SamplingStrata methodology

Giulio Barcaroli

29 August 2018

Problem definition

In a stratified sampling design with one or more stages, a sample is selected from a frame containing the units of the population of interest, stratified according to the values of one or more auxiliary variables (X) available for all units in the population.

For a given stratification, the overall size of the sample and the allocation in the different strata can be determinedon the basis of constraints placed on the expected accuracy of the various estimates regarding the survey variables (Y).

If the target survey variables are more than one the optimization problem is said to be **multivariate**; otherwise it is **univariate**.

For a given stratification, in the univariate case the optimization of the allocation is in general based on the **Neyman allocation**. In the univariate case it is possible to make use of the **Bethel algorithm**.

Problem definition

The criteria according to which stratification is defined are crucial for the efficiency of the sample.

With the same precision constraints, the overall size of the sample required to satisfy them may be significantly affected by the particular stratification chosen for the population of interest.

Optimal allocation for a given stratification

Given G survey variables, their sampling variance is:

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

If we introduce the following cost function:

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

Optimal allocation for a given stratification

the optimization problem can be formalized in this way:

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

under the constraints

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

where

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

The universe of stratifications

Given a population frame with m auxiliary variables $X_1, \ldots, X_M$ we define as **atomic stratification** the one that can be obtained considering the cartesian product of the definition domains 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 belong to the universe of stratifications. For example:

$$P_1 = \{(l_1, l_2, l_3)\} \qquad P_2 = \{(l_1), (l_2, l_3)\}$$
$$P_2 = \{(l_2), (l_1, l_3)\} \qquad P_4 = \{(l_{31}), (l_1, l_2)\}$$
$$P_5 = \{(l_1), (l_2), (l_k)\}$$

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

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

In concrete cases, it is therefore impossible to examine all the different possible alternative stratifications. The **Genetic Algorithm** allows to explore the universe of stratification 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

In planning a strafied sampling for a given survey, 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 CVs (coefficients of variation);
- in the available sampling frame build the **atomic stratification**, obtained as cartesian product of the domains of the auxiliary variables $X_1, \ldots, X_M$;

Use of the genetic algorithm in the optimization process

- in each atomic stratum report the distributional characteristics of the survey variables by calculating their **means** and **standard deviations** calculate the values of the population (directly or by using proxy variables);
- on the basis of these inputs, the Genetic Algorithm determines the **best solution** in terms of both frame **stratification**, **sample size** and **allocation** in optimized strata.

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

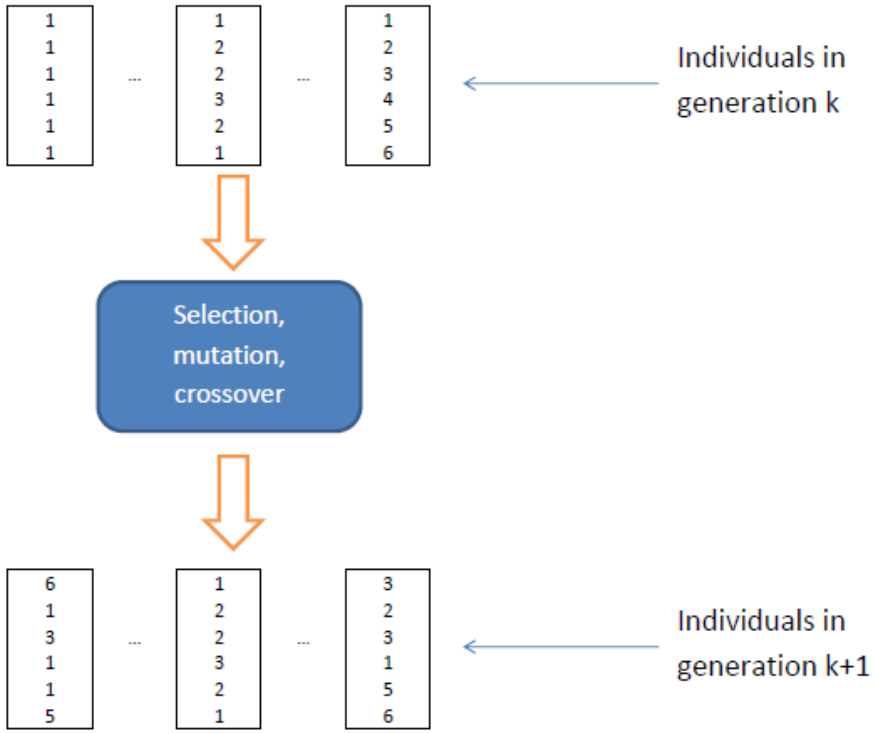
Individuals in a generation

Use of the genetic algorithm in the optimization process

The application of the genetic algorithm is based on the following steps:

- a given stratification is considered as an *individual* in a population (= *generation*) subject to *evolution*;
- each individual is characterized 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;
- each element in the vector is assigned a random value between 1 and K (maximum acceptable number of strata): the vector therefore indicates the way in which the individual atomic strata are aggregated together;

Use of the genetic algorithm in the optimization process



Use of the genetic algorithm in the optimization process

- for each individual (stratification) its *fitness* is calculated by solving the corresponding problem of optimal allocation by means of Bethel's algorithm;
- in passing from one generation to the next, *individuals with higher fitness are favored*;
- at the end of the process of evolution, ***the individual with the overall best fitness represents the optimal solution***.