ON EFFICIENCY OF THE SAMPLING WITH VARYING PROBABILITIES WITHOUT REPLACEMENT

By Daroga Singh

Indian Council of Agricultual Research, New Delhi

1. Introduction

It is generally felt that in a sub-sampling design the selection of primary units with varying probabilities without replacement leads to a more efficient estimate than sampling with varying probabilities with replacement. In this connection a paper by Narain (1951) deserves special mention. He compared the two systems of sampling in a two-stage sampling design where the primary units were selected with varying probabilities and the secondary units with equal probabilities without replacement in both cases. He showed the necessary condition, *viz.*, $\frac{n}{n-1}\lambda_{ij} < 2\lambda_i\lambda_j$ for the former being more efficient than the latter, where λ_i is the probability of ith primary unit being included in a sample of size n in a system of sampling with varying probabilities without replacement and λ_{ij} is the probability of both ith and jth units being included in a sample of size n. Durbin (1953) has also stated (without proof) in a recent paper that it is not difficult to find out some situation in which sampling without replacement will be less efficient as compared to that with replacement. This paper will be a study in that direction.

2. Evaluation of Probabilities

Before actually comparing the two systems we will evaluate probabilities and consider its behaviour in a sampling design without replacement. Let there be N primary units and p_i $(i = 1, 2, \ldots, N)$ be the initial probability that the ith unit will be drawn in the first draw. If $P_{ij\ldots mn}$ denote the probability that ith, jth,... and nth units will be included in a sample of size n,

$$P_{ij\ldots\ldots wu} = \sum \frac{p_i p_j \ldots p_m p_n}{(1-p_i)(1-p_i-p_j), \ldots, (1-p_i-p_j-\ldots p_m)} \tag{1}$$

(Summation over all the n ! terms which will be formed by exchanging the position of $i, j, \ldots, m$, and n)

For example, for $n = 2$, the probability that ith and jth units will be included in the sample is given by

$$P_{ij} = P_{ji} = p_i p_j \left(\frac{1}{1 - p_i} + \frac{1}{1 - p_j}\right) \tag{2}$$

Similarly for $n = 3$, the probability that ith, jth, and kth units will be included in the sample is given by

$$P_{ijk} = P_{ikj} = \dots = P_{kij}$$

$$= \sum \frac{p_i p_j p_k}{(1 - p_i)\ (1 - p_i - p_j)}$$

(Summation over all the 3 ! terms)

$$= p_i p_j p_k \left\{\frac{1}{1 - p_i - p_j}\left(\frac{1}{1 - p_i} + \frac{1}{1 - p_j}\right) + \frac{1}{1 - p_i - p_k}\right.$$

$$\left.\times \left(\frac{1}{1 - p_i} + \frac{1}{1 - p_k}\right) + \frac{1}{1 - p_j - p_k}\left(\frac{1}{1 - p_j} + \frac{1}{1 - p_k}\right)\right\} \tag{3}$$

Thus the exact expression for $P_{ij\dots a}$ may be written for any value of n although the work involved will be quite laborious for a large value of n as there will be n ! terms in all. Further it may be seen that

$$\sum_{i=1}^{N} \sum_{j=i+1} \dots \sum_{m=l+1} \sum_{n=m+1} P_{ij\dots mn} = 1 \tag{4}$$

Now the probability of any unit being included in a sample of n may easily be calculated from the above expression (1). For example, the probability that ith unit will be selected in a sample of n is given by

$$P_i(n) = \sum_{j}^{N} \sum_{k} \dots\dots\dots \sum_{n} P_{ijk\dots\dots n} \tag{5}$$
$$i \neq j \neq k \neq \dots\dots \neq n$$

Thus

$$P_i(2) = p_i + p_i \sum_{\substack{j=1 \\ i \neq j}}^{N} p_j/1 - p_j;$$

$$P_i(3) = P_i(2) + p_i \sum_{j}^{N} \sum_{l} \left[\frac{p_j p_l}{1 - p_j - p_l}\left(\frac{1}{1 - p_l} + \frac{1}{1 - p_j}\right)\right];$$
$$i \neq j \neq l$$

$$P_i(4) = P_i(3) + \sum_{\substack{j \\ i \neq j \neq k \neq l}}^{N} \sum_{k} \sum_{l} \Big[\frac{1}{1 - p_j - p_k} \Big(\frac{1}{1 - p_j} + \frac{1}{1 - p_k} \Big)$$

