SEQUENTIAL SAMPLE SELECTION METHODS

James R. Chromy
Research Triangle Institute

1. INTRODUCTION

Sequential sample selection methods are distinguished from conventional methods in the manner in which random numbers are used to determine the sample. Conventionally, a sample of size n is selected from a sampling frame of N sampling units by selecting n random numbers and mapping then into n of the labels in the sampling frame. Sequential methods require each sampling unit in the sampling frame to be considered in order and a probabilistic decision reached concerning its inclusion in the sample. Sequential methods can be implemented efficiently on computers and are particularly adaptable to selecting samples as a file is read or as a process takes place. A thorough discussion of several equal probability sequential selection procedures is given by Fan, Muller, and Rezucha (1962). Their method 1 for selecting a simple random sample of size n from N involves sequentially comparing conditional probabilities of selection for each unit labeled i = 1,2,...,N with a uniform random number

$$\frac{n-u}{N+1-i} \geq r_i$$

where u is the number selected from the first (i-1) units. Sampling unit i is included when the inequality is satisfied. This method can be programmed very easily, requires very little computer memory, and can be incorporated into other computer programs which read and process computer accessible files.

The purpose of this paper is to extend the concept of sequential selection to more general unequal probability sampling schemes.

2. DEFINITIONS

The following notation will be used:

- N = number of sampling units in the sampling frame;
- i = sampling unit label;
- $S(i)$ = size measure associated with sampling unit i;
- $S(+) = \sum_{i=1}^{N} S(i)$;
- n = total sample size; and
- $n(i)$ = number of sample hits at sampling unit i.

A probability proportional to size (PPS) sample design will be defined as one for which the expected number of sample hits at unit i is proportional to the size measure, S(i). More precisely, the following conditions will hold

$$En(i) = nS(i)/S(+)$$

and

$$\sum_{i=1}^{N} En(i) = n.$$

The process of sample selection can be considered as the process of determining the values of n(i) for i = 1,2,...,N. If a probability nonreplacement (PNR) sample design is employed, then n(i) will equal one for n values of i and will equal zero for all other values of i. If a probability replacement (PR) sample design is employed, the same sampling unit can be selected more than once, i.e., each n(i) can assume values 0 through n. In both cases, PNR and PR, it holds that

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

Both PR and PNR sample designs are well defined in the statistical sampling literature. The concept of a probability minimum replacement (PMR) sample design is defined for the purposes of this paper. A PMR sample design is defined as a PPS design in which each n(i) can assume at most two values: (1) the integer portions of nS(i)/S(+), and (2) the next largest integer. It can be verified that a PMR sample design is PPS if

$$\begin{aligned}\text{Prob}\{n(i) &= \text{Int}[nS(i)/S(+)] + 1\} \\ &= \text{Frac}\,[nS(i)/S(+)].\end{aligned}$$

Note that if $n\,S(i)/S(+) < 1$ for i = 1,2,...,N, then a PMR design is also a PNR design.

Special equal probability cases of the general probability sampling designs discussed above occur when all the size measures, S(i), are equal. Equal probability cases of PR, PNR, and PMR designs can be denoted by EPR, EPNR, and EPMR respectively.

3. A SEQUENTIAL ALGORITHM FOR A PMR DESIGN

Some additional terms must be defined in order to describe the algorithm proposed for achieving one particular PMR sample design. Assume the sampling units in the sample frame are ordered from 1 to N. Let

$$I(i) = \text{Int}\left\{\sum_{j=1}^{i} n\,S(i)/S(+)\right\},$$

and

$$F(i) = \text{Frac}\left\{\sum_{j=1}^{i} n\,S(i)/S(+)\right\}.$$

By definition, $I(0) = 0$ and $F(0) = 0$. With this convention the PMR algorithm can be described as a sequential selection scheme where

$$\text{Prob}\left\{\sum_{j=1}^{i} n(j) = I(i) + 1 \,\middle|\, \sum_{j=1}^{i-1} n(j)\right\}$$

is defined as a function of $F(i)$ and $F(i-1)$. Conditional sequential probabilities of selection for three possible cases are shown in Table 1.

To select the sample, $\sum_{j=1}^{i} n(j)$ is determined for each value of i by the calculation of the appropriate conditional probability and comparison of the probability with a uniform random number. If the random number is less than the appropriate conditional probability, then

$$\sum_{j=1}^{i} n(j) = I(i) + 1.$$

Otherwise it is set at $I(i)$. The values of $n(i)$ are determined as

$$n(i) = \sum_{j=1}^{i} n(j) - \sum_{j=1}^{i-1} n(j).$$

For PNR designs, the sequential selection table can be stated in terms of conditional probabilities of selection as shown in Table 2.

