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 $\Pi = \mathrm{E}(\mathbf{a}\mathbf{a}^\top)$. In many applications, inclusion probabilities are such that samples have a fixed size n. Let the set of all samples that have fixed size equal to n be defined by

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

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

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

Usually, some auxiliary information $\mathbf{x}_k^\top = (x_{k1}, x_{k2}, \dots, x_{kq}) \in \mathbb{R}^q$ regarding the population units is available. In the particular case of spatial sampling, a set of spatial coordinates $\mathbf{z}_k^\top = (z_{k1}, z_{k2}, \dots, z_{kp}) \in \mathbb{R}^p$ is supposed to be available, where p is the dimension of the considered space. A sampling design is said to be balanced on the auxiliary variables x_k if and only if it satisfies the balancing equations

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

3 Simulation

4 Discussion