SAMPLING EXPERIMENTS ON RATIO ESTIMATORS*

By: Alan Ross, University of Kentucky Medical Center

1. Introduction. An increasing variety of estimation formulas involving auxiliary variables is appearing in the statistical literature and elsewhere [4, 5, 6, 7, 8, 9]. A general tack in this development is a search for estimators that eliminate or suppress bias which is present in the standard ratio and regression estimators used in sample surveys. Small samples are the principal source of difficulty when one attempts a formal evaluation of the characteristics of these estimators, since formulas for the variance and bias have only been obtained as approximations valid for large samples. The purpose of this paper is to present some empirical evidence bearing on the performance of the ordinary ratio of means estimator for samples of size 2, 3, and 4, and to compare this estimator against an unbiased ratio-type estimator.

Let y denote the characteristic whose population mean $\bar{Y}$ is to be estimated, and let x be a variable with known mean $\bar{X}$. The formula

$$\tilde{\bar{y}} = \bar{y}\bar{X}/\bar{x}, \qquad (1)$$

where $\bar{y}$ and $\bar{x}$ are sample means, is the standard ratio of means estimator in its elementary form appropriate for simple random sampling. The ratio estimator $\tilde{\bar{y}}$ is known to be biased with the bias decreasing for increasing sample size. In survey designs involving small samples from many strata it is possible that the combined bias in $\tilde{\bar{y}}$ may assume serious proportions. For it can be shown that if the bias in $\tilde{\bar{y}}$ for each stratum has the same sign, the bias in the estimated population mean will be approximately constant and equal to the average bias for individual strata, whereas the standard deviation of the overall estimate decreases by a factor of $1/\sqrt{L}$ (where L is the total number of strata) [2].

The large sample variance formula generally used for $\tilde{\bar{y}}$ is

$$\mathrm{Var}(\tilde{\bar{y}}) = \frac{N-n}{N}\frac{\bar{Y}^2}{n}\left\{\frac{S_y^2}{\bar{Y}^2} + \frac{S_x^2}{\bar{X}^2} - 2\frac{S_{xy}}{\bar{Y}\bar{X}}\right\}, \qquad (2)$$

where the S^2's are mean squares and S_{xy} is the mean product of x and y. A consequence of (2) is that the ratio estimate $\tilde{\bar{y}}$ will have smaller variance than the simple mean $\bar{y}$ if

$$\rho_{xy} > \frac{1}{2}\left(\frac{S_x}{\bar{X}}\right)\Big/\left(\frac{S_y}{\bar{Y}}\right),$$

where ρ_{xy} is the correlation between x and y. (See, e.g., [2].)

A fairly simple unbiased estimator based on ratios was proposed in [4]. The estimator for simple random sampling is written

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

where $r_i = y_i/x_i$, $\bar{r} = \frac{1}{n}\sum r_i$, N is the number of units in the population, and n is the sample size. Robson [4] gave an exact formula for the variance of y' in terms of multivariate symmetric means. In terms of ordinary population mean squares and products the variance of y' is given in [8].

Let

$$\bar{R} = \frac{1}{N}\sum r_i$$

$$S_r^2 = \frac{1}{N-1}\sum (r_i - \bar{R})^2$$

$$S_{xr} = \frac{1}{N-1}\sum (x_i - \bar{X})(r_i - \bar{R})$$

with parallel notation for the mean squares and products involving y's and x's. Then the variance of y' can be written

$$\frac{Nn(n-1)}{(N-1)^2}V(y') = aS_y^2 + (N-1)bS_x^2S_r^2 + a\bar{R}^2S_x^2 + c\bar{x}^2S_r^2 + 2d\bar{x}S_{yr} + 2e\bar{R}S_{xy} + \frac{(N-1)}{N}fS_{xr}^2 + g\bar{x}\bar{R}S_{xr},$$

where a, b, ..., g are constant coefficients as follows:

$$a = \frac{n-1}{N} - \frac{(N-1)(n-2)(n-3)}{N(N-2)(N-3)}$$

$$b = \frac{1}{N^2} - \frac{2(n-2)}{N^2(N-2)} + \frac{(n-2)(n-3)}{(N-2)(N-3)}$$

$$c = \frac{(n-1)}{N} - \frac{(N-1)(n-2)(n-3)}{N(N-2)(N-3)} - \frac{(n-1)}{(N-1)} - \frac{(n-1)^2}{(N-1)^2} + \frac{2(n-1)(n-2)}{(N-1)(N-2)}$$

$$d = \frac{(n-2)(n-3)}{N(N-2)} - \frac{(n-1)(n-3)}{N(N-1)} + \frac{2(n-2)(n-3)}{N(N-2)(N-3)} - \frac{2(n-1)(n-2)}{N(N-1)(N-2)}$$

