ASYMPTOTIC EQUIVALENCE OF EFFICIENCY FOR THREE PROCEDURES OF UNEQUAL PROBABILITY SAMPLING WITHOUT REPLACEMENT

J. N. K. Rao
Iowa State University

I. Introduction

Let y_j denote a characteristic attached to the j^{th} unit of a finite population of N units with population total $Y = \Sigma y_j$. It is well known that the use of unequal probabilities in selecting a sample may bring about considerable reduction in variance as compared to equal probability sampling. For example, such a procedure may be useful when a 'measure of size' x_j is known for all units in the population and it is suspected that these known sizes x_j are correlated with the characteristic y_j. One method (though by no means the only method) of utilizing x_j is to draw units with probabilities proportional to sizes x_j (p.p.s.), a technique frequently used in surveys, particularly for selecting primary units in multistage designs. Now the theory of sampling with unequal probabilities is equivalent to multinomial sampling provided the units are drawn with replacement. On the other hand, we know from equal probability sampling that selection with replacement results in estimators which are less precise than those computed from samples selected without replacement, the proportional reduction in the variance being given by the sampling fraction n/N. It is, therefore, natural to investigate the efficiency of unequal probability sampling without replacement as compared to unequal probability sampling with replacement.

A general theory of unequal probability sampling without replacement is given by Horvitz and Thompson (1952). Their estimator of Y is

$$\hat{Y} = \sum_{1}^{n} y_j/P_j \qquad (1)$$

with variance

$$V(\hat{Y}) = \sum_{1}^{N} \frac{y_j^2}{P_j} + 2 \sum_{i'>i=1}^{N} \frac{P_{ii'}}{P_i P_{i'}}\, y_i y_{i'} - Y^2, \qquad (2)$$

where P_j is the probability for the j^{th} unit to be in a sample of size n and $P_{ii'}$ is the probability for the units i and i' both to be in the sample.* Now, when the P_j are proportional to the y_j, $\hat{Y}$ is constant and hence $V(\hat{Y})$ is zero which suggests that, by making P_j proportional to the known sizes x_j, considerable reduction in the variance will result if the x_j are approximately proportional to the y_j. Special devices are needed to satisfy this condition, namely $P_j = np_j$ where $p_j = x_j/\Sigma x_j$, when sampling with p.p.s. and without replacement. Yates and Grundy (1953) suggest an iterative procedure to obtain revised sizes x_j^* and draw the sample by selecting the first unit with probabilities proportional to the revised sizes, the second unit with probabilities proportional to the remaining (revised sizes), and so on, such that $P_j = np_j$. Let us call their method procedure 1. One can also use the iterative procedure of obtaining revised sizes for another well known sampling scheme which is as follows: n units are drawn with probabilities proportional to revised sizes with replacement. If any unit is selected more than once in the sample, reject the n selections and make further n selections with replacement and with probabilities proportional to revised sizes, the process being continued until n different units are selected in the sample. Let us call this method procedure 2 (e. g., Durbin (1953)). In order to draw a sample by procedure 1 or 2 it is convenient to use a method suggested by Lahiri (1951).* The application of Lahiri's method leads to the following scheme of drawing the sample for procedure 1: Let us consider, for illustration, n=2. Let x_0^* be a number not smaller than the largest of x_j^*. Select a random integer between 1 and N, say α where $1 \leq \alpha \leq N$. Select similarly a random number β, subject to the condition $0 < \beta \leq x_0^*$. If $\beta \leq x_\alpha^*$, then unit α is selected for the sample. If $\beta > x_\alpha^*$, unit α is not selected (at this 'draw') and the whole process has to be repeated with all the units until a selection is made. The probability of selecting a unit by this procedure is proportional to the revised sizes. After a unit is selected, repeat the whole process with the remaining x_j^* until another selection is made. For procedure 2, after a unit is selected by Lahiri's method, the whole process is repeated with all the x_j^* for the selection of the second unit. If the second unit selected happens to be the same as the first unit, reject both selections and repeat the whole process with all the units until two different selections are made.

There is, however, a well known procedure of drawing a sample with p.p.s. and without replacement which insures that $P_j = np_j$ with the original sizes x_j. This procedure, which we denote as procedure 3, is as follows: The N units in the population are listed in a random order and their x_j are cumulated and a systematic selection of n elements from a random start is then made on the cumulation (e.g., Goodman and Kish (1950)). It may be noted that for all the three procedures all x_j's should satisfy the necessary condition $P_j = np_j \leq 1$. Therefore, if there are some units for which $np > 1$, either select them automatically in the sample or use some other device like stratification or splitting a unit into smaller units, etc.

