NUMBER 9

SELECTION OF SIMPLE AND STRATIFIED RANDOM SAMPLES OF FIXED SIZE WITHOUT REPLACEMENT

Michael H. Bostron

Division of Disability Studies

JUNE 1979

Social Security Administration
Office of Policy
Office of Research and Statistics

Selection of Simple and Stratified Random Samples of Fixed Size Without Replacement*

Introduction

For the past few years, the Division of Disability Studies has been using simple random and stratified random sampling procedures for many of its studies. The beneficiary sample for the 1978 Survey of Disability and Work was a stratified random sample drawn from the Master Benefit Record. The samples used in the Study of Consistency and Validity of Initial Disability Decisions and the Trial Work Period Folder Study also used simple random sampling procedures. Simple random subsampling has been used to enable multivariate analysis to be performed on files which would otherwise have been too large for existing software.

Because of the Division of Disability Studies' wide use of simple and stratified random sampling designs, software was developed to efficiently accomplish these sampling schemes. This paper describes the algorithm and presents the computer programs which are currently being used in the Division.

---

*By Michael H. Bostron, Division of Disability Studies

Simple random sampling without replacement

The basic model for a simple random sampling design without replacement can be defined in the following manner. Let U be a finite population of N identifiable units labeled 1, 2, ..., N. Let L(U) be the set of all possible subsets of U. There are $\binom{N}{n}$ sets of size n in L(U). A simple random sampling design without replacement can then be defined as a set function P on L(U) with the following properties.

(1) $P(S) = 1/\binom{N}{n}$ for all sets $S \in L(U)$ such that S has exactly n elements

(2) $P(S) = 0$ for all sets $S \in L(U)$ such that S does not have exactly n elements

Note that $P(S) \geq 0$ for all $S \in L(U)$ and $\sum P(S) = 1$.

Among these $\binom{N}{n}$ possible samples, there are $\binom{N-1}{n-1}$ which contain any one specific unit.

Each unit has probability of selection equal to $\binom{N-1}{n-1}/\binom{N}{n} = \frac{n}{N}$.

A sequential access file may be defined as a population list in which one can only access one case at a time in order from beginning to end. A simple random sample of size n, without replacement, may be selected in one of two ways from a sequential access file.

(1) Randomly pick n integers without replacement from the set of integers $\{1, 2, \ldots, N\}$, sort these numbers in ascending order, and store them so that these case numbers can be used to match to the frame to select the sample.

(2) Use the algorithm described below to randomly select the n cases from the file sequentially. This algorithm is based upon properties of the hypergeometric distribution.

Since the first method becomes cumbersome quickly as the sample size increases because a set of n random numbers must be generated, sorted, and stored before the sample cases can be selected from the frame, the second method is recommended for both convenience and efficiency.

The Algorithm

The following algorithm can be used to select a simple random sample of size n, given a population of size N.

Step 1: Let $N^* = N$ and $n^* = n$

Step 2: Read a case from the frame

Step 3: Let $F = n^*/N^*$

Step 4: Draw a uniformly distributed random number, R, on the interval (0,1).

Step 5: If $F \geq R$ select this this case, decrement N^* and n^* by 1 and go to Step 2. If $F < R$, do not select this case, decrement N^* by 1 and go to Step 2.

Steps (2-5) are repeated until n sample cases have been selected.

To show that this procedure produces a simple random sample, one must prove that:

$P(S) = 1/\binom{N}{n}$ for all possible sequences of unique case numbers, S, of size n.

Given:

N = population size

n = sample size desired

Let:

$$U = \{u_1, u_2, \ldots, u_n\}$$

$$S = \{u_{j_1}, u_{j_2}, \ldots, u_{j_n}\} = \{s_1, s_2, \ldots, s_n\}$$

Then:

$$P(S) = P(s_1)\, P(s_2 | s_1)\, P(s_3 | s_1, s_2) \cdots P(s_n | s_1, s_2, \ldots, s_{n-1})$$

according to the algorithm.

$$P(s_1) = \left[\prod_{i=0}^{s_1 - 2} \left(1 - \frac{n}{N - i}\right)\right] \frac{n}{N - s_2 + 1}$$

$$P(s_2 | s_1) = \left[\prod_{i=s_1}^{s_2 - 2} \left(1 - \frac{n - 1}{N - i}\right)\right] \frac{n - 1}{N - s_2 + 1}$$

$$P(s_k | s_1, s_2, \ldots, s_{k-1}) = \left[\prod_{i=s_{k-1}}^{s_k - 2} \left(1 - \frac{n - (k-1)}{N - i} \right) \right] \frac{n - (k-1)}{N - s_k + 1}$$

$$P(S) = \prod_{k=1}^{n} \left[\prod_{i=s_{k-1}}^{s_k - 2} \left(\frac{N - n - i + (k-1)}{N - i} \right) \right] \frac{n - k + 1}{N - s_k + 1}, \text{ where } s_0 = 0$$

$$n! \prod_{k=1}^{n} \left[\frac{\frac{(N - n - s_{k-1} + k - 1)!}{(N - n - s_k + k)!}}{\frac{(N - s_{k-1})!}{(N - s_k + 1)!}} \cdot \frac{1}{N - s_k + 1} \right]$$

$$= \frac{n! \left[\prod_{k=1}^{n} \frac{(N - s_k)!}{(N - n - s_k + k)!} \right]}{\frac{N!}{(N-n)!} \prod_{k=2}^{n} \frac{(N - s_{k-1})!}{(N - n - s_{k-1} + k - 1)!}}$$

$$= \frac{n!}{\frac{N!}{(N-n)!}} \left[\frac{\prod_{k=1}^{n} \frac{(N - s_k)!}{(N - n - s_k + k)!}}{\prod_{k=1}^{n-1} \frac{(N - s_k)!}{(N - n - s_k + k)!}} \right]$$

$$= \frac{n!(N-n)!}{N!} = \frac{1}{\binom{N}{n}}$$

Therefore, this procedure produces a simple random sample without replacement of size n from a finite population of size N.

