The cube method

Yves Tillé
University of Neuchâtel

May 19, 2006

Idea and History

- Idea : Same means in the population and the sample for all the auxiliary variables.
- Balanced sampling $\neq$ purposive selection
- Random balanced sampling
- Yates (1949),Thionet (1953),Royall and Herson (1973),Deville, Grsbras and Roth (1988),Ardilly (1991)Hedayat and Majumar (1995),Brawer (1999)Deville and Tillé (2004), Deville and Tillé (2005),

Notation

- Auxiliary variables $x_1, \dots, x_p$, known for each unit of the population.
- $\mathbf{x}_k = (x_{k1}, \dots, x_{kp})'$, is known for all $k \in U$.
- The vector of totals $\mathbf{X} = \sum_{k \in U} \mathbf{x}_k$.
- The Horvitz-Thompson estimator of the vector of totals

$$\widehat{\mathbf{X}}_\pi = \sum_{k \in S} \frac{\mathbf{x}_k}{\pi_k}.$$

- The aim is always to estimate $\widehat{Y}_\pi = \sum_{k \in S} \frac{y_k}{\pi_k}$.

Definition

- Definition

  A sampling design $p(s)$ is said to be balanced on the auxiliary variables $x_1, ..., x_p$, if and only if it satisfies the balancing equations given by $\widehat{\mathbf{X}}_\pi = \mathbf{X}$, which can also be written

$$\sum_{k \in s} \frac{x_{kj}}{\pi_k} = \sum_{k \in U} x_{kj},$$

  for all $s \in \mathcal{S}$ such that $p(s) > 0$, and for all $j = 1, ..., p$, or in other words

$$\text{Var}\left(\widehat{\mathbf{X}}_\pi\right) = 0.$$

Example 1

- A sampling design of fixed sample size n is balanced on the variable $x_k = \pi_k, k \in U$. Indeed,

$$\sum_{k \in S} \frac{x_k}{\pi_k} = \sum_{k \in S} 1 = \sum_{k \in U} \pi_k = n.$$

Example 2

- Stratification with strata $U_h, h = 1, ..., H$, $\# U_h = N_h$
  Simple random sample of size n_h in each stratum
  The design is balanced on variables δ_{kh} of values

$$\delta_{kh} = \begin{cases} 1 & \text{if } k \in U_h \\ 0 & \text{if } k \notin U_h. \end{cases}$$

Indeed $\sum_{k \in S} \frac{\delta_{kh}}{\pi_k} = \sum_{k \in S} \delta_{kh} \frac{N_h}{n_h} = N_h$, for $h = 1, ..., H$.

Example 3

- $N = 10, n = 7, \pi_k = 7/10, k \in U,$
  $x_k = k, k \in U.$

  $$\sum_{k \in S} \frac{k}{\pi_k} = \sum_{k \in U} k,$$

  which gives that

  $$\sum_{k \in S} k = 55 \times 7/10 = 38.5,$$

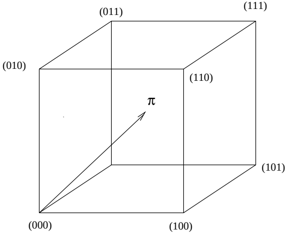
  IMPOSSIBLE: Rounding problem.
- Aim: find a sample approximately balanced!

Cube representation

- Geometric representation of a sampling design.

$$\mathbf{s} = (I[1 \in s] \ \dots \ I[k \in s] \ \dots \ I[N \in s])',$$

where $I[k \in s]$ takes the value 1 if $k \in s$ and 0 if not.



Possible samples in a population of size $N = 3$

Cube representation

- Geometrically, each vector $\mathbf{s}$ is a vertex of a N-cube.

$$E(\mathbf{s}) = \sum_{s \in \mathcal{S}} p(\mathbf{s})\mathbf{s} = \boldsymbol{\pi},$$

where $\boldsymbol{\pi} = [\pi_k]$ is the vector of inclusion probabilities.

