ k^2 V(\bar{x}R^*_m)$$

$$+ 2k\bar{X} \text{Cov}(\bar{y}, R^*_m)-2k\bar{X} \text{Cov}(R^*_m, \bar{x} R^*_m)-2k^2 \text{Cov}(\bar{y},\hat{\bar{x}} R^*_m). \quad (2.11)$$

Now write $\bar{y} = \bar{Y}+e_1$, $\bar{x} = \bar{X}+e_2$ and $R^*_m = \bar{R}_m + e_3$,

where $\bar{R}_m = \frac{1}{\binom{N}{n}} \sum^{\binom{N}{n}} R^*_m$, the summation being taken over all the possible samples;

so that $E(e_1)= E(e_2)=E(e_3)=0$

$$V(e_1)= V(\bar{y}),\ V(e_2) = V(\bar{x}),\ V(e_3) = V(R^*_m)$$

and $\text{Cov}(e_1 e_2)= \text{Cov}(\bar{y},\bar{x})$, $\text{Cov}(e_1,e_3) = \text{Cov}(\bar{y},R^*_m)$, $\text{Cov}(e_2,e_3)=\text{Cov}(\bar{x},R^*_m)$

Neglecting expectations of terms in e's of order 3 or more, we have then

$$V(\bar{x} R^*_m) = V(\bar{R}_m \bar{X} + e_3 \bar{X}+e_2 \bar{R}_m + e_2e_3)$$

$$\simeq \bar{X}^2 V(R^*_m) + \bar{R}^2_m V(\bar{x}) + 2\bar{X} \bar{R}_m \text{Cov}(\bar{x}, R^*_m) \quad (2.12)$$

Similarly, to the same order of approximation,

$$\text{Cov}(R^*_m, \bar{x} R^*_m)\simeq \bar{X} V(R^*_m) + \bar{R}_m \text{Cov}(\bar{x}, R^*_m), \quad (2.13)$$

and

$$\text{Cov}(\bar{y},\bar{x} R^*_m) = \bar{X} \text{Cov}(\bar{y},R^*_m) + \bar{R}_m \text{Cov}(\bar{y},\bar{x}). \quad (2.14)$$

Using (2.12), (2.13) and (2.14) in (2.11), we obtain

$$V(T^*_{1m}) = k^2V(\bar{y}) + k^2\bar{R}^2_m V(\bar{x}) + (k-1)^2 \bar{X}^2V(R^*_m)-2k^2\bar{R}_m \text{Cov}(\bar{y},\bar{x})$$

$$-2k(k-1)\bar{X} \text{Cov}(\bar{y},R^*_m)+2k(k-1)\bar{X} \bar{R}_m \text{Cov}(\bar{x}, R^*_m) \quad (2.15)$$

Expression (2.15) provides the variance of T^*_{1m} in large samples for any m less than n.

Efficiency of T^*_{1m} for small m, in large samples

Assume that m is so small as compared to n, so that $k = \frac{N-m}{N} \cdot \frac{n}{n-m} \approx 1$. Then, from (2.15), we have

$$V(T^*_{1m}) = V(\bar{y}) + \bar{R}_m^2 V(\bar{x}) - 2 \bar{R}_m \text{Cov}(\bar{y},\bar{x}), \tag{2.16}$$

$$= \frac{N-n}{Nn} \cdot \frac{1}{N-1} \sum_{j=1}^{N} \left[(y_j-\bar{Y}) - \bar{R}_m(x_j-\bar{X}) \right]^2. \tag{2.17}$$

In large samples, the corresponding expression for the variance of the usual biased ratio estimator $\bar{y}_R$ is given by

$$V(\bar{y}_R) = \frac{N-n}{Nn} \cdot \frac{1}{N-1} \left[(y_j-\bar{Y}) - R(x_j-\bar{X}) \right]^2. \tag{2.18}$$

Thus (2.17) and (2.18) show that T^*_{1m} is more precise than $\bar{y}_R$, in large samples for m small compared to n, if and only if the line $\bar{Y} + \bar{R}_m(x_j-\bar{X})$ fits the values y_j more closely than the line Rx_j; in other words, if the slope of the regression line of y on x is closer to $\bar{R}_m = \frac{1}{\binom{N}{n}} \sum^{\binom{N}{n}} R^*_m = \frac{1}{\binom{N}{m}} \sum^{\binom{N}{m}} R_m$ than to the population ratio $R = \bar{Y} / \bar{X}$.

In particular, when m=1, $T^*_{11} = \bar{y}_r'$, the Hartley and Ross unbiased estimator and $\bar{R}_m = \frac{1}{N} \sum_{j=1}^{N} y_j/x_j = \bar{r}_p$, so that we arrive at the conclusion of Goodman and Hartley (1958) that $\bar{y}_r'$ is more efficient than $\bar{y}_R$, in large samples, if and only if the slope of the regression line of y on x is closer to $\bar{r}_p$ than to R.

Efficiency of T^*_{1m} for large m

To obtain the variance of the estimator T^*_{1m} for m sufficiently large, we further evaluate the terms $V(R^*_m)$, $Cov(\bar{y}, R^*_m)$ and $Cov(\bar{x}, R^*_m)$ occurring inthe large sample variance expression (2.15) of the estimator T^*_{1m}. In their evaluation we suppose that terms of order$1/mn$ or $1/n^2$ can be neglected.

We have

$$V(R_m) = E[V(R_m/n)] + V[E(R_m/n)].$$

But since $E(R_m/n) = R^*_m$,

$$V(R^*_m) = V(R_m) - E[V(R_m/n)]$$

$$\simeq \frac{N-m}{Nm} R^2(C_y^2 + C_x^2 - 2C_{xy})$$

$$- E\, \frac{n-m}{nm} R_n^2 (c_y^2 + c_x^2 - 2c_{xy})$$

$$\simeq \frac{N-n}{Nn} R^2(C_y^2 + C_x^2 - 2C_{xy}), \qquad (2.19)$$

where$R_n = \bar{y}/\bar{x}$, $c_y^2 = s_y^2/\bar{y}^2$, $c_x^2 = s_x^2/\bar{x}^2$ and $c_{xy} = s_{xy}/\bar{x}\bar{y}$.

Here s_y^2, s_x^2, and s_{xy} represent the mean sums of squares and sum of products in the sample of size n.

Proceeding on similar lines with the help of the formulae

$$Cov(\bar{y}_m, R_m) = E[Cov(\bar{y}_m, R_m/n)] + Cov[E(\bar{y}_m/n), E(R_m/n)],$$

it can be shown that

$$\text{Cov}(\bar{y}, R_m^*) \simeq \frac{N-n}{Nn} R\bar{Y}(C_y^2 - C_{xy}). \tag{2.20}$$

Again, by a similar argument, to the same order of approximation, we have

$$\text{Cov}(\bar{x}, R_m^*) \simeq \frac{N-n}{Nn} R\bar{X} (C_{xy} - C_x^2). \tag{2.21}$$

Using the results (2.19), (2.20) and (2.21) in (2.15) and observing that $\bar{R}_m \simeq R$ when m is sufficiently large, we obtain, to the order of approximation 1/n,

$$V(T_{1m}^*) = \frac{N-n}{Nn} \bar{Y}^2 (C_y^2 + C_x^2 - 2C_{xy}). \tag{2.22}$$

This shows that, in large samples to the order of approximation 1/n, when m is sufficiently large, the unbiased ratio estimator T_{1m}^* and the conventional biased ratio estimator $\bar{y}_R$ are of equal precision.

<u>$V(T_{1m})$ and $V(T_{1m}^*)$ when m is large, in a large 'Bivariate normal' population</u>

Assuming that the population is large and follows a bivariate normal distribution, the variances of the estimators T_{1m} and T_{1m}^* are obtained here to the order of approximation $1/m^2$. In this sense the results, obtained here, may be considered as improved approximations over the corresponding results of above.

It has been shown by Sukhatme (1954) that in random samples of size n from a large population, following a bivariate normal distribution, the expected value, the variance and the mean square error (M.S.E.) of the usual biased ratio estimator $\bar{y}_R$, to the order of approximation $1/n^2$, are given by

$$E(\bar{y}_R) = \bar{Y}\left[1+\frac{1}{n}(C_x^2 - C_{xy})(1+\frac{3}{n}C_x^2)\right] \quad (2.23)$$

$$V(\bar{y}_R) = \bar{Y}^2\left[(C_y^2 + C_x^2 - 2C_{xy})(\frac{1}{n}+\frac{3}{n^2}C_x^2) + \frac{6}{n^2}(C_x^2 - C_{xy})^2\right] \quad (2.24)$$

and

$$M.S.E.(\bar{y}_R) = \bar{Y}^2\left[(C_y^2 + C_x^2 - 2C_{xy})(\frac{1}{n}+\frac{3}{n^2}C_x^2) + \frac{6}{n^2}(C_x^2 - C_{xy})^2\right] \quad (2.25)$$

For the finite population of size N, the effect will be approximately to write N-n/Nn for n in the above expressions.

<u>Variance of T_{1m}</u>

From (2.3), we have

$$V(T_{1m}) = \frac{1}{n-m} E\left[s^2_{N-m,y} + R_m^2 s^2_{N-m,x} - 2R_m s_{N-m,xy}\right] \quad (2.26)$$

when the finite population correction factor is assumed to be negligible.

Now since in samples from a bivariate normal population the sample means are independently distributed of the sample variances and covariances, from (2.26) we obtain

$$V(T_{1m}) = \frac{1}{n-m}\left[S_y^2 + E(R_m^2)S_x^2 - 2E(R_m)S_{xy}\right] \quad (2.27)$$

Also, based on the results (2.23) and (2.24), to the order of approximation $1/m^2$, we have

$$E(R_m) = R\left[1+\frac{1}{m}(C_x^2-C_{xy})(1+\frac{3}{m}C_x^2)\right] \quad (2.28)$$

and

$$E(R_m^2) = V(R_m) + \left[E(R_m)\right]^2$$

$$= R^2\left[1+\frac{2}{m}(C_x^2-C_{xy})+\frac{6}{m^2}(C_x^2-C_{xy})^2+\right.$$

$$\left.\frac{6}{m^2}(C_x^2-C_{xy})C_x^2+(\frac{1}{m}+\frac{3}{m^2}C_x^2)(C_y^2+C_x^2-2C_{xy})\right]. \quad (2.29)$$

Substituting (2.28) and (2.29) in (2.27), we have, after a bit of simplification,

$$V(T_{1m}) = \frac{\bar{Y}^2}{n-m}\left[(C_y^2+C_x^2-2C_{xy})(1+\frac{C_x^2}{m}+\frac{3C_x^4}{m^2})\right.$$

$$\left.+2(C_x^2-C_{xy})^2(\frac{1}{m}+\frac{6}{m^2}C_x^2)\right] \quad (2.30)$$

Expression (2.30) provides the variance of the unbiased ratio estimator T_{1m}, to the order of approximation $1/m^2$, in samples from a large bivariate normal population.

Variance of T_{1m}^*

Since $V(T_{1m}) = E\left[V(T_{1m}/n)\right] + V(T_{1m}^*)$,

we have $V(T_{1m}^*) = V(T_{1m}) - E\left[(\bar{X}-k\bar{x})^2 V(R_m/n)\right]$, (2.31)

where $k = \dfrac{(N-m)n}{N(n-m)}$.

