UNBIASED COMPONENT-WISE RATIO ESTIMATION[1]

By: D. S. Robson and Chitra Vithayasai, Cornell University

INTRODUCTION

The precision of a ratio-type estimator such as $\bar{y}\bar{X}/\bar{x}$ can sometimes be substantially increased if the correlated variables y and x can be expressed as sums of more highly correlated components, $y=y_1+\cdots+y_k$ and $x=x_1+\cdots+x_k$. An empirical example of this arises in the ratio estimation of total dry matter yield of corn in field plot experiments; when both the green weight x and oven-dry weight y are measured and estimated separately for ears and the vegetative parts of the plant the efficiency of estimation of plot total dry weight is increased by approximately 70%. An example from general sample survey methodology is the case of cluster sampling with unequal size clusters when the elements in each randomly selected cluster are stratified into k strata; the usual mean per cluster ratio estimate is then replaced by the sum of k such ratio estimates for the individual strata.

In this paper we are concerned primarily with the Hartley-Ross [1]type of unbiased component-wise ratio estimator, for which we present the exact variance formula and an unbiased estimator of the variance. The efficiency of component-wise ratio estimation is then examined empirically with the data from 39 corn plots of 10 hills each, and the bias of the conventional ratio estimate and its variance formula are evaluated.

THE VARIANCE OF THE HARTLEY-ROSS TYPE OF COMPONENT-WISE RATIO ESTIMATOR

The Hartley-Ross unbiased ratio estimator of the population total for a single component y takes the form

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

where $\bar{r}$ is the mean ratio of y to x in a random sample of size n from a population of size N, and X is the population total for x. Goodman and Hartley [2] gave the limiting form of the variance of this estimator as

$$\lim_{N\to\infty} \frac{1}{N^2}\,\mathrm{var}(Y') = \left[\frac{1}{n}\,\sigma_y^2 + \bar{R}^2\sigma_x^2 - 2\bar{R}\sigma_{x,y} + \frac{1}{n-1}(\sigma_r^2\sigma_x^2 - \sigma_{r,x}^2)\right]$$

where $\bar{R}$ is the population mean of the ratio $r=(y/x)$.

The exact variance for finite N was given by Robson [3] in terms of Tukey's [4] multivariate polykays and may be most conveniently expressed in the notation of Tukey's symbolic, dot-multiplication as

$$\mathrm{var}(Y') = \frac{N(N-n)}{n}\left[\sigma_y^2 + \bar{R}\cdot\bar{R}\cdot\sigma_x^2 - 2\bar{R}\cdot\sigma_{x,y} + \frac{1}{n-1}\left(\frac{N-1}{N}\,\sigma_r^2\cdot\sigma_x^2 + \frac{N-n}{N}\,\sigma_{r,x}\cdot\sigma_{r,x}\right)\right]$$

All variances and covariances appearing in this formula are understood to be defined in the usual manner for finite populations; for example

$$\sigma_{r,x} = \frac{1}{N-1}\left(\sum_{i=1}^{N} r_i x_i - N\bar{R}\bar{X}\right)$$

This definition, which arises naturally in the algebraic treatment of moments and cumulants of a finite population, also serves to illustrate what is meant by dot-multiplication, since

$$\sigma_{r,x} = \frac{1}{N}\sum_{i=1}^{N} r_i x_i - \frac{1}{N(N-1)}\sum_{i\neq j}^{N} r_i x_j$$

$$= \overline{RX} - \bar{R}\cdot\bar{X}$$

thus, the dot-product of two means is the mean of all possible crossproducts. The same is true for the dot-product of more than two means; for example,

$$\bar{R}\cdot\sigma_{x,y} = \bar{R}\cdot(\overline{XY} - \bar{X}\cdot\bar{Y})$$

$$= \bar{R}\cdot\overline{XY} - \bar{R}\cdot\bar{X}\cdot\bar{Y}$$

$$= \frac{1}{N(N-1)}\sum_{i\neq j}^{N} r_i x_j y_j - \frac{1}{N(N-1)(N-2)}\sum_{i\neq j\neq k}^{N} r_i x_j y_k$$

and, similarly,

$$\bar{R}\cdot\bar{R}\cdot\sigma_x^2 = \bar{R}\cdot\bar{R}\cdot(\overline{XX} - \bar{X}\cdot\bar{X})$$

$$= \bar{R}\cdot\bar{R}\cdot\overline{XX} - \bar{R}\cdot\bar{R}\cdot\bar{X}\cdot\bar{X}$$

$$= \frac{1}{N(N-1)(N-2)}\sum_{i\neq j\neq k}^{N} r_i r_j x_k^2 - \frac{1}{N(N-1)(N-2)(N-3)}\sum_{i\neq j\neq k\neq h}^{N} r_i r_j x_k x_h$$

As N gets large, of course, the dot-product of two or more moments approaches the ordinary product of the moments, provided the latter approach a limit, and so the limiting var(Y') of Goodman and Hartley is obtained.

