United States
Department of
Agriculture

National
Agricultural
Statistics Service

Washington, D.C.
20250

COMPUTING INCLUSION PROBABILITIES

James W. Mergerson
Charles R. Perry

Staff Report No. SSB8802
January 1988

COMPUTING INCLUSION PROBABILITIES. James W. Mergerson and Charles R. Perry, Jr., Research and Applications Division, National Agricultural Statistics Service, U.S. Department of Agriculture, Washington, D.C. 20250, January 1988, Staff Report No. SSB8802.

ABSTRACT

In applied statistics, authors sometimes give general procedures in forms which cannot be programmed for implementation on a digital computer or give formulas only for trival cases. Raj (1968) describes a general procedure for computing inclusion probabilities which is not in a form amenable to implementation on a digital computer. Many authors present a formula for computing inclusion probabilities for samples selected strictly proportional to size and strictly without replacement for a non-rotating sample design (Narain's procedure) for the case $n = 2$. No solution has been proposed for Narain's procedure for the case $n > 2$. This paper presents a general, computational formula, which is easily implemented on a digital computer, for computing inclusion probabilities for samples selected strictly proportional to size and strictly without replacement for a non-rotating sample design for the case $n > 2$.

KEY WORDS: Proportional to size, Without replacement, Non-Rotating.

CONTENTS

COMPUTING INCLUSION PROBABILITIES

James W. Mergerson and Charles R. Perry Jr.

INTRODUCTION

Suppose a sample of size n is selected strictly proportional to size and strictly without replacement for a non-rotating sample design from a finite population of size N. Let p_i (i=1,2,3, ... ,N) denote the probability that the unit U_i is selected in a sample of size one. For n=2, the probability (π_i) that the unit U_i is included in the sample is given by

$$\pi_i = p_i + p_i * \sum_{j \neq i}^{N} p_j / (1-p_j) .$$

For $n > 2$, a computational expression for computing π_i has not been explicitly expressed ([1],[2],[3],[4],[7]).

A formula for computing π_i for the case $n > 2$ was needed after the review of the design of a sampling plan for a Pesticide Use Profile Survey in Pennsylvania [6]. The survey was undertaken in keeping with a commitment by the Department of Agriculture to the pollution abatement program in the Chesapeake Bay to assure that pesticide use practices in key Pennsylvania watersheds do not lead to pesticide problems for the Bay. Three watersheds were selected to be surveyed. The watersheds were the West Conewago Creek Watershed in Adams and York Counties, the Elk Creek Watershed in Chester County and the Pequea Creek Watershed in Lancaster County. The watersheds were selected based on their high priority status in Pennsylvania's Chesapeake Bay Program, diversity of agricultural enterprises, differing soil types, average farm size and other criteria. The Area Frame Section of the National Agricultural Statistics Service developed an area frame and selected the sample for this survey.

A general formula for computing inclusion probabilities for samples selected strictly proportional to size and strictly without replacement for a non-rotating sample design for any size sample is presented in this paper. A C program for computing inclusion probabilities, for $n \leq 5$ and $N \leq 100$, has been implemented on a Zilog Z8000 microcomputer. The program can easily be modified for other sample and population sizes.

EXAMPLE

Consider an ordered sample $(U_1,U_2, \cdots ,U_n)$. The probability of obtaining this sample is

$P(U_1) * P(U_2 \mid U_1) * P(U_3 \mid U_1 , U_2), \ldots , P(U_n \mid U_1 , \ldots , U_{n-1})$.

The probability that any unit U_i is selected is obtained by summing the probabilities of all samples of size n containing U_i. A procedure for computing inclusion probabilities and the computational complexity will now be illustrated for a simple example.

Let N=3 and n=2. All possible samples of size two, labelled $S_1,S_2,S_3,\ldots,S_6$, are:

S_1 - (U_1,U_2) S_2 - (U_1,U_3) S_3 - (U_2,U_1)

S_4 - (U_2,U_3) S_5 - (U_3,U_1) S_6 - (U_3,U_2).

We can compute the probability of selecting each sample as follows:

$P(S_1) = P(U_1) * P(U_2|U_1) = p_1 * p_2/(1-p_1)$

$P(S_2) = P(U_1) * P(U_3|U_1) = p_1 * p_3/(1-p_1)$

$P(S_3) = P(U_2) * P(U_1|U_2) = p_2 * p_1/(1-p_2)$

$P(S_4) = P(U_2) * P(U_3|U_2) = p_2 * p_3/(1-p_2)$

$P(S_5) = P(U_3) * P(U_1|U_3) = p_3 * p_1/(1-p_3)$

$P(S_6) = P(U_3) * P(U_2|U_3) = p_3 * p_2/(1-p_3)$.

