# ON SELECTING SAMPLING UNITS WITH PROBABILITIES PROPORTIONAL TO SIZES AND WITHOUT REPLACEMENT

by

TONG-SANG LIU

B. S., National Taiwan University, 1960

---

A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1967

Approved by

A. M. Feyerherm
Major Professor

TABLE OF CONTENTS

# INTRODUCTION

The purpose of research in sampling surveys is to find and develop more efficient estimates of population characteristics with proper sampling methods. One method of estimation is said to be more efficient than another if the variance or mean square error of an estimate with the first method is less than that of the second, provided the cost of obtaining the data and results are the same for both.

Most survey designs take the selection of n units at random, with equal probabilities and without replacement drawn from a population of N units, as a basic sampling procedure. It sometimes happens that to select units with unequal probabilities will yield a gain in efficiency. For example, such a procedure may be found appropriate when a "measure of size" $x_i$ is known for all the units in the population ($i = 1, 2, \cdots, N$), and these known sizes $x_i$ are correlated with the characteristics $y_i$ for which the population total Y is to be estimated. This criterion was first suggested by Hansen and Hurwitz (1943) who considered a design in which a sample of one unit is drawn with probability proportional to size within each stratum. Horvitz and Thompson (1952) generalized the results to a sample of n units drawn with probability proportional to size and without replacement.

From the general formulas given by Horvitz and Thompson (1952), Yates and Grundy (1953), Des Raj (1956) and Hartley and Rao (1962) derived different sampling procedures in order to get more efficient estimates. There are some limitations, of varying importance, attached to all these methods. Recently, Rao, Hartley and Cochran (1962) introduced a new method attempting to avoid all disadvantages which occurred in the previous methods at the expense of a slight loss in efficiency.

The purpose of this report is to describe and discuss, with the aid of numerical examples, these sampling procedures. The general formulas given by Horvitz and Thompson (1952) are introduced. Following this, the sampling procedures of Yates and Grundy (1953), Des Raj (1956), Hartley and Rao (1962) and Rao, Hartley and Cochran (1962) are introduced successively. The report concludes with some numerical examples.

## GENERAL THEORY

Horvitz and Thompson (1952) give an account of the general theory. Suppose a population consists of N elements $y_1, y_2, \cdots\cdots, y_N$. A sample of size n is to be drawn without replacement using probabilities of selection proportional to measures of size. The probability of selection associated with the $i^{th}$ element of the population prior to the first draw is denoted by $p_{i1}$ $(i = 1, 2, \cdots, N)$, where

$$p_{i1} \geq 0 \quad , \qquad \sum_{i=1}^{N} p_{i1} = 1 \quad .$$

This defines a probability distribution (of selection) for the elements of the population for samples of size one. This is sampling without replacement so that prior to each succeeding draw one must define a new probability distribution for the remaining elements. For the $m^{th}$ draw designate the probabilities of selection by $p_{im}$ where, as above,

$$p_{im} \geq 0 \quad , \qquad \sum_{i=1}^{N-m+1} p_{im} = 1 \quad ,$$

but the summation now extends only over the N - m + 1 elements.

Knowing the probability distributions used at each draw, it is possible to compute the a priori probability that the $i^{th}$ element (i.e., $y_i$) will be included in a sample of size n. This probability will be designated by $\pi_i$ or $p(y_i)$. It is well known that

$$\sum_{i=1}^{N} \pi_i = \sum_{i=1}^{N} p(y_i) = n \quad , \tag{1.1}$$

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 population of N elements. It is assumed that at each stage of the draw all remaining undrawn elements have a probability greater than zero of being selected. When the order of draw is taken into account, there are $n! \cdot \binom{N}{n} = S$ possible samples (since each different sample could occur in n! different orders). Denote $s_n$ ($s = 1, 2, \cdots, S$) the $s^{th}$ 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 $y_i$, $y_j$, $\cdots$, $y_t$ drawn in that order, then

$$P_r(s_n) = p_{i1} \cdot p_{j2} \cdots p_{tn} \tag{1.2}$$

The probability, $\pi_i$ or $p(y_i)$, of including element $y_i$ in the sample plays a fundamental role in the theory of developing the estimators. For a sample of size n, $\pi_i$ reduces to a summation of the probabilities associated with the $n! \cdot \binom{N-1}{n-1} = S^{(i)}$ samples that contain $y_i$. Notationally,

$$\pi_i = \sum_{s}^{S^{(i)}} P_r\left[s_n^{(i)}\right] \tag{1.3}$$

where a specific sample of size n which includes $y_i$ is designated by $s_n^{(i)}$.

The extension to the a priori probabilities of including both the elements $y_i$ and $y_j$ in a sample of size n follows readily; that is,

$$P_{ij} = \sum_{\delta}^{S^{(ij)}} P_r \left[ \delta_n^{(ij)} \right] \tag{1.4}$$

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

Suppose now that one measures a characteristic Y for the n elements in the sample. The expected value of the sum of the observed values of Y in the sample is then

$$\begin{aligned} E\left( \sum_{i=1}^{n} y_i \right) &= \sum_{\delta=1}^{S} P_r(\delta_n) \left( \sum_{i=1}^{n} y_i \right)_{\delta_n} , \\ &= \sum_{i=1}^{N} y_i \sum_{\delta}^{S^{(i)}} P_r \left[ \delta_n^{(i)} \right] , \\ &= \sum_{i=1}^{N} \pi_i y_i \quad . \end{aligned} \tag{1.5}$$

