ON THE ESTIMATE OF THE VARIANCE IN SAMPLING WITH VARYING PROBABILITIES*

BY A. R. SEN

Economics and Statistics Department, Uttar Pradesh, India

Received September 3, 1953

INTRODUCTION

HANSEN AND HURWITZ[1] showed that for a finite population the use of varying probabilities for selecting the sample elements is generally more efficient than selection with equal probability. Their sampling scheme was, however, confined to the selection of a single primary sampling unit (p.s.u.) from a stratum. Working independently Midzuno[4] and the present author[5, 6] generalised the Hansen and Hurwitz scheme to sampling a combination of n elements from a stratum with probability proportional to size (p.p.s.) of the combination. It was proved by the authors that this scheme amounts to selecting the first unit with p.p.s. and the remaining units with equal probability, the selection being made without replacement. The present author[7] further generalised the scheme for obtaining an unbiased estimate of the population total when the first r units are selected with p.p.s. and the remaining $n-r$ units are selected with equal probability and without replacement, and also derived expressions[5,6,7] for estimate of the variance of the estimate.

Recently Horvitz and Thomson[2] presented another technique for dealing with the problem of selecting n p.s.u.'s without replacement and with varying probabilities from a finite population. Formulæ for obtaining unbiased estimate of the population total as well as of the variance of the estimate were presented. As was observed by the present author[8] the scheme suffers from certain disadvantages. One such disadvantage is the difficulty involved in the determination of the selection probabilities. In fact, although the probabilities of selection of the p.s.u.'s can be easily worked out for samples of size two, computations become extremely difficult and almost unwieldy for samples of size greater than two. The scheme is thus practicable for samples of size two only and hence restricted to populations of small size or where a population could be divided into strata of small size.

---

* Presented at the 41st Session of the Indian Science Congress at Hyderabad in January 1954.

Another disadvantage is that the estimate of the variance may assume negative values except for the special case of equal probabilities of selection for the elements remaining prior to each draw. The present author derived further results†[7,8] in this direction. One such result is an unbiased estimate of the variance which is proved to be free from this defect for the scheme where the first p.s.u. is selected with p.p.s. and subsequent units with equal probability and without replacement.

In the present paper expression for the unbiased estimate of the variance referred[8] to in the foregoing para will be derived. It will be shown that this estimate is always positive. A biased estimator for the sampling variance is also given, which is shown to be more efficient than Horvitz and Thomson's unbiased estimate of the variance.

Horvitz and Thomson's Scheme

Consider a population of N elements $U_1, U_2, \ldots, U_N$ with respective measures of size proportional to $X_1, X_2, \ldots, X_N$. For simplicity, we shall assume our population to consist of one stratum only. The results obtained for one stratum can be easily summed up over K strata. For convenience in writing the formulæ replace the actual measures of size $X_i\ (i = 1, 2, \ldots, N)$ by p_i where $\sum_i^N p_i = I$. Let a sample of size n be drawn without replacement such that the first p.s.u. is selected with p.p.s., and the second with p.p.s. to the size of the remaining units and, so on. It is easy to see that the probability of selecting the i and j units together is given by

$$p_i p_j \left(\frac{1}{1-p_i} + \frac{1}{1-p}\right) \tag{1}$$

and the probability that the i-th element is included in a sample of two is given by

$$p_i \left(1 + S - \frac{p_i}{1-p_i}\right) \tag{2}$$

---

† The negative aspect of the unbiased estimate of the variance provided by Horvitz and Thomson has also been observed very recently by Yates and Grundy[9] in a manuscript of the paper received by the present author. Yates and Grundy have also derived independently the same expression for the estimate of the variance. No proofs are, however, given for such an estimate being always positive although as the authors observe 'this appears to be the case when the usual method of selection is empioyed'.

where

$$S = \sum_{i=1}^{N} \frac{p_i}{1 + p_i}$$

Using the notation of Horvitz and Thomson let $P(U_i)$ and $P(U_iU_j)$ denote, in general, the probability that the i-th element (U_i) and i-th and j-th elements (U_i and U_j) be respectively selected in a sample of size n.

