Balanced sampling by means of the cube method

Yves Tillé
University of Neuchâtel

Euskal Estatistika Erakundea
XXIII Seminario Internacional de Estadística
November 2010

Table of Contents

1. Introduction and Motivations
2. The cube method
3. Applications of the Cube Method in Official Statistics
4. Variance and Variance Estimation
5. FAQ
6. Examples
7. Coordination of balanced samples

Introduction

Introduction and Motivations

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 $x_1, ..., x_p$, known for each unit of the population.
- $\mathbf{x}_k = (x_{k1}, ..., x_{kp})'$, 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

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

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 $\displaystyle\sum_{k \in S} \frac{\delta_{kh}}{\pi_k} = \sum_{k \in S} \delta_{kh} \frac{N_h}{n_h} = N_h$, 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!

Example 4

- Balance on the variable $x_k = 1, k \in U$. The balancing equations becomes
$$\sum_{k \in S} \frac{1}{\pi_k} = \sum_{k \in U} 1 = N.$$
or
$$\widehat{N}_\pi = N.$$
The population size is estimated without error.
- REMARK: there is always two free auxiliary variables
$$x_{k1} = \pi_k \text{ and } x_{k2} = 1, k \in U.$$
A sample should always be balanced on these variables.

The cube method

The cube method

General Remark

- All the problems of sampling can theoretically be solved by using a linear program.
- Define a cost for each sample $C(s)$. The cost is small if the sample is well balanced.
- Search the sampling design $p(s)$ that minimizes the expected cost

$$\sum_{s \subset U} C(s)p(s)$$

subject to

$$\sum_{s \subset U} p(s) = 1 \text{ and } \sum_{s \subset U, s \ni k} p(s) = \pi_k, k \in U.$$

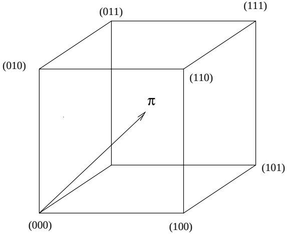
- Impossible in practice because of the combinatory explosion (2^N samples).
- The cube method is a shortcut that avoids the enumeration of the samples.

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 an affine subspace in $\mathbb{R}^N$ of dimension $N-p$ denoted Q.
- $Q=\boldsymbol{\pi}+Ker(A)$ where $\mathbf{A}=(\mathbf{a}_1,\dots,\mathbf{a}_k,\dots,\mathbf{a}_N)$. **The problem:** Choose a vertex of the N-cube (a sample) that remains on the sub-space Q.

System exactly verifiable

Example

$\pi_1 + \pi_2 + \pi_3 = 2.$
$x_k = \pi_k, k \in U$ and $\sum_{k \in U} s_k = 2.$

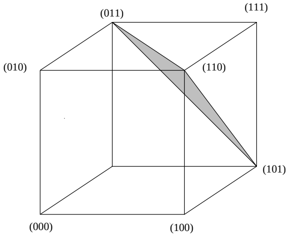


Figure: Fixed size constraint: all the vertices of K are vertices of the cube

System approximately verifiable

Example

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

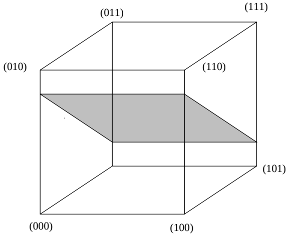


Figure: none of vertices of K are vertices of the cube

System sometimes verifiable

Example

$\pi_1 + 3 \times \pi_2 + \pi_3 = 4.$
$x_1 = \pi_1, x_2 = 3 \times \pi_2$ and $x_3 = \pi_3.$
$s_1 + 3s_2 + s_3 = 4.$

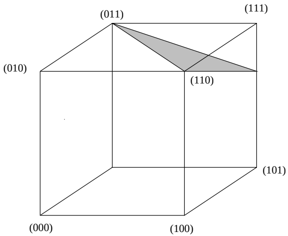


Figure: some vertices of K are vertices of the cube and others not

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 generalization of the splitting method. It 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.

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

Maximum $N - p$ steps.