Balancing equations

- The balancing equations
$$\sum_{k\in S}\frac{\mathbf{x}_k}{\pi_k}=\sum_{k\in U}\mathbf{x}_k,$$
can also be written
$$\sum_{k\in U}\mathbf{a}_k s_k=\sum_{k\in U}\mathbf{a}_k\pi_k \text{ with } s_k\in\{0,1\}, k\in U,$$
where $\mathbf{a}_k=\mathbf{x}_k/\pi_k, k\in U.$
- The balancing equations defines a linear subspace in $\mathbb{R}^N$ of dimension $N-p$ denoted Q.
- **The problem:** Choose a vertex of the N-cube (a sample) that remains on the linear sub-space Q.

System exactly verifiable

Example

$\pi_1 + \pi_2 + \pi_3 = 2.$

$x_k = \pi_k, k \in U$ and $\sum_{k \in U} s_k = 2.$

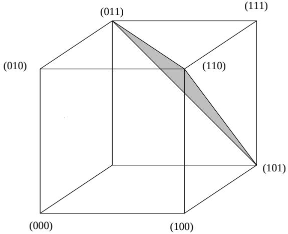
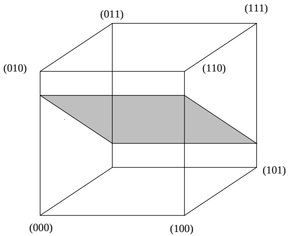


Figure: Fixed size constraint: all the vertices of K are vertices of the cube

System approximately verifiable

Example

- $6 \times \pi_2 + 4 \times \pi_3 = 5.$
- $x_1 = 0, x_2 = 6 \times \pi_2$ and $x_3 = 4 \times \pi_3.$



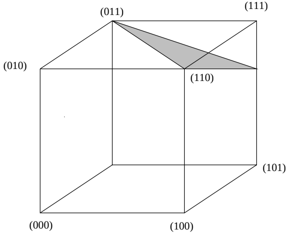
System sometimes verifiable

Example

$\pi_1 + 3 \times \pi_2 + \pi_3 = 4.$

$x_1 = \pi_1, x_2 = 3 \times \pi_2$ and $x_3 = \pi_3.$

$s_1 + 3s_2 + s_3 = 4.$



Cube methods: phases

- **Cube method** (Deville and Tillé, 2004)
  1. flight phase
  2. landing phase (needed only it there exists a rounding problem)
- **The flight phase** is a random walk that begins at the vector of inclusion probabilities and remains in the intersection of the cube and the constraint subspace.
  This random walk stops at a vertex of the intersection of the cube and the constraint subspace.
- **The landing phase** At the end of the flight phase, if a sample is not obtained, a sample is selected as close as possible to the constraint subspace.

Cube methods: examples

- Example

  The constraints is the fixed sample size. The flight phase transforms a vector of inclusion probabilities into a vector of 0 and 1.

$$\pi = \begin{pmatrix} 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \end{pmatrix} \rightarrow \begin{pmatrix} 0.6666 \\ 0.6666 \\ 0.6666 \\ 0 \end{pmatrix} \rightarrow \begin{pmatrix} 1 \\ 0.5 \\ 0.5 \\ 0 \end{pmatrix} \rightarrow \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix} = \mathbf{S}.$$

  Maximum $N - p$ steps.

Cube methods: examples

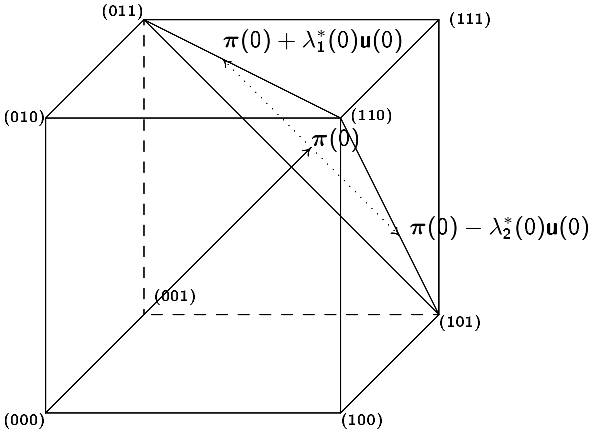
- Example

  If there exists a rounding problem, then some components cannot be put to zero.

