# UNBIASED RATIO-TYPE ESTIMATORS

by

REGINALD GERALD WORTHLEY

B. A., University of Maine, 1965

---

A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1967

Approved by:

A. M. Feyerherm
Major Professor

TABLE OF CONTENTS

## 1. INTRODUCTION

Ratio estimators have been used quite extensively in sample surveys, not only as estimators of population ratios, but as estimators of population means and totals. In the latter case they involve the use of an extra variable, correlated with the variable of interest. These ratio estimators, although known to be biased, have often been preferred over the traditional unbiased mean per unit estimator, since it has been demonstrated that in a great many situations the ratio estimator has a smaller variance. A major drawback to the ratio estimator is the fact that it is biased, although in large samples it has been demonstrated that the bias is negligible. In very small samples, or even moderate samples from a stratified population, no really convincing argument has been given for the negligibility of the bias, since no exact expression for it is available. Several authors have avoided this question of bias by developing methods which eliminate the bias while retaining the essential properties of a ratio estimator.

This paper reviews the usual ratio estimator, giving optimum conditions for its use. The bias is approximated and limits for the bias are given, as well as cases that might arise in which the bias might become an important factor. Methods are then considered which give rise to reduced bias estimators, as well as unbiased ratio-type estimators. The latter is divided into two major classes of development, (1) the elimination of bias through the use of commonly used sampling schemes, and (2) the elimination of bias through the use of certain modifications of sampling schemes making the usual biased estimator unbiased.

## 2. THE BIASED RATIO ESTIMATOR

The classic estimator for a population mean, $\bar{Y}$, or population total, Y, has been the sample mean, $\bar{y}$, and inflated sample mean, $N\bar{y}$, where $N$ is the finite population size. In the past quarter of a century, the ratio estimator, using a variable $x$ correlated with the variable of interest $y$ to estimate population means and totals, has come into prominence, especially in surveys. The usual simple ratio estimator is

$$R_1 = \bar{y}/\bar{x}$$

the ratio of the two sample means. Corresponding estimators of the population mean and total are; respectively,

$$\bar{Y}_{R_1} = \bar{y}/\bar{x} \cdot \bar{X}$$

and

$$Y_{R_1} = \bar{y}/\bar{x} \cdot X$$

where $X$ is the population total of the $x$ values. It is noted that except for estimating the ratio, the population total of the $x$ values has to be known.

Although these estimators are known to be biased except in certain situations, it is very common in practice, that they have a smaller variance than those based on the mean per unit estimator,

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

Cochran (2) explains, that for large samples, if the correlation coefficient between $y$ and $x$ is greater than one half of the coefficient of variation of $x$ divided by the coefficient of variation of $y$, the ratio estimate of $Y$ has a smaller variance than the simple expansion method, $N\bar{y}$.

This occurs very often in survey practice. One of the common uses of ratio estimators is when $x_i$ is the value $y_i$ at some previous time, and here the two coefficients of variation may be about equal. If the coefficient of correlation is greater than 0.5, in this case, the ratio estimate is superior.

Cochran (2) also applied the Gauss-Markov Theorem to show that if the regression of $y$ on $x$ is a straight line through the origin, and the variance of $y_i$ about this line is proportional to $x_i$, then the ratio estimate is a minimum-variance unbiased estimator. It is also known that in large samples the distribution of $R_1$, the simple ratio estimator, tends to a normal distribution, and since the bias is of order $1/n$, the bias tends to zero.

There are cases when the existence of a bias becomes an important factor. Goodman and Hartley (7) state there is one very important class of surveys in which the bias may become of vital interest. This arises when drawing small samples from a large number (k) of strata. It often occurs in sampling, that the bias in each sample will be of the same sign, therefore the bias in the estimate of the population total will be $k$ times the bias for a stratum total. Since the variance only multiplies by $k$, the mean square error of the estimate of the population total will be of order of magnitude $k^2$, whereas if unbiased the order of magnitude would be $k$. It is evident that an unbiased ratio estimator in this case would be of great advantage. Lahiri (18) emphasizes particularly the risk involved in using the usual (biased) ratio-estimator in small samples from many strata, so, since no such risk is involved in the unbiased ratio-type estimators, it is easily seen that more extensive stratification is possible.

Devices for reducing and eliminating the bias have mostly been developed since the early 1950's. Although many of the estimators arrived at seem very burdensome to calculate, this seems like an unimportant objection to their use, since much survey work is being done by computers.

## 3. THE ALMOST UNBIASED RATIO-TYPE ESTIMATOR

### 3.1. Early Work

Since the bias in the usual ratio estimator

$$R_1 = \bar{y}/\bar{x}$$

is, essentially, the product of two random variables, the exact expression for the bias cannot be obtained in a straightforward manner. The first practical method proposed for finding the bias used a Taylor's series expansion.

$$R_1 - R = \bar{y}/\bar{x} - R = \frac{\bar{y} - R\bar{x}}{\bar{x}}$$

$$= \frac{\bar{y} - R\bar{x}}{\bar{X}} \cdot \frac{\bar{X}}{\bar{x}} = \frac{\bar{y} - R\bar{x}}{X} \cdot \bar{X} \left(\frac{1}{\bar{X} + (\bar{x} - \bar{X})}\right)$$

$$= \frac{\bar{y} - R\bar{x}}{\bar{X}} \left(1 + \frac{\bar{x} - \bar{X}}{\bar{X}}\right)^{-1}$$

$$= \frac{\bar{y} - R\bar{x}}{\bar{X}} \left(1 - \frac{\bar{x} - \bar{X}}{\bar{X}} + \frac{(\bar{x} - \bar{X})^2}{\bar{X}^2} - \ldots\right)$$

Cochran (2) used the above expression to find the leading term in the bias, which is

$$\frac{(1 - n/N)}{n \bar{X}^2} \left(RV(x) - C(x,y)\right)$$

where

$V(x)$ is the population variance of $x$

$C(x,y)$ is the population covariance of $x$ and $y$.

These results were sometimes used to obtain checks on the size of the bias in a specific sample by substituting sample values, but until 1951, no serious thought was given to finding unbiased estimators of the ratio-type.

### 3.2. Koop's Estimator

Koop (17) in 1951, found Taylor's theorem to be an unsatisfactory method of expansion to find the bias of the simple ratio estimator since it uses the fact that $R_1$ is differentiable near $(\bar{Y}, \bar{X})$. Since $R_1$ is not continuous, it is therefore not differentiable. Koop (17) obtained an expression for the bias by using a binomial series expansion, then substituted sample values in the expression for the bias, reducing it to any desired degree. The following estimator due to Koop is unbiased to order $1/n^4$.

$$R_2 = \bar{y}/\bar{x} - 1/n \left(\frac{s^2(x)}{\bar{x}^2} - \frac{s_{11}(x,y)}{\bar{x}\,\bar{y}}\right) \left(\frac{N-n}{N-1}\right) - 1/n^2 \left(\frac{s_{12}(y,x)}{\bar{y}\,\bar{x}^2} - \frac{s_{03}(\bar{y},x)}{\bar{x}^3}\right) \frac{(n-n)(N-2n)}{(N-1)(N-2)}$$

$$- \frac{3(n-1)}{n^3} \left(\frac{(s^2(x))^2}{\bar{x}^2} - \frac{s_{11}(y,x)s^2(x)}{\bar{y}\,\bar{x}^3}\right) \frac{N(N-n)(N-n-1)}{(n-1)(N-2)(N-3)}$$

$$- 1/n^3 \left(\frac{s_{04}(y,x)}{\bar{x}^4} - \frac{s_{13}(y,x)}{\bar{y}\,\bar{x}^3}\right) \frac{(N-n)(N^2-6Nn + N + 6n^2)}{(N-1)(N-2)(N-3)}$$

where

$$s^2(x) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$

$$s^2(y) = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}$$

$$s_{ij}(y,x) = \frac{\sum_{k=1}^{n} (y_k - \bar{y})^i (x_k - \bar{x})^j}{n-1}$$

This formula was admitted by Koop to be a clumsey and crude method having possibly large sampling errors. However, the method used to obtain this estimate is of theoretical interest. Koops procedure was as follows:

$$\bar{y}/\bar{x} = \sum_{1}^{n} y_i \; / \sum_{1}^{n} x_i$$

$$= \left(\sum_{1}^{N} y_i - \sum_{1}^{N-n} y_k\right) / \left(\sum_{1}^{N} x_i - \sum_{1}^{N-n} y_k\right)$$

$$= \frac{N\bar{X} - (N-n)\bar{y}'}{N\bar{X} - (N-n)\bar{x}'}$$

where

$$\bar{y}' = \frac{\sum_{k=1}^{N-n} y_k}{N-n}$$

$$= \bar{Y}/\bar{X} \left(1 - \frac{N-n}{N} \frac{\bar{y}'}{Y}\right) \left(1 - \left(\frac{N-n}{N}\right) \frac{\bar{x}'}{\bar{X}}\right)^{-1}$$

Koop (17) states the conditions that must be satisfied to expand $\left(1 - \left(\frac{N-n}{N}\right) \frac{\bar{x}'}{\bar{X}}\right)^{-1}$ as a binomial series and shows the conditions are satisfied. For an exact proof, see Koop (17). He mentions another method for finding the bias which involves writing

$$\bar{y}/\bar{x} = \bar{y}/\bar{x} \left(1 + \frac{\bar{y} - \bar{X}}{\bar{y}}\right) \left(1 + \frac{\bar{x} - \bar{X}}{\bar{X}}\right)^{-1}$$

and finding its expected value by the expansion of the last term by a binomial series. This expansion resulted in the same expression for the bias as the previous method.

