SAMPLING WITH UNEQUAL PROBABILITIES AND WITHOUT REPLACEMENT

H. O. Hartley and J. N. K. Rao
Iowa State University

I. Introduction:

Most survey designs incorporate as a basic sampling procedure the selection of n units at random, with equal probability and without replacement drawn from a population of N units. It is, however, sometimes advantageous to select units with unequal probabilities. For example, such a procedure may be found appropriate when a 'measure of size' x_i is known for all units in the population (i=1, 2, .., N) and it is suspected that these known sizes x_i are correlated with the characteristic y_i for which the population total

$$Y = \sum_{i=1}^{N} y_i \tag{1}$$

is to be estimated. For example, the sales return for the past year (y_i) of a population of companies may be correlated with the (known) sales returns for the previous year (x_i). Or, again, if the total corn production (y_i) on the i^{th} farm is the characteristic whose population total Y is to be estimated this characteristic is very likely correlated with the total farm acreage (x_i) of the i^{th} farm. One method (though by no means the only method) of utilizing the auxiliary variates (measures of size) x_i is to draw units with probabilities proportional to sizes x_i (pps), a technique frequently used in surveys, particularly for the sampling of primary sampling units in multi-stage designs. Now the theory of sampling with unequal (prescribed) probabilities is equivalent to multinomial sampling provided units are drawn <u>with</u> replacement. On the other hand, it is well known from the theory of equal probability selection that sampling with replacement results in estimators which are less precise than those computed from samples selected without replacement, the proportional variance decrease being given by the sampling fraction $\frac{n}{N}$ (finite population correction'). It has therefore been felt for sometime that similar increases in precision should be reaped by switching to a selection without replacement in unequal probability sampling. However, the theory of sampling with unequal probability and without replacement involves certain mathematical and computational difficulties and has therefore not been fully developed.

The general theory of sampling with varying probabilities and without replacement has been first given by Horvitz and Thompson (1952). Since then, several papers have been published on this topic, but we shall review here only the papers relevant to the particular problem considered in this paper which will be stated later.

Let π_i denote the probability for i^{th} unit to be in a sample of size n. The statistic

$$\hat{Y} = \sum_{1}^{n} \frac{y_i}{\pi_i} \tag{2}$$

is then an unbiased estimate of the population total Y whilst its variance is given by

$$\text{Var } \hat{Y} = \sum_{i=1}^{N} y_i^2/\pi_i + 2\sum_{i<j}^{N} P_{ij} y_i y_j/\pi_i \pi_j - Y^2 \tag{3}$$

where P_{ij} denotes the probability for the i^{th} and the j^{th} unit to be both in the sample. Horvitz and Thompson have given an unbiased estimate of the variance, but it was shown by Yates and Grundy (1953) that it can often assume negative values and they proposed an alternative unbiased estimator of variance which is believed to take negative values less often, and is given by

$$\hat{\text{Var}}\ (\hat{Y}) = \sum_{j>i}^{n} \frac{\pi_i \pi_j - P_{ij}}{P_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 . \tag{4}$$

Most of the published theory is confined to samples of size n=2, owing to considerable computational difficulties involved in the extension to larger sample sizes. Now when the π_i are exactly proportional to the y_i, Var $\hat{Y}$ is zero. This suggests that if we make the probabilities π_i proportional to the 'size measures' x_i i.e. if we put

$$\pi_i = \sum_{j \neq i}^{N} P_{ij} = c\ x_i \tag{5}$$

that a considerable reduction in Var $\hat{Y}$ will result since 'x_i' is correlated with 'y_i'. Horvitz and Thompson, for the case n=2, propose two methods to satisfy (5) approximately, but their methods have some limitations. Yates and Grundy (1953) also deal with the case n=2, introduce 'modified probabilities p_i'' and set up P_{ij} in the form

$$P_{ij} = p_i'\ p_j' \left\{ \frac{1}{1-p_i'} + \frac{1}{1-p_j'} \right\} \tag{6}$$

