AN ESTIMATOR IN PPS SAMPLING FOR MULTIPLE CHARACTERISTICS*

By P. K. PATHAK

Indian Statistical Institute

SUMMARY. For pps sampling schemes when the characteristic under study and the selection probabilities are poorly correlated, it is shown how customary estimators of population total can be modified to yield estimators which are more precise than customary ones in the sense of having smaller expected variance under a super population model. The results established in this paper are extensions of similar results proved by Rao (1966). A result conjectured by Rao (1966) concerning Murthy's estimator (1957) in this connection is shown to be true.

1. Introduction and statement of the problem

Consider a population $\pi = (U_1, U_2, ..., U_j, ..., U_N)$ of N elements. Let P_j and Y_j respectively denote the values of some P and Y characteristics of U_j. Unless otherwise stated the subscript j runs from 1 through N. It is assumed that $\Sigma P_j = 1$. Consider now the problem of estimating $Y = \Sigma Y_j$ on the basis of a sampling scheme $\{S, P\} = \{(s, p(s)) : s \epsilon S\}$ defined on π where S denotes the set of samples to be selected from π and $p(s)$ denotes the probability of selection associated with the sample $s (s \epsilon S)$. We restrict our attention to the following class of unbiased estimators of Y.

$$t_\varphi(s) = \Sigma b_{sj}(Y_j/P_j) \qquad \ldots \quad (1)$$

where $E(b_{sj}) = P_j$ and b_{sj} is the coefficient of (Y_j/P_j) in $t_\varphi(s)$ $(s \epsilon S)$.

It is well-known that when Y_j and P_j are highly correlated, estimators like $t_\varphi(s)$ in sampling schemes with probabilities P_j lead to considerable gain in efficiency as compared with the customary estimators like $n^{-1}\Sigma y_i$ suitable for equal probability sampling. However, in surveys, where population parameters of several characteristics are to be estimated on the basis of some sampling scheme with probabilities P_j, it is likely that some characteristics will be highly correlated with P_j, while others may be poorly correlated and sometimes even uncorrelated with P_j. For example Rao (1966) considers an example of chicken population wherein the number of chickens in a farm and the farm size are very poorly correlated. In this case estimates of total number of chickens obtained on the basis of farms selected by probabilities proportional to farm size, are likely to be inefficient. For such situations Rao (1966) has suggested alternative estimators of Y under several sampling schemes. These estimators can be obtained by substituting NY_jP_j for y_j in $t_\varphi(s)$ given in (1). These

* This paper was written by Dr. P. K. Pathak after seeing the paper by Dr. J. N. K. Rao sent for publication in *Sankhyā* and which appears in this number. Dr. Pathak's paper is being published in the same number with the permission of Dr. J. N. K. Rao.—*Editor.*

estimators are likely to have smaller mean square error than the corresponding unbiased estimators $t_\phi(s)$ particularly in small samples. The bias of such estimators remains the same for all sample sizes unless Y_j and P_j are uncorrelated in which case they are unbiased. Under a super population model Rao (1966) has shown that the new estimators have smaller expected variance than the corresponding unbiased estimators of Y for some sampling schemes. For sampling with unequal probabilities without replacement of size two, Rao (1966) leaves open the problem of comparing Murthy's estimator (1957) and the new estimator obtained from it.

In this note we give a general treatment of the problem considered by Rao (1966), and solve the problem concerning Murthy's estimator (1957).

2. A NEW ESTIMATOR

Under the earlier described sampling scheme $(S, P) = \{(s, p(s)) : s \in S\}$ defined on π, consider the following estimator

$$t_\phi^*(s) = N\Sigma\, b_{sj} Y_j \qquad \ldots (2)$$

which is obtained by replacing Y_j by NY_jP_j in $t_\phi(s)$ given in (1).
Since

$$E\{t_\phi(s)\} = Y_j, \qquad \ldots (3)$$

we can easily get $E\{t_\phi^*(s)\}$ by substituting NY_jP_j for Y_j in the right side of (3).

Therefore

$$E\{t_\phi^*(s)\} = N\Sigma Y_jP_j. \qquad \ldots (4)$$

Thus the bias in $t_\phi^*(s)$, $B(t_\phi)$, is given by

$$\begin{aligned} B(t_\phi^*) &= N\Sigma Y_jP_j - \Sigma Y_j \\ &= N\Sigma(Y_j - N^{-1}\Sigma Y_j)(P_j - N^{-1}) \\ &= N^2 \operatorname{cov}(Y_j, P_j). \end{aligned} \qquad \ldots (5)$$

It is evident from (5) that $t_\phi^*(s)$ will be unbiased if Y_j and P_j are uncorrelated. It is also important to note that the bias remains the same for all sample sizes. Especially for small samples when Y_j and P_j are poorly correlated, the bias may be expected to be small relative to the standard error.

