ON THE SELECTION OF n PRIMARY SAMPLING UNITS FROM A STRATUM STRUCTURE ($n \geqq 2$)

A. R. SEN[1]

Economics and Statistics Department, Uttar Pradesh, Lucknow, India

1. Introduction and summary. Hansen and Hurwitz [1] showed for two-stage sampling that the selection of a single primary sampling unit (p.s.u.) with probability proportional to some measure of its size (p.p.s.) from each stratum is generally more efficient than selection with equal probability. Midzuno [5] generalised the Hansen and Hurwitz approach to sampling a combination of n elements from each stratum. Neither Hansen and Hurwitz nor Midzuno provided a method for estimating the between component of the total error from the sample. Recently Horvitz and Thomson [3] have also given a method for dealing with sampling without replacement when arbitrary probabilities of selection are used for elements remaining prior to each draw. Methods for obtaining an unbiased estimate of the population total as well as of the variance of the estimate are presented. The scheme, however, suffers from certain practical disadvantages. One such disadvantage is the difficulty involved in the determination of the selection probabilities. Another disadvantage is that Horvitz and Thomson's unbiased estimate of the variance has generally no practical application as it may assume negative values. This has been shown independently by the present author [9], [10] and Yates and Grundy [11]. The authors also derived an expression of the unbiased estimate which is free from this defect.

Working independently, the present author [7], [8] developed the theory when a combination of n p.s.u.'s are sampled from a stratum and applied it to the case when $n = 2$. In this paper an outline is given of the general theory of the selection and estimation procedure for obtaining unbiased estimates of the between component of total error where first r p.s.u.'s are selected with p.p.s. and the remaining $n - r$ are selected with equal probability, the selection being without replacement. An expression for the estimate of the variance of the estimated total is presented. It is shown that the unbiased estimate of the variance of the estimate is generally inefficient and may assume negative values for certain combinations of the sample values except for the special case when the measures of sizes are all equal. A biased estimate has, however, been derived which is always positive and is more efficient than the unbiased estimate. It is shown that for the particular case when $r = 1$ the unbiased estimate of the total reduces to a simple form which is useful in practice. It is proved that the selection of one p.s.u. with p.p.s. and the remaining $n - 1$ with equal probability is equiva-

Received August 31, 1954.

[1] This paper is based on a dissertation written under the supervision of Professor R. L. Anderson and submitted in 1952 as partial fulfillment of the requirements for the Ph.D. degree in Experimental Statistics at the University of North Carolina.

lent to selecting a combination of n p.s.u.'s, where the measure of size is the sum of the measures of the combination.

2. Probability functions for selection with probability proportionate to size. Consider a population $U_1, \cdots, U_N$ of units with respective measures of sizes proportional to $X_1, \cdots, X_N$. In particular, the X's may be previous census values of the characteristic for the units known exactly. Let $(i_1, \cdots, i_n)$ be any arbitrary combination of size n taken from $1, \cdots, N$. Obviously the total number of possible combinations is $\binom{N}{n}$.

Let $Q(i_1, \cdots, i_r)$ be the probability of selecting the units $U_{i_1}, \cdots, U_{i_r}$ from among the units $U_1, \cdots, U_N$ with p.p.s. to $X_{i_1}, \cdots, X_{i_r}$ the selection being without replacement, each element being selected p.p.s. from those remaining after the preceding selection.

THEOREM 1. *$Q(i_1, \cdots, i_r)$ is given by the recursive formula*

$$Q(i_1, \cdots, i_r) = \sum_{j=1}^{r} \frac{X_{i_j}}{X} Q_{i_j}(i_1, \cdots, i_{j-1}, i_{j+1}, \cdots, i_r) \tag{1}$$

where $X = \sum_{i=1}^{N} X_i$, and Q_{i_j} indicates that U_{i_j} is eliminated as a possible selection.