Computer Programs

Two versions of a FORTRAN subroutine are available to perform this type of sample selection procedure. The first program, PICKN, can select k simple random samples simultaneously and thus can be used for stratified random sampling with fixed sample sizes in each strata. When using PICKN one must dimension two arrays, A and B (see example), by k, where k is the number of random samples to be selected. PICKN must be CALLed k times to initialize the arrays A and B to the corresponding sample size and population counts for each of the k stratum before processing of the frame can begin. The other program, PICK, selects a single simple random sample.

A random number generator is built into both routines which is set up for a 36-bit computer word (UNIVAC 1100 series). Documentation is available from the author describing the random number generator and how to change it for use on different computer systems.

The two subroutines are called as follows:

1. CALL PICKN(ISAMP,IPOP,I,ISEED,J)

   J --pointer for the Jth strata (sample) for which selection is made.

   ISAMP--Sample size desired for sample J, used for initiating purposes only.

   IPOP --Population size corresponding to sample J, used for initiating purposes only.

   I --Returns the case number of next case to be selected. If subroutine returns I = -1, then the Jth sample

selection is complete. Note, if k samples are to be selected, then one must set I = 0 to initiate A and B to correct sample and population values in subroutine, to invoke selection procedures, and to return the first sample case number, for each of the k samples. That is, PICKN must be called k times to invoke selection for each of the k samples before processing of frame can be done.

ISEED--If ISEED = 0, random number generator will start itself, otherwise, enter a five digit (or more) odd integer. This is done so that one can duplicate a sample selection by setting ISEED to value from the first selection.

2. CALL PICK(ISAMP,IPOP,I,ISEED)

Subroutine arguments are the same as above. Initializing must be done only once and the arrays A and B are not dimensioned.

An Example

Suppose one wanted to select a simple random sample of 150 initial disability decisions from each of the following four strata:

Strata	Population Count
Medical Allowance	21,035
Vocational Allowance	5,186
Medical Denial	17,411
Vocational Denial	23,494

```
@ASG,A M*M.
@COPY,S M*M.PICKN
@FREE   M*M.
@FOR,S  PICKN
-2,2
      REAL  A(4),B(4)
@FOR,IS  MAIN
C SELECT SAMPLE FOR BHI/DDS UNIFORMITY STUDY
        INTEGER IREC(4),ICOUNT(4)
        INTEGER IPOP(4)/21035,5186,17411,23494/
        INTEGER ISAMP(4)/4*150/,I(4)/4*0/
C INITIALIZE POPULATION AND SAMPLE COUNTS
C IP WILL CONTAIN FIRST RECORD IN EACH STRATA TO SELECT
        CALL PICKN(ISAMP(1),IPOP(1),IP(1),ISEED,1)
        CALL PICKN(ISAMP(2),IPOP(2),IP(2),ISEED,2)
        CALL PICKN(ISAMP(3),IPOP(3),IP(3),ISEED,3)
        CALL PICKN(ISAMP(4),IPOP(4),IP(4),ISEED,4)
C READ RECORD
   10    READ(10,100,END = 200)IREC
  100    FORMAT(1X,I9,3I4)
C COMPUTE STRATA IN WHICH RECORD BELONGS
        J=0
        IF(IREC(2).EQ.3)J=1
        IF(IREC(2).EQ.1.AND.IREC(3).EQ.1)J=2
        IF(IREC(2).EQ.1.AND.IREC(4).EQ.2)J=3
        IF(IREC(2).EQ.2)J=4
C IF RECORD NOT IN FRAME, GET ANOTHER
        IF(J.EQ.0)GO TO 10
        ICOUNT(J)=ICOUNT(J)+1
C IF RECORD NOT IN SAMPLE, GET ANOTHER
        IF(ICOUNT(J).NE.IP(J))GO TO 10
C RECORD SELECTED
        WRITE (12,105)J,IREC
  105   FORMAT(1X,I3,I10,3I4)
S SELECT NEW RECORD NUMBER FOR (J)TH STRATA
        CALL PICKN(ISAMP(J),IPOP(J),IP(J),ISEED,J)
        GO TO 10
  200   ENDFILE 12
        END
@ASG,A   DATA.
@USE   10,DATA.
@ASG,UP  12.
@XQT
```

DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
Social Security Administration
Office of Policy
Office of Research and Statistics
Division of Disability Studies

ORS WORKING PAPERS SERIES

Report Number	Title	Author	Date
2	Disability Beneficiary Recovery	Ralph Treitel	February 1979
3	Disability Claimants Who Contest Denials and Win Reversals Through Hearings	Ralph Treitel	February 1979
4	A Measure of Functional Capacity	Sandy Duchnok	March 1979
5	A Causative Matrix Approach to Mobility	Barry Bye John Hennessey	April 1979
6	Selection of Simple and Stratified Randon Samples of Fixed Size Without Replacement	Michael H. Boston	June 1979