# THE RELATIVE EFFICIENCY OF SOME TWO-PHASE SAMPLING SCHEMES

By M. P. Singh

*Indian Statistical Institute, Calcutta*

**Introduction.** It is sometimes convenient and economical to use two-phase sampling where data on the auxiliary variable $x$ are collected on an initial first-phase sample and information on the study variable $y$ is collected only for a subsample of that sample. The following procedures can be used at the second-phase for utilising the data on $x$ collected in the first-phase: (a) stratification, (b) ratio or regression methods of estimation and (c) selection of the second-phase sample with probability proportional to size (pps), size being $x$. Procedures (a) and (b) are given in Cochran [1] and (c) has been considered by Saito [8], Des Raj [3] and Singh and Singh [9] using the pps with replacement scheme.

In the present note we consider the selection of $N'$ units from the $N$ population units in the first-phase by simple random sampling without replacement for collecting data on $x$, and sampling of $n$ of these $N'$ units in the second-phase for getting information on $y$ using the three pps without replacement schemes given by (i) Des Raj [2], (ii) Hartley and Rao [4], and (iii) Rao, Hartley and Cochran [7]. These schemes, as is well-known, have been considered by the above authors for single-phase sampling only. The authors of schemes (ii) and (iii) have given the sampling variances of the estimators considered by them, while Pathak [7] has obtained an upper bound for the variance of the estimator in scheme (i). Here we obtain the variances of the estimators in the two-phase sampling scheme and compare them under the super-population model considered by Cochran [1].

**2. Estimators and Variances.** An unbiased estimator $T$ of the population total $Y = \sum Y_j$ in the present scheme will be of the form $T = (N/N')t$, where $t$ is an unbiased estimator of the first-phase sample total $Y' = \sum_1 Y_j$, $\sum$ and $\sum_1$ denote summations over all the units in the population and the first-phase sample respectively. In particular $T$ for schemes (i), (ii) and (iii) respectively, will be of the form

$$(2.1) \qquad T_d = (N/nN')\sum_2 [y_1 + y_2 \cdots + y_{j-1} + y_j/P_j'(1 - P_1' \cdots P_{j-1}')],$$

$$T_{Hr} = (N/nN')\sum_2 y_j/P_j' \quad \text{and} \quad T_{rHc} = (N/N')\sum_2 Y_j\pi_j'/P_j',$$

where $\sum_2$ denotes summation over second-phase sample, $P_j' = X_j/\sum_1 X_j$ is the probability of selection of $j$th first-phase unit and $\pi_j' = \sum_{\text{group } j} P_j'$.

The sampling variance of the estimator $T$ can be expressed as

$$(2.2) \qquad V(T) = V_1E_2(T) + E_1V_2(T)$$

where $E_1$, $V_1$ are unconditional and $E_2$, $V_2$ are conditional (given the first-phase sample) expectations and variances respectively.

Received 19 February 1966; revised 3 January 1967.

It may be mentioned that the sampling variances obtained by the respective authors for the estimators in case of single-phase sampling represent the conditional variance of $t$ in the present situation. Hence by evaluating $E_1$ of the conditional variances of $(N/N')t$ and substituting in (2.2), we get the variances of the corresponding estimators in (2.1) as

$$(2.3)\quad V(T_d) \leqq V(T_{d'}) - ((n-1)/2n)N/N'[\sum P_j^2 \sum P_j((Y_i/P_j) - Y)^2 + ((N-N')/N')\sum Y_j^2 P_j + \sum P_j^2((Y_j/P_j) - Y)^2]$$

to the order of $O(N^{-1})$ and assuming $\max P_j' = O(N'^{-1})$,

$$(2.4)\quad V(T_{hr}) = V(T_{d'}) - ((n-1)/n)(N/N')\sum P_j^2((Y_j/P_j) - Y)^2$$

to the order of $O(N^{-1})$ and assuming $\max P_j' = O(N'^{-1})$ and

$$(2.5)\quad V(T_{rhc}) = V(T_{d'}) - ((n-1)/n)(N/(N-1)N')\sum P_j((Y_j/P_j) - Y)^2,$$

where

$$(2.6)\quad V(T_{d'}) = (N(N-N')N')S_y^2 + (N/(N-1))((N'-1)/N')(1/n)\sum P_j((Y_i/P_j) - Y)^2$$

is the sampling variance of the unbiased estimator in pps with replacement selection of the second-phase sample (Des Raj [3]). $V_1E_2(T)$ in (2.2) is given by first term in (2.6), and since it does not depend on the scheme used at the second-phase, remains the same for all three schemes considered here. $S_y^2$ has the usual definition given by $(N-1)^{-1}\sum (Y_j - Y/N)^2$ and $P_j = X_j/\sum X_j$.

It is evident that under these approximations all three schemes considered here, as is expected, are more efficient than the pps with replacement scheme considered by Des Raj [3] and others. The expressions (2.5) and (2.6) above are, however, exact.

