Generalization of systematic sampling

Raphaël Jauslin[a] and Yves Tillé[a]

Abstract

In this paper we propose a definition of multidimensional systematic sampling from a finite spatial population with equal or unequal probabilities.

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 $\mathbf{x}_k^\top = (x_{k1}, x_{k2}, \ldots, x_{kq}) \in \mathbb{R}^q$ regarding the population units is available. In the particular case of spatial sampling, a set of spatial coordinates $\mathbf{z}_k^\top = (z_{k1}, z_{k2}, \ldots, z_{kp}) \in \mathbb{R}^p$ 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 Systematic sampling

Algorithm 1 Algorithm for systematic sampling

Let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)$. Generate a random start u uniformly distributed $\mathcal{U}(0,1)$. Let $a, b \in \mathbb{R}$, for $k = 1, \ldots, N$ repeat the following procedure:

1. $b = a$.
2. $a = a + \pi_k$.
3. if $\lfloor a \rfloor \neq \lfloor b \rfloor$ select k.

4 Generalization of systematic sampling

4.1 Strata

In order to generalize a systematic sampling in more than one dimension, we need to properly define strata that will step by step cover the space. We want to find a clever procedure to select a strata that have a target unit at the edge. First of all let define a strata.

Definition 1 *Let define the* ***strata of the unit k*** *as the subset of units $S_k \subset U$ centered around k such that, starting from the unit k, gradually nearest neighbors are added such that the sum of the inclusion probabilities inside S_k exceed 1. Meaning that if we denote L the number of units in S_k, the inclusion probabilities satisfies:*

$$\sum_{\ell=1}^{L} \pi_\ell > 1 \text{ with } \sum_{\ell=1}^{L-1} \pi_\ell \leq 1.$$

In order to reach exactly 1, we cut the inclusion probability of the farthest unit such that the sum inside the strata reach exactly 1. The farthest unit is modified as the following way:

$$\pi_L = 1 - \sum_{\ell=1}^{L-1} \pi_\ell.$$

If all distances are different, the strata of the unit k has only one unit whose inclusion probability is modified. But we are looking for strata that have target unit and modified unit at the maximal distance. Suppose that the index of the target unit is labeled as j. We could then redefined the modified unit such that we take the farthest unit from the target unit j and not from the center k.

Definition 2 *Let define the* ***strata of the unit k targeted on j*** *as the subset $S_k^j \subset U$, such that instead of modify the farthest unit from the unit k, we modify the farthest unit of the unit j. This strata is well-defined only if the unit j is contained in S_k.*

We could then defined the union of the whole strata that contains in any way the targeted unit.

Definition 3 *Let's define the* ***raised strata*** *of the unit j as the set $R_j \subset U$ such that*

$$R_j = \bigcup_{k \in U} S_k^j.$$

It means that R_j contains also strata such that j is at the opposite edge of the modified unit in strata S_k^j. We could then defined the strata that are possible candidate.

Definition 4 *Define the* ***candidate units*** *as the units that has never been completely added in any strata in R_j.*

4.2 Tore and shifted distance

4.3 Implementation

We select a starting unit j with respect to the inclusion probabilities $\boldsymbol{\pi}$. We only select a part of the inclusion probability u distributed as a random variable $\mathcal{U}(0, \pi_j)$. Let's define $\boldsymbol{\pi}^*$ the vector of inclusion probabilities such that the π_j is replaced by u. We then looking at the raised strata R_j of the unit j

