A MODIFICATION OF POISSON SAMPLING

Dhiren Ghosh, Synectics for Management Decisions, Inc.
Andrew Vogt, Georgetown University
Andrew Vogt, Department of Mathematics, Georgetown University,
Washington, DC 20057-1233, vogt@math.georgetown.edu

Key Words: List selection probability, inclusion probability, Brouwer's Fixed Point Theorem.

1. Introduction

The list selection probability in a sampling scheme is the probability that a unit will be selected at some stage in the sampling regardless of whether it appears in the final sample. The inclusion probability is the probability that an individual unit will be in the final sample.

In Poisson sampling a list of population units is generated and successive units are independently included in the sample, the ith unit being chosen with probability p_i. This sampling scheme ensures that the list selection probabilities and inclusion probabilities are identical. Moreover, the joint inclusion probabilities are products of the individual inclusion probabilities since the selections are independent. These are the great advantages of Poisson sampling. The main shortcoming of Poisson sampling is that the sampling scheme may yield any sample size for the final sample from 0 to N, the population size.

In this paper we present a modification of Poisson sampling such that a fixed sample size n is achieved. This modification, however, destroys the equality of the list selection probabilities and the inclusion probabilities, and disrupts the simple multiplicative relationships of the joint inclusion probabilities. Nevertheless, given inclusion probabilities, list selection probabilities can be computed that will yield these inclusion probabilities. Also joint inclusion probabilities can be obtained by a similar computation. See Ghosh and Vogt (1998) for other related sampling schemes.

Poisson sampling and the modification both utilize the Horvitz-Thompson estimator for the population total. It can be shown that for the modification the usual Yates-Grundy-Sen estimator of the sampling variance (of the Horvitz-Thompson estimator) is always non-negative. For some sampling methods non-negativity has only been established empirically and/or for small sample sizes. See Cochran (1977), pp. 259-270.

2. The Modification

Our modification consists in applying list selection probabilities p_i to draw an initial sample, but rejecting the sample if its size differs from n, and drawing another sample using the same list selection probabilities. We continue this process until we obtain a sample of size n. The inclusion probability π_i of the i-th unit is a rational function of the list selection probabilities but in general is distinct from p_i. The novelty of this procedure is that list selection probabilities exist and can be found approximately which, through this rejection mechanism, yield any desired inclusion probabilities.

The equation below demonstrates the connection between the list selection probabilities and the inclusion probabilities for the modified procedure.

$$\pi_k = \frac{l_k \sum\{l_{i_1} l_{i_2} ... l_{i_{n-1}} : k \notin \{i_1, i_2, ..., i_{n-1}\} \subset \{1, ..., N\}\}}{\sum\{l_{i_1} l_{i_2} ... l_{i_n} : \{i_1, i_2, ..., i_n\} \subset \{1, ..., N\}\}}$$

$$= f_k(p_1, p_2, ..., p_N)$$

for $k = 1, 2, ..., N$. Here $l_i = \frac{p_i}{q_i}$ for $i = 1, 2, ..., N$ where p_i is the list selection probability of the i-th unit in the population, $q_i = 1 - p_i$, and π_k is the inclusion probability for the k-th unit in the population. Note that l_i is the odds ratio for list selection of the i-th unit.

For example, when the population size N is 6 and the desired sample size n is 3, then

$$\pi_1 = \frac{l_1 \sum\{l_i l_j : \{i, j\} \subset \{2, 3, 4, 5, 6\}\}}{\sum\{l_i l_j l_k : \{i, j, k\} \subset \{1, 2, 3, 4, 5, 6\}\}},$$

with similar formulas for $\pi_2, ..., \pi_6$.

To achieve desired inclusion probabilities π_k, an iterative process can be used to find the corresponding list selection probabilities p_k. We start with $p_k^{(1)} = \pi_k$. Then

$$\pi_k^{(1)} = f_k(p_1^{(1)}, ..., p_N^{(1)}).$$

We compare the $\pi_k^{(1)}$'s with the π_k's and make adjustments to the $p_k^{(1)}$'s to obtain $p_k^{(2)}$'s. According as $\pi_k^{(1)}$ is greater than or less than π_k, choose $p_k^{(2)}$ smaller than or larger than $p_k^{(1)}$. Do this for each k. Calculate the new $\pi_k^{(2)}$'s that result and repeat the process. In general the $\pi_k^{(j)}$'s converge to the π_k's as j increases and the $p_k^{(j)}$'s converge to the required list selection probabilities π_k. The iterative process can be performed by a computer program developed by the authors.

A mathematical proof that p_k's exist for each set of π_k's runs as follows. The function $f = (f_1, ..., f_N)$ defined above has three important properties:

i) f takes the set $S = \{p \in R^N : \sum_{k=1}^N p_k = n, 0 \leq p_k \leq 1 \text{ for each } k\}$ into itself;

ii) f is continuous;

iii) $f_k(p) = p_k$ whenever $p_k = 0$ or 1.

Assume that $1 \leq n \leq N$ and without loss of generality that π^* is a member of S with no component equal to 0 or 1. Consider the map $p \longmapsto p + \lambda(\pi^* - f(p))$. Here λ is the smaller of the numbers:

$$\inf\{\frac{p_i}{f_i(p) - \pi_i^*} : p \in S, f_i(p) > \pi_i^*, i = 1, 2, ..., N\}$$

and

$$\inf\{\frac{1 - p_i}{\pi_i^* - f_i(p)} : p \in S, \pi_i^* > f_i(p), i = 1, 2, ..., N\}.$$

It is easily seen from compactness of S and ii) and iii) that these two numbers are positive and thus that λ is likewise. Furthermore, the definition of λ guarantees that the map just defined takes S into S. Since this map is continuous, the Brouwer Fixed Point Theorem applies, and the map has a fixed point p^* in S. Accordingly, $f(p^*) = \pi^*$.

3. Some Properties

Under the modification the joint inclusion probability π_{jk} for the j-th and k-th population units, $j \neq k$, is:

$$\pi_{jk} =$$

$$\frac{l_j l_k \sum\{l_{i_1} l_{i_2} ... l_{i_{n-2}} : j, k \notin \{i_1, i_2, ..., i_{n-2}\} \subset \{1, ..., N\}\}}{\sum\{l_{i_1} l_{i_2} ... l_{i_n} : \{i_1, i_2, ..., i_n\} \subset \{1, ..., N\}\}}.$$

It can be shown by algebraic manipulation that this expression is strictly smaller than $\pi_j \pi_k$.

The Horvitz-Thompson estimator for the population total is:

$$\sum_{k=1}^N \frac{y_k}{\pi_k} I_k$$

where I_k is the variable taking the value 1 if the k-th unit is in the final sample and 0 otherwise, and y_k is the value of the variable of interest for the k-th unit. Then, as in Cochran (1977), p. 261, the Yates-Grundy-Sen estimator of the variance of the Horvitz-Thompson estimator is:

$$\sum_{1 \leq j < k \leq N} \frac{\pi_j \pi_k - \pi_{jk}}{\pi_j \pi_k} (\frac{y_j}{\pi_j} - \frac{y_k}{\pi_k})^2 I_j I_k.$$

It follows from the above that this is nonnegative.

References

Cochran, W. G. (1977). Sampling Techniques. John Wiley & Sons, Inc. New York, Third edition.

Ghosh, D. and A. Vogt (1998). Rectification of sample size in Bernoulli and Poisson sampling. Proceedings of the American Statistical Association, Survey Research Methods Section.