A GENERALIZATION OF SAMPLING WITHOUT REPLACEMENT FROM A FINITE UNIVERSE*

D. G. HORVITZ† AND D. J. THOMPSON
Iowa State College

This paper presents a general technique for the treatment of samples drawn without replacement from finite universes when unequal selection probabilities are used. Two sampling schemes are discussed in connection with the problem of determining optimum selection probabilities according to the information available in a supplementary variable. Admittedly, these two schemes have limited application. They should prove useful, however, for the first stage of sampling with multi-stage designs, since both permit unbiased estimation of the sampling variance without resorting to additional assumptions.

INTRODUCTION

WHEN sampling a finite universe in which we can identify the individual elements, we are free to assign in a completely arbitrary manner the probability of selecting an element on any particular draw. By appropriate assignment of the selection probabilities it is possible to reduce considerably the sampling variances of unbiased sample estimates over those obtained when sampling with equal probabilities throughout.

The possibility of using unequal probabilities for selecting the sample elements from the universe as a means of increasing precision perhaps received its first impetus for applied sampling from Hansen and Hurwitz [2] in 1943. They introduced the selection of primary units (in a subsampling scheme) with probabilities proportionate to some measure of their size and presented the appropriate theory. Their sampling scheme was confined (when sampling without replacement) to samples of one primary unit per stratum, however, the theory not having been extended beyond this point. More recently, Midzuno [6] has generalized the Hansen and Hurwitz approach to sampling a combination of n elements of the universe with probability proportionate to some measure of size of the combination. Madow [5] has made some contributions to the theory of the systematic selection of several clusters with probability proportionate to a measure of size.

* Journal Paper No. J 2139 of the Iowa Agricultural Experiment Station, Ames, Iowa, Project 1005. Presented to the Institute of Mathematical Statistics, March 17, 1951.

† Now at the University of Pittsburgh.

Research in the theory of sampling for surveys has been concerned with the development of more efficient sampling systems, the system including both the sample design and the method of estimation. One sampling system is said to be more efficient than another if the variance or mean square error of the estimate with the first system is less than that of the second, provided the cost of obtaining the data and results is the same for both. The development of stratified, multi-stage, multiphase, cluster, systematic, and other sample designs beyond simple or unrestricted random sampling, as well as alternative methods of estimation, have all resulted in increased efficiency in specific circumstances. As indicated above, the appropriate use of variable probabilities for the selection of the sample elements can lead to gains in efficiency over systems using equal probabilities of selection.

It is well known that if samples of size one are drawn with probabilities proportionate to the exact measure of the characteristic under observation, unbiased estimates of means or totals for the population exist which have zero sampling error. Similarly, Midzuno provides an unbiased estimator for his design which has zero sampling error when the samples are drawn with probabilities proportionate to the total measure of the elements in each for the characteristic observed. Since in practical situations the values of the characteristic under study are not known in advance, the problem arises of determining the selection probabilities (from any additional information available) which have optimum properties, i.e. maximize the efficiency. Midzuno [6] and Hansen and Hurwitz [3] have both considered this problem with some success.

A limitation of the Hansen and Hurwitz scheme is that an unbiased estimate of the sampling variance of their estimator cannot be obtained from the sample elements. This difficulty appears to exist in Midzuno's system as well, except in the trivial case of equal probability for each sample combination.

The purpose of this paper is twofold. First, it provides a general method for dealing with sampling without replacement from a finite universe when variable probabilities of selection are used for the elements remaining prior to each draw. An unbiased linear estimator for the population total of the characteristic measured is given, as well as the sampling variance of this estimator. An unbiased estimator for the sampling variance is also given. This is for a one-stage design. An extension of the use of this method for two-stage sampling is presented. Second, it examines and discusses some of the problems arising in the practical application of sampling with variable selection probabili-

ties. In this connection, two sampling schemes, for samples of size two, using unequal selection probabilities are presented. Although general use of these schemes is limited because of the small sample size, they should have wide application for the first stage of sampling with stratified two-stage designs. Either of the two schemes permits an unbiased estimate of the sampling variance to be made from the sample data without resorting to additional assumptions.

SAMPLING WITH ARBITRARY PROBABILITIES OF SELECTION

Let the universe, U, consist of N elements $u_1, u_2, \cdots, u_N$. A sample of size n is to be drawn without replacement using arbitrary probabilities of selection for each draw. We denote the probability of selection associated with the ith element of the universe prior to the first draw by p_{i_1} $(i=1, 2, \cdots, N)$, where

$$p_{i_1} \geqq 0, \qquad \sum_{i=1}^{N} p_{i_1} = 1.$$

This, in a sense, defines a probability distribution (of selection) for the elements of the universe for samples of size one. We are sampling without replacement so that prior to each succeeding draw we must define a new probability distribution for the remaining elements. These may be based on the initial probabilities or, in fact, can be a completely unrelated set. For the mth draw we shall designate the probabilities of selection by p_{i_m} where, as above,

$$p_{i_m} \geqq 0, \qquad \sum_{i} p_{i_m} = 1,$$

but the summation now extends only over the $N-m+1$ remaining elements.[1] We will denote the n sets of selection probabilities by

$$\{p_{i_m}\}, \qquad m = 1, 2, \cdots, n. \tag{1}$$

Knowing the probability distributions used at each draw, it is possible to compute the *a priori* probability that the ith element (i.e. u_i) will be included in a sample of size n. This probability will be designated notationally by $P(u_i)$. It is well known that

$$\sum_{i=1}^{N} P(u_i) = n \tag{2}$$

[1] Actually, sampling without replacement as considered here is the special case of sampling with replacement which arises when the elements once selected have probability zero of being chosen on any succeeding draw.

rather than one since we are not summing probabilities of mutually exclusive events, except for samples of size one.