Consequently to evaluate $(V(T_{1m}^*)$ to the order of approximation $1/m^2$, we need to evaluate:

$$E\left[(\bar{X}-k\bar{x})^2 V(R_m/n)\right] \quad (2.32)$$

to the order $1/m^2$ and use it in the formula (2.31) along with the result (2.30) for $V(T_{1m})$.

Now taking into account the finite population correction and applying the formula (2.24) for $V(R_m/n)$, we have

$$E\left[(\bar{X}-k\bar{x})^2\ V(R_m/n)\right] = \frac{n-m}{nm}\left[(\bar{X}^2-2k\bar{x}\bar{X}+k^2\bar{x}^2)R_n^2\left\{(c_y^2+c_x^2-2c_{xy})(1+3\frac{n-m}{nm}c_x^2)\right.\right.$$
$$\left.\left.+6\frac{n-m}{nm}(c_x^2-c_{xy})^2\right\}\right], \qquad (2.33)$$

where R_n, c_y^2, c_x^2, and c_{xy} refer to the total sample of size n.

Since we are interested in evaluating expression (2.32) to the order $1/m^2$ only, the expectations of terms with coefficient $(\frac{n-m}{nm})^2$ in the above equation can be replaced by the corresponding population terms. Thus the contribution of terms with coefficient $(\frac{n-m}{nm})^2$ to (2.32) is given by

$$G = \bar{Y}^2(\frac{n-m}{nm})^2(k-1)^2\left[3C_x^2(C_y{}^2+C_x{}^2-2C_{xy})+6(C_x^2-C_{xy})^2\right]. (2.33)$$

Now to obtain the contribution of the other term

$$\frac{n-m}{nm}E\left[(\bar{X}^2-2k\bar{x}\bar{X}+k^2\bar{x}^2)R_n^2\ (c_y^2+c_x^2-2c_{xy})\right] \qquad (2.34)$$

of the R.H.S. of (2.33), the expected value is evaluated here to the order $1/n$. Since we have assumed a large bivariate normal population, in this evaluation we take $N-n/Nn \doteq 1/n$ and observe that the sample means $\bar{y}$ and $\bar{x}$ are distributed independently of the sample mean squares s_y^2 and s_x^2 and the sample mean sum of products s_{xy}.

Thus term by term, to the order 1/n we have

$$E(R_n^2\, c_y^2) = R^2C_y^2\left[1+ \frac{3}{n} C_x^2 \right]$$

$$E(R_n^2\, c_x^2) = R^2C_x^2\left[1+ \frac{1}{n}(C_y^2 + 10C_x^2 - 8C_{xy})\right]$$

$$E(R_n^2\, c_{xy}) = R^2C_{xy}\left[1+ \frac{1}{n} (6C_x^2 -3C_{xy})\right]$$

$$E(\bar{x}\, R_n^2 c_y^2) = \bar{X}\, R^2\, C_y^2 \left[1+ \frac{1}{n} C_x^2 \right]$$

$$E(\bar{x}\, R_n^2 c_x^2) = \bar{X}\, R^2\, C_x^2 \left[1+ \frac{1}{n} (C_y^2 +6C_x^2 -6C_{xy})\right]$$

$$E(\bar{x}\, R_n^2 c_{xy}) = \bar{X}\, R^2\, C_{xy} \left[1+ \frac{1}{n}(3C_x^2 -2C_{xy})\right]$$

$$E(\bar{x}^2 R_n^2 c_y^2) = S_y^2$$

$$E(\bar{x}^2 R_n^2 c_x^2) = \bar{X}^2 R^2 C_x^2 \left[1+ \frac{1}{n}(C_y^2+3C_x^2-4C_{xy})\right]$$

and finally,

$$E(\bar{x}^2 R_n^2\, c_{xy}) = \bar{X}^2 R^2 C_{xy}\left[1+ \frac{1}{n} (C_x^2 - C_{xy})\right].$$

Making use of these results in (2.34) and simplifying, we obtain the contribution H of the terms with coefficient $\left(\frac{n-m}{nm}\right)$ as given by

$$H = \bar{Y}^2 \frac{n-m}{nm} \left[(C_y^2+C_x^2-2C_{xy}) \left\{(k-1)^2+ \frac{(k-2)^2 C_x^2}{n}\right\} + \frac{2(k-1)(k-3)}{n} (C_x^2 - C_{xy})^2 \right] \tag{2.35}$$

Thus to the order of approximation $1/m^2$, we have

$$\text{Expression } (2.32) \quad = \quad G + H \tag{2.36}$$

Now we observe $k = \frac{N-m}{N} \cdot \frac{n}{n-m} \;\simeq\; \frac{n}{n-m}$ in a large population and use the results (2.30) and (2.36) in the formula (2.31) to obtain

$$V(T^{*}_{1m}) \quad = \quad \bar{Y}^2 \Big[(C_y^2 + C_x^2 - 2C_{xy}) \Big(\frac{1}{n} + \frac{C_x^2}{n^2} + \frac{3C_x^4}{m^2(n-m)} \Big) + \Big(\frac{3}{n^2} + \frac{2}{mn} + \frac{12C_x^2}{m^2(n-m)} \Big)(C_x^2 - C_{xy})^2 \Big]. \tag{2.37}$$

Expression (2.37) provides the variance of T^{*}_{1m}, to the order of approximation $1/m^2$, in samples from a large bivariate normal population.

Efficiency of T^{*}_{1m}

We now compare the variance of T^{*}_{1m} given by (2.37) with the M.S.E. of $\bar{y}_R$ given by (2.25).

Thus

$$\text{M.S.E.}(\bar{y}_R) - V(T^{*}_{1m}) \quad = \quad \bar{Y}^2 \Big[(C_y^2 + C_x^2 - 2C_{xy}) \Big(\frac{2}{n^2} - \frac{3C_x^2}{m^2(n-m)} \Big) C_x^2 + \Big(\frac{5}{n^2} - \frac{2}{mn} - \frac{12C_x^2}{m^2(n-m)} \Big)(C_x^2 - C_{xy})^2 \Big]. \tag{2.38}$$

From this it follows that T^{*}_{1m} is more efficient than $\bar{y}_R$ if

$$C_x^2 < \text{Min.} \left[\frac{2m^2(n-m)}{3n^2}, \quad \frac{m(n-m)(5m-2n)}{12n^2} \right],$$

$$= \frac{m(n-m)\ (5m-2n)}{12n^2} \,. \qquad (2.39)$$

In particular for the choices $m = n/2$, $m = 3n/4$ and $m=n-1$, the conditions are obtained below. Numerically for $n=100$ the upper bounds for C_x^2 are also given.

m	Upper bound for C_x^2		For n = 100
$n/2$	$n/96$	=	1.04
$3n/4$	$7n/256$	=	2.74
$n-1$	$(n-1)(3n-5)/12n^2$	=	0.24

In fact, for a large sample and a choice of 'm' near to the total sample size we have approximately

$$\frac{m(n-m)(5m-2n)}{12n^2} = 0.25.$$

Thus in large samples, for a choice of m sufficiently near to the total sample size, T_{1m}^* will be more efficient than $\bar{y}_R$ if $C_x^2 < 0.25$.

3. PRECISION OF REGRESSION TYPE ESTIMATORS WITH ONE AUXILIARY VARIABLE

Mickey's unbiased regression type estimators, utilizing information on only one auxiliary variable, are T_{2m} and T^*_{2m}, given by (1.8) and (1.9) of Section 1. For each m, T^*_{2m} is never less efficient than T_{2m}, applied to any particular permutation of the sample elements. Among the unbiased regression type estimators T^*_{2m}, computationally the choice m=n-1 yields the most feasible estimator $T^*_{2(n-1)}$; for it is possible to express $T^*_{2(n-1)}$ in an alternative form:

$$T^*_{2(n-1)} = \left[\bar{y} - \bar{b}'(\bar{x} - \bar{X}) \right] - \frac{N-n}{Nn} \left[\sum_j^n x_j b_j' - n\bar{x}\bar{b}' \right], \quad (3.1)$$

where b_j' is the value of the regression coefficient if the j^{th} sample element is omitted

$$\text{(i.e.,)} \quad b_j' = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y}) - \frac{n}{n-1}(x_j - \bar{x})(y_j - \bar{y})}{\sum_i^n (x_i - \bar{x})^2 - \frac{n}{n-1}(x_j - \bar{x})^2}, \quad (3.2)$$

and $\bar{b}' = \frac{1}{n} \sum_j^n b_j'$.

The present investigation of the efficiency, confined to this most important practical case, shows that, to the order of approximation $1/n$, $T^*_{2(n-1)}$ is as efficient as the usual biased regression estimator $\bar{y}_{1r}$ given by

$$\bar{y}_{1r} = \bar{y} - b(\bar{x} - \bar{X}),$$

where b is the sample regression coefficient. The same result, though not formally established here, is expected to hold good for sufficiently large m, since a corresponding result has been obtained in Section 2 in the case of Mickey's unbiased ratio type estimators T^{*}_{1m} and the usual biased ratio estimator $\bar{y}_R$. It is interesting that this result leads to the conclusion that in large samples, for sufficiently large m, Mickey's unbiased regression type estimators T^{*}_{2m} are never less efficient than the unbiased ratio type estimators T^{*}_{1m}, since it is known that in large samplex the usual regression estimator $\bar{y}_{1r}$ is never less efficient than the usual ratio estimator $\bar{y}_R$.

We now formally establish that $V(T^{*}_{2(n-1)}) = V(\bar{y}_{lr})$, to the order of approximation 1/n.

Variance of $T^{*}_{2(n-1)}$

Neglecting the finite population correction factor, we have

$$T^{*}_{2(n-1)} = T' - m_{11},$$

where $T' = \bar{y} - \bar{b}'(\bar{x} - \bar{X})$ (3.3)

and $m_{11} = \frac{1}{n}\sum_{j}^{n}(x_j - \bar{x})b'_j$, the sample covariance between x_j and b'_j. (3.4)

Hence $$V(T^{*}_{2(n-1)}) = V(T')+V(m_{11}) - 2\,\mathrm{Cov}(T', m_{11}). \quad (3.5)$$

In the following we shall evaluate $V(T')$ to the order $1/n$, and show that $V(m_{11})$ and $Cov(T', m_{11})$ are at least of order $1/n^2$.

To evaluate $V(T')$, we write

$$T' = T-(\bar{b}'- B)(\bar{x} - \bar{X}), \tag{3.6}$$

where $$T = \bar{y} - B(\bar{x} - \bar{X}), \tag{3.7}$$

and B is the population regression coefficient of y on x. Then, since T is unbiased for the population mean $\bar{Y}$, we have

$$V(T') = E(T-\bar{Y})^2 + E[(\bar{b}'-B)(\bar{x}-\bar{X})]^2$$
$$-[E(\bar{b}'-B)(\bar{x}-\bar{X})]^2 - 2E[(T-\bar{Y})(\bar{b}'-B)(\bar{x}-\bar{X})]. \tag{3.8}$$

Now we show that all the terms except the first one on the R.H.S. of equation (3.8) are of order $1/n^2$.

For this, we note that

$$E(T-\bar{Y})^2 = O(n^{-1}),\ E(T - \bar{Y})^4 = O(n^{-2}),$$

$$E(\bar{x} -\bar{X})^2 = O(n^{-1}),\ E(\bar{x} - \bar{X})^4 = O(n^{-2}) \tag{3.9}$$

and $$E(\bar{b}'-B)^2 = O(n^{-1}),\ E(\bar{b}'- B)^4 = O(n^{-2}), \tag{3.10}$$

where $O(n^{-1})$ and $O(n^{-2})$ indicate that the terms are of order $1/n$ and $1/n^2$ respectively.