$$e = \frac{(n-2)(n-3)}{N(N-2)} - \frac{(n-1)}{N} + \frac{2(n-2)(n-3)}{N(N-2)(N-3)}$$

$$f = \frac{1}{N} + \frac{(n-1)^2}{N} + \frac{2(n-2)^2}{N(N-2)} + \frac{2(n-2)(n-3)}{N(N-2)(N-3)} + \frac{(n-1)}{N} - \frac{(N-1)(n-2)(n-3)}{N(N-2)(N-3)} - \frac{Nn(n-1)}{N(N-1)}$$

$$g = \frac{2}{N} + \frac{(n-1)(n-2)}{N} - \frac{n(n-1)}{(N-1)} + \frac{(n-2)(n-3)}{N(N-2)}$$

$$+ \frac{N(n-1)(n-2)}{(N-1)(N-2)} - \frac{(N-1)^2(n-2)(n-3)}{N(N-2)(N-3)}$$

2. The Sampling Experiments. It was indicated above that there is as yet no simple formulation of the characteristics of the distribution of the estimator for arbitrary (finite) populations in terms of population moments or equivalents. In order to gather some evidence on how effective $\tilde{y}$ is for small samples, two universes of 50 elements each were constructed, random samples (without replacement) of size 2, 3, and 4 were drawn, and the distributions of $\tilde{y}$ were estimated on the basis of the repeated sampling.

One universe was composed of the first 50 families included in a sample survey of the University of Kentucky faculty and staff [1]. For that survey data were recorded for each family on the number of persons in the family, number of physician visits in 1957 (home, office, University Health Service Dispensary), total charges to the family by physicians for these outpatient visits, and a host of other material.

In example A we regard the average number of physician visits as the quantity to be estimated from a sample, and we assume that the average number of persons per family is known for the universe. In formula (1) $\bar{X}$ is then the population mean number of persons per family (2.94 for this example), and $\bar{y}$ and $\bar{x}$ denote sample means of physician visits and number of persons per family, respectively.

The same universe of 50 families was used for Example B with physician charges (dollars) forming the y population and number of physician visits was the x population or concomitant variable.

The University of Kentucky Department of Agricultural Economics provided corn acreage data for 1952 and 1956 on a number of farms scattered throughout Kentucky. The first 50 of these farms were used as a universe for Example C. In this situation we shall estimate the average number of acres planted in corn in 1956 (the y variable) with assistance of the population mean acreage planted in 1952 (the x variable).

Sampling was accomplished by loading the x and y variables for a universe onto the drum of an IBM Type 650 Data Processing Machine and then feeding random numbers (between 1 and 50) to designate (x,y) pairs for a sample. The random numbers were introduced four at a time--the first two designated a sample of size 2, the first three numbers designated a sample of size 3, and the four numbers together designated a sample of size 4. The numbers in each set of four were distinct so that the method of selection was random sampling without replacement. No control was exercised over repeated pairs, triplets, or quadruplets of random numbers so that sets of two, three, or four elements were selected at random with replacement from populations consisting of pairs, triplets, and quadruplets of (x,y) sets.

The estimates $\tilde{y}$ and y' were computed for each sample designated as given above. Totals of $\tilde{y}$, $\tilde{y}^2$, $\tilde{y}^3$, $\tilde{y}^4$, y', y'^2, y'^3, and y'^4 were accumulated as the sampling proceeded to allow for computation of moments of the distributions of $\tilde{y}$ and y'. The frequencies were also accumulated in eleven equally spaced intervals for each distribution.

Example A--Physician Visits. Some of the properties of the populations are given in Table 1. Since the relation $\rho_{xy} > C_x/2C_y$ holds we should expect that $\tilde{y}$ will be more efficient than the simple mean $\bar{y}$, at least for large samples.

One thousand samples of size 2, size 3, and size 4 were drawn from this universe. The results of the samplings are summarized in Table 2. The entries for $\tilde{y}$ and y' were calculated from the samplings, while the variances of the simple mean $\bar{y}$ was computed from formula. The variance formula for y' was not used since there was some evidence that rounding error introduced by the machine inflated the estimated variance of both $\tilde{y}$ and y'.

The variance of $\bar{y}$ is given in the first line as a reference point. The estimated bias of $\tilde{y}$, shown in line 2, is seen to be negligible, although the bias demonstrated no tendency to decrease as the sample size increased from 2 to 4. The bias relative to the standard deviation of $\tilde{y}$, given in line 6, showed a tendency to increase. Line 8 indicates an increase in efficiency of the unbiased estimator y' relative to $\tilde{y}$ with increasing sample size.

Example B--Physician Charges. The populations and summary of results for this example are presented in Table 3 and Table 4. As before the characteristics of the distributions were estimated from 1000 samples of each size.