PROOF. The right hand side of (1) within the summation sign is the probability of selecting the r units $U_{i_1}, \cdots, U_{i_r}$ out of $U_{i_1}, \cdots, U_{i_N}$ such that the i_j^{th} unit U_{i_j} is first selected with p.p.s. to X_{i_j} and the remaining $r-1$ units $U_{i_1}, \cdots, U_{i_{j-1}}, U_{i_{j+1}}, \cdots, U_{i_r}$ are next selected from the remaining $N-1$ units $U_{i_1}, \cdots, U_{i_{j-1}}, U_{i_{j+1}}, \cdots, U_{i_N}$ with p.p.s. to $X_{i_1}, \cdots, X_{i_{j-1}}, X_{i_{j+1}}, \cdots, X_{i_N}$ and without replacement. The sum of all such probabilities where the i_j^{th} unit U_{i_j} may be any one of the units $U_{i_1}, \cdots, U_{i_j}, \cdots, U_{i_r}$ is equal to $Q(i_1, \cdots, i_r)$.

In particular,

$$Q(i_1, i_2) = \frac{X_{i_1} X_{i_2}}{X}\left(\frac{1}{X - X_{i_1}} + \frac{1}{X - X_{i_2}}\right)$$

THEOREM 2. *The probability of selecting a specified n units such that any r units $U_{i_1}, \cdots, U_{i_r}$ are first selected with* p.p.s. *to $X_{i_1}, \cdots, X_{i_r}$ and the remaining $n-r$ units with equal probability from among the remaining $N-r$ units, the selection being without replacement, is given by*

$$P(n, r, [i]) = \frac{1}{(N-r; n-r)} \sum_{(n;r)} Q(i_1, \cdots, i_r) \tag{2}$$

where $(N-r; n-r)$ is written for $\binom{N-r}{n-r}$ and $\sum_{(n;r)}$ denotes summation over all possible combinations $(i_1, \cdots, i_r)$ out of $(i_1, \cdots, i_r, \cdots, i_n)$. For simplicity write $P(n, r)$ for $P(n, r, [i])$. The proof is omitted.

Special Cases.

The following are some important special cases:

$r = 2$, $n = n$.

$$P(n, 2) = \sum_{(n;2)} \left[\frac{X_{i_1} X_{i_2}}{X}\left(\frac{1}{X - X_{i_1}} + \frac{1}{X - X_{i_2}}\right)\frac{1}{(N-2; n-2)}\right] \tag{3}$$

Thus $P(n, 2)$ is the probability of selecting n p.s.u.'s such that the first two units U_{i_1}, U_{i_2} are selected with p.p.s. to X_{i_1}, X_{i_2} and the remaining $n - 2$ units with equal probability, the selection being made without replacement.

$r = 1, n = n.$

$$P(n, 1) = \frac{1}{(N-1; n-1)} \frac{(X_{i_1} + \cdots + X_{i_n})}{X} \tag{4}$$

$$= \frac{X_{i_1}}{X} \cdot \frac{1}{(N-1; n-1)} + \frac{X_{i_2}}{X} \cdot \frac{1}{(N-1; n-1)} + \cdots + \frac{X_{i_n}}{X} \cdot \frac{1}{(N-1; n-1)} \tag{5}$$

From (4) it follows that $P(n, 1)$ is the probability of selecting a combination of n units $U_{i_1}, \cdots, U_{i_n}$ with probability proportional to the total measure of size of the combination. Also from (5) $P(n, 1)$ is the probability of selecting n p.s.u.'s $U_{i_1}, \cdots, U_{i_n}$ such that the first unit is selected with p.p.s. to measure of size and the remaining $n - 1$ units with equal probability but without replacement. Hence the selection of one p.s.u. with p.p.s. and the remaining $n - 1$ units with equal probability and without replacement is equivalent to selecting a combination of n p.s.u.'s with p.p.s. to measure of size of the combination.

3. Unbiased systems. Let T_p be any function of the observations on n elements selected by some probability system from a population consisting of N elements. In particular, let T_p be an estimate of the population value Y with regard to the probability function $P(n, r)$. Then $[P(n, r), T_p]$ will be defined as a sampling system.