4. PROPERTIES OF THE BASIC DESIGN

It can be shown that the sample design is PPS and further that it is PMR. The following lemma is required for the proofs.

Lemma: After each sequential sample selection step, the following condition holds:

$$\text{Prob}\left\{\sum_{j=1}^{i} n(j) = I(i) + 1\right\} = F(i).$$

Both the lemma and the PPS and PMR properties can be proved inductively since the algorithm is applied sequentially. All three cases in Table 1 must be considered at each step of the respective proofs.

5. AN UNBIASED ESTIMATE FOR THE POPULATION TOTAL

The population total, T, is defined as

$$T = \sum_{i=1}^{N} Y(i)$$

where $Y(i)$ is an observed value for sampling unit i.

Table 1. Conditional Sequential Probabilities for Cumulative Sample Counts

Case no.	Deterministic conditions	$\text{Prob}\left\{\sum_{j=1}^{i} n(j) = I(i) + 1 \,\middle|\, \sum_{j=1}^{i-1} n(j)\right\}$	
		$\sum_{j=1}^{i-1} n(j) = I(i-1)$	$\sum_{j=1}^{i-1} n(j) = I(i-1)+1$
(1)	$F(i) = 0$	0	0
(2)	$F(i)>F(i-1)\geq 0$	$[F(i)-F(i-1)]/[1-F(i-1)]$	1
(3)	$F(i-1)\geq F(i)>0$	0	$F(i)/F(i-1)$

Table 2. Conditional Sequential Probabilities of Selection (PNR Designs Only)

Case no.	Deterministic conditions	$\text{Prob}\left\{ n(i) = 1 \,\middle|\, \sum_{j=1}^{i-1} n(j) \right\}$	
		$\sum_{j=1}^{i-1} n(j) = I(i-1)$	$\sum_{j=1}^{i-1} n(j) = I(i-1)+1$
(1)	$F(i) = 0$	1	0
(2)	$F(i)>F(i-1)\geq 0$	$[F(i)-F(i-1)]/[1-F(i-1)]$	0
(3)	$F(i-1)\geq F(i)>0$	1	$F(i)/F(i-1)$

The sequential PMR selection algorithm always produces n hits, although they may sometimes be associated with fewer than n unique values of i. The sample can be represented in terms of an array of n labels

$$i_1, i_2, i_3, \ldots, i_h, \ldots, i_n.$$

The general form of estimator suggested for this design is

$$\hat{t} = \sum_{h=1}^{n} W(i_h)\, Y(i_h).$$

The term $\lambda(ih)$ is defined in order to find general conditions for unbiasedness as $\lambda(ih) = 1$ if sampling unit i is the h th element in the sample and as $\lambda(ih) = 0$ otherwise. (It is also used to facilitate variance estimation later in this paper.) Note that

$$n(i) = \sum_{h=1}^{n} \lambda(ih).$$

The estimator t can then be written as

$$t = \sum_{i=1}^{N} Y(i) \sum_{h=1}^{n} W(i_h)\, \lambda(ih),$$

and is seen to be an unbiased estimator of T if

$$E \sum_{h=1}^{n} W(i_h)\, \lambda(ih) = 1$$

for i = 1,2,...,N. The general condition for unbiasedness of t can be utilized to achieve self-weighting samples in multi-stage sample designs when the second-stage sample size is based on the selection h rather than being uniform over all selections. Ordinarily, $W(i_h)$ does not depend on h; under this circumstance, the condition for unbiasedness becomes

$$W(i)\, E \sum_{h=1}^{n} \lambda(i_h) = 1,$$

or

$$W(i)\, E\, n(i) = 1,$$

or

$$W(i) = 1/En(i),$$

and

$$t = \sum_{h=1}^{n} Y(i_h)/En(i_h).$$

For PPS designs,

$$W(i) = [S(+)/n][1/S(i)]$$

and

$$t = [S(+)/n] \sum_{h=1}^{n} y(i_h)/S(i_h).$$

The variance of t for the case $W(i) = 1/En(i)$ can be determined readily by noting that

$$t = \sum_{i=1}^{N} n(i)[Y(i)/En(i)] \quad .$$

Then since the n(i) are the only random variables in the above expression,

$$Var[t] = \sum_{i=1}^{N} [Y(i)/En(i)]^2\, Var[n(i)]$$

$$+ \sum_{i \neq j}^{N} \sum^{N} [Y(i)/En(i)][Y(j)/En(j)]\, Cov[n(i)\, n(j)] \quad .$$

Note that if this sample design is also PNR (i.e., if $En(i) < 1$ for each $i = 1,2,\ldots,N$), then the estimator t and its variance correspond to those developed by Horvitz and Thompson (1952). It should be noted that unless $En(i)n(j) > 0$ for all $i \neq j$, no unbiased estimator of the variance exists. Some useful biased estimators of the variance could perhaps be developed based on assumptions similar to those used for obtaining approximate variance estimates when systematic or one-draw-per-stratum stratified sample designs are used. Such approximations are not treated in this paper, since a fairly simple modification which resolves this problem is discussed in the next section.

