Construction d'un échantillonnage équilibré avec remise

Jean Rubin [1] & Guillaume Chauvet [2]

[1] *Insee, France, jean.rubin@insee.fr*
[2] *Ensai (Irmar), France, guillaume.chauvet@ensai.fr*

Résumé. L'équilibrage est une approche relativement naturelle pour tirer parti d'une information connue afin d'avoir de meilleures estimations à coût réduit (Deville and Tillé, 2004, 2005). Elle est notamment employée en France pour la conception de l'Échantillon Maître, qui est un échantillon géographique de premier degré, représentatif de l'ensemble du territoire, et utilisé entre autres pour le recensement de la population française. Une manière classique pour produire un échantillon équilibré consiste à utiliser la méthode dite "du cube" proposée par Deville et Tillé. La méthode du cube se concentre toutefois sur la construction d'échantillons sans remise. Nous présentons des méthodes permettant d'effectuer un tirage équilibré avec remise en généralisant la méthode du cube de plusieurs manières. A l'aide de ces méthodes, il serait possible d'envisager des estimations de variance bootstrap, simples d'utilisation, se basant sur un rééchantillonnage des observations. Nous étudierons les propriétés de ces méthodes par simulation.

Mots-clés. Méthode du Cube, Estimation de variance, Entropie, Bootstrap

Abstract. Balanced sampling is a relatively natural approach to use known information in order obtain better estimates at reduced cost (Deville and Tillé, 2004, 2005). It is used in particular for the design of the French Master Sample, which is a geographical sample representative of the entire territory and used for the census of the French population. The methods generally used for producing balanced samples are based on the so-called "cube" method proposed by Deville and Tillé, 2004. However, the cube method can only produce samples without replacement. This presentation therefore aims to showcase methods for producing with-replacement balanced samples, by generalizing the cube methods in several ways. With these methods, it could be possible to consider easy-to-use bootstrap variance estimates based on the resampling of units. We will study the properties of these methods by simulation.

Keywords. Cube method, Variance estimation, Entropy, Bootstrap

1 With-replacement cube method

Let U denote a finite population of size N. Let S denote a random sample of size n, selected by means of a probabilistic sampling design $p(\cdot)$. The sampling design may be with replacement, so that S may be with repetitions. Let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)^\top$ denote a vector of probabilities, where π_k stands for the expected number of times that the unit k is selected in the sample S and $\sum_{k\in U}\pi_k = n$. The Hansen-Hurvitz estimator is given by

$$\hat{t}_{y,HH} = \sum_{k\in U} w_k \frac{y_k}{\pi_k} = \sum_{k\in S} \frac{y_k}{\pi_k} \tag{1}$$

with w_k the number of selections of unit k in the sample S. If the sampling design is without-replacement, $\hat{t}_{y,HH}$ is simply the Horvitz-Thompson estimator.

Let $\mathbf{x}_k \in \mathbb{R}^q$ denote a q-vector of variables known for any unit $k \in U$. A sample is balanced on $\mathbf{x}_k$ if the balancing equations

$$\sum_{k\in U} w_k \frac{\mathbf{x}_k}{\pi_k} = \sum_{k\in U} \mathbf{x}_k \tag{2}$$

hold, and the whole sampling design is balanced on $\mathbf{x}_k$ if the balancing equations (2) hold p-almost surely. In general, it may not be possible to satisfy the constraints of equations (2), and a sample can be only approximately balanced. Note that the constraints for balanced sampling designs are the same for without-replacement designs, the only difference being that in this case the multiplicities correspond to indicators of selection, i.e. $w_k \in \{0, 1\}$.

The cube method is a well known approach developped by Deville and Tillé (2004) to generate without-replacement balanced samples. The first phase of the method is called the flight phase and is described in Algorithm 1. At the end of this stage, the balanced equations are satisfied but some units are left undecided since w_k is not an integer for at most q units. A second part of the algorithm called the "landing phase" will decide whether to chose or not the undecided units, while approximately satisfying the balanced equations.