Suppose now it is required to estimate the stratum total $Y (Y = \Sigma_i Y_i)$ where Y_i, the value of the i-th unit U_i is correlated with X_i. In particular, X may be the previous census value of the characteristic Y and is known exactly. Then as has been shown by Horvitz and Thomson

$$\sum_{i=1}^{n} \frac{y_i}{P(u_i)} \tag{3}$$

is an unbiased estimate of Y. It is shown[2] that the variance of the estimate is given by

$$\sum_{i=1}^{N} \frac{Y_i^2 (1 - P_{(u_i)})}{P_{(u_i)}} + \sum\sum_{i \neq j}^{N} Y_i Y_j \frac{P_{(u_iu_j)} - P_{(u_i)} \cdot P_{(u_j)}}{P_{(u_i)} \cdot P_{(u_j)}} \tag{4}$$

An unbiased estimate of the variance as presented by the authors is easily seen to be

$$\sum_{i=1}^{n} \frac{Y_i^2 (1 - P_{(u_i)})}{P^2_{(u_i)}} + \sum\sum_{i \neq j}^{n} Y_i Y_j \frac{P_{(u_iu_j)} - P_{(u_i)} \cdot P_{(u_j)}}{P_{(u_iu_j)} \cdot P_{(u_i)} \cdot P_{(u_j)}} \tag{5}$$

Weakness of the Estimate of the Variance

The estimate (5) above has an undesirable property that it may assume negative values for certain combinations of the sampling units except for the special case when $p_i = p_j$, etc., when (5) is positive and reduces[26] to

$$\frac{N(N-n)}{n^2(n-1)} \sum\sum_{i<j} (y_i - y_j)^2 \tag{6}$$

A necessary and sufficient condition for (5) to be always positive is that all the principal minors of the quadratic (5) are positive which

is not always true. Nevertheless, for specific populations (5) may be always positive, *i.e.*, for all pairs (y_i, y_j).

Biased Estimate which is Always Positive

A biased estimate of the variance of the estimate which is always positive is given by

(i) expression (5) when the estimate of variance is positive;
(ii) zero when (5) is negative. } (7)

It follows that the bias is always positive and that (7) has a lower mean square error than (5) except for populations for which (5) is always positive in which case the biased estimate is identically the same as the unbiased estimate (5).

Unbiased Estimate of the Variance which is Always Positive

Expression (4) may be written as

$$\sum_i \sum_{j \neq i} \frac{Y_i^2}{P^2_{(u_i)}} \left[P_{(u_i)} . P_{(u_j)} - P_{(u_i u_j)} \right] - \sum_{i \neq j}^{N} \sum \frac{Y_i Y_j}{P_{(u_i)} . P_{(u_j)}} \left[P_{(u_i)} . P_{(u_j)} - P_{(u_i u_j)} \right]$$

$$= \sum_{i<j}^{N} \sum \left[P_{(u_i)} . P_{(u_j)} - P_{(u_i u_j)} \right] \left[\frac{Y_i}{P_{(u_i)}} - \frac{Y_j}{P_{(u_j)}} \right]^2 \quad (8)$$

An unbiased estimate of (8) is, therefore, given by

$$\sum_{i<j}^{n} \sum \left[\frac{P_{(u_i)} . P_{(u_j)} - P_{(u_i u_j)}}{P_{(u_i u_j)}} \right] \left[\frac{y_i}{P_{(u_i)}} - \frac{y_j}{P_{(u_j)}} \right]^2 \quad (9)$$

We shall now state and prove two theorems regarding the positive nature of (9).

Theorem 1.—For the sampling system[3] in which the first p.s.u. is selected with p.p.s. and the remaining $n - 1$ units are selected with equal probability and without replacement, the unbiased estimate of the variance obtained by substituting the values of $P(u_i)$, $P(u_j)$ and $P(u_i u_j)$ in (9) is always positive.

Proof.—For such a sampling system it can be shown by combinatorial analysis that

$$P_{(u_i)} = \frac{p_i \lfloor n-1}{(N-1)\ldots(N-n+1)}\binom{N-1}{n-1} + \sum_{j \neq i} p_j \frac{n-1}{N-1}$$

$$= \frac{N-n}{N-1} p_i + \frac{n-1}{N-1} \tag{10}$$

and

$$P_{(u_i u_j)} = \frac{n-1}{N-1}\left[\frac{N-n}{N-2}(p_i + p_j) + \frac{n-2}{N-2}\right] \tag{11}$$

Therefore

$$P_{(u_i)} \cdot P_{(u_j)} - P_{(u_i u_j)} = \left(\frac{N-n}{N-1}\right)^2 p_i p_j + \frac{(n-1)(N-n)}{(N-1)^2(N-2)}(1 - p_i - p_j) \tag{12}$$

which is always positive since $p_i + p_j < 1$. Hence, by substitution (9) is always positive.

