ON SUCCESSIVE SAMPLING AND FIXED INCLUSION PROBABILITIES

CHRISTOPHER R. H. HANUSA

Abstract. In the method of successive sampling, a sample of \( n \) distinct units is drawn from a population of \( N \) units one at a time. Each unit \( i \) is drawn with a fixed selection probability \( p_i \) and any repeated units are ignored. When the \( p_i \) are proportional to a given size measure, this is called probability proportional to size without replacement (PPSWOR).

The first order inclusion probabilities \( \pi_i \) of each unit in a PPSWOR scheme are not proportional to the size measure, as proved in Kochar and Korwar (2001). This article discusses a modification to the selection probabilities which creates a successive sampling scheme where first order inclusion probabilities are proportional to size (IPPS) in the case when \( n = 2 \).

1. Background

A sample of size \( n \) is to be taken from a finite population of \( N \) distinct units, \( \{u_1, \ldots, u_N\} \). In the method of successive sampling, we draw units one at a time from the population, with replacement, where each unit \( u_i \) is chosen with a fixed selection probability \( p_i \) \((\sum_{i=1}^{N} p_i = 1)\). Repeated elements are ignored and the draws stop when \( n \) distinct elements have been chosen. An example is provided in Section 2. For more on the theory of successive sampling and other sampling techniques, see Hájek’s [3]. We denote by \( p = (p_1, \ldots, p_N) \) the vector of selection probabilities.

For each unit \( u_i \) of the population, \( \pi_i \) denotes the probability that \( u_i \) is a member of a chosen sample, the first order inclusion probability of \( u_i \). Because each sample is of size \( n \), it follows that \( \sum_{i=1}^{N} \pi_i = n \). We denote by \( \pi = (\pi_1, \ldots, \pi_N) \) the vector of first order inclusion probabilities. Kochar and Korwar [4] show that \( p \) is equal to \( \pi / n \) only in the case when \( p_i \equiv 1/n; \) in all cases the vector \( \pi / n \) is majorized by \( p \). That is, after sorting \( \pi / n \) and \( p \) from largest to smallest, the respective partial sums satisfy \( \sum_{i=1}^{j} (\pi_i / n) \leq \sum_{i=1}^{j} p_i \) for \( 1 \leq j \leq N \).

An interesting case of successive sampling is when the probabilities \( p_i \) are proportional to some size measure \( x_i > 0 \) for each unit \( u_i \); we have \( p_i = x_i / \sum_{i=1}^{N} x_i \). This sampling technique called probability proportional to size without replacement (PPSWOR). Kochar and Korwar’s result shows that in all non-trivial examples of PPSWOR, the first order inclusion probabilities are not proportional to the size measure; this is borne out in our example in Section 2. Rao et al.’s [5] investigates the relationship between \( p \) and \( \pi \) when slight modifications are made to certain selection probabilities \( p_i \).

Sampling schemes in which the first order inclusion probabilities are proportional to the size measures are called IPPS, or inclusion probability proportional to size. One example
of an IPPS sampling scheme is the modified rejective sampling method of Sampford [7], a generalization of the ideas of Durbin [2]. Sampford’s method first selects one unit $u_i$ with probability $\pi_i/n$; all successive units $u_i$ are drawn with replacement with probabilities proportional to $\frac{\pi_i}{1-\pi_i}$. While the first order inclusion probabilities are proportional to the given size measure, the modification of selection probabilities after drawing the first unit seems less natural than the method of successive sampling.

Of recent interest also include the order sampling schemes (including the Pareto design) introduced by Rosén [6], in which the first order inclusion probabilities $\pi_i$ approximate desired inclusion probabilities. Traat et al. [8] discuss and compare Sampford, Pareto, and other sampling schemes, including their first order and second order inclusion probabilities and design correlations.

The origins of this letter are in mathematical models of social networks. The goal was a simple algorithm to model the arrival of a newcomer into a network; the newcomer would become mutual friends with a fixed-size subset of existing members with inclusion probability proportional to their popularity, as suggested by Barabási and Albert [1] and others. In this letter we investigate whether it is possible to modify the unit selection probabilities to guarantee that that the method of successive sampling gives samples of size two with certain first order inclusion probabilities. In Section 3 we give explicit formulas for modified probabilities in the case of $N = 3$, and in Section 4 we give a method which allows for calculation of such probabilities for $N > 3$.