$$\boldsymbol{\pi} = \begin{pmatrix} 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \end{pmatrix} \rightarrow \begin{pmatrix} 0.625 \\ 0 \\ 0.625 \\ 0.625 \\ 0.625 \end{pmatrix} \rightarrow \begin{pmatrix} 0.5 \\ 0 \\ 0.5 \\ 1 \\ 0.5 \end{pmatrix} \rightarrow \begin{pmatrix} 1 \\ 0 \\ 0.25 \\ 1 \\ 0.25 \end{pmatrix} \rightarrow \begin{pmatrix} 1 \\ 0 \\ 0.5 \\ 1 \\ 0 \end{pmatrix} = \boldsymbol{\pi}^*.$$

In this case, the flight phase let one non-integer components.

Balancing martingale



Balancing martingale

- Definition

  A discrete time stochastic process $\boldsymbol{\pi}(t) = [\pi_k(t)], t = 0, 1, \ldots$ in $\mathbb{R}^N$ is said to be a balancing martingale for an inclusion probability vector $\boldsymbol{\pi}$ and the auxiliary variables $x_1, \ldots, x_p$, if

  1. $\boldsymbol{\pi}(0) = \boldsymbol{\pi}$,
  2. $E\left[\boldsymbol{\pi}(t)|\boldsymbol{\pi}(t-1), \ldots., \boldsymbol{\pi}(0)\right] = \boldsymbol{\pi}(t-1), t = 1, 2, \ldots$
  3. $\boldsymbol{\pi}(t) \in K = \left\{[0,1]^N \cap (\boldsymbol{\pi} + \text{Ker } \mathbf{A})\right\}$, where $\mathbf{A}$ is the $p \times N$ matrix given by $\mathbf{A} = (\mathbf{x}_1/\pi_1 \ldots \mathbf{x}_k/\pi_k \ldots \mathbf{x}_N/\pi_N)$.
  4. In other words, a balancing martingale is such that $\boldsymbol{\pi}(t-1)$ is in the center of the following possible values of $\boldsymbol{\pi}(t)$.

Balancing martingale

- If $\pi(t)$ is a balancing martingale, then
  (i) $E[\pi(t)] = E[\pi(t-1)] = \dots = E[\pi(0)] = \pi$.
  (ii) $\sum_{k \in U} \mathbf{a}_k \pi_k(t) = \sum_{k \in U} \mathbf{a}_k \pi_k = \mathbf{X}, t = 0, 1, 2, \dots$.
  (iii) When the balancing martingale reaches a face of K, it remains "stuck" on this face.

Flight Phase

First initialize with $\boldsymbol{\pi}(0) = \boldsymbol{\pi}$. Next, at time $t = 1, ..., T$,

1. Generate any vector $\mathbf{u}(t) = [u_k(t)] \neq 0$ such that
   (i) $\mathbf{u}(t)$ is in the kernel of matrix $\mathbf{A}$
   (ii) $u_k(t) = 0$ if $\pi_k(t)$ is integer.
2. Compute $\lambda_1^*(t)$ and $\lambda_2^*(t)$, the largest values such that
   $0 \leq \boldsymbol{\pi}(t) + \lambda_1(t)\mathbf{u}(t) \leq 1,$
   $0 \leq \boldsymbol{\pi}(t) - \lambda_2(t)\mathbf{u}(t) \leq 1.$
3. Compute
   $$\boldsymbol{\pi}(t) = \begin{cases} \boldsymbol{\pi}(t-1) + \lambda_1^*(t)\mathbf{u}(t) & \text{with a proba } q_1(t) \\ \boldsymbol{\pi}(t-1) - \lambda_2^*(t)\mathbf{u}(t) & \text{with a proba } q_2(t), \end{cases}$$
   where $q_1(t) = \lambda_2^*(t)/\{\lambda_1^*(t) + \lambda_2^*(t)\}$ and $q_2(t) = 1 - q_1(t)\}$.