Note that for sample sums, $y_i$ refers to the value for the element selected on the $i^{th}$ draw. It follows readily that

$$E\left( \sum_{i=1}^{n} y_i^{g} \right) = \sum_{i=1}^{N} \pi_i y_i^{g} \tag{1.6}$$

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

$$\begin{aligned} E\left( \sum_{i \neq j}^{n} y_i y_j \right) &= \sum_{\delta=1}^{S} P_r(\delta_n) \left( \sum_{i \neq j}^{n} y_i y_j \right) , \\ &= \sum_{i \neq j}^{N} y_i y_j \sum_{\delta}^{S^{(ij)}} P_r \left[ \delta_n^{(ij)} \right] , \\ &= \sum_{i \neq j}^{N} P_{ij} \cdot y_i y_j \quad . \end{aligned} \tag{1.7}$$

## Estimation of the Population Total

Only unbiased linear estimators for the population total (Y) will be considered. Actually, a number of linear estimators exist. Horvitz and Thompson (1952) restricted themselves to using

$$\hat{Y} = \sum_{i=1}^{n} B_i y_i$$

where n is the size of the sample and each $B_i$ (i = 1, 2, ···, N) is a constant to be used as a weight for the $i^{th}$ element whenever it is selected for the sample. Also, the B coefficients depend on the particular sample selected.

In order that Y be unbiased it must be true that

$$E(\hat{Y}) = Y = \sum_{i=1}^{N} y_i \tag{1.8}$$

and, hence, from equation (1.5)

$$E(\hat{Y}) = E\left(\sum_{i=1}^{n} \beta_i y_i\right) = \sum_{i=1}^{N} \pi_i \beta_i y_i \quad .$$

In order for the equality of equation (1.8) to hold, it is necessary that

$$\pi_i \beta_i = 1 \qquad \text{for all } i.$$

Therefore,

$$\hat{Y} = \sum_{i=1}^{n} y_i / \pi_i \tag{1.9}$$

is the only unbised linear estimator possible for consideration.

Note that if

$$\pi_i = \frac{n y_i}{Y},$$

$\hat{Y}$ will have zero variance.

Variance and the Estimate of Variance for $\hat{Y}$

By definition the variance of $\hat{Y}$ is

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

$$= E\left(\sum_{i=1}^{n} \frac{y_i}{\pi_i} - Y\right)^2,$$

$$= E\left[\left(\sum_{i=1}^{n} \frac{y_i^2}{\pi_i^2} + \sum_{i \neq j}^{n} \frac{y_i y_j}{\pi_i \pi_j}\right) - 2Y\left(\sum_{i=1}^{n} \frac{y_i}{\pi_i}\right) + Y^2\right],$$

$$= E\left(\sum_{i=1}^{n} \frac{y_i^2}{\pi_i^2}\right) + E\left(\sum_{i \neq j}^{n} \frac{y_i y_j}{\pi_i \pi_j}\right) - 2Y \cdot E\left(\sum_{i=1}^{n} \frac{y_i}{\pi_i}\right) + Y^2.$$

Using the results obtained in equation (1.5), (1.6), and (1.7);

$$V(\hat{Y}) = \sum_{i=1}^{N} \frac{y_i^2}{\pi_i} + \sum_{i \neq j}^{N} \frac{P_{ij}}{\pi_i \pi_j} y_i y_j - Y^2. \qquad (1.10)$$

This formula applies only when $\pi_i > 0$ for all i.

An unbiased estimator of the variance of $\hat{Y}$ is also obtainable, provided n is greater than one. Thus,

$$v(\hat{Y}) = \sum_{i=1}^{n} y_i^2 \cdot \frac{(1 - \pi_i)}{\pi_i^2} + \sum_{i \neq j}^{n} y_i y_j \frac{(P_{ij} - \pi_i \pi_j)}{P_{ij} \cdot \pi_i \pi_j}. \qquad (1.11)$$

Both $\pi_i$ and $P_{ij}$ are greater than zero for all i and j.

In the estimating functions (1.9), (1.10) and (1.11) it will be noticed that it is the quantities $\pi_i$ and $P_{ij}$ that can be controlled by the sampler. Assume a "measure of size" $x_i$ is known for all the units in the population (i = 1, 2, $\cdots$, N) and it is suspected that these known sizes $x_i$ are correlated with the characteristics $y_i$ for which the population total Y is to be estimated. The sampler may wish to utilize the information in $x_i$ in assigning the selection probabilities such that the resulting $\pi_i$ and $P_{ij}$ will lead to a reduction in variance. There are many papers discussing this kind of problem. In the next three sections three sampling procedures will be discussed.

## METHODS OF YATES AND GRUNDY

### Description of the Sampling Procedure

Yates and Grundy (1953) attack the problem of assigning selection probabilities as follows. The first unit in the sample is selected with probabilities proportional to the revised sizes $x_i^*$ which are obtained by an iteration method to be explained in the next section, the second unit with probabilities proportional to the remaining revised sizes, and so on. It is possible to determine revised size measures from the original sizes measures.

