

UNBIASED ESTIMATION

By: W. H. Williams, McMaster University

1. INTRODUCTION

A favourite method in sampling theory of increasing the precision of estimates is the utilization of auxiliary information. Analytically, we have a random sample of n pairs (y_i, x_i) drawn from a population of size N and the problem is to estimate the population mean μ_Y relative to the assumption that the population mean μ_X is known exactly. Ratio and regression estimators have been designed for this problem and they are described in detail along with illustrations in the textbooks on the subject, see for example Cochran [1953]. Additional contributions have been made by Hartley and Ross [1954], Nieto [1958] and Robson [1957].

First, we specify that an estimator is of the regression type if it is invariant under location and scale changes in x and undergoes the same location and scale changes as the y variate. A ratio estimator has these properties for scale changes only.

The two common ratio estimators are known to be biased. These estimators are the ratio of means estimator $\hat{y} = \bar{y}\mu_X/\bar{x}$ 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 $r_i = y_i/x_i$. The classical regression estimator is obtained by evaluating the least squares line of best fit 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.

Some exactly unbiased ratio and regression estimators are presented in this paper.

2. A PROCEDURE FOR UNBIASED ESTIMATION

The following sampling procedure can be used to derive unbiased estimators. The scheme consists of two steps. First select with equal probability one of all possible splits of the population in mutually exclusive groups of size n/k . Let s be the number of groups and assume that n/k divides N so that s is an integer. At the second step select randomly without replacement k of the groups from the total number of groups (s) obtained in step one. Thus a sample of size n is obtained.

Next consider the conditional distribution for a fixed set of s groups. These groups have y and x means which are denoted $\bar{y}^{(i)}$ and $\bar{x}^{(i)}$ ($i = 1, 2, \dots, s$); also let $b^{(i)}$ denote an as yet 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}^1 and \bar{x}^1 $i = 1, 2, 3, \dots, k$ are μ_Y and μ_X respectively; furthermore

$$(1 - \frac{n}{N}) \frac{1}{k(k-1)} \sum_{i=1}^k (b^i - \bar{b})(\bar{x}^i - \bar{x}) \quad (1)$$

is an unbiased estimator of $\text{Cov}(\bar{b}, \bar{x})$ where $\bar{b} =$

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

Hence $Eg = \mu_Y - \text{Cov}(\bar{b}, \bar{x})$ where $g = \bar{y} + \bar{b}(\mu_X - \bar{x})$ showing that

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

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

T_k remains unbiased for any defined form of the coefficients $b^{(i)}$. It is classified as a regression estimator if $b^{(i)}$ has a form which is invariant under linear x and y transformation (for example least squares form). If $b^{(i)} = \bar{y}^{(i)}/\bar{x}^{(i)}$ (say) then T_k falls into the class of a ratio estimator.

3. SOME ILLUSTRATIONS

It is natural to consider $b^{(i)}$ in the least squares form shown in equation (3).

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

$$(i) = 1, 2, \dots, s.$$

In this case T_k is similar to \hat{y}_b but possesses an additional term which compensates for possible bias in \hat{y}_b . One also wonders about the efficiency of T_k when compared with \hat{y}_b , for when the linear model assumption is valid \hat{y}_b possesses certain optimum variance properties. However, \hat{y}_b is then also unbiased and the advantage of T_k is unbiasedness in situations in which \hat{y}_b is not unbiased. It would be desirable that the variance of T_k compare favourably with \hat{y}_b even under the assumption of a linear model. A discussion of this case showing that the loss in efficiency is $O(n^{-1})$ can be found in Williams [1958].

Another possible choice is $b^{(i)} = \frac{\sum_{j=1}^{n/k} y_j x_j}{\sum_{j=1}^{n/k} x_j^2}$. 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 $b^{(i)} = \bar{y}^{(i)}/\bar{x}^{(i)} = \bar{r}^{(i)}$, T_k will reduce to the form

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

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].

Finally, consider $b^i = r^i = \frac{k n_j}{n} r_j^i$, $r_j = y_j/x_j$ 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.

Other forms could be considered.

4. EXTENSION TO STRATIFIED SAMPLING

Unbiased estimators are important in stratified sampling as a bias may be magnified relative to the standard deviation. Their separate use 'within strata' requires exact knowledge of the population strata means but is straightforward. A 'combined' stratified estimator can also be developed.

