fast cube implementation

Raphaël Jauslin[a], Esther Eustache[a] and Yves Tillé[a]

Abstract

Key words: optimal design, spread sampling, stratification

[a]Institute of statistics, University of Neuchatel, Av. de Bellevaux 51, 2000 Neuchatel, Switzerland
(E-mail: raphael.jauslin@unine.ch)

1 Introduction

2 Notation

Consider a finite population U of size N whose units can be defined by labels $k \in \{1, 2, \ldots, N\}$. Let $\mathcal{S} = \{s | s \subset U\}$ be the set of all possible samples. A sampling design is defined by a probability distribution $p(.)$ on $\mathcal{S}$ such that

$$p(s) \geq 0 \text{ for all } s \in \mathcal{S} \text{ and } \sum_{s \in \mathcal{S}} p(s) = 1.$$

A random sample S is a random vector that maps elements of $\mathcal{S}$ to an N vector of 0 or 1 such that $\mathrm{P}(S = s) = p(s)$. Define $a_k(S)$, for $k = 1, \ldots, N$:

$$a_k = \begin{cases} 1 & \text{if } k \in S \\ 0 & \text{otherwise.} \end{cases}$$

Then a sample can be denoted by means of a vector notation: $\mathbf{a}^\top = (a_1, a_2, \ldots, a_N)$. For each unit of the population, the inclusion probability $0 \leq \pi_k \leq 1$ is defined as the probability that unit k is selected into sample S:

$$\pi_k = \mathrm{P}(k \in S) = \mathrm{E}(a_k) = \sum_{s \in S | k \in s} p(s), \text{ for all } k \in U.$$

Let $\boldsymbol{\pi}^\top = (\pi_1, \ldots, \pi_N)$ be the vector of inclusion probabilities. Then, $\mathrm{E}(\mathbf{a}) = \boldsymbol{\pi}$. Let also $\pi_{k\ell}$ be the probability of selecting the units k and ℓ together in the sample, with $\pi_{kk} = \pi_k$. The matrix of second-order inclusion probabilities is given by $\Pi = \mathrm{E}(\mathbf{a}\mathbf{a}^\top)$. The sample is generally selected with the aim of estimating some population parameters. Let y_k denote a real number associated with unit $k \in U$, usually called the variable of interest. For example, the total

$$Y = \sum_{k \in U} y_k$$

can be estimated by using the classical Horvitz-Thompson estimator of the total defined by

$$\widehat{Y}_{HT} = \sum_{k \in U} \frac{y_k a_k}{\pi_k}. \tag{1}$$

3 Balanced Sampling

Usually, some auxiliary information $\mathbf{x}_k^\top = (x_{k1}, x_{k2}, \ldots, x_{kq}) \in \mathbb{R}^q$ regarding the population units is available. A sampling design is said to be balanced on the

auxiliary variables x_k if and only if it satisfies the balancing equations

$$\widehat{\mathbf{X}} = \sum_{k \in S} \frac{\mathbf{x}_k}{\pi_k} = \sum_{k \in U} \frac{\mathbf{x}_k a_k}{\pi_k} = \sum_{k \in U} \mathbf{x}_k = \mathbf{X}.$$

Sometimes it is not possible to select a sample that satisfies exactly the constraint. We write $\widehat{\mathbf{X}} \approx \mathbf{X}$ to notice that the sample is approximately balanced. In many applications, inclusion probabilities are such that samples have a fixed size n. A sampling design of fixed size can be viewed as balanced on only one auxiliary variable $x_k = \pi_k$. Indeed, we have mathematically,

$$\sum_{k \in S} \frac{x_k}{\pi_k} = \sum_{k \in S} \frac{\pi_k}{\pi_k} = n_S.$$

Let denote the set of all samples that have fixed size equal to n by

$$\mathcal{S}_n = \left\{ \mathbf{a} \in \{0,1\}^N \;\middle|\; \sum_{k=1}^{N} a_k = n \right\}.$$

More generally, we write the problem of selecting a balanced sample by the following linear system :

$$\begin{cases} \displaystyle\sum_{k \in U} \frac{\mathbf{x}_k a_k}{\pi_k} = \sum_{k \in U} \frac{\mathbf{x}_k}{\pi_k} \pi_k \\ a_k \in \{0,1\}, \; k \in U. \end{cases}$$

Or also written in matrix form,

$$\mathbf{A}\mathbf{a} = \mathbf{A}\boldsymbol{\pi}, \tag{2}$$

where $\mathbf{A} = \left(\frac{\mathbf{x}_1}{\pi_1}, \ldots, \frac{\mathbf{x}_N}{\pi_N} \right)$. The aim consist then of obtaining a sample $\mathbf{a}$ that satisfies the constraints.

