6-1991

Generating Unbiased Ratio and Regression Estimators

William (Bill) H. Williams
CUNY Hunter College

Reprinted from
BIOMETRICS
THE BIOMETRIC SOCIETY, Vol. 17, No. 2, June 1961

GENERATING UNBIASED RATIO AND REGRESSION ESTIMATORS

W. H. WILLIAMS

Bell Telephone Laboratories, Incorporated
Murray Hill, New Jersey, U. S. A.

1. INTRODUCTION

Information collected on a concomitant variate is often used in finite sampling theory to create more precise estimators of population characteristics. This supplementary information is obtained in addition to the characteristic under study and some aspects of it may be derived from sources other than the sample itself. It may be either qualitative or quantitative. For example, suppose that the variate under consideration in a sample survey is the number of dairy cattle per farm y and that at the time of the survey the number of grazing acres per farm x is also obtained. It may then be known from census data that the total number of grazing acres in the entire area is $N\mu_X$ and the mean per farm is μ_X. Analytically, we have a random sample of n pairs (y_i, x_i), $i = 1, \cdots, n$, from a population of size N and the population x-mean is known exactly. The problem is to estimate the population mean μ_Y.

A general class of estimators designed to utilize this supplementary information includes ratio and regression estimators. These estimators are described in textbooks on the subject, see for example Cochran [1953]. Additional developments have been presented by Hartley and Ross [1954], Nieto [1958] and Robson [1957].

The two classical ratio estimators are the ratio of means estimator $\tilde{y} = (\bar{y}/\bar{x})\mu_X$ and the mean of ratios estimator $\hat{y} = \mu_X \sum_{i=1}^{n} r_i/n$ where $\bar{y}$ and $\bar{x}$ are sample means and $r_i = y_i/x_i$. It is well known that these estimators are biased. The usual regression estimator is obtained by evaluating the least squares line of best fit $y = \bar{y} + b(x - \bar{x})$ at the point μ_X giving $\hat{y}_b = \bar{y} + b(\mu_X - \bar{x})$ as a regression estimator of μ_Y. This estimator is biased if the assumption of a linear model is not valid.

The generation of some exactly unbiased ratio and regression estimators is discussed in this paper. Specifically, we classify an estimator as of the regression type if it is invariant under location and scale changes in x and if it undergoes the same location and scale changes

as the y variate. A ratio estimator has analogous properties but for scale changes only.

2. DERIVATION OF UNBIASED ESTIMATORS FOR SIMPLE RANDOM SAMPLING

To generate unbiased estimators, consider the following sampling procedure. At step one, select with equal probability one of all possible splits of the population into s mutually exclusive groups of size[1] n/k, i.e., $N = sn/k$. At the second stage, select randomly without replacement k of the groups from the total number of groups s of that particular split of the population. This gives a sample of size n.

Now consider the conditional distribution for a particular set of s groups. Attached to each of these groups there are characteristics[2] $\bar{y}^{(i)}$, $\bar{x}^{(i)}$, $b^{(i)}$, $i = 1, \cdots, s$, where $\bar{y}^{(i)}$ and $\bar{x}^{(i)}$ are means of the n/k units in the group and $b^{(i)}$ is as yet an unspecified function of the y and x of that group. For a given split and a random selection of groups, the expectations of $\bar{y}^i$ and $\bar{x}^i$, $i = 1, \cdots, k$, are μ_Y and μ_X respectively; that is, they are conditionally unbiased. Furthermore,

$$\left(1 - \frac{k}{s}\right) \frac{1}{k(k-1)} \sum_{i=1}^{k} (b^i - \bar{b})(\bar{x}^i - \bar{x}) \tag{1}$$

is an unbiased estimator of Cov $(\bar{b}, \bar{x})$ where $\bar{b} = \sum_{i=1}^{k} b^i/k$.

Hence if $g = \bar{y} + \bar{b}(\mu_X - \bar{x})$ then $E(g) = \mu_Y - \text{Cov}\ (\bar{b}, \bar{x})$ and

$$T_k = \bar{y} + \bar{b}(\mu_X - \bar{x}) + \left(1 - \frac{n}{N}\right) \frac{1}{k(k-1)} \sum_{i=1}^{k} (b^i - \bar{b})(\bar{x}^i - \bar{x}) \tag{2}$$

is a conditionally unbiased estimator of μ_Y. It is then unbiased unconditionally.