$$+ \frac{1}{1 - p_j - p_l} \Big(\frac{1}{1 - p_j} + \frac{1}{1 - p_l} \Big)$$

$$+ \frac{1}{1 - p_k - p_l} \Big(\frac{1}{1 - p_k} + \frac{1}{1 - p_l} \Big) \Big]$$

$$\times \frac{p_j p_k p_l}{1 - p_j - p_k - p_l} .$$

Similarly the probability of ith and jth units being included in the sample is given by

$$P_{ij}(n) = \sum_{k}^{N} \sum_{l} \ldots \sum_{m} \sum_{n} P_{ijk \cdots mn}$$

$$i \neq j \neq k \neq \ldots \neq m \neq n$$

Thus

$$P_{ij}(2) = p_i p_j \Big(\frac{1}{1 - p_i} + \frac{1}{1 - p_j} \Big) \tag{6}$$

$$P_{ij}(3) = P_{ij}(2) + p_i p_j \sum_{\substack{k \\ i \neq j \neq k}}^{N} \Big[\frac{1}{1 - p_i - p_k} \Big(\frac{1}{1 - p_i} - \frac{1}{1 - p_k} \Big)$$

$$+ \frac{1}{1 - p_k - p_j} \Big(\frac{1}{1 - p_j} - \frac{1}{1 - p_k} \Big) \Big] p_k ;$$

$$P_{ij}(4) = P_{ij}(3) + \sum_{\substack{k \\ i \neq j \neq k \neq l}}^{N} \sum_{l} \Big[\frac{1}{(1 - p_i)(1 - p_i - p_l - p_k)}$$

$$\times \Big(\frac{1}{1 - p_i - p_k} + \frac{1}{1 - p_i - p_l} \Big) + \frac{1}{(1 - p_j)(1 - p_j - p_k - p_l)}$$

$$\times \Big(\frac{1}{1 - p_j - p_k} + \frac{1}{1 - p_j - p_l} \Big) + \frac{1}{1 - p_i - p_k - p_l}$$

$$\times \Big\{ \frac{1}{(1 - p_l)(1 - p_i - p_l)} + \frac{1}{(1 - p_k)(1 - p_i - p_k)} \Big\}$$

$$+ \frac{1}{1-p_j-p_k-p_l} \left\{ \frac{1}{(1-p_l)\,(1-p_l-p_j)} + \frac{1}{(1-p_k)\,(1-p_k-p_j)} \right\} + \frac{1}{1-p_l-p_k} \left(\frac{1}{1-p_l} + \frac{1}{1-p_k} \right)$$

$$\times \left(\frac{1}{1-p_l-p_k-p_i} + \frac{1}{1-p_l-p_k-p_j} \right) \Bigg] \, p_l \, p_k .$$

Now we will prove the following:—

$$P_i(n)\, P_j(n) > P_{ij}(n) \text{ for all values of } N \text{ and}$$

$$p_i < 1 \ (i = 1, 2, \ldots, N), \text{ if } n = 2.$$

For simplicity of notation we will show that

$$P_1(2)\ P_2(2) > P_{12}(2),$$

and this result will follow for all i and j $(i, j = 1, 2, \ldots, N)$.

$$P_1(2) = \sum_{i=2}^{N} P_{1i}$$

and

$$P_2(2) = \sum_{\substack{i=1 \\ i \neq 2}}^{N} P_{2i}$$

$$\therefore \quad P_1(2)\, P_2(2) = (P_{12} + \sum_{i=3}^{N} P_{1i})\ (P_{12} + \sum_{i=3}^{N} P_{2i})$$

$$= P_{12}\ (P_{12} + \sum_{i=3}^{N} P_{1i} + \sum_{i=3}^{N} P_{2i})$$

$$+ \ (\sum_{i=3}^{N} P_{1i})\ (\sum_{i=3}^{N} P_{2i})$$

$$= P_{12}\ [1 - \sum_{i=3}^{N} \sum_{j=i+1} P_{ij}] + (\sum_{i=3}^{N} P_{1i})\ (\sum_{j=3}^{N} P_{2i})$$

$$\therefore \quad P_1(2)\ P_2(2) - P_{12}(2) = (\sum_{i=3}^{N} P_{1i})\ (\sum_{i=3}^{N} P_{2i}) - P_{12} \sum_{i=3}^{N} \sum_{j=i+1} P_{ij}$$

[since $P_{12}(2) = P_{12}$]

