A NOTE ON THE DETERMINATION OF OPTIMUM PROBABILITIES IN SAMPLING WITHOUT REPLACEMENT

By DES RAJ
Indian Statistical Institute, Calcutta

1 Introduction

It is now well-known that the use of varying probabilities in selecting a sample may bring about considerable reduction in the sampling variance of the estimate as compared to the case when the units are selected with equal probabilities. This technique was first suggested by Hansen and Hurwitz (1943) who considered a design in which one first stage unit is selected with probability proportional to size within each stratum. Horvitz and Thompson (1952) generalised it to the selection of n units without replacement within strata. Their estimator of the population (or stratum) total is

$$\hat{y} = \sum_{1}^{n} \frac{y_i}{\pi_i} \qquad \text{... (1.1)}$$

with variance
$$V(\hat{y}) = \sum_{1}^{N} \frac{y_i^2}{\pi_i} + 2\sum{}' \pi_{ij} \frac{y_i}{\pi_i}\frac{y_j}{\pi_j} - Y^2 \qquad \text{... (1.2)}$$

where π_{ij} = probability with which two units u_i and u_j enter the sample,

π_i = probability that the unit u_i is selected in the sample,

y_i = value of the unit u_i for the character y,

and Σ' denotes summation over the ${}^{N}C_2$ pairs of units.

An important problem then arises: How should the sample be selected so that the variance of the estimate is made smallest? It is easy to see that the estimator has zero variance if $\pi_i \propto y_i$. This result is not of practical interest because if the y_i were known in advance, the sample would be unnecessary. The result, however, suggests that if the values of the units for a known auxiliary character x are reasonably proportional to y, it would be advantageous to select the sample so that $\pi_i \propto x_i$.

Now there may be several methods of drawing such a sample. Each such method will lead to certain π_{ij} whose values are going to affect the variance of the estimate. It is then of interest to choose a sampling scheme from all such schemes such that the π_{ij} associated with this scheme should minimise $V(\hat{y})$ given by (1.2) The object of this note is to obtain such an optimum scheme.*

2. Solution of the problem

Considering the practically useful case of $n = 2$ (and the method is unlikely to be convenient for larger sample sizes), the problem consists in the determination of ${}^{N}C_2$ probabilities of selection of pairs, π_{ij}, such that

$$\pi_{ij} \geqslant 0, \qquad \text{... (2.1)}$$

$$\sum_{j \neq i} \pi_{ij} = \pi_i \; (i = 1, 2, \ldots, N) \qquad \text{... (2.2)}$$

and
$$\sum{}' \pi_{ij} \frac{y_i}{\pi_i}\frac{y_j}{\pi_j} \qquad \text{... (2.3)}$$

is minimised.

Stated thus, this is a familiar problem in linear programming. But the difficulty involved is that the coefficients of π_{ij} in (2.3) are unknown. We shall make the assumption that

$$y = \alpha+\beta x \qquad \dots \quad (2.4)$$

i.e. the relation between y and x is a straight line (and the method is unlikely to be useful if the relation is not linear). We shall, however, not assume any knowledge of the actual values of α and β. The quantity to be minimised then is found to be

$$\sum{}' \frac{\pi_{ij}}{\pi_i\pi_j} . \qquad \dots \quad (2.5)$$

This result follows from the observation that

$$\sum{}' \pi_{ij} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} = \tfrac{1}{2} \sum_{i=1}^{N} \sum_{j\neq i} \frac{\pi_{ij}}{\pi_i\pi_j} (\alpha^2+\alpha\beta x_i+\alpha\beta x_j+\beta^2 x_i x_j)$$

$$= \frac{\alpha^2}{2} \sum\sum \frac{\pi_{ij}}{\pi_i\pi_j} + \frac{\left(\sum\limits_1^N x_i\right)}{4} \alpha\beta \left(\sum\sum \frac{\pi_{ij}}{\pi_j} + \sum\sum \frac{\pi_{ij}}{\pi_i}\right)$$

$$+ \frac{\left(\sum\limits_1^N x_i\right)^2}{8} \beta^2 \, \Sigma\Sigma\pi_{ij}$$

$$= \frac{\alpha^2}{2} \sum\sum \frac{\pi_{ij}}{\pi_i\pi_j} + \frac{N}{2} \alpha\beta \left(\sum_1^N x_i\right) + \frac{\left(\sum\limits_1^N x_i\right)^2}{4} \beta^2$$

since
$$\pi_i = \frac{2x_i}{\sum\limits_1^N x_i}, \; \pi_j = \frac{2x_j}{\sum\limits_1^N x_j}, \; \sum_1^N \pi_i = \sum_1^N \pi_j = 2,$$

$$\sum_{j\neq i} \pi_{ij} = \pi_i, \; \sum_{i\neq j} \pi_{ij} = \pi_j.$$

Then the problem reduces to the determination of π_{ij} such that

$$\left.\begin{aligned} &\pi_{ij} \geqslant 0, \\ &\sum_{j\neq i} \pi_{ij} = \pi_i \; (i = 1, 2, \dots, N) \\ \text{and} \quad &\sum{}' \pi_{ij}/(\pi_i\pi_j) \text{ is minimised.} \end{aligned}\right\} \qquad .. \quad (2.6)$$

As stated before, this is a problem in linear programming and can be solved by the simplex method given in Charnes and others (1953).