An alternate expression for the variance corresponding to the one developed for PNR sample designs by Yates and Grundy (1953) can be written as

$$Var[t] = \sum_{i < j} \sum [Y(i)/En(i) - Y(j)/En(j)]^2$$

$$[En(i)En(j) - En(i)n(j)] \quad .$$

Some additional notation is useful for evaluating $En(i)En(j) - En(i)n(j)$.

Let

$$\pi(i) = Frac[En(i)]$$

$$= Prob\{n(i) = Int[En(i)] + 1\},$$

and

$$\pi(ij) = Prob\{n(i) = Int[En(i)] + 1,\ n(j) = Int[En(j)] + 1\} \quad .$$

Note that these quantities correspond to unit and pairwise probabilities in a PNR sample design. Then, it can be shown that

$$En(i)En(j) - En(i)n(j) = \pi(i)\pi(j) - \pi(ij) \quad .$$

Reductions in variance associated with systematic,

stratified, or zone sampling from meaningfully ordered lists also accrue to the proposed PMR selection scheme since for units adequately far apart on the sampling frame listing $En(i)n(j) = En(i)En(j)$ and their contribution to the variance of t can be seen to be zero from examination of the Yates-Grundy analogue to the variance formula.

6. AN ALGORITHM MODIFICATION FOR UNBIASED VARIANCE ESTIMATION

Most sampling frames can be viewed as a closed loop. Although stratification can be achieved by classification, it is more often based on ordering or on some combination of classification and ordering which ultimately results in an ordered list of N sampling units. The algorithm modification required to insure that

$$En(i)n(j) > 0$$

for all $i \neq j$ requires the following steps:

(1) Develop an ordered sampling frame of N sampling units;

(2) Select a unit with probability proportional to its size to receive the label 1;

(3) Continue labeling serially to the end of the sampling frame;

(4) Assign the next serial label to the first unit at the beginning of the list and continue until all sampling units are labeled;

(5) Apply the sequential PMR sample selection algorithm starting with the sampling unit labeled 1.

With this modification, an unbiased variance estimator can be obtained for sample designs with

$$\sum_{i=1}^{n} Frac[En(i)] \geq 2.$$

7. EXAMPLES AND APPLICATIONS

Example 1: An equal probability sample of 2 out of 5. The working probabilities worksheet assuming a random start of 1 is illustrated in Table 3.

For a fixed start to the sequential selection process, the values of $En(i)n(j)$ in matrix form are

$$1/20 \begin{bmatrix} 0 & 0 & 2 & 3 & 3 \\ 0 & 0 & 2 & 3 & 3 \\ 2 & 2 & 0 & 2 & 2 \\ 3 & 3 & 2 & 0 & 0 \\ 3 & 3 & 2 & 0 & 0 \end{bmatrix}$$

By allowing a random start and taking expectation over the 5 equally probable random starts, we get $En(i)n(j)$ in matrix form as

$$\begin{bmatrix} 0 & .07 & .13 & .13 & .07 \\ .07 & 0 & .07 & .13 & .13 \\ .13 & .07 & 0 & .07 & .13 \\ .13 & .13 & .07 & 0 & .07 \\ .07 & .13 & .13 & .07 & 0 \end{bmatrix}$$

This structure can be summarized as follows:

(1) $En(i) = 2/5$ for $i = 1,2,\ldots,5$;

(2) For $i<j$ and $Min\{j-i, 5+i-j\} = 1$,

$En(i)n(j) = .07$,

and

$[En(i)En(j) - En(i)n(j)]/En(i)n(j) = 9/7.$

For other $i \neq j$

$En(i)n(j) = .13$,

and

$[En(i)En(j) - En(i)n(j)]/En(i)n(j) = 3/13.$

Table 3. Working Probability Worksheet for Example 1

i	En(i)	$\sum_{j=1}^{i} En(i)$	I(i)	F(i)	$Prob\{\sum_{j=1}^{i} n(i) = I(i)+1 \mid \sum_{j=1}^{i-1} n(j)\}$	
					$\sum_{j=1}^{i-1} n(i) = I(i-1)$	$\sum_{j=1}^{i-1} n(i) = I(i-1)+1$
1	.4	.4	0	.4	.4	1
2	.4	.8	0	.8	2/3	1
3	.4	1.2	1	.2	0	1/4
4	.4	1.6	1	.6	1/2	1
5	.4	2.0	2	0	0	0

The variance estimator is

$$v(t) = (9/7)[5Y(i)/2 - 5Y(j)/2]^2$$

for $\text{Min}\{ji, 5+i-j\} = 1$, or

$$v(t) = (3/13)[5Y(i)/2 - 5Y(j)/2]^2$$

for other $i \neq j$.

