SamplingStrata methodology

Marco Ballin, Giulio Barcaroli

20 May 2019

Problem definition

Accordingly to Sarndal, Swensson and Wretman

in a stratified sampling design the population is divided into nonoverlapping subpopulations called strata. A probability sample is selected in each stratum. The selection in the different strata are indipendent.

A stratification is a partition of the population units and it is usually defined according to the values of one or more auxiliary variables (X) available for all units of the population.

Problem definition

If the stratification is given, a typical problem that has to be solved is the definition of the total sample size and its allocation among the strata in order the expected accuracies of the sample estimates are below some fixed thresholds.

If we have only one variable of interest (univariate case) the problem can be solved by **Neyman allocation** criteria, while in the multivariate case it is possible to make use of the **Bethel algorithm**.

These criteria define the accuracy in terms of variances of Horvitz Tompson estimator:

$$Var(\hat{Y}_g) = \sum_{h=1}^{H} N_h^2 (1 - \frac{n_h}{N_h}) \frac{S_{h,g}^2}{n_h} \qquad g = 1, \ldots, G$$

Optimal allocation for a given stratification

If we introduce the following cost function:

$$C(n_1, \ldots, n_H) = C_0 + \sum_{h=1}^{H} C_h n_h$$

the optimization problem can be formalized in this way:

$$min = C_0 + \sum_{h=1}^{H} C_h n_h$$

under the constraints

$$\begin{cases} CV(\hat{Y}_1) < U_1 \\ CV(\hat{Y}_2) < U_2 \\ \ldots \\ CV(\hat{Y}_G) < U_G \end{cases}$$

where

$$CV(\hat{Y}_g) = \frac{\sqrt{Var(\hat{Y}_g)}}{mean(\hat{Y}_g)}$$

Optimal allocation for a given stratification

In the univariate case it is possible to obtain an analytic solution

$$n_h = \frac{W_h S_h \sum_{h=1}^{H} W_h S_h}{U}$$

while in the multivariate case solution has to be searched by numeric algorithm.

Since the previous problem is equivalent to search for a minimun of a convex function under linear constraints, such solution always exist.

The universe of stratifications

If the stratification is not given, the solution has to be searched possibly using some optimal criteria.

Coherently with previous definitions, we define the best stratification as the one that support

$$\sum_{h=1}^{H} n_h = \min$$

for a given *U*.

SamplingStrata looks for the best stratification among all possible stratifications

The universe of stratifications

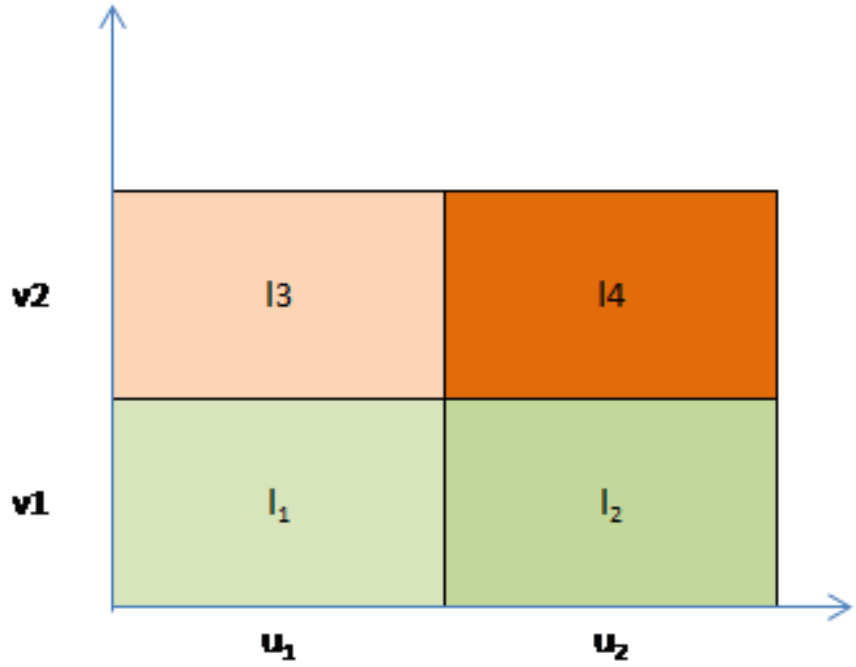
Given a population frame with M auxiliary qualitative variables $X_1, \ldots, X_M$ we define as **atomic stratification** the one that can be obtained considering the cartesian product of the M variables.

$$L = \{(l_1), (l_2), \ldots, (l_k)\}$$

Starting from the atomic stratification, it is possible to generate all the different stratifications that constitute the universe of stratifications.