There are $\binom{N}{n}$ different samples when n elements are drawn without replacement from a finite universe of N elements (assuming that at each stage of the draw all remaining undrawn elements have a probability greater than zero of being selected). Consider now the number of possible samples when the order of draw is taken into account. Since each different sample could occur in $n!$ different orders there are $n!\binom{N}{n}=S$ possible samples, considering order. Denote by $s_n(s=1, \cdots, S)$ the sth such sample of size n. The probability that s_n will be drawn is given by the product of the probabilities of selection of the elements in the sample considering the order of the draw. Thus, if s_n contains the elements $u_i, u_j, \cdots, u_t$ drawn in that order, then

$$\Pr(s_n) = p_{i_1}p_{j_2}\cdots p_{t_n}. \tag{3}$$

The probability, $P(u_i)$, of including element u_i in the sample plays the fundamental role in the theory developed in the following sections. For a sample of size n, $P(u_i)$ reduces to a summation of the probabilities associated with the $n!\binom{N-1}{n-1}=S^{(i)}$ samples that contain u_i. Notationally, we have

$$P(u_i) = \sum_{s}^{S^{(i)}} \Pr\left[s_n{}^{(i)}\right] \tag{4}$$

where we are designating a specific sample of size n which includes u_i by $s_n{}^{(i)}$.

The extension to the *a priori* probabilities of including both the elements u_i and u_j in a sample of size n follows readily. Thus

$$P(u_iu_j) = \sum_{s}^{S^{(ij)}} \Pr\left[s_n{}^{(ij)}\right] \tag{5}$$

since there will be $n!\binom{N-2}{n-2}=S^{(ij)}$ such samples, $s_n{}^{(ij)}$ designating a specific one.

EXPECTED VALUES OF SUMS AND PRODUCT-SUMS

Suppose now that we are to measure a characteristic X for the n elements in the sample. Denote by X_i the value of X assumed by element

u_i. The X_i's are not necessarily all different, of course. The expected value of the sum of the observed values of X in the sample is then

$$E\left(\sum_{i=1}^{n} x_i\right) = \sum_{s=1}^{S} \Pr\,(s_n)\left(\sum_{i=1}^{n} x_i\right)_{s_n}.$$

Factoring the x_i common to the ith element u_i and summing over the population, we have

$$\begin{aligned} E\left(\sum_{i=1}^{n} x_i\right) &= \sum_{i=1}^{N} X_i \sum_{s}^{S^{(i)}} \Pr\,[s^{(i)}{}_n] \\ &= \sum_{i=1}^{N} P(u_i)X_i. \end{aligned}$$

Note that for sample sums, x_i refers to the value of X for the element selected on the ith draw. It follows readily that

$$E\left(\sum_{i=1}^{n} x^q{}_i\right) = \sum_{i=1}^{N} P(u_i)X^q{}_i.$$

The expected value of the sum of cross products x_ix_j, $i \neq j$, is given by

$$\begin{aligned} E\left(\sum_{i\neq j}^{n} x_ix_j\right) &= \sum_{s=1}^{S} \Pr\,(s_n)\left(\sum_{i\neq j}^{n} x_ix_j\right)_{s_n} \\ &= \sum_{i\neq j}^{N} X_iX_j \sum_{s}^{S^{(ij)}} \Pr\,[s^{(ij)}{}_n] \\ &= \sum_{i\neq j}^{N} P(u_iu_j)X_iX_j. \end{aligned}$$

Also, of course,

$$E\left(\sum_{i\neq j}^{n} x^q{}_ix^r{}_j\right) = \sum_{i\neq j}^{N} P(u_iu_j)X^q{}_iX^r{}_j.$$

It is to be noted that the process of taking expected values of sums and product-sums reduces to summing the product of the particular function of the observed values by the appropriate *a priori* probability over the elements of the universe. The extension to triple product-sums and higher should now be clear.

ESTIMATION OF THE POPULATION TOTAL

The question of what to use for the estimation of population characteristics when sampling with arbitrary probabilities of selection at

each draw naturally arises. We will restrict ourselves here to using an unbiased linear estimator of a certain class for the population total of the characteristic X.

Actually a number of subclasses of linear estimators exist when sampling a finite universe without replacement. For example, for estimating (from a sample of size n) the population total of X, i.e.

$$T = \sum_{i=1}^{N} X_i,$$

we could consider using either

$$\hat{T}_1 = \sum_{i=1}^{n} \alpha_i x_i,$$

where α_i $(i=1, \cdots, n)$ is a constant to be used as a weight for the element selected on the ith draw; or

$$\hat{T}_2 = \sum_{i=1}^{n} \beta_i x_i,$$

where β_i $(i=1, \cdots, N)$ is a constant to be used as a weight for the ith element whenever it is selected for the sample; or

$$\hat{T}_3 = \gamma_{s_n}\left(\sum_{i=1}^{n} x_i\right)_{s_n},$$

where γ_{s_n} is a constant to be used as a weight whenever the s_nth sample is selected. It should be noted that the α coefficients are independent of the particular sample that is selected. However, the β and γ coefficients, although known constants for a specified sampling procedure, depend on the particular sample selected.

It is the usual procedure, whenever a linear function of n independent random variables is desired as an estimator of some population parameter, to choose the one which has the smallest variance among those that are unbiased. The resulting estimator is then classed as the best linear unbiased estimator. We have indicated above only three of the possible subclasses of linear estimators of T when sampling a finite universe without replacement. The determination of the unbiased estimator which has minimum variance within each of these subclasses is straightforward. The general solution to the problem of determining the best linear unbiased estimator, however, when sampling a finite universe without replacement and with arbitrary probabilities of selection has not been considered by the authors. We observe here, if

(i) there is only one-stage of sampling,
(ii) the individual elements of the universe can be identified in advance,
(iii) information on any supplementary variables for use in the estimation process is lacking, and
(iv) there is no advance knowledge of the values of the characteristic to be measured,

that a general solution is lacking even in the case of equal probabilities of selection. In connection with this remark, although it can be easily shown, when sampling with equal probabilities of selection for each draw, that the α's, β's, and γ's are all equal to N/n for the best linear unbiased estimators of T for each of the three subclasses, this is certainly not sufficient to claim

$$\hat{T} = \frac{N}{n} \sum_{i=1}^{n} x_i$$

as the "best" among all possible linear unbiased estimators of T.