There are several possibilities to apply balanced sampling at Step 2 of Algorithm 1:

- The sample may be selected by applying the ordered cube method, i.e. by applying the cube method to the population U ranked in a specific order (e.g., option `order=2` in the `samplecube()` function of the `sampling` package from Tillé and Matei, 2023).
- The sample may be selected by applying the randomized cube method, i.e. by applying the cube method to the population U which is randomly ordered (e.g., option `order=1` in the `samplecube()` function of the `sampling` package).

One simple approach to extend this algorithm to produce a with-replacement version of the cube method is to apply the cube method from Deville and Tillé, 2004 to a population with duplicated individuals, as described in Algorithm 2. Following Algorithm 2, it is straightforward to verify that we have

$$\forall k \in U,\ E_p(w_k) = n\frac{\pi_k}{n} = \pi_k,$$

and

$$\sum_{k\in S} \frac{\mathbf{x}_k}{\frac{\pi_k}{n}} = n\sum_{k\in U} \mathbf{x}_k, \tag{3}$$

Algorithm 1 Deville-Tillé Cube method (Flight phase)

1. Initialize $\boldsymbol{\pi}(0) = \boldsymbol{\pi}$.

2. At time $t = 1, \dots, T$, repeat the following three steps:

   (a) Generate any vector $\mathbf{u}(t) \in \mathbb{R}^N$, such that $\mathbf{u}(t)$ is in the kernel of the $q \times N$ matrix $\boldsymbol{A} = (\mathbf{x}_1/\pi_1, \dots, \mathbf{x}_N/\pi_N)$, and $\forall k \in U,\ u_k(t) = 0$ if $\pi_k(t-1)$ is an integer. The vector $\mathbf{u}(t)$ can be chosen randomly or deterministically but $\mathbf{u}(t)$ must be linearly independent of $\boldsymbol{\pi}(t-1), \dots, \boldsymbol{\pi}(1)$.

   (b) Compute $\lambda_1^*(t) \in \mathbb{R}$ and $\lambda_2^*(t) \in \mathbb{R}$, the largest values of $\lambda_1(t)$ and $\lambda_2(t)$ such that

   $$\forall k \in U,\ 0 \le \pi_k(t-1) + \lambda_1(t) u_k(t) \le 1, \quad 0 \le \pi_k(t-1) - \lambda_2(t) u_k(t) \le 1$$

   Note that $\lambda_1(t) > 0$ and $\lambda_2(t) > 0$.

   (c) Select

   $$\boldsymbol{\pi}(t) = \begin{cases} \boldsymbol{\pi}(t-1) + \lambda_1^*(t)\mathbf{u}(t) & \text{with probability } \alpha(t), \\ \boldsymbol{\pi}(t-1) - \lambda_2^*(t)\mathbf{u}(t) & \text{with probability } 1-\alpha(t), \end{cases}$$

   where $\alpha(t) = \lambda_2^*(t)/[\lambda_1^*(t) + \lambda_2^*(t)]$.

3. The procedure stops at time T when it is no longer possible to carry out Step 2. $\mathbf{w} = \boldsymbol{\pi}(T)$ is the result of the procedure. Note that some units are left undecided.

---

and the balancing equations (2) are therefore respected. This method is a generalization of multinomial sampling (Tillé, 2011, section 5.4), which is a particular case obtained when $\mathbf{x}_k = \pi_k$. This method has also multiple variants, depending on the option (population with a fixed or random order) used to apply the cube method. Furthermore, the sample may be selected by applying stratified balanced sampling (Chauvet, 2009; Hasler and Tillé, 2014): when applying the cube algorithm, the population U^n is stratified in n strata, each stratum being associated to one duplication of the population U^n. One flight phase is first applied in each stratum, and a final flight phase followed by a landing phase are applied on the remaining units (e.g., function `balancedstratification()` of the `sampling` package).

Algorithm 2 With-replacement cube method (duplication approach)

1. Create a product population U^n by duplicating n times each unit $k \in U$.

2. Select a sample S in U^n, with inclusion probabilities $(\pi_k/n)_{k\in U}$, by balanced sampling on the vectors $(\mathbf{x}_k)_{k\in U}$.

In the following sections, we focus on the flight phase of the cube algorithm, which is the central part of balanced sampling. This means that we allow the multiplicities w_k to take non-integer values, even though it is supposed to represent the number of selections of a unit.