A sampling system $[P(n, r), T_p]$ is said to be unbiased [4], [8] if

$$E_{P(n,r)}(T_p) = Y \tag{6}$$

We will now consider the class of unbiased estimates where the value of the auxiliary variate X correlated with that of the characteristic Y is known beforehand from a complete census. In particular, X may be the value of the characteristic Y at a previous census, not subject to any sampling error. For simplicity of notation we will consider only one stratum.

Consider now a population total Y as the total of the population of different units. Also let Y_i be the population total for the i^{th} unit. Consider a two-stage sampling system in which n p.s.u.'s are selected out of N units from the stratum such that the first r units are selected with p.p.s. and the remaining $n - r$ units with equal probability and without replacement. Let the selected n units be next independently subsampled at random without replacement and the unbiased estimated totals based on the subsamples be $y'_{i_1}, \cdots, y'_{i_n}$.

THEOREM 3.

$$\left[P(n, r), \frac{\sum_{j=1}^{n} y'_{i_j}}{(N-1; n-1)P(n, r)}\right] \tag{7}$$

is an unbiased sampling system.

PROOF. By Theorem 14 ch. 3 of [2] on conditional expectation,

$$E\left[\frac{\sum_{j=1}^{n} y'_{i_j}}{(N-1; n-1)P(n, r)}\right] = \frac{1}{(N-1; n-1)} E\left[\frac{1}{P(n, r)} \cdot E_1 \sum_{j=1}^{n} y'_{i_j}\right]$$

where $E_1(\sum_{j=1}^{n} y'_{i_j})$ is the conditional expectation of $\sum_{j=1}^{n} y'_{i_j}$ holding the first stage units constant. By theorem 6 ch. 3 of [2],

$$E_1\left(\sum_{j=1}^{n} y'_{i_j}\right) = \sum_{j=1}^{n} E_1 y'_{i_j} = \sum_{j=1}^{n} Y_{i_j}.$$

Hence

$$E\left[\frac{\sum_{j=1}^{n} y'_{i_j}}{(N-1; n-1)P(n, r)}\right] = \sum_{(N;n)} \left[\frac{1}{(N-1; n-1)} \sum_{j=1}^{n} Y_{i_j}\right] = Y. \tag{8}$$

By Theorem 15, ch. 3 of [2] on conditional variance,

$$\operatorname{Var}\left[\frac{\sum_{j=1}^{n} y'_{i_j}}{(N-1; n-1)P(n, r)}\right] = E\left[E_1\left\{\frac{\sum_{j=1}^{n} (y'_{i_j} - Y_{i_j})}{(N-1; n-1)P(n, r)}\right\}^2\right] + E\left[\left\{\frac{(Y_{i_1} + \cdots + Y_{i_n})}{(N-1; n-1)} - Y\right\}^2\right]$$

where $E_1\{\sum_{j=1}^{n}(y'_{i_j} - Y_{i_j}) / (N-1; n-1)P(n, r)\}^2$ is the conditional variance of $\sum_{j=1}^{n}(y'_{i_j}) / (N-1; n-1)P(n, r)$ holding the first stage units constant. Hence

$$\operatorname{Var}\left[\frac{\sum_{j=1}^{n} y'_{i_j}}{(N-1; n-1)P(n, r)}\right] = \underbrace{\frac{1}{(N-1; n-1)^2} \sum_{i_1} \cdots \sum_{i_n} \frac{(Z_{i_1} + \cdots + Z_{i_n})}{P(n, r)}}_{\text{Within variance}} + \underbrace{\frac{1}{(N-1; n-1)^2} \sum_{i_1} \cdots \sum_{i_n} \frac{(Y_{i_1} + \cdots + Y_{i_n})^2}{P(n, r)} - Y^2}_{\text{Between variance}} \tag{9}$$

where $Z_i = M_i(M_i - m_i)\frac{\sigma_i^2}{m_i}$, m_i is the number of subsampling units sampled from M_i units and σ_i^2 is the within variance of the i^{th} p.s.u. sampled.

Special Cases.