2. Motivating Example

Suppose we wish to select a sample of size two from the population $\{u_1, u_2, u_3\}$, where $x_1 = 4$, $x_2 = 3$, and $x_3 = 2$. If we use successive sampling with probability vector $p = \left(\frac{4}{9}, \frac{3}{9}, \frac{2}{9}\right)$, then the first order inclusion probabilities are as follows.

- First, $u_1$, second, $u_2$: $\pi_1 = \frac{10}{21}$
- First, $u_2$, second, $u_1$: $\pi_1 = \frac{10}{21}$
- First, $u_3$, second, $u_2$: $\pi_1 = \frac{10}{21}$
- First, $u_2$, second, $u_3$: $\pi_1 = \frac{10}{21}$

We remark that $\pi_1/2$ is indeed majorized by $p$.

If instead we use successive sampling with the modified selection probabilities $p'_1 \approx 0.572$, $p'_2 \approx 0.263$, and $p'_3 \approx 0.165$, then the new first order inclusion probabilities are $\pi'_1 = \frac{8}{9}$, $\pi'_2 = \frac{8}{9}$, and $\pi'_3 = \frac{8}{9}$, as desired. The exact values for $p'_1$, $p'_2$, and $p'_3$ are roots of certain polynomials which are explicitly provided in Section 3.

3. Exact Modified Selection Probabilities When $n = 2, N = 3$

The existence of $p'_1$, $p'_2$, and $p'_3$ in the previous example is not a coincidence. When $N = 3$ and $n = 2$, it is always possible to find modified selection probabilities $p'_1$, $p'_2$, and $p'_3$ such that when successive sampling is performed with these selection probabilities, the desired first order inclusion probabilities are achieved.

**Theorem 1.** Suppose that $q_1$, $q_2$, and $q_3$ are the desired inclusion probabilities, real numbers between zero and one which satisfy $q_1 + q_2 + q_3 = 2$. Then there exist modified selection probabilities $p'_1$, $p'_2$, and $p'_3$, exact values for which are given below, such that when applying successive sampling to the population $\{u_1, u_2, u_3\}$ with these modified selection probabilities, the first order inclusion probabilities are indeed $\pi_1 = q_1$, $\pi_2 = q_2$, and $\pi_3 = q_3$. 
On Successive Sampling and Fixed Inclusion Probabilities

Theorem 2. Suppose that \( q_1 \) through \( q_N \) are the desired inclusion probabilities, real numbers between zero and one which satisfy \( q_1 + \cdots + q_N = 2 \). Then through the method described below, we can calculate modified selection probabilities \( p'_1 \) through \( p'_N \) such that when applying successive sampling to the population \( \{u_1, \ldots, u_N\} \) with these modified selection probabilities, the first order inclusion probabilities are indeed \( \pi_i = q_i \) for all \( i \).

Proof. Notice that for all \( i \), \( \pi_i = p'_i(1 + \sum_{j \neq i} \frac{p'_j}{1-p'_j}) \). If we let \( S = 1 + \sum_{i=1}^{N} \frac{p'_i}{1-p'_i} \), then we see that \( S = \frac{\pi_i}{p'_i} + \frac{1}{1-p'_i} \) for all \( i \). From this, we have \( N-1 \) equations \( \frac{\pi_i}{p'_i} + \frac{1}{1-p'_i} = \frac{\pi_j}{p'_j} + \frac{1}{1-p'_j} \).
for $2 \leq i \leq N$. Solving each equation for $p'_i$ and substituting the result into the equation
\[ p'_1 + \sum_{i=2}^{N} \left[ \left( \frac{p'_i + \pi_i}{2} + p'_i\pi_i - p'_i\pi_i - p'_i^2\pi_i}{2(p'_i + \pi_i - p'_i\pi_i)} \right) \right] = 1. \]

Back substitution gives us the values of $p'_i$ for $2 \leq i \leq N$. \hfill \square

For example, if $\pi = (\frac{6}{11}, \frac{4}{11}, \frac{4}{11}, \frac{3}{11}, \frac{3}{11}, \frac{2}{11})$, then solving Equation (4) for $p'_1$ gives $p'_1 \approx 0.2988$; consequently, $p' \approx (0.2988, 0.1788, 0.1788, 0.1297, 0.1297, 0.0842)$. We see once again that $\pi/2$ is majorized by $p'$.