Equation (6) can be realized exactly by making a first draw with probability proportional to the p_i' and a second draw with probability proportional to the p_i' of the remaining (n-1) units. To satisfy (5) they substitute (6) in (5) and solve the resulting system of N nonlinear equations for the p_i' by iteration. This method becomes cumbersome when N becomes large. Des Raj (1956) employs (5) as a set of N equations for the $\frac{1}{2}$ N(N-1) probabilities P_{ij} and determines the latter by minimising (3) subject to (5). This leads to a "linear programming problem" for the $\frac{1}{2}$ N(N-1) positive P_{ij} satisfying (5). The 'objective function' (the variance) involves the population values y_i (which are unknown) and these are replaced by the 'sizes' x_i it being assumed that

$$y_i = \alpha + \beta\ x_i \tag{7}$$

exactly. Even if these assumptions are accepted the method is clearly unmanageable when n > 2 and /or for large N.

In this paper, we adopt a <u>particular</u>

procedure of drawing the sample for which it is easy to show that equation (5) is satisfied. The probabilities P_{ij} are derived directly from the sampling scheme. Although this scheme which will be described later, is well known to survey practitioners and is for example described by Horvitz and Thompson, page 678, no formulas for the P_{ij} in terms of the π_i are available in the literature, due to mathematical difficulties. These mathematical difficulties were resolved by us and compact expressions for $V(\hat{Y})$ were obtained for moderate size populations N. The sampling procedure is particularly simple for any sample size n and the asymptotic variance formulas derived permit an evaluation of the merits of both the sampling scheme as a design and the estimation (2). It should be pointed out that this method and the results (unlike the procedures previously published) cover the case of general sample size n.

The mathematical derivations of P_{ij} and $V(\hat{Y})$ in terms of the π_i will be published elsewhere. Here we shall confine ourselves to describing the sampling procedure with an example, and to stating the formulas for P_{ij}, $\mathrm{Var}(\hat{Y})$ and $\widehat{\mathrm{Var}}(\hat{Y})$. Finally we shall give an example to illustrate the efficiency comparisons.

II. The Sampling Scheme:

It can be shown easily that a necessary condition for a sampling scheme to satisfy (5) is that

$$c = n \Big/ \sum_{i=1}^{N} x_i$$

and

$$\pi_i = n\, x_i \Big/ \sum_{j=1}^{N} x_j \leqslant 1 \quad . \tag{8}$$

Henceforth, we shall only consider such 'sizes x_i' and associated probabilities $p_i = x_i \Big/ \sum_{j=1}^{N} x_j$ which satisfy the necessary condition (8). The following sampling scheme is now considered:

a. Arrange the units in random order and denote (without loss of generality) by j=1, 2,.., N this random order and by

$$\Pi_j = \sum_{i=1}^{j} \pi_i \, , \Pi_o = 0 \tag{9}$$

the progressive totals of the π_i in that order.

b. Select a 'random start' i.e. select a 'uniform variate' x with $0 < x < 1$. Then the n selected units are those whose index, j, satisfies

$$\Pi_{j-1} < x + k \leqslant \Pi_j \tag{10}$$

for some integer k between 0 and (n-1). Since $\pi_i \leqslant 1$ every one of the n integers k=0, 1, .., n-1 will 'select' a different sampling unit j.

Numerical Example:

Consider the population of N = 8 units j=1, 2, .., 8 arranged in random order and with 'sizes x_j' shown in the second column of Table 1. A sample of n=3 units is to be drawn with probabilities proportional to size (pps) and without replacement. The 'total size' is

$$X = \sum_{j=1}^{8} x_j = 300 \quad .$$

Instead of computing the $p_i = x_i/X$ and $\pi_i = 3p_i$ we scale all computations up by a factor of $X/n = 300/3 = 100$. Thus we compute the progressive sums of the x_i and these are shown in column three of Table 1. and correspond to the quantities $(X/n)\,\Pi_j$ defined by (9). Then we select a random integer (start) between 1 and (X/n) i.e. between 1 and 100

Table 1. An example of the selection of n = 3 units from a population of N = 8 units with probabilities proportional to size and without replacement.