Cube methods: examples

Example

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

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

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

Idea of the algorithm

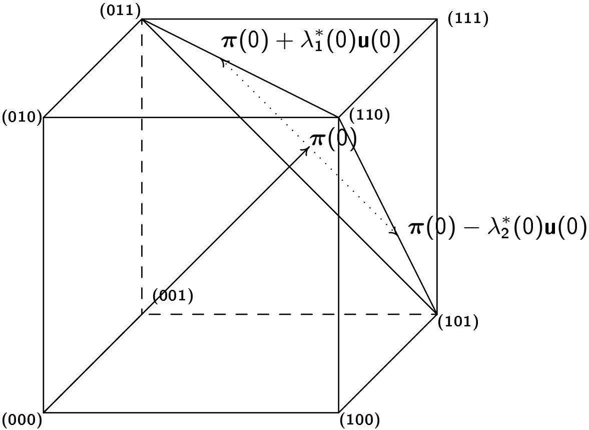

(011)
(111)
$\pi(0) + \lambda_1^*(0)\mathbf{u}(0)$
(010)
(110)
$\pi(0)$
$\pi(0) - \lambda_2^*(0)\mathbf{u}(0)$
(001)
(101)
(000)
(100)

Figure: Flight phase in a population of size $N = 3$ with a sample size constraint $n = 2$

The algorithm

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

1. Generate any vector $\mathbf{u}(t) = [u_k(t)] \neq 0$ such that
   (i) $\mathbf{u}(t)$ is in the kernel of matrix $\mathbf{A} = (\mathbf{x}_1/\pi_1, \dots, \mathbf{x}_k/\pi_k, \dots, \mathbf{x}_N/\pi_N))$
   (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
   $0 \leq \boldsymbol{\pi}(t) + \lambda_1(t)\mathbf{u}(t) \leq 1,$
   $0 \leq \boldsymbol{\pi}(t) - 1 - \lambda_2(t)\mathbf{u}(t) \leq 1.$
3. Compute

$$\boldsymbol{\pi}(t+1) = \begin{cases} \boldsymbol{\pi}(t) + \lambda_1^*(t)\mathbf{u}(t) & \text{with a proba } q_1(t) \\ \boldsymbol{\pi}(t) - \lambda_2^*(t)\mathbf{u}(t) & \text{with a proba } q_2(t), \end{cases}$$

where $q_1(t) = \lambda_2^*(t)/\{\lambda_1^*(t) + \lambda_2^*(t)\}$ and $q_2(t) = 1 - q_1(t)\}$.

Chauvet Tillé Implementation

- Chauvet and Tillé (2005b,a, 2006, 2007); Tillé and Matei (2007)
- Fast algorithm. Execution time $O(N \times p^2)$.
- Apply each step on the algorithm only on the first $p + 1$ units with non integer $\pi_k(t)$.
- If the only constraint is the fixed sample size: pivotal method.
- The order of the file change the samling design.
- Two solutions:
  - Random order of the sample. (Increasing of the randomness of the sample, of the entropy).
  - Decreasing order of size (reduce the rounding problem).

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

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

- The aim of the landing phase is to find a sample $\mathbf{S}$ such that $E(\mathbf{S}|\boldsymbol{\pi}^*) = \boldsymbol{\pi}^*$, and that is almost balanced.

Landing Phase 1

- Solution: linear program defined only on $q \leq p$ units.
- Search the sampling design on U^* that minimize

$$\sum_{s^* \subset U*} p(s^*) C(s^*)$$

subject to

$$\sum_{s^* \subset U^*, s^* \ni k} p(s^*) = \pi_k^* \text{ and } \sum_{s^* \subset U^**} p(s^*) = 1.$$

- $C(s^*)$ is the cost of sample s^* (for instance the distance between the sample and the subspace of constraints).

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.

Applications

Applications of the Cube Method in Official Statistics

France's New Census

- Selection of the rotation groups for the French census.
- **Small municipalities (<10,000 inhabitants)**
  Five non-overlapping rotation groups were selected using a balanced sampling design with equal inclusion probabilities (1/5). Each year, a fifth of the municipalities are surveyed.
- **Big municipalities (>10,000 inhabitants)**
  Five non-overlapping balanced samples of addresses are selected with inclusion probabilities 1/8. So, after 5 years, 40% of the addresses are visited.
- The balancing variables are socio-demographic variables of the last Census.