$r = 1, n = n.$

$$\text{System:}\left[P(n, 1), \frac{\sum_{j=1}^{n} y'_{i_j}}{\sum_{j=1}^{n} X_{i_j}} X\right]. \tag{10}$$

$$\begin{aligned}\text{Var}&\left[\frac{\sum_{j=1}^{n} y'_{i_j}}{\sum_{j=1}^{n} X_{i_j}} X\right] \\ &= \frac{1}{(N-1; n-1)} \sum_{i_1 < \cdots <} \sum_{i_n} \frac{(Y_{i_1} + \cdots + Y_{i_n})^2}{(X_{i_1} + \cdots + X_{i_n})} X - Y^2 \\ &\quad + \frac{1}{(N-1; n-1)^2} \sum_{i_1 < \cdots <} \sum_{i_n} \frac{(Z_{i_1} + \cdots + Z_{i_n})}{P(n, 1)}.\end{aligned} \tag{11}$$

$r = 1, n = 2.$

$$\text{System:}\left[\frac{(X_{i_1} + X_{i_2})}{(N-1)X}, \quad \frac{(y'_{i_1} + y'_{i_2})}{(X_{i_1} + X_{i_2})} X\right], \tag{12}$$

and

$$\begin{aligned}\text{Var}&\left[\frac{(y'_{i_1} + y'_{i_2})}{(X_{i_1} + X_{i_2})} X\right] = \sum_{i_1 < i_2} \sum \frac{(Y_{i_1} + Y_{i_2})^2}{(X_{i_1} + X_{i_2})(N-1)} X - Y^2 \\ &\quad + \sum_{i_1 < i_2} \sum \frac{(Z_{i_1} + Z_{i_2})X}{(N-1)(X_{i_1} + X_{i_2})}.\end{aligned} \tag{13}$$

4. Two other cases.

CASE 1. $r = 0, n = n$. In this case the n units are selected with equal probability and without replacement.

$$P(n, 0) = \frac{1}{(N; n)}. \tag{14}$$

The unbiased system and its variance are given by substituting $P(n, 0)$ for $P(n, r)$ in (7) and (9) respectively.

CASE 2. $r = n, n = n$. In this case the n units are selected with p.p.s. and without replacement. Substituting n for r in (2)

$$P(n, n) = Q(i_1, \cdots, i_n) \tag{15}$$

The unbiased system and its variance are given by substituting $P(n, n)$ from (15) for $P(n, r)$ in (7) and (9) respectively. A practical case of interest is when $n = 2$. The unbiased system is

$$\left[P(2, 2), \frac{(y'_{i_1} + y'_{i_2})}{(N-1; n-1)P(2, 2)}\right] \tag{16}$$

where

$$P(2,2) = Q(i_1, i_2) = \frac{X_{i_1} \cdot X_{i_2}}{X}\left[\frac{1}{X - X_{i_1}} + \frac{1}{X - X_{i_2}}\right]$$

The variance of (16) is obtained by substituting the value of $P(2, 2)$ for $P(n, r)$ in (9).

5. Unbiased estimate of the between p.s.u. component of variance.

THEOREM 4. *An unbiased estimate of the between* p.s.u. *component of variance in* (9) *is given by*

$$G'_n = Q'_n - \sum_{j=1}^{n} (a_{i_j} \cdot Z'_{i_j}) \tag{17}$$

where

$$Q'_n = \sum_{j=1}^{n} a_{i_j} y'^2_{i_j} + 2 \sum_{i_j < i_k} \sum c_{i_j \cdot i_k} y'_{i_j} y'_{i_k} \tag{18}$$

$$\begin{aligned} a_{i_j} &= \frac{1}{(N-1; n-1)^2 P^2(n, r)} - \frac{1}{(N-1; n-1) P(n, r)} \\ c_{i_j \cdot i_k} &= \frac{1}{(N-1; n-1)^2 P^2(n, r)} - \frac{(N; 2)}{(n; 2)(N; n) P(n, r)} \end{aligned} \tag{19}$$

and Z'_{i_j} *is an unbiased estimate of* $M_{i_j}(M_{i_j} - m_{i_j})\sigma^2_{i_j} / m_{i_j}$.

