

THEMIS SIGNAL ANALYSIS STATISTICS RESEARCH PROGRAM

GENERALIZATION OF RATIO ESTIMATE FOR POPULATION TOTAL

by

John E. Walsh

Technical Report No. 16  
Department of Statistics THEMIS Contract

October 1, 1968

Research sponsored by the Office of Naval Research  
Contract N00014-68-A-0515  
Project NR 042-260

Reproduction in whole or in part is permitted  
for any purpose of the United States Government.

DEPARTMENT OF STATISTICS  
Southern Methodist University

# GENERALIZATION OF RATIO ESTIMATE FOR POPULATION TOTAL

John E. Walsh

Southern Methodist University\*

## ABSTRACT

It is desired to estimate the total  $Y$  for a finite population (of size  $N$ ) with unknown values. A population with corresponding known values is available. Also, a simple random sample of size  $n$  is taken from the unknown population. Let  $\bar{y}$  be the mean of the sample while  $\bar{x}$  is the mean of the  $n$  values from the known population that correspond to these sample values. The ratio estimate of  $Y$  is generalized to the form  $N\bar{y}/[A\bar{x} + (1-A)\bar{X}]$ , where  $\bar{X}$  is the mean of the known population. Use of  $A = A_0$  suitably chosen, yields an estimate that approximately (terms of order  $n^{-2}$  in  $n$  neglected) is unbiased and has as small a variance as is attainable for a linear regression estimate. When  $A_0$  is unknown in advance, it can be estimated (denoted by  $A'_0$ ). Use of  $A'_0$  and a mild modification yields an estimate with the favorable properties occurring for  $A_0$ . An estimate is developed (terms of order  $n^{-5/2}$  neglected) for the standard deviation of the estimate using  $A'_0$ . Also, an estimate is obtained (terms of order  $n^{-2}$  neglected) for the standard deviation of this estimate for the standard deviation. Finally, some comparisons are made with the linear regression estimate.

---

\* Based on some work done for Mobil Research and Development Corporation. Associated with ONR Contract N00014-68-A-0515.

## INTRODUCTION

The ratio estimate is well established and has a good intuitive basis. However, its variance is (approximately) greater than or equal to that for a suitable linear regression estimate and can be much greater (for example, see Cochran, 1963). It would seem worthwhile to develop an estimate which has a ratio form similar to that of the ratio estimate and variance properties that are not inferior to those attainable with a linear regression estimate. This can be accomplished by a suitable modification of the denominator.

The total  $Y$  of a finite population of size  $N$  is to be estimated. This is done on the basis of a simple random sample (without replacement) of size  $n$  and complete knowledge about a population of  $N$  values that correspond to the values of the population with unknown total  $Y$ . That is, to each value  $Y_i$  of the unknown population there corresponds a known value  $X_i$  from the other population ( $i = 1, \dots, N$ ). Let  $\bar{y}$  be the arithmetic average of the sample values while  $\bar{x}$  is the arithmetic average of the  $X_i$  that correspond to the values of this sample. Then, the ratio estimate is determined by the intuitively plausible relations

$$Y/X = Y/N\bar{X} \doteq N\bar{y}/N\bar{x} = \bar{y}/\bar{x},$$

so that this estimate of  $Y$  equals  $N\bar{X}\bar{y}/\bar{x}$ , where  $X$  is the total and  $\bar{X}$  is the mean of the population of the  $X_i$ .

The same kind of plausible relations hold if  $\bar{x}$  is replaced by a quantity of the form  $A\bar{x} + (1-A)\bar{X}$ , and this is the basis for use of

$$N\bar{X}\bar{y}/[A\bar{x} + (1-A)\bar{X}].$$

as the form of the generalized ratio estimate.

Suitable choice of  $A$  yields an estimate with a standard deviation that (approximately) is as small as that attainable for the linear regression estimate. Specifically, the choice is

$$A_0 = \rho_{xy} (S_y/\bar{Y})(S_x/\bar{X})^{-1},$$

where  $\bar{Y}$  is the (unknown) mean of the population of the  $Y_i$ ,  $S_y^2$  is the (unknown) variance of the  $Y_i$ ,  $\rho_{xy}$  is the (unknown) correlation between the two populations, and  $S_x^2$  is the (known) variance of the  $X_i$ . The resulting estimate is unbiased when terms of order  $N/n^2$  are neglected and its standard deviation equals

$$[N(N-n)/n]^{1/2} S_y (1-\rho_{xy}^2)^{1/2}$$

plus terms of order  $N^2/n^2$ .