We[2] will restrict ourselves here to the subclass of linear estimators for the population total of X given by $\hat{T}_2$. In order that $\hat{T}_2$ be unbiased we must have

$$E(\hat{T}_2) = T$$

and, hence,

$$\sum_{i=1}^{N} P(u_i)\beta_i X_i = \sum_{i=1}^{N} X_i.$$

In order for this equality to hold whatever be the values of the unknown X's, we must have

$$P(u_i)\beta_i = 1$$

for all i. Therefore,

$$\hat{T} = \sum_{i=1}^{n} \frac{x_i}{P(u_i)} \tag{6}$$

is the only unbiased linear estimator possible in the subclass under consideration and hence is "best" for that subclass. Note that if

$$P(u_i) = \frac{nX_i}{T}, \tag{7}$$

$\hat{T}$ will have zero variance and the sampling will be optimum.

[2] Midzuno uses an estimator belonging to the subclass specified by $\hat{T}_1$.

Using the results obtained in the section on expected values the variance of $\hat{T}$, say $V(\hat{T})$, follows readily. Thus,

$$V(\hat{T}) = E(\hat{T} - T)^2$$

$$= \sum_{i=1}^{N} \frac{X^2_i}{P(u_i)} + \sum_{i \neq j}^{N} \frac{P(u_iu_j)}{P(u_i)P(u_j)} X_iX_j - T^2 \tag{8}$$

$$= \sum_{i=1}^{N} X^2_i \frac{1 - P(u_i)}{P(u_i)} + \sum_{i \neq j}^{N} X_iX_j \frac{P(u_iu_j) - P(u_i)P(u_j)}{P(u_i)P(u_j)}. \tag{9}$$

This formula applies only when every element has a positive probability of inclusion in the sample, however (i.e. $P(u_i) > 0$ for all i).

An unbiased estimator of the variance of $\hat{T}$ in the general sampling procedure is also readily obtainable, provided n is greater than one. Thus,

$$\hat{V}(\hat{T}) = \sum_{i=1}^{n} x^2_i \frac{1 - P(u_i)}{P^2(u_i)} + \sum_{i \neq j}^{n} x_ix_j \frac{P(u_iu_j) - P(u_i)P(u_j)}{P(u_iu_j)P(u_i)P(u_j)}. \tag{10}$$

Again, this formula is restricted to those sampling schemes which yield positive probabilities of inclusion for every element and every pair of elements (i.e. both $P(u_i)$ and $P(u_iu_j)$ greater than zero for all i and j). Alternatively, we may write (10) in the form

$$\hat{V}(\hat{T}) = \hat{T}^2 - \sum_{i=1}^{n} \frac{x^2_i}{P(u_i)} - \sum_{i \neq j}^{n} \frac{x_ix_j}{P(u_iu_j)}. \tag{11}$$

If an unbiased estimate of the population mean is desired, it is sufficient to divide the unbiased estimator of the population total, (6), by N. The sampling variance of this estimator is the same as (8) except for an additional factor of $1/N^2$.

APPLICATION TO KNOWN SAMPLING DESIGNS

The general nature of this approach to sampling a finite universe without replacement will be illustrated by considering the estimator and its sampling variance, as derived in the preceding section, for simple random, systematic, and stratified random sampling procedures.

With simple random sampling or equal probabilities of selection for the elements remaining prior to each draw, we have

$$P(u_i) = \frac{n}{N} \qquad (i = 1, 2, \cdots, N)$$

$$P(u_iu_j) = \frac{n(n-1)}{N(N-1)} \qquad (i, j = 1, 2, \cdots, N, i \neq j).$$

Substitution of these inclusion probabilities in formulas (6) and (8) yields the estimator

$$\hat{T} = \frac{N}{n} \sum_{i=1}^{n} x_i$$

with sampling variance

$$V(\hat{T}) = \frac{N(N-n)}{n(N-1)} \sum_{i=1}^{N} \left(X_i - \frac{T}{N} \right)^2 .$$

In addition, the estimator for the variance of $\hat{T}$ in the general case, formula (10), reduces to

$$\hat{V}(\hat{T}) = \frac{N(N-n)}{n(n-1)} \sum_{i=1}^{n} \left(x_i - \frac{\hat{T}}{N} \right)^2 ,$$

These derived expressions agree with the formulas usually prescribed for the respective quantities when the sample is selected at random without replacement.

To illustrate the application of the general results to systematic samples, we consider the simplified case of a universe of $N=kn$ elements. A systematic sample is obtained by selecting every kth element following the choice of a random starting point among the elements numbered 1 through k. The measured value, of the characteristic of interest, associated with the jth element in the ith possible sample is denoted by X_{ij}.