Now in order to compare the two estimators, $t_\phi^*(s)$ and $t_\phi(s)$, we consider the following model (Rao, 1966).

$$Y_j = m + e_j \qquad \ldots (6)$$

where e_j are mutually independent and satisfy $\mathfrak{e}(e_j \mid P_j) = 0$ and $\mathfrak{v}(e_j \mid P_j) = a\,(a > 0)$, $\mathfrak{e}$ and $\mathfrak{v}$ respectively denoting the expectation and the variance under the given model.

It is easy to see that under the model (6), we have

$$\begin{aligned}\mathscr{E}\{V(t_\varphi(s))\} &= \mathscr{E}\{E[\Sigma(b_{sj}-Eb_{sj})Y_j/P_j]^2\} \\ &= E\{V[\Sigma(b_{sj}-Eb_{sj})Y_j/P_j]\}+E\{[\mathscr{E}(\Sigma(b_{sj}-Eb_{sj})Y_j/P_j)]^2\} \\ &= a\Sigma V(b_{sj})/P_j^2+m^2V[\Sigma b_{sj}/P_j] \qquad \ldots (7)\end{aligned}$$

where E and V respectively denote the expectation and the variance under the given sampling scheme.

It is also straightforward to verify that

$$\mathscr{E}\{V(t_\varphi^*(s))\} = aN^2\Sigma V(b_{sj})+m^2N^2V[\Sigma b_{sj}] \qquad \ldots (8)$$

The following theorem asserts that under some regularity conditions, satisfied by many estimators in practice, $\mathscr{E}\{V(t_\varphi^*(s))\} \leqslant \mathscr{E}\{V(t_\varphi(s))\}$.

Theorem 1 : *Let the estimator $t_\varphi(s)$ given in (1) satisfy $\Sigma b_{sj} = 1$ for all samples and $\Sigma V(b_{sj})/P_j^2 \geqslant N^2\Sigma V(b_{sj})$. Then under the model (6)*

$$\mathscr{E}\{V(t_\varphi^*(s))\} \leqslant \mathscr{E}\{V(t_\varphi(s))\} \qquad \ldots (9)$$

where $t_\varphi^*(s)$ is the estimator obtained by replacing Y_j by NY_jP_j in $t_\varphi(s)$.

Proof : Evidently

$$\begin{aligned}&\mathscr{E}\{V(t_\varphi(s))\}-\mathscr{E}\{V(t_\varphi^*(s))\} \\ &= a\Sigma V(b_{sj})/P_j^2-aN^2\Sigma V(b_{sj})+m^2V(\Sigma b_{sj}/P_j)-m^2N^2V(\Sigma b_{sj}) \\ &\geqslant m^2V(\Sigma b_{sj}/P_j) \geqslant 0 \qquad \ldots (10)\end{aligned}$$

since by assumption $V(\Sigma b_{sj}) = 0$ and

$$\Sigma V(b_{sj})/P_j^2 \geqslant N^2\Sigma V(b_{sj}).$$

This completes the proof.

Corollary 1 : $\mathscr{E}\{V(t_\varphi^*(s))\} \leqslant \mathscr{E}\{V(t_\varphi(s))\}$ *if* $V(b_{sj}) = c_1P_j-c_2P_j^2$, *where c_1 and c_2 are non-negative scalars.*

Remark : In practice the verification of the inequality $\Sigma V(b_{sj})/P_j^2 \geqslant N^2\Sigma V(b_{sj})$ will be greatly facilitated if we know $V(b_{sj})$. An easy way of computing $V(b_{sj})$ when we know $V(t_\varphi(s))$, is to let $Y_j = P_j$ and $Y_{j'} = 0$ for all $j' \neq j$ in $V(t_\varphi(s))$.

We mention without going into details that

(i) $$V(b_{sj}) = P_j(1-P_j)/n$$

in sampling with unequal probabilities with replacement,

(ii) $$V(b_{sj}) = (1-(n-1)/n(N-1)+k(n-k)/(N-1))P_j(1-P_j)$$

in the Rao-Hartley-Cochran sampling scheme (1962) and

(iii) $$V(b_{sj}) = P_j/n-P_j^2$$

for the Horvitz-Thompson estimator (1952) in pps sampling without replacement of size n. As a consequence of Corollary 1, we have

$$\{\varepsilon V(t_m^*(s))\} \leqslant \varepsilon\{V(t_m(s))\}$$

in all these cases.

In the next section we prove the conjecture by Rao (1966) concerning Murthy's estimator (1957).

3. MURTHY'S ESTIMATOR

In pps sampling without replacement of size two, Murthy's estimator (1957) of the population total is given by

$$t_m(s) = \frac{1}{(2-p_1-p_2)}\{(1-p_2)y_1/p_1+(1-p_1)y_2/p_2\} \qquad \ldots \quad (11)$$

where the symbols have their usual meanings.