THEOREM 2.—For a sample of size two where the first p.s.u. is selected with p.p.s. and the second with p.p.s. of the remaining units, the estimate of the variance as given by (9) is always positive.

Proof.

$$\frac{P_{(u_i)} \cdot P_{(u_j)} - P_{(u_i u_j)}}{P_{(u_i u_j)}} = \frac{K^2 + K\left(\frac{1}{1-p_i} + \frac{1}{1-p_j}\right) + \frac{p_i + p_j - 1}{(1-p_i)(1-p_j)}}{\frac{1}{1-p_i} + \frac{1}{1-p_j}} \tag{13}$$

where

$$K = \frac{p_k}{1-p_k} + \ldots + \frac{p_s}{1-p_s} = \sum_{l \neq i,j} \frac{p_l}{1-p_l}$$

Now

$$K^2 + K\left(\frac{1}{1-p_i} + \frac{1}{1-p_j}\right) + \frac{p_i + p_j - 1}{(1-p_i)(1-p_j)}$$

$$\geqslant K_m^2 + K_m\left(\frac{1}{1-p_i} + \frac{1}{1-p_j}\right) + \frac{p_i + p_j - 1}{(1-p_i)(1-p_j)} \tag{14}$$

where K_m is the minimum value of K for a given p_i, p_j. It is easy to see that

$$K_m = \frac{(N-2)\alpha}{N-2-\alpha}$$

where $\alpha = 1 - p_i - p_j$. Substituting for K_m the right-hand side of the inequality (14)

$$= K_m^2 + \frac{(N-1)\alpha^2}{(N-2-\alpha)(1-p_i)(1-p_j)}$$

which is always positive for $N > 2$ since $2 < 2 + \alpha < 3$. Hence (9) is always positive for $n = 2$ and $N > 2$. It appears that the theorem is generally true for samples of size $n\,(n > 2)$ although no formal proof could be derived so far. This, however, is not important since as has already been stated in the introduction, computations of $P_{(ui)}$, $P_{(uiuj)}$ and hence the unbiased estimate of the total as also of the variance become extremely complicated for samples of size exceeding 2.

Practical Application

This section will be devoted to illustrate some of the results discussed in the foregoing sections, by means of a numerical example. In Table I is given a population of five units, the data being presented in columns 2 and 3. The correlation between p and y for the population is 0·8. In columns 4 and 5 are given the values of $P_{(ui)}$ according to formulæ (2) and (10) respectively.

Table I

Unit	p	y	Formulæ (2) P (u_i)	Formulæ (10) P (u_i)
(1)	(2)	(3)	(4)	(5)
1	0·10	3·0	0·2188	0·3250
2	0·15	2·0	0·3184	0·3625
3	0·20	7·0	0·4099	0·4000
4	0·25	5·0	0·4916	0·4375
5	0·30	8·0	0·5613	0·4750
	1·00	25·0	2·0000	2·0000

The calculations of the values of the estimates of the variances for the schemes (*a*) when the first unit is selected with p.p.s. and the second with p.p.s. from the remaining units and (*b*) when the first element is selected with p.p.s. and the second with equal probability from the remaining units are presented in Table II.

Comparisons of Efficiency (Ignoring Cost)

The expected values of the estimates of the variances as well as the variance of the estimates are presented at the bottom of the table. A comparison of the efficiencies of error would show that the

TABLE II

Units	ESTIMATE OF ERROR						
	Scheme (*a*)			Scheme (*b*)			Scheme (*c*)
	Formulæ 5	Formulæ 9	Formulæ 7	Formulæ 5	Formulæ 9	Formulæ 7	Formulæ 6
1	2	3	4	5	6	7	8
(1, 2)	− 3·98	56·95	0	−13·22	12·20	0	3·75
(1, 3)	−102·71	10·22	0	4·34	50·14	4·34	60·00
(1, 4)	− 12·78	9·52	0	− 0·89	3·02	0	15·00
(1, 5)	− 3·11	0·18	0	37·36	31·50	37·36	93·75
(2, 3)	28·92	92·39	28·92	76·25	94·36	76·25	93·75
(2, 4)	− 5·25	10·04	0	18·98	20·47	18·98	33·75
(2, 5)	22·05	33·36	22·05	69·73	68·05	69·73	135·00
(3, 4)	30·30	26·68	30·30	35·00	20·48	35·00	15·00
(3, 5)	51·01	3·44	51·01	26·15	0·22	26·15	3·75
(4, 5)	45·52	5·53	45·52	25·53	14·99	25·53	33·75
Expected error	19·7	19·7	26·2	30·3	30·3	31·2	48·7
Variance of estimate of error	1206	594	438*	640	800	574*	1844
Relative Efficiency (%) (Ignoring cost)	153	310	421	288	230	321	100

* Mean Square (Variance + Bias).