The probability (π_1) that U_1 is included in the sample is the sum of the selection probabilities of all samples containing U_1.

$$\pi_1 = p_1 * p_2/(1-p_1) + p_1 * p_3/(1-p_1) +$$
$$p_2 * p_1/(1-p_2) + p_3 * p_1/(1-p_3)$$
$$\pi_1 = p_1 * ((p_2 + p_3)/(1-p_1)) + p_1 * (p_2/(1-p_2) + p_3/(1-p_3))$$
$$\pi_1 = p_1 * ((p_2 + p_3)/(p_2 + p_3)) + p_1 * (p_2/(1-p_2) + p_3/(1-p_3))$$
$$\pi_1 = p_1 + p_1 * \sum_{j\neq 1}^{3} p_j / (1-p_j).$$

The probability (π_2) that U_2 is included in the sample is the sum of the selection probabilities of all samples containing U_2.

$$\pi_2 = p_1 * p_2/(1-p_1) + p_2 * p_1/(1-p_2) +$$
$$p_2 * p_3/(1-p_2) + p_3 * p_2/(1-p_3)$$
$$\pi_2 = p_2 * ((p_1 + p_3)/(1-p_2)) + p_2 * (p_1/(1-p_1) + p_3/(1-p_3))$$
$$\pi_2 = p_2 * (p_1 + p_3)/(p_1 + p_3) + p_2 * \sum_{j\neq 2}^{3} p_j / (1-p_j)$$
$$\pi_2 = p_2 + p_2 * \sum_{j\neq 2}^{3} p_j / (1-p_j).$$

The probability (π_3) that U_3 is included in the sample is the sum of the selection probabilities of all samples containing U_3.

$$\pi_3 = p_1 * p_3/(1-p_1) + p_2 * p_3/(1-p_2) +$$
$$p_3 * p_1/(1-p_3) + p_3 * p_2/(1-p_3)$$
$$\pi_3 = p_3 * ((p_1 + p_2)/(1 - p_3)) + p_3 * (p_1/(1-p_1) + p_2/(1-p_2))$$
$$\pi_3 = p_3 * ((p_1 + p_2)/(p_1 + p_2)) + p_3 * \sum_{j\neq 3}^{3} p_j / (1-p_j).$$
$$\pi_3 = p_3 + p_3 * \sum_{j\neq 3}^{3} p_j / (1-p_j)$$

We can rewrite the π_i's (i=1,2,3) as follows:

$$\pi_i = p_i + p_i * \sum_{j\neq i}^{3} p_j / (1-p_j).$$

THEOREM

Let S be a sample of size n, drawn strictly without replacement and strictly with probabilities proportional to size (pps) from the finite population $U=\{U_1,U_2, \cdots ,U_N\}$, where the initial probabilities of selection are p_i for i=1,2, ... ,N (that is; $p_i = x_i / \sum_{i=1}^{N} x_i$ for i=1,2, ... N, where x_i is the size associated with the unit U_i). Then the probability π_i that the unit U_i is included in S is given by:

$$\pi_i = p_i + p_i * \sum_{\substack{i_2 \in I \\ i_2 \neq i}} p_{i_2}/(1 - p_{i_2}) +$$

$$p_i * \sum_{\substack{i_2 \in I \\ i_2 \neq i}} p_{i_2}/(1 - p_{i_2}) * \sum_{\substack{i_3 \in I \\ i_3 \neq i \\ i_3 \neq i_2}} p_{i_3}/(1 - p_{i_2} - p_{i_3})$$

$$+ \cdots +$$

$$p_i * \sum_{\substack{i_2 \in I \\ i_2 \neq i}} p_{i_2}/(1 - p_{i_2}) * \sum_{\substack{i_3 \in I \\ i_3 \neq i \\ i_3 \neq i_2}} p_{i_3}/(1 - p_{i_2} - p_{i_3}) * \cdots *$$

$$\sum_{\substack{i_n \in I \\ i_n \neq i \\ i_n \neq i_2 \\ i_n \neq i_3 \\ \vdots \\ i_n \neq i_{n-1}}} p_{i_n}/(1 - p_{i_2} - p_{i_3} - \cdots - p_{i_n}). \qquad (1)$$

where I is the index set {1, 2, 3, ... , N}.