Recently (Hartley and Rao (1959), (1962)), the mathematical difficulties involved in evaluating the probabilities $P_{ii'}$, for procedure 3, are resolved with the help of an asymptotic

*It may be pointed out that in unequal probability sampling there are several different classes of linear estimators. The estimator (1) belongs to one of these classes and we concentrate here only on the estimator (1). For a detailed discussion of different classes of linear estimators the reader is referred to Koop (1957).

*Professor Leslie Kish suggested that I point out how Lahiri's method can be used to draw a sample by procedure 1 or 2.

theory, and compact expressions for the variance of the estimate of the population total together with variance estimates have been obtained. The following approach is used in developing the asymptotic theory: In sampling with p.p.s. and with replacement we have

$$\hat{Y}^{1} = \sum_{n} y_j/np_j, \tag{3}$$

where $\sum_{n}$ denotes the summation over the n units selected with replacement, as the estimator of Y. Also the variance of $\hat{Y}^{1}$ is

$$V(\hat{Y}^{1}) = \sum_{1}^{N} P_j\left(\frac{y_j}{P_j} - \frac{Y}{n}\right)^2 \tag{4}$$

where $P_j = np_j$. Therefore, by assuming that each P_j is of the order of N^{-1} for large N, it is seen that $V(\hat{Y}^{1})$ is of order N^2. In sampling without replacement the term of order N^2 will be the leading term in $V(\hat{Y})$ and hence the next lower order terms, namely terms of order N^1, will represent the gain in precision due to sampling without replacement. Therefore, Hartley and Rao evaluated (for their procedure 3) $V(\hat{Y})$ to order N^1. This is equivalent to evaluating $P_{ii'}$ to order N^{-3} and substituting it in the variance formula (2). Also, for the benefit of smaller size populations, they evaluated $P_{ii'}$ to order N^{-4} and hence $V(\hat{Y})$ to order N^0.

The purpose of the present paper is to present compact expressions for the variance and estimated variance together with simplified formulas for revised sizes in terms of the original sizes for procedures 1 and 2, obtained by using the asymptotic theory. The mathematical derivations are not given here and will be published elsewhere. It is shown that the three procedures have exactly the same value of $P_{ii'}$ to order N^{-3} and hence identical $V(\hat{Y})$ to order N^1. Since the terms of order N^1 are the important terms in $V(\hat{Y})$, which contribute to the gain in precision of sampling without replacement over sampling with replacement for large N, it follows that the three procedures have practically the same efficiency. However, with procedure 3 there is no need to compute the revised sizes, and this procedure, therefore, circumvenes an operation which may be cumbersome as N becomes large. It is also shown for the case n = 2 that, to order N^0, the estimator from procedure 1 has smaller variance than that from procedure 2 and that from procedure 2 has smaller variance than that from procedure 3. However, as N increases, the contribution from terms of order N^0 becomes negligible.

II. The case n = 2

Let $p_j^* = x_j^*/\Sigma x_j$. Then, for procedure 1, we find

$$p_j^* = p_j\left[1 + \frac{1}{4}\left(2p_j - 2\sum_1^N p_t^2 + 4p_j^2 - 3p_j\sum_1^N p_t^2 + 3\left(\sum_1^N p_t^2\right)^2 - 4\sum_1^N p_t^3\right)\right] \tag{5}$$

to order N^{-3}. If the sample is selected by procedure 1 using the revised sizes from (5), we obtain

$$v^{(1)}(\hat{Y}) = \sum_1^N P_j\left(1 - \frac{P_j}{2}\right)\left(\frac{y_j}{P_j} - \frac{Y}{2}\right)^2 - \frac{1}{2}\sum_1^N \left(P_j - \frac{P_j^2\sum_1^N P_t^2}{4}\right)\left(\frac{y_j}{P_j} - \frac{Y}{2}\right)^2 + \frac{3}{32}\left(\sum_1^N P_t y_t - \frac{Y\sum_1^N P_t^2}{2}\right)^2 \tag{6}$$

and

$$v^{(1)}(\hat{Y}) = \left[1 - (P_i + P_{i'}) + \frac{1}{2}\sum_1^N P_t^2 - \frac{1}{2}(P_i^2 + P_{i'}^2) + \frac{5}{8}P_iP_{i'} - \frac{3}{32}\left(\sum_1^N P_t^2\right)^2 - \frac{1}{16}(P_i + P_{i'})\sum_1^N P_t^2 + \frac{1}{2}\sum_1^N P_t^3\right] \cdot \left(\frac{y_i}{P_i} - \frac{y_{i'}}{P_{i'}}\right)^2 \tag{7}$$

to order N^0, where $P_j = 2p_j$. Here $v^{(1)}(\hat{Y})$ and $v(1)(\hat{Y})$ denote the variance and the estimated variance respectively for procedure 1, and i and i' are the two units included in the sample. For procedure 2 we have