### 3.3. Quenouille's Estimator

Quenouille (25) in 1956, developed a method for reducing bias in a large class of estimators. He considered the general problem of estimating an unknown parameter T, from a function $t_n(x_1, x_2, \ldots, x_n)$ of a series of observations taken in random order, the estimator can often be written as a function of the unbiased estimates of the cumulants, $k_1, k_2, \ldots, k_m$. Quenouille noted that the moments of the estimates of the cumulants are power series in $1/n$ and therefore the bias in $t_n$ could be expressed as a power series in $1/n$, if the following conditions hold:

(1) m is independent of n

(2) $t_n$ can be expanded by a Taylor's series

(3) $t_n$ is consistent

If the above conditions hold, then

$$E(\text{bias}) = a_{1/n} + a_{2/n^2} + \cdots .$$

If one considers an estimator $t'_n$,

where

$$t'_n = nt_n - (n-1)t_{n-1},$$

then

$$E(t'_n) = T - a_{2/n^2} - \frac{a_2 + a_3}{n^3} - \cdots$$

and therefore $t'_n$ is unbiased to order $1/n^2$. See Quenouille (25) for proof of this. Also $t''_n$,

where

$$t''_n = \frac{n^2 t'_n - (n-1)^2 t'_{n-1}}{n^2 - (n-1)^2}$$

is biased to order $1/n^3$ only, and so on. He also stated that any subset of the observations may be used to correct for bias. Another result was the estimator $t'_{2p}$

$$t'_{2p} = 2t_{2p} - t_p$$

which is free from bias to order $1/n^2$.

Quenouille worried somewhat about loss of efficiency in a procedure like this, but stated that if the average of all possible sets of $n-1$ observations, $\bar{t}_{n-1}$, is used in place of $t_{n-1}$, little loss of efficiency should result.

Durbin (6) in 1959, applied Quenouille's findings to ratio estimators, finding that if the regression of $y$ on $x$ is linear and $x$ is normal, that Quenouille's device actually decreased the variance. The estimate Durbin considered was

$$R_3 = 2R_1 - 1/2(R_{11} + R_{12})$$

where

$R_1$ is the simple ratio estimate from a sample of size $n$

$R_{11}$, $R_{12}$ are the simple ratio estimates from the two halves of the sample.

The following example from Deming (4) illustrates its use.

| Characteristic | Sample 1 | Sample 2 | Both |
|---|---|---|---|
| Total Rent | \$2720 | \$2350 | \$5070 |
| Total number of delinquents | 33 | 31 | 64 |
| Average rent | \$82.42 | \$75.81 | \$79.22 |

$$R_3 = 2(79.22 - 1/2\ (82.42 + 75.81)$$

$$= \$79.33\ .$$

Durbin (7) also considered the case where $x$ has a gamma distribution and found that, although the variance is increased by using $R_3$, the mean square error is decreased. For proofs of these cases, see Durbin (6).

Kish, Namboodiri, and Pillai (15) also look at Quenouille's results and were dissatisfied with it, saying the degree of reduction in bias didn't warrent the increased cost in computation, and that there were no practical methods for estimating its variance.

The general form for Quenouille's method as applied to ratio estimates was discussed by Rao (32). This form is

$$R_4 = gR_1 - \frac{g-1}{g} \sum_{j=1}^{g} R_{1j}$$

where

$R_1$ is the usual biased ratio estimator,

$R_{ij}$ is the usual ratio estimator omitting the j-th group,

$g$ is the number of groups of equal size into which the sample of size $n$ is split.

This form with $g=2$, reduces to the estimator $R_3$ considered by Durbin. Rao, assuming the regression of $y$ on $x$ was linear and that $x$ was normally distributed, found the variance of $R_4$ for general $g$ to order $n^{-3}$. He showed that both the bias and the variance of $R_4$ were decreasing functions of $g$, and therefore the optimum choice for $g$ would be $n$. The estimate

$$R_5 = nR_1 - \frac{n-1}{n} \sum_{j=1}^{n} R_{1j}$$

may be preferred to others.

$R_{1j}$ is the estimate obtained by omitting the j-th observation.

Tin (38) compares Quenouille-based estimators with others, discussed later in this paper, but also considers two extensions. One extension led to the same result previously considered by Rao,

$$R_4 = gR_1 - \frac{g-1}{g} \sum_{j=1}^{g} R_{1j} .$$

He states that as $g$ is increased, the variance becomes smaller as Rao proves assuming normality, but Tin also says it becomes more biased. He also states a condition for the efficiency of $R_4$ to be greater than the efficiency of $R_1$.

For

$$n > 12 \left(\frac{k_{20}}{\bar{X}^2}\right)$$

where

$k_{ij}$ is the $ij$ cumulant of $x$ and $y$. $R_4$ is less biased and more efficient than $R_1$ if $n$ is chosen between 2 and $n\, k_{20}/\bar{X}^2$. For a discussion of this, see Tin (38).

Tin's other extension was to divide the sample into two halves; and then divide each of these further in two halves. He then obtains the estimator

$$R_6 = 8/3\, R_1 - (R_{11} + R_{12}) + 1/12\, (R_{111} + R_{112} + R_{121} + R_{122})$$

where

$R_1$ is the usual ratio estimate

$R_{1j}$ is the usual ratio estimate calculated from the j-th half of the sample

$R_{1jk}$ is the usual ratio estimate calculated from the k-th half of the j-th half of the sample.

This was shown by Tin to be less biased, but also less efficient than both the simple ratio estimator and Durbins estimator. As Cochran (2) mentions, these estimators derived from Quenouille's general method can not

be expected to be of help when small samples are taken within strata, of course this is when an unbiased or reduced bias estimator would be of the most help. These estimators are useful however in another respect, when taking only moderate samples from a population having wide variation in the x variate.

### 3.4. Beale's Estimator

Beale (1) derived an asymptotic expansion for both the bias and the variance of the simple ratio estimator in terms of the coefficient of variation. Using this he obtained the following estimator

$$R_7 = R_1 \frac{1 + (\frac{1}{n} - \frac{1}{N}) \frac{S(x,y)}{\bar{x}\,\bar{y}}}{1 + (\frac{1}{n} - \frac{1}{N}) \frac{S^2(x)}{\bar{x}^2}}$$

where

$$R_1 = \bar{y}/\bar{x},$$

$S(x,y)$ = sample covariance,

$S^2(x)$ = sample variance of x.

This estimator removes the leading term in the bias and also decreases its asymptotic variance. Beale also mentioned that the extra cost is negligible if one wanted to estimate the variance, since the above quantities are needed for this. This appeared to Tin (38), to be one of the better ratio-type estimators, from the standpoint of degree of bias and efficiency.

Tin (38), in an effort to reduce the bias in the simple ratio estimate, developed the following estimator

$$R_8 = R_1 \left(1 + \left(\frac{1}{n} - \frac{1}{N}\right) \left(\frac{S(x,y)}{\bar{x}\,\bar{y}} - \frac{S^2(x)}{\bar{x}^2}\right)\right)$$

where the symbols are defined as in $R_7$. This has the same general form as Beale's estimator when neglecting terms of order $1/n^2$. This estimator, also less biased than the simple ratio estimator, is more efficient, a surprising result to Tin. He proved that this is not true since, by continually decreasing the bias, there is a point when the estimator starts becoming less efficient. Tin (38) also compared $R_7$ and $R_8$ above, with $R_3$, Durbin's application of Quenouille's method to ratio estimates, and $R_1$, the simple ratio estimate. Tin showed that Beale's estimator was the least biased, followed by Tin's modified ratio estimator which was less biased than Quenoville's method as applied by Durbin. A comparison between Durbins estimator and the usual estimator has already shown Durbin's to be superior in most cases. The variances were then compared and, to order $1/n^2$ or $1/n^3$, the modified ratio estimator $R_8$ was the most efficient followed by Beale's estimator and then Durbin's estimator. Tin also showed that there is little difference in their approach to normality in large samples, but for small samples (n=50) the modified ratio estimator appears to be the best in regard to bias, efficiency, and approach to normality, followed by Beale's, Durbin's, and the simple ratio estimator in that order.