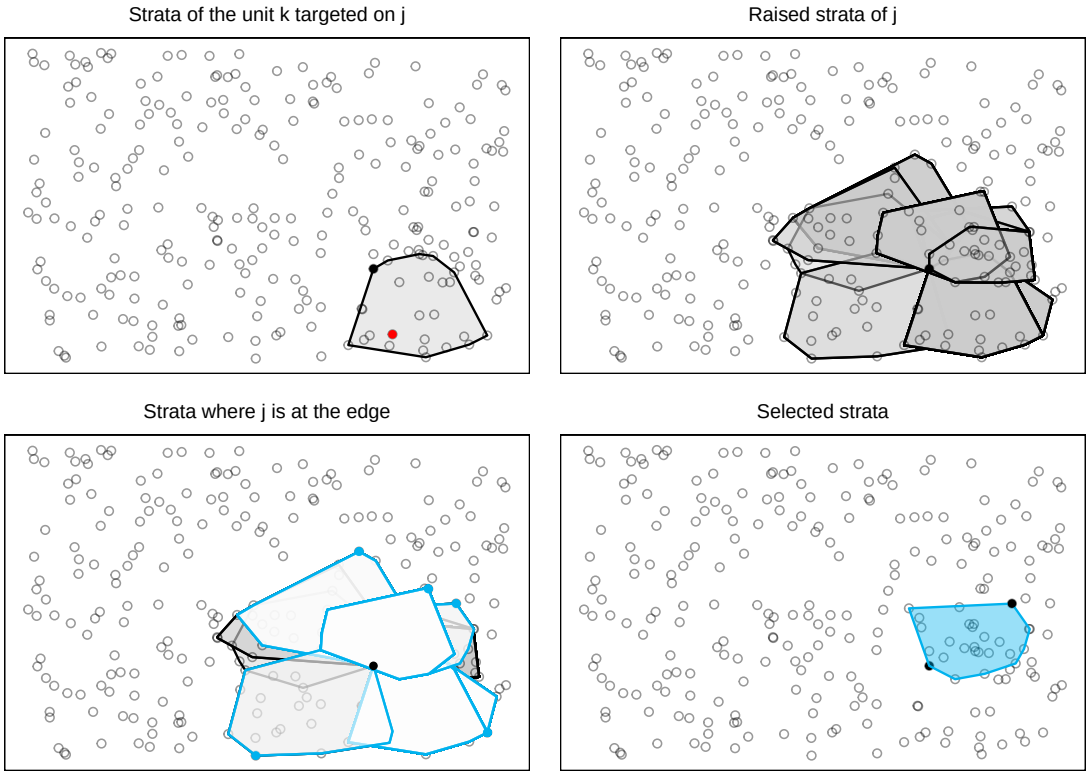

Figure 1: Illustration of the convex hull of the four definition of the different strata.

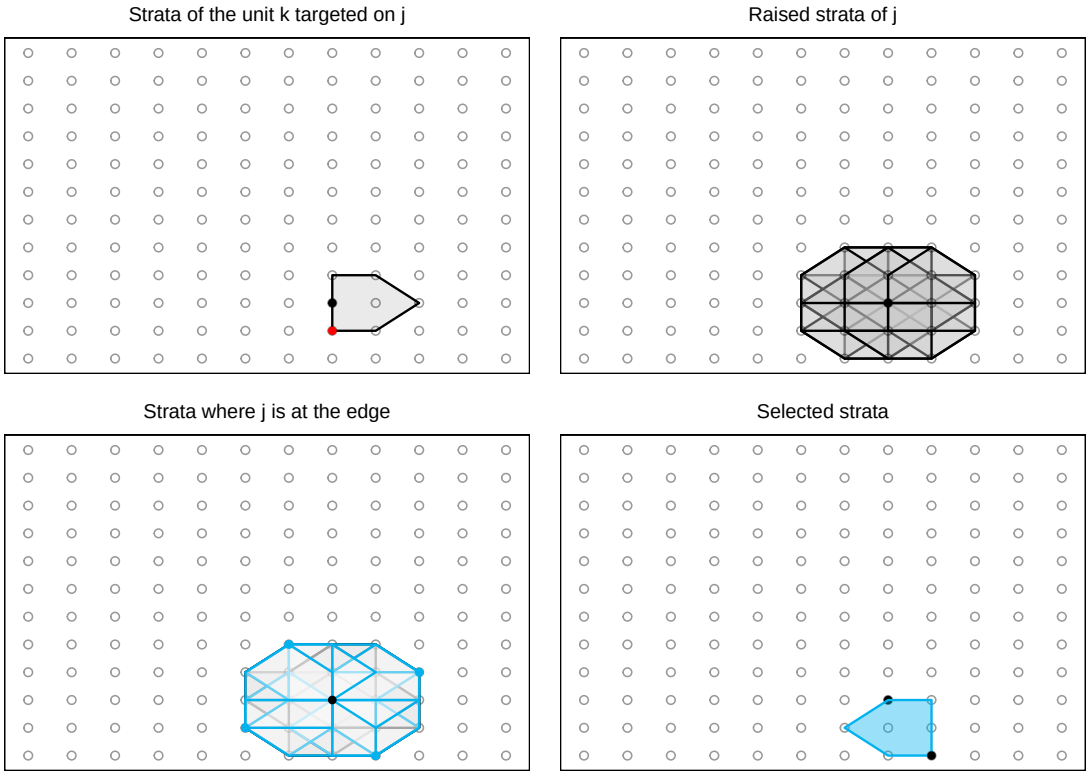

Figure 2: Illustration of the convex hull of the four definition of the different strata.

Algorithm 2 Algorithm for systematic sampling

Let $\boldsymbol{\pi}_0 = (\pi_1, \ldots, \pi_N) = \boldsymbol{\pi}$ for the initialization step. Select a starting unit j with respect to $\boldsymbol{\pi}$. Generate a random start u distributed as a uniform variable $\mathcal{U}(0, \pi_j)$.

1. Define $\boldsymbol{\pi}^* = (\pi_1, \ldots, \pi_{j-1}, u, \pi_{j+1}, \ldots, \pi_N)$
2. Compute the raised strata R_j using $\boldsymbol{\pi}^*$.
3. Find the candidate unit that have the minimum distance from the unit j and the modified unit in the strata. Let i be the index of the modified unit of the selected strata. Let r_i be the quantity not selected in the strata of the unit i.
4. Update
$$\boldsymbol{\pi}^* = (\pi_1, \ldots, \pi_{i-1}, r_i, \pi_{i+1}, \ldots, \pi_{j-1}, \pi_j - u, \pi_{j+1}, \ldots, \pi_N)$$
5. Repeat 2. to 4. for the non-zero modified inclusion probabilities until you selected the right number of units.

5 Spatial Balance

6 Simulation

7 Discussion