To obtain probabilities using the original size measures, let $y_i$ denotes a characteristic attached to the $i^{th}$ unit of a finite population of N units. Suppose $x_i$ is a known size measure related to the $i^{th}$ unit. For convenience in writing the formulas we may replace the actual measures of size $x_i$ by proportions $p_i$ such that $p_i = \frac{X_i}{\sum_{i=1}^{N} x_i}$ and $\sum_{i=1}^{N} p_i = 1$. The probability of selecting units i and j in that order is then

$$p_i \cdot \frac{p_j}{1-p_i}$$

The total probability of selecting units i and j when a sample of two units is taken is therefore

$$p'_{ij} = p_i p_j \left( \frac{1}{1-p_i} + \frac{1}{1-p_j} \right) . \qquad (2.1)$$

The total probability of selecting unit i, which we may denote by $2p_i'$, is given by the sum of the probabilities of selecting unit i first and the probability of selecting unit i second after having selected some other unit. If

$$A = \sum_{i=1}^{N} p_i / (1 - p_i) ,$$

we have

$$2p_i' = p_i + \sum_{j(\neq i)}^{N} \frac{p_i p_j}{1 - p_j} = \left[ 1 + A - \frac{p_i}{1 - p_i} \right] \quad (2.2)$$

where $p_i'$ may be termed the effective relative probability of selection of unit i. These probabilities are easily calculated. All that is necessary is to calculate all $p_i/(1 - p_i)$ and their sum.

Determine Revised Size Measures

Let new effective relative probabilities be denoted by $q_i$. From equation (2.2), substituting $p_i$ for $p_i'$, and $q_i$ for $p_i$, we have the following N equations for determining the $q_i$.

$$2p_i = q_i \left[ 1 + A' - \frac{q_i}{1 - q_i} \right], \quad i = 1, 2, \cdots N \quad (2.3)$$

where

$$A' = \sum_{i=1}^{N} \frac{q_i}{1 - q_i} .$$

As a first approximation we may put

$$1 + A' - \frac{q_i}{1 - q_i} \doteq 1 + A - \frac{p_i}{1 - p_i} \quad (2.4)$$

From equation (2.2) the right hand side of equation (2.4) equals $2p_i'/p_i$ . Hence, a first approximation is given by modifying equation (2.3) such that

$$2p_i = [q_i]_1 \left( 2p_i'/p_i \right) ,$$

$$[q_i]_1 = \frac{2p_i^2}{2p_i'} = \frac{p_i^2}{p_i'} . \tag{2.5}$$

It will be necessary to make small adjustments in the $[q_i]_1$ in order to make them add up to unity. The new effective relative probabilities of selection $[q_i']_1$ can now be calculated from these $[q_i]_1$ in the same manner as the $p_i'$ were calculated from $p_i$ by equation (2.2).

A second approximation is given by

$$2p_i = [q_i]_2 \left( 2[q_i']_1/[q_i]_1 \right) .$$

Therefore

$$[q_i]_2 = \frac{[q_i]_1 \cdot p_i}{[q_i']_1}$$

with adjustment as before.

Then the $n^{th}$ approximation will be

$$[q_i]_n = \frac{[q_i]_{n-1} \cdot p_i}{[q_i']_{n-1}} . \tag{2.6}$$

## Numerical Example

In order to illustrate the practical utility of the above formulas we take a population with N = 4 as shown in Table 1. Assume the original size measures are known.

Table 1. Successive Approximations to Required Probabilities

| Unit | $P_i$ | $[\pi_i]_1$ | $[\pi'_i]_1$ | $[\pi_i]_2$ | $[\pi'_i]_2$ |
|---|---|---|---|---|---|
| 1 | .1 | .084 | .101 | .081 | .100 |
| 2 | .2 | .180 | .207 | .173 | .203 |
| 3 | .3 | .293 | .308 | .283 | .305 |
| 4 | .4 | .443 | .382 | .463 | .392 |
| Total | 1.0 | 1.000 | .998 | 1.000 | 1.000 |

## Variance and the Estimate of Variance of the Estimator of Population Total

Only the case n = 2 is considered. In terms of the notation in previous section,

$$\pi_i = 2\,[\pi'_i]_n \quad , \qquad P_{ij} = P'_{ij} \quad .$$

Then

$$\hat{Y} = \sum_{i=1}^{n} \frac{y_i}{\pi_i} \quad .$$

A more suitable form of the variance may be derived from equation (1.10). We have

$$\sum_{i=1}^{N} \pi_i = n \quad , \qquad \sum_{i>j} P_{ij} = \frac{1}{2} n (n-1).$$

It is also easily established that

$$\sum_{j(\neq i)}^{N} P_{ij} = (n-1)\pi_i \ ,$$

$$\sum_{j(\neq i)} (P_{ij} - \pi_i \pi_j) = (n-1)\pi_i - \pi_i (n - \pi_i) = -\pi_i (1-\pi_i) .$$