This approach is valid for any defined form of the coefficient $b^{(i)}$; T_k will remain unbiased. If $b^{(i)}$ has a form which is invariant under linear x and y transformation (say least squares form) then T_k is classified as a regression estimator. If $b^{(i)} = \bar{y}^{(i)}/\bar{x}^{(i)}$ (say), then T_k falls into the class of a ratio estimator.

This procedure is used to generate the unbiased estimators; in practice a simple random sample would be drawn and to compute T_k it would be split randomly into groups, see Section 7 for an example. The latter operation is equivalent to the generating procedure which allows a particular split-sample to arise in $\binom{s}{k}(N - n)!/[(n/k)!]^{s-k}$

[1]It is assumed that this relationship is true in terms of integers.

[2]Superscripts will be used to specify the groups. They will be used with parentheses when the reference is to the entire population of s groups and without parentheses when referring to the sample of k groups.

ways while splitting the simple random sample allows a particular split-sample to arise in only one way. The unbiasedness is preserved by either procedure.

The argument is easily generalized to p auxiliary variates.

3. SPECIFIC ILLUSTRATIONS IN SIMPLE RANDOM SAMPLING

A form of interest is

$$b^{(i)} = \frac{\sum_{j=1}^{n/k} (y_j - \bar{y}^{(i)})(x_j - \bar{x}^{(i)})}{\sum_{j=1}^{n/k} (x_j - \bar{x}^{(i)})^2}, \qquad i = 1, \cdots, s, \tag{3}$$

the least squares slope form.

In this case T_k bears much resemblance to $\hat{y}_b$ and might be thought of as possessing an additional component which is required to compensate for possible bias in $\hat{y}_b$. This is not exactly true of course, because the first two terms of T_k are not exactly the two terms of $\hat{y}_b$. However, in this case (3), it seems natural to make some remarks on the efficiency of T_k.

The variance of T_k depends very much on the form of the $b^{(i)}$ coefficients. In fact, until the form of $b^{(i)}$ is specified little can be said about the variance of T_k. One can imagine choices which would lead to poor efficiency indeed. However, T_k in this case has coefficients in the least squares slope form and it is natural to ask how it compares with $\hat{y}_b$ when a linear model is assumed, for then $\hat{y}_b$ has optimum variance properties. But with this assumption, $\hat{y}_b$ also possesses unbiasedness and the advantage of T_k is unbiasedness in situations in which $\hat{y}_b$ is not unbiased. However, one would like the efficiency of T_k to compare favorably even in this linear model case. So by assuming a linear model and a normal x-distribution, it is easily found that $V(\hat{y}_b)/V(T_k) = (n-2)(n-6)/(n-3)(n-4)$, $k = 2$ and $n > 6$. This expression is less than one but approaches one as n gets larger and, for example, when $n = 15, 25$ is equal to 0.89 and 0.95. Thus we see that one does not lose all the efficiency brought about by the use of an auxiliary variate and that $[V(T_k) - V(\hat{y}_b)]/V(\hat{y}_b)$ is $O(n^{-1})$.

Furthermore, the role of k also depends upon the choice of the $b^{(i)}$. For example, in the special case of the previous paragraph, if the number of groups is regarded as variable, $V(\hat{y}_b)/V(T_k)$ will be found to have a maximum at $k = \sqrt{n/3}$. Thus for this form of the $b^{(i)}$, the optimum number of groups is $\sqrt{n/3}$. Other forms of the $b^{(i)}$ would yield other results.

Another possible choice is

$$b^{(i)} = \sum_{j=1}^{n/k} y_j x_j \Big/ \sum_{j=1}^{n/k} x_j^2 .$$

In this form T_k is a ratio estimator and it is unbiased even if the linear relationship of y and x does not pass through the origin. But characteristically the variance will be inflated by such a relationship.

Next, if $b^{(i)} = \bar{y}^{(i)}/\bar{x}^{(i)} = r^{(i)}$, T_k will reduce to the form

$$T_k = \bar{r}\mu_X + \frac{Nk - n}{N(k-1)}(\bar{y} - \bar{r}\bar{x}) \tag{4}$$

where $\bar{b}$ is denoted $\bar{r}$. It will be noted that when $k = n$, $T_k = y'$, the unbiased ratio estimator presented by Hartley and Ross [1954]. The efficiency of this form of T_k has been examined in detail by Goodman and Hartley [1958] and Robson [1957]. Robson presents an exact variance formula for finite populations.