Ordinarily, the value for  $A_0$  is not known in advance with sufficient accuracy. However, it can be estimated (denoted by  $A'_0$ ) and, with a slight modification, an estimate is obtained that has the same bias and variance properties that were stated for the case of  $A_0$  known. The modification consists in multiplying  $\bar{y}$  by a quantity that should not differ much from unity.

The standard deviation of the estimate using  $A'_0$  is estimated by a statistic that is unbiased when terms of order  $N/n^{5/2}$  are neglected. The standard deviation of this statistic is estimated by another statistic that is unbiased when terms of order  $N/n^2$  are neglected.

These results should ordinarily be usable even when  $n$  is not very large (say,  $n \geq 20$ ). That is, the terms neglected should be acceptably small in many cases where  $n$  is of only moderate size. This is an advantage of developing estimates whose expectations include the terms that are of order  $n^{-1}$  and  $n^{-3/2}$  in  $n$ .

Sometimes a conservative estimate for the standard deviation of the ratio estimate using  $A'_0$  is desired. This can be obtained, say, by adding twice the estimate of its own standard deviation to the (approximately) unbiased estimate of the standard deviation for the ratio estimate.

The ratio estimate using  $A'_0$  can be simplified by not modifying  $\bar{y}$ . However, its expectation then equals  $Y$  plus terms of order  $N/n$ , so that much larger  $n$  is needed to justify usability. The standard deviation is the same as when  $\bar{y}$  is modified if terms of order  $N/n^{3/2}$  are neglected.

Generalized ratio estimates are, of course, also applicable for stratified random sampling situations. That is, a generalized ratio estimate is obtained for each stratum and a weighted average of these estimates is obtained. Here, the weight used for a stratum might be the sample size for this stratum divided by the total sample size (proportional allocation method).

The fact that the generalized ratio estimate (using  $A_0$  or  $A'_0$ ) has properties similar to those attainable with a linear regression estimate indicates that these two kinds of estimates may be strongly related. Examination shows that this is the case.

The next section contains a complete statement of the notation. This is followed by a statement of the ratio estimates using  $A'_0$  and the two estimates of standard deviations. Some properties of these estimates are also considered. The next to last section is devoted to a comparison of the generalized ratio estimate with the linear regression estimate. The final section contains an outline of the derivations for the stated material (including the case of  $A_0$  known).

#### NOTATION

$N$  = population size

$Y_i$  = value of  $i$ -th element in population whose total is to be estimated ( $i = 1, \dots, N$ )

$X_i$  = known value that corresponds to  $Y_i$ . That is,  $(X_i, Y_i)$  are considered as a pair ( $i = 1, \dots, N$ )

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^N Y_i, \quad Y = N\bar{Y}$$

$$\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i$$

$$S_y^2 = \frac{1}{N} \sum_{i=1}^N (Y_i - \bar{Y})^2 / (N - 1)$$

$$S_x^2 = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2 / (N - 1)$$

$$\rho_{xy} = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y}) / (N-1) S_x S_y$$

$n$  = size of sample randomly drawn from the  $Y_i$  (without replacement) so that each set of  $n$  different  $Y_i$  has the same probability of occurring

$y_j$  =  $j$ -th of the  $n$  sample values ( $j = 1, \dots, n$ )

$x_j$  = value of  $X_i$  paired with the value obtained for  $y_j$

$$\bar{y} = \frac{1}{n} \sum_{j=1}^n y_j, \quad \bar{x} = \frac{1}{n} \sum_{j=1}^n x_j$$

$$s_y^2 = \frac{1}{n} \sum_{j=1}^n (y_j - \bar{y})^2 / (n-1)$$

$$c_{xy} = \frac{1}{n} \sum_{j=1}^n (x_j - \bar{X})(y_j - \bar{y}) / (n-1)$$

$$E[(x_j - \bar{X})^a (y_j - \bar{y})^b] = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^a (Y_i - \bar{Y})^b / (N-1)$$