Unit number	Size	Progressive sum	Start=36
			Step=X/n=300/3=100

j	x_j	300 $_j$ =	
1	15	15	
2	81	96	k=o,100x=36
3	26	122	
4	42	164	k=i,100x+100=36
5	20	184	
6	16	200	
7	45	245	k=2,100x+200=236
8	55	300	

and this corresponds to the quantity X x/n. In our example this integer turned out to be 36 and the selection of the three units in accordance with (10) is shown in the 4th column: We must find the lines(j) where the column $300\,\Pi_j$ passes through the levels 100x = 36(for k=o), 100x + 100 = 136 (for k=1) and 100x + 200 = 236 (for k=2). The units j=2, 4, and 7 are thereby selected. This procedure (either with or without the initial randomization a.) has been frequently used but, in the absence of a better theory, is usually treated approximately as a pps sample drawn with replacement.

It can be easily proved that for this sampling scheme, for any ordered arrangement of the N units

$$\Pr\left\{ j^{\text{th}} \text{ unit in sample} \right\} = \pi_j \tag{11}$$

thus satisfying condition (5). It may be remarked that the randomization of the arrangement in step 'a' of the scheme is not required to prove (11). However this is required for obtaining a variance formula for the estimate of Y which does not depend on any particular arrangement of the units.

From equation (3), we see that in order to evaluate Var $\hat{Y}$ in terms of π_i's, we have to find P_{ij} in terms of the π_i . Now we have succeeded in evaluating directly from the sampling scheme the leading term of P_{ij} as well as the term which is of next lower order of magnitude in powers of N^{-1}. This second term represents the gain in precision due to the finite population correction. In particular we have shown that to order N^{-3}, for the case n=2,

$$P_{ij} = \frac{N-2}{N-1} \frac{\pi_i \pi_j}{2-\pi_{i-} \pi_j} \left(1- \frac{S^2(N-2)}{(2-\pi_i - \pi_j)2} \right) \qquad (12)$$

where

$$S^2 = \frac{1}{N-3} \left(\sum_{t=1}^{N} \pi_t^2 - \pi_i^2 - \pi_j^2 \right) - \frac{(2-\pi_i - \pi_j)^2}{(N-2)} . \qquad (13)$$

By substituting P_{ij} from equation (12) into (3) and simplifying, we obtained

$$\mathrm{Var}(\hat{Y}) = \sum_{i=1}^{N} \pi_i \left(1- \frac{\pi_i}{2}\right)\left(\frac{y_i}{\pi_i} - \frac{Y}{2}\right)^2 \qquad (14)$$

to order N^{-1}. This formula is satisfactory for moderately large populations N. Further improvements in $\mathrm{Var}(\hat{Y})$ have been made by taking terms of order N^0 also into account, but these details will not be given here. We notice that for sampling with replacement,

$$\mathrm{Var}'(\hat{Y}) = \sum_{i=1}^{N} \pi_i \left(\frac{y_i}{\pi_i} - \frac{Y}{2} \right)^2 \qquad (15)$$

so that sampling without replacement and pps is more efficient than sampling with replacement and pps, since the weight factors $1- \frac{\pi_i}{2}$ in (14) are all less than 1.