1.1 A general form of with-replacement flight phase

A general form of with-replacement flight phase that does not require to increase the size of the population is described in Algorithm 3.

Algorithm 3 With-replacement cube method (no dimensional increase, flight phase)

---

1. Initialize $\boldsymbol{\pi}(0) = \boldsymbol{\pi}$.
2. At time $t = 1, \ldots, T$, repeat the following four steps:
   (a) Generate any vector $\mathbf{u}(t) \in \mathbb{R}^N$, such that $\mathbf{u}(t)$ is in the kernel of the matrix $\boldsymbol{A} = (\mathbf{x}_1/\pi_1, \ldots, \mathbf{x}_N/\pi_N)$, and $u_k(t) = 0$ if $\pi_k(t-1)$ is an integer. The vector $\mathbf{u}(t)$ can be chosen randomly or deterministically but $\mathbf{u}(t)$ must be linearly independent of $\boldsymbol{\pi}(t-1), \ldots, \boldsymbol{\pi}(1)$.
   (b) Consider every $\lambda_m(t) \in \mathbb{R}$ such that $\boldsymbol{\pi}(t-1)+\lambda_m(t)\mathbf{u}(t)$ obtain a new non-negative integer coordinate. The number of such $\lambda_m(t)$ can vary with t. We will index them as $\{\lambda_m(t)\}_{m=1}^{M(t)}$.
   (c) Select $\boldsymbol{\pi}(t) = \boldsymbol{\pi}(t-1) + \lambda_m(t)\mathbf{u}(t)$ with probability $\alpha_m(t)$. These probabilities $\alpha_m(t)$ must satisfy the martingale constraint:

$$\sum_{m=1}^{M(t)} \lambda_m(t)\alpha_m(t) = 0 \tag{4}$$

   and $\sum_{m=1}^{M(t)} \alpha_m(t) = 1$.
3. The procedure stops at time T when it is no longer possible to carry out Step 2a. $\mathbf{w} = \boldsymbol{\pi}(T)$ is the result of the procedure. Note that some units are left undecided.

This algorithm corresponds in fact to a family of algorithms, in the sense that there are many degrees of freedom in the way to realise it:

- In step 2.a, many choices of vector $\mathbf{u}(t)$ are valid. The standard cube algorithm does not specify precisely a choice of $\mathbf{u}(t)$ either, so this option is also let as general as possible in Algorithm 3.
- For the same reasons for which there exists two constants $\lambda_1(t)$ and $\lambda_2(t)$ in Step 2.b of Algorithm 1, there will always exists $M(t) \geq 2$ and a set of probabilities $(\alpha_m(t))$ to carry out Steps 2.b and 2.c in Algorithm 3. In general, there are actually many ways to select $\alpha_m(t)$ when $M(t) \geq 3$. If $\alpha_m(t)$ is strictly positive for only two λ_m, where one is positive and the other negative, we recover the weights of the cube algorithm, i.e. $\alpha_m(t)$ must be proportional to $|\lambda_m(t)|^{-1}$. It is also possible that there is in fact an infinite number of $\lambda_m(t)$, i.e. $M(t)$ could be infinite. One can check that if the inclusion probability is included in the balancing variables, then $M(t)$ is finite at each step.

When there are multiple ways to select the $\alpha_m(t)$, one natural choice would be the ones that maximize entropy, i.e. which are the solution of the following optimization problem :

$$\max \left\{ -\sum_{m=1}^{M(t)} \alpha_m(t) \ln[\alpha_m(t)] \right\},$$
$$\text{with } \sum_{m=1}^{M(t)} \alpha_m(t) = 1,$$
$$\text{and } \sum_{m=1}^{M(t)} \lambda_m(t)\alpha_m(t) = 0.$$

We can prove that the solution of this optimization problem is such that

$$\alpha_m(t) \propto \exp\left\{-\gamma(t)\lambda_m(t)\right\},$$

where $\gamma(t)$ is chosen such that

$$\sum_{m=1}^{M(t)} \lambda_m(t) \exp\left\{-\gamma(t)\lambda_m(t)\right\} = 0$$

1.2 With-replacement cube sampling by relaxing upper constraints

Algorithm 4 is a particular case of Algorithm 3, which is very close from the original without-replacement cube method described in Deville and Tillé, 2004. The idea is to only keep the constraint of having the multiplicities be positive, thus removing the constraint of them being less than one.