Now using the relation (6),

$$P_{ij} = p_i\, p_j \left(\frac{1}{1-p_i} + \frac{1}{1-p_j}\right)$$

we find that

$$P_1(2)\ P_2\ (2) > P_{12}\,(2) \tag{7}$$

The inequality $P_1\,(n)\,P_2\,(n) > P_{12}\,(n)$ will generally hold good even for $n > 2$ if the p_i's are not very heterogeneous. But if some of the values of p_i's are so large as to dominate the entire population the above inequality may not hold good. In such cases it will generally

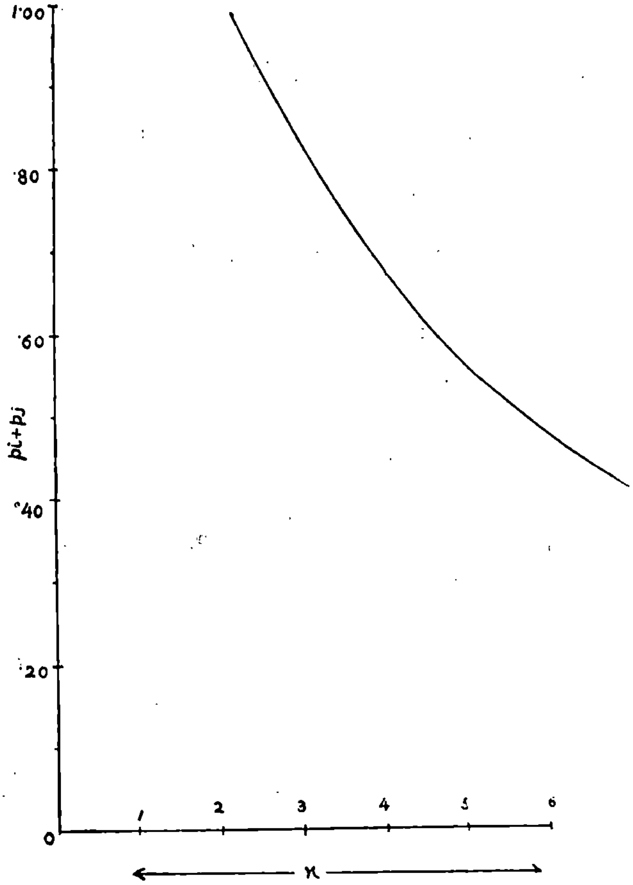

DIAGRAM 1.

$P_i(n)\,P_j(n) - P_{ij}(n) = 0$

(i) $P_i(n)\,P_j(n) > P_{ij}(n)$ below the curve

(ii) $P_i(n)\,P_j(n) < P_{ij}(n)$ above the curve

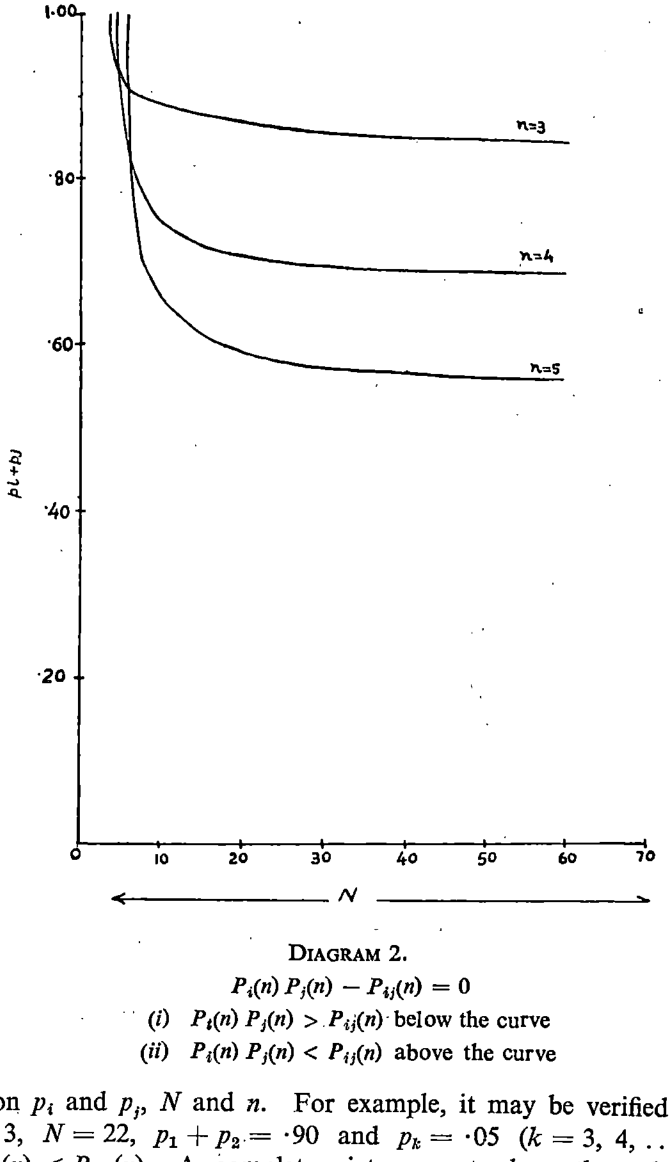

DIAGRAM 2.