$$s_R^2 = s_y^2 - \left( \frac{nc_{xy}}{(n-1)S_x} \right)^2 + 2 \sum_{j=1}^n (x_j - \bar{X})^2 (y_j - \bar{y})^2 / (n-1) S_x^2$$

$$\epsilon'_n = c_{xy}^2 / n \bar{y}^2 S_x^2 - \sum_{j=1}^n (x_j - \bar{X})^2 (y_j - \bar{y})^2 / (n-1) \bar{y}^2 S_x^2$$

$$\epsilon'_R = \frac{1}{8(n-1)s_R^3} \left\{ \sum_{j=1}^n (y_j - \bar{y})^4 / (n-2) - (s_y^2 - 2c_{xy}^2 / S_x^2)^2 + 4c_{xy}^2 \sum_{j=1}^n (x_j - \bar{X})^2 (y_j - \bar{y})^2 / (n-1) S_x^4 - 4c_{xy} \sum_{j=1}^n (x_j - \bar{X})(y_j - \bar{y})^3 / (n-1) S_x^2 \right\}$$

$$A'_0 = \bar{X} c_{xy} / \bar{y} S_x^2$$

$$s_v^2 = 2s_R \epsilon'_R$$

ESTIMATES AND PROPERTIES

The generalized ratio estimate of  $Y$  that uses  $A'_0$  and a modification of  $\bar{y}$  is

$$y_R = N\bar{X}\bar{y}(1 - \epsilon'_n)/[A'_0\bar{x} + (1 - A'_0)\bar{X}].$$

The expected value of  $y_R$  equals  $Y$  plus terms of order  $N/n^2$ . Its standard deviation is

$$\left[\frac{N(N-n)}{n}\right]^{1/2} \left\{ S_y^2 - \left(1 - \frac{1}{n-1}\right) \rho_{xy}^2 S_y^2 + E[(x_j - \bar{X})^2(y_j - \bar{y})^2]/(n-1)S_x^2 + O(1/n^2) \right\}^{1/2}.$$

This standard deviation is unbiasedly estimated, with terms of orders  $N/n^{5/2}$  neglected, by

$$\left[\frac{N(N-n)}{n}\right]^{1/2} (s_R + \epsilon'_R).$$

Finally, the standard deviation of this estimate of the standard deviation, whose square equals

$$\frac{N(N-n)}{n(n-1)} [4S_y^2(1 - \rho_{xy}^2)]^{-1} \{ E(y_j - \bar{Y})^4 - S_y^4(1 - 2\rho_{xy}^2)^2 + 4\rho_{xy}^2 S_y^2 E[(x_j - \bar{X})^2(y_j - \bar{y})^2]/S_x^2 - 4\rho_{xy} S_y E[(x_j - \bar{X})(y_j - \bar{Y})^3]/S_x + O(1/n) \},$$

is unbiasedly estimated by

$$\left[\frac{N(N-n)}{n}\right]^{1/2} s_v$$

if terms of order  $N/n^2$  are neglected.

A simplified estimate of Y is obtained by setting  $\epsilon'_n$  equal to zero. Then, the ratio estimate is not necessarily unbiased unless terms of order  $N/n$  are neglected, and its standard deviation is

$$\left[\frac{N(N-n)}{n}\right]^{1/2} \{S_y^2(1 - \rho_{xy}^2) + o(1/n)\}^{1/2}.$$

This standard deviation is unbiasedly estimated by the square root of

$$\sum_{j=1}^n [y_j - \bar{y} - A'_0(x_j - \bar{X})]^2 / (n-1),$$

if terms of order  $N/n^{3/2}$  are neglected. The standard deviation for this estimate of the standard deviation is of order  $N/n$ , which is the order of terms that are being neglected.

#### COMPARISON WITH LINEAR REGRESSION ESTIMATE

First, let us consider the form and properties of linear regression estimates. The form is

$$N[\bar{y} - B(\bar{x} - \bar{X})].$$

When B is a specified constant, this estimate of Y is unbiased for all n. Also, it has smallest standard deviation for B equal to

$$B_0 = \rho_{xy} S_y / S_x,$$

and this standard deviation, for all n, is

$$\left[\frac{N(N-n)}{n}\right]^{1/2} S_y (1 - \rho_{xy}^2)^{1/2}$$

(for example, see Cochran, 1963).

