RECTIFICATION OF SAMPLE SIZE IN BERNOULLI AND POISSON SAMPLING

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

Key Words: Sampling Schemes, Inclusion Probabilities, Π PS Sampling.

Abstract:

When a sample of size n is to be drawn from a population of size N so that the i-th unit has probability p_i of inclusion, where $p_1 + ... + p_N = n$, there is some prospect of obtaining a sample that is either too small or too large. Techniques are examined for ensuring that the sample size is n or close to n while retaining advantages of Bernoulli and Poisson sampling.

1. Introduction

The objective of this paper is to examine modifications of Bernoulli and Poisson sampling that restrict the sample size. The advantages of Bernoulli and Poisson sampling are that the inclusion probabilities are known for each member of the population and these probabilities are pairwise independent. Independence permits easy calculation of variances of sampling statistics. In a population of size N, however, the possible sample size ranges from 0 to N for both kinds of sampling. In some cases the resulting sample will be unacceptable: too small for reasonable inferences or too large for available resources. If the sample size is restricted to values close to a fixed integer n - this can be accomplished in various ways, we study the effect on the inclusion probabilities and independence.

Examples of ways to restrict the sample size include the following:

1) Use the Bernoulli/Poisson sampling technique - i. e., include the i-th member of the population in the sample when a random number drawn uniformly from 0 to 1 is less than or equal to the preassigned inclusion probability for that member. But modify this technique by discarding the resulting sample if the sample size does not satisfy a given constraint. Then resample in the same manner until a sample is achieved that meets the requirement.

2) Use the Bernoulli/Poisson technique, but if the sample is too large, randomly select some units from the sample and remove them. If the sample size is too small, discard the sample.

3) Use 2) but if the sample size is too small, instead of discarding the sample, randomize the order of the unselected units in the population and randomly choose some of them with probabilities proportional to the original inclusion probabilities until the sample size is satisfactory.

We show below that the two chief characteristics of Bernoulli and Poisson sampling - namely, predetermined inclusion probabilities and pairwise independence - imply that sample size is a variable with a calculable variance. However, this does not preclude very strong restrictions on the permissible sample sizes.

Let I_k for $1 \leq k \leq N$ be indicator variables for the members of the population - i. e., $I_k = 1$ or 0 according as element k is or is not in the sample. Let π_k be the probability of inclusion of the k-th member of the population in the sample. Without loss of generality we can assume that each π_k satisfies: $0 < \pi_k < 1$. Let $\hat{n}$ be the actual sample size. Then $\hat{n} = \sum_{k=1}^{N} I_k$. Since $E(I_k) = \pi_k$ for each k, $E(\hat{n}) = \sum_{k=1}^{N} \pi_k$. We denote this sum by n. It is the desired sample size and is usually taken to be a positive integer. In the case of Bernoulli sampling π_k is the same for all k and equals $\frac{n}{N}$.

It is easily seen that any restriction on sample size prohibits independence of the indicator variables.

PROPOSITION 1: Suppose that the variable $\hat{n}$ cannot take some integer value. Then the variables $I_1, I_2, ..., I_N$ are dependent.

Proof: Let n_0 be a forbidden value of n. If the variables are independent, their joint distribution is the product of the individual distributions. Then $0 = P(I_1 = 1, I_2 = 1, ..., I_{n_0} = 1, I_{n_o+1} = 0, ..., I_N = 0) = P(I_1 = 1)P(I_2 = 1)...P(I_{n_0} = 1)P(I_{n_0+1} = 0)...P(I_N = 0) = \pi_1\pi_2...\pi_{n_0}(1 - \pi_{n_0+1})...(1 - \pi_N)$.

But this is impossible. Q. E. D.

2. Generalized Bernoulli Sampling

A sampling scheme on a population of size N is called a generalized Bernoulli sampling scheme provided: i) $E(I_k)$ is the same for all k and this number is in the open interval $(0,1)$; ii) $E(I_k/\hat{n})$ is the same for all k; iii) for each k $E(I_j/I_k,\hat{n})$ is the same for all j, $j \neq k$; and iv) for each j,k, $j \neq k$, I_j and I_k are pairwise independent.

Simple random sampling satisfies i), ii), and iii). Ordinary Bernoulli sampling satisfies i) through iv). Because the variables I_j and I_k only assume two values, condition iv) is equivalent to: $P(I_j = 1 \text{ and } I_k = 1) = P(I_j = 1)P(I_k = 1)$. The latter in turn is equivalent to: $E(I_jI_k) = E(I_j)E(I_k)$.

PROPOSITION 2: For a sampling scheme
i) holds if and only $E(I_k) = \frac{n}{N}$.
ii) holds if and only if $E(I_k/\hat{n}) = \frac{\hat{n}}{N}$.
iii) holds if and only if $E(I_j/I_k,\hat{n}) = \frac{\hat{n}-I_k}{N-1}$.