Another modification of $R_7$ was obtained by Tin, by subtracting an estimate of the bias, to obtain a less biased but also less efficient estimator than $R_7$. The estimator was

$$R_9 = R_1 \left(1 + \left(\frac{1}{n} - \frac{1}{N}\right) \left(\frac{S(x,y)}{\bar{x}\,\bar{y}} - \frac{S^2(x)}{\bar{x}^2}\right) \left\{1 - 3\left(\frac{1}{n} - \frac{1}{N}\right) \left(\frac{S^2(x)}{\bar{x}^2}\right)\right\}\right)$$

where the symbols are as defined previously.

An estimate of the variance of $R_1$, $R_3$, $R_7$, or $R_8$ to order 1/n, supplied by Tin (37) is

$$(\frac{1}{n} - \frac{1}{N}) R_1^2 (\frac{S^2(x)}{\bar{x}^2} + \frac{S^2(y)}{\bar{y}^2} - 2 \frac{S(x,y)}{\bar{x}\,\bar{y}})$$

which does not involve much extra computation, since $S^2(x)$ and $S(x,y)$ are needed in the estimates, $R_7$ and $R_8$.

### 3.5. Jones Method For Correction of Bias

Jones (14) wrote about a graphic procedure used by Tukey to get an estimate of the bias and correct for it by using replicated samples. Since the bias contains the factor 1/n, it is obvious that as the sample size increases the bias decreases rapidly. If it is inconvenient in some way, or costly to take large samples, one may use the following procedure to get an estimate of the ratio one would obtain by increasing indefinitely the size of the sample. The procedure is as follows. Divide the sample into g subsamples, calculate the simple ratio estimator for each of the g subsamples, and average them. Next combine the g subsamples in equal groups of size $m_i$ obtaining $g/m_i$ groups for each choice of $m_i$. Find the average of the simple ratio estimator calculated for each of the $g/m_i$ groups. To illustrate this part of the procedure, let us consider the case g=10. Here the possible choices for $m_i$, are $m_1 = 2$, $m_2 = 5$, $m_3 = 10$, yielding 5,2, and 1 groups respectively. This gives i+1 average ratios. The second step is to plot these on coordinate paper against the number of subsample estimates used to compute the average. For g=10, the averages would be plotted 10,5,2, and 1 unit away; respectively, where the length of

the unit is immaterial. The third step is to draw the line of best fit. Extrapolation to zero gives a quick estimate, $R_{10}$, of the ratio one would obtain by increasing indefinitely the size of the sample. This procedure should also be useful when relationship between the bias and the reciprocal of the sample size is not linear. An example of the use of this process follows.

A sample of size 50 was taken from a population with $\bar{Y}=40$, $\bar{X}=80$. The sample was randomly divided into 10 subgroups. $\bar{y}/\bar{x}$ was computed for each of the 10 subgroups and their average found. The average for 5,2, and 1 subgroups were also found by combining the 10 subgroups. The following results were obtained.

Table 1. Sample Data for Jone's Graphical Method

| | | | | |
|---|---|---|---|---|
| Average for | 10 | groups | - | .5048 |
| " | " | 5 " | - | .4980 |
| " | " | 2 " | - | .5009 |
| " | " | 1 " | - | .4996 |

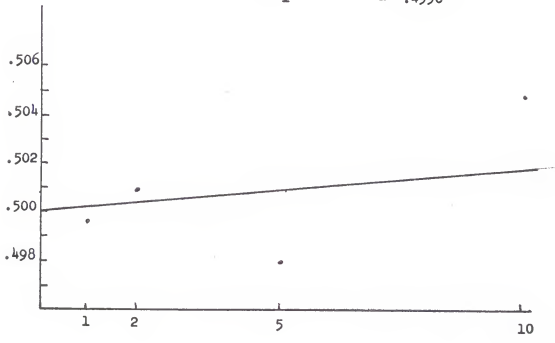


Fig. 1. Illustration of Jone's Method

The resulting estimate of the ratio one would obtain by increasing the sample size indefinitely is .5000.

### 3.6. Murthey and Nanjamma's Estimator

Murthey and Nanjamma (21) developed a technique to estimate the bias of the simple ratio estimate. This was used to obtain an almost unbiased estimate by using a correction factor. The simple ratio estimate is

$$R_1 = \bar{y}/\bar{x} .$$

Another biased estimate often used is the average of the sum of ratios $y_i/x_i$,

$$R_n = \frac{1}{n} \sum_{i=1}^{n} y_i/x_i .$$

This latter estimator is often used when a ratio estimator seems appropriate, but the variance of $y$ doesn't increase linearly with $x$.

Murthey and Nanjamma (21), using a series expansion and neglecting terms of degree greater than two, expressed the bias of $R_1$ as

$$B_1 = \frac{1}{\bar{x}^2} \left(RV(\bar{x}) - C(\bar{x},\bar{y})\right) ,$$

and the bias of $R_n$ as

$$B_n = \frac{1}{n} \sum_{i=1}^{n} B(y_i/x_i),$$

where

$B(y_i/x_i)$ is the bias of $y_i/x_i$,

$V(\bar{x})$ is the population variance of $\bar{x}$,

$C(\bar{x},\bar{y})$ is the population covariance of $\bar{x}$ and $\bar{y}$.

The quantity $B_1$ can be written,

$$B_1 = \frac{1}{x^2}\left(RV\left(\frac{\sum_{i=1}^{n} x_i}{n}\right) - C\left(\frac{\sum_{i=1}^{n} x_i}{n}\right)\right) = \frac{1}{n^2}\sum_{i=1}^{n} B(y_i/x_i) .$$

Therefore to the second degree of approximation

$$B_n = n\,B_1,$$

and

$$E(R_n - R_1) = B_n - B_1 = (n-1)B_1 .$$

So an unbiased estimate of the bias of $R_1$ to the second degree of approximation is

$$\frac{R_n - R_1}{n-1} .$$

This is used to correct $R_1$ for its bias obtaining,

$$R_{11} = \frac{n\,R_1 - R_n}{n-1} .$$

Another estimator, unbiased to the third degree of approximation is

$$R_{11A} = \frac{2n\,R_1}{n-1} - \frac{n\,R_2}{n-2} + \frac{2\,R_n}{(n-1)(n-2)}$$

where the sample was split into two parts and

$$R_2 = \bar{y}_1/\bar{x}_1 + \bar{y}_2/\bar{x}_2 ,$$

$$R_1 = \bar{y}/\bar{x} ,$$

$$R_n = \frac{1}{n}\sum y_i/x_i .$$

Unbiased Ratio-Type Estimator

The fact that the standard ratio estimators used are biased estimators has led to the exploration and development of unbiased ratio-type estimators. These estimators, though having the desirable properties of a ratio estimator are unbiased. Research in this field can be classified into two broad categories. The first, the development of an unbiased estimator through the use of commonly used sampling schemes, has been explored by Hartley and Ross (1954), Robson (1957), Goodman and Hartley (1958), Mickey (1959), Robson and Vithayasai (1961), and Williams (1961), among others. The second class of development was concerned with developing and modifying certain sampling schemes, so that under these schemes, the usual ratio estimator becomes unbiased. Major contributions here have been Lahiri (1951), Midyuno (1952), Horvety and Thompson (1952), Raj (1954), Mickey (1959), Nanjamma, Murthey, and Sethi (1960), Williams (1961), and Pathak (1964). Both of these classes will be reviewed in this report with some comparisons between these and the previously mentioned reduced-bias estimators.