PROOF.

$$\begin{aligned} E[G'_n] &= E[Q'_n] - E\left[\sum_{j=1}^{n} (a_{i_j} Z'_{i_j})\right] \\ &= \sum_{i_1 < \cdots < i_n} \sum \left[P(n, r)\left(\sum_{j=1}^{n} a_{i_j} Y^2_{i_j} + 2 \sum_{i_j < i_k} \sum c_{i_j \cdot i_k} Y_{i_j} \cdot Y_{i_k}\right)\right]. \end{aligned} \tag{20}$$

Also if $Q'_n - \sum_{j=1}^{n} (a_{i_j} \cdot Z'_{i_j})$ is an unbiased estimate of the between p.s.u. component of variance in (9)

$$\begin{aligned} E\left[Q'_n - \sum_{j=1}^{n} (a_{i_j} \cdot Z'_{i_j})\right] &= \frac{1}{(N-1; n-1)^2} \sum_{i_1 < \cdots < i_n} \sum \frac{(Y_{i_1} + \cdots + Y_{i_n})^2}{P(n, r)} \\ &\quad - \left[\sum_{i_1 < \cdots < i_n} \sum \frac{(Y^2_{i_1} + \cdots + Y^2_{i_n})}{(N-1; n-1)}\right. \\ &\quad \left. + 2 \sum_{i_1 < \cdots < i_n} \sum \frac{(Y_{i_1} \cdot Y_{i_2} + \cdots Y_{i_k} Y_{i_j} + \cdots)(N; 2)}{(n; 2)(N; n)}\right]. \end{aligned} \tag{21}$$

Hence by comparing coefficients of the terms of the type $Y^2_{i_1}$, $Y_{i_1} \cdot Y_{i_2}$ etc., in (20) and (21) we have a_{i_j}, $c_{i_j \cdot i_k}$ as in (19). For unistage sampling, an unbiased estimate of the variance of the estimate is given by

$$\sum_{j=1}^{n} a_{i_j} Y^2_{i_j} + 2 \sum_{i_j < i_k} \sum c_{i_j \cdot i_k} Y_{i_j} \cdot Y_{i_k}. \tag{22}$$

This follows as a special case of (17) when the within component of variance is zero.

Further, the necessary and sufficient condition that the quadratic form Q_n' is nonnegative is that all principal minors of the quadratic form are nonnegative. Considering only the first and second order principal minors we have as a necessary condition for the existence of an unbiased estimate of the between p.s.u. component of variance

$$\frac{1}{(N-1;n-1)} \geqq P(n, r) \tag{23}$$

and

$$\frac{2(n-1)}{(N+n-2)(N-1;n-1)} \geqq P(n, r). \tag{24}$$

Condition (24) implies (23) but is not generally true. In fact, if $n = 2$, the inequality (24) reduces to $2 / N(N-1) \geqq P(2, r)$ which holds only when all the elements are selected with equal probability but without replacement.

Special Cases.

$r = 1$, $n = n$. An unbiased estimate of the between p.s.u. component of variance is given by (17) where

$$\begin{aligned} a_{i_j} &= \left(\frac{X}{\sum_{j=1}^{n} X_{i_j}}\right)^2 - \left(\frac{X}{\sum_{j=1}^{n} X_{i_j}}\right), \\ c_{i_j \cdot i_k} &= \left(\frac{X}{\sum_{j=1}^{n} X_{i_j}}\right)^2 - \left(\frac{N-1}{n-1}\right)\left(\frac{X}{\sum_{j=1}^{n} X_{i_j}}\right). \end{aligned} \tag{25}$$

$r = 1$, $n = 2$. An unbiased estimate of the between p.s.u. component of variance is given by (17) where

$$\begin{aligned} a_{i_1} = a_{i_2} &= \left(\frac{X}{\sum_{j=1}^{2} X_{i_j}}\right)^2 - \left(\frac{X}{\sum_{j=1}^{2} X_{i_j}}\right), \\ c_{i_1 \cdot i_2} &= \left(\frac{X}{\sum_{j=1}^{2} X_{i_j}}\right)^2 - (N-1)\left(\frac{X}{\sum_{j=1}^{2} X_{i_j}}\right). \end{aligned} \tag{26}$$