Suppose that there are L strata with N_t units in the t th stratum, $t = 1, 2, \dots, L$ with $\sum_{t=1}^L N_t = N$. The sampling is again done in two stages. First select with equal probability one of all possible splits of each stratum into s groups of size n_t/k . 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, μ_Y and μ_X are estimated unbiasedly by $\bar{y}_{st} = \sum_{t=1}^L N_t \bar{y}_t^i / N$ and $\bar{x}_{st} = \sum_{t=1}^L N_t \bar{x}_t^i / N$ 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

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

(i) = 1, 2, . . . s

is an overall slope estimator.

Next, notice that

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

where \bar{y}_t^i is the mean of the n_t observations in the t th stratum (similarly for \bar{x}_{st}^i) and that a conditionally unbiased estimator of $\text{Cov}(\bar{b}_{st}, \bar{x}_{st})$ is given by $(1 - \frac{n}{N}) \frac{1}{k(k-1)} \sum_{i=1}^k (\bar{x}_{st}^i - \bar{x}_{st})(\bar{b}_{st}^i - \bar{b}_{st})$. Therefore, if $g = \bar{y}_{st} + \bar{b}_{st}(\mu_X - \bar{x}_{st})$ then $Eg = \mu_Y - \text{Cov}(\bar{b}_{st}, \bar{x}_{st})$ and

$$T_k(st) = \bar{y}_{st} + \bar{b}(\mu_X - \bar{x}_{st}) \quad (7)$$

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

is a combined stratified unbiased estimator of μ_Y . Note that $N_t = sn_t/k$ implies $k/s = n/N$.

Again the generalization to p auxiliary variates is straightforward.

To illustrate the stratified estimator take first $b_{st}^{(i)} = \bar{y}_{st}^i / \bar{x}_{st}^i = r_{st}^i$. Then $T_k(st)$ reduces to

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

When $N_t = \bar{N}$, $n_t = \bar{n}$ for all t and $k = \bar{n}$, $s = \bar{N}$ then

$$T_k(st) = \bar{r}_{st} \mu_X + \frac{(\bar{N}-1)\bar{n}}{(\bar{n}-1)\bar{N}} (\bar{y}_{st} - \bar{r}_{st} \bar{x}_{st}) \quad (9)$$

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 , $t = 1, 2, \dots, L$. For this, the coefficient is taken in the form $b_{st}^i = r_{st}^i = \sum_{t=1}^L \frac{N_t}{N} r_t^i$ where $r_t^i = \frac{k}{n_t} \sum_{j=1}^{n_t/k} \frac{y_{tj}^i}{x_{tj}^i}$. Therefore, $\bar{r}_{st} = \sum_{i=1}^k r_{st}^i / k = \sum_{t=1}^L \frac{N_t}{N} \sum_{j=1}^{n_t/k} \frac{y_{tj}^i}{x_{tj}^i} / n_t = \sum_{t=1}^L \frac{N_t}{N} \bar{r}_t$ and some

algebraic reduction will show that $T_k(st)$ averaged over all possible splits is equal to

$$T_k^*(st) = \bar{r}_{st} \mu_X + (\bar{y}_{st} - \bar{r}_{st} \bar{x}_{st}) \quad (10)$$

$$+ (1 - \frac{n}{N}) \sum_{t=1}^L \frac{N_t^2}{N^2} \frac{(\bar{y}_t - \bar{r}_t \bar{x}_t)}{(n_t - 1)}$$

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 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 M/k . Then with equal probability and without replacement draw k of the groups so that the sample size is 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{y}_n , the population mean of the n selected primaries.

$$T_{k(M)} = \bar{y} + b(\mu_X - \bar{x}) \quad (11)$$

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

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

$\frac{1}{n\bar{m}} \sum_{t=1}^n \sum_{j=1}^{\bar{m}} y_{tj}$ 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{y}_n over all possible primary selections; this is μ_Y and $T_{k(M)}$ is unbiased in multistage 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.

This delivered paper was submitted to *Biometrics* in February, 1960.

REFERENCES

- Cochran, W. G. [1953]. *Sampling Techniques*. New York: John Wiley and Sons.
- Hartley, H. O. and Ross, A. [1954]. Unbiased ratio estimators. *Nature* 174, 270.
- Nieto, J. [1958]. Unbiased ratio estimators in stratified sampling. Contributed paper, *Inst. of Math. Stat., Ames, Iowa*.
- Robson, D. S. [1957]. Application of multivariate polykays to the theory of unbiased ratio-type estimation. *J. Amer. Stat. Assoc.* 52: 511.
- Williams, W. H. [1958]. Unbiased regression estimators. Contributed paper, *Inst. of Math. Stat., Ames, Iowa*.