Finally, consider $b^{(i)} = r^{(i)} = (k/n) \sum_{j=1}^{n/k} r_j$, $r_j = y_j/x_j$, then $\bar{b} = \bar{r} = \sum_{j=1}^{n} r_j/n$ which does not depend upon the particular split of the population. Now if, after substitution of this form into T_k , the estimator is averaged over all possible splits of the sample into groups of size n/k it will be found that the result is again the Hartley-Ross unbiased ratio estimator. This averaging process is indicated by a star, i.e., T_k^* .

Other forms could, of course, be considered.

4. STRATIFIED SAMPLING

Since a bias may be magnified relative to the standard deviation, stratified sampling may perhaps be regarded as the most important application of unbiased estimators. Their separate use within strata requires exact knowledge of the population strata means but is straightforward. We now develop a combined stratified estimator.

Consider L strata of size N_t , $t = 1, \cdots, L$ with $\sum_{t=1}^{L} N_t = N$, and again consider the sampling in two stages. At the first stage select with equal probability one of the possible splits of each stratum into s groups of size n_t/k, $t = 1, \cdots, L$. Then $N_t = sn_t/k$. At the second stage select k groups with equal probability and without replacement from each of the strata, giving a sample of size n_t in the t-th stratum, $\sum_{t=1}^{L} n_t = n$.

For a given split and a random selection of groups

$$\bar{y}_{st}^{i} = \sum_{t=1}^{L} (N_t/N)\bar{y}_t^i \quad \text{and} \quad \bar{x}_{st}^{i} = \sum_{t=1}^{L} (N_t/N)\bar{x}_t^i$$

are unbiased estimators of μ_Y and μ_X respectively, where $\bar{y}_t^i$ and $\bar{x}_t^i$ denote means of the i-th group in the t-th stratum. Also we can consider

a coefficient $b_{st}^{(i)}$ which is as yet unspecified in form but utilizes the set of elements in the i-th group of all strata. For example,

$$b_{st}^{(i)} = \frac{\sum_{t=1}^{L}\sum_{j=1}^{n_t/k}(y_{tj} - \bar{y}_t^{(i)})(x_{tj} - \bar{x}_t^{(i)})}{\sum_{t=1}^{L}\sum_{j=1}^{n_t/k}(x_{tj} - \bar{x}_t^{(i)})^2}, \qquad i = 1, \cdots, s \tag{5}$$

is an over-all slope estimator.

Next we note that

$$\bar{y}_{st} = \sum_{t=1}^{L}(N_t/N)\bar{y}_t = \sum_{i=1}^{k}\bar{y}_{st}^{i}/k \tag{6}$$

where $\bar{y}_t$ is the mean of the n_t observations in the t-th stratum (similarly for $\bar{x}_{st}$) and finally that a conditionally unbiased estimator of Cov $(\bar{b}_{st}, \bar{x}_{st})$ is given by

$$(1 - n/N)\frac{1}{k(k-1)}\sum_{i=1}^{k}(\bar{x}_{st}^{i} - \bar{x}_{st})(b_{st}^{i} - \bar{b}_{st}).$$

Consequently, if $g = \bar{y}_{st} + \bar{b}_{st}(\mu_X - \bar{x}_{st})$ then $E(g) = \mu_Y -$ Cov $(\bar{b}_{st}, \bar{x}_{st})$ and therefore

$$T_{k(st)} = \bar{y}_{st} + \bar{b}_{st}(\mu_X - \bar{x}_{st}) + \left(1 - \frac{n}{N}\right)\frac{1}{k(k-1)}\sum_{i=1}^{k}(\bar{x}_{st}^{i} - \bar{x}_{st})(b_{st}^{i} - \bar{b}_{st}) \tag{7}$$

is a combined stratified unbiased estimator of μ_Y. Note that since $N_t = sn_t/k$, $k/s = n/N$. Nieto [1958] discussed the efficiency of the estimator (7) (for sampling with replacement) in detail.

Again the generalization to p auxiliary variates is straightforward.