PROOF

The theorem is clearly true for samples of size one, since π_i reduces to p_i for n=1. Using the Principle of Mathematical Induction we will assume the theorem holds for samples of size n and derive a formula for samples of size n+1.

The probability that the unit U_i is included in a sample of size n+1 from the set $U=\{U_1,U_2,U_3,\cdots,U_N\}$ is equivalent to the probability that U_i is selected on the first draw from U, plus the sum over $j \in (I-\{i\})$ of the product of the probabilities that U_j is drawn on the first draw from U and the probability that U_i is included in a sample of size n from $U-\{U_j\}$. Symbolically,

π_i = Pr[U_i is drawn on the first draw from U] +

$$\sum_{j \in (I-\{i\})} Pr[U_j \textit{ is drawn on the first draw from } U] *$$

$$\Pr[U_i \text{ is included in a sample of size n from } U-\{U_j\}]. \quad (2)$$

The revised selection probabilities associated with the set $U-\{U_j\}$ are

$$\frac{p_1}{1-p_j}, \frac{p_2}{1-p_j}, \cdots, \frac{p_{j-1}}{1-p_j}, \frac{p_{j+1}}{1-p_j}, \cdots, \frac{p_N}{1-p_j}.$$

The probability that U_i is included in a sample of size n from $U-\{U_j\}$ is obtained by setting $p_{i_x} = \frac{p_{i_x}}{1-p_j}$ in (1) to compute the last component of (2).

$$\pi_i = p_i + \sum_{j \in (I-\{i\})} p_j * [\frac{p_i}{(1-p_j)} + \frac{p_i}{(1-p_j)} * (\sum_{\substack{i_2 \in (I-\{j\}) \\ i_2 \neq i}} \frac{p_{i_2}}{(1-p_j)} / (1 - \frac{p_{i_2}}{(1-p_j)})) +$$

$$\frac{p_i}{(1-p_j)} * (\sum_{\substack{i_2 \in (I-\{j\}) \\ i_2 \neq i}} \frac{p_{i_2}}{(1-p_j)} / (1 - \frac{p_{i_2}}{(1-p_j)}) *$$

$$(\sum_{\substack{i_3 \in (I-\{j\}) \\ i_3 \neq i \\ i_3 \neq i_2}} \frac{p_{i_3}}{(1-p_j)} / (1 - \frac{p_{i_2}}{(1-p_j)} - \frac{p_{i_3}}{(1-p_j)})))$$

$$+ \cdots +$$

$$\frac{p_i}{(1-p_j)} * (\sum_{\substack{i_2 \in (I-\{j\}) \\ i_2 \neq i}} \frac{p_{i_2}}{(1-p_j)} / (1 - \frac{p_{i_2}}{(1-p_j)}) *$$

$$(\sum_{\substack{i_3 \in (I-\{j\}) \\ i_3 \neq i \\ i_3 \neq i_2}} \frac{p_{i_3}}{(1-p_j)} / (1 - \frac{p_{i_2}}{(1-p_j)} - \frac{p_{i_3}}{(1-p_j)}) *$$

$$(...(\sum_{\substack{i_n \in (I-\{j\}) \\ i_n \neq i \\ i_n \neq i_2 \\ \vdots \\ i_n \neq i_{n-1}}} \frac{p_{i_n}}{(1-p_j)} / (1 - \frac{p_{i_2}}{(1-p_j)} - \frac{p_{i_3}}{(1-p_j)} - \cdots \frac{p_{i_n}}{(1-p_j)}))) \cdots)]. \quad (3)$$

Rewriting (3),

$$\pi_i = p_i + p_i * \sum_{j \in (I-\{i\})} \frac{p_j}{(1-p_j)} +$$

$$p_i * \sum_{j \in (I-\{i\})} \frac{p_j}{(1-p_j)} * \Big(\sum_{\substack{i_2 \in (I-\{j\}) \\ i_2 \neq i}} p_{i_2}/(1 - p_j - p_{i_2}) \Big) +$$

$$p_i * \sum_{j \in (I-\{i\})} \frac{p_j}{(1-p_j)} * \Big(\sum_{\substack{i_2 \in (I-\{j\}) \\ i_2 \neq i}} p_{i_2}/(1 - p_j - p_{i_2}) *$$

$$\Big(\sum_{\substack{i_3 \in (I-\{j\}) \\ i_3 \neq i \\ i_3 \neq i_2}} p_{i_3}/(1 - p_j - p_{i_2} - p_{i_3}) \Big) \Big) \quad + \cdots +$$