$P_i(n)\,P_j(n) - P_{ij}(n) = 0$

(*i*) $P_i(n)\,P_j(n) > P_{ij}(n)$ below the curve

(*ii*) $P_i(n)\,P_j(n) < P_{ij}(n)$ above the curve

depend on p_i and p_j, N and n. For example, it may be verified that for $n = 3$, $N = 22$, $p_1 + p_2 = \cdot 90$ and $p_k = \cdot 05$ $(k = 3, 4, \ldots, N)$ $P_1(n)\,P_2(n) < P_{12}(n)$. A complete picture as to how the value of $P_1(n)\,.\,P_2(n) - P_{12}(n)$ will depend on p_i and p_j, N and n is shown in the diagrams 1 and 2. It may be seen from diagram 1 that the entire space $P_1(n)\;P_2(n) - P_{12}(n)$ may be divided into two parts, in one $P_1(n)\;P_2(n)$ is always greater than $P_{12}(n)$ whereas in other it is otherwise. The diagram is indicative of the minimum value $\phi(n)$

of $P_1(n)\,P_2(n) - P_{12}(n)$ for given $p_1 + p_2$. For the actual values of $P_1(n)\,P_2(n) - P_{12}(n)$ the dividing curve may move slightly up. The minimum value $\phi(n)$ is calculated by assuming

$$p_k = \frac{1 - p_1 - p_2}{N - 2} \quad (k = 3, 4, \ldots, N).$$

EFFICIENCY

3. We shall consider a sub-sampling system where a given probability for being included in the sample is assigned to each primary unit in a stratum. The sub-sampling of secondary units within a primary unit shall be always considered with equal probabilities without replacement. For simplicity of notation we shall consider sampling within a single stratum, the results being easily capable of generalization for any number of strata. Let

N = number of p.s.u.'s in the population;

M_i = number of secondary units in the ith p.s.u.;

$M = \sum_{i=1}^{N} M_i$,

Y_{ij} = the value of jth secondary unit in the ith p.s.u. ;

$\bar{Y}_i = \frac{1}{M_i} \sum_{j=1}^{Mi} Y_{ij}$ = population mean per secondary unit in the ith p.s.u. ;

$\bar{Y} = \sum_{i=1}^{N} \sum_{j=1}^{Mi} Y_{ij}/M$ = population mean to be estimated;

$T_i = \sum_{j=1}^{Mi} Y_{jj} = M_i \bar{Y}_i$ = total for the ith p.s.u.;

$T = \sum_{i=1}^{N} T_i$ = total for the population;

n = number of p.s.u. to be selected in the sample (in case of sampling with replacement all n, p.s.u. need not be distinct);

m = number of secondary units to be sampled from a selected p.s.u.;

$$\sigma_i^2 = \frac{\sum\limits_{i=1}^{M_i}(Y_{ij}-\bar{Y}_i)^2}{M_i-1} = \text{variance within the } i\text{th p.s.u.};$$

and p_i has got the same meaning as given in Section 2.

In case of sampling with replacement it is not very difficult to see that (Cochran, 1953)

$$\bar{y} = \frac{1}{nM}\sum_{i=1}^{n}\frac{M_i}{p_i}\left(\sum_{j=1}^{m}\frac{y_{ij}}{m}\right)$$

$$= \frac{1}{nM}\sum_{i=1}^{n}\frac{M_i}{p_i}\bar{y}_i$$

is an unbiassed estimate of $\bar{Y}$, and

$$V_1 = V(\bar{y}) = \frac{1}{M^2n}\left[\sum_{i=1}^{N}\frac{M_i}{p_i}\,\frac{M_i-m}{m}\,\sigma_i^2 + \sum_{i=1}^{N}\frac{T_i^2}{p_i} - T^2\right] \quad (8)$$

In case of sampling without replacement it is easy to see that (Horvitz and Thompson, 1952)

$$\bar{y}' = \frac{1}{M}\sum_{i=1}^{n}\frac{M_i}{P_i(n)}\left(\sum_{j=1}^{m}\frac{y_{ij}}{m}\right)$$

$$= \frac{1}{M}\sum_{i=1}^{n}\frac{M_i}{P_i(n)}\bar{y}_i$$

is an unbiassed estimate of $\bar{Y}$.