## 4. THE UNBIASED ESTIMATOR (COMMONLY USED SAMPLING SCHEMES)

### 4.1. Hartley and Ross's Estimator

The first developments in unbiased-ratio-type estimators employing commonly used sampling schemes were by Hartley and Ross (12) in 1954. In brief they considered

$$R_n = \frac{1}{n} \sum y_i/x_i \ ,$$

one of the standard biased estimators, and connected it for bias by examining the population covariance of $y/x$ and $x$.

$$\text{Cov } (y/x,x) = E(y/x \cdot x) - E(y/x)E(x)$$

and so

$$E(y/x) = \frac{E(y)}{E(x)} - \frac{\text{Cov } (y/x,x)}{E(x)}$$

$$= \bar{Y}/\bar{X} - \frac{1}{\bar{X}} \text{ Cov } (y/x,x) \ .$$

Since

$$E(R_n) = E(y/x)$$

the bias in $R_n$ is given by $-\frac{1}{\bar{X}}$ Cov $(y/x,x)$, an exact expression. An unbiased estimate of this covariance is

$$\frac{N-1}{N(n-1)} \sum_{i=1}^{n} (r_i - \bar{r})(x_i - \bar{x}) = \frac{(N-1)n}{N(n-1)} \ (\bar{y} - \bar{r}\ \bar{x})$$

where

$$r_i = y_i/x_i$$

$R_n$ connected for bias becomes

$$R_{12} = R_n + \frac{n(N-1)}{(n-1)N\ \bar{X}} \ (\bar{y} - \bar{r}\ \bar{x}) \ .$$

Hartley and Ross (12) gave an approximate variance, for large samples, as

$$V(R_{12}) \doteq \frac{1}{n} \ (V(y) + R^2 v(x) - 2RC \ (x,y))$$

where

$V(y)$ = population variance of $y$,

$v(x)$ = population standard deviation of $x$,

$C(s,y)$ = population covariance of $x$ and $y$,

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

They state that this is also the approximate variance of $R_n$ if terms up to and including the quadratic are considered. Therefore they conclude that while the bias is eliminated, the variance has not increased to any degree. They also state that similar results for bias elimination in $R_1$ may also be applied. If this is done, we obtain

$$R_{13} = R_1 - \frac{\hat{C}(R_1,\bar{x})}{\bar{X}}$$

An exact formula for the variance of $R_{12}$ is given for any size sample by Goodman and Hartley (7) if the finite population correction may be omitted, as

$$V(R_{12}) = \frac{1}{n\bar{x}^2}\left(V(y) + R_p^2 V(x) - 2R_p C(x,y) + \frac{1}{n-1}(V(r)V(x) + C(r,x)\right)$$

where

$R_p$ is the population mean of the $R_i$'s ,

$V(r)$ is the population variance of the $r_i$'s ,

$C(r,x)$ is the population covariance of $r_i$ and $x_i$ .

An exact formula obtained through using multivariate polykays was obtained by Robson (34).

Goodman and Hartley develop an extremely cumbersome formula for an unbiased estimate of the population variance, (see Goodman and Hartley (7)). In the same paper they developed a much simpler, also unbiased but with

larger sampling error, estimate of the population variance by modifying the sampling scheme. The procedure is as follows. First draw a random sample of m pairs $(x_i, y_i)$ without replacement, then replace the sample and draw another sample of m pairs. This method makes the two sample independent, whereas the random splitting of a sample of size n=2m, into two halves will not. If the two samples are identical, reject the second and draw another. If $n << N$, the two samples will usually have no elements in common. The estimator

$$\bar{Y}_{R_{14}} = \frac{1}{2}\left(\bar{Y}_{R_{14_1}} + \bar{Y}_{R_{14_2}}\right)$$

is an unbiased estimator of $\bar{Y}$ and an unbiased estimate of the variance of $\bar{Y}_{R_{14}}$ is

$$s^2(\bar{Y}_{R_{14}}) = \frac{1}{4}\left(\bar{Y}_{R_{14_1}} - \bar{Y}_{R_{14_2}}\right)^2 \frac{\binom{N}{m} - 2}{\binom{N}{m}}$$

where

$$\bar{Y}_{R_{14_i}} = \bar{X}\bar{r}_i + \left(\frac{N-1}{N}\right)\left(\frac{m}{m-1}\right)\left(\bar{y}_i - \bar{r}_i\,\bar{x}_i\right)$$

and $\bar{x}_i$, $\bar{r}_i$, $\bar{y}_i$ are the sample means from the i-th sample. The unbiased estimate of the variance is based on only one degree of freedom, and if more degrees of freedom are desired, k samples of size n/k could be drawn. In stratified sampling the disadvantage of the one degree of freedom is eliminated to a certain degree. The following example illustrates the use of this method. In this example N=400 and $\bar{X}$=2. Two samples of size n=2 were drawn.

| 1st Sample | | | 2nd Sample | | |
|---|---|---|---|---|---|
| x | y | r | x | y | r |
| 1 | 3 | 3 | 4 | 8 | 2 |
| 2 | 6 | 3 | 1 | 2 | 2 |
| $\bar{x}_1 = 1.5$ | $\bar{y}_1 = 4.5$ | $\bar{r}_1 = 3$ | $\bar{x}_2 = 2.5$ | $\bar{y}_2 = 5$ | $\bar{r}_2 = 2$ |

In an example like this, the finite population corrections $(\frac{N-1}{N})$ and $(\frac{N}{m}) - 2/(\frac{N}{m})$ can usually be replaced by 1.

$$\bar{Y}_{R_{14_1}} = 2(3) + 2(4.5 - 3(1.5)) = 6$$

$$\bar{Y}_{R_{14_2}} = 2(2) + 2(5 - 2(2.5)) = 4$$

$$\bar{Y}_{R_{14}} = \frac{6+4}{2} = 5$$

$$S^2(\bar{Y}_{R_{14}}) = \frac{1}{4}(6-4)^2 = 1 \quad .$$

This example was due to Goodman and Hartley (7).

Goodman and Hartley state that in large samples, where the approximate formula for $V(R_n)$ is applicable, $V(R_n)$ will be smaller than $V(R_{12})$ in most cases. Raj (30) showed that present comparisons are not valid for small samples since the approximate variance formula definitely understates the true variance. If $\bar{x}$ were symmetrical the understatement as a proportion of the approximate variance exceeds three times the relative variance of $\bar{x}$ with a higher underestimation if the distribution of $\bar{x}$ is negatively skewed.

Goodman and Hartley point out a special case where the variance of the unbiased estimator is always smaller than the usual one. This is when the conditional variance of $r$ given $x$ is decreasing with $x$, i.e., the array variance of $r$ decreases with increasing $x$ in the scatter diagram $(x,r)$. For this kind of data, the unbiased ratio estimator proposed by Hartley and Ross (12) is better than the simple ratio method.

Olkin (23) estended Hartley and Ross's estimator to the case where multi-auxiliary variables are used to increase precision. Considering the case of $p$ such auxiliary variables $x_1, x_2, \ldots, x_p$, Olkin developed the estimator

$$R_{15} = \sum_{i=1}^{p} w_i \bar{r}_i \bar{X}_i + \frac{(N-1)n}{N(n-1)} \left(\bar{y} - \sum_{i=1}^{p} w_i \bar{r}_i \bar{x}_i\right);$$

an unbiased estimator of $\bar{Y}$,

where

$$n\bar{r}_i = \sum_{i=1}^{n} y_i / x_{ij}$$

and $w_i$ is chosen to minimize the variance of $R_{15}$. Common choices of $w_i$ would be $1/x_{ij}$, if the variance of $y$ increases with the square of $x$, or 1 if the variance of $y$ appears to increase linearly with $x$. For a full discussion of optimum choices of weights, see Raj (31).

## 4.2 Robson's Estimator

Robson (35), in 1957, applied the results of multivariate polykays to obtain the previously mentioned exact variance formula for Hartley and Ross's unbiased estimator. He also obtained Hartley and Ross's estimator

by using multivariate polykays. For a discussion of this see Robson (35). In this same paper, Robson adjusted another standard biased ratio estimator $\frac{\bar{x}\,\bar{y}}{\bar{X}^2}$ which has greater precision than $R_1$ or $R_n$ if the correlation is negative between $x$ and $y$, to obtain a corresponding unbiased estimator. The bias of $\frac{\bar{x}\,\bar{y}}{\bar{X}}$, an estimate of $\bar{Y}$ is

$$E\left(\frac{\bar{x}\,\bar{y}}{\bar{X}} - \bar{Y}\right) = \frac{1}{\bar{X}}\left(E(\bar{x}\,\bar{y}) - \bar{X}\,\bar{Y}\right)$$

$$= \frac{1}{\bar{X}}\,\mathrm{Cov}\,(\bar{x}, \bar{y})$$

Therefore, an adjusted unbiased estimator of the ratio is

$$R_{16} = \frac{\bar{x}\,\bar{y}}{\bar{X}^2} - \frac{1}{\bar{X}^2} \cdot \frac{(N-n)}{nN(n-1)} \sum_{i=1}^{n} (x_i - \bar{x})\, y_i$$

or

$$R_{16} = \frac{1}{\bar{X}^2}\left(\frac{n(N-1)}{N(n-1)}\,\bar{x}\,\bar{y} - \frac{N-n}{N(n-1)}\,\frac{\sum_{i=1}^{n} x_i\, y_i}{n}\right) .$$

Again using multivariate polykays, Robson (35) found for an unbiased estimate of the variance of $R_{16}$, as the sample size becomes large,

$$s^2(R_{16}) = \frac{\bar{y}^2}{n\,\bar{X}^2}\left(\frac{s^2(y)}{\bar{y}^2} + \frac{s^2(x)}{\bar{x}^2} + 2\,\frac{s(x,y)}{\bar{x}\,\bar{y}} + \frac{1}{n-1}\left(\frac{s^2(x)\,s^2(y) + \left(s(x,y)\right)^2}{\bar{x}^2\,\bar{y}^2}\right)\right) .$$

This was obtained by substituting the above sample estimates for population values in the population variance.