A minimum variance unbiased estimator of $\mathrm{var}(\hat{Y})$ is easily constructed using the fact that polykays, or dot-products of sample cumulants, are minimum variance unbiased estimators of the corresponding polykays of the finite population. Thus, for example, the minimum variance unbiased estimator of $\bar{R}\cdot\sigma_{x,y}$ is

$$\bar{r}\cdot s_{x,y}=\bar{r}\cdot\overline{xy}-\bar{r}\cdot\bar{x}\cdot\bar{y}$$

$$=\frac{1}{n(n-1)}\sum_{i\neq j}^{n} r_i x_j y_j-\frac{1}{n(n-1)(n-2)}\sum_{i\neq j\neq k}^{n} r_i x_j y_k$$

or, expressed in more convenient computational form,

$$\bar{r}\cdot s_{x,y}=\frac{1}{n(n-1)(n-2)}\left[(n-1)\sum_1^n x_i y_i \sum_1^n r_i - n\sum_1^n y_i^2+\sum_1^n x_i \sum_1^n y_i r_i\right.$$

$$\left.+\left(\sum_1^n y_i\right)^2+\sum_1^n x_i \sum_1^n y_i \sum_1^n r_i\right]$$

The other components of var(Y') are similarly estimated; computing formulas for the estimates are given by Robson [5] and will follow as special cases of the more general formulas given next for the component-wise ratio estimator.

The general case we wish to consider is

$$Y'=\sum_{i=1}^{k} Y_i'=\sum_{i=1}^{k}\left[X_i\bar{r}_i+\frac{n(N-1)}{n-1}(\bar{y}_i-\bar{x}_i\bar{r}_i)\right]$$

where now

$$\mathrm{var}(Y')=\sum_{i=1}^{k}\mathrm{var}(Y_i')+2\sum_{i<j}^{k}\mathrm{cov}(Y_i',Y_j')$$

Since the individual terms var(Y_i') take the form indicated earlier for a single component estimator, the only new algebraic problem is the computation of cov(Y_i',Y_j'), and by the same methods used earlier this may be shown to take the analogous form

$$\mathrm{cov}(Y_i',Y_j')=\frac{N(N-n)}{n}\left[\sigma_{y_i,y_j}+\bar{R}_i\cdot\bar{R}_j\cdot\sigma_{x_i,x_j}-\bar{R}_i\cdot\sigma_{x_i,y_j}\right.$$

$$-R_j\cdot\sigma_{y_i,x_j}+\frac{1}{n-1}\left(\frac{N-1}{N}\sigma_{r_i,r_j}\cdot\sigma_{x_i,x_j}\right.$$

$$\left.\left.+\frac{N-n}{N}\sigma_{r_i,x_j}\cdot\sigma_{x_i,r_j}\right)\right]$$

Computing formulas for the minimum variance unbiased estimators of the terms in this covariance formula are shown in Table 1 for the case i=1, j=2; sample means are expressed in the manner indicated earlier as, for example,

$$\overline{y_1y_2}=\frac{1}{n}\sum_{i=1}^{n} y_{1i}y_{2i}$$

and all products represent ordinary products, as

$$\bar{y}_1\bar{y}_2=\frac{1}{n^2}\left(\sum_1^n y_{1i}\right)\left(\sum_1^n y_{2i}\right)$$

In addition, the abbreviation $(n)_m$ is used for $n(n-1)\cdots(n-m+1)$. Computing formulas for estimating the components of var($\bar{Y}_i'$) may be obtained from putting $(r_1,x_1,y_1)=(r_2,x_2,y_2)=(r_i,x_i,y_i)$.