Hence

$$V(\hat{Y}) = \sum_{i=1}^{N} \frac{y_i^2}{\pi_i} + \sum_{i\neq j}^{N} \frac{P_{ij}}{\pi_i \pi_j} y_i y_j - Y^2 \ ,$$

$$= \sum_{i=1}^{N} \pi_i \cdot \frac{y_i^2}{\pi_i^2} + \sum_{i\neq j}^{N} P_{ij} \cdot \frac{y_i y_j}{\pi_i \pi_j} - \left( \sum_{i=1}^{N} \pi_i \cdot \frac{y_i}{\pi_i} \right)^2 ,$$

$$= \sum_{i=1}^{N} \pi_i \cdot \frac{y_i^2}{\pi_i^2} + \sum_{i\neq j}^{N} P_{ij} \frac{y_i y_j}{\pi_i \pi_j} - \sum_{i=1}^{N} \pi_i^2 \cdot \frac{y_i^2}{\pi_i^2}$$

$$- \sum_{i\neq j}^{N} \pi_i \pi_j \frac{y_i y_j}{\pi_i \pi_j} \ ,$$

$$= \sum_{i=1}^{N} \pi_i (1-\pi_i) \frac{y_i^2}{\pi_i^2} + \sum_{i\neq j}^{N} (P_{ij} - \pi_i \pi_j) \cdot \frac{y_i y_j}{\pi_i \pi_j} \ ,$$

$$= \sum_{i=1}^{N} \sum_{j(\neq i)}^{N} (\pi_i \pi_j - P_{ij}) \frac{y_i^2}{\pi_i^2} - \sum_{i\neq j}^{N} (\pi_i \pi_j - P_{ij}) \frac{y_i y_j}{\pi_i \pi_j} \ ,$$

$$= \sum_{i\neq j}^{N} (\pi_i \pi_j - P_{ij}) \left( \frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 .$$

An unbiased estimate of variance is given by

$$v(\hat{Y}) = \sum_{i \neq j}^{n} \frac{(\pi_i \pi_j - P_{ij})}{P_{ij}} \left( \frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 .$$

## METHOD OF DES RAJ

### Description of the Sampling Procedure

The following method is described by Des Raj (1956) and is relatively simple. Out of the totality of $\binom{N}{2}$ groups with two units each (this method is likely to be inconvenient for larger sample size), one selects one group. But, the restriction is that its given probabilities should be assigned in an optimum way; i.e.,

$$P_{ij} \geq 0 \quad , \tag{3.1}$$

$$\sum_{j(\neq i)} P_{ij} = \pi_i \quad ( i = 1, 2, \cdots, N ), \tag{3.2}$$

and

$$\sum_{i<j}^{N} P_{ij} \frac{y_i y_j}{\pi_i \pi_j} \tag{3.3}$$

is minimized.

It is assumed that

$$y = \alpha + \beta x$$

where x is a "measure of size" and y represents the characteristic of the population. $\alpha$ and $\beta$ are constants.

By modification

$$\sum_{i<j}^{N} P_{ij} \frac{y_i y_j}{\pi_i \pi_j} = \frac{1}{2} \sum_{i \neq j}^{N} \frac{P_{ij}}{\pi_i \pi_j} ( \alpha + \beta x_i )( \alpha + \beta x_j ),$$

$$= \frac{1}{2} \sum_{i \neq j}^{N} \frac{P_{ij}}{\pi_i \pi_j} ( \alpha^2 + \alpha \beta x_i + \alpha \beta x_j + \beta^2 x_i x_j ).$$

(3.4)

Since

$$\pi_i = \frac{2\,x_i}{\sum_{i=1}^{N} x_i}\,, \qquad \pi_j = \frac{2\,x_j}{\sum_{j=1}^{N} x_j}$$

then

$$x_i = \frac{\pi_i \cdot \sum_{i=1}^{N} x_i}{2}\,, \qquad x_j = \frac{\pi_j \cdot \sum_{j=1}^{N} x_j}{2}\,.$$

Furthermore,

$$\sum_{i=1}^{N} \pi_i = \sum_{j=1}^{N} \pi_j = 2\,,$$

$$\sum_{j(\neq i)}^{N} P_{ij} = \pi_i\,, \qquad \sum_{i(\neq j)}^{N} P_{ij} = \pi_j\,.$$

So

$$\sum_{i<j}^{N} P_{ij} \cdot \frac{y_i\, y_j}{\pi_i \pi_j} = \frac{1}{2} \sum_{i\neq j}^{N} \frac{P_{ij}}{\pi_i \pi_j} \Big( \alpha^2 + \alpha\beta \cdot \frac{(\sum_{t=1}^{N} x_t)\cdot \pi_i}{2} + \alpha\beta \cdot \frac{\pi_j \cdot \sum_{t=1}^{N} x_t}{2} + \beta^2 \cdot \frac{\pi_i \cdot \sum_{t=1}^{N} x_t}{2} \cdot \frac{\pi_j \cdot \sum_{t=1}^{N} x_t}{2} \Big),$$

$$= \frac{\alpha^2}{2} \sum_{i\neq j}^{N} \frac{P_{ij}}{\pi_i \pi_j} + \frac{\alpha\beta \cdot (\sum_{t=1}^{N} x_t)}{4} \Big( \sum_{i\neq j}^{N} \frac{P_{ij}}{\pi_j} + \sum_{i\neq j}^{N} \frac{P_{ij}}{\pi_i} \Big) + \frac{(\sum_{t=1}^{N} x_t)^2}{8} \beta^2 \sum_{i\neq j}^{N} P_{ij}\,,$$

$$= \frac{\alpha^2}{2} \sum_{i\neq j}^{N} \frac{P_{ij}}{\pi_i \pi_j} + \frac{N}{2} \alpha\beta \Big( \sum_{t=1}^{N} x_t \Big) + \frac{(\sum_{t=1}^{N} x_t)^2}{4} \cdot \beta^2\,. \qquad (3.5)$$

In equation (3.5), only $\sum_{i \neq j}^{N} \frac{P_{ij}}{\pi_i \pi_j}$ is variable. The problem reduces to the determination of $P_{ij}$ such that

$$\left.\begin{aligned} & P_{ij} \geq 0 \ , \\ & \sum_{j(\neq i)}^{N} P_{ij} = \pi_i \quad (i = 1, 2, \cdots, N) \end{aligned}\right\} \qquad (3.6)$$

and $\sum_{i<j}^{N} P_{ij} / \pi_i \pi_j$ is minimized.