It follows readily that

$$P(u_{ij}) = \frac{1}{k}$$

for all i and j,

$$P(u_{ij}u_{i'j'}) = \frac{1}{k}$$

for $i=i'$, $j \neq j'$, and

$$P(u_{ij}u_{i'j'}) = 0$$

for all other pairs of elements. Formula (6) again yields the usual estimator

$$\hat{T} = k \sum_{j=1}^{n} X_{ij} = \frac{N}{n} \sum_{j=1}^{n} X_{ij}$$

for the population total, the subscript i denoting the sample chosen. The sampling variance of $\widehat{T}$, as given by (8), namely

$$V(\widehat{T}) = k\sum_{i=1}^{k}\sum_{j=1}^{n} X^2_{ij} + k\sum_{i=1}^{k}\sum_{j\neq j'}^{n} X_{ij}X_{ij'} - T^2,$$

is also equivalent to the usual formula. This is most easily seen by expanding the particular form

$$V(\widehat{T}) = \frac{N^2}{k}\sum_{i=1}^{k}\left(\mu_i - \frac{T}{N}\right)^2,$$

where

$$\mu_i = \frac{1}{n}\sum_{j=1}^{n} X_{ij},$$

of the usual variance formula for systematic samples. (See, for example, L. H. Madow [4]). Since certain pairs of elements have no chance of being included together in a sample with this systematic design, formula (10) cannot be used to estimate the sampling variance of $\widehat{T}$ from the sample data.

The variance formula (8), left in its expanded form, provides an interesting method for examining the conditions under which one sampling system will be more efficient than another disregarding costs. To examine the efficiency of a systematic sample versus a random sample we note that the respective variance formulas differ only in the middle term of (8). Thus a systematic sample will be more efficient (the particular estimator chosen will have a smaller variance for systematic samples than for random samples) if

$$\sum_{i=1}^{k}\sum_{j\neq j'}^{n} X_{ij}X_{ij'} < \frac{n-1}{N-1}\sum_{i\neq i'}^{k}\sum_{j\neq j'}^{n} X_{ij}X_{i'j'}.$$

Following some algebraic manipulation, this condition reduces to

$$\frac{n}{k-1}\sum_{i=1}^{k}\left(\mu_i - \frac{T}{N}\right)^2 < \sum_{i=1}^{k}\frac{\sigma^2_i}{k}, \tag{12}$$

where μ_i is the mean of the ith sample as defined above and

$$\sigma^2_i = \frac{\sum_{j=1}^{n}(X_{ij} - \mu_i)^2}{n-1}.$$

Essentially, then, a systematic sample will be more efficient than a random sample if the variation between the possible systematic samples is less than the average variation within the possible samples. It should be noted in particular that (12) is exactly equivalent to the well-known condition of an intraclass (intrasample, in this case) correlation coefficient less than $-1/(N-1)$ for an efficient systematic sample relative to a random sample. Further examination of (12) leads rapidly to several of the other known conditions for gains with a systematic sample.

When the universe elements have been classified into K strata and a random sample selected from each stratum, substitution of the inclusion probabilities for individual elements and pair of elements in (6) and (8) again yields the usual formulas for the appropriate linear unbiased estimator and its sampling variance. If u_{ij} denotes the jth element in the ith stratum, N_i the number of elements in the ith stratum, and n_i the number of elements selected for the sample from that stratum, the inclusion probabilities are

$$P(u_{ij}) = \frac{n_i}{N_i}$$

for all j

$$P(u_{ij}u_{i'j'}) = \frac{n_i(n_i - 1)}{N_i(N_i - 1)}$$

for $i=i'$, $j\neq j'$, and

$$P(u_{ij}u_{i'j'}) = P(u_{ij})P(u_{i'j'}) = \frac{n_i n_{i'}}{N_i N_{i'}}$$

for all j and j', $i\neq i'$.

Whereas the above results point out that for the schemes considered the possible estimators $\hat{T}_1$ and $\hat{T}_2$ are equivalent, this will not be true in general. It should be noted that for each of these schemes the probability of including a particular element in a sample is the same either for all the elements of the universe or for all the elements of the same sub-universe.

EXTENSION TO A TWO-STAGE SAMPLING DESIGN

The extension of the use of arbitrary probabilities of selection for each draw to designs involving more than one stage has been examined for a special case only. The universe now consists of K primary sampling

units with the ith such unit containing N_i secondary or subsampling units. Let X_{ij} be the value of some characteristic X of the jth subsampling unit of the ith primary sampling unit. The population total

$$T = \sum_{i=1}^{K} \sum_{j=1}^{N_i} X_{ij} \tag{13}$$

is to be estimated from a sample of k primary units, n_i subsampling units to be drawn from the ith primary unit if it is in the sample. The primary units are drawn without replacement using arbitrary probabilities of selection for each draw. An over-all sampling rate, say t, is specified in advance and the n_i determined from the relation

$$n_i = \frac{tN_i}{P(u_i)}, \qquad (i = 1, 2, \cdots, k), \tag{14}$$

where $P(u_i)$ now denotes the *a priori* probability that the ith primary unit will be included in a sample of k such units. The subsampling units are to be drawn without replacement with equal probabilities of selection for those remaining prior to each draw, i.e. at random. This sampling procedure is entirely analogous to that specified by Hansen and Hurwitz [2] (ignoring area substratification) when a single primary unit is drawn with probability proportionate to its estimated size.

One difficulty that arises with this scheme concerns the n_i as determined by (15). In most practical applications this relation will not yield integral subsampling sizes. We will neglect the bias introduced by choosing the closest integral value for n_i in what follows. It should also be noted that the N_i need not be known in advance of the primary unit selection stage of the draw.

Since every subsampling unit will have the same chance of being included in the sample, it follows that

$$\hat{T} = \frac{1}{t} \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij} \tag{15}$$

will provide an unbiased estimate of T. The variance of this estimator (provided $P(u_i) > 0$) follows readily from the previous results. Thus

$$V(\hat{T}) = \sum_{i=1}^{K} T^2{}_i \frac{1 - P(u_i)}{P(u_i)} + \sum_{i \neq j}^{K} T_i T_j \frac{P(u_i u_j) - P(u_i)P(u_j)}{P(u_i)P(u_j)} + \sum_{i=1}^{K} \frac{N_i(N_i - n_i)}{n_i P(u_i)} \sigma^2{}_i, \tag{16}$$

where

$$T_i = \sum_{j=1}^{N_i} X_{ij} = N_i \mu_i$$

and

$$\sigma^2{}_i = \sum_{j=1}^{N_i} \frac{(X_{ij} - \mu_i)^2}{(N_i - 1)} .$$

The first two terms of the right member of (16) make up the usual between primary unit component of variance, the last term being the within component.