AN EMPIRICAL EVALUATION OF COMPONENT-WISE RATIO ESTIMATION OF CORN PLOT TOTAL DRY WEIGHT

Crop yield in agronomic experiments with silage corn is ordinarily measured in terms of total dry matter production per plot. Dry weight can be measured accurately only by drying the harvested plant material in ovens and there are, of course, distinct limitations on the amount of material which can be handled in this manner. Green, or fresh weight of the production from a plot, however, can be measured directly in the field as the material is harvested, and since green and dry weight are highly correlated the total dry weight for the plot can be accurately estimated by determining the dry matter percentage in a sample from the plot and applying this sample dry matter per cent to the measured total green weight. For the purpose of measuring the sampling error in this method of estimation, green and dry weight determinations were made on 390 individual hills of corn in an experiment containing an early, medium, and a late maturing variety arranged in plots of 10 hills.* These weight determinations were made separately for the ears and stovers of each hill (stovers =husks+stalks+leaves), thus providing an opportunity also to examine the efficiency of a component-wise estimator of plot total dry weight. The separate and combined components of hill green and dry weights are summarized graphically in Figure 1, showing that a somewhat higher green weight-dry weight correlation exists for the separate components, ears and stovers, than for the total, ears + stovers. Average within-plot correlations between green and dry weight of ears, stovers, and ears + stovers were .953, .932, and .824, respectively.

For each plot the efficiency of the unbiased component-wise ratio estimator $Y'=Y_{stover}'+Y_{ear}'$ relative to the unbiased combined ratio estimator $Y_{stover+ear}'$ was computed for samples of n hills, n=2,3,···,9. These efficiencies, in the form of a variance ratio $\mathrm{var}(Y_{s+e}')/\mathrm{var}(Y_s'+Y_e')$, were relatively constant for all n, and the average efficiencies over all 39 plots as shown in Table 2. The two components of the estimator, Y_{stover}' and Y_{ear}', were correlated in this experiment, but to a much lesser degree than the green and dry weights within each component (Table 3).

The variances var(Y_{s+e}'), var($Y_s'+Y_e'$), var(Y_s'), var(Y_e') and cov(Y_s',Y_e') employed in the above evaluation were computed directly from the formulas given earlier. In addition to this evaluation, however, the data provided an opportunity to compare the sampling error of the unbiased ratio estimator with the error mean square of the more conventional, but biased, ratio estimator $\hat{Y}=\bar{y}X/\bar{x}$. This was accomplished by enumerating all possible samples of size n for each plot of N=10 hills, computing the conventional ratio estimate for each such sample, and then averaging the squared error, (estimate-known plot dry weight)^2, over all $\binom{N}{n}$ samples. Averaged over all 39 plots, the error mean squares (EMS) for the three estimators $\hat{Y}_{s+e}$, $\hat{Y}_s$, $\hat{Y}_e$ compared to the variances of the corresponding unbiased ratio estimators as shown in Table 4.

The bias of the conventional estimator is negligible in this case, even for small samples, and its sampling error is the same as that of the unbiased ratio estimator. In practice, of course, the conventional estimator offers the advantage that individual hills in the sample need not be weighed and dried separately but may be handled in bulk.

Finally, the actual error mean square of the biased estimator $\hat{\bar{Y}}=\bar{y}\bar{X}/\bar{x}$ can be compared to variance approximation

$$\operatorname{var}(\hat{\bar{Y}}) \doteq \frac{N(N-n)}{n}\bar{Y}^2\left[\frac{\sigma_y^2}{\bar{Y}^2}+\frac{\sigma_x^2}{\bar{X}^2}-\frac{2\sigma_{x,y}}{\bar{X}\,\bar{Y}}\right]$$

This comparison is shown graphically in Figure 2. A tendency for this approximation to underestimate the true error mean square decreases as sample size increases since the actual error mean square decreases at a faster rate than the function N-n/n.

Table 1. Computing formulas for the estimation of $\operatorname{cov}(Y_1', Y_2')$

$$s_{y_1,y_2}=[\overline{y_1y_2}-\bar{y}_1\bar{y}_2]n^2/(n)_2$$

$$\bar{r}_1\cdot\bar{r}_2\cdot s_{x_1,x_2}=\Big\{n^2[2(\overline{y_1y_2}-\bar{x}_1\overline{r_1y_2}-\bar{x}_2\overline{y_1r_2})-(n-1)(\bar{r}_1\overline{x_1y_2}+\bar{r}_2\overline{y_1x_2})$$
$$-(n-2)(\overline{x_1x_2r_1r_2})-(\bar{y}_1\bar{y}_2+\overline{r_1y_2}\,\overline{y_1r_2})]+n^3[(n-2)\bar{r}_1\bar{r}_2\overline{x_1x_2}+\bar{r}_1\bar{x}_1\bar{y}_2$$
$$+\bar{r}_1\bar{x}_2\overline{x_1r_2}+\bar{r}_2\bar{x}_1\overline{r_1x_2}+\bar{r}_2\bar{x}_2\bar{y}_1+\bar{x}_1\bar{x}_2\overline{r_1r_2}]-n^4\bar{r}_1\bar{r}_2\bar{x}_1\bar{x}_2\Big\}/(n)_4$$