Results (3.9) follow from the fact that T and $\bar{x}$ are arithmetic means, based on a simple random sample of size n, without replacement. Also, since

$$E(b_j' - B)^2 = E(b_j' - \bar{b}')^2 + E(\bar{b}' - B)^2,$$

We have

$$E(\bar{b}' - B)^2 \leq E(b_j' - B)^2$$
$$= O(n^{-1}),$$

b_j' being the regression coefficient based on a simple random sample of size n-1.

Thus $E(\bar{b}' - B)^2 = O(n^{-1})$.

Similarly it can be shown that

$$E(\bar{b}' - B)^4 = O(n^{-2}).$$

To show that the three terms except the first one on the R.H.S. of equation (3.8) are of order $1/n^2$, we repeatedly make use of the inequality

$$E(uv) \leq [E(u^2)E(v^2)]^{\frac{1}{2}} \qquad (3.11)$$

where u and v are any two random variables having finite second moments.

Thus $E[(\bar{b}'-B)(\bar{x}-\bar{X})]^2 \leq [E(\bar{b}'-B)^4 E(\bar{x}-\bar{X})^4]^{\frac{1}{2}}$

$= [O(n^{-2})\ O(n^{-2})]^{\frac{1}{2}}$ from (3.9) and (3.10).

Hence $E[(\bar{b}'-B)(\bar{x}-\bar{X})]^2 = O(n^{-2}) \qquad (3.12)$

Again, since

$$E[(\bar{b}'-B)(\bar{x}-\bar{X})] \leq [E(\bar{b}'-B)^2 E(\bar{x}-\bar{X})^2]^{\frac{1}{2}}$$

$= [O(n^{-1})\ O(n^{-1})]^{\frac{1}{2}}$ from (3.9) and (3.10),

we obtain

$$[E(\bar{b}'-B)(\bar{x}-\bar{X})]^2 = O(n^{-2}). \qquad (3.13)$$

Also $E\left[(T-\bar{Y})(\bar{b}'-B)(\bar{x}-\bar{X})\right] \leq \left[E(\bar{b}'-B)^2\, E\left((T-\bar{Y})(\bar{x}-\bar{X})\right)^2\right]^{\frac{1}{2}}$

$$\leq \left[E(\bar{b}'-B)^2\right]^{\frac{1}{2}}\left[E(T-\bar{Y})^4 E(\bar{x}-\bar{X})^4\right]^{\frac{1}{4}}$$

$$= \left[O(n^{-1})\right]^{\frac{1}{2}}\left[O(n^{-4})\right]^{\frac{1}{4}} \text{ from (3.9) and (3.10),}$$

$$= O(n^{-3/2}).$$

But expectations of products and ratios of arithmetic means must have integer orders. So it follows that

$$E\left[(T-\bar{Y})(\bar{b}'-B)\ (\bar{x}-\bar{X})\right] = O(n^{-2}). \qquad (3.14)$$

Consequently from (3.8), (3.12), (3.13) and (3.14) we obtain

$$V(T') = E(T-\bar{Y})^2 + O(n^{-2}). \qquad (3.15)$$

Returning to equation (3.5), it remains to evaluate $V(m_{11})$ and $Cov(T', m_{11})$.

Using the large sample theory, it can be shown that

$$V(m_{11}) = \frac{1}{n}\left(\mu_{22} - \mu_{11}^2\right), \qquad (3.16)$$

where μ_{22} and μ_{11} are the parent central moments of the joint distribution of x_j and b_j'.

Now if B' denotes the population value corresponding to b_j', we have

$$E(b_j' - B')^2 = O(n^{-1}) \text{ and } E(b_j' - B')^4 = O(n^{-2}). \qquad (3.17)$$

Hence $\mu_{11} = E\left[(b_j'-B')(x_j-\bar{X})\right]$

$$\leq \left[E(b_j'-B')^2\, E(x_j-\bar{X})^2\right]^{\frac{1}{2}}$$

$$= \left[O(n^{-1})\right]^{\frac{1}{2}} \text{ from (3.17).}$$

Thus $\mu_{11}^2 = O(n^{-1})$ (3.18)

Again

$$\mu_{22} = E\left[(b_j' - B')(x_j - \bar{X})\right]^2 \leq \left[E(b_j' - B')^4 E(x_j - \bar{X})^4\right]^{\frac{1}{2}}$$

$$= \left[O(n^{-2})\right]^{\frac{1}{2}} \text{ from (3.17).}$$

Thus $\mu_{22} = O(n^{-1})$ (3.19)

Consequently from (3.16), (3.18) and (3.19) we have

$$V(m_{11}) = O(n^{-2}). \qquad (3.20)$$

Also from (3.15) and (3.20) by an application of the formula (3.11), it is easy to see that

$$\text{Cov}(T', m_{11}) \leq O(n^{-3/2}).$$

But since m_{11} and T' are built up as arithmetic means of products and ratios of arithmetic means, $\text{Cov}(T', m_{11})$ must have an integer order. Thus

$$\text{Cov}(T', m_{11}) = O(n^{-2}). \qquad (3.21)$$

Now from (3.5), (3.15), (3.20) and (3.21) it follows that

$$V(T^*_{2(n-1)}) = E(T - \bar{Y})^2 + O(n^{-2}),$$

$$= V(\bar{y}) + B^2 V(\bar{x}) - 2B \, \text{Cov}(\bar{y}, \bar{x}) + O(n^{-2}).$$

(3.22)

From this it is concluded that, to the order of approximation $1/n$, $T^*_{2(n-1)}$ and the usual biased regression estimator $\bar{y}_{1r}$ are of equal precision.

4. EFFICIENCY OF RATIO TYPE ESTIMATORS WITH TWO AUXILIARY VARIABLES

When information on two auxiliary variables x_1 and x_2 is available, for the choice

$$a_1(z_m) = \frac{\bar{y}_m}{\bar{x}_{1m}} = R_m(x_1) \text{ and } a_2(z_m) = \frac{\bar{y}_m}{\bar{x}_{2m}} = R_m(x_2),$$

the unbiased estimator T_m expressed in the form (1.5) provides an unbiased ratio-type estimator:

$$T_{1m}(x_1,x_2) = T_{1m}(x_1) + T_{1m}(x_2) + \frac{(N-n)m}{N(n-m)}\bar{y}_m - \frac{(N-m)n}{N(n-m)}\bar{y}, \tag{4.1}$$

where $T_{1m}(x_1) = R_m(x_1)\bar{X}_1 + \frac{(N-m)n}{N(n-m)}\left[\bar{y} - R_m(x_1)\bar{x}_1\right]$ (4.2)

and $T_{1m}(x_2) = R_m(x_2)\bar{X}_2 + \frac{(N-m)n}{N(n-m)}\left[\bar{y} - R_m(x_2)\bar{x}_2\right]$ (4.3)

are the unbiased ratio type estimators obtained by using information on x_1 and x_2 separately.

Averaging $T_{1m}(x_1,x_2)$ over all permutations of the sample elements, we have the other unbiased ratio type estimator

$$T^*_{1m}(x_1,x_2) = T^*_{1m}(x_1) + T^*_{1m}(x_2) - \bar{y} \tag{4.4}$$

where $T^*_{1m}(x_1) = R^*_m(x_1)\bar{X}_1 + \frac{(N-m)n}{N(n-m)}\left[\bar{y} - R^*_m(x_1)\bar{x}_1\right]$ (4.5)

and $T^*_{1m}(x_2) = R^*_m(x_2)\bar{X}_2 + \frac{(N-m)n}{N(n-m)}\left[\bar{y} - R^*_m(x_2)\bar{x}_2\right].$ (4.6)

For any m less than n, $T^*_{1m}(x_1,x_2)$ is never less efficient than $T_{1m}(x_1,x_2)$. In this section we shall discuss the relative efficiency of the estimator $T^*_{1m}(x_1,x_2)$ with respect to Olkin's weighted ratio estimator in large samples for the two cases:

(i) m is small as compared to n and (ii) m is sufficiently large. Also we shall investigate, when m is sufficiently large, whether there is an increase in precision by using $T^*_{1m}(x_1,x_2)$ instead of $T^*_{1m}(x_1)$ or $T^*_{1m}(x_2)$.

Olkin's weighted Ratio estimator

When information on two auxiliary variables x_1 and x_2 is available, the weighted ratio estimator, suggested by Olkin (1958), is given by

$$\bar{y}_w = W_1\bar{y}_{R_1} + W_2\,\bar{y}_{R_2}, \tag{4,7}$$

where
$$\bar{y}_{R_1} = \frac{\bar{y}}{\bar{x}_1}\bar{X}_1, \quad \bar{y}_{R_2} = \frac{\bar{y}}{\bar{x}_2}\bar{X}_2,$$

$$W_1 = \frac{V(\bar{y}_{R_2}) - Cov(\bar{y}_{R_1},\bar{y}_{R_2})}{V(\bar{y}_{R_1} - \bar{y}_{R_2})} \tag{4.8}$$

and
$$W_2 = \frac{V(\bar{y}_{R_1}) - Cov(\bar{y}_{R_1},\bar{y}_{R_2})}{V(\bar{y}_{R_1} - \bar{y}_{R_2})}. \tag{4.9}$$

The estimator $\bar{y}_w$ is biased but consistent. The large sample variance of $\bar{y}_w$ is given by

$$V(\bar{y}_w) = V(\bar{y} - W_1R_1\bar{x}_1 - W_2R_2\bar{x}_2), \tag{4.10}$$

where $R_1 = \bar{Y} / \bar{X}_1$ and $R_2 = \bar{Y} / \bar{X}_2$.

Variance of $T^*_{1m}(x_1,x_2)$ in large samples

From (4.4), we have

$$V[T^*_{1m}(x_1,x_2)] = V[T^*_{1m}(x_1)] + V[T^*_{1m}(x_2)] + 2Cov[T^*_{1m}(x_1), T^*_{1m}(x_2)] + V(\bar{y}) - 2Cov[\bar{y}, T^*_{1m}(x_1)] - 2Cov[\bar{y}, T^*_{1m}(x_2)]. \quad (4.11)$$

Now from result (2.15) of Section 2, to the order of approximation $1/n$,

$$V[T^*_{1m}(x_i)] = k^2V(\bar{y}) + k^2\bar{R}^2_m(x_i)V(\bar{x}_i) + (k-1)^2\bar{X}^2_i V[R^*_m(x_i)] - 2k^2\bar{R}_m(x_i)Cov(\bar{y},\bar{x}_i) - 2k(k-1)\bar{X}_i Cov[\bar{y}, R^*_m(x_i)] + 2k(k-1)\bar{X}_i\bar{R}_m(x_i)Cov[\bar{x}_i, R^*_m(x_i)], \; i=1,2, \quad (4.12)$$

where $k = \frac{(N-m)n}{N(n-m)}$.