An estimate of this sampling variance may be computed from the elements in the sample, the estimator provided here having the property of unbiasedness. Thus (when the $P(u_i)$ and $P(u_i u_j)$ are all greater than zero)

$$(17) \quad \hat{V}(\hat{T}) = \hat{T}^2 - \sum_{i=1}^{k} \frac{\hat{T}_i{}^2}{P(u_i)} - \sum_{i \neq j}^{k} \frac{\hat{T}_i \hat{T}_j}{P(u_i u_j)} + \frac{1}{t} \sum_{i=1}^{k} (N_i - n_i) s^2{}_i,$$

where

$$\hat{T}_i = N_i \bar{x}_i = \frac{N_i}{n_i} \sum_{j=1}^{n_i} x_{ij}$$

and

$$s_i{}^2 = \sum_{j=1}^{n_i} \frac{(x_{ij} - \bar{x}_i)^2}{(n_i - 1)} .$$

An unbiased estimate of the between component of variance may be obtained by using $\hat{V}(\hat{T})$ in conjunction with an unbiased estimate of the within component, the latter being given by the quantity

$$\frac{1}{t} \sum_{i=1}^{k} \frac{N_i - n_i}{P(u_i)} s^2{}_i.$$

SOME ASPECTS OF THE PRACTICAL APPLICATION OF THE THEORY

The remaining sections of this paper will be devoted to examining some aspects of the problems arising in connection with attempts to utilize the preceding theory in practical applications. To simplify the exposition, attention will be confined to a one stage sampling scheme; however, various extensions of the results to multi-stage stratified designs are evident.

In the estimating functions (6) and (15) as well as the corresponding

expressions for the variance and sample estimates of the variance, it will be noticed that it is the quantities $P(u_i)$ and $P(u_iu_j)$ that can be controlled by the sampler. Assume that a variable Y reasonably correlated with X is known for each element of the universe, and that the sampler wishes to utilize the information in Y in assigning the selection probabilities (1), so that the resulting $P(u_i)$ and $P(u_iu_j)$ will lead to a reduction in variance. The three main problems that arise in this connection consist of

(i) determining the quantities $P(u_iu_j)$ that will minimize the variance (9) (specifying the $P(u_iu_j)$ determines the $P(u_i)$ since $\sum_j P(u_iu_j) = (n-1)P(u_i)$, as may be easily verified),

(ii) defining the sets (1) to achieve the $P(u_iu_j)$ thus determined, and

(iii) investigating the conditions required on the relationship between Y and X to obtain gains in efficiency over sampling systems employing the information in Y in alternative ways.

These three problems are not independent and a general solution has not been reached by the authors. Some progress has been made in particular cases, however, and it is hoped that a discussion of these cases will provoke interest and stimulate others to investigate sampling systems of the type here considered.

Considering first the problem of assigning the $P(u_iu_j)$, as a first approximation to an "optimum" assignment, we may require only that

$$\frac{1}{n-1}\sum_{j}^{N} P(u_iu_j) = P(u_i) = nY_i \Big/ \sum_{i=1}^{N} Y_i. \tag{18}$$

If the X_i are approximately proportional to the Y_i, this assignment may be expected to lead to an estimator with small variance, as may be seen by assuming strict proportionality between X and Y and noting that (6) is then identically T. From examination of (9), it appears that the assignment of the $P(u_iu_j)$ (in terms of the Y_i) that leads to minimum variance depends upon the joint distribution of X and Y. This complicates the problem. In the examples discussed in a later section, however, it will be demonstrated that a substantial reduction in variance can be achieved by an assignment of the type indicated in (18).

The problem of determining the sets of selection probabilities (1) that will yield preassigned "optimum" values of $P(u_iu_j)$, or $P(u_i)$ as in (18) above, can be illustrated by a sample example. Suppose a sample of size 2 is to be drawn without replacement from the universe of 6 elements given in Table 1 in such a manner that $P(u_i) = 2Y_i/\sum_{i=1}^{6} Y_i$. In the notation of the previous sections, n sets $\{p_{i_m}\}$, $(m = 1, 2, \cdots, n)$, each satisfying

$$0 \leqq p_{i_m} \leqq 1$$

$$\sum_{i=1}^{N-m+1} p_{i_m} = 1,$$

must be defined, so that when the n draws have been performed according to these sets, the probability that u_i will be included in the sample will be the assigned probability $P(u_i)$. The notation adopted for the sets of selection probabilities is not entirely satisfactory since it does not indicate the dependence of the set used for selecting the mth element in the sample on the results of the previous $m-1$ draws. In Table 1, columns 5 and 6, this dependence is explicitly indicated by using the notation commonly employed for a conditional probability. Thus,

$$\{p_{3_2} | \ u_1 \in s\}$$

is the probability assigned to the selection of u_3 on the second draw, given that u_1 has been obtained on the first draw.

TABLE 1

(1) u_i	(2) Y_i	(3) $P(u_i)$	(4) $\{p_{i_1}\}$	(5) $\{p_{i_2} \mid u_1 \in s\}$	(6) $\{p_{i_2} \mid u_2 \in s\}$
1	32	.64	.6	0	.1
2	23	.46	.4	.1	0
3	17	.34	0	.3	.4
4	13	.26	0	.3	.2
5	10	.20	0	.2	.2
6	5	.10	0	.1	.1
	100	2.00	1.0	1.0	1.0

It may be easily verified that the selection probabilities defined in columns 4, 5, and 6 of Table 1 do achieve the inclusion probabilities in column 3. The solution indicated is one of an infinity of solutions and was chosen primarily for its simplicity. The authors are not aware of general methods for examining the consistency of systems of equations of the type used in obtaining columns 4, 5, and 6 or of finding simultaneous positive solutions when they exist. For the solution given $P(u_iu_j)=0$ for $i \neq j=3$, 4, 5, 6, and formula (10) is therefore not applicable.