French Master Sample

- The primary units are geographical areas that are selected using a balanced sampling design.
- Self-weighted multi-stage sampling.
- So the primary units are selected with unequal probabilities proportional their sizes.
- The balancing variables are socio-demographic variables of the least census.

Other applications

- **Argentina's Master Sample**
  Simulations to evaluate the interest of balanced sampling for the master sample.
- **Statistics Canada's Business Survey**
  Use of a balanced sampling design to select a sample of businesses. (Canadian unincorporated businesses).
- **France's Sample of Subsidized Job Recipients**
  Use of a balanced sampling design for selecting a sample of beneficiaries of subsidized job (from a register).

Other applications

- **Balanced Sampling for Non-response**
  Deville proposed to use balanced sampling to make random controlled imputations.
- **Selection of Times Series for Archiving**
  At Électricité de France (EDF), new electricity meters allow electricity consumption for each household to be measured on a continuous basis. The amount of collected information is so large that it is impossible to archive all the data. Dessertaine proposed to select times series by using balanced sampling.
- **Italy's Consumer Index**
  Use of balanced sampling to improve the quality of consumer index in Italy.
- **Estimation in Small Domain**
  Use of a balanced sampling design to estimate totals in small domains.

Variance and Variance Estimation

Variance and Variance Estimation

Variance Approximation by a Residual Technique

- Deville and Tillé (2005) proposed an approximation:
- Idea: the balanced sampling design is a Poisson sampling design conditionally to the balanced constraints.
- Assuming that $(\widehat{Y}_\pi, \widehat{\mathbf{X}}'_\pi)'$ has a multivariate normal distribution, a simple reasonning allows us to compute:

$$\mathrm{Var}_p(\widehat{Y}_\pi) = \mathrm{Var}_{\tilde{p}}(\widehat{Y}_\pi | \widehat{\mathbf{X}}_\pi = \mathbf{X}),$$

where $\tilde{p}(.)$ is the Poisson design and $p(.)$ is the balanced design.

Variance Approximation by a Residual Technique

- Approximation of the variance

$$\mathrm{Var}_p(\widehat{Y}_\pi) \cong \mathrm{Var}_{app}(\widehat{Y}_\pi) = \sum_{k \in U} b_k \frac{(y_k - \mathbf{x}_k'\mathbf{b})^2}{\pi_k^2}, \tag{1}$$

where

$$\mathbf{b} = \left(\sum_{k \in U} b_k \frac{\mathbf{x}_k \mathbf{x}_k'}{\pi_k^2} \right)^{-1} \sum_{k \in U} b_k \frac{\mathbf{x}_k y_k}{\pi_k^2},$$

and the b_k are the solution of the nonlinear system

$$\pi_k(1-\pi_k) = b_k - \frac{b_k \mathbf{x}_k'}{\pi_k} \left(\sum_{\ell \in U} b_\ell \frac{\mathbf{x}_\ell \mathbf{x}_\ell'}{\pi_\ell^2} \right)^{-1} \frac{b_k \mathbf{x}_k}{\pi_k}, k \in U. \tag{2}$$

other solution $b_k = \frac{N}{N-1}\pi_k(1-\pi_k)$.

Estimation of Variance

- Deville and Tillé (2005) proposed a family of variance estimators

$$\widehat{\text{Var}}(\widehat{Y}_\pi) = \sum_{k\in S} c_k \frac{\left(y_k - \mathbf{x}_k'\widehat{b}\right)^2}{\pi_k^2}, \qquad (3)$$

where

$$\widehat{b} = \left(\sum_{\ell\in S} c_\ell \frac{\mathbf{x}_\ell\mathbf{x}_\ell'}{\pi_\ell^2}\right)^{-1} \sum_{\ell\in S} c_\ell \frac{\mathbf{x}_\ell y_\ell}{\pi_\ell^2}$$

and the c_k are the solutions of the nonlinear system

$$1 - \pi_k = c_k - \frac{c_k\mathbf{x}_k'}{\pi_k}\left(\sum_{\ell\in S} c_\ell \frac{\mathbf{x}_\ell\mathbf{x}_\ell'}{\pi_\ell^2}\right)^{-1} \frac{c_k\mathbf{x}_k}{\pi_k}, \qquad (4)$$