$r = 2$, $n = 2$. An unbiased estimate of the between p.s.u. component of variance is given by (17) where

$$\begin{aligned} a_{i_1} = a_{i_2} &= \frac{1}{(N-1)^2 P^2(2, 2)} - \frac{1}{(N-1)P(2, 2)} \\ c_{i_1 \cdot i_2} &= \frac{1}{(N-1)^2 P^2(2, 2)} - \frac{1}{P(2, 2)}. \end{aligned} \tag{27}$$

6. A biased estimate of the between p.s.u. component of variance.

THEOREM 5. *A biased estimate of the between* p.s.u. *component of variance which is more efficient than the unbiased estimate* (17) *is given by*

(28) (i) *Zero where* (17) *is negative*;
(ii) (17) *where* (17) *is positive and where* a_{i_j}*'s and* $c_{i_j \cdot i_k}$*'s are given by* (19).

PROOF.

$$\text{Let Var}\left[\frac{\sum_{j=1}^{n} y'_{i_j}}{(N-1;n-1)P(n,r)}\right] = E\left[Q'_n - \sum_{j=1}^{n}(a_{i_j} \cdot Z'_{i_j})\right] = \beta \text{ where } \beta \geqq 0.$$

Denote the estimate (28) by R'_n. Then $R'_n = G'_n$ for the set of points for which $G'_n \geqq 0$ and $R'_n = 0$ for the set of points for which $G'_n < 0$, i.e., $G'_n \equiv -H'_n$ (say) and

$$E[R'_n - \beta]^2 = \sum_1 [(G'_n - \beta)^2 P(n,r)] + \sum_2 \beta^2 P(n,r)$$

$$E[G'_n - \beta]^2 = \sum_1 [(G'_n - \beta)^2 P(n,r)] + \sum_2 (H'_n + \beta)^2 P(n,r)$$

where $\sum_1$ denotes summation for nonnegative values of G'_n and $\sum_2$ denotes summation for negative values of G'_n. Then $E[R'_n - \beta]^2 < E[G'_n - \beta]^2$ if G'_n assumes negative values with positive probability. Hence (28) is more efficient than (17).

REFERENCES

[1] M. H. HANSEN AND W. N. HURWITZ, "On the theory of sampling from finite populations," *Ann. Math. Stat.*, Vol. 14 (1943), pp. 333–362.

[2] M. H. HANSEN, W. N. HURWITZ AND W. G. MADOW, *Sample Survey Methods*, Vol. II. John Wiley and Sons, New York, 1953.

[3] D. G. HORVITZ AND D. J. THOMPSON, "A generalisation of sampling without replacement from a finite universe," *J. Amer. Stat. Assn.*, Vol. 47 (1952), pp. 663–685.

[4] H. MIDZUNO, "An outline of the theory of sampling systems," *Ann. Inst. Stat. Math.*, Tokyo, Vol. I (1950), pp. 149–156.

[5] H. MIDZUNO, "On the sampling system with probability proportionate to sums of sizes." *Ann. Inst. Stat. Math.*, Tokyo, Vol. III (1952), pp. 99–107.

[6] R. D. NARAIN, "On sampling without replacement with varying probabilities," *J. Ind. Soc. Agr. Stat.*, Vol. 3 (1951), pp. 169–174.

[7] A. R. SEN, "Present status of probability sampling and its use in the estimation of farm characteristics," (abstract) *Econometrica*, Vol. 20 (1952), p. 103.

[8] A. R. SEN, "Further developments of the theory and application of the selection of p.s.u.'s with special reference to N. C. agricultural populations," Thesis, Library N. C. State College, 1952.

[9] A. R. SEN, "Recent advances in sampling with varying probabilities," *Calcutta Stat. Assn. Bull.*, Vol. 5 (1953), pp. 1–15.

[10] A. R. SEN, "On the estimate of the variance in sampling with varying probabilities," *J. Ind. Soc. Agr. Stat.*, Vol. 5 (1953), pp. 119–127.

[11] F. YATES AND P. M. GRUNDY, "Selection without replacement from within strata with probability proportional to size," *J. Roy. Stat. Soc., Ser. B*, Vol. 15 (1953), pp. 235–261.