It should be noted that the $P(u_i)$ are probabilities satisfying

$$0 < P(u_i) < 1$$

$$\sum_{i=1}^{N} P(u_i) = n,$$

so that if the "measures of size" are such that for any element, say u_t the quantity $nY_t/\sum_{i=1}^{N} Y_i$ is greater than unity, no method of drawing the sample exists which will give $P(u_t)/\sum_{i=1}^{N} Y_i$. This situation can be obviated in various ways (stratification, subdivision of the elements of the universe, etc.) so that it need cause no difficulty.

A secondary consideration of practical importance in defining the sets $\{p_{i_m}\}$ by the general method indicated above is that particular solutions may not facilitate the computation of the quantities $P(u_iu_j)$ required for estimating the variance (10). To calculate this expression the quantities $P(u_iu_j)$ must be determined for the $\binom{n}{2}$ combinations of the sample elements. For n of any considerable size the direct calculation of the $P(u_iu_j)$ by summing the probabilities associated with the samples containing u_i and u_i is impractical. It would thus seem advisable to restrict further the choice of selection schemes to those schemes that permit ready calculation of $P(u_iu_j)$.

A selection scheme that obviates the problem of explicitly defining the set (1) yet satisfies (18) is mentioned by Goodman and Kish [1]. The N universe elements are listed in a random order and their measures of size are cumulated. A systematic selection of n elements from a random start is then made on the cumulation so that $P(u_t) = nY_t/\sum_{i=1}^{N} Y_i$. This selection is easily performed, but there does not appear to be any simple way to determine the $P(u_iu_j)$.

Sampling Scheme 1

A method of defining the set $\{p_{i_m}\}$ that yields an exact solution under certain conditions can be developed in the following way. Consider drawing a sample of n elements from a universe of N elements without replacement, where the first element is selected according to the set p_{i_1}, $(i = 1, 2, \cdots, N)$ $\sum_{i=1}^{N} p_{i_1} = 1$, $p_{i_1} > 0$. At the second and all remaining $(n-1)$ stages of the draw equal probabilities are assigned to the elements remaining, i.e. the set p_{i_2} consists of $N-1$ equal elements $1/(N-1)$, the set p_{i_2}, $N-2$ equal elements $1/(N-2)$, etc.[3] By simple combinatorial analysis we find that

[3] Midzuno suggested using this scheme for drawing the sample in connection with his sampling system on a recent visit to the Statistical Laboratory, Iowa State College. It may be mentioned that with this method of drawing, each of the possible $\binom{N}{n}$ different samples has a probability of being selected proportional to the total of the measures of size for the elements in the combination, which is desirable in his system.

$$P(u_i) = \frac{p_i}{(N-1)\cdots(N-n+1)}\binom{N-1}{n-1}(n-1)! + \sum_{j\neq i}^{N} \frac{p_j}{(N-1)\cdots(N-n+1)}\binom{N-1}{n-1}(n-1)!\frac{n-1}{N-1},$$

which upon simplification becomes

$$P(u_i) = \frac{N-n}{N-1}p_i + \frac{n-1}{N-1}, \qquad (i = 1, 2, \cdots, N). \tag{19}$$

Similarly, it may be shown that for this case

$$P(u_iu_j) = \frac{n-1}{N-1}\left[\frac{N-n}{N-2}(p_i + p_j) + \frac{n-2}{N-2}\right], \qquad (i \neq j: i, j = 1, 2, \cdots, N). \tag{20}$$

Solving (19) for p_i in terms of the $P(u_i)$,

$$p_i = \frac{N-1}{N-n}P(u_i) - \frac{n-1}{N-n}, \qquad (i = 1, 2, \cdots, N). \tag{21}$$

It should be noted that the p_i are subject to the two conditions,

(i) $p_i \geqq 0$, for all i,

(ii) $\sum_{i=1}^{N} p_i = 1.$

In a particular case when the $P(u_i)$ have been assigned such that one or more of the inequalities $P(u_i) < (n-1)/(N-1)$ are satisfied, the corresponding solutions of (21) will be negative. This restriction is rather severe, in general, and limits the usefulness of this method. For a small sample size, however, the method may be satisfactory, as will be demonstrated in the example that follows, and approximate solutions based upon it for larger sample sizes can be obtained easily. For the case when all solutions of (21) are positive with $P(u_i) = nY_i/\sum_{i=1}^{N} Y_i$, the first element would be drawn according to the set of solutions of (21) and equal probabilities would be used for the remaining draws.

Sampling Scheme 2

There are undoubtedly many other ways of defining sets of selection probabilities such that condition (18) will be satisfied approximately. To be practical, the necessary computations should remain simple, however. At the same time, although such schemes yield only an ap-

proximate solution to the preassigned "optimum" values of the $P(u_i)$, they should include exact expressions for the actual values of these quantities and for the $P(u_iu_j)$ as well. The scheme proposed here satisfies these requirements but is restricted to samples of size 2. It may lead to a better approximation to the desired $P(u_i)$ when the condition for an exact solution with Sampling Scheme 1 is not satisfied.