## 4.3. Mickey's Estimator

Mickey (19) developed a method for producing a broad class of unbiased ratio-type estimators, by using the fact that $\bar{y} - a(\bar{x}-\bar{X})$ is an unbiased estimator of $\bar{Y}$ for any choice of a. He also used the fact that for any choice m of the n sampling units, the n-m remaining units can be considered a random sample of n-m from the N-m units derived by omitting the m given units. Mickey then chooses a as a function of the m selected units and uses $\bar{y} - a(\bar{x}-\bar{X})$ to get an unbiased estimate of the population of N-m units which leads to an unbiased estimate for the whole population by utilizing the relationship between the two populations determined by m, N, and the m selected units. Since $\bar{y}$ is a biased estimator, $a(\bar{x}-\bar{X})$ is an estimate of the bias obtained by using the form of the biased estimator to the subsample in estimating the sample mean, $\bar{y}$. Mickey uses the following formula to generate his estimators.

$$R_m = a(Z_m)\bar{X} + \frac{N-m}{N(n-m)}\left(Y(n) - a(Z_m)X(n)\right) - \frac{N-n}{N(n-m)}\left(Y(m) - a(Z_m)X(m)\right)$$

where

$Z_m$ is the ordered set of observations on the first m sample elements

$1 \leq m < n$, $a(Z_m)$ is a function of these observations to be determined, X(m), Y(m) are the sums of the first m sample elements, X(n), Y(n) are the sample totals. Particular estimators are generated by the choice of $a(Z_m)$, and a general class of estimators is constructed by including all estimators of the form above applied to any permutation of the ordering of the sample, weighted averages of such estimators, and estimators obtained from subsamples of the given sample. A knowledge of the population one is sampling from

helps in choosing functions. When the variance in $y$ increases as the square of $x$, Mickey's techniques lead to the estimator, $R_{12}$, Hartley and Ross's estimator. When the variance increases linearly with $x$, Mickey's estimator is

$$R_{17} = \frac{\bar{y}(m)}{\bar{x}(m)} + \frac{(N-m)n}{N(n-m)\bar{X}} \left(\bar{y} - \frac{\bar{y}(m)}{\bar{x}(m)} \cdot \bar{x}\right)$$

where

$\bar{y}(m)$, $\bar{x}(m)$ are sample means of the first $m$ observations. For $m=n-1$, $R_{17}$ becomes

$$R_{18} = \bar{R}_{n-1} + \frac{(N-n+1)n}{N(\bar{X})} \left(\bar{y} - \bar{R}_{n-1}\bar{x}\right)$$

where

$$\bar{R}_{n-1} = \frac{1}{n} \sum_{j=1}^{n} \frac{n\bar{y} - y_j}{n\bar{x} - x_j} .$$

Mickey goes on to develop another estimator for which he also develops an easy formula to estimate its variance. Let $R(m,n)$ denote an estimator $R_m$ based on a sample of size $n$. Suppose also there are $k+1$ integers $0 < m_1 < \ldots m_{k+1} = n$, and consider the $k$ estimators

$R(m_1,m_2)$, $R(m_2,m_3)$, ..., $R(m_k,n)$. The estimator Mickey developed was

$$R_{19} = \frac{1}{k} \sum_{j=1}^{k} R(m_j, m_{j+1}) .$$

He states an unbiased, non-negative estimator of the variance is

$$S^2(R_{19}) = \frac{1}{k(k-1)} \sum_{j=1}^{k} \left(R(m_j, m_{j+1}) - R_{19}\right)^2 .$$

There is a great deal of flexibility since the $R(m_j, m_{j+1})$ may be chosen as Hartley and Ross's estimator, $R_{17}$, $R_{18}$, or other similar estimators. The precision of $R_{19}$ could be improved by averaging with respect to a random sample or all possible orderings of the sample elements. To clarify the previous discussion two examples will be considered.

Example 1. The first example involves a table constructed by Cochran (2, table 6.1). He gives values of $x$ and $y$ for 49 cities, where $y$ is the number of inhabitants of a city in 1930 and $x$ is the corresponding number for 1920. The unit of count is 1000 individuals. A random sample of size 5 was selected and $R_{19}$ was calculated using

$$R(m_j, m_{j+1}) = R(i-1,i) = R(i-1) + \frac{N - i+1}{N} \left(Y(i) - R(i-1)X(i)\right)$$

and

$$R_{19} = \frac{1}{k} \sum_{j=1}^{k} R(j-1,j)$$

where

$$Y(i) = \sum_{j=1}^{i} y_j ,$$

$$X(i) = \sum_{j=1}^{i} x_j ,$$

$$R(i) = Y(i)/X(i) .$$

The five elements sampled in the order drawn were: (63,37), (58,50), (80,76), (53,45), and (113,121).

Table 2. Illustration of Computations for Estimator $R_{19}$

| i | Y(i) | X(i) | R(i-1) | Y(i)-R(i-1)X(i) | $\frac{N-i+1}{N}$ | R(i-1,i) |
|---|---|---|---|---|---|---|
| 1 | 63 | 37 | | | | |
| 2 | 121 | 87 | 1.7027 | -27.135 | .9796 | 1.4498 |
| 3 | 201 | 163 | 1.3908 | -25.700 | .9592 | 1.1518 |
| 4 | 254 | 208 | 1.2331 | -2.485 | .9388 | 1.2105 |
| 5 | 367 | 329 | 1.2212 | -34.775 | .9184 | .9116 |

$$R_{19} = \frac{1.4498 + 1.1518 + 1.2105 + .9116}{4} = 1.1809$$

$$S^2(R_{19}) = \frac{(1.4498)^2 + \ldots\ldots + (.9116)^2 - 4(1.1809)^2}{4(3)} = .0119923$$

$$S(R_{19}) = .1095$$

Example 2. This time the population is the entire 196 cities considered by Cochran and the sample is the 49 cities listed. Computation can be lessened by using $m_j$ equals some number larger than 1. Choosing k=4 as in the previous example, let $m_1$=5, $m_2$=19, $m_3$=31, $m_4$=41, $m_5$=49. These are strictly arbitrary.

Table 3. Illustration of Computations for Estimator $R_{19}$

| i | $m_i$ | $Y(m_i)$ | $X(m_i)$ | $\frac{Y(m_i)-R(m_{i-1})X(m_i)}{R(m_{i-1})}$ | | $\frac{N-m_{i-1}}{N(m_i-m_{i-1})}$ | $R(m_{i-1}, m_i)$ |
|---|---|---|---|---|---|---|---|
| 1 | 5 | 804 | 691 | | | | |
| 2 | 19 | 3103 | 2522 | 1.163531 | 168.574818 | .096606 | 1.49478 |
| 3 | 31 | 4154 | 3368 | 1.2303736 | 10.103736 | .075255 | 1.23687 |
| 4 | 41 | 5334 | 4306 | 1.233373 | 23.095862 | .084184 | 1.25000 |
| 5 | 49 | 6262 | 5054 | 1.238737 | 1.423302 | .098852 | 1.23994 |

$$R_{19} = 1.305397$$

$$s^2(R_{19}) = \frac{(1.49478)^2 + \ldots + (1.23994)^2 - 4(1.305397)}{4(3)} = .136881$$

### 4.4 Robson and Vithayasai's Estimator

Robson and Vithayasai (36) develop a more efficient estimator for certain types of populations by using Hartley and Ross's correction for bias. The type of population under consideration was when $x$ and $y$ could be expressed as the sum of $k$ corresponding components, and when the components were more highly correlated than $x$ and $y$. In this case a componentwise ratio estimator such as

$$\sum_{j=1}^{k} \bar{y}_j / \bar{x}_j$$

is generally more efficient, although it is biased. By using Hartley and Ross's estimator, Robson and Vithayasai obtained an unbiased componentwise ratio-type estimator

$$R_{20} = \sum_{j=1}^{k} \bar{r}_j + \frac{n_j(N_j-1)}{(n_j-1)\bar{x}_j}(\bar{y}_j - \bar{x}_j \bar{r}_j)$$

where

$\bar{r}_j$, $\bar{x}_j$, $\bar{y}_j$ are the means of the k components,

$N_j$ is the population size of the j-th components,

$\bar{X}_j$ is the population mean of the j-th component.

An example for its use from general sample survey theory is the case of cluster sampling with post stratification, x representing the number of elements in a cluster and y the cluster total for some measured character. If the x elements in a randomly chosen cluster are partitioned into k strata of size $x_j$, $X_j$ known, then the above estimator may be much more efficient than the non-stratified estimator.