or

$$c_k \doteq \frac{n}{n-q}(1-\pi_k).$$

FAQ

Frequently Asked Questions

FAQ 1

- **What Are the Particular Cases of Balanced Sampling?**
  1. **Unequal probability sampling** (Brewer method, pivotal method, corrected Sunter method).
  2. **Stratification** The balancing variables are the indicators of the strata.
  3. **Overlapping strata**
  4. **Systematic sampling** can be seen as a balanced sampling design on the order statistic related to the variable on which the population is ordered.
- **Is Balanced Sampling Better Than Other Sampling Techniques?** This is not a good question, because almost all the other sampling techniques are particular case of balanced sampling (except multistage sampling). In fact, balanced sampling is simply more general.

FAQ 2

- **What Are the Main Drawbacks of Balanced Sampling?**
  Difficulty to co-ordinate the samples, because

$$\sum_{k\in S} \frac{\mathbf{x}_k}{\pi_k} = \sum_{k\in U} \mathbf{x}_k$$

does not imply that

$$\sum_{k\in U\backslash S} \frac{\mathbf{x}_k}{1-\pi_k} = \sum_{k\in U} \mathbf{x}_k.$$

(Tillé and Favre, 2005)

- **Is Balancing not Contradictory with Random Sampling?**
  No, the cube method randomly selects a sample that exactly satisfies the inclusion probabilities. This could looks contradictory, but a stratified sample is also balanced and is random.

FAQ 3

- **How Accurate is the Approximation With the Cube Method? Is the Rounding Problem Important?** It is possible to prove, under realistic assumptions (see Deville and Tillé, 2004), that

$$\left|\frac{\widehat{X}_j - X_j}{X_j}\right| < O(p/n),$$

where p is the number of variables, while in simple random sampling

$$\left|\frac{\widehat{X}_j - X_j}{X_j}\right| = O_p(\sqrt{1/n}).$$

- **Why Not Use Calibration Instead of Balancing?**
Stratification: particular case of balancing.
Poststratification: particular case of calibration.
With balancing the weight are random.

FAQ 4

- **Can Balancing and Calibration be Used Together?**
  Yes, but one should recalibrate on the balancing variable.
- **Is it Possible to Balance on a Non-linear Statistic?**
  Yes (see Lesage, 2007). A simple trick consists in balancing on the linearized variable.
- **What Are the Main Implementations of Balanced Sampling?**
  1. A SAS-IML done by three students of the ENSAI (Bousabaa et al., 1999).
  2. SAS-IML Official INSEE version done by Tardieu (2001) Rousseau and Tardieu (2004).
  3. University of Neuchâtel version Chauvet and Tillé (2005b, 2006); **?**
  4. In the R sampling package (Tillé and Matei, 2007)
- **What Are the Limits of the Population Size and the Number of Variables**
  40 balanced variables, no limit for N.

FAQ 4

- **How to select the balancing variables?**
  They must be correlated to the interest variables.
- **Question for Ray Chambers: What the matter if you are model-based?**
  Balanced sampling is recommended under a linear model (robustness argument).
- **How to choose the inclusion probabilities?**
  In some cases the π_k are mandatory (self-weighted two-stage sampling design, equal inclusion probabilities). In some other cases, you can choose the inclusion probabilities. Theoretically, they should be proportional to the standard deviations of the error terms of the linear model. (Nedyalkova and Tillé, 2009)

Examples

Examples

The Annual County Population

- The "CO-EST2006-alldata": is the Annual County Population Estimates and Estimated Components of Change: April 1, 2000 to July 1, 2006.
- County level.
- Source : web site of the Census Bureau in csv format.