Proceeding on similar lines as in the derivation of $V[T^*_{1m}(x_i)]$, it can be seen that,

$$Cov[T^*_{1m}(x_1), T^*_{1m}(x_2)] = k^2V(\bar{y}) + k^2\bar{R}_m(x_1)\bar{R}_m(x_2)Cov(\bar{x}_1,\bar{x}_2) + (k-1)^2\bar{X}_1\bar{X}_2 Cov[R^*_m(x_1), R^*_m(x_2)] - k^2\bar{R}_m(x_1)Cov(\bar{y},\bar{x}_1) - k^2\bar{R}_m(x_2)Cov(\bar{y},\bar{x}_2) - k(k-1)\bar{X}_1 Cov[\bar{y}, R^*_m(x_1)] - k(k-1)\bar{X}_2 Cov[\bar{y}, R^*_m(x_2)]$$

$$+k(k-1)\bar{X}_2\bar{R}_m(x_1)\mathrm{Cov}[\bar{x}_1,R^*_m(x_2)]$$

$$+k(k-1)\bar{X}_1\bar{R}_m(x_2)\mathrm{Cov}[\bar{x}_2,R^*_m(x_1)] \qquad (4.13)$$

Similarly, to the order of approximation $1/n$, for $i = 1,2$, we have

$$\mathrm{Cov}[\bar{y},T^*_{1m}(x_i)] = kV(\bar{y})-k\bar{R}_m(x_i)\mathrm{Cov}(\bar{y},\bar{x}_i)-(k-1)\bar{X}_i\mathrm{Cov}[\bar{y},R^*_m(x_i)] \qquad (4.14)$$

Substituting in (4.11) for the various terms from (4.12), (4.13) and (4.14) and simplifying we have

$$\begin{aligned} V[T^*_{1m}(x_1,x_2)] = {} & (2k-1)^2V(\bar{y})+k^2V[\bar{x}_1\bar{R}_m(x_1)+\bar{x}_2\bar{R}_m(x_2)] \\ & +(k-1)^2V[\bar{X}_1R^*_m(x_1)+\bar{X}_2R^*_m(x_2)] \\ & -2k(2k-1)[\bar{R}_m(x_1)\mathrm{Cov}(\bar{y},\bar{x}_1)+\bar{R}_m(x_2)\mathrm{Cov}(\bar{y},\bar{x}_2)] \\ & -2(k-1)(2k-1)\Big\{\bar{X}_1\mathrm{Cov}[\bar{y},R^*_m(x_1)]+ \\ & \bar{X}_2\mathrm{Cov}[\bar{y},R^*_m(x_2)]\Big\} \\ & +2k(k-1)\Big\{\bar{X}_1\bar{R}_m(x_1)\mathrm{Cov}[\bar{x}_1,R^*_m(x_1)] \\ & +\bar{X}_2\bar{R}_m(x_2)\mathrm{Cov}[\bar{x}_2,R^*_m(x_2)] \\ & +\bar{X}_1\bar{R}_m(x_2)\mathrm{Cov}[\bar{x}_2,R^*_m(x_1)] \\ & +\bar{X}_2\bar{R}_m(x_1)\mathrm{Cov}[\bar{x}_1,R^*_m(x_2)]\Big\}. \end{aligned} \qquad (4.15)$$

Expression (4.15) provides the variance of $T^*_{1m}(x_1,x_2)$, to the order of approximation $1/n$, for any m less than n.

__Efficiency of $T^*_{1m}(x_1,x_2)$ for small m, in large samples__

When m is small as compared to n, $k = \frac{N-m}{N}\cdot\frac{n}{n-m} \doteq 1$ and consequently the variance expression (4.15) reduces to

$$V\left[T^{*}_{1m}(x_1,x_2) \right] = V\left[\bar{y} - \bar{R}_m(x_1)\bar{x}_1 - \bar{R}_m(x_2)\bar{x}_2 \right]. \quad (4.16)$$

A comparison of (4.10) and (4.16) shows that, when m is small as compared to n, $T^{*}_{1m}(x_1,x_2)$ is more efficient than Olkin's biased estimator $\bar{y}_w$, if and only if the plane

$\bar{Y} + \bar{R}_m(x_1) \left[x_{1j} - \bar{X}_1 \right] + \bar{R}_m(x_2) \left[x_{2j} - \bar{X}_2 \right]$ fits the values y_j more closely than the plane

$$\bar{Y} + W_1R_1 \left[x_{1j} - \bar{X}_1 \right] + W_2R_2 \left[x_{2j} - \bar{X}_2 \right] = W_1R_1x_{1j} + W_2R_2x_{2j},$$

where W_1 and W_2 are the optimum weights given by (4.8) and (4.9).

Efficiency of $T^{*}_{1m}(x_1,x_2)$ for large m

From result (2.19) of section 2, when m is large, to the order of approximation 1/n,

$$V\left[R^{*}_m(x_i) \right] = \frac{1}{\bar{X}_i^2} \left[V(\bar{y}) + R_i^2 V(\bar{x}_i) - 2R_i \operatorname{Cov}(\bar{y},\bar{x}_i) \right], \quad i=1,2. \quad (4.17)$$

Similarly

$$\operatorname{Cov}\left[R^{*}_m(x_1), R^{*}_m(x_2) \right] = \frac{1}{\bar{X}_1\bar{X}_2} \left[V(\bar{y}) + R_1R_2\operatorname{Cov}(\bar{x}_1,\bar{x}_2) - R_1\operatorname{Cov}(\bar{y},\bar{x}_1) - R_2\operatorname{Cov}(\bar{y},\bar{x}_2) \right]. \quad (4.18)$$

Again from result (2.20) of section 2, we have

$$\operatorname{Cov}\left[\bar{y}, R^{*}_m(x_i) \right] = \frac{1}{\bar{X}_i} \left[V(\bar{y}) - R_i \operatorname{Cov}(\bar{x}_i, \bar{y}) \right], \quad i=1,2. \quad (4.19)$$

From result (2.21), we have

$$\text{Cov}[\bar{x}_i, R_m^*(x_i)] = \frac{1}{\bar{Y}}[R_i \text{Cov}(\bar{x}_i,\bar{y}) - R_i^2 V(\bar{x}_i)], i=1,2. \quad (4.20)$$

Similarly, it can be seen that

$$\text{Cov}[\bar{x}_i, R_m^*(x_j)] = \frac{1}{\bar{Y}}[R_j \text{Cov}(\bar{x}_i,\bar{y}) - R_j^2 \text{Cov}(\bar{x}_i,\bar{x}_j)] \text{ for } i \neq j. \quad (4.21)$$

Making use of the results (4.17) to (4.21) in (4.15) and observing that in this case $\bar{R}_m(x_i) \simeq R_i$, we obtain on simplification,

$$V[T_{1m}^*(x_1,x_2)] = V[\bar{y} - R_1\bar{x}_1 - R_2\bar{x}_2]. \quad (4.22)$$

Now from (4.10) and (4.22), we have

$$\begin{aligned} V[\bar{y}_w] - V[T_{1m}^*(x_1,x_2)] = {} & W_1[2R_2\text{Cov}(\bar{y},\bar{x}_2) - R_2^2 V(\bar{x}_2)] \\ & + W_2[2R_1\text{Cov}(\bar{y},\bar{x}_1) - R_1^2 V(\bar{x}_2)] \\ & + W_1W_2[2R_1R_2\text{Cov}(\bar{x}_1,\bar{x}_2) - R_1^2 V(\bar{x}_1) - R_2^2 V(\bar{x}_2)] \\ & - 2R_1R_2\text{Cov}(\bar{x}_1,\bar{x}_2). \end{aligned}$$

From this it follows that $T_{1m}^*(x_1,x_2)$ is more efficient, equally efficient or less efficient than Olkin's estimator according as

$$2[W_1C_{yx_2} + W_2C_{yx_1} + (W_1W_2-1)C_{x_1x_2}] \gtreqless W_1(1+W_2)C_{x_1}^2 + W_2(1+W_1)C_{x_2}^2 \quad (4.23)$$

where $C_{x_1}^2 = s_{x_1}^2 / \bar{X}_1^2$, etc.

Inequality (4.23) is difficult to interpret. However, if we assume

$$C_{x_1}^2 = C_{x_2}^2 = C_x^2 \text{ and } C_{yx_1} = C_{yx_2} = C_{yx},$$

then $W_1 = W_2 = 1/2$ and (4.23) reduces to the inequality

$$\rho_{xy} \gtrless \frac{1}{4} \frac{C_x}{C_y} (1+\rho_{12}) ,$$

where ρ_{xy} is the correlation between an auxiliary variable and the principle variable y and ρ_{12} is the correlation between x_1 and x_2.

Thus for sufficiently large m, under the conditions

$$C^2_{x_1} = C^2_{x_2} = C^2_x , \qquad C_{yx_1} = C_{yx_2} = C_{yx} ,$$

$T^*_{1m}(x_1,x_2)$ is more precise or equally precise or less precise than Olkin's estimator according as

$$\rho_{xy} \gtrless \frac{1}{4} \frac{C_x}{C_y} (1+\rho_{12}). \tag{4.24}$$

<u>Effect on the precision by introducing a new auxiliary variable</u>

Since the weights W_1 and W_2, used in Olkin's estimator, are determined by minimising the variance of $W_1\bar{y}_{R_1} + W_2\bar{y}_{R_2}$ with respect to W_1 and W_2, subject to the condition $W_1+W_2 = 1$; for any combination of the weights other than the optimum combination the variance of $W_1\bar{y}_{R_1} + W_2\bar{y}_{R_2}$ is more than the variance of Olkin's estimator. Thus, in particular since $W_1 = 1$, $W_2 = 0$, gives $\bar{y}_{R_1}$, the introduction of a new auxiliary variable always results in obtaining a more precise ratio estimator.

But in Mickey's unbiased ratio type estimators, $T^*_{1m}(x_1,x_2)$ is not constructed as the optimum weighted average of

$T^*_{1m}(x_1)$ and $T^*_{1m}(x_2)$, but is formed simply as

$$T^*_{1m}(x_1,x_2) \quad = \quad T^*_{1m}(x_1) + T^*_{1m}(x_2) \quad - \bar{y}.$$

As such, we cannot say without any reservation that $T^*_{1m}(x_1,x_2)$ always provides a more efficient estimator than $T^*_{1m}(x_1)$ or $T^*_{1m}(x_2)$.

In fact, when m is sufficiently large, we have

$$V\left[T^*_{1m}(x_1)\right] \quad - \quad V\left[T^*_{1m}(x_1,x_2)\right] \quad \gtreqless \quad 0,$$

according as

$$2C_{yx_2} \quad - \quad C^2_{x_2} \quad - \quad 2C_{x_1x_2} \gtreqless \quad 0.$$

Thus, in particular, when $C^2_{x1} = C^2_{x2} = C^2_x$ and $C_{x_1y} = C_{x_2y} = C_{xy}$, $T^*_{1m}(x_1,x_2)$ is more precise or equally precise or less precise than $T^*_{1m}(x_1)$ according as

$$\rho_{xy} \gtreqless \frac{1}{2}\frac{C_x}{C_y}(1 + 2\rho_{12}). \qquad (4.28)$$

This result shows the need for caution in introducing a new auxiliary variable in the case of Mickey's unbiased ratio type estimators.