For example, using two dichotomous variables X_1 and X_2 whose modalities are $\{(u_1), (u_2)\}$ and $\{(v_1), (v_2)\}$ respectivelly, the cartesian product supports four elements

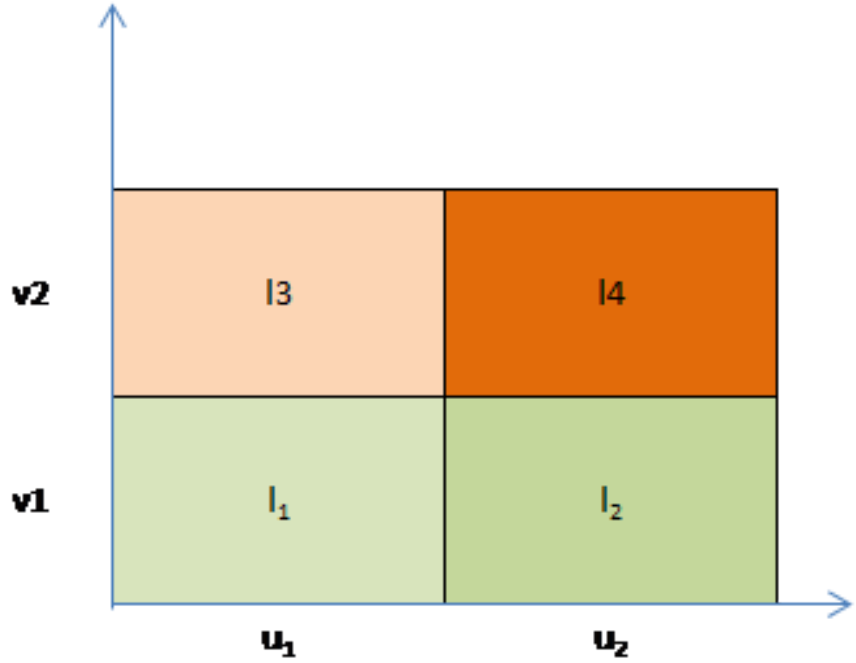
$$L = \{(l_1), (l_2), (l_3), (l_4)\}$$


v2
v1
I3
I4
I1
I2
u1
u2

The universe of stratifications

In such conditions, the universe of stratifications is constituted by the following elements:

$$P_1 = \{(l_1, l_2, l_3, l_4)\} \qquad P_2 = \{(l_1), (l_2), (l_3), (l_4)\}$$
$$P_3 = \{(l_1), (l_2, l_3, l_4)\} \qquad P_4 = \{(l_2), (l_1, l_3, l_4)\}$$
$$P_5 = \{(l_3), (l_1, l_2, l_4)\} \qquad P_6 = \{(l_4), (l_1, l_2), l_3\}$$
$$P_7 = \{(l_1, l_2), (l_3, l_4)\} \qquad P_8 = \{(l_1, l_3), (l_2, l_4)\}$$
$$P_9 = \{(l_1, l_4), (l_2, l_3)\}$$
$$P_{10} = \{(l_1, l_2), (l_3), (l_4)\} \qquad P_{11} = \{(l_1, l_3), (l_2), (l_4)\}$$
$$P_{12} = \{(l_1, l_4), (l_2), (l_4)\} \qquad P_{13} = \{(l_2, l_3), (l_1), (l_4)\}$$
$$P_{14} = \{(l_2, l_4), (l_1), (l_3)\} \qquad P_{15} = \{(l_3, l_4), (l_1), (l_2)\}$$



The universe of stratifications

If we would like to determine which is the best stratification among these fifteen it should be sufficient to solve the optimal allocation problem for each one of them and finally to choose the one that support the minimum total sample size.