Landing Phase 1

- Let $\boldsymbol{\pi}^* = [\pi_k^*]$ the vector obtained at the last step of the flight phase.
Inclusion probabilities	Flight Phase	Landing phase
$\boldsymbol{\pi}$	$\rightarrow \boldsymbol{\pi}^*$	$\rightarrow S$
- It is possible to proof that

$$\operatorname{card} U^* = \operatorname{card}\left\{k \in U | 0 < \pi_k^* < 1\right\} = q \leq p.$$

- The aim of the landing phase is to find a sample **S** such that $E(\mathbf{S}|\boldsymbol{\pi}^*) = \boldsymbol{\pi}^*$, and that is almost balanced.
- Solution: linear program defined only on $q \leq p$ units.

Landing Phase 2

- If the number of auxiliary variables is too large for the linear program to be solved by a simplex algorithm, $q > 13$ then, at the end of the flight phase, an auxiliary variable can be dropped.
- Next, one can return to the flight phase until it is no longer possible to 'move' within the constraint subspace. The constraints are thus relaxed successively.

Main applications

- New French census
  - For the mulicipalities < 10000 inhab., selection of 5 rotations groups of municipalities.
  - For the mulicipalities > 10000 inhab., selection of 5 rotations groups of addresses.
- Master sample in France: selection of the primary units.

Example: 245 municipalities of the Swiss Ticino canton

Table: Balancing variables of the population of municipalities of Ticino

POP	number of men and women
ONE	constant variable that takes always the value 1
ARE	area of the municipality in hectares
POM	number of men
POW	number of women
P00	number of men and women aged between 0 and 20
P20	number of men and women aged between 20 and 40
P40	number of men and women aged between 40 and 65
P65	number of men and women aged between 65 and over
HOU	number of households

Example: sampling design

- Inclusion probabilities proportional to size.

- Big municipalities are always in the sample Lugano, Bellinzona, Locarno, Chiasso, Pregassona, Giubiasco, Minusio, Losone, Viganello, Biasca, Mendrisio, Massagno.
- Sample size = 50.
- the population totals for each variable X_j,
- the estimated total by the Horvitz-Thompson estimator $\widehat{X}_{j\pi}$,
- the relative deviation in % defined by

$$\text{RD} = 100 \times \frac{\widehat{X}_{j\pi} - X_j}{X_j}.$$

Example: Results

Table: Quality of balancing

Variable	Population total	HT-Estimator	Relative deviation in %
POP	306846	306846.0	0.00
ONE	245	248.6	1.49
HA	273758	276603.1	1.04
POM	146216	146218.9	0.00
POW	160630	160627.1	-0.00
P00	60886	60653.1	-0.38
P20	86908	87075.3	0.19
P40	104292	104084.9	-0.20
P65	54760	55032.6	0.50
HOU	134916	135396.6	0.36

FAQ

- Why not use calibration in place of balancing?
  Stratification is a particular case of balancing, post-stratification is a particular case of calibration. In stratification and balancing, the weights does not become random.
- How accurate is the approximation with the cube method?

$$\left|\frac{\widehat{X}_j - X_j}{X_j}\right| < O(p/N) \leq O_p(\sqrt{1/n}).$$

- What is the limit for the size of the population? *If depends on the program: N=200000, p=40 is possible.*
- How to estimate the variance?
  By a residual technique see Deville and Tillé (2005)
- What is the best strategy, balancing of calibration? *Both techniques can be used together.*

References

Deville, J.-C. and Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, 91:893–912.

Deville, J.-C. and Tillé, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference*, 128:411–425.

Tillé, Y. (2006). *Sampling algorithms*. Springer-Verlag, New York.