4 Cube Method

Deville and Tillé (2004) developed the cube method. It selects a sample that is balanced and respect the inclusion probabilities. The method can take equal or unequal inclusion probabilities. A each step , vector $\boldsymbol{\pi}$ is randomly modified. The subspace induced by the linear system (2) is defined by the following,

$$\begin{aligned} \mathcal{A} &= \left\{ \mathbf{a} \in \mathbb{R}^N | \mathbf{A}\mathbf{a} = \mathbf{A}\boldsymbol{\pi} \right\} \\ &= \boldsymbol{\pi} + \mathrm{Null}(\mathbf{A}), \end{aligned}$$

where $\text{Null}(\mathbf{A}) = \{u \in \mathbb{R}^N | \mathbf{Au} = 0\}$. The idea is then to use a vector of the null space of $\mathbf{A}$ such that we ensure to have martingale property of the updated inclusion probabilities. More specifically we have the following equation,

$$\text{E}_p(\boldsymbol{\pi}^t | \boldsymbol{\pi}^{t-1}) = \text{E}_p(\boldsymbol{\pi}^{t-1}), \text{ for all t} = 1,\ldots, \text{N}.$$

At each step, at least one component is set to 0 or 1. Matrix A is updated from the new inclusion probabilities. This step is repeated until there is only one component that is not equal to 0 or 1. Algorithm 1 present the full picture of the method. Chauvet and Tillé (2006) have improved the time consuming cost by using a sub-matrix of smaller size to find a vector that is inside of the null space of $\mathbf{A}$. In the next section we present the proposed strategy to improved even more this cost.

5 Reduction

6 Simulation

7 Discussion

References

Chauvet, G. and Tillé, Y. (2006). A fast algorithm of balanced sampling. *Journal of Computational Statistics*, 21:9–31.

Deville, J.-C. and Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, 91:893–912.

Algorithm 1 fast flight phase of the cube Method

Calculate at first i the number of inclusion probabilities that are not equal to 0 or 1. Let $\boldsymbol{\pi}$ be equal to the i corresponding inclusion probabilities and initializing $\boldsymbol{\pi}^1$ by $\boldsymbol{\pi}$. For $t = 1, \ldots, N$, we repeat :

1. Find $\widetilde{\boldsymbol{\pi}}^t$ the first J entries of the inclusion probabilities $\boldsymbol{\pi}^t$, where $J = \min(p+1, i)$. Define $\mathbf{B}$ as the J corresponding rows of the matrix A. Notice that the matrix $\mathbf{B}$ is either a $(p+1) \times p$ matrix or a $i \times p$ matrix.

2. Find a non null vector $\widetilde{\mathbf{u}}^t$ inside of the null space of $\mathbf{B}$. Define $\mathbf{u}^t$ as the expanded null vector such that $u_k^t = 0$ for all entry that is not equal to the corresponding J values.

3. Calculate $\widetilde{\lambda}_1^t$ and $\widetilde{\lambda}_2^t$ the two greater value such that

$$\begin{array}{rcl} 0 & \leqslant & \pi_k^t + \lambda_1^t u_k^t \leqslant 1, \\ 0 & \leqslant & \pi_k^t - \lambda_2^t u_k^t \leqslant 1, \end{array} \quad \text{for all } k \in U$$

   Observe that λ_1^t and λ_2^t are both greater than 0.

4. Update the inclusion probabilities using the rules :

$$\boldsymbol{\pi}^{t+1} = \begin{cases} \boldsymbol{\pi}^t + \widetilde{\lambda}_1^t \mathbf{u}^t & \text{with probability} \quad q_1^t \\ \boldsymbol{\pi}^t - \widetilde{\lambda}_2^t \mathbf{u}^t & \text{with probability} \quad q_2^t \end{cases}$$

   where $q_1^t = \widetilde{\lambda}_2^t / (\widetilde{\lambda}_1^t + \widetilde{\lambda}_2^t)$ and $q_2^t = \widetilde{\lambda}_1^t / (\widetilde{\lambda}_1^t + \widetilde{\lambda}_2^t)$.

5. Update i the number of inclusion probabilities not equal to 0 or 1.

We repeat these steps until it is no more possible to find a vector $\widetilde{\mathbf{u}}^t$ that is inside of the null space.