$$\bar{r}_1\cdot s_{x_1,y_2}=\Big\{n^2[(n-1)\bar{r}_1\overline{x_1y_2}-\overline{y_1y_2}+\bar{x}_1\overline{r_1y_2}+\bar{y}_1\bar{y}_2]-n^3\bar{r}_1\bar{x}_1\bar{y}_2\Big\}/(n)_3$$

$$\bar{r}_2\cdot s_{y_1,x_2}=\Big\{n^2[(n-1)\bar{r}_2\overline{y_1x_2}-\overline{y_1y_2}+\bar{x}_2\overline{y_1r_2}+\bar{y}_1\bar{y}_2]-n^3\bar{r}_2\bar{y}_1\bar{x}_2\Big\}/(n)_3$$

$$s_{r_1,r_2}\cdot s_{x_1,x_2}=\Big\{n^2[(n^2-3n+1)\overline{r_1r_2}\,\overline{x_1x_2}+(n-1)(\bar{x}_1\overline{r_1y_2}-\overline{y_1y_2}+\bar{x}_2\overline{y_1r_2}+\bar{r}_1\overline{x_1y_2}$$
$$+\bar{r}_2\overline{y_1x_2}+\bar{y}_1\bar{y}_2+\overline{r_1x_2}\,\overline{x_1r_2}]-n_3[(n-2)(\overline{r_1r_2}\bar{x}_1\bar{x}_2+\bar{r}_1\bar{r}_2\overline{x_1x_2})$$
$$+\bar{r}_1\bar{x}_1\bar{y}_2+\bar{r}_1\bar{x}_2\overline{x_1r_2}+\bar{r}_2\bar{x}_1\overline{r_1x_2}+\bar{r}_2\bar{x}_2\bar{y}_1]+n^4\bar{r}_1\bar{r}_2\bar{x}_1\bar{x}_2\Big\}/(n)_4$$

$$s_{r_1,x_2}\cdot s_{x_1,r_2}=\Big\{n^2[(n^2-3n+1)\overline{r_1x_2}\,\overline{x_1r_2}+(n-1)(\bar{x}_1\overline{r_1y_2}-\overline{y_1y_2}+\bar{r}_2\overline{y_1x_2}+\bar{r}_1\overline{x_1y_2}$$
$$+\bar{x}_2\overline{y_1r_2})+\bar{y}_1\bar{y}_2+\overline{r_1r_2}\,\overline{x_1x_2}]-n^3[(n-2)(\overline{r_1x_2}\bar{x}_1\bar{r}_2+\bar{r}_1\bar{x}_2\overline{x_1r_2})$$
$$+\bar{r}_1\bar{x}_1\bar{y}_2+\bar{r}_1\bar{r}_2\overline{x_1x_2}+\bar{x}_1\bar{x}_2\overline{r_1r_2}+\bar{r}_2\bar{x}_2\bar{y}_1]+n^4\bar{r}_1\bar{r}_2\bar{x}_1\bar{x}_2\Big\}/(n)_4$$