$$p_i * \sum_{j \in (I-\{i\})} \frac{p_j}{(1-p_j)} * \Big(\sum_{\substack{i_2 \in (I-\{j\}) \\ i_2 \neq i}} p_{i_2}/(1 - p_j - p_{i_2}) *$$

$$\Big(\sum_{\substack{i_3 \in (I-\{j\}) \\ i_3 \neq i \\ i_3 \neq i_2}} p_{i_3}/(1 - p_j - p_{i_2} - p_{i_3}) *$$

$$\Big(\cdots \Big(\sum_{\substack{i_n \in (I-\{j\}) \\ i_n \neq i \\ i_n \neq i_2 \\ \vdots \\ i_n \neq i_{n-1}}} p_{i_n}/(1 - p_j - p_{i_2} - p_{i_3} - \cdots - p_{i_n}) \Big) \Big) \cdots \Big) . \quad (4)$$

Now make the substitutions $j=i_2$, $i_2=i_3$, $i_3=i_4$, $\cdots$,$i_n=i_{n+1}$ in (4). We have for samples of size n+1:

$$\pi_i = p_i + p_i * \sum_{\substack{i_2 \in I \\ i_2 \neq i}} p_{i_2}/(1-p_{i_2}) +$$

$$p_i * \sum_{\substack{i_2 \in I \\ i_2 \neq i}} p_{i_2}/(1-p_{i_2}) * \left(\sum_{\substack{i_3 \in I \\ i_3 \neq i \\ i_3 \neq i_2}} p_{i_3}/(1-p_{i_2}-p_{i_3}) \right) +$$

$$p_i * \sum_{\substack{i_2 \in I \\ i_2 \neq i}} p_{i_2}/(1-p_{i_2}) * \left(\sum_{\substack{i_3 \in I \\ i_3 \neq i \\ i_3 \neq i_2}} p_{i_3}/(1-p_{i_2}-p_{i_3}) * \right.$$

$$\left(\sum_{\substack{i_4 \in I \\ i_4 \neq i \\ i_4 \neq i_2 \\ i_4 \neq i_3}} p_{i_4}/(1-p_{i_2}-p_{i_3}-p_{i_4}) \right)) \quad + \qquad +$$

$$p_i * \sum_{\substack{i_2 \in I \\ i_2 \neq i}} p_{i_2}/(1-p_{i_2}) * \left(\sum_{\substack{i_3 \in I \\ i_3 \neq i \\ i_3 \neq i_2}} p_{i_3}/(1-p_{i_2}-p_{i_3}) * \left(\sum_{\substack{i_4 \in I \\ i_4 \neq i \\ i_4 \neq i_2 \\ i_4 \neq i_3}} p_{i_4}/(1-p_{i_2}-p_{i_3}-p_{i_4}) * \right. \right.$$

$$\left(\cdots \left(\sum_{\substack{i_{n+1} \in I \\ i_{n+1} \neq i_2 \\ i_{n+1} \neq i_3 \\ \vdots \\ i_{n+1} \neq i_n}} p_{i_{n+1}}/(1-p_{i_2}-p_{i_3}- \qquad p_{i_{n+1}}) \right) \right) \cdots \Big) \Big) \cdots \Big) .$$

SUMMARY

We have presented a general, computationally-efficient formula for computing inclusion probabilities when a sample of size n is selected strictly proportional to size and strictly without replacement for a non-rotating sample from a finite population. The formulation has been implemented on a Zilog Z8000 microcomputer for $n \leq 5$ and $N \leq 100$. The program is easy to modify for other sample and population sizes.

REFERENCES

[1] Brewer, K.R.W., and Hanif, M. (1983), Sampling With Unequal Probabilities, New York: Springer-Verlag.

[2] Cochran, W.G., (1977), Sampling Techniques, New York: John Wiley and Sons.

[3] Deming, W.E., (1960), Sample Design in Business Research, New York: John Wiley & Sons.

[4] Hansen, M.H., Hurwitz, W.N., and Madow, W.G., (1953), Sample Survey Methods and Theory, New York: John Wiley & Sons.

[5] Jessen, R.J. (1978), Statistical Survey Techniques, New York: John Wiley & Sons.

[6] Pennsylvania Agricultural Statistics Service, (1986), "1985 Pesticide Use Profile Survey," Harrisburg, Pa.

[7] Raj, D., (1968), Sampling Theory, New York: McGraw-Hill.