Table 1. Properties of the populations used in Example A.

y = number of physician visits
$\bar{Y} = 18.42$
$S_y^2 = 243.06$
$C_y = .85$
x = number of persons in family
$\bar{X} = 2.94$
$S_x^2 = 2.43$
$C_x = .53$
$\rho_{xy} = .47$
$C_x/2C_y = .31$

Table 2. Summary of Example A

	Sample Size		
	n=2	n=3	n=4
1. $\text{var}(\bar{y})$	116.67	76.14	55.90
2. $\text{bias}(\tilde{\bar{y}})$	.09 (0.5%)	.27 (1.5%)	.30 (1.6%)
3. m.s.e.$(\tilde{\bar{y}})$	75.05	53.28	41.06
4. $\text{var}(\tilde{\bar{y}})$	75.04	53.21	40.97
5. $\sqrt{\text{var}(\tilde{\bar{y}})}$	8.66	7.29	6.40
6. $\text{bias}(\tilde{\bar{y}})/\sqrt{\text{var}(\tilde{\bar{y}})}$	1.1%	3.7%	4.7%
7. var(y')	111.93	63.47	47.65
8. m.s.e.$(\tilde{\bar{y}})$/var(y')	67.0%	83.8%	86.2%

In Example B the bias in $\tilde{\bar{y}}$ was more pronounced than in the previous example. The estimated biases for the three sample sizes were significantly different from zero with probabilities less than .01. Line 8 shows increasing efficiency of y' relative to $\tilde{\bar{y}}$ as the sample size increased from 2 to 4.

Example C--Acres in Corn. Tables 5 and 6 give the information pertaining to this example. In this example characteristics of the distributions were estimated from 500 samples of each size.

The estimated biases for $\tilde{\bar{y}}$ in Table 6 are significantly different from zero at the 1% level for samples of size 2 and size 3 and at the 5% level for samples of size 4.

Table 3. Properties of the populations used in Example B.

y = physician charges
$\bar{Y} = 84.92$
$S_y^2 = 6659.71$
$C_y = .96$
x = number physician visits
$\bar{X} = 18.42$
$S_x^2 = 243.06$
$C_x = .85$
$\rho_{xy} = .74$
$C_x/2C_y = .44$

Table 4. Summary of Example B

	Sample Size		
	n=2	n=3	n=4
1. $\text{var}(\bar{y})$	3196.66	2086.71	1531.73
2. $\text{bias}(\tilde{\bar{y}})$	4.63 (5.5%)	3.25 (3.8%)	4.45 (5.2%)
3. m.s.e.$(\tilde{\bar{y}})$	2078.03	1318.37	1022.94
4. $\text{var}(\tilde{\bar{y}})$	2056.60	1307.78	1003.16
5. $\sqrt{\text{var}(\tilde{\bar{y}})}$	45.35	36.16	31.68
6. $\text{bias}(\tilde{\bar{y}})/\sqrt{\text{var}(\tilde{\bar{y}})}$	10.2%	9.1%	14.0%
7. var(y')	2833.48	1427.33	1026.56
8. m.s.e.$(\tilde{\bar{y}})$/var(y')	73.3%	92.4%	99.6%

Table 5. Properties of the populations used in Example C.

y = acres in corn (1956)
$\bar{Y} = 25.28$
$S_y^2 = 237.31$
$C_y = .61$
x = acres in corn (1952)
$\bar{X} = 36.56$
$S_x^2 = 399.37$
$C_x = .55$
$\rho_{xy} = .70$
$C_x/2C_y = .45$

3. Discussion. Perhaps the most consistent finding for the three examples was the tendency for the efficiency of the unbiased estimator y' to increase relative to $\tilde{\bar{y}}$ as sample size was increased from 2 to 4. The bias in $\tilde{\bar{y}}$ was persistent but not overwhelming in two of the examples when compared to the standard error (line 6 in Tables 2, 4, and 6). Judging from the examples presented here it appears that when a ratio estimate is appropriate according to the ρ_{xy} criterion the $\tilde{\bar{y}}$ estimator maintains its efficiency relative to $\bar{y}$ even for samples of size 2. The unbiased estimator y' compared favorably with $\tilde{\bar{y}}$ on accuracy for samples of size 4, and we note that an unbiased estimate of the variance of y' can be computed from a sample of four or more elements; but a further result of the sampling experiments indicates that a variance estimate for $\tilde{\bar{y}}$ for samples of size 4 computed by substituting sample values in (2) may be an underestimate by 15% to 35%.