When  $B_0$  is not sufficiently well known in advance, it can be estimated by

$$B'_0 = c_{xy} / S_x^2.$$

The estimate of  $y$  with  $B = B'_0$  is unbiased if terms of order  $N/n$  are neglected. Its standard deviation equals that occurring when  $B_0$  is used if terms of order  $N/n^{3/2}$  are neglected.

Now, let us consider the generalized ratio estimate with  $A = A_0$ . This can be written as

$$\begin{aligned} & N\bar{y}[1 + A_0(\bar{x} - \bar{X})/\bar{X}]^{-1} \\ &= N\bar{y}[1 - A_0(\bar{x} - \bar{X})/\bar{X} + A_0^2(\bar{x} - \bar{X})^2/\bar{X}^2 - \dots]. \end{aligned}$$

If terms with expectations of order  $N/n$  are neglected, this can be expressed as

$$N[\bar{y} - A_0(\bar{Y}/\bar{X})(\bar{x} - \bar{X})] = N[\bar{y} - B_0(\bar{x} - \bar{X})],$$

since  $A_0(\bar{Y}/\bar{X}) = B_0$ . Thus, to this accuracy, the two types of estimates are the same.

Finally, consider the generalized ratio estimate with  $A = A'_0$  and the modification involving  $\epsilon'_n$ . In  $\epsilon'_n$ , the term  $c_{xy}^2/n\bar{y}^2S_x^2$  has the same expectation as the more variable quantity  $A_0'^2(\bar{x} - \bar{X})^2/\bar{X}^2$ . Let  $1 - \epsilon'_n$  be replaced by

$$[1 - A_0'^2(\bar{x} - \bar{X})^2/\bar{X}^2]\{1 + \sum_{j=1}^n (x_j - \bar{X})^2(y_j - \bar{y})/(n-1)^2\bar{y}S_x^2\},$$

whose expectation equals that of  $1 - \epsilon'_n$  if terms of order  $N/n^2$  are neglected. The resulting ratio estimate can be expressed as

$$N\bar{y}[1 - A_0'^2(\bar{x} - \bar{X})^2/\bar{X}^2]\{1 + \sum_{j=1}^n (x_j - \bar{X})^2(y_j - \bar{y})/(n-1)^2\bar{y}S_x^2\},$$

divided by  $1 + A'_0(\bar{x} - \bar{X})$ . This equals

$$\begin{aligned} N\bar{y}[1 - A'_0(\bar{x} - \bar{X})/\bar{X}] &= N[\bar{y} - A'_0(\bar{y}/\bar{X})(\bar{x} - \bar{X})] \\ &= N[\bar{y} - B'_0(\bar{x} - \bar{X})] \end{aligned}$$

if terms of order  $N/n$  are neglected, so that the two types of estimates are the same to this level of accuracy. Agreement except for terms with expectation  $N/n^2$  occurs if the modification

$$N[\bar{y} - B'_0(\bar{x} - \bar{X}) + \sum_{j=1}^n (x_j - \bar{X})^2 (y_j - \bar{y}) / (n-1)^2 S_x^2]$$

is made in the linear regression estimate. Analysis similar to that for  $A = A_0$  shows that the simplified generalized ratio estimate with  $A = A'_0$  is the same as the linear regression estimate with  $B = B'_0$  if terms with expectation  $N/n$  are neglected.

#### OUTLINE OF DERIVATIONS

Verification of the properties stated for the various estimates is tedious but straightforward. The method is to first clear fractions in the expression for the ratio estimate minus  $Y$ . This yields

$$N[1 + A(\bar{x} - \bar{X})/\bar{X}]^{-1} [\bar{y} - \bar{Y} - A(\bar{x} - \bar{X})],$$

where  $\bar{y}$  is replaced by  $\bar{y} - \epsilon'_n \bar{y}$  when  $A = A'_0$  and the modification involving  $\epsilon'_n$  is used. Replacement of  $[1 + A(\bar{x} - \bar{X})/\bar{X}]^{-1}$  by  $1 - A(\bar{x} - \bar{X})/\bar{X}$  provides an equivalent expression if terms with expectations of order  $N/n^2$  are neglected. Thus, it is sufficient to determine properties for

$$N[\bar{y} - \bar{Y} - A(\bar{x} - \bar{X}) - A(\bar{x} - \bar{X})(\bar{y} - \bar{Y})/\bar{X} + A^2(\bar{x} - \bar{X})^2/\bar{X}^2]. \quad (1)$$

Examination shows that this quantity has zero expectation when

$A = A_0$  and when  $A = A'_0$  with  $\bar{y}$  replaced by  $\bar{y} - \epsilon'_n \bar{y}$ . Its expectation is of order  $N/n$  for the simplified estimate with  $A = A'_0$ .