## 4.5. Willimas' Estimators

Williams (39) considered the generation of some unbiased ratio and regression estimators, differentiating between the two as follows. He classified an estimator as a regression type if it was invariant under location and scale changes in x and if it underwent the same location and scale changes in y. He classified an estimator as a ratio type if the above properties hold for scale changes only.

The following procedure was considered by Williams. First he selected with equal probability one of all possible splits of the population into s groups of size n/k, N = Sn/k. Second he selected at random without replacement k of the groups from the s groups of that split, yielding a sample of size n. Williams considered the conditional distribution for a particular set of s groups, eventually deriving the unconditionally unbiased estimate of R

$$R_{21} = \frac{1}{\bar{x}} \left[ \bar{y} + \bar{b} \, (\bar{X}-\bar{x}) + \left(1 - \frac{n}{N}\right) \frac{1}{k(k-1)} \sum_{i=1}^{k} (b_i-\bar{b})(\bar{x}_i-\bar{x}) \right.$$

where

$\bar{x}_i$ is the mean of the n/k units in the i-th group

$b_i$ is as yet unspecified function of the y and x of the i-th group, to make $R_{21}$ a ratio estimator.

$$\bar{b} = \sum_{i=1}^{k} b_i/k$$

This approach insures that $R_{21}$ is an unbiased estimator for any choice of the $b_i$.

In practice a sample of size n is taken and split randomly into groups. Williams states that this also preserves the unbiasedness of the estimator. For

$$b_i = \sum_{j=1}^{n/k} y_{ij}x_{ij} / \sum_{j=1}^{n/k} x_{ij}^2 ,$$

Williams gets

$$R_{22} = \frac{1}{\bar{X}} \left(\bar{y} + \frac{1}{k} \sum_{i=1}^{k} \left( \sum_{j=1}^{n/k} y_{ij}x_{ij} / \sum_{j=1}^{n/k} x_{ij}^2 \right) (\bar{X}-\bar{x}) + (1-\frac{n}{N}) \frac{1}{k(k-1)} \sum_{i=1}^{k} \left( \frac{\sum_{j=1}^{n/k} y_{ij}x_{ij}}{\sum_{j=1}^{n/k} x_{ij}^2} \right) \left( \sum_{j=1}^{n/k} x_{ij} \right) \right.$$

$$\left. - \frac{\bar{x}}{k} \sum_{i=1}^{k} \left( \frac{\sum_{j=1}^{n/k} y_{ij}x_{ij}}{\sum_{j=1}^{n/k} x_{ij}^2} \right) \right) .$$

For

$$b_i = \bar{y}_i/\bar{x}_i = r_i ,$$

$R_{21}$ becomes

$$R_{23} = \frac{\bar{r}\bar{x}}{\bar{X}} + \frac{1}{\bar{X}} \frac{N(k-n)}{N(k-1)} \quad (y-\bar{r}\bar{x})$$

When k=n, $R_{23}$ is identical to Hartley and Ross's unbiased ratio estimator.

For

$$b_i = r_i = k/n \sum_{j=1}^{n/k} r_{ij}$$

$$r_{ij} = y_{ij}/x_{ij}$$

$$\bar{b} = \bar{r} = \frac{1}{k} \sum_{i=1}^{k} r_i,$$

Williams again gets Hartley and Ross's estimator upon substitution into $R_{21}$, when averaged over all possible splits of the sample into groups of size n/k. For clarification of Williams estimator a simple example follows.

A simple random sample of four pairs $(y_i, x_i)$ were drawn from a population of size 100 with $\bar{X} = 2.0$. The sample was split randomly in 2 groups; (2,1) and (3,2) in the first group, (1,1) and (4,2) in the second. For $R_{22}$ we have the following

$$b_1 = \frac{2+6}{1+4} = 1.6$$

$$b_2 = \frac{1+8}{1+4} = 1.8$$

$$\sum_j x_{1j} = 3$$

$$\sum_j x_{2j} = 3$$

$$R_{22} = \frac{1}{2}\{(2.5+\frac{1}{2}(1.6+1.8)(2-1.5))+\frac{96}{100}\cdot\frac{1}{2}((1.6)3+1.8(3))-\frac{2.5}{2}(1.6+1.8)\}$$

$$= \frac{1}{2}\{(3.35) + \frac{48}{100}(10.2) - (2.5)(1.7)\}$$

$$= \frac{1}{2}(3.996) = 1.998 .$$

## 5. THE UNBIASED ESTIMATOR (MODIFICATION OF SAMPLING SCHEMES)

This section will be concerned with a presentation of various sampling schemes and modification of sampling schemes to make the ordinary simple ratio estimators unbiased. Theoretical results will be minimized to clarify the actual methods in the following section.

### 5.1. Lahiri's Methods

Lahiri (18) in 1951, showed if a sample was drawn with probability proportional to the sum of the $x$ elements in the sample, the ordinary ratio estimate $\bar{y}/\bar{x}$ was unbiased. An exact result would involve forming cumulative totals for all possible samples of size $n$, an almost impossible task in most cases. Lahiri then developed some procedures, which while yielding an unbiased estimator, involved procedures which greatly reduced the amount of work in sampling. The first was drawing a sample of size $n$, unit by unit, when the largest $x$ value is known. This involves sampling proportional to the $x$ values. To select the first unit in the sample, choose a random value between 0 and $x_{max.}$, the largest value. Now choose at random one of the units in the population. If it is greater than or equal to the random value chosen, retain it; if not, reject it. In either case a new random.

value is chosen, and a new unit is chosen from the population each time until a sample of the desired size is chosen. This process results in a sample of size n proportional to the x's. The unbiased estimator is

$$R_{24} = R_n = \frac{1}{n} \sum_{i=1}^{n} y_i/x_i \quad .$$

The variance of this estimator under this sampling scheme was given by Raj (27) to be

$$V(R_{24}) = \frac{1}{n\bar{X}} \sum x_i (y_i/x_i - R)^2$$

and estimated by

$$S^2(R_{24}) = \frac{1}{n(n-1)} \sum_{1}^{n} (y_i/x_i - R_{24})^2 \quad .$$

This sampling procedure can involve many rejections, which may be costly. To reduce the number of rejections, Lahiri considered several alternative schemes. The first involved using some large unit $x^1$ max. Now a unit is chosen, say $x_1$. If $x_1$ is larger than $x^1$ max, keep it and look at $x_1/x^1$ max. = Q+R where Q is an integer. The unit is listed 1+Q times, the first of size R, the rest of size $x^1$ max. An alternative device may be used if there are a small number of extraordinarily large sizes and it consists of dividing the population into two groups, one made up of the large units, the second, the remaining units. A set of three random numbers is utilized which:

(1) decides which group the selection is to be made from,

(2) fixes the unit which is to be accepted or rejected on the basis of three,

(3) chooses the random value between 0 and x max,

The second type of procedure Lahiri employed was to choose the entire sample with probability proportional to the sum of the observations of x in the sample, $\sum x_i$. His practical method was to:

(1) choose a set of n elements at random (with or without replacement) and find $\sum x_i$,

(2) choose a random value between 0 and $\sum x_i$ = say V,

(3 now choose another sample and if $\sum x_i$ for this sample is greater than or equal to V, keep it. If $\sum x_i$ is less than V, replace it and begin the process anew. Find another random number V and draw another sample, until the sample satisfies the criterion.

The estimator used by Lahiri in this case was

$$R_{25} = R_1 = \bar{y}/\bar{x} \quad .$$

Raj (27) in his investigation of Lahiri's procedure, derived the variance of $R_{25}$ as

$$V(R_{25}) = \frac{1}{\binom{N-1}{n-1}} \sum_j' \frac{\{(\sum y_i)_{s_j}\}^2}{(\sum x_i)_{s_j}} - R^2$$

where $\sum'$ denotes summation over all possible samples; $(\sum y_i)_{s_j}$, $(\sum x_i)_{s_j}$ are totals of the j-th sample. He also obtained an unbiased estimate of the variance as

$$s^2(R_{25}) = R_{25}^2 - \frac{\binom{N-1}{n-1}}{\sum x_i} \frac{\sum y_i^2}{\binom{N-1}{n-1}} + 2 \frac{\left(\sum_{j>i=1}^{n} y_i y_j\right)}{\binom{N-2}{n-2}}$$

## 5.2. Midzuno's Method

Midzuno (20) and Sen have independently given a simple procedure for obtaining a sample with probability proportional to size, thereby making the simple ratio estimate $\bar{y}/\bar{x}$ unbiased. Their method involved the following procedure

(1) Select the first unit in the sample with probability proportional to size as follows: Choose a random number between 0 and the largest x value, now choose a random x value. If it is greater than or equal to the random number, keep it; otherwise, start the procedure again.