The particular scheme suggested here requires the prior determination only of the set p_{i_1}, $(i=1, 2, \cdots, N)$; that is the set of selection probabilities to be used on the first draw. The set to be used for the second draw depends on the first element selected. Thus, if element u_j is selected on the first draw, then

$$p_{i_2} = \frac{p_{i_1}}{1 - p_{i_1}}, \qquad \text{for } i \neq j$$

$$p_{i_2} = 0, \qquad \text{for } i = j$$

defines the set of selection probabilities p_{i_2}, $(i=1, 2, \cdots, N)$. In practice the second element may be selected after adjusting the selection probabilities used on the first draw or the same set may be used throughout, the selection process continuing until two different elements have been chosen. Since only the set of selection probabilities for the first draw needs to be determined in advance, the subscript indicating the draw will be dropped.

If a sample of size 2 is to be drawn using one set of selection probabilities and with replacement, then the probability that element u_i will be selected only once in the two draws is $2p_i(1-p_i)$. If the conditions are such that sampling without replacement is not much different than sampling with replacement, this probability will be approximately equal to $P(u_i)$, the inclusion probability for sampling without replacement. This suggests that a set of selection probabilities which will lead to an approximation of the desired $P(u_i)$ with the prescribed sampling procedure may be determined from the solution to the system of equations

$$(22) \qquad 2p^2_i - 2p_i + 2Y_i \Big/ \sum_{i=1}^{N} Y_i = 0, \qquad (i = 1, 2, \cdots, N)$$

where, of course, the common coefficient 2 may be cancelled. Again, the solution must be such that

$$p_i \geqq 0, \qquad \text{for all}$$

and

$$\sum_{i=1}^{N} p_i = 1.$$

In practice, a satisfactory solution to this system may be obtained by solving each of the N quadratic equations separately, taking the smaller of the two roots. Since the selection probabilities must sum to unity, a simple adjustment is then made by dividing each of the p_i obtained in this manner by their sum. This method was used in the example which follows where it proved quite adequate. It should be noted that this procedure for solving the system (22) breaks down if any of the desired $P(u_i) \geqq \frac{1}{2}$, since the solutions for the particular quadratic equations will then be imaginary.

The accuracy of whole method depends on the original assumption that a formula based on sampling with replacement will be adequate even though the sampling is without replacement. Although no detailed investigation has been made on this point, it appears that for N at least as large as 10 and the desired $P(u_i)$ not dominated entirely by one or two elements reasonable success will result.

One additional point is necessary. Whatever the set of selection probabilities adopted with this sampling procedure, the exact formulas for the $P(u_i)$ and $P(u_iu_j)$ are

$$P(u_i) = p_i + p_i \sum_{j \neq i}^{N} \frac{p_j}{1 - p_j}, \tag{23}$$

$$P(u_iu_j) = p_ip_j\left(\frac{1}{1 - p_i} + \frac{1}{1 - p_j}\right). \tag{24}$$

Example:

The universe to be investigated consists of 20 blocks in Ames, Iowa, the data being given in columns 1 to 3 in Table 2. These data are taken from a survey conducted by the Statistical Laboratory of Iowa State College. The estimated number of households (column (3)) was obtained by a team of observers who drove through the portion of the city of Ames represented by these 20 blocks and made rapid eye-estimates of the number of households on each block.

We shall consider drawing a sample of 2 blocks with probability proportionate to this measure of size (eye-estimated households) according to the two selection schemes previously developed. The exact values which the selection schemes are designed to achieve are listed in column (4) of Table 2, and the corresponding results of the two proposed selection schemes are shown in columns (6) and (8). Columns (5)

and (7) give the selection probabilities for the first draw with scheme 1 and all draws with the second scheme respectively. When the sample has been drawn according to either of these schemes, the $P(u_iu_j)$ required in (10) to estimate the variance of $\hat{T}$ are easily computed from (20) or (24) respectively.

TABLE 2

Block	Number of households on ith block	Eye-estimated number of households on ith block		Selection Scheme 1		Selection Scheme 2	
(1) u_i	(2) X_i	(3) Y_i	(4) $2Y_i/\sum Y_i$	(5) p_i*	(6) $P(u_i)$	(7) p_i	(8) $P(u_i)$
1	19	18	.091	.040	.090	.045	.091
2	9	9	.046	.003	.055	.022	.045
3	17	14	.071	.019	.070	.035	.070
4	14	12	.061	.008	.060	.029	.060
5	21	24	.122	.072	.121	.061	122
6	22	25	.127	.077	.126	.064	.127
7	27	23	.117	.066	.116	.058	.117
8	35	24	.122	.072	.121	.061	.122
9	20	17	.086	.035	.085	.042	.086
10	15	14	.071	.019	.070	.035	.070
11	18	18	.091	.040	.090	.045	.091
12	37	40	.203	.157	.201	.108	.209
13	12	12	.061	.008	.060	.029	.060
14	47	30	.152	.104	.151	.078	.154
15	27	27	.137	.088	.136	.069	.138
16	25	26	.132	.082	.131	.067	.133
17	25	21	.107	.056	.106	.053	.106
18	13	9	.046	.003	.055	.022	.045
19	19	19	.096	.045	.096	.048	.096
20	12	12	.061	.008	.060	.029	.060
Totals	434	394	2.0	1.0	2.0	1.0	2.0

* The two blocks (2 and 18) with the smallest eye-estimated size were arbitrarily assigned an eye-estimated value of 11 households to satisfy the condition $2Y_i/\sum_{i=1}^{N}Y_i > 1/(N-1)$ in obtaining these results.

COMPARISONS OF EFFICIENCY (IGNORING COST)

A primary purpose of this paper has been to extend the theory of finite sampling with unequal probabilities to permit unbiased estimation of the sampling error without resorting to additional assumptions.

As mentioned previously, however, it is of interest to compare the relative effectiveness of using the supplementary quantitative variable Y in alternative ways, e.g. stratification, other estimators, etc. No simple expressions for the relative efficiency of sampling systems of the type herein developed to alternative systems have as yet been obtained. As indicative of the type of results obtainable when the relationship between Y and X is one of approximate proportionality, the empirical comparisons with a number of alternatives included in Table 3 are of interest.