In case of a fixed-size sampling design, i.e. if π_k belongs to the balancing variables, we have $\sum_{k\in U} u_k(t) = 0$ and there are therefore at least one unit k such that $u_k(t) < 0$, and one unit l such that $u_l(t) > 0$ at each time t. The quantities $\lambda_1^*(t)$ and $\lambda_2^*(t)$ are therefore well defined. Moreover, since $\sum_{k\in U} \pi_k(t) = n$ and $\pi_k(t) \geq 0$ for any $k \in U$ from (5), the components of $\boldsymbol{\pi}(t)$ are no greater than n. Note that equation 5 implies that one of the new coordinates must be 0, i.e. at each step the algorithm randomly decides which unit **not** to choose.

1.3 Variance approximation

1.3.1 Theoretical assumptions

We propose a variance approximation for with-replacement balanced sampling, which is based on the following assumptions/conjectures:

C1: For any vector $\mathbf{p} = (p_1, \ldots, p_N)^\top$ such that $\sum_{k\in U} p_k = 1$, let $Mult(\mathbf{p}, n)$ denote a multinomial sampling of size n with drawing probabilities $\mathbf{p}$ (see Tillé, 2011). We suppose

Algorithm 4 With-replacement cube method (exhaustion approach, flight phase)

1. Initialize at $\boldsymbol{\pi}(0) = \boldsymbol{\pi}$ and $\mathbf{w}(0) = (0, \ldots, 0)^\top$.

2. At time $t = 1, \ldots, T$, repeat the following four steps:

   (a) Generate any vector $\mathbf{u}(t) \in \mathbb{R}^N$, such that $\mathbf{u}(t)$ is in the kernel of the matrix $\boldsymbol{A} = (\mathbf{x}_1/\pi_1, \ldots, \mathbf{x}_N/\pi_N)$, and $u_k(t) = 0$ if $\pi_k(t-1)$ is an integer. The vector $\mathbf{u}(t)$ can be chosen randomly or deterministically but $\mathbf{u}(t)$ must be linearly independent of $\boldsymbol{\pi}(t-1), \ldots, \boldsymbol{\pi}(1)$.

   (b) Compute $\lambda_1^*(t)$ and $\lambda_2^*(t)$ the largest values of $\lambda_1(t)$ and $\lambda_2(t)$ such that

   $$\forall k \in U,\ 0 \leq \pi_k(t-1) + \lambda_1(t) u_k(t) \quad \text{and} \quad 0 \leq \pi_k(t-1) - \lambda_2(t) u_k(t). \tag{5}$$

   Note that

   $$\lambda_1^*(t) = \min_{k; u_k(t) < 0} \left\{ -\frac{\pi_k(t-1)}{u_k(t)} \right\} > 0 \quad \text{and} \quad \lambda_2^*(t) = \min_{k; u_k(t) > 0} \left\{ \frac{\pi_k(t-1)}{u_k(t)} \right\} > 0.$$

   (c) Select

   $$\boldsymbol{\phi}(t) = \begin{cases} \boldsymbol{\pi}(t-1) + \lambda_1^*(t)\mathbf{u}(t) & \text{with probability } \alpha(t), \\ \boldsymbol{\pi}(t-1) - \lambda_2^*(t)\mathbf{u}(t) & \text{with probability } 1 - \alpha(t), \end{cases}$$

   where $\alpha(t) = \lambda_2^*(t)/[\lambda_1^*(t) + \lambda_2^*(t)]$.

   (d) Take

   $$\begin{aligned} \boldsymbol{\pi}(t) &= \boldsymbol{\phi}(t) - \lfloor \boldsymbol{\phi}(t) \rfloor, \\ \mathbf{w}(t) &= \mathbf{w}(t-1) + \lfloor \boldsymbol{\phi}(t) \rfloor, \end{aligned}$$

   where $\lfloor \cdot \rfloor$ stands for the coordinate-wise integer part.