From (4.24) and (4.28) it follows that, when m is sufficiently large, and $C^2_{x1} = C^2_{x2} = C^2_x$, $C_{x_1y} = C_{x2y} = C_{xy}$, $T^*_{1m}(x_1,x_2)$ is more precise than Olkin's estimators as well as $T^*_{1m}(x_1)$, if

$$\rho_{xy} > \text{Max.}\left[\frac{3}{4}\frac{C_x}{C_y}(1+\rho_{12}),\ \frac{1}{2}\frac{C_x}{C_y}(1+2\rho_{12})\right]$$
$$= \frac{3}{4}\frac{C_x}{C_y}(1+\rho_{12}). \qquad (4.30)$$

In the following, a table of values of the function $\frac{3}{4}\ \frac{C_x}{C_y}(1+\rho_{12})$ for different values of $(\frac{C_x}{C_y})$ and ρ_{12} is given to see how much the correlation ρ_{xy} should be in order to make an efficient use of the estimator $T^*_{1m}(x_1,x_2)$.

$\frac{C_x}{C_y}$ \ ρ_{12}	0.10	0.20	0.30	0.40	0.50	0.75
0.25	0.206	0.225	0.244	0.263	0.281	0.328
0.50	0.413	0.450	0.488	0.525	0.563	0.656
0.75	0.619	0.675	0.731	0.788	0.844	0.984
1.00	0.825	0.900	0.975	N.A.*	N.A.	N.A.

* N.A.- denotes that the value is not admissible, being greater than 1.

5. EXTENSION TO DOUBLE SAMPLING

In constructing the general class of unbiased estimators, Mickey has assumed that the population means of all the auxiliary variables are known. When, however, the population means of the auxiliary variables are not known in advance, using the technique of 'double sampling', we shall develop in this section a general class of unbiased estimators of which 'unbiased ratio and regression type' estimators are special cases. This section also gives unbiased estimators of the variance of the proposed estimators and a discussion concerning the efficiency of unbiased ratio and regression type estimators.

Preliminaries

Consider a finite population of size N, represented by the set of (p+1) vectors:

$$(y_j, x_{1j}, x_{2j}, \ldots\ldots, x_{pj}) \quad j = 1,2, \ldots, N.$$

Let a simple random sample of size n' be drawn without replacement from the population and observations be made on all the auxiliary characteristics. Let $\bar{x}_i'$, i=1,2,....,p, represent the means of the auxiliary variables, based on the sample of size n'. Now let a sub-sample of size n be drawn with equal probabilities without replacement from the sample of size n' to observe the variable y under study. Further, let $\bar{y}$ and $\bar{x}_i$ i=1,2,....,p, denote the means based on the sub-sample.

For any choice of m of the sub-sample elements (z_m), suppose $\bar{y}_m$ and $\bar{x}_{im}$ $(i=1,2,\ldots,p)$ are the means based on z_m. Let $a_i(z_m)$ be some known real valued functions of z_m. Further, define

$$\bar{y}_{n-m} = \frac{n\bar{y}-m\bar{y}_m}{n-m},\quad \bar{x}_{i\,n-m} = \frac{n\bar{x}_i - m\bar{x}_{im}}{n-m},\ \text{and}\ \bar{x}'_{i\,n'-m} = \frac{n'\bar{x}'_i - m\bar{x}_{im}}{n'-m},$$

$$i = 1,2,\ldots,p.$$

Finally, let

$$U_{md} = \bar{y}_{n-m} - \sum_{i=1}^{p} a_i(z_m)(\bar{x}_{i\,n-m} - \bar{x}'_{i\,n'-m}), \quad (5.1)$$

and

$$T_{md} = \frac{(n'-m)U_{md} + m\bar{y}_m}{n'}. \quad (5.2)$$

<u>A general class of unbiased estimators</u>

By an argument similar to the one used in section 1, it can be shown that,

$$E\,(T_{md} / m, n') = \bar{y}',$$

where $\bar{y}'$ is the mean based on the sample of size n'.

Consequently $E\,(T_{md}) = E(\bar{y}') = \bar{Y}$, the population mean. Thus

$$T_{md} = \frac{(n'-m)U_{md} + m\bar{y}_m}{n'}$$

$$= \bar{y} - \sum_{i=1}^{p} a_i(z_m)(\bar{x}_i - \bar{x}'_i) - \frac{m(n'-n)}{(n-m)n'}\left[\bar{y}_m - \bar{y} - \sum_{i=1}^{p} a_i(z_m)(\bar{x}_{im} - \bar{x}_i)\right] \quad (5.3)$$

$$= \frac{(n'-m)n}{n'(n-m)}\left[\bar{y} - \sum_{i=1}^{p} a_i(z_m)(\bar{x}_i - \bar{x}'_i)\right]$$

$$- \frac{(n'-n)m}{n'(n-m)}\left[\bar{y}_m - \sum_{i=1}^{p} a_i(z_m)(\bar{x}_{im} - \bar{x}'_i)\right] \quad (5.4)$$

$$= \sum_{i=1}^{p} a_i(z_m)\bar{x}_i' + \frac{(n'-m)n}{n'(n-m)}(\bar{y} - \sum_{i=1}^{p} a_i(z_m)\bar{x}_i)$$
$$- \frac{(n'-n)m}{n'(n-m)}(\bar{y}_m - \sum_{i=1}^{p} a_i(z_m)\bar{x}_{im}), \qquad (5.5)$$

is an unbiased estimator of the population mean $\bar{Y}$.

Now a general class of unbiased estimators may be constructed by including all estimators of the form T_{md}, applied to all possible permutations of the sub-sample elements and weighted averages of such estimators. Of all the estimators of the class, T^*_{md}, obtained as an average of the estimators T_{md}, applied to all the possible permutations of the sub-sample elements, is of more interest since the variance of T^*_{md} is never greater than the variance of T_{md}.

It may be noted that, by putting n' = N in (5.3), (5.4) and (5.5), we obtained the general class of unbiased estimators given by Mickey (1959).

<u>Unbiased ratio type estimators</u>

When information on only one auxiliary variable is taken, for the choice $a(z_m) = R_m = \bar{y}_m / \bar{x}_m$, $1 \leq m \leq n-1$, we obtain the unbiased ratio type estimators, given by

$$T_{1md} = R_m\bar{x}' + \frac{(n'-m)n}{n'(n-m)}(\bar{y} - R_m\bar{x}) \qquad (5.6)$$

and $$T^*_{1md} = R^*_m\bar{x}' + \frac{(n'-m)n}{n'(n-m)}(\bar{y} - R^*_m\bar{x}), \qquad (5.7)$$

where R^*_m is the average of R_m, taken over all the permutations of the sub-sample elements.

In particular, when m = 1, $R^*_m = \frac{1}{n}\sum_{j}^{n}\frac{y_j}{x_j} = \bar{r}_n$, and

$$T^*_{11d} = \bar{r}_n \bar{x}' + \frac{(n'-1)n}{n'(n-1)} (\bar{y} - \bar{r}_n\bar{x}). \qquad (5.8)$$

The estimator T^*_{11d} is a modified form of Hartley and Ross unbiased ratio estimator and has been studied by Sukhatme (1962).

Unbiased Regression type estimators

Using information on only one auxiliary variable, with the choice $a(z_m) = b_m$, we obtain the unbiased regression type estimators given by,

$$T_{2md} = [\bar{y}-b_m(\bar{x}-\bar{x}')] - \frac{m(n'-n)}{(n-m)n'} [\bar{y}_m-\bar{y}-b_m(\bar{x}_m-\bar{x})] \qquad (5.9)$$

and

$$T^*_{2md} = [\bar{y}-b^*_m(\bar{x}-\bar{x}')] + \frac{m(n'-n)}{(n-m)n'} \cdot \frac{1}{\binom{n}{m}} \sum^{\binom{n}{m}} b_m(\bar{x}_m-\bar{x}), \qquad (5.10)$$

where b_m is the regression coefficient based on z_m and b^*_m is the average of b_m over all permutations of the sub-sample elements.

Estimation of variance of T_{md}

We have

$$V(T_{md}) = E[V(T_{md} / n')] + V[E(T_{md} / n')]$$

$$= E[V(T_{md} / n')] + V(\bar{y}'). \qquad (5.11)$$

From result (1.12) of section 1, a non-negative unbiased estimator of $V(T_{md} / n')$ is given by

$$\frac{(n'-n)(n'-m)}{n'^2(n-m)} \cdot \frac{1}{(n-m-1)} \sum_{j}^{n-m} [(y_j-\bar{y}_{n-m}) - \sum_{i=1}^{p} a_i(z_m)(x_{ij}-\bar{x}_{i\,n-m})]^2, \qquad (5.12)$$

where the summation is taken over all sub-sample elements excluding z_m.

Also a non-ve unbiased estimator of $V(\bar{y}')$, based on sub-sample elements only, is given by

$$\left(\frac{N-n'}{Nn'}\right) \cdot \frac{1}{n-1} \sum_{j}^{n} (y_j - \bar{y})^2 \qquad (5.13)$$

From (5.11), (5.12) and (5.13), it follows that a non-ve unbiased estimator of the variance of T_{md} is provided by

$$\text{Est.}V(T_{md}) = \frac{(n'-n)(n'-m)}{n'^2(n-m)} \cdot \frac{1}{(n-m-1)} \sum_{j}^{n-m} \left[(y_j - \bar{y}_{n-m}) - \sum_{i=1}^{p} a_i(z_m)(x_{ij} - \bar{x}_{in-m}) \right.$$

$$+ \left(\frac{N-n'}{Nn'}\right) \cdot \frac{1}{n-1} \sum_{j}^{n} (y_j - \bar{y})^2 \qquad (5.14)$$

<u>Estimator of the variance of T^*_{md}</u>

Following on similar lines, as in single-phase sampling, section 1, it can be shown that an estimator of the variance of T^*_{md} is given by

$$\text{Est.}V(T^*_{md}) = \frac{1}{\binom{n}{m}} \sum^{\binom{n}{m}} \left[\text{Est.}V(T_{md}) - (T_{md} - T^*_{md})^2 \right], \qquad (5.15)$$

where the summation is taken over all the possible $\binom{n}{m}$ estimators of the form T_{md} for a given sub-sample.

<u>Efficiency of Ratio type estimators T^*_{1md}</u>

If $\bar{y}_{Rd}$ denotes the usual biased ratio estimator, $\frac{\bar{y}}{\bar{x}}\,\bar{x}'$, in double sampling, then to the order of approximation $1/n$, we have

$$V(\bar{y}_{Rd}) = \left(\frac{1}{n} - \frac{1}{n'}\right)(S_y^2 + R^2 S_x^2 - 2RS_{xy}) + \frac{N-n'}{Nn'} \cdot S_y^2.$$

(5.16)

Also we can write

$$V(T^*_{1md}) = E[V(T^*_{1md} / n')] + V[E(T^*_{1md} / n')]$$

$$= E[V(T^*_{1md} / n')] + V(\bar{y}'). \qquad (5.17)$$

Now from results (2.16) and (2.22) of section 2, to the order of approximation 1/n, we have

$$V[T^*_{1md} / n'] = \left(\frac{1}{n} - \frac{1}{n'}\right)(s^2_{y,n'} + \bar{R}^2_{m,n'}\, s^2_{x,n'} - 2\bar{R}_{m,n'}\, s_{xy,n'}),$$

when m is small; (5.18)

$$= \left(\frac{1}{n} - \frac{1}{n'}\right)(s^2_{y,n'} + R^2_{n'}\, s^2_{x,n'} - 2R_{n'}\, s_{xy,n'}),$$

when m is sufficiently large;

(5.19)

where $\bar{R}_{m,n'} = E(R_m/n')$, $R_{n'} = \dfrac{\bar{y}'}{\bar{x}'}$,

and $s^2_{y,n'}$, $s^2_{x,n'}$, and $s_{xy,n'}$ are the mean sums of squares and sum of products based on the sample of size n'.