(2) Select the rest of the sample with equal probability without replacement from the remaining units of the population.

The following proof showing that

$$R_{26} = \bar{y}/\bar{x}$$

is unbiased for this procedure is due to Cochran (2).

The probability that a sample of size $n$ with a fixed value of $\sum x_i$ is drawn is

$$P = \frac{\sum x_i}{\binom{N-1}{n-1} X}$$

since the total of $\sum x_i$ added over all simple random samples of size $n$ is $\binom{N-1}{n-1} X$.

For the estimator $\bar{y}/\bar{x} = \dfrac{\sum y_i}{\sum x_i}$

$$E(\bar{y}/\bar{x}) = \sum_{\text{all } S} (P) \left(\frac{\sum y_i}{\sum x_i}\right)$$

where $\sum_{\text{all } s}$ represents a summing over all possible simple random samples

$$E(\bar{y}/\bar{x}) = \sum_{\text{all } s} \frac{\sum x_i}{\binom{N-1}{n-1} X} \frac{\sum y_i}{\sum x_i}$$

$$= \frac{\binom{N-1}{n-1} Y}{\binom{N-1}{n-1} X} = \frac{Y}{X} ,$$

showing $\bar{y}/\bar{x}$ is unbiased for this method of selection.

An unbiased estimate of the variance of $R_{26}$ was given by Nanjamma, Murthey, and Sethi to be

$$S^2(R_{26}) = R_{26}^2 - \frac{\sum_{1}^{n} y_i^2 + 2 \frac{N-1}{n-1} \sum_{i>j}^{n} y_i y_j}{Nn \bar{x}\bar{X}}$$

They also state that the efficiency of the unbiased estimate will be greater than, or equal to, or less than correlation coefficient of $(\bar{y}^2/\bar{x}, x) \lessgtr 0$ .

### 5.3. Nanjamma, Murthey, and Sethi's Methods

Nanjamma, Murthey, and Sethi (22) in 1960, modified many of the selection procedures commonly used, equal probability sampling, varying probability sampling, stratified sampling, and multi-stage sampling to make the usual simple ratio estimator unbiased. The procedure is similar to other methods considered previously, that is, selecting one unit with probability proportional to size of the correlated x-variable and the remaining units according to the original scheme of sampling. Variance estimators were given by Nanjamma, Murthey and Sethi for some of the more important sampling schemes.

UNSTRATIFIED SAMPLING WITH EQUAL PROBABILITY AND WITH REPLACEMENT

(1) Select one unit with probability proportional to size of the x variate, using Lahiri's (18) or Midzuno's (20) method.

(2) Select the rest of the sample with equal probability with replacement.

Then

$$R_{27} = R_1 = \bar{y}/\bar{x}$$

is an unbiased estimate of $R = \bar{y}/\bar{x}$. The probability of getting a particular sample was shown by Nanjamma, Murthey and Sethi to be

$$P(S) = \frac{1}{N^n} \cdot \frac{n!}{\prod_{i=1}^{v} L_i!} \cdot \frac{\bar{x}}{\bar{X}}$$

where $L_i$ is the number of repititions of the i-th unit and v is the number of distinct units in the sample. The estimated variance of $R_{27}$ was given to be

$$s^2(R_{27}) = R_{27}^2 - \frac{\sum_{1}^{v} L_i(L_{i-1})y_i^2 + 2 \sum_{i>j}^{v} L_i L_j y_i y_j}{n(n-1)\bar{x}\bar{x}}$$

UNSTRATIFIED SAMPLING WITH EQUAL PROBABILITY SYSTEMATICALLY

Here the authors considered each unit as made up of n sub-unit of the i-th unit having the size $X_i/n$ where $X_i$ is the total of the i-th unit. Now a sub-unit is chosen with probability proportional to size of the x values. The others are then determined by proceeding to select the remainder of the sample systematically with the sub-unit selected first as the random start. The probability of a particular sample s, is

$$P(S) = \bar{x}/x$$

and an unbiased estimator of the population ratio is

$$R_{28} = \bar{y}/\bar{x} .$$

Nanjamma, Murthey, and Sethi state that it is impossible to get an unbiased estimate of the population variance from a single sample.

VARYING PROBABILITY SAMPLING PROBABILITY PROPORTIONAL TO SIZE WITH REPLACEMENT SCHEME

(1) Select first one unit with probability proportional to $x$ and replace it.

(2) Select the rest of the sample with probability proportional to $Z$ with replacement, where $Z$ is some measure of size under consideration.

An unbiased estimate of $R$ is then given by

$$R_{29} = \frac{1/n \sum_{i=1}^{n} y_i/p_i}{1/n \sum_{1}^{n} x_i/p_i}$$

where

$$P_i = Z_i / \sum_{i=1}^{n} Z_i$$

VARYING PROBABILITY SAMPLING PROBABILITY PROPORTIONAL TO SIZE WITHOUT REPLACEMENT SCHEME.

This is in general not a practical scheme since it involves very heavy computations but two special cases were considered by Nanjamma, Murthey, and Sethi. The first involves a sample of size two, the first element taken with probability proportional to x, the second probability proportional to Z. An unbiased ratio estimate is given as

$$R_{30} = \frac{y_1/p_1\ (1-p_2) + y_2/p_2\ (1-p_1)}{x_1/p_1\ (1-p_2) + x_2/p_2\ (1-p_1)} \ ,$$

where

$$P_1 = x_1/X \ ,$$

$$P_2 = Z_2/Z \ .$$

The second involved the first two steps above, and then drawing n-2 other elements with equal probability, thus obtaining a ratio estimate

$$\frac{\sum_{i=1}^{n} \frac{y_i}{(1-p_i)} \left( \sum_{j=i}^{n} p_j \right)}{\sum_{i=1}^{n} \left( \frac{x_i}{1-p_i} \right) \left( \sum_{j=i}^{n} p_j \right)}$$

which is unbiased.

## 6. CONCLUDING REMARKS

Since there has been little discussion in this report on extensions of the ratio estimators considered to sampling schemes other than simple random sampling, a brief list of the more important papers in certain areas of sampling follows. The interested reader is referred to these articles.

In the area of two-stage and multi-stage sampling, unbiased ratio-type estimators have been investigated by Nanjamma, Murthey, and Sethi (22), Pathak (25), Raj (27), Raj (28), Raj (29), Sukhatme (37), and Williams (39). Although many of the estimators can be directly applied to stratified sampling schemes, for a more extensive discussion of these techniques see Raj (27), and Williams (39). For a discussion of unbiased ratio-type estimation applied to systematic sampling, see Nanjamma, Murthey, and Sethi (22). Since this report has been concerned primarily with a single variate correlated with the variate of interest, the reader is referred to Olkin (23), Raj (31), and Williams (40) for use of multi-auxiliary information.

Although Tin (38) has made a fairly thorough comparative study of several of the reduced-bias estimators, there seems to be little available to the reader interested in a more extensive comparison involving the usual biased estimators, reduced-bias estimators, and both classes of unbiased ratio-type estimators. One of the major reasons is that some of the variance formulas involved are not known, and some are only large sample approximations. Exact expressions for variances are usually mathematically cumbersome and difficult to compare.

The following study involves three small populations (n=6) with samples of size (n=4) taken from each. All possible samples were taken from each population, so the bias and variance could be found exactly for each population.

Table 4. Computer Study One

Population 1. (0,2), (1,3), (2,5), (4,9), (8,14), (9,15); $\bar{X} = 8.0$

| Estimator | Bias | Variance | M.S.E. |
|---|---|---|---|
| $\bar{y}$ | 0.0000 | 1.1666 | 1.1666 |
| $R_1$ | 0.0627 | .1203 | .1242 |
| $R_n$ | 0.8673 | .2571 | 1.0033 |
| $R_7$ | 0.0195 | .1152 | .1156 |
| $R_8$ | 0.1168 | .1046 | .1182 |
| $R_{11}$ | 0.2056 | .1235 | .1657 |
| $R_{12}$ | 0.0000 | .2328 | .2328 |
| $R_{15B}$ | 0.2166 | 3.5129 | 3.5598 |
| $R_{15}$ | 0.0000 | 3.4364 | 3.4364 |
| $R_5$ | .3216 | .1217 | .2251 |
| $R_{17}, R_{19}$ | 0.0000 | .0711 | .0711 |
| $R_{22}$ | 0.0000 | .0220 | .0220 |