5. Discussion

If we wish to evaluate our sampling scheme, we might aim to compare it alongside those in Traat et al. [8] in which samples of size 3 are chosen from a population of size 6, and where $\pi = (\frac{2}{3}, \frac{2}{3}, \frac{2}{3}, \frac{1}{3}, \frac{2}{3}, \frac{1}{3})$. By solving the appropriate system of equations, we can determine that a successive sampling scheme with modified selection probabilities of approximately $p' = (0.239, 0.239, 0.239, 0.095, 0.095, 0.095)$ will generate samples with the desired first order inclusion probabilities.

As such, in the comparison with the Sampford and Pareto designs, the first order inclusion probabilities for our and Sampford’s designs are exact, while for the Pareto design, $\pi_1 = \pi_2 = \pi_3 \approx 0.67232$ and $\pi_4 = \pi_5 = \pi_6 \approx 0.32768$. The second order inclusion probabilities are compared in Table 1 and the design correlations are compared in Table 2. We can see that all these comparison statistics for our method are in excellent agreement with those for the Sampford and Pareto schemes.

<table>
<thead>
<tr>
<th>$i, j = 1, 2, 3; i \neq j$</th>
<th>this article</th>
<th>Sampford</th>
<th>Pareto</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.404122</td>
<td>0.40252</td>
<td>0.41124</td>
</tr>
<tr>
<td>$i, j = 4, 5, 6; i \neq j$</td>
<td>0.070789</td>
<td>0.06918</td>
<td>0.06661</td>
</tr>
<tr>
<td>$i = 1, 2, 3; j = 4, 5, 6$</td>
<td>0.175030</td>
<td>0.17610</td>
<td>0.17405</td>
</tr>
</tbody>
</table>

Table 1. A comparison between our method and the Sampford and Pareto methods, of second order inclusion probabilities for samples $\{u_i, u_j\} \subset \{u_1, \ldots, u_6\}$.

<table>
<thead>
<tr>
<th>$i, j = 1, 2, 3; i \neq j$</th>
<th>this article</th>
<th>Sampford</th>
<th>Pareto</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-0.18145</td>
<td>-0.18868</td>
<td>-0.18506</td>
</tr>
<tr>
<td>$i, j = 4, 5, 6; i \neq j$</td>
<td>-0.18145</td>
<td>-0.18868</td>
<td>-0.18506</td>
</tr>
<tr>
<td>$i = 1, 2, 3; j = 4, 5, 6$</td>
<td>-0.21237</td>
<td>-0.20755</td>
<td>-0.20996</td>
</tr>
</tbody>
</table>

Table 2. A comparison between our method and the Sampford and Pareto methods, of design correlations for samples $\{u_i, u_j\} \subset \{u_1, \ldots, u_6\}$.

The calculations in Sections 3 and 4 show that it is possible to create a successive sampling scheme which will generate desired first order inclusion probabilities when $n = 2$. The calculations required are more computationally demanding than those of Sampford, which
ON SUCCESSIVE SAMPLING AND FIXED INCLUSION PROBABILITIES

means that a decision to implement this method should be weighed against the increase in computation time.

The author expects that future research in this subject will provide a successive sampling scheme which ensures desired first order inclusion probabilities when \( n > 2 \), as summarized by the following conjecture.

**Conjecture 3.** Suppose that \( \pi_1, \ldots, \pi_N \) are real numbers between zero and one which satisfy \( \pi_1 + \cdots + \pi_N = n \). Then there exist modified selection probabilities \( p'_1, \ldots, p'_N \) such that when applying successive sampling to the population \( \{u_1, \ldots, u_N\} \) with these modified selection probabilities, the first order inclusion probabilities are exactly \( \pi_1, \ldots, \pi_N \).

6. ACKNOWLEDGMENTS

I would like to thank my students Dana Warmsley and Michael Pandazis—Dana for asking a motivating question and Michael for questioning the motivation.

REFERENCES


Department of Mathematics, Queens College (CUNY), 6530 Kissena Blvd., Flushing, NY 11367, U.S.A., Phone: +1-718-997-5964

E-mail address: chanusa@qc.cuny.edu