Numerical Example

As an illustration of the practical utility of the method the three populations given by Yates and Grundy (1953) are taken into consideration. The object is to estimate the population total by selecting two units with probabilities of inclusion $\pi_i$ proportional to the following $p_i$

| Unit | p | $y_i$ |
|---|---|---|
| 1 | 0.1 | 0.5 |
| 2 | 0.2 | 1.2 |
| 3 | 0.3 | 2.1 |
| 4 | 0.4 | 3.2 |

We have to find $p_{ij}$ according to the following restrictions:

$$p_{ij} \geq 0$$

$$p_{12} + p_{13} + p_{14} = 0.2,$$
$$p_{21} + p_{23} + p_{24} = 0.4, \qquad \text{where } p_{ij} = p_{ji}$$
$$p_{31} + p_{32} + p_{34} = 0.6,$$
$$p_{41} + p_{42} + p_{43} = 0.8,$$

and $G = 12.5p_{12} + 8.3333p_{13} + 6.25p_{14} + 4.1667p_{23} + 3.125p_{24} + 2.0833p_{34}$ is minimized.

The optimum assignment of $p_{ij}$, obtained by the simplex method, is given in Table 2 below.

Table 2. Optimum Assignment of $p_{ij}$

| $u_i$ \ $u_j$ | 1 | 2 | 3 | 4 | Total |
|---|---|---|---|---|---|
| 1 | --- | 0.0 | 0.0 | 0.2 | 0.2 |
| 2 | 0.0 | --- | 0.2 | .0.2 | 0.4 |
| 3 | 0.0 | 0.2 | --- | 0.4 | 0.6 |
| 4 | 0.2 | 0.2 | 0.4 | --- | 0.8 |
| Total | 0.2 | 0.4 | 0.6 | 0.8 | 2.0 |

Estimates of the mean and variance of Y are obtained by using the standard formula given by Horvitz and Thompson (1952). (See 1.9 and 1.11). Thus

$$\hat{Y} = \sum_{i=1}^{n} y_i / \pi_i \quad ,$$

$$V(\hat{Y}) = \sum_{i=1}^{N} \frac{y_i^2}{\pi_i} + \sum_{i \neq j}^{N} \frac{p_{ij}}{\pi_i \pi_j} y_i y_j - Y^2$$

and

$$v(\hat{Y}) = \sum_{i=1}^{n} y_i^2 \cdot \frac{(1-\pi_i)}{\pi_i^2} + \sum_{i \neq j}^{n} y_i y_j \frac{(p_{ij} - \pi_i \pi_j)}{p_{ij} \cdot \pi_i \pi_j}$$

is an unbiased estimator of the variance of $\hat{Y}$.

## METHOD OF H. O. HARTLEY AND J. N. K. RAO

### Description of the Sampling Procedure

In the method proposed by Hartley and Rao (1962), it is assumed that

$$\pi_i = n p_i \leq 1 \tag{4.1}$$

where

$$p_i = x_i \Big/ \sum_{j=1}^{N} x_j .$$

Arrange the units of the population in a random order. Then give notation j = 1, 2, ···, N to this random order and denote the progressive totals of the $(np_i)$ in that order by

$$M_j = \sum_{i=1}^{j} (n p_i) \quad , \quad M_0 = 0 .$$

Selected a "random start," i.e., select a "uniform variate" d with $0 \leq d < 1$. Then the n selected units are those whose index, j, satisfies

$$M_{j-1} \leq d + k < M_j \tag{4.2}$$

for some integer k between 0 and n - 1. Since $np_i \leq 1$, every one of the n integers k = 0, 1, 2, ··· n - 1 will select a different sampling unit j.

### Numerical Example

Consider a population of N = 8 units arranged in a random order and with sizes $x_j$ shown in the second column of Table 3. A sample of n = 3 is to be drawn using this sampling

procedure. Instead of computing the quantities $np_i$, scale all computations up by a factor of $\sum_{i=1}^{N} x_i/n = \frac{300}{3} = 100$. Then compute the progressive sums of the $x_i$ and these are shown in column 3 of Table 3 and correspond to the quantities $M_j \cdot \sum_{i=1}^{N} x_i/n$. Then select a random integer between 1 and 100 and this corresponds to the quantity $d \cdot \sum_{i=1}^{N} x_i/n$. In this example the integer turned out to be 58 and the selection of the three units in accordance with (4.2) is shown in column 4. We may find the lines (j) where the column $100M_j$ passes through the levels 100d = 58 (for k = 0), 100d + 100 = 158 (for k = 1) and 100d + 200 = 258 (for k = 2). So the units j = 2, 4, 8 are selected.