**3. Efficiency Comparisons.** For making efficiency comparisons among the above estimators, we regard the finite population as being drawn from an infinite super-population in which $y$ and $x$ are correlated. The results obtained do not apply to any single finite population, but to the average of all finite populations that can be drawn from the super-population. This technique has been used by Cochran [1] and others for making efficiency comparisons, since a direct comparison of the variances involves the unknown $y$-values. The super-population model often used for the comparisons between the expected variances of the estimators is

$$(3.1)\quad \begin{aligned} Y_j &= \beta X_j + e_j, \qquad j = 1, 2, \cdots N; \\ \mathcal{E}(e_j \mid X_j) &= 0, \qquad \mathcal{E}(e_j^2 \mid X_j) = aX_j^g; \\ \mathcal{E}(e_j e_{j'} \mid X_j, X_{j'}) &= 0, \qquad a > 0,\ g \geqq 0; \end{aligned}$$

where $\mathcal{E}$ denotes expectation under the infinite super-population. Furthermore, in practice $g$ may be expected to lie between 1 and 2. Denoting $\mathcal{E}V(T)$ by $V'(T)$ the expected variances of the estimators under the model (3.1) are

$$V'(T_d) \leqq V'(T_{d'}) - ((n-1)/2n)(N/N')a[\sum X_j^{g-1}(X\sum P_j^2 + X_j) + O(N^{-2})], \tag{3.2}$$

$$V'(T_{hr}) = V'(T_{d'}) - ((n-1)/n)(N/N')a[\sum X_j^g + O(N^{-2})], \tag{3.3}$$

$$V'(T_{rhc}) = V'(T_{d'}) - ((n-1)/n)(a/N')[\sum X_j^{g-1}(X - X_j) + O(N^{-2})]. \tag{3.4}$$

It may be mentioned that the pps with replacement estimator, which is less efficient than the pps without replacement estimators considered here, has been shown to be more efficient than the ratio estimator in two-phase sampling whenever $g > 1$, by Des Raj [3]. Now comparing the three estimators among themselves under the model (3.1) we get the following results.

(i)

$$V'(T_{hr}) - V'(T_d) = ((n-1)/2n)(a/N'\bar{X})\sum_{j\neq j'} X_j^{g-1}X_{j'}(X_{j'} - X_j) > 0 \tag{3.5}$$

if all $X_j$ is are not equal, i.e., at least one $X_j \neq \bar{X}$ and $g < 2$.

(ii)

$$V'(T_{rhc}) - V'(T_d) = ((n-1)/n)(a/N')[(N+2)\sum X_j^g - X\sum X_j^{g-1}] > 0 \tag{3.6}$$

when $\max P_j = O(N^{-1})$ and $\sum P_j^2 > N^{-1}$ if some $P_j \neq N^{-1}$ and $g > 1$.

(iii)

$$V'(T_{rhc}) - V'(T_{hr}) = ((n-1)/n)(a/N') \cdot [N\sum X_j^g - X\sum X_j^{g-1} + \sum X_j^g] > 0 \tag{3.7}$$

for $g > 1$. Hence we can summarise the results as follows: Under the model (3.1), if (i) $N$ and $N'$ are sufficiently large (ii) $\max P_j' = O(N'^{-1})$ and $\max P_j = O(N^{-1})$ with at least one $P_j \neq N^{-1}$ and (iii) $g$ lies between 1 and 2, then

$$V'(T_d) < V'(T_{hr}) < V'(T_{rhc}) < V'(T_{d'}) < V'(T_r) \tag{3.8}$$

where $T_r$ is the ratio estimator in two-phase sampling.

REMARK 1. It may be mentioned that in the Hartley and Rao scheme, when used in single-phase sampling, it is necessary to arrange the population units in random order before selection of the sample and in the Rao-Hartley-Cochran scheme the population units are required to be grouped in $n$ random groups before selection of the sample of $n$ units. This randomisation in both cases is a

difficult task especially when the number of units in the population is large. However the application of these schemes at the second-phase in two-phase sampling becomes quite simple, for the first-phase sample is a random sample and the units within it are automatically arranged in random order. Hence application of these schemes may be even simpler than the pps with replacement scheme in some cases.

REMARK 2. Next by considering Murthy's estimator [5] (the unordered Des Raj pps without replacement estimator) in two-phase sampling it will be observed that this estimator is more efficient even than the Des Raj estimator but the gain in efficiency achieved should be very small for large values of $N'$ and $N$.

The author wishes to express his thanks to Dr. M. N. Murthy and to the referee for their valuable suggestions.

## REFERENCES

[1] COCHRAN, W. G. (1963). Sampling techniques. *Wiley, New York.*

[2] DES RAJ (1956). Some estimators in sampling with varying probabilities without replacement. *J. Amer. Statist. Assoc.* **51** 269–284.

[3] DES RAJ (1964). On double sampling for PPS estimation. *Ann. Math. Statist.* **35** 900–902.

[4] HARTLEY, H. O. and RAO, J. N. K. (1962). Sampling with unequal probabilities without replacement. *Ann. Math. Statist.* **33** 350–374.

[5] MURTHY, M. N. (1957). Ordered and unordered estimators in sampling without replacement. *Sankhya*, **18** 379–390.

[6] PATHAK, P. K. (1965). Recent advances in sampling theory; mimeographed lecture note. *RTS. Indian Statist. Inst.*

[7] 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 482–491.

[8] SAITO, K. (1958–59). Theoretical consideration on the use of supplementary information in sample survey design III. *Sophia. Econ. Review* **86**.

[9] SINGH, D. and SINGH, B. D. (1965). Some contribution to two-phase sampling. *Austr. J. Statist.* **7** 2 45–47.