THEOREM 1: For any generalized Bernoulli sampling scheme,

$$V(\hat{n}) = E(\hat{n})(1 - \frac{E(\hat{n})}{N}). \quad (1)$$

Furthermore, if a sampling scheme satisfies i), ii), iii) and (1), then it satisfies iv) and is a generalized Bernoulli sampling scheme.

COROLLARY: In generalized Bernoulli sampling there must be at least two sample sizes.

Indeed, given a mean μ and a standard deviation σ, the integer-valued variable $\hat{n}$ can be taken to assume three values n_1, n_2, n_3 with $n_1 < n_2 < n_3$ provided $n_1 = \mu - d, n_2 = \mu + f$, and $n_3 = \mu + e$ where d and e are positive and $de > \sigma^2 > \max\{df, -ef\}$. This means that sample size can be restricted to three values - roughly, $n - \sqrt{n}, n, n + \sqrt{n}$ - when the finite population correction factor is ignored.

It is even possible to restrict to two sample sizes: $n_1 = \mu - d, n_2 = \mu + e$ with $ed = \sigma^2$ although this condition cannot always be fulfilled with integer values.

The strategy then is to choose sample sizes as above and then conditional upon the sample size devise a sampling scheme based on the probabilities in Proposition 2. Once the sample size is determined, simple random sampling can be done since simple random sampling with given sample size $\hat{n}$ yields the probabilities in Proposition 2. However, there are other choices. For example, when the population size is 7 and the sample size is 3, randomly choosing one of the seven samples $123, 145, 167, 246, 257, 347, 356$ out of 35 possible samples is a sampling scheme consistent with Proposition 2.

3. Generalized Poisson Sampling

A sampling scheme on a population of size N is called a generalized Poisson sampling scheme provided: i) $E(I_k) = \pi_k$ for all k where π_k is in the interval $(0,1)$ for all k; and ii) I_j and I_k are independent for all j,k, $j \neq k$.

Generalized Bernoulli schemes are included in this family, and so is ordinary Poisson sampling.

PROPOSITION 3: In a generalized Poisson sampling scheme the sample size $\hat{n}$ satisfies:

$$\hat{n} \leq \min\{\frac{E(\hat{n})}{\pi_k} : k = 1, ..., N\}.$$

Furthermore, for $j \neq k$, $P(I_j = 1/I_k,\hat{n}) = f(k, I_k, \hat{n}, \pi_k)$ if and only if all π_k's are the same.

THEOREM 2: For any generalized Poisson sampling scheme

$$V(\hat{n}) = \sum_{k=1}^{N} \pi_k(1 - \pi_k).$$

COROLLARY: In generalized Poisson sampling there must be at least two sample sizes.

As in the case of Bernoulli sampling one may in principle realize a generalized Poisson sampling scheme using two or three sample sizes grouped around $n = E(\hat{n})$. We omit details.

4. A Variant of Poisson Sampling

If one ignores the injunction against a single sample size and attempts to implement a scheme in which Poisson sampling is performed with the sample discarded unless the sample size equals n, one obtains results that we now describe.

This time let $\{p_k\}$ denote the inclusion probabilities of the original Poisson scheme. They form a

vector $p = (p_1, ..., p_N)$. If all samples are discarded except those of size n, then

$$\pi_k = P(I_k = 1) = \frac{p_k \sum\{P_S(1-P)_T : (S,T) \text{ satisfies } (N)\}}{\sum\{P_S(1-P)_T : (S,T) \text{ satisifies } (D)\}}$$

where $P_S = \prod_{k \in S} p_k$; $(1-P)_T = \prod_{k \in T}(1-p_k)$; (N) is the condition that $|S| = n-1$, $S \cap T = \emptyset$, and $S \cup T = \{1, ..., \hat{k}, ..., N\}$; (D) is the condition that $|S| = n$, $S \cap T = \emptyset$, and $S \cup T = \{1, ..., N\}$; and $\hat{k}$ indicates that k is omitted. The resultant inclusion probability vector $\pi = (\pi_1, ..., \pi_N)$ satisfies $\sum_k \pi_k = n$, and this is true even if $\sum_k p_k \neq n$.

THEOREM 3: The map $p \mapsto \pi$ has the following properties:

i) if $p_j < p_k$ then $\pi_j < \pi_k$;

ii) $\frac{\partial \pi_k}{\partial p_k} > 0$ for all k; and

iii) this map takes the set

$$\{p : 0 \leq p_k \leq 1 \text{ for all } k, \sum_k p_k = n\}$$

onto itself.

Thus computers can be used to generate a prescribed π at least approximately, perhaps by starting with a value of p near π and tweaking the vector. Although our previous theorems indicate that I_j amd I_k are dependent, formulas akin to the formulas above for π in terms of p can be used to compute the covariances $V(I_j, I_k)$ and the variances of the sample statistics. There is even the prospect of reduced overall variance.

REFERENCES

K. R. W. Brewer and M. Hanif, Sampling with Unequal Probabilities, Lecture Notes in Statistics 15, Springer-Verlag, New York, 1983.

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