SAMPLING METHODS RELATED TO BERNOULLI AND POISSON SAMPLING

Dhiren Ghosh
Synectics for Management Decisions
1901 North Moore Street, Suite 900
Arlington, VA 22209

Andrew Vogt
Department of Mathematics
Georgetown University
Washington, DC 20057-1233
vogt@math.georgetown.edu

KEYWORDS:, Generalized Binomial, Generalized Bernoulli, Generalized Poisson, Rejective Sampling, Scaling

Introduction

Bernoulli sampling is a method of sampling in which all members of the population have the same probability of selection and the inclusion variables are jointly independent. Poisson sampling removes the restriction of equiprobability, allowing the inclusion probabilities for each member to be distinct. Because of the joint independence it is easy to calculate variances for estimators based on Bernoulli and Poisson sampling. However, sample size is a variable under these methods and can vary in principle from 0 to the population size N.

Ways around this difficulty are discussed here. One of them is rejective sampling, or so-called conditional Poisson sampling, in which Poisson sampling is performed but the sample is rejected unless the desired sample size is achieved. Rejective sampling can be performed in such a way as to obtain stipulated inclusion probabilities for each population member. Another alternative is to reject the sample if the size is smaller than desired and otherwise randomly trim some elements from the obtained sample to achieve the desired size. Another method is to adjust the inclusion probabilities, for example, by a scale factor, to enlarge or shrink a previously determined sample. A number of other methods go under such names as modified Poisson, collocated sampling, list sequential sampling, sequential updating, and the methods of Rao and Hajek.

The object of these methods is to regularize the sample size. But regularization disrupts joint independence since if the sample size is at all restricted, the inclusion variables cannot be jointly independent. The ostensible reason for seeking joint independence is to make variance calculations straight-forward. However, joint independence is sufficient but not necessary. Straight-forward variance calculations can be done with only pairwise independence of the inclusion variables.

Thus we propose to require that the inclusion probabilities have whatever values are stipulated but that the inclusion variables only need be pairwise independent. In the case of equal inclusion probabilities, this requirement is easily satisfied, along with other natural requirements, and yields a sampling methodology that we call generalized Bernoulli sampling. This type of sampling can be done with as few as two or three permitted sample sizes. In the case of unequal inclusion probabilities, we call our method generalized Poisson sampling, and it too allows reduction of sample size.

Bernoulli and Poisson sampling

In Bernoulli sampling each element of the population has the same probability of selection π and the inclusion variables I_k, k = 1, ..., N are independent. (I_k = 1 if the k-th element is in the sample, 0 otherwise.) Indeed, $E(I_k) = P(I_k = 1) = \pi$. The sample size $\underline{n} = \Sigma\, I_k$ is itself an integer-valued random variable. In fact it is a binomial random variable with mean $N\pi$ and variance $N\pi(1-\pi)$.

Poisson sampling allows the inclusion probabilities to be variable, and thus $E(I_k) = \pi_k$ for k = 1, ..., N. Then the sample size $\underline{n}$ has mean $\Sigma\, \pi_k$ and variance $\Sigma\pi_k(1 - \pi_k)$. A common method of realizing a Poisson sample is to calculate the values of N independent uniform random variables U_1, ..., U_N on [0,1] and to include the k-th population element in the sample if and only if the value of U_k is less than or equal to π_k.

Conditional Poisson sampling

In conditional Poisson sampling, one applies the Poisson methodology but rejects the results unless the desired sample size is achieved. Hence this is often called rejective sampling. If one starts with inclusion probabilities p_k and rejects the sample unless the sample size is n (where n perhaps is the integer closest to $\Sigma\, p_k$), then the effective inclusion probabilities are:

$$\pi_k = \frac{s_k \Sigma_{i1,i2,\ldots,i_{n-1}} \{s_{i1}s_{i2}\cdots s_{i_{n-1}} : k \notin \{i_1,i_2,\ldots,i_{n-1}\} \subset \{1,2,\ldots,N\}\}}{\Sigma_{i1,i2,\ldots,i_n} \{s_{i1}s_{i2}\cdots s_{i_n} : \{i_1,i_2,\ldots,i_n\} \subseteq \{1,2,\ldots,N\}\}}$$
$$= f_k(p_1,p_2,\ldots,p_N)$$

where $s_k = p_k/(1 - p_k)$ for k = 1, ..., N. The Brouwer Fixed Point Theorem can be applied to show that there exist p_i's yielding any desired π_i's (the latter assumed to be between 0 and 1 and summing to n). An iterative process can be used to find these p_i's. We start with $p_i^{(1)} = \pi_i$. Then

$$\pi_i^{(1)} = f_i(p_1^{(1)}, \ldots, p_N^{(1)})$$

We compare the $\pi_i^{(1)}$'s with the π_i's and make adjustments to the $p_i^{(1)}$'s to obtain $p_i^{(2)}$'s. According as $\pi_i^{(1)}$ is greater than or less than π_i, choose $p_i^{(2)}$ smaller than

or larger than $p_i^{(1)}$. Do this for each i. Calculate the new $\pi_i^{(2)}$'s that result and repeat the process. In general the $\pi_i^{(j)}$'s converge to the π_i's as j increases, and the $p_i^{(j)}$'s converge to the corresponding p_i's. The iterative process can be performed by a computer program developed by the authors that works well on small populations.