Table 2. Average relative efficiency of the unbiased component-wise estimator

n=2	n=3	n=4	n=5	n=6	n=7	n=8	n=9
1.67	1.68	1.69	1.69	1.69	1.69	1.69	1.69

Table 3. Average correlation between the two components of the estimator

n=2	n=3	n=4	n=5	n=6	n=7	n=8	n=9
.237	.240	.241	.241	.242	.242	.242	.241

Figure 1. Summary of data on 390 hills of corn

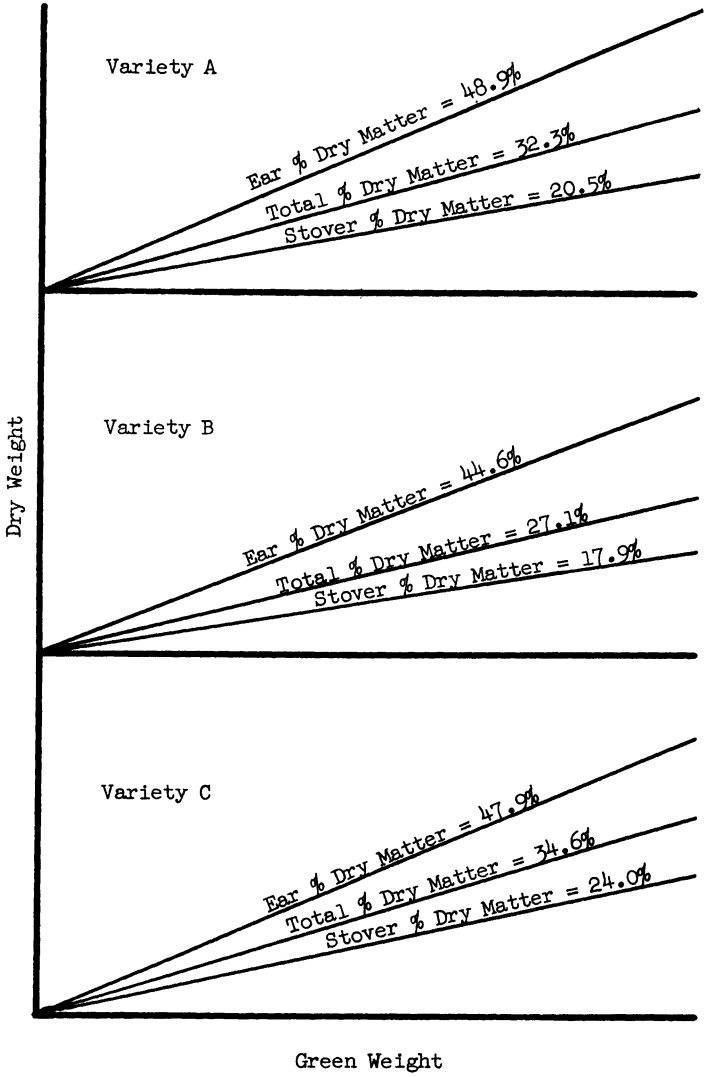

Figure 2. Bias of variance approximation

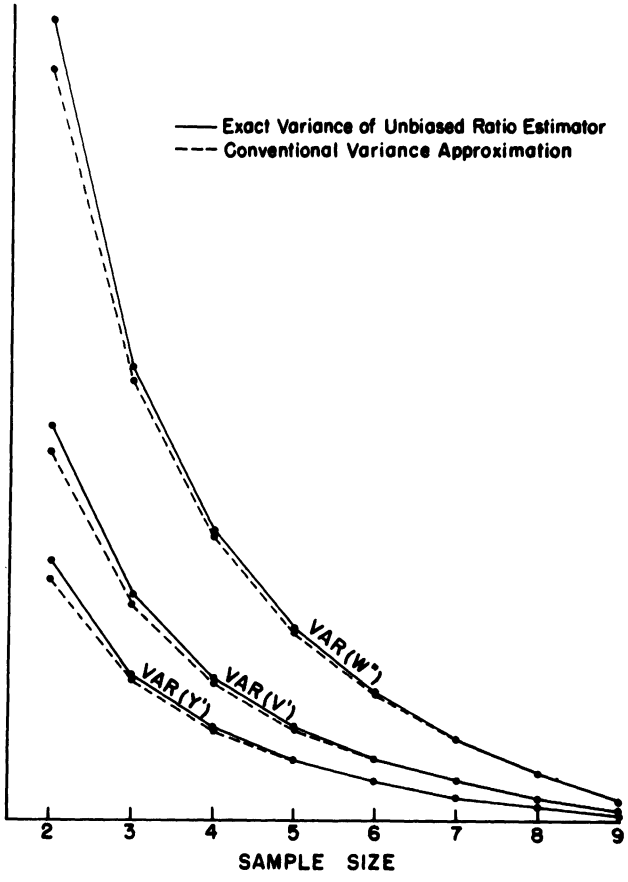

Table 4. Comparison of error mean squares of biased and unbiased ratio estimators

	n=2	n=3	n=4	n=5	n=6	n=7	n=8	n=9
EMS($\hat{Y}_{s+e}$)	45,603	26,102	16,665	--	7,352	4,718	2,748	1,220
var(Y'_{s+e})	46,233	26,275	16,744	11,113	7,389	4,742	2,762	1,226
EMS($\hat{Y}_s$)	9,612	5,362	3,382	--	1,477	945	549	243
var(Y'_s)	9,466	5,317	3,374	2,235	1,484	952	554	246
EMS($\hat{Y}_e$)	14,222	8,102	5,169	--	2,286	1,468	856	380
var(Y'_e)	14,458	8,197	5,217	3,461	2,300	1,476	860	382

[1] Prepared in connection with research sponsored by the National Science Foundation

REFERENCES

1) Hartley, H. O. and A. Ross. Unbiased ratio estimators. Nature 174:270, 1954.

2) Goodman, L. A. and H. O. Hartley. The precision of unbiased ratio-type estimators. Jour. Amer. Stat. Assoc. 53:451-509, 1958.

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

4) Tukey, J. W. Keeping moment-like sampling computations simple. Annals Math. Stat. 27: 37-54, 1956.

5) Robson, D. S. An unbiased sampling and estimation procedure for creel censuses of fishermen. Biometrics 16, No. 2, 1960.