In order to find an estimate of the variance of $\hat{Y}$, we substitute P_{ij} from equation (12) in equation (4) and after simplification we obtain to order N^{-1},

$$\hat{\mathrm{Var}}(\hat{Y}) = \left(1-(\pi_1+\pi_2) + \frac{\sum^{N} \pi_t^2}{2}\right)\left(\frac{y_1}{\pi_1} - \frac{y_2}{\pi_2}\right)^2 . \qquad (16)$$

A check on our formulas is obtained by making all π_i equal i.e. equal to $\frac{2}{N}$ when it will be seen that, to order N^{-1}, equations (14) and (16) reduce to the well known formulas for the equal probability case available for the variance of $\hat{Y}$ and for the estimate of variance of $\hat{Y}$. In the case of general n, we have shown that

$$\mathrm{Var}(\hat{Y}) = \sum_{1}^{N} \pi_i \left(1-\frac{(n-1)}{n} \pi_i\right)\left(\frac{y_i}{\pi_i} - \frac{Y}{n} \right)^2 \qquad (17)$$

to order N^{-1} assuming n is relatively small compared to N. The details of the case of general n will be given in a seperate paper.

III. A numerical example for the evaluation of the variance formulas:

Horvitz and Thompson (1952)(pp. 681-3) give an example of a small population of size N=20 for which the data are reproduced in Table 2. below: For N=20 blocks in Ames, Iowa, are given

y_i = Number of households in ith block
x_i = 'Eye-estimate' of number of households in ith block

Table 2. Data for population of size N=20

i=	1	2	3	4	5	6	7	8	9	10
y_i=	19	9	17	14	21	22	27	35	20	15
x_i=	18	9	14	12	24	25	23	24	17	14
i=	11	12	13	14	15	16	17	18	19	20
y_i=	18	37	12	47	27	25	25	13	19	12
x_i=	18	40	12	30	27	26	21	9	19	12

Now the probability π_i for the ith unit to be in the sample is taken proportional to the 'eye-estimated' number of households x_i i.e. we take

$$\pi_i = 2x_i \Big/ \sum_{j=1}^{20} x_j \quad . \qquad (18)$$

In Table 3. below we give the evaluation of the variance of our estimator from formula (14) correct to order N^1 and shown below this is the value obtained from an improved formula (not shown here) which is correct to order N^0. Above this value we give the variance formula for sampling with probabilities proportional to the π_i but with replacement. And finally (on top) the variance of $N\bar{y}$ i.e. of the estimator when sampling is with equal probabilities and without replacement.

Table 3. Variances of various estimates of the total of the y_i population shown in Table 2.

Sampling Scheme	Form of Estimator	Numerical value of variance of estimator
Equal probability sampling without replacement	$N\bar{y}$	16,219
Probabilities proportional to size x_i, with replacement	$\sum_{i=1}^{2} y_i/\pi_i$	3,241
Probabilities proportional to size, xi; without replacement	$\sum_{i=1}^{2} y_i/\pi_i$	3,025 (14) 3,007

A comparison of the variances in Table 3. shows that sampling proportional to an approximate measure of size is vastly superior to sampling with equal probabilities. It must not be forgotten, however, that there are other devices of decreasing the variance in the latter ease with the help of the known x_i values: Ratio and regression estimation, for example, may be employed. About 7 per cent (235/3241) are gained in precision through sampling without replacement. The two results for our variance viz. 3025 (correct to order N^1) and 3007 (correct to order N^0) are in good agreement and illustrate the convergence of our formulas.

References

1. Des Raj (1956), "A Note On the Determination of Optimum Probabilities in Sampling Without Replacement," Sankhya, p. 197.
2. Horvitz, D. G. and Thompson, D. J. (1952), "A Generalization of Sampling Without Replacement From a finite Universe", Journal of the American Statistical Association, p. 663.
3. Yates, F. and Grundy, P. M. (1953), "Selection Without Replacement from Within Strata With Probability Proportional to Size", Journal of the Royal Statistical Society, Series B, p. 253.

Editorial Note: The authors inform us that they are submitting a paper giving full details of the mathematical proofs to the ANNALS OF MATHEMATICAL STATISTICS.