Table: Metropolitan and Micropolitan Statistical Area Population Estimates File for Internet Display: source Census Bureau

Variable	Description
CTYNAME	County name
STNAME	State name
CENSUS2000POP	4/1/2000 resident Census 2000 population
POPESTIMATE2005	7/1/2005 resident total population estimate
POPESTIMATE2006	7/1/2006 resident total population estimate
BIRTHS2000	Births 4/1/2000 to 7/1/2000
BIRTHS2006	Births 7/1/2005 to 7/1/2006
DEATHS2000	Deaths 4/1/2000 to 7/1/2000
DEATHS2006	Deaths 7/1/2005 to 7/1/2006
INTERNATIONALMIG2000	Net international migration 4/1/2000 to 7/1/2000
INTERNATIONALMIG2006	Net international migration 7/1/2005 to 7/1/2006

The sampling design

- A sample of size $n = 400$ from the population of $N = 3141$ counties.
- Inclusion probabilities proportional to the variable popestimate2006.
- All the auxiliary variables are used as balancing variable.
- The R sampling package (see Tillé and Matei, 2007) allows the sample to be selected directly. The function samplecube(X, pik)-

```
# loading the sampling package library(sampling)
# reading of the file
V=read.csv("C://CO-EST2006-ALLDATA.csv",  header = TRUE, sep=",", dec=".", fill = T
attach(V)
# definition of the matrix of balancing variables
X=cbind(
  popestimate2006,
  births2000,births2006,
  deaths2000,deaths2006,
  internationalmig2000,internationalmig2006,
  one=rep(1,length(popestimate2006))
  )
# selection of the counties X=X[sumlev==50,]
# definition of the vector of inclusion probabilities
pik=inclusionprobabilities(X[,1],400)
# selection of the sample
s=samplecube(X,pik)
```

```
BEGINNING OF THE FLIGHT PHASE
The matrix of balanced variable has 8  variables and  3141  units
The size of the inclusion probability vector is  3141
The sum of the inclusion probability vector is  400
The inclusion probability vector has  3038  non-integer elements
Step 1

BEGINNING OF THE LANDING PHASE
At the end of the flight phase, there remain  8 non integer probabilities
The sum of these probabilities is  4
This sum is  integer
The linear program will consider  70  possible samples
The mean cost is  0.005863816
The smallest cost is  0.001072027
The largest cost is  0.00987782
The cost of the selected sample is 0.002267229

QUALITY OF BALANCING    TOTALS HorvitzThompson_estimators Relative_deviation
popestimate2006      299398484               2.993985e+08      -2.189894e-13
births2000              989020               9.897919e+05       7.804513e-02
births2006             4151889               4.154413e+06       6.079854e-02
deaths2000              560891               5.599909e+05      -1.604724e-01
deaths2006             2464633               2.462009e+06      -1.064653e-01
internationalmig2000    364221               3.658406e+05       4.446634e-01
internationalmig2006   1204167               1.209396e+06       4.342091e-01
one                       3141               3.183646e+03       1.357711e+00
```

The Selected Sample : Virginia, Washington, Wisconsin, Wyoming

	Name	State	Incl. prob.	Pop2006
371	Fairfax County	Virginia	1.000000000	1010443
372	Giles County	Virginia	0.030210273	17403
373	Greensville County	Virginia	0.019105572	11006
374	Hanover County	Virginia	0.171826896	98983
375	Loudoun County	Virginia	0.466645695	268817
376	Rockingham County	Virginia	0.125965539	72564
377	Stafford County	Virginia	0.208605903	120170
378	Alexandria city	Virginia	0.237776359	136974
379	Danville city	Virginia	0.079133800	45586
380	Hampton city	Virginia	0.251738390	145017
381	Norfolk city	Virginia	0.397720860	229112
382	Richmond city	Virginia	0.334882172	192913
383	Suffolk city	Virginia	0.140733038	81071
384	Virginia Beach city	Virginia	0.756201174	435619
385	King County	Washington	1.000000000	1826732
386	Pierce County	Washington	1.000000000	766878
387	Snohomish County	Washington	1.000000000	669887
388	Spokane County	Washington	0.775447356	446706
389	Yakima County	Washington	0.404652402	233105
390	Marshall County	West Virginia	0.058840856	33896
391	Dane County	Wisconsin	0.805166363	463826
392	Kenosha County	Wisconsin	0.281221311	162001
393	Langlade County	Wisconsin	0.035813834	20631
394	Lincoln County	Wisconsin	0.052339824	30151
395	Milwaukee County	Wisconsin	1.000000000	915097
396	Outagamie County	Wisconsin	0.299852976	172734
397	Shawano County	Wisconsin	0.071868961	41401
398	Wood County	Wisconsin	0.129801929	74774
399	Laramie County	Wyoming	0.148220075	85384
400	Sweetwater County	Wyoming	0.067289595	38763