TABLE 3

(1) Sampling system	(2) Method of selection	(3) Method of estimation	(4) Variance of the estimator	(5) Relative efficiency (%)
1.	Unrestricted random.	$N\bar{x}$	16,219	100
2.	Unrestricted random.	$(\bar{x}/\bar{y})\sum_{i=1}^{N} Y_i$	3,280§	497
3.	Stratified random; one element from each of 2 strata with equal probability.	$N\bar{x}$	7,873	206
4.	Stratified; one element with probability proportionate to measure of size from each of 2 strata.	$\sum x_i/P_i$*	3,934	412
5.	Systematic sample; every kth from random start.	$N\bar{x}$	10,224	159
6.	Midzuno; pair of elements with probability proportionate to the sum of the measures for the pair.	$(\bar{x}/\bar{y})\sum_{i=1}^{N} Y_i$	3,579	453
7.	Scheme 1	$\sum x_i/P_i$†	3,095	524
8.	Scheme 2	$\sum x_i/P_i$‡	3,075	527

* P_i proportional to Y_i.
† P_i as given in column 6, Table 2.
‡ P_i as given in column 8, Table 2.
§ The bias for this estimator equals 1.17 which has been neglected here.

The quantity under estimate is the total number of households on the 20 blocks in Table 2. A sample of size 2 is considered. For this small universe and sample size it was feasible to compute the exact variance of the estimator employed in each sampling system directly from the definition, so that, for example, the mean square error of the so called "ratio estimate" (line 2, Table 3) is not the usual approximation.

For the sampling systems 3 and 4 the blocks are ranked according to

the measure of size Y, and the ten largest blocks were taken as stratum 1 with the remaining 10 in stratum 2. The systematic sample is also to be considered as drawn from the blocks after ranking from large to small. It is of interest to note that the "ratio estimate" used in sampling system 2 is identical with the estimator in Midzuno's system (sampling system 6), and that it is an unbiased estimator for his method of selection.

The data presented in Table 3 are, of course, far from conclusive, but they do indicate that substantial reductions in variance can be obtained through the use of unequal probabilities without forfeiting an unbiased estimate of the sampling variance.

It is the opinion of the authors that the techniques suggested by this paper may be of greatest utility in specialized enquiries where the characteristics under measurement are few and related, or where selection with unequal probability arises naturally. The estimator (6) from a computational point of view is at a serious disadvantage when compared with self-weighting estimators. The estimated variance (9) has similar disadvantages when compared with designs that permit estimation of error by the use of an analysis of variance or other simple technique. When an unbiased estimator of high precision and an unbiased sample estimate of its variance are required, however, the sampling system employing unequal probabilities, with the selection of two or more units at each stage of sampling, may be particularly appropriate. This is particularly true when the universe (at any stage) is small and the alternative use of the information in Y is a ratio-estimator (based on Y with equal probability selection) with its possible bias and unknown error.

A modified formulation of the theory in connection with the technique suggested by Hartley and Politz-Simmons [7] for the problem of the "not-at-homes" in an interview survey is possible along the lines suggested by this paper. It also appears that the technique of control beyond stratification suggested by Goodman and Kish [1] is closely related to the problem of the optimal assignment of the $P(u_iu_j)$.

Finally, the possibility of employing the sampling systems considered here in connection with "point sampling" is of considerable interest. By "point sampling" we have reference to the selection of farms in an agricultural survey by locating points at random on a map of the area to be surveyed, and including as sample elements the farms within whose boundaries the points happen to fall. (See, for example, F. Yates [8].) It is clear that the size of the farms will be related to their probability of inclusion in the sample, and that unbiased estimates are

possible from samples drawn in this manner. The details for this case and similar cases in other types of investigations remain to be work out.

The authors are grateful to Dr. R. J. Jessen for kindling our interest in this problem and to Professor O. Kempthorne and Dr. P. C. Tang for their helpful criticisms and advice in the preparation of this paper.

REFERENCES

[1] Goodman, Roe, and Kish, Leslie, "Control beyond stratification; a technique in probability sampling," *Journal of the American Statistical Association*, 45 (1950), 350–72.

[2] Hansen, Morris H., and Hurwitz, William N., "On the theory of sampling from finite populations," *Annals of Mathematical Statistics*, 14 (1943), 333–62.

[3] ———, "On the determination of optimum probabilities in sampling," *Annals of Mathematical Statistics*, 20 (1949), 426–32.

[4] Madow, Lillian H., "Systematic sampling and its relation to other sampling designs," *Journal of the American Statistical Association*, 41 (1946), 204–17.

[5] Madow, William G., "On the theory of systematic sampling, II," *Annals of Mathematical Statistics*, 20 (1949), 333–54.

[6] Midzuno, Hiroshi, "An outline of the theory of sampling systems," *Annals of the Institute of Statistical Mathematics* (Japan), 1 (1950), 149–56.

[7] Politz, Alfred, and Simmons, Willard, "An attempt to get the 'Not at Homes' into the sample without callbacks," *Journal of the American Statistical Association*, 44 (1949), 9–31.

[8] Yates, Frank, *Sampling Methods for Censuses and Surveys*, London: Charles Griffen and Co., Ltd. (1949), 167–69.

ERRATUM: THE EFFECTIVENESS OF QUALITY CONTROL CHARTS

LEO A. AROIAN AND HOWARD LEVENE

The following corrections should be made in the article published under the above title in this journal (Vol. 45, 1950, pp. 520–529).

1) Page 521: 6th line from top, insert "it" between "that" and "is".

2) Page 524: equation (9), replace "$\prod_{i=1}^{N}$" by "$\prod_{i=1}^{N-1}$" and add "for $i=1, f(N)=\gamma_1$".

3) Page 525: 6th line from top, replace "m_n" by "m_N".