ON THE DETERMINATION OF OPTIMUM PROBABILITIES IN SAMPLING

By Morris H. Hansen and William N. Hurwitz

Bureau of the Census

1. Summary. In a previous paper [2] it was shown that it is sometimes profitable to select sampling units with probability proportionate to size of the unit. This note indicates a method of determining the probabilities of selection which minimize the variance of the sample estimate at a fixed cost. Some approximations that have practical applications are given.

2. Introduction. Neyman has shown that it is possible to reduce the sampling variance of an estimate by dividing a population into sub-populations (called strata) and varying the proportions of units included in the sample from stratum to stratum [1]. His treatment presumed that the units within any stratum would be drawn with equal probability. In many practical sampling problems, the use of constant probabilities is neither necessary nor desirable. Not only is it possible to obtain unbiased or consistent estimates with varying probabilities of selection of the sampling units, but also it is possible to reduce the variance of sample estimates by appropriate use of this device.

It has been shown [2] that in a subsampling system, the selection of primary units with probabilities proportionate to the number of elements included in the primary unit may bring about marked reductions in sampling variances over sampling with equal probabilities. In this note, we shall indicate a method of determining the optimum probabilities under certain conditions, and also some approximations to the optima that have practical applications.

By optimum probabilities, we mean the set of probabilities of selection that will minimize the variance for a fixed cost of obtaining sample results, or alternatively that will minimize the cost for a fixed sampling error.

3. Optimum probability with a subsampling system. Consider, for example, the simple subsampling system where primary units are first drawn for inclusion in the sample and then a sample of elements is drawn from the selected primary units. We shall suppose, for simplicity of notation, that the sampling is done without stratification. The conclusions indicated below will be similar if stratified sampling is used, and they will hold even if only one unit is drawn from each stratum. Suppose that a population contains M primary units, and that the sampling of primary units is to be done with replacement. Sampling with replacement is assumed in order to simplify the mathematics. We wish to estimate the ratio

$$\frac{X}{Y} = \frac{\sum_{i=1}^{M} \sum_{j=1}^{N_i} X_{ij}}{\sum_{i=1}^{M} \sum_{j=1}^{N_i} Y_{ij}}$$

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Mathematical Statistics.* www.jstor.org

where X_{ij} and Y_{ij} are the values of two characteristics of the jth element within the ith primary unit, and N_i is the number of elements in the ith primary unit. A consistent estimate of X/Y is given by

$$r = \frac{\sum_{i=1}^{m} \frac{N_i}{P_i} \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}}{\sum_{i=1}^{m} \frac{N_i}{P_i} \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}} \tag{1}$$

where

P_i = The probability of selecting the ith primary unit on a single draw.

n_i = The total number of elements included in the sample from the ith unit if it is drawn. If a particular unit happens to be included in the sample more than once the subsampling will be independently carried through each time it is drawn.

m = The total number of primary units included in the sample.

It will be assumed that a self-weighting sample is to be used, i.e., that although the probabilities of selecting primary units will vary, the subsampling rate within the ith selected primary unit, $\frac{n_i}{N_i}$, will be such that $P_i \frac{n_i}{N_i} = k$. Note that, with this condition, k is the probability that an element will be included in the sample by making a single draw of a primary unit, and by carrying out the specified subsampling within the selected primary unit. It follows that mkN is the expected total number of elements included in a sample of primary units, where

$$N = \sum_{i=1}^{M} N_i.$$

The method can be extended to cover situations where other conditions are imposed.

We shall express the variance of r in terms of P_i, m, and k, and also express the cost in terms of these same quantities. The optimum values of P_i, m, and k will then be determined.

The variance of the sample estimate. To terms of order $1/m$ of the Taylor expansion of a ratio, the sampling variance of the estimate (1) is approximately

$$\sigma_r^2 \doteq \frac{\sum_{i=1}^{M} \frac{N_i^2}{P_i} \Delta_i^2 + \sum_{i=1}^{M} \frac{N_i^2}{P_i} \frac{N_i - n_i}{N_i n_i} \sigma_i^2}{mY^2} \tag{2}$$

where

$$\Delta_i^2 = \bar{Y}_i^2 \left(\frac{\bar{X}_i}{\bar{Y}_i} - \frac{X}{Y} \right)^2, \qquad \bar{Y}_i = \frac{\sum_{j=1}^{N_i} Y_{ij}}{N_i}, \qquad \bar{X}_i = \frac{\sum_{j=1}^{N_i} X_{ij}}{N_i},$$

$$\sigma_i^2 = \sigma_{ix}^2 + \frac{\bar{X}^2}{\bar{Y}^2}\sigma_{iy}^2 - 2\frac{\bar{X}}{\bar{Y}}\sigma_{ixy},$$

$$\sigma_{ix}^2 = \frac{\sum_{j=1}^{N_i}(X_{ij} - \bar{X}_i)^2}{N_i - 1},$$

$$\sigma_{iy}^2 = \frac{\sum_{j=1}^{N_i}(Y_{ij} - \bar{Y}_i)^2}{N_i - 1},$$

$$\sigma_{ixy} = \frac{\sum_{j=1}^{N_i} X_{ij} - \bar{X}_i)(Y_{ij} - \bar{Y}_i)}{N_i - 1}.$$