<u>Case i) m is small as compared to n</u>

In this case, from (5.17) and (5.18) to the order 1/n, we have

$$V(T^*_{1md}) = \left(\frac{1}{n} - \frac{1}{n'}\right) E(s^2_{y,n'} + \bar{R}^2_{m,n'} s^2_{x,n'} - 2\bar{R}_{m,n'} s_{xy,n'}) + V(\bar{y}').$$

$$= \left(\frac{1}{n} - \frac{1}{n'}\right)(S^2_y + \bar{R}^2_m S^2_x - 2\bar{R}_m S_{xy}) + \left(\frac{1}{n'} - \frac{1}{N}\right) S^2_y,$$

where $\bar{R}_m = E(R_m)$. (5.20)

In particular, when m=1, estimator T^*_{11d} is a modified form of Hartley and Ross unbiased estimator and its variance is given by

$$V(T^*_{11d}) = \left(\frac{1}{n} - \frac{1}{n'}\right)\left(S_y^2 + \bar{r}_N^2 S_x^2 - 2\bar{r}_N S_{xy}\right) + \left(\frac{1}{n'} - \frac{1}{N}\right)S_y^2 , \quad (5.21)$$

where

$$\bar{r}_N = \frac{1}{N}\sum_{j=1}^{N}\frac{y_j}{x_j} = \frac{1}{N}\sum_{j=1}^{N} r_j .$$

Neglecting the finite population correction factor, Sukhatme (1962) has given the variance of T^*_{11d} in the form:

$$V(T^*_{11d}) = \frac{1}{n}\left(\sigma_y^2 + \bar{r}_N^2 \sigma_x^2 - 2\bar{r}_N \sigma_{xy}\right) + \frac{1}{n'}\left\{\bar{r}_N^2 \sigma_x^2 + 2\bar{r}_N \bar{X}\sigma_{rx} + 2\bar{r}_N E(\Delta^2 x \, \Delta r)\right\} \quad (5.22)$$

where

$$\Delta x = x - \bar{X}, \quad \Delta r = r - \bar{r}_N,$$

$\sigma_y^2 = V(y)$, $\sigma_x^2 = V(x)$, $\sigma_{xy} = Cov(y,x)$ and $\sigma_{rx} = Cov(r,x)$.

When the finite population correction factor is ignored, it can be easily shown that expressions (5.21) and (5.22) are identical by making use of the identity

$$\sigma_{xy} = \bar{r}_N^2 \sigma_x^2 + \bar{X}\sigma_{rx} + E(\Delta^2 x \, \Delta r). \quad (5.23)$$

The advantage of the form (5.21) and in general (5.20) is that it is easily comparable with the variance of the biased ratio estimator $\bar{y}_{Rd}$, given by (5.16). Thus a comparison of (5.16) and (5.20) shows that the unbiased ratio estimator T^*_{1md} is more efficient than the biased ratio estimator $\bar{y}_{Rd}$, if and only if the population regression coefficient of y on x is nearer to $\bar{R}_m = \frac{1}{\binom{N}{m}}\sum^{\binom{N}{m}} R_m$ than to the population ratio $R = \bar{Y} / \bar{X}$.

Case ii) m is sufficiently large

In this case from (5.17) and (5.19), to the order of approximation 1/n, we have

$$V(T^*_{1md}) = (\frac{1}{n} - \frac{1}{n'})E(s^2_{y,n'} + R^2_{n'} s^2_{x,n'} - 2R_{n'} s_{xy,n'}) + (\frac{1}{n'} - \frac{1}{N})S^2_y$$

$$= (\frac{1}{n} - \frac{1}{n'})(S^2_y + R^2 S^2_x - 2R S_{xy}) + (\frac{1}{n'} - \frac{1}{N})S^2_y \quad (5.24)$$

Results (5.16) and (5.24) establish that, when m is sufficiently large, T^*_{1md} is as efficient as the biased ratio estimator $\bar{y}_{Rd}$.

Efficiency of Regression type estimators T^*_{2md}

In double sampling, the usual biased regression estimator is given by $\bar{y}_{1rd} = \bar{y} + b(\bar{x} - \bar{x}')$, where b is the regression coefficient based on the sub-sample of size n.

Also it is known that, to the order of approximation 1/n,

$$V(\bar{y}_{1rd}) = (\frac{1}{n} - \frac{1}{n'})(S^2_y + B^2 S^2_x - 2B S_{xy}) + (\frac{1}{n'} - \frac{1}{N})S^2_y, \quad (5.25)$$

where B is the population regression coefficient.

For the unbiased regression estimator T^*_{2md} we write

$$V(T^*_{2md}) = E[V(T^*_{2md} / n')] + V[E(T^*_{2md} / n')].$$

Here substituting for $V(T^*_{2md} / n')$ from the results of section 4, we have, when m is sufficiently large, to the order 1/n,

$$V(T^*_{2md}) = E[(\frac{1}{n} - \frac{1}{n'})(s^2_{y,n'} + b^2_{n'} s^2_{x,n'} - 2b_{n'} s_{xy,n'})]$$
$$+ (\frac{1}{n'} - \frac{1}{N})S^2_y$$

$$= \left(\frac{1}{n} - \frac{1}{n'}\right) (S_y^2 + B^2 S_x^2 - 2B\, S_{xy}) + \left(\frac{1}{n'} - \frac{1}{N}\right) S_y^2, \tag{5.26}$$

where $b_{n'}$ is the regression coefficient based on the sample of size n'.

Thus from (5.25) and (5.26) it follows that, when m is sufficiently large, the unbiased regression estimator T^*_{2md} is as efficient as the biased regression estimator $\bar{y}_{lrd}$. Also a comparison of (5.24) and (5.26) proves that the unbiased regression estimators T^*_{2md} are never less efficient than the unbiased ratio estimators T^*_{1md}, when m is sufficiently large.

Finally, it is interesting to note that the results concerning the relative efficiency of the unbiased ratio and regression estimators with respect to the usual biased ratio and regression estimators, in double sampling, are exactly the same as those obtained in the single-phase sampling.

6. APPLICATION TO STRATIFIED POPULATIONS

In this section, a stratified population with one auxiliary variable is considered. Assuming that the strata means of the auxiliary variable are known, two sets of 'combined' and 'separate' unbiased ratio type estimators based on a stratified simple random sample, drawn without replacement, are obtained together with unbiased estimators of their precision.

Preliminaries

Let the finite population of size N be divided into L strata with N_h units in the h^{th} stratum, for $h = 1,2,\ldots,L$. Let (y_{hi}, x_{hi}) denote the observations on the principle variable y and the auxiliary variable x for the i^{th} unit in the h^{th} stratum. Suppose $\bar{Y}_h$ and $\bar{X}_h$ represent the h^{th} stratum means. Define the population means $\bar{Y} = \sum_{h=1}^{L} P_h \bar{Y}_h$ and $\bar{X} = \sum_{h=1}^{L} P_h \bar{X}_h$, where $P_h = N_h/N$.

Let a simple random sample of size n_h be drawn without replacement from the h^{th} stratum $(h=1,2,\ldots,L)$ with $\sum_{h=1}^{L} n_h = n$. Further, let $\bar{y}_h$ and $\bar{x}_h$ denote the means based on the sample from the h^{th} stratum and $\bar{y} = \frac{1}{n}\sum_{h=1}^{L} n_h \bar{y}_h$, $\bar{x} = \frac{1}{n}\sum_{h=1}^{L} n_h \bar{x}_h$, be the means based on the total stratified sample.

Suppose z_{m_h}, $1 \leq m_h \leq n_h - 1$, represent a set of m_h elements, chosen out of the n_h sample elements from the h^{th} stratum, with $\sum_{h=1}^{L} m_h = m$. Based on the set z_{m_h}, let $\bar{y}_{m_h}$ and $\bar{x}_{m_h}$ represent the means and R_{m_h}, the ratio $\bar{y}_{m_h} / \bar{x}_{m_h}$.

Further, let, $\bar{y}_m = \frac{1}{m} m_h \bar{y}_{m_h}$, $\bar{x}_m = \frac{1}{m} m_h \bar{x}_{m_h}$ and $R_m = \bar{y}_m / \bar{x}_m$.

Finally, define

$$\bar{y}_{n_h-m_h} = \frac{n_h \bar{y}_h - m_h \bar{y}_{m_h}}{n_h - m_h}, \quad \bar{Y}_{N_h-m_h} = \frac{N_h \bar{Y}_h - m_h \bar{y}_{m_h}}{N_h - m_h},$$

and

$$\bar{x}_{n_h-m_h} = \frac{n_h \bar{x}_h - m_h \bar{x}_{m_h}}{n_h - m_h}, \quad \bar{X}_{N_h-m_h} = \frac{N_h \bar{X}_{N_h} - m_h \bar{x}_{m_h}}{N_h - m_h}.$$

<u>Separate unbiased ratio type estimators</u>

As usual, a separate unbiased ratio type estimator is formed by estimating the strata means $\bar{Y}_h$ $(h = 1,2,\ldots,L)$ with the help of unbiased ratio type estimators within different strata.

For the h^{th} stratum mean $\bar{Y}_h$, Mickey's unbiased ratio type estimators are given by

$$T_{1m_h} = R_{m_h} \bar{X}_h + \frac{(N_h-m_h)n_h}{N_h(n_h-m_h)} \left[\bar{y}_h - R_{m_h} \bar{x}_h \right], \tag{6.1}$$

$$\text{and } T^*_{1m_h} = \overset{*}{R}_{m_h} \bar{X}_h + \frac{(N_h-m_h)n_h}{N_h(n_h-m_h)} \left[\bar{y}_h - \overset{*}{R}_{m_h} \bar{x}_h \right], \tag{6.2}$$

$$\text{where } \overset{*}{R}_{m_h} = \frac{1}{\binom{n_h}{m_h}} \sum^{\binom{n_h}{m_h}} R_{m_h}.$$

Consequently

$$T_{1m(s)} = \sum_{h=1}^{L} P_h T_{1m_h}, \tag{6.3}$$

$$\text{and } \overset{*}{T}_{1m(s)} = \sum_{h=1}^{L} P_h \overset{*}{T}_{1m_h}, \tag{6.4}$$

provide separate unbiased ratio type estimators for the population mean $\bar{Y}$.

Now since m_h elements of the n_h sample elements can be chosen in $\binom{n_h}{m_h}$ ways, for a given stratified random sample, we have $\prod_{h=1}^{L}\binom{n_h}{m_h}$ estimators of the form $T_{1m(s)}$.

Averaging over all such estimators also we obtain

$$\frac{1}{\prod_{h=1}^{L}\binom{n_h}{m_h}} \sum^{\prod_{h=1}^{L}\binom{n_h}{m_h}} T_{1m(s)} = T^{*}_{1m(s)} \cdot$$

As the stratified random sample plays the role of a sufficient statistic, from this it follows that $T^{*}_{1m(s)}$ is never less efficient than $T_{1m(s)}$.

Combined unbiased ratio type estimators

Following Mickey's principle, we now obtain combined unbiased ratio type estimators.