$$p_j^* = p_j\left[1 + \left(p_j - \sum_1^N p_t^2 + 2p_j^2 - 2p_j\sum_1^N p_t^2 + 2\left(\sum_1^N p_t^2\right)^2 - 2\sum_1^N p_t^3\right)\right] \tag{8}$$

to order N^{-3}. If the sample is selected by procedure 2 using the revised sizes from (8), we find

$$v^{(2)}(\hat{Y}) = \sum_1^N P_j\left(1 - \frac{P_j}{2}\right)\left(\frac{y_j}{P_j} - \frac{Y}{2}\right)^2 - \frac{1}{2}\sum_1^N \left(P_j^3 - \frac{P_j^2\sum_1^N P_t^2}{2}\right)\left(\frac{y_j}{P_j} - \frac{Y}{2}\right)^2 + \frac{1}{8}\left(\sum_1^N P_t y_t - \frac{Y\sum_1^N P_t^2}{2}\right)^2 \tag{9}$$

and

$$v^{(2)}(\hat{Y}) = \left[1 - (P_i + P_{i'}) + \frac{1}{2}\sum_1^N P_t^2 - \frac{1}{2}(P_i^2 + P_{i'}^2) + \frac{1}{2}P_iP_{i'} - \frac{1}{8}\left(\sum_1^N P_t^2\right)^2 + \frac{1}{2}\sum_1^N P_t^3\right]\left(\frac{y_i}{P_i} - \frac{y_{i'}}{P_{i'}}\right)^2 \tag{10}$$

to order N^0, where $V^{(2)}(\hat{Y})$ and $v^{(2)}(\hat{Y})$ denote the variance and the estimated variance respectively for procedure 2. Finally, for procedure 3, denoting the variance and the estimated variance by $V^{(3)}(\hat{Y})$ and $v^{(3)}(\hat{Y})$ respectively, we find

$$v^{(3)}(\hat{Y}) = \sum_1^N P_j\left(1 - \frac{P_j}{2}\right)\left(\frac{y_j}{P_j} - \frac{Y}{2}\right)^2 - \frac{1}{2}\sum_1^N \left(P_j^3 - \frac{P_j^2 \sum_1^N P_t^2}{2}\right)\left(\frac{y_j}{P_j} - \frac{Y}{2}\right)^2 + \frac{1}{4}\left(\sum_1^N P_t y_t - \frac{Y \sum_1^N P_t^2}{2}\right)^2 \quad (11)$$

and

$$v^{(3)}(\hat{Y}) = \left[1 - (P_i + P_{i'}) + \frac{1}{2}\sum_1^N P_t^2 - \frac{1}{2}(P_i^2 + P_{i'}^2) - \frac{1}{4}\left(\sum_1^N P_t^2\right)^2 + \frac{1}{4}(P_i + P_{i'})\sum_1^N P_t^2 + \frac{1}{2}\sum_1^N P_t^3\right] \cdot \left(\frac{y_i}{P_i} - \frac{y_{i'}}{P_{i'}}\right)^2 \quad (12)$$

to order N^0. From (6), (9) and (11) it follows that $V^{(1)}(\hat{Y}) < V^{(2)}(\hat{Y}) < V^{(3)}(\hat{Y})$, to order N^0. However, the three procedures have exactly the same variance to order N^1, namely

$$V(\hat{Y}) = \sum_1^N P_j\left(1 - \frac{P_j}{2}\right)\left(\frac{y_j}{P_j} - \frac{Y}{2}\right)^2. \quad (13)$$

Also, the three procedures have exactly the same estimated variance to order N^1, namely

$$v(\hat{Y}) = \left[1 - (P_i + P_{i'}) + \frac{1}{2}\sum_1^N P_t^2\right] \cdot \left(\frac{y_i}{P_i} - \frac{y_{i'}}{P_{i'}}\right)^2. \quad (14)$$

Equation (13) when compared with the variance in sampling with replacement, namely (4), clearly demonstrates the reduction in variance achieved by sampling without replacement through the 'finite population corrections' $(1 - P_j/2) < 1$.

Since the three procedures have exactly the same variance to order N^1 and since the terms of order N^1 are the important terms that contribute to the reduction in the variance achieved by sampling without replacement for large N, one can conclude that in most of the practical situations there is little to choose between the three procedures on the basis of efficiency alone. However, since there is no need to compute revised sizes with procedure 3, this may be preferred to other procedures. If one is using procedure 1 and is satisfied with the variance to order N^1, then the revised sizes x_j^* are obtained from the simplified formula

$$p_j^* = p_j\left[1 + \frac{1}{2}\left(p_j - \sum_1^N p_t^2\right)\right] \quad (15)$$

to order N^{-2}. Similarly, for procedure 2, the revised sizes x_j^* are obtained from

$$p_j^* = p_j\left[1 + \left(p_j - \sum_1^N p_t^2\right)\right] \quad (16)$$

to order N^{-2}.