Table 3. Selection of n = 3 units from Population of N = 8 Units (p.p.s.)

| Unit Number j | Size $x_j$ | Progressive Sum $100\,M_j$ | |
|---|---|---|---|
| 1 | 15 | 15 | |
| 2 | 81 | 96 | k=0, 100d=58 |
| 3 | 26 | 122 | |
| 4 | 42 | 164 | k=1, 100d+100=158 |
| 5 | 20 | 184 | |
| 6 | 16 | 200 | k=2, 100d+200=258 |
| 7 | 45 | 245 | |
| 8 | 55 | 300 | |

With the help of asymptotic theory, compact expressions for the variance of the estimate of the population total are derived by H. O. Hartley and J. N. K. Rao (1962). The estimates of variance of the population total are derived by them, too. These formulas are applicable for moderate values of N.

For the case, n = 2, the variance of the estimate of the population total $\hat{Y} = \sum_{i=1}^{n} y_i / \pi_i$ is given by

$$V_1(\hat{Y}) = \sum_{i=1}^{N} \pi_i \left(1 - \frac{\pi_i}{2}\right)\left(\frac{y_i}{\pi_i} - \frac{Y}{2}\right)^2 - \frac{1}{2}\sum_{i=1}^{N}\left[\pi_i^3 - \frac{\pi_i^2 \sum_{i=1}^{N} \pi_i^2}{4}\right]\left(\frac{y_i}{\pi_i} - \frac{Y}{2}\right)^2 + \frac{1}{4}\left[\sum_{i=1}^{N} \pi_i y_i - \frac{Y \sum_{i=1}^{N} \pi_i^2}{2}\right]^2 \qquad (4.3)$$

to terms of $O(N^0)$, and

$$V_2(\hat{Y}) = \sum_{i=1}^{N} \pi_i \left(1 - \frac{\pi_i}{2}\right)\left(\frac{y_i}{\pi_i} - \frac{Y}{2}\right)^2 \qquad (4.4)$$

to terms of $O(N^1)$.

The estimate of $V_1(\hat{Y})$ is

$$v_1(\hat{Y}) = \left[1 - (\pi_i + \pi_j) + \frac{1}{2}\sum_{i=1}^{N} \pi_i^2 - \frac{1}{2}(\pi_i^2 + \pi_j^2) - \frac{1}{4}\left(\sum_{i=1}^{N} \pi_i^2\right)^2 + \frac{1}{4}(\pi_i + \pi_j)\sum_{i=1}^{N} \pi_i^2 + \frac{1}{2}\sum_{i=1}^{N} \pi_i^3\right]\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2, \qquad (4.5)$$

while the estimate of $V_2(\hat{Y})$ is

$$v_2(\hat{Y}) = \left[1 - (\pi_i + \pi_j) + \frac{1}{2}\sum_{i=1}^{N} \pi_i\right]\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2. \qquad (4.6)$$

The choice between $V_1(\hat{Y})$ and $V_2(\hat{Y})$ depends on the size of the population. For moderately large N, $V_2(\hat{Y})$ will contribute enough to the reduction in the variance. However, for smaller N, it may be necessary to take $V_1(\hat{Y})$ into account.

For the general case n > 2, the variance becomes

$$V(\hat{Y}) = \sum_{t=1}^{N} \pi_t \left[1 - \frac{(n-1)}{n} \cdot \pi_t\right] \left(\frac{y_t}{\pi_t} - \frac{Y}{n}\right)^2 \qquad (4.7)$$

to terms of $O(N^1)$ and this is estimated by

$$v(\hat{Y}) = \frac{1}{n-1} \sum_{i<j}^{n} \left[1 - (\pi_i + \pi_j) + \frac{1}{n}\pi_t^2\right] \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2. \qquad (4.8)$$

## A SIMPLE PROCEDURE GIVEN BY RAO, HARTLEY AND COCHRAN

### Description of the Sampling Procedure

The following procedure was discussed by Rao, Hartley and Cochran (1962). Let $p_t$ be the probability for drawing the $t^{th}$ unit in the first draw from the whole population. For example, suppose we are sampling with probability proportional to size of $x_t$, where $p_t = x_t / \sum_{t=1}^{N} x_t$. The sampling procedure consists of the following two stages:

(1) Split the population at random into n groups of sizes $N_1, N_2, N_3, \cdots, N_n$ where $N_1 + N_2 + \cdots + N_n = N$.

(2) Draw a sample of size one with probabilities proportional to $p_t$ from each of these n groups independently.

If the $t^{th}$ unit falls in group i, the actual probability that it will be selected is $p_t / \pi_i$ where

$$\pi_i = \sum_{\text{Group } i} p_t \qquad (5.1)$$

### Numerical Example

Take the same population of N = 8 units shown in Table 3. Arrange these units in a random order. Break them into three groups (i.e. n = 3) with $N_1 = 3$ units, $N_2 = 3$ units and $N_3 = 2$ units according to the first stage introduced by this

sampling procedure. Then perform the procedure in stage two to find $\pi_i$ and $p_s/\pi_i$.

Table 4. Relative Probabilities from Population of N = 8

| Unit Number j | Size $x_t$ | Relative Probability $p_t = x_t / \sum_{t=1}^{N} x_t$ |
|---|---|---|
| 1 | 15 | 0.050 |
| 2 | 81 | 0.270 |
| 3 | 26 | 0.087 |
| 4 | 42 | 0.140 |
| 5 | 20 | 0.067 |
| 6 | 16 | 0.053 |
| 7 | 45 | 0.150 |
| 8 | 55 | 0.183 |
| Total | 300 | 1.000 |

Let $N_1$ consists of units 1, 4 and 8, then

$$\pi_1 = \sum_{\text{Group1}} p_t = 0.050 + 0.140 + 0.183 = 0.373$$

and

$$p_8/\pi_1 = \frac{0.183}{0.373} = 0.490$$

if unit 8 is drawn in the sample of n = 3.