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

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

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

Coordination

Co-ordination of balanced samples

The main problem of co-ordination

- If a sample is balanced on $\mathbf{x}_k$, the complementary sample $\overline{S} = U \backslash S$ is generally not balanced

$$\sum_{k \in S} \frac{\mathbf{x}_k}{\pi_k} = \sum_{k \in U} \mathbf{x}_k \not\Rightarrow \sum_{k \in \overline{S}} \frac{\mathbf{x}_k}{1 - \pi_k} = \sum_{k \in U} \mathbf{x}_k.$$

- If the inclusion probabilities are equal, it works

$$\sum_{k \in S} \frac{\mathbf{x}_k}{\pi} = \sum_{k \in U} \mathbf{x}_k \Rightarrow \sum_{k \in \overline{S}} \frac{\mathbf{x}_k}{1 - \pi} = \sum_{k \in U} \mathbf{x}_k.$$

  Application in the French new census (five balanced groups of rotation of municipalities)

- Question: How to balance several samples with unequal probabilities.

Solution 1: all the samples are selected together

- Tillé and Favre (2004)
- T samples must be selected with inclusion probabilities $\pi_k^1, \pi_k^2, \dots, \pi_k^t, \cdots, \pi_k^T$.
- Let $\pi_k = \pi_k^1 + \pi_k^2 + \cdots + \pi_k^t + \cdots + \pi_k^T$.
- Suppose that $\pi_k^t / \pi_k = C_t$ for all t. (The inclusion probabilities are proportional).
- Select a balanced 'big' sample S with inclusion probabilities $\pi_k = \pi_k^1 + \pi_k^2 + \cdots + \pi_k^t + \cdots + \pi_k^T$.
$$\sum_{k \in S} \frac{\mathbf{x}_k}{\pi_k} = \sum_{k \in U} \mathbf{x}_k.$$
- Select the small samples S_t from S with inclusion probabilities π_k^t / π_k and balance on $\mathbf{z}_k = \frac{\mathbf{x}_k}{\pi_k}$. Indeed,
$$\sum_{k \in S_t} \frac{\mathbf{x}_k}{\pi_k^t} = \sum_{k \in S} \frac{\mathbf{x}_k}{\pi_k} = \sum_{k \in U} \mathbf{x}_k,$$
implies that
$$\sum_{k \in S_t} \frac{\mathbf{z}_k}{\pi_k^t / \pi_k} = \sum_{k \in S} \mathbf{z}_k.$$
- If the inclusion probabilities are not proportional, the complementary of S_t in S must also be balanced (see next slide).

Solution 2: balance the complementary

- Select a balanced sample in such a way that the complementary is also balanced.
$$\sum_{k\in S} \frac{\mathbf{x}_k}{\pi_k} = \sum_{k\in \overline{S}} \frac{\mathbf{x}_k}{1-\pi_k} = \sum_{k\in U} \mathbf{x}_k .$$
- $\displaystyle\sum_{k\in \overline{S}} \frac{\mathbf{x}_k}{1-\pi_k} = \sum_{k\in U} \mathbf{x}_k \Leftrightarrow \sum_{k\in S} \frac{\pi_k \mathbf{x}_k}{1-\pi_k}\frac{1}{\pi_k} = \sum_{k\in U} \left(\frac{\pi_k \mathbf{x}_k}{1-\pi_k}\right)$
- Thus one must balance on $\mathbf{x}_k$ and on $\frac{\pi_k \mathbf{x}_k}{1-\pi_k}$.
- Since the complementary is balanced, then one can select a balanced subsample in $\overline{S}$.
- Disadvantage: the number of balanced variables has doubled.