An example for the evaluation of the variance formulas

Horvitz and Thompson (1952, Table 2) give the data of a population of N = 20 blocks in Ames, Iowa. Here y_i and x_i denote the number of households and the 'eye-estimated' number of households respectively in the i^{th} block. Using their data the following values are obtained:

$V(\hat{Y}^1) = 3{,}241$ using (4)

$V(\hat{Y}) = 3{,}025$ to order N^1 using (13).

Incidentally, for this example, $\sum_{20} P_t y_t = 49.63$ and $Y \sum_{20} P_t^2/2 = 49.95$. Therefore, we find that, to the nearest integer and to order N^0, $V^{(1)}(\hat{Y}) = V^{(2)}(\hat{Y}) = V^{(3)}(\hat{Y}) = 3{,}007$. It may be of interest to exhibit the nature of convergence of the various approximations to the variance by regarding the variance formula for sampling with replacement as an approximation to order N^2 as set out in Table 1 below.

Table 1. Terms in the approximations to the variance formula

Order of approximation	Sampling procedure	Variance	Difference
$O(N^2)$	with replacement	3,241	216
$O(N^1)$	procedures 1, 2 and 3	3,025	18
$O(N^0)$	procedures 1, 2 and 3	3,007	

The convergence in this example appears to be satisfactory although the population size (N = 20) is much smaller than those usually encountered in survey work. The variance in equal probability sampling without replacement for this example is 16,219. Therefore, all these procedures of p.p.s. sampling are vastly superior compared to equal probability sampling. It must not be forgotten, however, that there are other devices of decreasing the variance in the latter case with the help of the known x_i values (e.g., ratio estimation). The gain in precision through unequal probability sampling without replacement as compared to sampling with replacement is about 7% (235/3241).

III. The general case $n \geq 2$

Most of the published literature on unequal probability sampling deals only with the case n=2 and does not have anything to offer for $n > 2$ due to difficulties in evaluating $P_{ii'}$, and hence $V(\hat{Y})$ and $v(\hat{Y})$. Though the case n = 2 is important, particularly in stratified designs, situations often arise when n is greater than 2. A striking feature of our asymptotic approach is that it permits an easy evaluation of $V(\hat{Y})$ and $v(\hat{Y})$ for $n > 2$. It can be shown that the three procedures have exactly the same variance and estimated variance to order N^1, namely

$$V(\hat{\bar{Y}}) = \sum_{1}^{N} P_j \left[1 - \frac{(n-1)}{n} P_j\right] \left(\frac{y_j}{P_j} - \frac{Y}{n}\right)^2 \quad (17)$$

and

$$v(\hat{\bar{Y}}) = (n-1)^{-1} \sum_{i>i'}^{n} \left[1 - (P_i + P_{i'}) + \frac{1}{n}\sum_{1}^{N} P_t^2\right] \left(\frac{y_i}{P_i} - \frac{y_{i'}}{P_{i'}}\right)^2, \quad (18)$$

where $P_j = np_j$. The assumption that P_j is of order N^{-1} implies that n is relatively small compared to N. For procedure 1 the revised sizes x_j^* are obtained from

$$p_j^* = p_j \left[1 + \frac{(n-1)}{2} p_j - \frac{(n-1)}{2} \sum_{1}^{N} p_t^2\right], \quad (19)$$

and for procedure 2 the x_j^* are found from

$$p_j^* = p_j \left[1 + (n-1) p_j - (n-1) \sum_{1}^{N} p_t^2\right] \quad (20)$$

to order N^{-2}.

The extension of the theory, presented here for simple sampling, to multi-stage sampling, etc., is fairly straightforward and will not be discussed here.

References

Durbin, J. (1953). Journal of the Royal Statistical Society, Series B, p. 253.

Goodman, R. and Kish, L. (1950). Journal of the American Statistical Association, p. 350.

Hartley, H. O. and Rao, J. N. K. (1959). Proceedings of the Social Statistics Section, American Statistical Association, 1960, p. 56.

Hartley, H. O. and Rao, J. N. K. (1962). Annals of Mathematical Statistics (in press).

Horvitz, D. G. and Thompson, D. J. (1952). Journal of the American Statistical Association, p. 663.

Koop, J. C. (1957). Institute of Statistics, Mimeo Series No. 296, North Carolina.

Lahiri, D. B. (1951). Bulletin of the International Statistical Institute, Part II, p. 133.

Yates, F. and Grundy, P. M. (1953). Journal of the Royal Statistical Society, Series B, p. 235.