Balanced Sampling

Laurent Costa & Thomas Merly-Alpa

***Résumé* — The aim of this methodological note is to describe balanced sampling and when it is to be applied. The cube method and its implementation are also discussed, and variance estimation will also be briefly touched upon.**

I. Introduction

When establishing a sample, the question of the sampling plan and its effectiveness must be addressed. The aim is to obtain a sample that best reflects the heterogeneity of the population surveyed by reducing the variance of the estimators and limiting costs. The classical sampling designs which aim at reducing variance are stratified designs and sampling designs with unequal probabilities of inclusion. Nevertheless, it is not always desirable to perform stratification if n is low or if you do not wish to compute allocations for rounding issues, for example.

The idea of balanced sampling is based on using available information correlated with the variable of interest when developing the plan. The precision of a sampling design is based on balancing properties : the sample is selected so as to comply with known information. For example :

- compliance with age-gender structure ;
- distribution according to number of employees.

Where a selected sample accurately reflects the information available in accordance with what is actually found within the population, it will reflect the information concerning the variable of interest well thanks to the correlation between the two types of information. This explains the ability of the balanced sampling plan to improve the efficiency of the estimators.

Despite the difficulty in applying a general method in an algorithmic manner that complies with both the balancing constraints and the random selection of the sample [1], we will see that the CUBE method, developed by Deville and Tillé en 2004, makes it possible to draw samples that are approximately balanced.

II. Definition of Balanced Sampling

A sample is said to be balanced on one or more of the variables available within the sampling frame when, for each of them, the Horvitz-Thompson estimator of the total precisely matches the actual total from the sampling frame.

1. The balancing may prove to be so constrained that it would lead to a deterministic selection. However, the selection must remain random in order for the statistical properties of sampling bias and variance to remain meaningful and in order to comply with the inclusion probabilities.

By way of a reminder, the following shows the definition of the unbiased Horvitz-Thompson estimator of the total for a variable x written $\hat{t}_{x\pi}$ for a sample S :

$$\hat{t}_{x\pi} = \sum_{i \in S} \frac{x_i}{\pi_i}$$

where π_i is the probability of the individual i being included in the sample S.

A sample S from a population U balanced with the control variable x therefore complies with the following constraint :

$$\sum_{i \in S} \frac{x_i}{\pi_i} = \sum_{i \in U} x_i \quad \text{so} \quad \hat{\mathrm{t}}_{\mathrm{x}\pi} = \mathrm{t}_{\mathrm{x}}$$

This therefore acts as a form of calibration at the level of the sampling plan for the auxiliary variables. By design, the estimator of the total for x is unbiased and of zero variance.
Let us examine a working model of type :

$$y_i = \beta x_i + \varepsilon_i$$

which can be rewritten by dividing by π_i then aggregating each individual i in the form :

$$\hat{t}_{y\pi} = \beta \hat{t}_{x\pi} + \hat{t}_{\varepsilon\pi}$$

Since the sampling plan is balanced using the variable x correlated to y,its total is perfectly estimated [2], this therefore gives :

$$\hat{t}_{y\pi} = \beta t_x + \hat{t}_{\varepsilon\pi}$$

And we obtain the following [3] :

$$V(\hat{t}_{y\pi}) = V(\hat{t}_{\varepsilon\pi})$$

It can therefore be seen that :

- Compliance with the probabilities of inclusion results in an unbiased estimate $\rightarrow E(\hat{t_{y\pi}})=t_y$;
- Restricting the support from the sampling plan to balanced samples allows the first term variability in x to be cancelled out ;
- The variance is now only indicated by the residuals of the model.

It is also possible to deduce certain properties :
Assuming that $x_i=\pi_i$, i.e. balancing takes place on the basis of the probabilities of inclusion.
The balancing equation implies that

$$\hat{t}_{x\pi} = \sum_{i \in S} \frac{x_i}{\pi_i} = \sum_{i \in S} \frac{\pi_i}{\pi_i} = n(s)$$

$$\textit{and} \quad t_x = \sum_{i \in U} \pi_i = E(n(S))$$

2. The variance of $\hat{t}_{x\pi}$ is zero.
3. Where the expression βt_x is constant.

However $\hat{t}_{x\pi}$=t_x hence

$$n(s) = E(n(S))$$

The sampling design then implies a fixed size of the sample.

Assuming that x_i=1, i.e. balancing takes place on the basis of a constant variable of 1.
The balancing equation implies that

$$\sum_{i\in S} \frac{x_i}{\pi_i} = \sum_{i\in S} \frac{1}{\pi_i} = \hat{N}_\pi = \sum_{i\in U} 1 = N$$

The size of the population has therefore been perfectly estimated.