1. Columns (2), (3) and (4) give the estimates of error using formulæ (1) and (2) for values of $P(u_i u_j)$ and $P(u_i)$.

2. Columns (5), (6) and (7) give the estimates of error using formulæ (10) and (11) for values of $P(u_i)$ and $P(u_i u_j)$.

biased estimates of error (columns 4, 7) are more efficient than the corresponding unbiased estimates by Horvitz and Thomson which have the additional weakness of assuming negative values. For the scheme (*a*) when both the units are selected with p.p.s. and without replacement, Horvitz and Thomson's estimate of error is very inefficient compared to the unbiased estimate (column 3) and the biased estimate of error (column 4). For the selection scheme (*b*) where the first unit is selected with p.p.s. and the second with equal probability, the relative position is much better. In fact Horvitz and Thomson's estimate of error is more efficient than the unbiased estimate of error (column 6) but less efficient than the biased estimate (column 7).

The bias is rather large for scheme (*a*) but small for scheme (*b*). The bias is, however, not so serious because it is positive for both the schemes and this is always true. The larger bias for scheme (*a*) is

because the values of $P_{(ui)}$ are more variable for scheme (a) than for scheme (b). In fact the bias is zero in the limiting case when the $P_{(ui)}$'s are all equal which amounts to all p_i's being equal. In this case, of course, both the unbiased estimates of the total and of the variance of the estimate are most inefficient.

SUMMARY

It is shown that Horvitz and Thomson's estimate of the variance of the estimated total has an undesirable property that it may assume negative values. An unbiased estimate has been derived which is proved to be free from this defect for the practical case $n = 2$ when the first unit is selected with p.p.s. and the second with p.p.s. of the remaining units. It is also proved that the unbiased estimate is always positive for the general case when the first unit is selected with p.p.s. and the remaining $n - 1$ units with equal probability and without replacement.

For the practical use Horvitz and Thomson's estimate of the variance can assume only positive values since the variance cannot take a negative value which, if it occurs, has to be substituted by 0. The estimate of variance thus modified is biased, but more efficient than the original estimate. For populations for which Horvitz and Thomson's estimate of the variance is always positive, the biased estimate becomes identically the same as Horvitz and Thomson's unbiased estimate. For the scheme when the first unit is selected with p.p.s. and the second with equal probability of the remaining units Horvitz and Thomson's estimate of the variance is generally more efficient than the corresponding unbiased estimate when the first unit is selected with p.p.s. and the second with p.p.s. of the remaining units. These results have been illustrated with a numerical example.

REFERENCES

1. Hansen, M. H. and Hurwitz, W. N. — "On the theory of sampling from finite populations," *Ann. Math. Stat.*, 1943, **14**, 333–362.

2. Horvitz, D. G. and Thomson, D. J. — "A generalisation of sampling without replacement from a finite universe," *J.A.S.A.*, 1952, **47**, 663–685.

3. Midzuno, H. .. "An outline of the theory of sampling systems," *Annals of the Inst. of Stat. Math.* (Tokyo), 1950, **1**, 149–156.

4. ———— .. "On the Sampling system with probability proportional to sums of sizes," *ibid.*, 1952, **3**, 99–107.

5. Sen, A. R. .. "Present status of probability sampling and its use in the estimation of Farm Characteristics," Joint meeting of Econometric Society with Inst. Math. Stat.,

Minneapolis, Minn., Sept. 1951 (Abstract in the *Jour. Econ. Soc.*, 1952, **20**, 103).

6. Sen, A. R. .. "Further developments of the theory and application of the selection of primary sampling units with special reference to N. C. Agricultural Population," *Thesis*, Library, N. C. State College, 1952.

7. ———— .. "On the selection of n primary sampling units from a stratum structure ($n \geqslant 2$)" (to be published) 1952.

8. ———— .. "Recent Advances in Sampling with varying probabilities," *Cal. Stat. Assoc. Bull.*, 1953, **5**, 1–15.

9. Yates, F. and Grundy, P. M. "Selection without replacement from within strata with probability proportional to size," *J.R.S.S.*, **(B 15)**, 1953, 235–261.