Survey Methodology, December 1990
Vol. 16, No. 2, pp. 279–282
Statistics Canada

Forgot the Sampling Scheme at the Estimation Stage?

SHIBDAS BANDYOPADHYAY[1]

ABSTRACT

For a class of linear unbiased estimators in a class of sampling schemes, it is shown that one can forget the weights used for sample selection while estimating a population ratio by a ratio of two unbiased estimators, respectively of the numerator and the denominator defining the population ratio. This class of schemes includes commonly used sampling schemes such as unequal probability sampling with or without replacement, stratified proportional allocation sampling with unequal selection probabilities and without replacement in each stratum, *etc.*

KEY WORDS: Ratio of unweighted totals; Symmetric sampling.

1. INTRODUCTION

Let m be the number of adult literates among t adult members in a sample of n families drawn from a given population. Let the population adult literacy rate R be estimated as $r = m/t$. Similarly, for a two- way table giving percentage distribution of persons by age-group and sex, let a cell entry be estimated by a ratio (multiplied by 100 to make it a percentage) of the number of persons classified into the cell to the total number of persons, in the sample of n families.

Irrespective of the method of selection of the families, this simple ratio of two unweighted totals for estimating a ratio or a percentage distribution is acceptable to many non-statistical users. Indeed, in some survey reports, tables giving percentage distributions or rates are so computed, as if the sampling scheme had been a self-weighting one.

If, however, the sampling scheme for selecting the n families had been a (single stage) PPSWOR, one is expected to go about finding weighted totals for obtaining unbiased estimators of numerators and respective denominators before computing a ratio or a percentage distribution.

This study shows that, for sampling schemes such as a single stage PPSWOR but without any further assumptions,

(i) a ratio of two unweighted totals estimates the corresponding population ratio, as a ratio of an *unbiased estimator* of the numerator to an *unbiased estimator* of the respective denominator;

(ii) there is a class of sampling schemes, other than self-weighting designs, for which (i) holds. This class includes one stage unequal probability, with or without replacement, sampling schemes and stratified proportional allocation sampling with unequal probability without replacement selection in each stratum.

2. SYMMETRIC SAMPLING SCHEMES

Consider a finite population consisting of N units $U_1, U_2, \ldots, U_N$. Let Y_i and X_i, denote the values of two study variables, Y and X respectively, associated with the unit U_i, $i = 1, 2, \ldots, N$.

[1] Shibdas Bandyopadhyay; Applied Statistics, Surveys and Computing Division, Indian Statistical Institute, 203 Barrackpore Trunk Road, Calcutta 700 035, India.

The problem is to estimate a rate or a ratio $R = T(Y)/T(X)$ where $T(Y) = Y_1 + Y_2 + \ldots + Y_N$, and $T(X)$ is similarly defined with the variable X.

The usual procedure is to estimate $T(Y)$ and $T(X)$ unbiasedly and take their ratio to estimate R. The aim of this paper is to follow the same procedure in such a way that the ratio becomes free of the selection probabilities of the sample units.

Fix a sampling scheme.

Let S denote the set consisting of all possible samples such that $p(s) > 0$, where $p(s)$ denotes the probability of drawing the sample s, and $\sum_{s \in S} p(s) = 1$.

For s in S and $i = 1, 2, \ldots, N$,

$n(i,s)$ = the number of times U_i is included in s, and $\alpha_i = \sum_{s \in S} n(i,s)$, the number of times U_i is included in all possible samples.

S, $p(s)$, α_i depend on the sampling scheme.

Definition 2.1. A sampling scheme is said to be symmetric if $\alpha_i = \alpha$, for all $i = 1, 2, \ldots, N$.

The following estimator, based on the sample s, in the class of linear unbiased estimators of Godambe (1955) for $T(Y)$, was studied by Bandyopadhyay *et al.* (1977).

$$T(Y,s) = \sum_{i=1}^{N} Y_i n(i,s) \alpha_i^{-1} p^{-1}(s). \tag{2.1}$$

Clearly, $T(Y,s)$ is unbiased for $T(Y)$. An estimator of the ratio $R = T(Y)/T(X)$, as a ratio of an unbiased estimator of $T(Y)$ to an unbiased estimator of $T(X)$, based on a sample s, is

$$R(s) = T(Y,s)/T(X,s) = \sum_{i=1}^{N} Y_i n(i,s) \alpha_i^{-1} \Big/ \sum_{i=1}^{N} X_i n(i,s) \alpha_i^{-1}. \tag{2.2}$$

For symmetric sampling schemes, $\alpha_i = \alpha$ for all i and (2.2) becomes

$$R(s) = \sum_{i=1}^{N} Y_i n(i,s) \Big/ \sum_{i=1}^{n} X_i n(i,s) =$$

$$\frac{\text{unweighted total of } Y \text{ values in the sample}}{\text{unweighted total of } X \text{ values in the sample}} \tag{2.3}$$

and the above observations are summarized in the following theorem.

Main theorem. For a symmetric sampling scheme, a ratio of two unweighted totals estimates the corresponding population ratio as a ratio of an unbiased estimator of the numerator to an unbiased estimator of the respective denominator, but the estimated ratio does not involve the selection probabilities of the population units in the sample.

It may be noted that the inclusion probabilities of the units in the sample need not be equal for symmetric sampling schemes. Thus, symmetric sampling schemes need not be self-weighting. Self-weighting designs require constancy of $\alpha_i p(s)$ for all i and s, and constancy of $\alpha_i p(s)$ for all i and s does not make the sampling scheme symmetric.