For this we define

$$z_m = \{ z_{m_h} \}, \; h = 1,2,\ldots,L,$$

and

$$U_m = \sum_{h=1}^{L} Q_h \bar{y}_{n_h-m_h} - R_m \left[\sum_{h=1}^{L} Q_h(\bar{x}_{n_h-m_h} - \bar{X}_{N_h-m_h}) \right], \tag{6.5}$$

where

$$Q_h = \frac{N_h - m_h}{N-m}.$$

Now since the sampling within different strata is independent, for a given set z_m, the $n_h - m_h$ elements, obtained by excluding the set z_{m_h} from the n_h sample elements of the h^{th} stratum, constitute a simple random sample without replacement from the $N_h - m_h$ elements of h^{th} stratum, ($h=1,2,\ldots,L$).

Consequently, for a given set z_m, we have

$$E(U_m / m) = \sum_{h=1}^{L} Q_h \, E(\bar{y}_{n_h-m_h} / m) - R_m \left[\sum_{h=1}^{L} Q_h E(\bar{x}_{n_h-m_h}/m) - \sum_{h=1}^{L} Q_h \, \bar{X}_{N_h-m_h} \right]$$

$$= Q_h \, \bar{Y}_{N_h-m_h},$$

$$= \frac{N\bar{Y} - m\bar{y}_m}{N - m}. \qquad (6.6)$$

Thus

$$T_{1m(c)} = \frac{(N-m) \; U_m + m\bar{y}_m}{N} \qquad (6.7)$$

is conditionally and hence unconditionally unbiased for the population mean $\bar{Y}$.

Substituting for U_m from (6.5) in (6.7), it can be seen that the combined unbiased ratio type estimator $T_{1m(c)}$ is also given by

$$T_{1m(c)} = R_m \bar{X} + \frac{1}{N} \sum_{h=1}^{L} (N_h - m_h) \; (\bar{y}_{n_h-m_h} - R_m \, \bar{x}_{n_h-m_h}). \qquad (6.8)$$

Averaging over all the possible $\prod_{h=1}^{L} \binom{n_h}{m_h}$ estimators of the form $T_{1m(c)}$, for a given stratified random sample we obtain the combined unbiased ratio type estimator

$$\overset{*}{T}_{1m(c)} = \frac{1}{\prod_{h=1}^{L} \binom{n_h}{m_h}} \sum^{\prod_{h=1}^{L} \binom{n_h}{m_h}} T_{1m(c)}, \qquad (6.9)$$

which is never less efficient than $T_{1m(c)}$.

Further, if for a given m, the m_h (h=1,2,...,L) are so chosen that $\dfrac{N_h - m_h}{n_h - m_h}$ = constant, (i.e.) $\dfrac{N_h - m_h}{n_h - m_h} = \dfrac{N-m}{n-m}$, then $T_{1m(c)}$ and $\overset{*}{T}_{1m(c)}$ assume simpler forms given by

$$T_{1m(c)} = R_m \bar{X} + \frac{(N-m) \, n}{N(n-m)} \left[\bar{y} - R_m \, \bar{x} \right], \qquad (6.10)$$

and

$$T^{*}_{1m(c)} = R^{*}_m \bar{X} + \frac{(N-m) \, n}{N(n-m)} \left[\bar{y} - R^{*}_m \; \bar{x} \right], \qquad (6.11)$$

where $R_m^{*} = \frac{1}{\prod_{h=1}^{L}\binom{n_h}{m_h}} \sum^{\prod_{h=1}^{L}\binom{n_h}{m_h}} R_m.$

It is interesting to note that estimators (6.10) and (6.11) are remarkably similar in form to the unbiased ratio type estimators, based on an unstratified random sample of size n. Further, in proportional allocation (i.e, when $n_h = nP_h$), for a given m, the condition

$\frac{N_h - m_h}{n_h - m_h} = \frac{N-m}{n-m}$ is satisfied if the choice of m_h is also proportional to P_h (i.e., $m_h = mP_h$).

Thus in proportional allocation for the choice $m_h = mP_h$, estimators (6.10) and (6.11) provide combined unbiased ratio type estimators.

Estimation of variance

We first give unbiased estimators of the variance of the separate ratio type estimators and then obtain unbiased estimation of the variance of the combined estimators.

(i) Separate unbiased ratio type estimators

From result (1.15) of section 1, a non-ve unbiased estimator of the variance of Mickey's unbiased ratio type estimator T_{1m_h} in the h^{th} stratum is given by

$$Est.V(T_{1m_h}) = \frac{(N_h-n_h)(N_h-m_h)}{(n_h-m_h)N_h^2} \cdot \frac{1}{(n_h-m_h-1)} \sum_{i}^{n_h-m_h} \left[(y_{hi} - R_{m_h} x_{hi}) - (\bar{y}_{n_h-m_h} - R_{m_h}\bar{x}_{n_h-m_h}) \right]^2,$$

except for the choice $m_h = n_h - 1$. (6.12)

Consequently, a non-ve unbiased estimator of the variance of the separate unbiased ratio type estimator $T_{1m(s)}$ is provided by

$$\text{Est.}V(T_{1m(s)}) = \sum_{h=1}^{L} P_h^2 \ \text{Est.}V(T_{1m_h}). \qquad (6.13)$$

Also, since for a given stratified random sample,

$$E\left[T_{1m(s)} \,/\, n_h, h = 1,2,\ldots\ldots,L \right] = \overset{*}{T}_{1m(s)}$$

an unbiased estimator of the variance of the separate unbiased ratio type estimator $\overset{*}{T}_{1m(s)}$ is given by

$$\text{Est.}V(\overset{*}{T}_{1m(s)}) = \frac{1}{\prod_{h=1}^{L}\binom{n_h}{m_h}} \sum^{\prod_{h=1}^{L}\binom{n_h}{m_h}} \left[\text{Est.}V(T_{1m(s)}) - (T_{1m(s)} - \overset{*}{T}_{1m(s)})^2 \right]. \quad (6.14)$$

(ii) <u>Combined unbiased ratio type estimators</u>

To obtain a non-ve unbiased estimator of the variance of the combined ratio type estimator $T_{1m(c)}$, we note that

$$\begin{aligned} V(T_{1m(c)}) &= E\left[V(T_{1m(c)} / m) \right] + V\left[E(T_{1m(c)}/m) \right] \\ &= \frac{(N-m)^2}{N^2} \, E\left[V(U_m/m) \right], \text{ from (6.7)} \\ &= \frac{(N-m)^2}{N^2} \, E\left[\sum_{h=1}^{L} Q_h^2 \, V(\bar{y}_{n_h - m_h} - R_m \bar{x}_{n_h-m_h}/m) \right], \\ &\qquad \text{from (6.5).} \qquad (6.15) \end{aligned}$$

Now, clearly, a non-ve unbiased estimator of

$V(\bar{y}_{n_h - m_h} - R_m \bar{x}_{n_h - m_h} / m)$ is provided by

$$\frac{(N_h - n_h)}{(n_h-m_h)(N_h-m_h)} \cdot \frac{1}{(n_h-m_h-1)} \sum_{i}^{n_h-m_h} \left[(y_{hi}-R_m x_{hi}) - (\bar{y}_{n_h-m_h} - R_m \bar{x}_{n_h-m_h}) \right]^2, \qquad (6.16)$$

except for the choice $m_h = n_h - 1$.

Consequently from (6.15) and (6.16), a non-ve unbiased estimator of $T_{lm(c)}$ is given by

$$Est.V(T_{lm(c)}) = \sum_{h=1}^{L} \frac{(N_h-n_h)(N_h-m_h)}{N^2(n_h-m_h)} \cdot \frac{1}{(n_h-m_h-1)} \sum_{i}^{n_h-m_h} \left[(y_{hi}-R_m x_{hi}) - (\bar{y}_{n_h-m_h} - R_m \bar{x}_{n_h-m_h}) \right]^2 \qquad (6.17)$$

Again, since for a given stratified random sample,

$$E \left[T_{lm(c)} / n_h,\ h=1,2,\ldots\ldots,L \right] = T^*_{lm(c)}$$

an unbiased estimator of the variance of $T^*_{lm(c)}$ is provided by

$$Est.V\left[T_{lm(c)}\right] = \frac{1}{\prod_{h=1}^{L} \binom{n_h}{m_h}} \sum^{\prod_{h=1}^{L} \binom{n_h}{m_h}} \left[Est.V(T_{lm(c)}) - (T_{lm(c)} - T^*_{lm(c)})^2 \right] \qquad (6.18)$$

7. NUMERICAL EXAMPLES

The theoretical investigation of the efficiency of Mickey's unbiased ratio and regression type estimators presented many difficult problems in view of the complexity of the estimators themselves and in fact no theoretical appraisal of their performance in small samples has been possible. Even the verification of the results, obtained in respect of their efficiency in large samples, involves heavy computations and is possible only with the help of the electronic computer. Also one of the interesting problems still remained unsolved is the behaviour of these estimators for increasing values of 'm'. In this section, however, a few numerical examples have been taken up in these directions for the unbiased ratio type estimators. Unless extensive comparisons are made, no general conclusions can be drawn regarding their performance in small samples; for increasing values of m; etc.

The first example demonstrates the construction of an exact unbiased estimator of the variance of Hartley and Ross estimator, as has been suggested by the results (1.15) and (1.16) of section 1, for a sample of size 9. In the second example, a sample of size 15 has been taken to obtain consistent estimators of the variance of the unbiased ratio type estimators T^{*}_{1m} (m = 1,2 and 3), with help of the result (2.16) of section 2. In the third example, a sample of size 100, with known population coefficients of variation and covariation, is taken up to study the variance of T^{*}_{1m}, for the choice of m ranging from 75 to 99, with the help of the bivariate normal approximation formula (2.37) of section 2. Finally, an artificial stratified population consisting of 3 strata, given by Cochran (II edition, page 179), has been considered to study the efficiency of the combined unbiased ratio type estimator $T^{*}_{i1(c)}$ of section 6.

Example I:

The data for this example come from a simple random sample of size 9, drawn without replacement from the 91 villages of the Venkatagiri Taluq in Nellore district in order to study the yield and cultivation practices of Lime. In Table 7.1, y_j represents the number of bearing trees and x_j the area (in acres) reported initially under Lime, for the j^{th} village. The problem is to estimate the average number of bearing trees per village in Venkatagiri stratum.

Table 7.1

Sampled village Code number	No. of bearing trees (y_j)	Area (in acres) reported initially
1	291	5.60
2	163	2.95
3	78	7.05
4	261	5.36
5	1302	13.50
6	504	6.59
7	1403	16.81
8	1703	12.63
9	554	5.68

Population size $N = 91$ Population mean of $x = \bar{X} = 11.731$ acres.

Sample size $n = 9$ $n\bar{y} = y(n) = 6259$,

$n\bar{x} = x(n) = 76.17$ $s_y^2 = 369677.75$

$s_x^2 = 21.73$ $s_{xy} = 2528.21$

M.S.E.$(\bar{y}_R) = 10104$,

where, $\bar{y}_R$ is the usual biased ratio estimator.