Solution 3: rebalance an unbalanced sample

- Suppose that a sample S_1 is selected with inclusion probabilities π_k^1.
- Even if S_1 is balanced, $\overline{S}_1$ is not balanced.
- We want to supplement S_2 by a non-overlapping sample S_2 selected with inclusion probabilities π_k^2 in such a way that

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

where $S = S_1 \cap S_2$ and $\pi_k = \pi_k^1 + \pi_k^2$

- Other formulation

$$\sum_{k \in S_1} \frac{\mathbf{x}_k}{\pi_k} = T(S_1),$$

where

$$T(S_1) = \sum_{k \in U} \mathbf{x}_k - \sum_{k \in S_2} \frac{\mathbf{x}_k}{\pi_k}$$

Solution 3: rebalance an unbalanced sample

- Solution: Select a balanced sample S_2 from $U\backslash S_1$ with conditional inclusion probabilities $\widetilde{\pi}_{kb|S_1}$ and balancing variables

$$z_k = \frac{\mathbf{x}_k \widetilde{\pi}_{kb|S_1}}{\pi_k}.$$

The probabilities $\widetilde{\pi}_{kb|S_1}$ are defined as

$$\widetilde{\pi}_{kb|S_1} = \begin{cases} \pi_{kb} + \{T(S_1) - V(S_1)\}' \left(\sum_{\ell \in U\backslash S_1} \frac{\mathbf{x}_\ell \mathbf{x}_\ell' w_\ell}{\pi_\ell^2} \right)^{-1} \frac{\mathbf{x}_k w_k}{\pi_k} & k \notin S_1 \\ 0 & k \in S_1, \end{cases}$$

where

$$\pi_{kb} = \begin{cases} \frac{\pi_{k2}}{1 - \pi_{k1}} & \text{if } k \notin S_1 \\ 0 & \text{if } k \in S_1, \end{cases} \quad \text{and } V(S_1) = \sum_{k \in U\backslash S_1} \frac{\mathbf{x}_k \pi_{kb}}{\pi_k},$$

and the w_k's are weights (choose $w_k = \pi_{kb}(1 - \pi_{kb})$).

Solution 3: rebalance an unbalanced sample

It is possible to proof that

- $$\mathrm{E}\left(\sum_{k\in S_2}\frac{\mathbf{x}_k}{\pi_k}\middle| S_1\right)=T(S_1).$$
- $$\sum_{k\in S_2\cup S_1}\frac{\mathbf{x}_k}{\pi_k}=\sum_{k\in U}\mathbf{x}_k.$$
- $$\mathrm{E}\left(\widetilde{\pi}_{kb|S_1}\right)=\pi_{k2}+\mathrm{E}\left\{O_p\left(\frac{\mathbf{x}_k}{n}\right)\right\}.$$

Conclusion

Conclusion

- It is always better to anticipate the problem of co-ordination.
- The best solution is to select all the samples together.
- If it it not possible, try to balance the complementary sample.

References

Bousabaa, A., Lieber, J., and Sirolli, R. (1999). La macro cube. Technical report, ENSAI, Rennes.

Chauvet, G. and Tillé, Y. (2005a). *Fast SAS Macros for balancing Samples: user's guide*. Software Manual, University of Neuchâtel.

Chauvet, G. and Tillé, Y. (2005b). New SAS macros for balanced sampling. In *Journées de Méthodologie Statistique, INSEE*, Paris.

Chauvet, G. and Tillé, Y. (2006). A fast algorithm of balanced sampling. *Journal of Computational Statistics*, 21:9–31.

Chauvet, G. and Tillé, Y. (2007). Application of the fast sas macros for balancing samples to the selection of addresses. *Case Studies in Business, Industry and Government Statistics*, 1:173–182.

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*, 128:569–591.

Nedyalkova, D. and Tillé, Y. (2009). Optimal sampling and estimation strategies under linear model. *Biometrika*, 95:521–537.

Tillé, Y. (2006). *Sampling Agorithms*. Springer, New York.

Tillé, Y. and Favre, A.-C. (2004). Co-ordination, combination and extension of optimal balanced samples. *Biometrika*, 91:913–927.

Tillé, Y. and Favre, A.-C. (2005). Optimal allocation in balanced sampling. *Statistics and Probability Letters*, 74:31–37.

Tillé, Y. and Matei, A. (2007). *The R Package Sampling*. The Comprehensive R Archive Network, http://cran. r-project. org/, Manual of the Contributed Packages.