and

$$V_2 = V(\bar{y}') = \frac{1}{M^2}\left[\sum_{i=1}^{N}\frac{M_i(M_i-m)}{P_i(n)\cdot m}\sigma_i^2 + \sum_{i=1}^{N}\frac{T_i^2}{P_i(n)}\right.$$

$$\left. + \sum_{\substack{i,j\\ i\neq j}}^{N}\frac{P_{ij}(n)\cdot T_i\cdot T_j}{P_i(n)\;P_j(n)} - T^2\right] \quad (9)$$

Now

$$V_1 - V_2 = \frac{1}{M^2}\left[\left\{\sum_{i=1}^{N}\left(\frac{M_i(M_i - m_i)}{m}\sigma_i^2 + T_i^2\right)\right.\right.$$

$$\left.\left.\times\left(\frac{1}{np_i} - \frac{1}{p_i(n)}\right)\right\} + \left\{T^2 - \frac{T^2}{n} - \sum_{\substack{i,j\\ i\neq j}}^{N}\frac{P_{ij}(n)\,T_i\,T_j}{P_i(n)\,p_j(n)}\right\}\right] \tag{10}$$

The first term in the above expression

$$\sum_{i=1}^{N}\left\{\frac{M_i(M_i - m)}{m}\sigma_i^2 + T_i^2\right\}\left\{\frac{1}{np_i} - \frac{1}{P_i(n)}\right\}$$

may be negligible in view of the fact that

$$\sum_{i=1}^{N} np_i = \sum_{i=1}^{N} P_i(n)$$

and $P_i(n)$ approximates to np_i. Yates (1953) has stated that bias is negligible if np_i is taken for $P_i(n)$ and also $P_i(n) \gtrless np_i$ according as $p_i \lessgtr 1/N$. Consider the second term. In order that the second term may always be non-negative irrespective of the character under study it is necessary and sufficient that all the principal minors of the discriminant of the quadratic

$$T^2 - \frac{T^2}{n} - \sum_{\substack{i,j\\ i\neq j}}\frac{P_{ij}(n)}{P_i(n)\;P_j(n)}\,T_i\,T_j \tag{11}$$

should be non-negative.

Considering the second order minors we have the necessary condition

$$\frac{n}{n-1}\,P_{ij}(n) \leqslant 2\,P_i(n)\,P_j(n) \tag{12}$$

The inequality (12) will always be satisfied for all values of n if

$$P_{ij}(n) \leqslant P_i(n)\,P_j(n)$$

But we have seen in the previous section that this inequality is always satisfied for $n = 2$; for $n > 2$, cases may arise when $P_i(n)\,P_j(n) < P_{ij}(n)$ which has already been observed in Section 2. In such cases the discriminant (11) may even be negative and sampling without replacement may lead to a less efficient estimate than the one obtained if the sampling with replacement is adopted. One thing appears very clear from the diagrams that as long as $np_i \leqslant 1$ $(i = 1, 2, \ldots, N)$

$$P_{ij}(n) \leqslant P_i(n) P_j(n).$$

Summary

In a sub-sampling design the selection of primary units with varying probabilities without replacement often leads to a more efficient estimate than the one where the sampling of primary units is carried out with varying probabilities with replacement if n is 2 (the number of primary units in the sample). But if $n > 2$, the former system of sampling need not always lead to a more efficient estimate than the one under the latter system and in such cases the efficiency will generally depend, besides n, on the values of p_i (probability of the ith p.s.u. being selected) and N.

At the end I gratefully thank Dr. G. R. Seth who helped me in preparing this paper.

References

1. Narain, R. D. .. "On sampling without replacement with varying probabilities," *J. Ind. Soc. Agri. Stat.*, 1951, **3**, 169.

2. Durbin, J. .. "Some results in sampling theory when units are selected with unequal probabilities," *J.R.S.S.*, 1953, **15** (2).

3. Horvitz, G. D. and Thompson, D. J. "A generalisation of sampling without replacement from a finite universe," *J. Amer. Stat. Ass.*, 1952, **47**, 663.

4. Yates, F. and Grundy, P. M. "Selection without replacement from within strata with probability proportional to size," *J.R.S.S.*, 1953, **15** (2), 253.

5. Cochran, W. G. .. *Sampling Techniques*, John Wiley and Sons, Inc. New York, 1953.