There is also reason to believe, and it could be proven that, for a stratified design with proportional allocation, the size of the strata can be perfectly estimated using the binary variables indicating strata affiliation as balancing variables. Balanced sampling is therefore a random sampling method that ensures that, in the end, once the sample has been drawn, the proportions of individuals, in the population and in the sample respectively, will be equal for each classes.

During the conduct of a survey, a non-response phenomenon is generally observed within the sample, which upsets the balance. Balanced sampling is therefore of particular interest for a first degree of sampling or when a low non-response rate is anticipated. On can mention, for example [4] :

- Drawing Primary Units from the Master Sample ;
- Drawing Census Rotation Groups.

III. The Cube Method

A. Principe

The algorithm proposed by Deville and Tillé (2004) [2] [5] has a general framework and enables balanced samples to be selected from any number of variables, with a given set of probabilities of inclusion $\boldsymbol{\pi}$=$(\pi_1,...,\pi_N)$. A sample s is seen as a vertex $(s_1,...,s_N) \in \{0,1\}^N$ of the N-cube C=$[0,1]^N$.The algorithm consists of a random walk from the probabilities of inclusion vector $\boldsymbol{\pi}$ to the selection indicator vector s $\mathbf{I}$ by randomly rounding the π_i to 0 or 1.
This gives :

$$\hat{t}_{x\pi} = \sum_{i\in U} \frac{x_i}{\pi_i} I_i = t_x = \sum_{i\in U} x_i = \sum_{i\in U} x_i \frac{\pi_i}{\pi_i}$$

$$\text{so} \quad \sum_{i\in U} \frac{x_i}{\pi_i}(I_i - \pi_i) = 0 \quad \text{or} \quad \mathbf{A}(\mathbf{I} - \boldsymbol{\pi}) = 0$$

where $\mathbf{A} = (\frac{x_1}{\pi_1},...,\frac{x_N}{\pi_N})$; $\mathbf{I} = (I_1,...,I_N)^T$is the selection indicators vector and $\boldsymbol{\pi} = (\pi_1,...,\pi_N)^T$ is the probabilities of inclusion vector. It can therefore be seen that I must be within the space of the constraints $\boldsymbol{\pi} + Ker(\mathbf{A})$ which represents the subspace in which the balancing conditions are met.

t is therefore easy to represent this model in a three-dimensional space for a population made up of three units : i.e. a cube. Taking the example of a simple random survey without replacement of size 2, and by assigning the same probabilities of inclusion to each of the units (π_i=2/3), it can be seen that precise balancing is always achieved by balancing to the constant variable equal to 1. It is therefore known that there are 3 balanced samples composed of 2 distinct units : these are the vertices (0,1,1) ; (1,1,0) and (1,0,1).

Fig. 1. Graphical representation for a population of 3 units for a simple random survey without replacement, balanced with the variable "1".

B. Details of the Algorithm

We will now provide details of the algorithm, which is implemented in two phases : the flight phase and the landing phase.

α. The flight phase

We start with $\boldsymbol{\pi}^{(0)}$=$\boldsymbol{\pi}$. At stage t, we have $\boldsymbol{\pi}^{(t)}$=$\boldsymbol{\pi}^{(t-1)}$+$\boldsymbol{\delta}^{(t)}$ where

$$\boldsymbol{\delta}^{(t)} = \begin{cases} \lambda_1(t)\boldsymbol{u}(t) & \text{with proba} \quad \lambda_2(t)/(\lambda_1(t)+\lambda_2(t)) \\ -\lambda_2(t)\boldsymbol{u}(t) & \text{with proba} \quad \lambda_1(t)/(\lambda_1(t)+\lambda_2(t)) \end{cases}$$

- $\lambda_1(t)$,$\lambda_2(t)$>0
  $\rightarrow$ ensures that at least one unit is selected or definitively rejected ;
- $\boldsymbol{u}(t) \in Ker(\mathbf{A})$
  $\rightarrow$ ensures that the balancing equations are complied with precisely ;
- The random selection ensures that the probabilities of inclusion are complied with precisely.

