A NOTE ON A BIASED ESTIMATOR IN SAMPLING WITH PROBABILITY PROPORTIONAL TO SIZE WITH REPLACEMENT

BY M. T. SUBRAHMANYA

Indian Statistical Institute, Calcutta

A finite population of N units $U_1, U_2, \cdots, U_N$ is considered. Let Y_i denote the value of the characteristic under study for the ith unit. It is desired to estimate the total $Y = \sum_{i=1}^{N} Y_i$ on the basis of a sample.

When sampling with equal probability, with replacement after each draw, the total Y may be estimated unbiasedly by

$$\hat{Y}' = (N/n) \sum_{i=1}^{n} y_i,$$

where y_i is the value recorded at the ith draw. Basu (1958) has shown that the estimator $\hat{Y}_d'$ based on distinct units in the sample, namely,

$$\hat{Y}_d' = (N/n_d) \sum_{i=1}^{n_d} y_{(i)}$$

is unbiased for Y and has on the average a smaller mean-square error than $\hat{Y}'$. Here n_d stands for the number of distinct units in a sample of size n and the suffix (i) indicates the ith distinct unit in the sample.

When sampling with probability proportional to size, with replacement after each draw (pps for brevity), the total Y may be estimated unbiasedly by

$$\hat{Y} = n^{-1} \sum_{i=1}^{n} (y_i/p_i),$$

where p_i is the probability of selecting the unit occurring at the ith draw. Basu (1958) presents for this design an unbiased estimator Y_d superior to $\hat{Y}$, which makes use only of the values recorded for the n_d distinct units. This estimator $\hat{Y}_d$ is—contrary to what is stated in Dalenius (1962) and in some other papers—*not* identical with

$$\hat{y}_d = n_d^{-1} \sum_{i=1}^{n_d} [y_{(i)}/p_{(i)}].$$

We stress here that $\hat{y}_d$ is not the same as $\hat{Y}_d$, except in some special cases. For instance, when $n_d = n - 1$,

$$\hat{Y}_d = n^{-1} \sum_{i=1}^{n-1} y_{(i)} / \sum_{i=1}^{n-1} p_{(i)} + [(n-1)/n]\hat{y}_d.$$

In fact, $\hat{y}_d$ is in general not unbiased. In view of the relative simplicity with which $\hat{y}_d$ may be computed, we will study the properties of $\hat{y}_d$ in some detail.

To begin with, we make the following remarks:

(a) $\hat{Y}$, $\hat{Y}_d$ and $\hat{y}_d$ are identical when $n = 1$ or 2 or when $Y_i/P_i = Y$ for $i = 1, 2, \cdots, N$.

(b) $\hat{Y}_d$ and $\hat{y}_d$ are identical when $p_i = N^{-1}$ for $i = 1, 2, \cdots, N$.

Received 13 October 1965; revised 4 February 1966.

TABLE 1

unit	probability	values of the characteristic (Y_i) for indicated populations		
	p_i	I	II	III
U_1	0.1	5	5	10
U_2	0.2	12	8	10
U_3	0.3	21	15	12
U_4	0.4	32	12	20
Total	1.0	70	40	52

TABLE 2

Squared bias (sb) *and mean square error* (mse) *for* $\hat{Y}$, $\hat{Y}_d$, $\hat{y}_d$ *and* $\hat{Y}'$ ($N = 4$, $n = 3$)

design	estimator	I population		II population		III population	
		sb	mse	sb	mse	sb	mse
pps with replacement	$\hat{Y}$	zero	33.333	zero	26.667	zero	92.000
	$\hat{Y}_d$	zero	29.680	zero	22.857	zero	84.596
	$\hat{y}_d$	0.250	31.833	0.090	21.500	0.250	105.033
simple random sampling with replacement	$\hat{Y}'$	zero	545.333	zero	77.333	zero	90.667

(c) $\hat{y}_d$ is biased for Y, except in the trivial cases mentioned above.

(d) It is interesting to note that the bias of $\hat{y}_d$ is doubled when the sample size is increased from 3 to 4.

PROOF. Denoting the bias of $\hat{y}_d$ by $B(\hat{y}_d)$ we see that

$$\begin{aligned} B(\hat{y}_d) &= E(\hat{y}_d - \hat{y}) \\ &= E(\hat{y}_d - \hat{Y}) \text{ since } E(\hat{Y}) = Y \\ &= \textstyle\sum_S (\hat{y}_{ds} - \hat{Y}_s)p(s), \end{aligned}$$

where S indicates the sum extended over all samples.

Here $\hat{y}_{ds}$ and $\hat{Y}_s$ are the values of the estimators $\hat{y}_d$ and $\hat{Y}$, as computed from the sample s, and $p(s)$ is the probability of selecting the sample s.

The terms corresponding to $n_d = 1$ or n, in which cases each distinct unit is repeated exactly n/n_d times, do not contribute to the above expression, because in those cases $\hat{y}_{d_s}$ and $\hat{Y}_s$ are the same. This fact helps us to calculate the bias.

When $n = 3$, we get

$$B_3(\hat{y}_d) = \tfrac{1}{2}\textstyle\sum_{i=1}^{N} \sum_{j>i}^{N} b_{ij},$$

where

$$b_{ij} = p_i p_j (p_j - p_i)(Y_i/p_i - Y_j/p_j).$$

This can also be written in the form $B_3(\hat{y}_d) = \frac{1}{2}\sum_{i=1}^{N} p_i(Yp_i - Y_i)$.

When $n = 4$, it can be shown that

$$B_4(\hat{y}_d) = \sum_{i=1}^{N} \sum_{j>i}^{N} (p_i + p_j)b_{ij} + \sum_{i=1}^{N} \sum_{j>i}^{N} \sum_{k>j}^{N} \{p_i b_{jk} + p_j b_{ki} + p_k b_{ij}\}$$
$$= 2B_3(\hat{y}_d).$$

For higher values of n also, the bias can be expressed as a linear combination of the b_{ij}. It may be noted that the b_{ij} vanish, independently of the actual values of the units, when one of the following conditions are satisfied:

(1) $$Y_i/p_i = Y, \quad i = 1, \cdots, N;$$

(2) $$p_i = N^{-1}, \quad i = 1, \cdots, N.$$

(e) While it is true that for a given pps design, $\hat{Y}_d$ is better than $\hat{Y}$, in the sense that the mean-square error of $\hat{Y}_d$ is less than or equal to that of $\hat{Y}$ for all values of the population total Y, it is certainly wrong to say that $\hat{y}_d$ is better than either $\hat{Y}_d$ or $\hat{Y}$. In fact, for the same pps design, the mean-square error of $\hat{y}_d$ can be (i) less than that of $\hat{Y}_d$, (ii) between those of $\hat{Y}_d$ and $\hat{Y}$ or (iii) greater than that of $\hat{Y}$, depending upon the values Y_i in the population. This point is illustrated by the following tables.

It is suggested that a study of the mean-square error of $\hat{y}_d$ be made. Some empirical studies carried out by the author indicated that the mean-square error of $\hat{y}_d$ is likely to be less than that of $\hat{Y}$, when $\hat{Y}$ is efficient compared to $\hat{Y}'$.

The author wishes to express his gratitude to Dr. M. N. Murthy of the Indian Statistical Institute for helpful suggestions. Thanks are also due to the referee for his suggestions concerning the presentation.

REFERENCES

[1] Basu, D. (1958). On sampling with and without replacement. *Sankhyā* **20** 287–294.
[2] Dalenius, Tore (1962). Recent advances in sample survey theory and methods. *Ann. Math. Statist.* **33** 325–349.