As a specific illustration consider the case in which

$$b_{st}^{(i)} = \bar{y}_{st}^{(i)}/\bar{x}_{st}^{(i)} = r_{st}^{(i)}.$$

Then $T_{k(st)}$ reduces to

$$T_{k(st)} = \bar{r}_{st}\mu_X + \frac{Nk - n}{N(k-1)}(\bar{y}_{st} - \bar{r}_{st}\bar{x}_{st}). \tag{8}$$

In the special case that $N_t = \bar{N}$, $n_t = \bar{n}$ for all t and $k = \bar{n}$, $s = \bar{N}$ then

$$T_{k(st)} = \bar{r}_{st}\mu_X + \frac{(\bar{N} - 1)\bar{n}}{(\bar{n} - 1)\bar{N}}(\bar{y}_{st} - \bar{r}_{st}\bar{x}_{st}), \tag{9}$$

which is a generalized Hartley-Ross estimator.

Finally, we again consider an averaging of T_k over all possible splits of the sample into groups of size n_t/k, $t = 1, \cdots, L$. For this, the

coefficient is taken in the form $b_{st}^{(i)} = r_{st}^{(i)} = \sum_{t=1}^{L} (N_t/N) r_t^{(i)}$ where $r_t^{(i)} = (k/n_t) \sum_{j=1}^{n_t/k} (y_j/x_j)$. Therefore,

$$\bar{r}_{st} = \sum_{i=1}^{k} r_{st}^{i} \Big/ k = \sum_{t=1}^{L} (N_t/N) \sum_{j=1}^{n_t} (y_j/x_j) \Big/ n_t = \sum_{t=1}^{L} (N_t/N) \bar{r}_t$$

and some algebraic reduction will show that $T_{k(st)}$ averaged over all possible splits is equal to

$$T_{k(st)}^{*} = \bar{r}_{st}\mu_X + (\bar{y}_{st} - \bar{r}_{st}\bar{x}_{st}) + \left(1 - \frac{n}{N}\right) \sum_{t=1}^{L} \frac{N_t^2}{N^2} \frac{(\bar{y}_t - \bar{r}_t\bar{x}_t)}{(n_t - 1)} , \qquad (10)$$

which does not quite reduce to a form similar in appearance to Equation (8) and the Hartley-Ross estimator.

As before other selections of coefficients will yield other unbiased estimators.

5. MULTISTAGE SAMPLING

We consider a population with N primaries of equal size $\bar{M}$ and the following sampling scheme. First select n primaries from the N available with equal probability with or without replacement. Then select with equal probability one of the splits of each of the primaries into s groups of size $\bar{m}/k$. Then with equal probability and without replacement draw k of the groups so that the sample size is $\bar{m}$ in each selected primary.

Consider now the conditional distribution for a fixed set of primaries and a fixed split of the primaries into s groups each. Then by Section 4, Equation (11) is an unbiased estimator of $\bar{\bar{Y}}_n$, the population mean of the n selected primaries.

$$T_{k(M)} = \bar{\bar{y}} + \bar{b}(\mu_X - \bar{\bar{x}}) + \left(1 - \frac{\bar{m}}{\bar{M}}\right) \frac{1}{k(k-1)} \sum_{i=1}^{k} (b^i - \bar{b})(\bar{x}^i - \bar{\bar{x}}) \qquad (11)$$

where

$$\bar{y}^i = (k/n\bar{m}) \sum_{t=1}^{n} \sum_{j=1}^{\bar{m}/k} y_{tj}^{i} , \qquad \bar{\bar{y}} = (1/k) \sum_{i=1}^{k} \bar{y}^i = (1/n\bar{m}) \sum_{t=1}^{n} \sum_{j=1}^{\bar{m}} y_{tj}$$

and similarly for x. The coefficient $b^{(i)}$ is again arbitrary in form.

Finally, the expectation of $T_{k(M)}$ over all possible primary selections is the average of $\bar{\bar{Y}}_n$ over all possible primary selections; this is μ_Y and $T_{k(M)}$ is unbiased in multi-stage sampling.

Again the selection of the coefficients yields estimators of different types. For example, an unbiased ratio estimator of the Hartley-Ross type generalized to multistage sampling can be obtained.