The expectation of the square of expression (1), which is also the variance of (1) to the accuracy considered, is found to be

$$[N(N - n)/n][S_y^2(1 - \rho_{xy}^2) + O(1/n)]$$

when  $A = A_0$ , and is also of this form when  $A = A'_0$  for the simplified estimate. For  $A = A'_0$  and  $\epsilon'_n$  used, the expression is

$$[N(N - n)/n]\{S_y^2 - [1 - 1/(n - 1)]\rho_{xy}^2 S_y^2 + E[(x_j - \bar{X})^2(y_j - \bar{Y})^2]/(n - 1)S_x^2 + O(1/n^2)\}.$$

The square roots of these expressions provide the corresponding standard deviations.

For  $A = A_0$ , or  $A = A'_0$  and the simplified estimate, it can be verified that, to the stated accuracy,

$$E\left\{\sum_{j=1}^n [y_j - \bar{y} - A(\bar{x} - \bar{X})]^2/(n - 1)\right\}^{1/2}$$

equals the corresponding standard deviation. For  $A = A_0$  and  $\epsilon'_n$  used,

it is first noticed that

$$\begin{aligned} Es_R &= E\{(Es_R^2)^{1/2} [1 + (1/2)(s_R^2 - Es_R^2)/Es_R^2 - (1/8)(s_R^2 - Es_R^2)^2/(Es_R^2)^2 + \dots]\} \\ &= (Es_R^2)^{1/2} - (1/8)E(s_R^2 - Es_R^2)^2/(Es_R^2)^{3/2} + O(1/n^2); \end{aligned}$$

also that

$$\begin{aligned}
\epsilon'_R &= E(s_R^2 - Es_R^2)^2 / (Es_R^2)^{3/2} + O(1/n^2) \\
&= [8(n-1)S_y^3(1-\rho_{xy}^2)^{3/2}]^{-1} \{E(y_j - \bar{Y})^4 - S_y^4(1-2\rho_{xy}^2)^2 \\
&\quad + 4\rho_{xy}^2 S_y^2 E(x_j - \bar{X})^2 (y_j - \bar{Y})^2\} / S_x^2 \\
&\quad - 4\rho_{xy} S_y E[(x_j - \bar{X})(y_j - \bar{Y})^3] / S_x + O(1/n)\}.
\end{aligned}$$

These relations, combined with the fact that  $[N(N-n)/n]Es_R^2$  equals the variance of the estimate when terms of order  $N/n^3$  are neglected, shows that the expectation of  $[N(N-n)/n]^{1/2}(s_R + \epsilon'_R)$  equals the standard deviation of the ratio estimate ( $A = A'_0$  and using  $\epsilon'_R$ ) when terms of order  $N/n^{5/2}$  are neglected.

Finally, consider the standard deviation of  $[N(N-n)/n]^{1/2}(s_R + \epsilon'_R)$  and an estimate of this standard deviation. The variance of  $s_R + \epsilon'_R$  equals

$$\begin{aligned}
E(s_R + \epsilon'_R)^2 - (Es_R + E\epsilon'_R)^2 \\
= 2E(s_R \epsilon'_R) + O(1/n^2),
\end{aligned}$$

since  $(Es_R + E\epsilon'_R)^2 = Es_R^2$  plus terms of order  $1/n^2$  and  $E(\epsilon'_R)^2$  is of order  $1/n^2$ . Thus, the standard deviation sought equals the square root of

$$2[N(N-n)/n][S_y(1-\rho_{xy}^2)^{1/2}E\epsilon'_R + O(1/n^2)].$$

This standard deviation, evidently, is estimated by

$$\{2[N(N-n)/n]s_R \epsilon'_R\}^{1/2},$$

whose expectation equals the standard deviation plus terms of order  $N/n^2$ .

REFERENCE

Cochran, William G., Sampling techniques (2nd Edition), John Wiley and Sons, 1963