3. The procedure stops at time T when it is no longer possible to carry out Step 2. We take $\mathbf{w} = \mathbf{w}(T) + \boldsymbol{\pi}(T)$ as the result of the procedure. Note that some units are left undecided.

that for any vector $\boldsymbol{\pi}$, there exists some vector $\mathbf{p}$ such that $Mult(\mathbf{p}, n)$ conditioned by equations (2), which we denote as $Mult(\mathbf{p}, n \left| \hat{t}_{\mathbf{x},HH} = t_{\mathbf{x}} \right.)$, leads to $\boldsymbol{\pi}$ as the vector of expected numbers of draws.

We also suppose that these probabilities are such that:

$$\forall k \in U, \ p_k \ \simeq \ \pi_k/n. \tag{6}$$

C2: Under $Mult(\mathbf{p}, n)$, the vector $(\hat{t}_{y,HH}, \hat{t}_{\mathbf{x},HH}^\top)^\top$ is approximately normally distributed with covariance matrix

$$V \ = \ \begin{pmatrix} V_y & C_{xy}^\top \\ C_{xy} & V_x \end{pmatrix}, \tag{7}$$

where

$$\begin{aligned} V_y &= \sum_{k \in U} np_k \left(\frac{y_k}{np_k} - \frac{t_y}{n} \right)^2, \\ V_x &= \sum_{k \in U} np_k \left(\frac{\mathbf{x}_k}{np_k} - \frac{t_{\mathbf{x}}}{n} \right) \left(\frac{\mathbf{x}_k}{np_k} - \frac{t_{\mathbf{x}}}{n} \right)^\top, \\ C_{xy} &= \sum_{k \in U} np_k \left(\frac{\mathbf{x}_k}{np_k} - \frac{t_{\mathbf{x}}}{n} \right) \left(\frac{y_k}{np_k} - \frac{t_y}{n} \right). \end{aligned} \tag{8}$$

Also,

$$\begin{aligned} V\left(\hat{t}_{y,HH} \left| \hat{t}_{\mathbf{x},HH} = t_{\mathbf{x}} \right.\right) &\simeq V_y - C_{xy}^\top V_x^{-1} C_{xy} \\ &= \sum_{k \in U} np_k \left(\frac{E_k}{np_k} - \frac{t_E}{n} \right)^2, \end{aligned} \tag{9}$$

where $E_k = y_k - B^T \mathbf{x}_k$ and $B = V_x^{-1} C_{xy}$.

The first part of $C1$ could be stated as an assumption, but we could also study if the result holds generally for any $\boldsymbol{\pi}$. A similar result was proved by Chen, Dempster, and Liu (1994) for unequal probability without-replacement sampling of fixed size, and Deville and Tillé, 2005 for without-replacement balanced sampling. The second part on the approximation of the p_k's in equation (6) could be proved by using a multivariate Edgeworth expansion, but this is quite intricate. Concerning $C2$, the asymptotic normality holds true under some regularity assumptions on the probabilities p_k and the vector of variables $(y_k, \mathbf{x}_k^\top)$. Equation (8) is a general result for normal distributions, so it should hold approximately (see Fuller, 2011). Under these hypotheses, we therefore obtain from equation (9) a variance approximation $\mathbf{V}_{\text{multi}}$ for with-replacement balanced sample, which has a similar form as an approximation proposed by Deville and Tillé (2004), which is reminded below:

$$\begin{aligned} \mathbf{V}_{\text{Deville-Tille}} &= \frac{N}{N-p} \sum_{k \in U} \frac{1-\pi_k}{\pi_k} \tilde{E}_k^2 \\ \text{with } \tilde{E}_k &= y_k - \tilde{B}^T \mathbf{x}_k \\ \text{and } \tilde{B} &= \left(\sum_{k \in U} \frac{1-\pi_k}{\pi_k} \mathbf{x}_k \mathbf{x}_k^T \right)^{-1} \sum_{k \in U} \frac{1-\pi_k}{\pi_k} \mathbf{x}_k y_k. \end{aligned}$$