6. VARIANCE ESTIMATION

It is interesting to notice that the same two-stage sampling scheme can be used to form an estimate of the variance of T_k. First assume a negligible n/N (or k/s) and a fixed set of uncorrelated groups. T_k can now be written

$$T_k = \bar{y} - \frac{1}{k(k-1)} \sum_{i \neq j}^{k} b^i(\bar{x}^j - \mu_X) \tag{12}$$

and its conditional variance can be expressed in terms of the variances and covariances of the components in (12). Since T_k is conditionally unbiased this variance has expectation equal to the over-all variance. Substitution of unbiased estimators for each of the terms of the variance (plus some terms of zero expectation) yields (13) as an unbiased estimator of the variance of T_k.

$$v(T_k) = T_k^2 - \frac{1}{k(k-1)} \sum_{i \neq j}^{k} \bar{y}^i \bar{y}^j. \tag{13}$$

Although this procedure is unbiased it can be subject to high sampling error, particularly for small k.

TABLE 1

A SIMPLE EXAMPLE OF THE ESTIMATORS

Sample Number	Pairs in Sample	$\bar{y}$	T_2			T_2^*
			Split 1	Split 2	Split 3	
1	$P_1P_2P_3P_4$	3.500	7.167	6.667	6.500	6.778
2	$P_1P_2P_3P_5$	5.250	8.917	8.250	7.917	8.361
3	$P_1P_2P_3P_6$	7.500	11.000	10.167	9.667	10.278
4	$P_1P_2P_4P_6$	8.750	11.917	10.250	9.917	10.694
5	$P_1P_2P_4P_5$	6.500	10.000	8.667	8.500	9.056
6	$P_2P_3P_4P_5$	7.500	8.167	7.667	7.500	7.778
7	$P_2P_3P_5P_6$	11.500	10.000	8.667	8.500	9.056
8	$P_2P_3P_4P_6$	9.750	9.417	8.750	8.417	8.861
9	$P_1P_3P_4P_5$	7.250	9.417	8.750	8.417	8.861
10	$P_1P_3P_4P_6$	9.500	11.000	10.167	9.500	10.222
11	$P_1P_3P_5P_6$	11.250	11.917	10.250	9.917	10.694
12	$P_3P_4P_5P_6$	13.500	7.167	6.667	6.500	6.778
13	$P_1P_4P_5P_6$	12.500	11.000	10.167	9.667	10.278
14	$P_2P_4P_5P_6$	12.750	8.917	8.250	7.917	8.361
15	$P_1P_2P_5P_6$	10.500	13.167	10.667	10.500	11.444

7. NUMERICAL ILLUSTRATION

To illustrate T_k a small population consisting of the six pairs $P_i = (y_i , x_i)$, $i = 1, 2, \cdots , 6$, with $y = x^2$ and $x_i = 0, 1, 2, \cdots , 5$ was completely examined. Table 1 presents the values of $\bar{y}$, T_2 and T_2^* [using Equation (3)] for all possible samples of size four. For T_2 , each of the possible samples was split into two groups of two in all possible ways, and the value of T_2 was computed for each. The three distinct values of T_2 for each sample are presented in the table. The numbering of the splits within a sample is of course arbitrary. It is readily verified that the average value of each of $\bar{y}$, T_2 and T_2^* is the population mean $\mu_Y = 9.167$. Furthermore, the exact population variances of $\bar{y}$, T_2 and T_2^* are 7.914, 2.281 and 1.886 respectively.

As a second example, the six pairs (y_i , x_i) were taken as follows: (0,2), (1, 3), (2, 5), (4, 9), (8, 14), (9, 15). All possible samples of size four were drawn and for each sample $\bar{y}$, y', $\hat{y}_b$, T_k (for all possible splits) and T_k^* were computed. A summary of the computations is presented in Table 2.

TABLE 2

ILLUSTRATION OF RELATIVE EFFICIENCIES

	Estimator				
	$\bar{y}$	y'	$\hat{y}_b$	T_k	T_k^*
Expectation	3.937	4.000	3.961	4.000	4.000
Bias	−0.063	0.000	−0.039	0.000	0.000
Variance	0.120	0.233	0.027	0.033	0.022
Mean Square Error	0.124	0.233	0.029	0.033	0.022

REFERENCES

Cochran, W. G., [1953]. *Sampling Techniques.* New York, John Wiley and Sons.

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

Hartley, H. O., and Ross, A., [1954]. Unbiased ratio estimators. *Nature 174,* 270–271.

Nieto, J., [1958]. *Unbiased ratio estimators in stratified sampling.* Unpublished M.S. Thesis, Iowa State University, Ames, Iowa.

Robson, D. S., [1957]. Application of multivariate polykays to the theory of unbiased ratio-type estimation. *J. Amer. Stat. Assoc.* 52. 511–522.