Let $N_2$ consists of units 2, 5 and 6, then

$$\pi_2 = 0.270 + 0.067 + 0.053 = 0.390$$

and

$$p_2/\pi_2 = \frac{0.270}{0.390} = 0.692$$

if unit 2 is drawn in the sample of n = 3.

Let $N_3$ consists of units 3 and 7, by the same method

$$\pi_3 = 0.087 + 0.150 = 0.237$$

$$p_7/\pi_3 = \frac{0.150}{0.237} = 0.633$$

assuming unit 7 is drawn from this group.

Variance of the Estimate of Y

The estimator of the population total Y is

$$\hat{Y} = \frac{\sum_{i=1}^{n} y_i}{p_i/\pi_i} \tag{5.2}$$

where the suffixes 1, 2, ···, n denote the n units selected from the n groups separately.

The variance of $\hat{Y}$ is

$$V(\hat{Y}) = \frac{n\left(\sum_{i=1}^{n} N_i^2 - N\right)}{N(N-1)} \left( \sum_{t=1}^{N} \frac{y_t^2}{n p_t} - \frac{Y^2}{n} \right) . \tag{5.3}$$

Now the estimator of Y in sampling with replacement is

$$\hat{Y}' = \sum_{i=1}^{n} y_i / n p_i \tag{5.4}$$

where $\sum_{i=1}^{n}$ denotes the summation over the n units drawn with replacement, with variance

$$V(\hat{Y}') = \sum_{i=1}^{N} \frac{y_i^2}{n p_i} - \frac{Y^2}{n} \tag{5.5}$$

(See Cochran (1963), p. 253).

Therefore

$$V(\hat{Y}) = \frac{n\left(\sum_{i=1}^{n} N_i^2 - N\right)}{N(N-1)} V(\hat{Y}') . \tag{5.6}$$

From equation (5.6) it is seen that $V(\hat{Y})$ will be minimized if we choose $N_1 = N_2 = \cdots = N = R$. Therefore, if $\frac{N}{n} = R$, where R is a positive integer, then $N_1 = N_2 = \cdots = N_n = R$, and equation (5.6) becomes

$$\begin{aligned} V(\hat{Y}) &= \frac{n\left[\sum_{i=1}^{n} \left(\frac{N}{n}\right)^2 - N\right]}{N(N-1)} V(\hat{Y}'), \\ &= \frac{n\left[n \cdot \frac{N^2}{n^2} - N\right]}{N(N-1)} V(\hat{Y}'), \\ &= \frac{N^2 - Nn}{N(N-1)} V(\hat{Y}'), \\ &= \frac{N-n}{N-1} V(\hat{Y}'), \\ &= \frac{N-n+1-1}{N-1} V(\hat{Y}'), \\ &= \left(1 - \frac{n-1}{N-1}\right) V(\hat{Y}'). \end{aligned} \tag{5.7}$$

Equation (5.7) clearly shows the reduction in the variance as compared to sampling with replacement. If N is not a multiple of n, we have $N = nk + k$ where $0 < k < n$ and R is a positive

integer. Then choose

$$N_1 = N_2 = \cdots = N_k = R + 1 ,$$
$$N_{k+1} = N_{k+2} = \cdots = N_n = R ,$$

and equation (5.6) reduces to

$$V_2(\hat{Y}) = \left\{ 1 - \frac{n-1}{N-1} + \frac{k(n-k)}{N(N-1)} \right\} V(\hat{Y}') . \qquad (5.8)$$

For k = 1 or n - 1, equation (5.8) reduces to

$$V(\hat{Y}) = \left(1 - \frac{n-1}{N}\right) V(\hat{Y}') . \qquad (5.9)$$

The unbiased estimator of V(Y) is

$$v(\hat{Y}) = \frac{\left(\sum_{i=1}^{n} N_i^2 - N\right)}{\left(N^2 - \sum_{i=1}^{n}\right)} \left\{ \sum_{i=1}^{n} \pi_i \left( \frac{y_i}{p_i} - \hat{Y} \right)^2 \right\} \qquad (5.10)$$

with $N_1 = N_2 = \cdots\cdots = N_k = R + 1$; $N_{k+1} = N_{k+2} = \cdots\cdots = R$.

Analogous to equation (5.9), one obtains

$$v_2(\hat{Y}) = \frac{N^2 + k(n-k) - Nn}{N^2(n-1) - k(n-k)} \left\{ \sum_{i=1}^{n} \pi_i \left( \frac{y_i}{p_i} - \hat{Y} \right) \right\} \qquad (5.11)$$

## EXAMPLES AND DISCUSSION

### Numerical Example

In order to compare the efficiency of these different procedures, consider the three populations introduced by Yates and Grundy (1953). The three populations have the same set of $p_i$ values and are given in Table 5.

Variances for the five procedures and the three populations are given in Table 6. Variances for procedures 1 and 2 are taken from Des Raj (1956). For procedure 3 the variance is taken from Hartley and Rao (1962). And variances for 4 and 5 are obtained from equation (5.7) and (5.5) respectively.