Any procedure that restricts the sample size affects the joint inclusion probabilities and makes them no longer equal to the product of the individual inclusion probabilities. In particular, the pairwise inclusion probabilities $\pi_{jk} = P(I_j = 1 \text{ and } I_k = 1)$ are given by:

$$\pi_{jk} = \frac{s_j s_k \sum_{i_1,i_2,\ldots,i_{n-2}} \{s_{i_1} s_{i_2} \cdots s_{i_{n-2}} : j,k \notin \{i_1,i_2,\ldots,i_{n-2}\} \subset \{1,2,\ldots,N\}\}}{\sum_{i_1,i_2,\ldots,i_n} \{s_{i_1} s_{i_2} \cdots s_{i_n} : \{i_1,i_2,\ldots,i_n\} \subseteq \{1,2,\ldots,N\}\}}$$

where s_j is as above in terms of the p_i's that correspond to the π_i's .

Generalized Bernoulli and generalized Poisson sampling

Generalized Bernoulli sampling is a sampling method in which: (1) $E(I_k) = \pi$ for each k, (2) the inclusion variables are pairwise independent, (3) the conditional mean $E(I_k/\underline{n})$ is independent of k, and (4) the conditional mean $E(I_k/(I_j \text{ and } \underline{n}))$ is independent of k ($k \neq j$). It can be shown that these four conditions are independent of each other. In effect they constrain binary relationships between inclusion variables to ensure the maximum symmetry among population elements. Furthermore, it can be shown that $E(I_k/\underline{n}) = \underline{n}/N$ and that $E(I_k/(I_j \text{ and } \underline{n})) = (\underline{n} - I_j)/(N-1)$. For generalized Bernoulli sampling, $E(\underline{n}) = N\pi$ and $V(\underline{n}) = N\pi(1-\pi)$. Indeed, $\underline{n}$ is a *generalized binomial random variable*, i.e., a variable whose possible values are integers in the set {0, 1, 2, ..., N} and whose mean and variance are the same as those of a binomial random variable.

A generalized Bernoulli sampling scheme is easily realized. Take a generalized binomial random variable. This assigns probabilities to each sample size from 0 to N (and one can take the permissible sample sizes to be as small as 2 or 3 provided $|2\pi - 1| \leq [1 - (4/N)]^{(1/2)}$). Choose a sample size according to these probabilities, and then given the sample size $\underline{n}$, choose a random sample of size $\underline{n}$ from the population.

For a *generalized Poisson sampling* scheme, we require that (1) $E(I_k) = \pi_k$ for k = 1, ..., N and (2) the inclusion variables are pairwise independent. It can be shown in this case, as in ordinary Poisson sampling, that the sample size $\underline{n}$ has mean $\Sigma \pi_k$ and variance $\Sigma \pi_k(1 - \pi_k)$. Since generalized Bernoulli schemes are special cases of generalized Poisson schemes, we have nontrivial examples.

Because of pairwise independence, it follows that $\pi_{jk} = \pi_j \pi_k$. Then the usual estimator for a population total $\sum$ Yj, namely, $\sum$ (YjIj)/π_j , which has zero bias, is easily seen to have variance of the form:

$$\sum_{j=1}^{N} \frac{Y_j^2 (1-\pi_j)}{\pi_j}.$$

If a population consists of subpopulations of sizes N_1, N_2, ..., N_k and in the i-th subpopulation the inclusion probability is π_i for each element, then generalized Bernoulli sampling can be performed on each subpopulation to accomplish generalized Poisson sampling on the combined population.

Poisson Scaling

Another way to modify Poisson sampling to achieve one sample size (but note that this also sacrifices pairwise independence of inclusion variables) is by introduction of a scale factor.

As noted above, one way to perform Poisson sampling is to select a random number U_j uniformly from [0,1] and include the j-th element of the population if and only if $U_j \leq \pi_j$ for j = 1, 2, ..., N.

If we record the values of U_1, U_2, ..., U_N and discover that our sample size is either too large or too small, we can introduce a scale factor λ and revisit our selections, this time including the j-th element if and only if $U_j \leq \lambda\pi_j$, where λ is chosen to give us exactly the desired sample size. Indeed λ is an order variable and is given by:

λ = n-th smallest of U_1/π_1, U_2/π_2, ..., U_N/π_N.

References

Chaudhuri, A. and Vos, J. W. E. (1988). Unified Theory and Strategies of Survey Sampling. North Holland. Amsterdam.

Ghosh, D. and Vogt, A. (1998). Rectification of sample size in Bernoulli and Poisson sampling. Proceedings of the American Statistical Association, Survey Research Methods Section, 448-450.

Ghosh, D. and Vogt, A. (1999). A modification of Poisson sampling. Proceedings of the American Statistical Association, Survey Research Methods Section, 198-199.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). Model Assisted Survey Sampling. Springer-Verlag. New York.