The modification of $t_m(s)$ leads to the following estimator

$$t_m^*(s) = \frac{N}{(2-p_1-p_2)}\{(1-p_2)y_1+(1-p_1)y_2\}. \qquad \ldots \quad (12)$$

Rao (1966) has conjectured that $t_m^*(s)$ has smaller expected variance than $t_m(s)$ under the model (6). We proceed now to prove this conjecture.

For Murthy's estimator (1957) the condition $\Sigma b_{sj} = 1$ is satisfied. Further we have

$$V(b_{sj}) = \sum_{j'=1\neq j}^{N} P_j P_{j'}(1-P_j-P_{j'})/(2-P_j-P_{j'}). \qquad \ldots \quad (13)$$

Thus
$$\varepsilon\{V(t_m^*(s))\} \leqslant \varepsilon\{V(t_m(s))\}$$

if
$$\sum_{j}\sum_{j'=1\neq j}^{N} P_j P_{j'}(1-P_j-P_{j'})(2-P_j-P_{j'})^{-1}(P_j^{-2}-N^2) \geqslant 0. \qquad \ldots \quad (14)$$

To prove (14), we make use of the following lemma.

Lemma 1 : *Let* $(U_1, V_1), \ldots, (U_j, V_j), \ldots, (U_N, V_N)$ *be* N *pairs of real numbers.*

Then
$$\Sigma U_j V_j \geqslant N^{-1}(\Sigma U_j)(\Sigma V_j) \qquad \ldots \quad (15)$$

if $U_l - U_k \geqslant 0$ *if and only if (iff)* $V_l - V_k \geqslant 0$ *for all* l, k $(l, k = 1, 2, \ldots, N)$.

The proof of the lemma is omitted.

To prove (14), it suffices to prove that

$$\sum_{j}\sum_{j'=1\neq j}^{N} P_{j'}(1-P_j-P_{j'})(2-P_j-P_{j'})^{-1}(P_j^{-1}-N^2P_j) \geqslant 0. \qquad \ldots (16)$$

Letting

$$U_j = P_j^{-1}-N^2P_j$$

and

$$V_j = \sum_{j'=1\neq j}^{N} P_{j'}(1-P_j-P_{j'})(2-P_j-P_{j'})^{-1},$$

we find that

$$U_j-U_k \geqslant 0 \quad \text{iff } P_k-P_j \geqslant 0$$

and

$$V_j-V_k = \Big[\sum_{j'=1\neq j\neq k}^{N} P_{j'}(P_k-P_j)(2-P_j-P_{j'})^{-1}(2-P_k-P_{j'})^{-1}\Big]$$

$$+(P_k-P_j)(1-P_j-P_k)(2-P_j-P_k)^{-1} \geqslant 0$$

iff

$$P_k-P_j \geqslant 0.$$

Therefore,

$$U_j-U_k \geqslant 0 \quad \text{iff} \quad V_j-V_k \geqslant 0.$$

Hence, from Lemma 1, we have

$$\sum_{j}\sum_{j'=1\neq j}^{N}\{P_{j'}(1-P_j-P_{j'})(2-P_j-P_{j'})^{-1}(P_j^{-1}-N^2P_j)\}$$

$$\geqslant N^{-1}\{\sum_{j}\sum_{j'=1\neq j}P_{j'}(1-P_j-P_{j'})(2-P_j-P_{j'})^{-1}\}\sum_j(P_j^{-1}-N^2P_j) \geqslant 0 \qquad \ldots (17)$$

since $\Sigma P_j^{-1} \geqslant N^2$.

Thus Rao's conjecture that

$$e\{V(t_m^*(s))\} \leqslant e\{V(t_m(s))\}$$

is indeed true.

Incidentally for estimating $V(t_g^*(s))$, it is trivial to verify that if $e(t_g(s))$ is an unbiased estimator of $V(t_g(s))$, then an unbiased estimator of $V(t_g^*(s))$ can be got by replacing Y_j by NY_jP_j in $e(t_g(s))$.

For interesting applications of Theorem 1 in a variety of situations, the reader may refer to the paper by Rao (1966).

ACKNOWLEDGEMENT

The author is greatly indebted to Professor C. R. Rao for suggesting him to work on the problem discussed in this paper.

REFERENCES

HORVITZ, D. G. and THOMPSON, D. J. (1952) : A generalization of sampling without replacement from a finite universe. *J. Amer. Stat. Assoc.*, 47, 663-685.

MURTHY, M. N. (1957) : Ordered and unordered estimators in sampling without replacement. *Sankhyā*, 18, 379-390.

RAO, J. N. K. (1966) : Alternative estimators in pps sampling for multiple characteristics. *Sankhyā*, Series A, 28, 47-60.

RAO, J. N. K., HARTLEY, H. O. and COCHRAN, W. G. (1962) : On a simple procedure of unequal probability sampling without replacement. *J. Roy. Stat. Soc.*, B, 24, 482-491.

Paper received : September, 1965.