Unfortunately, the number of feasible stratifications is exponential with respect to the number of initial atomic strata:

$$B_4 = 15 \qquad B_{10} = 115975 \qquad B_{100} \approx 4.76 \times 10^{115}$$

Consequently, in concrete cases, it is impossible to examine all alternative stratifications.

The **Genetic Algorithm** allows to explore the universe of stratifications in a very efficient way in order to find the optimal (or close to optimal) solution.

Use of the Genetic Algorithm in the optimization process

The basic implementation of **SamplingStrata** proceed as follows:

- given the survey variables $Y_1, Y_2, \ldots, Y_p$, set **precision constraints** on their estimates in the different domains, expressed in terms of coefficients of variations;
- using the auxiliary variables $X_1, \ldots, X_M$ recorded in the frame build the **atomic stratification**;
- for each atomic stratum report the **means** and **standard deviations** of the variables of interest (usually by using proxy variables);
- on the basis of these inputs, the Genetic Algorithm determines the **best solution** in terms of
  - ***stratification***,
  - ***sample size***,
  - ***strata allocation***.

Use of the genetic algorithm in the optimization process

The *genetic algorithm* looks for the best solution exploiting the principles that support the *natural evolutionary mechanism* of a population passing through successive *generations* formed by *individuals*.

In the stratification problem:

- a stratification is an *individual*;
- a set of *individuals* (that is a set of *stratifications*) at each iteration of the algorithm is a *generation* subject to *evolution*;
- each individual is characterized (identified) by a *genome* represented by a vector of dimension equal to the number of atomic strata;
- the position of each element in the vector identifies an atomic stratum;
- to each element of the vector is assigned a label between 1 and K (maximum acceptable number of strata): the sequence of labes indicates the way in which the individual atomic strata have to be aggregated together;
- each stratification is characterized by a particular sequence of labels.

Use of the genetic algorithm in the optimization process

Id	Gender	Income_class	Savings
1	M	2	1000
2	F	3	2500
3	M	1	1300
4	F	4	3500
5	M	1	1800
6	F	2	2800
7	M	3	1400
8	F	4	3000

Sampling Frame

STRATUM	N	Mean(Savings)	Stdev(Savings)	Gender	Income_class
1*1	2	1550	250	1	1
1*2	1	1000	0	1	2
1*3	1	1400	0	1	3
2*2	1	2800	0	2	2
2*3	1	2500	0	2	3
2*4	2	3250	250	2	4