The cost function. Now suppose that the total cost of the sampling procedure involves a fixed cost attached to each primary unit included in the sample, a cost of listing the elements within each selected primary unit (this listing may be necessary in order to draw a subsample), and a cost of obtaining information from each of the elements selected for inclusion in the sample. Under these circumstances the total expected cost of the survey will be:

$$C = C_1 m + C_2 m \sum_{i=1}^{M} P_i N_i + C_3 mkN \tag{3}$$

where

C_1 = The fixed cost per primary unit,
C_2 = The cost of listing one element in a selected primary unit and other costs that vary with the number of elements to be listed,
C_3 = The cost of obtaining the required information from one element in the sample,
$\sum_{i=1}^{M} P_i N_i$ = Expected number of elements in the sample per primary unit in the sample,
mk = The over-all sampling ratio, and
$N = \sum_{i=1}^{M} N_i$ = The total number of elements in the population.

It will be noted that although the values of P_i and m may be fixed in advance, the number of elements to be listed, $\sum_{i=1}^{m} N_i$, remains a chance variable. It is for this reason that we consider the expected cost rather than the actual cost.

The optimum values of P_i, m, and k. The values of P_i, m, and k which minimize the variance (2) subject to the conditions that:

$$C \text{ is fixed}, \qquad \frac{n_i}{N_i} P_i = k, \qquad \sum_{i=1}^{M} P_i = 1,$$

are given by

$$P_i = \frac{\sqrt{\frac{N_i^2 \delta_i}{C_1 + C_2 N_i}}}{\sum_{i=1}^{M} \sqrt{\frac{N_i^2 \delta_i}{C_1 + C_2 N_i}}}, \tag{4}$$

$$k = \frac{\sqrt{\frac{\sum_{1=1}^{M} N_i \sigma_i^2}{N}}}{\sum_{i=1}^{M} \sqrt{\frac{N_i^2 \delta_i}{C_1 + C_2 N_i}}\, C_3}, \tag{5}$$

$$m = \frac{C}{C_1 + C_2 \sum_{i=1}^{M} P_i N_i + C_3 kN}, \tag{6}$$

where

$$\delta_i = \Delta_i^2 - \frac{\sigma_i^2}{N_i}.$$

Ordinarily δ_i will be positive although it will often be found to be negative for some i. For a great many populations, such negative values can be avoided by classifying the primary units into size groups or other significant groups and then requiring that the probability of selection be P_α for every primary unit in the α-th group.

In actual practice, however, in advance of designing a sample one does not have the data to compute the optima and uses methods of approximating the optimum probabilities. Methods of approximating the optimum probabilities are given below.

4. Some rules for approximating the optimum probabilities. In another paper [2] considerations were presented from which it follows that δ_i tends to decrease with increasing size of unit, but seldom as fast as the size of unit increases. The rate of decrease is often small relative to the increase in N_i, and empirical data for a number of problems indicate that even the assumption of δ_i being fairly constant with increasing size of unit may not lead one far astray from the optimum probabilities. Under this assumption ($\delta_i = \delta$ for all i) the probabilities depend only on N_i, C_1, and C_2, and lead to the following results:

(a) When $C_1 > 0$ and $C_2 = 0$, probability proportionate to size will be the optimum.

(b) When $C_1 = 0$ and $C_2 > 0$, probability proportionate to the square root of the size will be the optimum.

If we go to the other extreme (extreme not in terms of mathematically possible values but in terms of most practical populations), and assume that δ_i decreases at the same rate that N_i increases, the results would be:

(a) When $C_1 > 0$ and $C_2 = 0$, probability proportionate to the square root of the size will be the optimum.

(b) When $C_1 = 0$ and $C_2 > 0$, equal probability will be the optimum.

The minimum is broad in the neighborhood of the optimum and the results for either of these extremes and the values in between often will give results reasonably close to the minimum. This leads to the following useful approximations:

(a) When $C_2 \Sigma P_i N_i$, the expected cost per primary unit of listing and related operations, is small in relation to C_1, the fixed cost per primary unit, the optimum probabilities will be between probability proportionate to size and probability proportionate to the square root of size, and either of these will be reasonably close to the optimum.

(b) When C_1 is small compared to $C_2 \Sigma P_i N_i$, the optimum probability will be between equal probability and probability proportionate to the square root of size, and either of these will be reasonably close to the optimum.