The flight phase operates in successive iterations where each step decides the fate of at least one individual and also selects a random direction within the constraints area. We will follow this until it leads to one of the sides of the cube. The flight phase allows decisions to be made for at least $N-p$ individuals [6]and allows the balancing constraints and probabilities of inclusion to be complied with.
If, at the end of this phase, we have arrived at one of the vertices of the cube, this means that the sample is perfectly balanced ; otherwise it is impossible to precisely comply with all of the constraints and we will find ourselves "stuck" on one side of the cube : the landing phase should therefore be initiated. This will allow decisions to be made for the remaining individuals while complying precisely with the probabilities of inclusion and roughly complying with the balancing constraints.

β. The landing phase

There are three possibilities for this phase of the algorithm.
The first is to release the constraints one by one. We therefore introduce a degree of freedom at each step, which allows us to continue sampling. This is the most general option in the sense that it allows us to work with any number of balancing variables. However, the first variables released may be poorly

4. See Part V. for examples of application.

5. The CUBE macro is available, together with its documentation, on the INSEE website : https ://www.insee.fr/fr/information/2021904

6. where p is the number of balancing constraints.

balanced.

The second involves the definition of a sampling plan for the remaining units :

- complying with the initial probabilities of inclusion ;
- minimising (on average) the deviation from the balance, using a criterion of type :

$$\min \quad E||\hat{t}_{x\pi} - t_x||^2$$

This option allows a good overall balance to be achieved ; however, it is necessary to fully define a sampling plan for a population of p individuals, which is not possible if p is large [7].

The third is identical to the second, but also complies with the fixed size constraint. In order to do so, it is necessary to achieve a balance with the possibility of inclusion.

γ. General example

Once again, we position ourselves within our cube, so within a population of 3 units to which we assign the same probabilities of inclusion (π_i=2/3)by way of a more general example, this would mean that the balance is sometimes exact [8]. ere we will examine the case of a random survey without replacement balanced with the order number [9] of the individuals.

Fig. 2. Graphical representation for a population of 3 units for a random survey without replacement, balanced with the order number of the units.

The balance is only precise at the vertex (1,0,1) that represents the only intersection of a vertex within the cube with the constraints space.
At the end of the flight phase, the algorithm will have selected this vertex [10] or will have led to one of the other three points of intersection between the cube and the constraints space : the landing phase will then need to be initiated. Depending on the optimality criterion, this phase will lead to the selection of one of the 5 samples (which are therefore approximately balanced) positioned on the vertices of the edges corresponding to the points of intersection [11].
The optimality criterion selected for the landing phase increases the chances of retaining the samples that are "closest" to the balance. Nevertheless, in order to preserve the random nature of the sampling, the method cannot guarantee that a unique, perfectly balanced sample will be obtained.

IV. Variance Estimation

Deville et Tillé (2005) proposed a class of variance estimators under the following assumptions :

- the sampling plan is exactly balanced ;
- the sampling design is at maximum entropy [12] among designs balanced with the same variables $\mathbf{x}_i$ and with the same probabilities of inclusion $\boldsymbol{\pi}$.

Therefore, under these two conditions, the balanced design can be viewed as a Poisson sampling design conditional to $\hat{t}_{x\pi}$=t_x. The resulting approximate variance is given by :

$$V_{app}(\hat{t}_{y\pi}) \quad = \quad \sum_{i\in U} b_i(\frac{y_i}{\pi_i} - \frac{\mathbf{x}_i^T}{\pi_i}\mathbf{B})^2$$

where B is the regression coefficient vector [13] of $\frac{y_i}{\pi_i}$ for the balancing variables $\frac{\mathbf{x}_i}{\pi_i}$ and the b_i's are solutions from a non-linear system, a first approximation of which is given in the article by Deville and Tillé as b_i=π_i(1-π_i).
Using the principle of expansion, we then obtain the Deville and Tillé variance estimator.

However, the two conditions described above are generally not verified, firstly due to the landing phase (but reasonable if the number of balancing variables p is small with respect to N) and secondly due to the difficulty of making a plan as random as possible (for example if the file is sorted according to an auxiliary variable, the result is a stratification effect that affects the entropy).

V. Applications

A. Balancing Within the Master Sample

The Master Sample (MS) is a sample of zones used as a reserve of dwellings for household surveys. The 1999 Master Sample (MS99) was used for the surveys conducted between 1999 and 2009. Each of these zones was entrusted to a collector who was "stable over time and relatively nearby". These are referred to as Collector Action Zones (ZAE). The switch to Census Surveys from 2004 onwards has resulted in changes having to be made to the sampling system used for the MS.

In each large municipality (more than 10,000 inhabitants), stratification took place according to the address type, which was divided into 5 rotation groups. In the case of small municipalities, stratification took place based on the region and the municipalities were divided into 5 rotation groups by means of random sampling performed using the Cube method.