Atomic strata

1	1	1
1	2	2
1	2	3
1	3	4
1	2	5
1	1	6

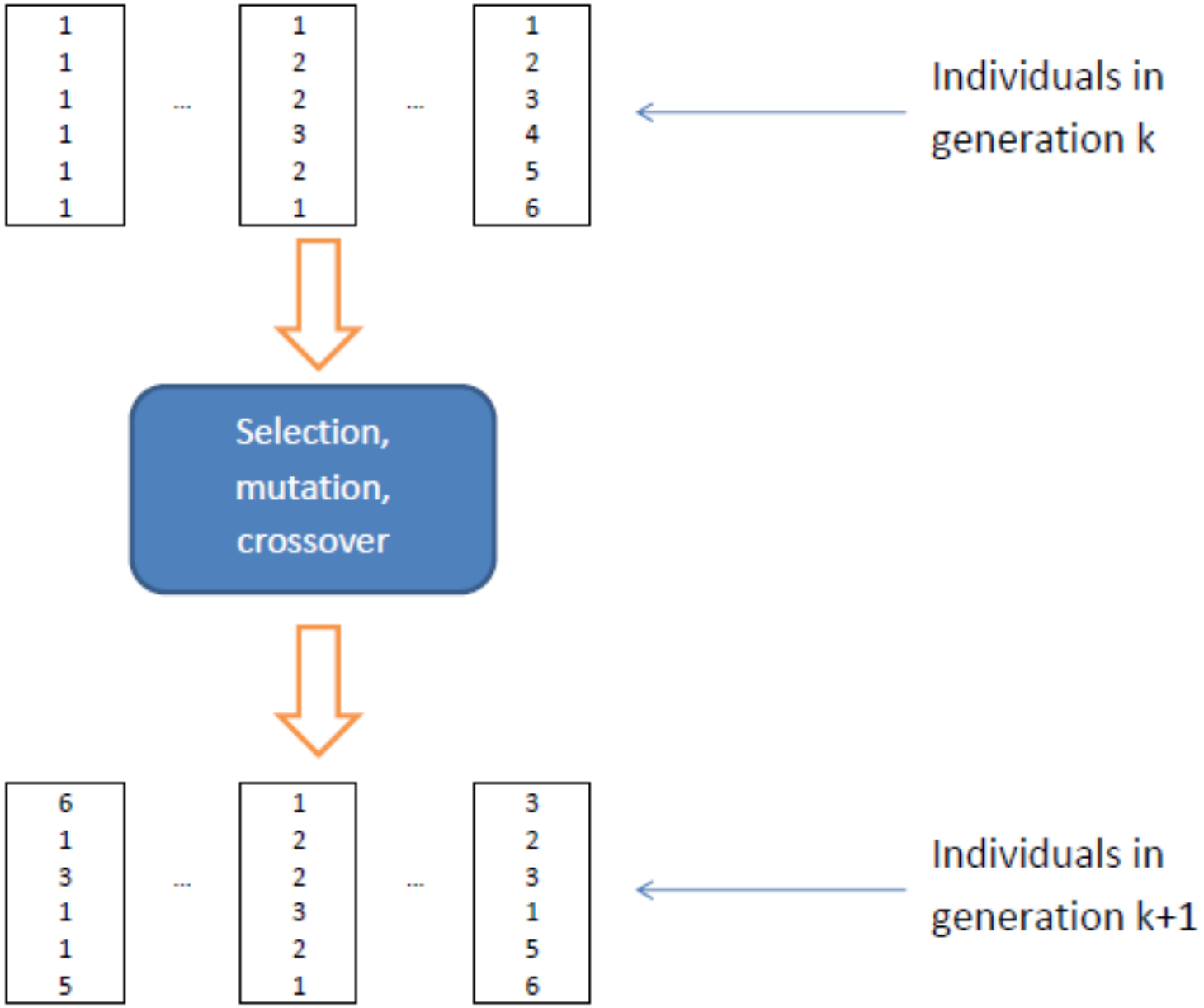
... ...

Individuals in a generation

Use of the genetic algorithm in the optimization process

- the *fitness* (that is the total cost or the total sample size if $C_h = 1$) of each individual (a stratification) is calculated by solving the corresponding optimization problem by means of Bethel's algorithm;
- in passing from one generation to the next, individuals with higher fitness (minimum cost or sample size) are favored by *natural evolutionary mechanisms*; such mechanisms are
  - *selection (individuals with higher fitness have greater probability to contribute to next generation)*
  - *crossover (some individuals are formed by crossing genoma of individuals with higher fitness)*
  - *mutation (random mutation to the genoma of infividuals with higher fitness)*
- the evolution process end after several generation (iteration) and ***the individual with the overall best fitness represents the optimal solution***.

Use of the genetic algorithm in the optimization process


1 1 1 1 1 1
...
1 2 2 3 2 1
...
1 2 3 4 5 6
Individuals in generation k
Selection, mutation, crossover
6 1 3 1 1 5
...
1 2 2 3 2 1
...
3 2 3 1 5 6
Individuals in generation k+1

The case of quantitative variables

Up to now, the principles below *SamplingStrata* have been illustrated assuming X are qualitatitive variables or assuming that quantitative variables have been previously categorized (*SamplingStrata* has a function that suggests a split of a quantitative variable in a given number of classes).

If we want to stratify using two or more quantitative variables we can use a more efficient version of algorithm.

The gain of efficiency can be achieved thanks to some constraints and assumptions:

- the optimal stratification is searched in the universe of stratifications where each strata is built by contiguous values of the stratification variables;
- each stratification variable is split in the same number (*k*) of intervals;
- each stratum is '7' shaped.