For a non-symmetric scheme, (2.2) is easy to compute as α_i's are easy to compute in most cases and there is no need to compute inclusion probabilities.

For without replacement sampling of n units, there are $\binom{N-1}{n-1}$ (un-ordered) samples containing a given unit U_i, so $\alpha_i = \binom{N-1}{n-1}$ for all i and thus, in particular, PPSWOR is symmetric. It may be noted that not all PPSWOR schemes result in $\binom{N}{n}$ possible samples. As noted in Connor (1966), in some cases systematic PPS samples in a pre-determined order or randomized PPS systematic sampling may result in zero probability for some set of n units. The result applies if the PPSWOR scheme is such that no joint inclusion probability of any set of n units is zero.

For with replacement sampling of n units, there are N^n (ordered) samples and so $\alpha_i = nN^{n-1}$ for all i and thus, in particular, PPSWR is symmetric.

For PPSWOR in each of k strata, the α-value for each unit in the jth stratum is

$$\alpha_j = \frac{n_j}{N_j} \prod_{i=1}^{K} \binom{N_i}{n_i}$$

which becomes a constant when allocation is proportional and if no joint probability of any set of units in any stratum is zero, where N_j and n_j are respectively the population and sample sizes for the jth stratum, $j = 1, 2, \ldots, k$. Similar allocation may be made to make a multistage sampling scheme symmetric.

For PPSWR sampling, it may be noted that the unbiased estimator of $T(Y)$ given by (2.1) is inadmissible. This estimator can be improved upon by putting $n^*(i,s)$ and α_i^* respectively for $n(i,s)$ and α_i, where $n^*(i,s)$ is 1 if $n(i,s)$ is at least 1 and $n^*(i,s)$ is zero if $n(i,s)$ is zero, and α_i^* is α defined with $n^*(i,s)$. Here, $\alpha_i^* = N^n - (N-1)^n$, the number of (ordered) samples containing a given unit U_i. It has not been possible to obtain a mathematical expression for relative efficiency in a closed form for comparison, even with respect to PPSWR schemes.

Among the possibilitities for comparison of relative bias and relative efficiency, an empirical study is included for comparison with PPSWOR scheme. Another attractive possibility is to study large sample variance and bias using Taylor series expansions.

It is clear that it is not possible to estimate the variance of $R(s)$ without the weights or further assumptions. However, if s_1 and s_2 are two half-samples drawn by the same symmetric sampling scheme (like two independent PPSWOR samples of equal size), R is estimated as $[R(s_1) + R(s_2)]/2$, and its unbiased variance estimator is $[R(s_1) - R(s_2)]^2/4$.

If $T(X)$ is known, a ratio-type estimator for $T(Y)$ is $T(X)T(Y,s)/T(X,s)$, which may be improved as in Bandyopadhyay (1980) depending on whether or not the sampling fraction is more than half.

When the population units are divided into k non-overlapping clusters and the selection probability of the jth cluster is p_j then the design become symmetric with $\alpha_i = 1$ for all units in all the clusters. It may be noted that the sample size is the size of the selected cluster and so, the symmetric sampling schemes need not be fixed sample size designs.

3. EMPIRICAL STUDY ON BIAS AND MEAN SQUARE ERROR

Yates and Grundy (1953) considered the following three hypothetical populations, each with 4 population units.

	Population A				Population B				Population C			
X	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4
Y	0.5	1.2	2.1	3.2	0.8	1.4	1.8	2.0	0.2	0.6	0.9	0.8

The sampling scheme is to draw a sample of size $n = 2$ by PPSWOR using X-values as size measure. It is proposed to compare bias and mean square error of $R(s)$ with those of $R_{HT}^{(s)}$ where $R_{HT}^{(s)}$ is the ratio of the Horvitz-Thompson (1952) estimator of $T(Y)$ to that of $T(X)$. The result of the comparison is presented below.

Populations:	A	B	C
Relative bias of $R(s)$	0.02456	−0.02785	−0.00496
Relative bias of $R_{HT}(s)$	−0.00379	0.00552	0.00232
MSE of $R(s)$	0.2946	0.2946	0.0824
MSE of $R_{HT}(s)$	0.3159	0.3642	0.0690
Relative efficiency of $R(s)$ to $R_{HT}(s)$	1.0723	1.2362	0.8374

Though the absolute bias of $R(s)$ relative to R is more than that of $R_{HT}^{(s)}$ for the three populations, differences are small. $R(s)$ is a more efficient estimator in populations A and B and $R_{HT}(s)$ is more efficient in population C.

Since the above three populations are more extreme than the situations usually met with in practice, it is anticipated that $R(s)$ may be useful when the sampling scheme is not available at the estimation stage.

ACKNOWLEDGEMENT

The author sincerely appreciates active and constructive comments from the referees leading to the final form of this paper.

REFERENCES

BANDYOPADHYAY, S., CHATTOPADHYAY, A.K., and KUNDU, S.C. (1977). On estimation of population total. *Sankhyā*, Ser. C, 39, 28-42.

BANDYOPADHYAY, S. (1980). Improved ratio and product estimators. *Sankhyā*, Ser. C, 42, 45-49.

CONNOR, W.S. (1966). An exact formula for the probability that two specificed sampling units will occur in a sample drawn with unequal probabilities and without replacement. *Journal of the American Statistical Association*, 61, 384-396.

GODAMBE, V.P. (1955). A unified theory of sampling from finite population. *Journal of the Royal Statistical Society*, Ser. B, 17, 269-278.

HORVITZ, D.G., and THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

YATES, F., and GRUNDY, P.M. (1953). Selections without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society*, Ser. B, 15, 253-261.