Table 5. Computer Study Two

Population 2. (5,1), (4,2), (4,5), (10,8), (12,11), (16,15); $\bar{X} = 7.0$

| Estimator | Bias | Variance | M.S.E. |
|---|---|---|---|
| $\bar{y}$ | 0.0000 | 2.0583 | 2.0583 |
| $R_1$ | 0.1327 | .5452 | .5628 |
| $R_n$ | 4.5755 | 10.2863 | 31.2215 |
| $R_7$ | 0.0248 | .4488 | .4549 |
| $R_8$ | 0.1798 | .2186 | .2509 |
| $R_{11}$ | 1.3482 | .3274 | 2.1450 |
| $R_{12}$ | 0.0000 | 2.5924 | 2.5924 |
| $R_{15B}$ | 0.3047 | 10.5797 | 10.6727 |
| $R_{15}$ | 0.0000 | 10.2666 | 10.2666 |

Table 6. Computer Study Three

Population 3. (0,0), (1,1), (4,2), (9,3), (16,4), (25,5): $\bar{X} = 2.5$

| Estimator | Bias | Variance | M.S.E. |
|---|---|---|---|
| $\bar{y}$ | 0.0000 | 7.9139 | 7.9139 |
| $R_1$ | 0.1893 | 1.6185 | 1.6543 |
| $R_n$ | 2.9166 | 1.8229 | 10.3294 |
| $R_7$ | .0783 | 1.6311 | 1.6372 |
| $R_8$ | .2764 | 1.7133 | 1.7897 |
| $R_{11}$ | .7191 | 2.0253 | 2.5424 |
| $R_{12}$ | 0.0000 | 2.9167 | 2.9167 |
| $R_{15B}$ | .5833 | 22.1271 | 22.4673 |
| $R_{15}$ | 0.0000 | 21.5059 | 21.5059 |
| $R_{22}$ | 0.0000 | 1.8860 | 1.8860 |

ACKNOWLEDGEMENT

I am indebted to Dr. A. M. Feyerherm for his advice in the preparation of this report and to William Cash for programming the computer survey.

REFERENCES

(1) Beale, E. M. L., "Some uses of computers in operational research," Industrielle Organisation, 31 (1962), 27-28.

(2) Cochran, W. G., Sampling Techniques, New York: Wiley, 1965.

(3) Das, A. C., "On two-phase sampling and sampling with varying probabilities," International Statistical Institute Bulletin, 33, II, 105-112.

(4) Deming, W. E., Some Theory of Sampling, New York: Wiley, 1950.

(5) Deming, W. E., Sample Design in Business Research, Canada: Wiley, 1960.

(6) Durbin, J., "A note on the application of Quenouille's method of bias reduction to the estimation of ratios," Biometrika, 46 (1959), 3 and 4, 477-480.

(7) Goodman, L. A., and Hartley, H. O., "The precision of unbiased ratio-type estimators," Journal of the American Statistical Association, 53 (1958), 491-508.

(8) Hansen, M. H., Hurwitz, W. N., and Gurney, M., "Problems and methods of the sample survey of business," Journal of the American Statistical Association, 41 (1946), 173-189.

(9) Hansen, M. H., Hurwitz, W. N., and Madow, W. G., Sample Survey Methods and Theory, Vol. I, II, New York: Wiley, 1953.

(10) Hansen, M. H. and Hurwitz, W. N., "On the theory of sampling from finite populations," Annals of Mathematical Statistics, 14, 333-362.

(11) Hartley, H. O., and Rao, J. N. K., "Sampling with unequal probabilities and without replacement," Annals of Mathematical Statistics, 33 (1962), 352, 370.

(12) Hartley, H. O., and Ross, A., "Unbiased ratio estimators," Nature, 174 (1954), 270-1.

(13) Horvitz, D. G., and Thompson, D. J., "A generalization of sampling without replacement from a finite universe," Journal of American Statistical Association, 47, 663-85.

(14) Jones, Howard L., "Investigating the properties of a sample mean by employing random subsample means," Journal of the American Statistical Association, 51 (1956), 77.

(15) Kish, L., Namboodiri, N. K., and Pillai, R. K., "The ratio bias in survey," Journal of the American Statistical Association, 57, 863-876.

(16) Kish, L., and Hess, I., "On variances of ratios and their differences in multi-stage samples," Journal of the American Statistical Association, 54 (1959), 416-446.

(17) Koop, J. C., "A note on the bias of the ratio estimate," International Statistical Institute Bulletin, 33 (1951), II, 141-6.

(18) Lahiri, D. B., "A method of sample selection providing unbiased ratio estimates," International Statistical Institute Bulletin, 33 (1951), II, 133-140.

(19) Mickey, M. R., "Some finite population unbiased ratio and regression estimators," Journal of the American Statistical Association, 54 (1959), 594-612.

(20) Midyuno, H., "On the sampling system with probability proportional to sum of sizes," Annals of the Institute of Statistical Mathematics, 2, 99-108.

(21) Murthey, M. N., and Nanjamma, N. S., "Almost unbiased ratio estimates based on interpenetrating sub-samples," Sankhya, The Indian Journal of Statistics, 21 (1960), 381-92.

(22) Nanjamma, N. S., Murthey, M. N., and Sethi, V. K., "Some sampling systems providing unbiased ratio estimators," Sankhya, The Indian Journal of Statistics, 21 (1960), 299-314.

(23) Olkin, I., "Multivariate ratio estimation for finite populations," Biometrika, 45 (1958), 154-165.

(24) de Pascual, J. N., "Unbiased ratio estimators in stratified sampling," Journal of the American Statistical Association, 56, 70-87.

(25) Pathak, P. K., "On sampling schemes providing unbiased ratio estimators," Annals of Mathematical Statistics, 35 (1964), 222-231.

(26) Quenouille, M. H., "Notes on bias in estimation," Biometrika, 43 (1956), 3 and 4, 353-360.

(27) Raj, D., "Ratio estimation in sampling with equal and unequal probabilities," Journal of the Indian Society of Agricultural Statistics, 6 (1954), 2, 127-138.

(28) Raj, D., "On double sampling for PPS estimation," Annals of Mathematical Statistics, 35 (1964), 2, 900-902.

(29) Raj, D., "On sampling over two occasions with probability proportionate to size," Annals of Mathematical Statistics, 36 (1965), 1, 327-330.

(30) Raj, D., "A note on the variance of the ratio estimate," Journal of the American Statistical Association, 59, 895-8.

(31) Raj, D., "On a method of using multi-auxiliary information in sample surveys," Journal of the American Statistical Association, 60, 270-277.

(32) Raj, D., "Variance estimation in randomized systematic sampling with probability proportionate to size," Journal of the American Statistical Association, 60, 278-284.

(33) Rao, J. N. K., A note on estimation of ratios by Quenouille's method," Biometrika, 52 (1965), 647-9.

(34) Rao, J. N. K., and Webster, J. T., "On two methods of bias reduction in the estimation of ratios," Annals of Mathematical Statistics, 37, 554-5.

(35) Robson, D. S., "Applications of multivariate polykays to the theory of unbiased ratio-type estimation," Journal of the American Statistical Association, 52 (1957), 511-522.

(36) Robson, D. S., and Vithayasai,
Journal of the American Statistical Association, 56, 350-8.

(37) Sukhatme, B. V., "Some ratio-type estimators in two-phase sampling," Journal of the American Statistical Association, 57, 628-632.

(38) Tin, M., "Comparisons of some ratio estimators," Journal of American Statistical Association, 60, 294-307.

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

(40) Williams, W. H., "On two methods of unbiased estimation with auxiliary variates," Journal of the American Statistical Association, 57, 184-186.

# UNBIASED RATIO-TYPE ESTIMATORS

by

REGINALD GERALD WORTHLEY

B. A., University of Maine, 1965

---

AN ABSTRACT OF A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1967

Ratio estimators have been used quite extensively in sample surveys, not only as estimators of population ratios, but as estimators of population means and totals. It has been demonstrated that in a great many situations the ratio estimator has a smaller variance than the traditional mean per unit estimator. A major drawback to the ratio estimator is the fact that it is biased, although in large samples it has been demonstrated that the bias is negligible. In very small samples, or even moderate samples from a stratified population, no really convincing argument has been given for the negligibility of the bias, since no exact expression for it is available. Several authors have avoided this question of bias by developing methods which eliminate the bias while retaining the essential properties of a ratio estimator.

This report reviews the usual ratio estimator, giving optimum conditions for its use. The bias is approximated and limits for the bias are given, as well as cases that might arise in which the bias might become an important factor. Methods are then considered which give rise to reduced bias estimators, as well as unbiased ratio-type estimators. The reduced bias estimators involve the use of expansions, approximations and a graphical method to obtain reduced bias estimators. The latter estimators are divided into two major classes of development, (1) the elimination of bias through the use of commonly used sampling schemes, and (2) the elimination of bias through the use of certain modifications of sampling schemes making the usual biased estimator unbiased.

Finally a small computer survey is presented in which several of the estimators are compared with respect to bias and efficiency.