(c) When both C_1 and $C_2 \Sigma P_i N_i$ are of significant size, i.e., when the costs vary substantially both with the number of primary units in the sample and the size of the units, then probability proportionate to the square root of the size will be a reasonably good approximation to the optimum.

(d) When units of small size are used and all of the subunits in the selected primary units are included in the sample (that is, there is no subsampling) equal probability is close to the optimum. It should be noted that this rule does not follow directly from the above analysis based on subsampling, but from a separate analysis in which no subsampling is involved.

For whatever system of probabilities is used, and with the cost function given by (3), the optimum value of k is given by:

$$k = \sqrt{\frac{\sum_{i=1}^{M} N_i \sigma_i^2 \left(C_1 + C_2 \sum_{j=1}^{M} P_j N_j \right)}{C_3 N \left(\sum_{i=1}^{M} \frac{N_i^2 \Delta_i^2}{P_i} - \sum_{i=1}^{M} \frac{N_i \sigma_i^2}{P_i} \right)}},$$

which can be approximated, in application, from prior experience or preliminary studies. The corresponding optimum value for m is obtained by substitution in the cost function.

The above results should not be accepted, of course, as the optima for every cost function or every sampling system. Either past experimental data may be available or pilot tests made to determine the cost function and the appropriate approximations that should be used in various practical situations.

An illustration. An illustration may be of interest. A characteristic published for city blocks in the 1940 Census of Housing is the number of dwelling units that are in need of major repairs or that lack a private bath. Suppose we

were sampling to estimate the proportion of the dwelling units having this characteristic for the Bronx in New York City, at the time of the 1940 Census. Let us assume that once we selected a system of probabilities we used the optimum numbers of blocks and the optimum sampling ratios appropriate to these probabilities, that is, the optimum values of k and m. For each of several cost functions the following Table 1 shows the sampling variances of each system, rela-

TABLE 1

Unit costs			Average cost per primary unit of listing and related operations ($C_2\Sigma P_i N_i$)			Variances relative to equal probability		
C_1	C_2	C_3	Equal probability	Probability proportionate to square root of size	Probability proportionate to size	Equal probability	Probability proportionate to square root of size	Probability proportionate to size
5	.10	1	13.49	21.15	27.63	100	92	104
5	.05	1	6.75	10.58	13.82	100	88	97
5	.02	1	2.70	4.23	5.53	100	83	87
5	0	1	0	0	0	100	75	73
2	.10	1	13.49	21.15	27.63	100	96	111
2	.05	1	6.75	10.58	13.82	100	93	106
2	.02	1	2.70	4.23	5.53	100	90	97
2	0	1	0	0	0	100	79	77
1	.10	1	13.49	21.15	27.63	100	97	114
1	.05	1	6.75	10.58	13.82	100	96	110
1	.02	1	2.70	4.23	5.53	100	93	103
1	0	1	0	0	0	100	82	81
0	.10	1	13.49	21.15	27.63	100	99	117
0	.05	1	6.75	10.58	13.82	100	99	115
0	.02	1	2.70	4.23	5.53	100	99	113

tive to the variance of sampling with equal probability. It also shows values of $C_2\Sigma P_i N_i$ for comparison with C_1.

Some of the costs given in the table do not have unreasonable relationships in terms of the situations encountered in practice in various types of jobs. The comparisons are not affected by the absolute magnitudes of the costs but only by their relative magnitudes. The results are consistent with the rough rules of thumb given above. It is worth noting that in each of the above instances probability proportionate to the square root of the size yields a comparatively low variance.

5. Sampling with or without replacement. In this paper the sampling with varying probabilities was assumed to be carried out with replacement which ordinarily would not be advisable in practice. When sampling is done without replacement the optimum probabilities and their approximations will be about the same as for sampling with replacement in at least those instances where the proportion of the population in the sample is small. Further investigation is needed for large sampling rates.

6. Conclusion. In summary, it is not essential and may not be desirable to give each element in the population (or stratum) the same chance of being drawn in order to avoid bias or to have a consistent estimate. Estimate (1) is a consistent estimate no matter what probabilities of selection are assigned to these units. The use of variable probabilities of selection is another device to be added to those already in the literature, such as stratification and efficient methods of estimation, which make it possible to achieve the objectives of a sample survey at reduced costs. Reference [2] gives another illustration of reductions in sampling variance achieved through the use of varying probabilities in accordance with the rules suggested above for approximating the optimum probabilities.

REFERENCES

[1] JERZY NEYMAN, "On the two different aspects of the representative method of purposive selection," *Roy. Stat. Soc. Jour.*, New Series, Vol. 97 (1934), pp. 558–606.

[2] MORRIS H. HANSEN AND WILLIAM N. HURWITZ, "On the theory of sampling from finite populations," *Annals of Math. Stat.*, Vol. 14 (1943), pp. 333–362