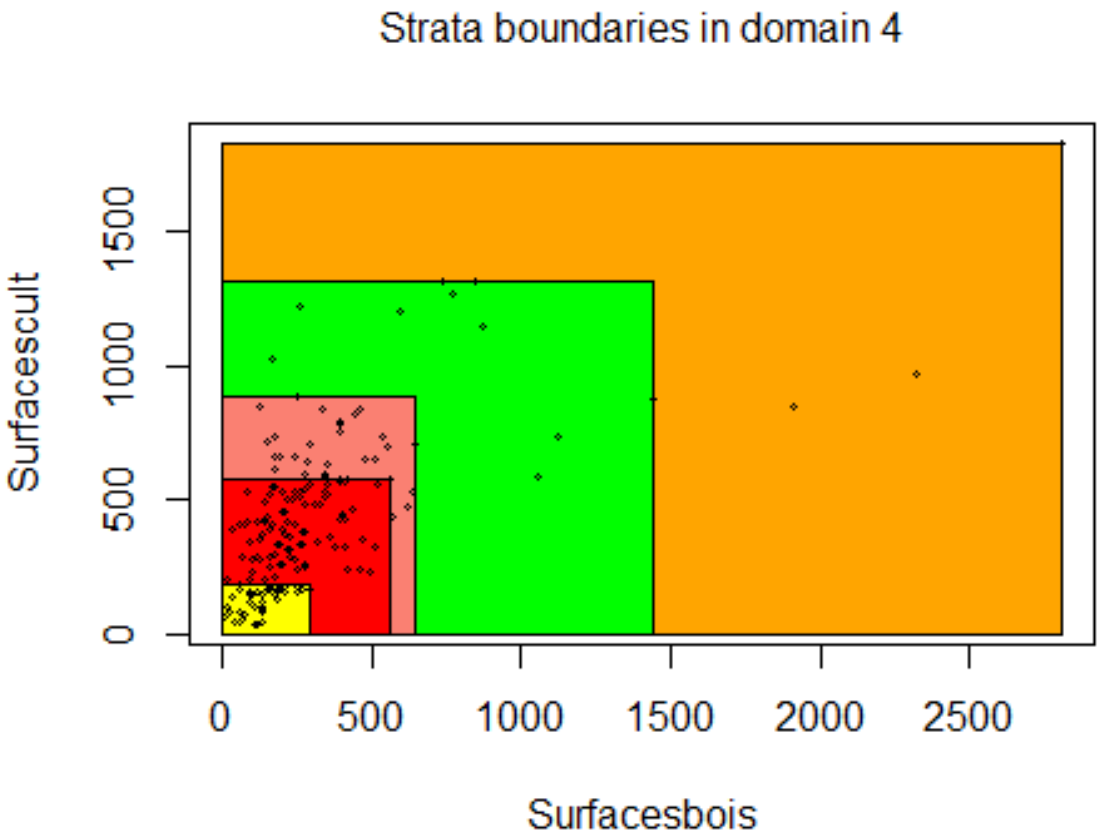
The first restriction to the universe of stratifications is very common and in most practical cases it supports the overall optimum stratification.

The second point could be easily avoided but it is essential to satisfy the third condition.

The '7' shape of the strata has been introduced to meet requests by personnel in charge of the business surveys who look for simple stratification rules.

The case of quantitative variables

In this case we don't need to build up the initial atomic stratification because each individual of the initial generation of stratifications is obtained by $(k-1)$ random cuts of the range of each stratification variable and then by combining them as exemplified in the following figure.



The principles of the genetic algorithm is now applied to the boundaries of each stratum.

Working with anticipated variance

Another important assumption that we have made until here is that the variance and the mean of each variable of interest can be computed for each atomic stratum using data beloging to a previous surveys or available in the frame.

Such assumption are clearly very strong and some times it can't be accepeted.

A weaker assumption is the one that assume a relationship between our variable of interest Y and a proxy variable Z available in the frame.

A typical relantinship is the linear model $Y_i = \beta * Z_i + \epsilon_i$

where $\epsilon \sim \ N(0, Z^\gamma * \sigma)$

If $\gamma = 0$ the model is homoscedastic otherwise it is heteroscedastic.

(SamplingStrata can handle linear and logistic models)

Working with anticipated variance

If the assumptions about the model hold the variance of HT estimator for Y can be writtes as

$$Var(\hat{Y}) = \sum_{h=1}^{H} \frac{N_h^2}{n_h}(1 - \frac{n_h}{N_h})\{\sum_{i=1}^{N_h} Z_i^{2\gamma}\sigma^2 + \sum_{i=1}^{N_h}(Z_i\beta - \overline{Z}\beta)^2\}$$

As it can be seen, the previous formula for variance can be obtained when $\sigma = 0$ and $\beta = 1$