Table 5. Three Populations of Size N = 4

| Unit Number | $p_i$ | Population A $y_i$ | Population B $y_i$ | Population C $y_i$ |
|---|---|---|---|---|
| 1 | 0.1 | 0.5 | 0.8 | 0.2 |
| 2 | 0.2 | 1.2 | 1.4 | 0.6 |
| 3 | 0.3 | 2.1 | 1.8 | 0.9 |
| 4 | 0.4 | 3.2 | 2.0 | 0.8 |

Table 6. Comparative Efficiency of Five Sampling Procedures

| Procedure | Population A | | Population B | | Population C | |
|---|---|---|---|---|---|---|
| | Var. | Eff.% | Var. | Eff.% | Var. | Eff.% |
| 1. Des Raj | 0.200 | 100.0 | 0.200 | 100.0 | 0.100 | 100.0 |
| 2. Yates & Grundy | 0.323 | 61.9 | 0.269 | 74.3 | 0.057 | 175.4 |
| 3. Hartley & Rao | 0.367 | 54.5 | 0.367 | 54.5 | 0.033 | 333.3 |
| 4. Rao, Hartley & Cochran | 0.333 | 60.1 | 0.333 | 60.1 | 0.083 | 120.5 |
| 5. With Replacement | 0.500 | 40.0 | 0.500 | 40.0 | 0.125 | 80.0 |

## Discussion

It is seen from Table 6 that procedures 1, 2, 3 and 4 are more efficient than sampling with replacement.

Des Raj's procedure is the most efficient on populations A and B because A and B fairly well satisfy the linear model $y = \alpha + \beta x$. For population C the model is not appropriate so that considerable loss in efficiency results from Des Raj's procedure. This procedure is not convenient for large sample size. It is usually applied to the case $n = 2$. Though this procedure involves heavy computation through the simplex method in linear programming, we may use a computer to solve it.

So far as efficiency is concerned, procedures 2, 3 and 4 are almost the same. Procedure 4 may have a slightly loss

in efficiency when N is moderate and is not a multiple of n.

Procedure 2 requires a cumbersome evaluation of revising the measures by an iteration method. Besides, it is impractical for use in the case $n > 2$.

Procedure 3 gives only asymptotic variance for the estimates of Y. It is impractical when N is small. For large and moderate size populations it provides a convenient process for sampling analysis.

Procedure 4 does not need heavy computations. It is convenient when either sample size $n = 2$ or sample size $n > 2$ is applied. In comparison to procedure 3 this procedure will, in many situations, lead to an estimator with a slightly larger variance.

## Conclusion

When a population approximately satisfies the linear model $y = \alpha + \beta x$ and the sample size is $n = 2$, we may either take procedure 1 to get high efficiency or use procedure 4 for easy calculation.

When N is moderate or large and the sample size is $n > 2$, procedure 3 is preferred.

When N is small and the sample size is $n > 2$, procedure 4 is preferred.

ACKNOWLEDGEMENT

The writer wishes to express his deep appreciation and gratitude to his advisor, Dr. A. M. Feyerherm, for suggestion of this topic and for his advise and guidance during the preparation of this report.

BIBLIOGRAPHY

1. Cochran, W. G. "Sampling Technique." New York: Wily & Sons, 1963.

2. Des Raj. "A Note on the Determination of Optium Probabilities in Sampling without Replacement." Journal of Sankhyā, V17 (1956), 197-200.

3. Hansen, Morris H. and Hurwitz, W. N. "On the Theory of Sampling from Finite Populations." Annals of Mathematical Statistics, V 20 (1943), 333-362.

4. Hartley, H. O. and Rao, J. N. K. "Sampling with Unequal Probabilities without Replacement." Annals of Mathematical Statistics, V 33 (1962), 350-374.

5. Horvitz, D. G. and Thompson, D. J. "A Generalization of Sampling without Replacement from a Finite Universe." Journal of American Statistical Association, V 47 (1952), 663-685.

6. Rao, J. N. K., Hartley, H. O. and Cochran, W. G. "On a Simple Procedure of Unequal Probability Sampling without Replacement." Journal of Royal Statistical Association, V 24 (1962), Series B, 482-491.

7. Yates, F. and Grundy, P. M. "Selection without Replacement from within Strata with Probability Proportional to size." Journal of Royal Statistical Association, V 15 (1953), Series B, 253-261.

# ON SELECTING SAMPLING UNITS WITH PROBABILITIES PROPORTIONAL TO SIZES AND WITHOUT REPLACEMENT

by

TONG-SANG LIU

B. S., National Taiwan University, 1960

---

AN ABSTRACT OF A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1967

ABSTRACT

Given a finite population of N units with values of a characteristic represented by $y_i$ ($i = 1, 2, \cdots, N$), this report deals with the problem of estimating the sum of the $y_i$'s when measures of size $x_i$, which are positively correlated with the values $y_i$, are known for all N units in the population. General theory and four important methods of selecting sampling units with probabilities proportional to sizes and without replacement are discussed.

The general theory is derived by Horvitz and Thompson (1952), and the four methods are given by Yates and Grundy (1953), Des Raj (1956), Hartley and Rao (1962) and Rao, Hartley and Cochran (1962), respectively.

Results from a number of numerical examples indicated that all four sampling methods are more efficient than sampling with replacement. When a population approximately satisfies the linear model $y = \alpha + \beta x$ and the sample size is $n = 2$, Des Raj's method is the most efficient. However, in more general situations, the method given by Rao, Hartley and Cochran (1962) is preferred for easier calculation.