Table 6. Summary of Example C

	Sample Size		
	n=2	n=3	n=4
1. $\text{var}(\bar{y})$	113.91	74.36	54.58
2. $\text{bias}(\tilde{\bar{y}})$	1.33(5.3%)	.98 (3.9%)	.59 (2.3%)
3. $\text{m.s.e.}(\tilde{\bar{y}})$	92.39	51.91	36.31
4. $\text{var}(\tilde{\bar{y}})$	90.62	50.95	35.96
5. $\sqrt{\text{var}(\tilde{\bar{y}})}$	9.52	7.13	5.99
6. $\text{bias}(\tilde{\bar{y}})/\sqrt{\text{var}(\tilde{\bar{y}})}$	14.0%	13.8%	9.9%
7. var(y')	115.62	57.11	40.79
8. $\text{m.s.e.}(\tilde{\bar{y}})/\text{var}(y')$	79.9%	90.9%	89.0%

Table 7. Coefficients of skewness and kurtosis for the distributions of $\tilde{\bar{y}}$ and y'.

		Sample Size					
		n=2		n=3		n=4	
		g_1	g_2	g_1	g_2	g_1	g_2
Physician Visits	$\tilde{\bar{y}}$	.676	3.412	.552	.241	.425	-.066
	y'	.192	1.56	.471	.236	.447	.052
Physician Charges	$\tilde{\bar{y}}$	1.167	2.358	1.92	1.311	.700	.921
	y'	-.176	4.094	.286	1.215	.342	.370
Acres in Corn	$\tilde{\bar{y}}$	.716	1.574	.241	.355	.169	.323
	y'	-.087	.936	.131	1.120	.146	.659

An additional set of descriptive measures for the sampling distributions of $\tilde{y}$ and y' is shown in Table 7 where estimates of γ_1 and γ_2 are given. The estimates of all 18 sampling distributions were unimodal with varying degrees of asymmetry and kurtosis. The pair of histograms in Figure 1 is typical and illustrative.

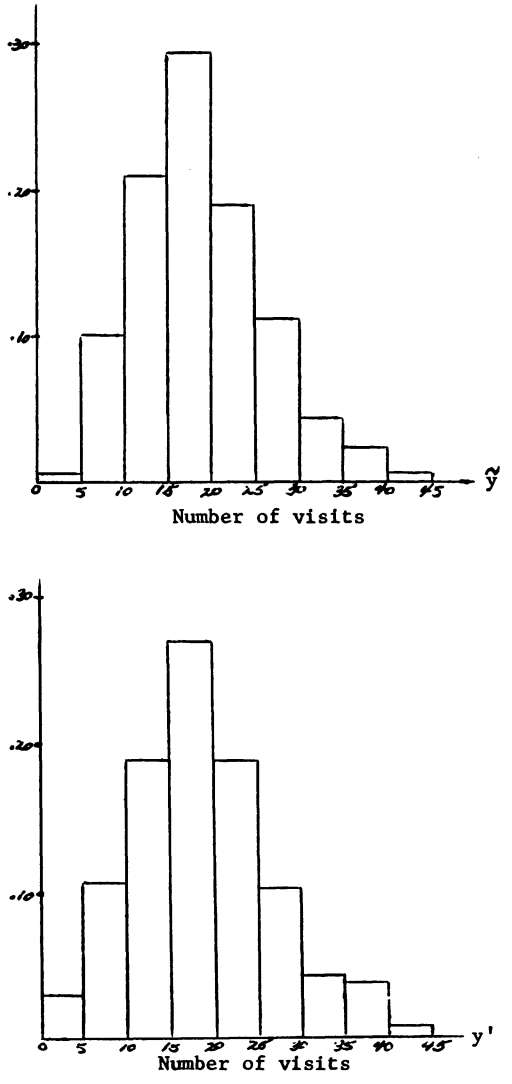

Figure 1. Histograms of distributions of $\tilde{y}$ and y' for physician visits for samples of size 3.

*The courtesy of the University of Kentucky Computing Center, whose facilities were used for the computations in this report, is gratefully acknowledged.

References

[1] Bost, H. L. and Ross, A. University of Kentucky employees and their families. Report no. 1 and report no. 3.(in preparation) on the health and insurance study. Lexington, Kentucky, University of Kentucky, Medical Center. 1959.

[2] Cochran, W. G. Sampling techniques. New York, John Wiley and Sons. 1953.

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

[4] Hartley, H. O. and Ross, A. Unbiased ratio estimators. Nature. 174:270-271. 1954.

[5] Mickey, M. R. Some finite population unbiased ratio and regression estimators. Jour. Amer. Stat. Assoc. 54:594-612. 1959.

[6] Nieto de Pascual, J. Unbiased ratio estimators in stratified sampling. Unpublished report. Ames, Iowa. 1959.

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

[8] Ross, A. On two problems in sampling theory: unbiased ratio estimators and variance estimates in optimum sampling designs. Unpublished Ph.D. thesis. Ames, Iowa, Iowa State University Library. 1960.

[9] Williams, W. H. Unbiased regression estimators. Unpublished report. Ames, Iowa. 1959.