Table 7.2

Illustration of computation for unbiased estimator of the variance of Hartley and Ross estimator T_{11}^* specified by (1.16) of section 1. (m=1)

Village No.	R_1	R_1^2	$\sum_j^{n-1} y_j^2$	$\sum_j^{n-1} x_j^2$	$\sum_j^{n-1} x_j y_j$	$\frac{[y(n)-R_1 x(n)]^2}{8}$	Est.V(T_{11})
1.	51.964	2700.26	7225528	787.169	71568.11	661769	20029
2.	55.254	3053.00	7283640	809.826	72716.86	525468	19015
3.	11.064	128.39	7304125	768.826	72647.81	3666977	33871
4.	48.694	2371.11	7242088	789.799	71798.75	812798	20842
5.	96.444	9301.45	5615005	636.279	55620.71	147734	10456
6.	76.480	5849.19	7056193	775.101	69876.35	23492	13975
7.	83.462	6965.91	5341800	535.953	49613.28	1208	12610
8.	134.838	18181.29	4410000	659.012	51688.82	2011627	7016
9.	97.535	9513.08	7003293	786.267	70050.99	171183	10298

$$\bar{R}_1 = \frac{1}{9}\sum R_1 = 72.859 ,$$

$$\frac{1}{9}\sum \text{Est.V}(T_{11}) = 16457, \quad (\bar{X} - \frac{(N-1)n}{N(n-1)}\bar{\bar{x}})^2 \frac{1}{9}\sum(R_1-\bar{R}_1)^2 = 6100.$$

$$\text{Est.V}(T_{11}^*) = 16457 - 6100 = 10357.$$

Estimated relative efficiency of Hartley and Ross unbiased estimator with respect to $\bar{y}_R$, is, therefore, given by $\frac{10104}{10357} \times 100 = 97.56\%$. T^*_{11} is hence preferable in view of its property of unbiasedness.

Example II

The data for this example are based on a simple random sample of size 15, drawn without replacement from the population of 91 villages of example I. Unbiased ratio type estimators T^*_{1m}, m = 1,2 and 3 are compared with $\bar{y}_R$ by calculating consistent estimates of their variance, obtained from the formula (2.16) of section 2.

Table 7.3

Sampled village Code number	No. of bearing Trees (y_j)	Area reported initially (x_j)
1	698	6.15
2	1403	16.81
3	873	8.72
4	212	5.34
5	558	4.60
6	78	7.05
7	1302	13.50
8	571	7.05
9	0	5.01
10	1302	13.50
11	291	5.60
12	307	12.54
13	1063	13.23
14	168	2.60
15	1470	14.77

For this sample $R_m = 74.896$, $R_1^* = 70.810$

$R_2^* = 72.843$, $R_3^* = 73.620$

Also $s_y^2 = 266245$, $s_x^2 = 20.706$

$s_{xy} = 1896.94$, $b_n = 91.612$,

where, b_n is the sample regression coefficient.

Using the formula (2.16) of section 2, for small values of m, a consistent estimate of the variance of T_{1m}^* is given by

$$\text{Est.V}(T_{1m}^*) = \text{Est.V}(\bar{y}) + R_m^{*2}\,\text{Est.V}(\bar{x}) - 2R_m^*\,\text{Est.Cov.}(\bar{y},\bar{x}).$$

Also a consistent estimate of the variance of $\bar{y}_R$ is given by

$$\text{Est.V}(\bar{y}_R) = \text{Est.V}(\bar{y}) + R_n^2\,\text{Est.V}(\bar{x}) - 2R_n\,\text{Est.Cov.}(\bar{y},\bar{x}).$$

Consequently, since in this example

$$n < R_1^* < R_2^* < R_3^* < R_n < b_n ,$$

we expect the inequality

$$\text{Est.V}(\bar{y}_R) < \text{Est.V}(T_{13}^*) < \text{Est.V}(T_{12}^*) < \text{Est.V.}(T_{11}^*).$$

Table 7.4 gives the estimates of the variance and the relative efficiencies compared to $\bar{y}_R$.

Table 7.4

Estimator	Estimate of the variance	Relative efficiency
$\bar{y}_R$	5470.393	100.00
Hartley and Ross T_{11}^*	5647.148	96.87
T_{12}^*	5554.387	98.49
T_{13}^*	5521.452	99.08

Although the unbiased estimators are all less efficient than the biased estimator, their variance is not significantly more than that of $\bar{y}_R$. Thus they compare satisfactorily with $\bar{y}_R$, from the point of view of efficiency. On the other hand computation of the biased estimator is the easiest. Among the unbiased estimators, T^{*}_{13} is the best from the point of view of efficiency, but its computation is relatively difficult. For this particular example, for increasing values of m from below, there is an increase in the efficiency of the unbiased estimator.

Example III

The data for this example are taken from page 171, second ~~volume~~ edition of Cochran's 'Sampling Techniques'. From a census of all the 2010 farms in Jefferson County in respect of the acreage under the corn crop (y) and the total acreage (x) of the farm, the following are the population means and the coefficients of variation and covariation.

$$\bar{Y} = 26.30, \quad \bar{X} = 117.28$$
$$C_y^2 = 0.896355, \quad C_x^2 = 0.553924,$$
$$C_{xy} = 0.471071.$$

A simple random sample of size 100 is assumed to be drawn from this population. The bivariate normal approximations (2.25) and (2.37) of section 2, are calculated respectively for the M.S.E. of $\bar{y}_R$ and the variance of T^{*}_{1m}, m ranging from 75 to 99. The variances together with the relative efficiencies are tabulated in Table 7.5.

Table 7.5

Estimator	M.S.E./variance	Relative efficiency
$\bar{y}_R$	35.7597	100.000
T^*_{1m}, m = 75	35.3848	101.059
80	35.3862	101.055
85	35.3910	101.042
90	35.4014	101.012
95	35.4353	100.915
99	35.7189	100.114

The results show a steady decrease in the efficiency of T^*_{1m} as m increases from 75 to 99, although the decrease is not quite significant. For all m ranging from 75 to 99, T^*_{1m} is more efficient than $\bar{y}_R$, although once again the gain in efficiency is not significant. The results confirm that to the first order of approximation, when m is sufficiently large, T^*_{1m} is as efficient as the biased ratio estimator $\bar{y}_R$.

Example IV

In this example, an artificial stratified population of 3 strata, constructed by Cochran (~~vol.~~ Edn.II, page 179), is considered. Each stratum contains 4 units out of which 2 units are selected at random without replacement. Thus the allocation of the total sample size n=6 is proportional to the strata sizes N_h (h=1,2 and 3). The population was constructed in such a way that (a) R_h varies markedly from stratum to stratum, thus favouring a separate ratio estimator, and (b) the ratio estimator within each stratum is badly biased. The choice of m_h is equal to 1 in each stratum, so that the averaged 'separate' and 'combined' unbiased ratio type estimators for the population total Y are respectively given by the separate Hartley and Ross estimator $N\,T^{*}_{13(s)}$ of (6.4) and the combined unbiased ratio type estimator $N\,T^{*}_{13(c)}$ of (6.11).

Five methods of estimating the population total are compared.

1. Simple expansion	:	$\sum_{h=1}^{L} N_h \bar{y}_h$.
2. The combined biased ratio estimator	:	$(\bar{y} / \bar{x})X$.
3. The separate biased ratio estimator	:	$\sum_{h=1}^{L} (\bar{y}_h / \bar{x}_h) X_h$.
4. The separate Hartley and Ross unbiased ratio estimator	:	$T^{*}_{13(s)}$ of (6.4)
5. The combined unbiased ratio type estimator	:	$T^{*}_{13(c)}$ of (6.11).

There are $6^3 = 216$ possible samples. The biases and variances are exact, since all possible samples are taken into account.

Table 7.6

A small artificial population

	Stratum					
	I		II		III	
	y	x	y	x	y	x
	2	2	2	1	3	1
	3	4	5	4	7	3
	4	6	9	8	9	4
	11	20	24	23	25	12
R_h	0.625		1.111		2.200	

Table 7.7

Results for the different estimators of Y

Method	Variance	$|Bias|^2$	M.S.E.
Simple expansion ..	820.3	0.0	820.3
Combined biased ratio ..	262.8	6.5	269.3
Separate biased ratio ..	35.9	24.1	60.0
Separate Hartley and Ross	153.6	0.0	153.6
Combined unbiased ratio	142.4	0.0	142.4

Irrespective of the extreme conditions, the contribution of the $(bias)^2$ to the mean square error of the combined biased ratio estimator is trivial. Because of considerable variation in R_h, the separate biased ratio estimator is much more accurate than the combined biased ratio estimator, but it is badly biased. The Hartley and Ross separate unbiased ratio estimator is superior to the combined biased ratio estimator, but inferior to the combined unbiased ratio estimator as well as to the separate biased ratio estimator, as judged by the M.S.E. of the latter. The combined unbiased ratio estimator is more efficient than any other estimator except the separate biased ratio estimator.

Cochran has included separate Lahiri unbiased ratio estimator also in the comparisons, but in the author's view it is not comparable as it is based entirely on a different probability sampling scheme. The five estimators compared here are all based on stratified simple random sampling without replacement.

REFERENCES

Lahiri, D.B. (1951). A method of sample selection providing unbiased ratio estimators. Bull. Inst. Internat. Statist., 33,2, 133-140.

Midzuno, H. (1952). On the sampling system with probability proportionate to sum of sizes. Ann. Inst. Statist. Math., 3, 99-107.

Sen, A.R. (1952). Further developments of the theory and application of the selection of primary units with special reference to the North Carolina agricultural population. Ph.D. Thesis, Univ. North Carolina.

Des Raj (1954). Ratio estimation in sampling with equal and unequal probabilities. J. Indian Soc. Agric. Statist., 6, 127-138.

Hartley, H.O. and Ross, A. (1954). Unbiased ratio estimators. Nature, 174, 270-271.

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

Robson, D.S. (1957). Applications of multivariate polykays to the theory of unbiased ratio type estimation. Jour. Amer. Stat. Assoc., 52, 511-522.

Goodman, L.A. and Hartley, H.O. (1958). The precision of unbiased ratio type estimators. Jour. Amer. Stat. Assoc., 53, 491-508.

Mickey, M.R. (1959). Some finite population unbiased ratio and regression estimators. Jour. Amer. Stat. Assoc., 54, 594-612.

Murty, M.N. and Nanjamma, N.S. (1959). Almost unbiased ratio estimates based on interpenetrating sub-sample estimates. Sankhya, 21, 381-392.

____________________ and Sethi, V.K. (1959). Some sampling systems providing unbiased ratio estimators. Sankhya, 21, 299-314.

Williams, W.H. (1961). Generating unbiased ratio and regression estimators. Biometrics, 17, 267-74.

Jose Nieto Pascual (1961). Unbiased ratio estimators in stratified sampling. Jour. Amer. Stat. Assoc., 56, 70-87.

Sukhatme, B.V. (1962). Some ratio type estimators in two-phase sampling. Jour. Amer. Stat. Assoc., 57, 628-632.

______________ (1962). Generalized Hartley-Ross unbiased ratio type estimator. Nature, vol.196, 1238.

Ravindra Singh (1962). Some estimation procedures in fruit surveys. Unpublished Thesis for Diploma of I.C.A.R.,New Delhi.

Williams, W.H. (1963). The precision of some unbiased regression estimators. Biometrics, 19, 352-361.

Pathak, P.K. (1964). On sampling schemes providing unbiased ratio estimators. Ann. Math. Stat., 35, 222-231.

Sukhatme, P.V. (1953). Sampling theory of surveys with applications. Iowa State College Press.

Hansen, M.H., Hurwitz. W.N. and Madow, W.G. (1953). Sample survey methods and theory. vol.II; John Wiley and Sons.

Cochran, W.G. (1963). Sampling Techniques. Second edition, John Wiley and Sons.

-------:0:------