Example 2: An unequal probability sample of 2 out of 4. Assuming a random start at the unit labeled 1, the working probability worksheet is illustrated below. Since this is a PNR design, the conditional probabilities are stated in terms of probabilities of selection rather than probabilities relating to the accumulated sample count. (See Table 4).

A method of computing pairwise probabilities can be developed if Table 5 is constructed based on Table 4.

Note that selection probabilities can be computed for each unit i as

$$\text{Prob}\{n(i) = 1\} = \sum_{u=0}^{1} \text{Prob}\{n(i) = 1 \mid u\} \; \text{Prob}\{\sum_{j=1}^{i-1} n(j) = u\} \; .$$

Pairwise probabilities for unit 1 and all other units can be obtained by constructing a worksheet for $i > 1$ conditional on the selection of unit 1 as shown in Table 6.

Now for each $i > 1$, the joint probability of unit i and unit 1 can be computed as

$$\text{Prob}\{n(1) = 1, n(i) = 1\} = \sum_{u=0}^{1} \text{Prob}\{n(i) = 1 \mid u\} \; \text{Prob}\{n(1) = 1, \sum_{j=1}^{i-1} n(j) = u\} \; .$$

Note that for PNR designs,

$$En(i)n(j) = \text{Prob}\{n(i) = 1, n(j) = 1\} \; .$$

The same procedure can be applied for computing $En(2)\, n(i)$ for $i > 2$ and for $En(3)\, n(4)$. If the procedure is repeated for all possible random starts, unconditonal values of $En(i)n(j)$ may be determined and are shown in Table 7.

8. PRACTICAL CONSIDERATIONS FOR VARIANCE ESTIMATION

Most applications of this method are associated with first-stage selection of multi-stage samples. In such applications, simpler variance

Table 4. Working Probability Worksheet for Example 2

i	En(i)	$\sum_{j=1}^{i} n(j)$	I(i)	F(i)	Prob $\{ n(i) = 1 \mid \sum_{j=1}^{i-1} n(j) \}$	
					$\sum_{j=1}^{i-1} n(i) = I(i-1)$	$\sum_{j=1}^{i-1} n(i) = I(i-1)+1$
1	.2	.2	0	.2	.2	-
2	.4	.6	0	.6	.5	0
3	.6	1.2	1	.2	1	1/3
4	.8	2.0	2	0	1	0

Table 5. Alternate Form of Working Probability Worksheet for Example 2

i	Prob $\{ \sum_{j=1}^{i-1} n(j) = u \}$			Prob $\{ n(i) = 1 \mid u = 0 \}$	Prob $\{ n(i) = 1 \mid u = 1 \}$
	u = 0	u = 1	u = 2		
1	1	0	0	.2	-
2	.8	.2	0	.5	0
3	.4	.6	0	1	1/3
4	0	.8	.2	0	1

Table 6. Worksheet for Computing En(1)n(i)

	Prob $\{n(1) = 1, \sum_{j=1}^{i-1} n(j) = u\}$			Prob $\{ n(i) = 1 \mid u = 0 \}$	Prob $\{ n(i) = 1 \mid u = 1 \}$
i	u = 0	u = 1	u = 2		
2	0	.2	0	.5	0
3	0	.2	0	1	1/3
4	0	.13333	.06667	1	0

Table 7. Unconditional Pairwise Expeditious

		En(i)n(j) for (ij) =					
Start	P(start)	(12)	(13)	(14)	(23)	(24)	(34)
1	.1	0	.06667	.13333	.13333	.26667	.4
2	.2	.08	.12	0	0	.32	.48
3	.3	0	.06667	.13333	.13333	.26667	.4
4	.4	.08	.12	0	0	.32	.48
Expected value		.048	.09867	.05333	.05333	.29867	.448

formulations which do not require computation of En(i)n(j) would usually be recommended. If pairwise expectations, En(i)n(j), are used for developing unbiased variance estimates, they only need to be computed for the pairs of sampling units in the sample.

REFERENCES

Fan, C. T., Muller, Mervin E., and Rezucha, Ivan (1962), "Development of Sampling Plans by Using Sequential (Item by Item) Selection Techniques and Digital Computers," Journal of the American Statistical Association, 57, 387-402.

Horvitz, D. G., and Thompson, D. J. (1952), "A Generalization of Sampling From a Finite Universe," Journal of the American Statistical Association, 47, 663-685.

Yates, F., and Grundy, P. M. (1953), "Selection Without Replacement From Within Strata and With Probability Proportional to Size," Journal of the Royal Statistical Society, B(15), 253-261.