1.3.2 Experimental Setup

Population We fix a population size of $N = 1000$, and generate iid observations of $(\mathbf{x}_k, \varepsilon_k)_{k\in U}$ such that:

$$\begin{aligned} x_k^1 &\sim \mathcal{N}(5, 1^2) \\ x_k^2 &\sim \mathcal{N}(3, 1^2) \\ \varepsilon_k &\sim \mathcal{N}(0, \eta^2) \end{aligned}$$

using varying level of noises $\eta \in \{0.1, 0.5, 1, 2, 5, 10\}$, thus creating 6 different sets of population. We then construct a variable of interest y_k such that

$$y_k = 2x_k^1 + \varepsilon_k$$

The R^2 associated to the linear regression of y_k against x_k^1 are displayed in Table 1.

Sampling procedure For a fixed population, we choose the inclusion probabilities $\pi_k = 1/10$ for every unit $k \in U$ and we generate samples (with eventually undecided units) $S \sim f(\boldsymbol{X}; \boldsymbol{\pi})$ using different types of sampling procedure f that can use $\boldsymbol{X} = (\mathbf{1}, \boldsymbol{\pi}, \mathbf{x})$ as auxiliary variables. More precisely, the sampling procedure f could be one of {SRSWOR, Base, Base Flight, WR Exhaustion, WR Copy, WR Entropy}, respectively corresponding to a sampling without replacement of fixed size $n = 100$, the basic cube method of Deville and Tillé, 2004(flight and landing phase), the basic cube method (only the flight phase), algorithm 4, algorithm 2 and algorithm 3 maximizing entropy at each step[1]. We then compute the variances of $\hat{t}_y$ by a Monte-Carlo procedure, using $T = 1000$ samples.

Likewise, we compute approximation of variance among one of $\{\mathbf{V}_{\text{Deville-Tille}}, \mathbf{V}_{\text{multi}}\}$. Results are in Table 1.

2 Future work

We provided three different approaches to produce with-replacement balanced samples, as well as a variance approximation for these types of sampling. More empirical tests are needed to conclude about the viability of each method for bootstrap variance estimation.

[1]The SRSWOR does not use the auxiliary variables in $\mathbf{X}$, but it could be interpreted as a balanced sample on the inclusion probabilities.

Table 1: Influence of noise on variance

			Standard Cube method			With-replacement Cube method			
η	R^2	SRSWOR	Base	Base Flight	$\mathbf{V}_{\text{Deville-Tille}}$	WR Exhaustion	WR Copy	WR Entropy	$\mathbf{V}_{\text{multi}}$
0.1	0.9970	32,900	519	102	98.6	103	151	119	106
0.5	0.9360	35,200	2,610	2,120	2,350.0	2540	2,640	2,910	2,570
1.0	0.7870	39,300	10,200	9,210	9,660.0	10,300	10,400	11,200	10,200
2.0	0.4960	75,100	36,100	35,900	36,300.0	42,900	40,600	40,700	39,300
5.0	0.1170	272,000	235,000	223,000	237,000.0	281,000	249,000	280,000	257,000
10.0	0.0358	855,000	844,000	811,000	927,000.0	1,140,000	963,000	1,100,000	943,000

References

Chauvet, Guillaume (2009). "Stratified balanced sampling". In: *Survey Methodology* 35.1, pp. 115–119.

Chen, Xiang-Hui, Arthur P Dempster, and Jun S Liu (1994). "Weighted finite population sampling to maximize entropy". In: *Biometrika* 81.3, pp. 457–469.

Deville, Jean-Claude and Yves Tillé (2004). "Efficient balanced sampling: the cube method". In: *Biometrika* 91.4, pp. 893–912.

Deville, Jean-Claude and Yves Tillé (2005). "Variance approximation under balanced sampling". In: *Journal of statistical planning and Inference* 128.2, pp. 569–591.

Fuller, Wayne A (2011). *Sampling statistics*. Vol. 560. John Wiley & Sons.

Hasler, Caren and Yves Tillé (2014). "Fast balanced sampling for highly stratified population". In: *Computational Statistics & Data Analysis* 74, pp. 81–94.

Tillé, Yves (2011). *Sampling algorithms*. Springer.

Tillé, Yves and Alina Matei (2023). *sampling: Survey Sampling*. R package version 2.10.