The new Octopusse Master Sample was presented as follows :
At the level of the large municipalities :

- 1 ZAE = 1 large municipality ;
- drawing a sample of large municipality ZAEs (Cube method) ;

7. if $p = 19$, for example, there are approximately 500,000 possible samples.

8. You will recall that, in the previous example, the balance was always precise and the algorithm was able to end during the flight phase by establishing a sample.

9. Variable corresponding to the value of the line on which the individual is positioned within the file.

10. And in this case, it will end here, having precisely complied with all of the balancing constraints.

11. This therefore relates to the following vertices : (0,1,0), (0,1,1), (0,0,1), (1,1,0) and (1,1,1).

12. The entropy of a sampling design p is defined by the following :

$$L(p) = -\sum_{s\in U} p(s) \quad ln(p(s))$$

This is a measure of disorder : the greater it is, the more the design allows for the selection of a large number of samples (and therefore leaves plenty of scope for randomness).

13. If we consider E_i=y_i-$\mathbf{x}_i^T\mathbf{B}$ within the variance formula, it can therefore be seen that E_i represents the regression residuals.

- for a survey conducted during year t +1, drawing a sample of dwellings from among those included in the Census surevy in year t.

→ Two-degree sampling.

At the level of the small municipalities :

- 1 ZAE = group of small municipalities, containing at least 300 principal residences from each of the rotation groups ;
- drawing a sample of small municipality ZAEs (Cube method) ;
- for a survey conducted during year t +1, drawing a sample of dwellings from among those included in the Census survey in year t.

→ Two-degree sampling.

The drawing of the ZAEs (in small municipalities or in large municipalities) was balanced, using the CUBE method, with the number of principal residences within the ZAE for each rotation group, the disaggregated tax income for each rotation group, the number of principal residences by space type in the 1999 Census survey [14].

B. Sampling for the CARE-I Survey

The Institution CARE survey conducted among elderly people living in institutions by the Directorate for Research, Studies, Assessment, and Statistics (DREES) is intended to supplement the CARE (Capacities, Aids and REsources) survey conducted among elderly people living in ordinary households, which pursues the same objectives : monitoring changes in dependency and measuring the involvement of family and friends in caring for the elderly person. The sampling of the institutions takes place in 2 phases. First, 30 departments are drawn and then 1000 institutions are drawn from within those departments. The departments are selected based on the number of residents they have living in institutions for elderly people. Indeed, the departments are not all equivalent to one another : they all have more or fewer institutions and/or residents. In order to take account of those differences, sampling takes place with unequal probabilities. The departments were drawn from within 3 groups of homogeneous departments, which were classified following an HCA, which took account of the type of institution and capacity bracket variables. These three groups form the sampling strata.
During the second phase, the aim is to achieve a distribution of senior citizens in the sample that is identical to the distribution of all residents in institutions (in the field selected). Samples are drawn from all of the departments previously selected. To enable representative sampling of the institutions, balanced sampling was performed using the SAS CUBE macro according to the institution category and legal status variables, taking account of the probability of the institution being included in the sample, while also complying with the fixed size sampling constraint (same number of establishments per department within the sample).

14. In the Île-de-France region, this balance is supplemented by other variables (concerning demographic structure and dwelling type in particular).

REFERENCES

[1] Deville, J.-C. & Tillé, Y. (2004). Efficient Balanced Sampling : The Cube Method. *Biometrika*, Vol 91, No 4, pp 893-912.

[2] Deville, J.-C. & Tillé, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference*, Vol 128, pp 569-591.

[3] Tillé, Y. (2011). Dix années d'échantillonnage équilibré par la méthode du cube : une évaluation. *Techniques d'enquête*, Vol. 37, No 2, pp. 233-246.

[4] Rousseau, S. & Tardieu, F. (2004). La macro SAS CUBE d'échantillonnage équilibré, Documentation de l'utilisateur. *Rapport technique*, Insee, Paris.

[5] Ardilly, P. (2006). *Les techniques de sondages*. Éditions Technip, Paris.

[6] Chauvet, G. (2012). Estimation de variance pour le nouvel Échantillon-Maître, *Journées de Méthodologie Statistique*, Paris.

[7] Christine, M. & Faivre, S. (2009). OCTOPUSSE : un système d'Échantillon- Maître pour le tirage des échantillons dans la dernière EAR, *Journées de Méthodologie Statistique*, Paris.

Département des méthodes statistiques
Version n° 1, diffusée le 21 juin 2017.