3. Selection of the sample

With regard to the actual procedure of drawing the sample ensuring $\pi_i \propto x_i$, various methods are now available. Horvitz and Thompson (1952) have suggested two methods which are of limited applicability. Narain's (1951) method requires the solution of an equation by graphical numerical methods. Yates and Grundy (1953) obtain revised size-measures (by an iterative process) and obtain the sample by selecting the first unit with probabilities proportional to the revised sizes and the second unit with probabilities proportional to the remaining sizes. Goodman and Kish (1950) have devised a very convenient method of selection by cumulating the sizes and drawing a systematic sample from the cumulated sizes, the π_{ij} depending on the order in which the units are listed. One characteristic common to all these methods is that they seek to arrive at some π_{ij} which may by no means be optimum. The method of drawing the sample, considered in this note, is however very simple. Out of the totality of ${}^{N}C_2$ groups (of two units each), one has to select one group with given probabilities assigned in an optimum way.

4. An illustration

As an illustration of the practical utility of the method we consider the three populations A, B and C given by Yates and Grundy (1953). These populations were deliberately chosen by them as being more extreme than will normally be encountered in practice. The object is to estimate the population total by selecting two units with probabilities of inclusion π_i proportional to the following p_i.

unit	p
1	0.1
2	0.2
3	0.3
4	0.4

We have to find π_{ij} such that

$$\pi_{ij} \geqslant 0,$$

$$\pi_{12}+\pi_{13}+\pi_{14} = 0.2, \quad \pi_{21}+\pi_{23}+\pi_{24} = 0.4,$$

$$\pi_{13}+\pi_{32}+\pi_{34} = 0.6, \quad \pi_{41}+\pi_{42}+\pi_{43} = 0.8,$$

and $$G = 12.5\pi_{12}+8.3333\pi_{13}+6.25\pi_{14}+4.1667\pi_{23}+3.125\pi_{24}+2.0833\pi_{34}$$

is minimised.

The optimum assignment of π_{ij}, obtained by the simplex method, is given in Table 1 below.

TABLE 1. OPTIMUM ASSIGNMENT OF π_{ij}

u_i \ u_j	1	2	3	4	total
1	—	0.0	0.0	0.2	0.2
2	0.0	—	0.2	0.2	0.4
3	0.0	0.2	—	0.4	0.6
4	0.2	0.2	0.4	—	0.8
total	0.2	0.4	0.6	0.8	2.0

Yates and Grundy's assignment of probabilities is given in Table 2. In this case revised size-measures, based on three successive approximations, were used.

TABLE 2. YATES AND GRUNDY'S ASSIGNMENT OF π_{ij}

u_i u_j	1	2	3	4	total
1	—	.032	.059	.113	.204
2	.032	—	.122	.246	.400
3	.059	.122	—	.428	.600
4	.113	.246	.428	—	.787
total	.204	.400	.600	.787	2.000

Denoting by V_{opt} and V_{YG} the variances of the estimate $\hat{y}$, when the π_{ij} are taken from Tables 1 and 2 respectively, the following results (given in Table 3) are obtained. It may be observed that the relation between y and x may be assumed to be approximately linear for populations A and B but not so for population C. It is also found that the set of π_{ij} minimising (2.5) is the same as the set minimising (2.3) for populations A and B while it is not so for population C.

TABLE 3. COMPARISON OF SAMPLING VARIANCES

	V_{YG}	V_{opt}	% reduction in variance
population A	.323	.200	38.1
population B	.269	.200	25.7
population C	.057	.100	−75.4

5. Conclusion

These results show that if the relation between y and x can be assumed to be linear, the optimum assignment of π_{ij}, suggested in this note, can bring about marked reduction in variance as compared to the usual method of assignment. But this refinement is conveniently obtainable only when the number of units in a stratum is small. The method is expected to be of considerable use if the population is stratified to the maximum extent possible and the selection of two first stage units within strata is considered to be adequate.

References

Charnes, A., Cooper, W. W. and Henderson, A. (1953): "*An Introduction to Linear Programming*" John Wiley & Sons, New York.

Goodman, R. and Kish, L. (1950): Controlled selection—a technique in probability sampling. *J. Amer. Stat. Ass.*, **45**, 350-372.

Hansen, M. H. and Hurwitz, W. N. (1943): On the theory of sampling from finite populations. *Ann. Math. Stat.*, **14**, 333-362.

Horvitz, D. G. and Thompson, D. J. (1952): A generalisation of sampling without replacement from a finite universe. *J. Amer. Stat. Ass.*, **47**, 663-685.

Narain, R. D. (1951): On sampling without replacement with varying probabilities. *J. Ind. Soc. Agri. Stat.*, **3**, 169-174.

Yates, F. and Grundy, P. M. (1953): Selection without replacement from within strata with probabilities proportionate to size. *J. Roy. Stat. Soc.*, B, **15**, 253-261.

Paper received : May, 1955.