February-March 2015 – ELSTAT

Main sampling techniques

Martin Chevalier (INSEE)

General framework of sampling

Goal For some variable Y, to estimate parameters (total, mean, median...) defined on a population U.

Problem Y is NOT known on U.

Solution Randomly select a sample s of U where Y is measured and use this information to estimate the parameters on U.

Sampling design Probability distribution of all subsets of U. For each observation i of U, a sampling design defines a first-order inclusion probability

$$\pi_i = P(i \in s) = \sum_{s|i \in s} p(s)$$

and a set of second-order inclusion probabilities

$$\pi_{ij} = P(i \in s \quad \text{AND} \quad j \in s)$$

General framework of sampling

Horvitz-Thompson estimator

The total of variable Y in the population U is estimated by

$$\hat{T}(Y) = \sum_{i \in s} \frac{y_i}{\pi_i}$$

- Unbiased estimator under the sampling design
- The variance can be computed as a function of the π_{ij} :

$$V(\hat{T}(Y)) = \sum_{i \in U} \sum_{j \in U} (\pi_{ij} - \pi_i \pi_j) \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} \quad \text{with} \quad \pi_{ii} = \pi_i$$

General framework of sampling

Simple random sampling (SRS)

A simple random sampling is a sampling technique yielding samples with fixed size n that assures that every sample of size n has the same probability of being drawn.

$$\pi_i = \frac{n}{N} \quad \text{and} \quad \pi_{ij} = \frac{n(n-1)}{N(N-1)}$$

$$\hat{T}_{SRS}(Y) = \sum_{i \in s} \frac{y_i}{n/N} = N\bar{y} \quad \text{and} \quad \hat{V}(\hat{T}_{SRS}(Y)) = N^2(1-f)\frac{s_Y^2}{n}$$

with $f = \frac{n}{N}$ and $s_Y^2 = \frac{1}{n-1} \sum_{i \in s} (y_i - \bar{y})^2$.

Where do we go from there?

SRS only requires a list of the population of interest (which is already a lot!).

But when this list exists, there is often also information about the units to sample (at least derived from contact information): **auxiliary variables**.

The common aim of the various sampling techniques presented is this session is to **take advantage of this auxiliary information** in order to improve the sampling design.

Where do we go from there?

Over-represent some sub-population of interest: **two-phases sampling**.

Adjust the total of some variables in the sample to their counterpart in the sampling frame: **balanced sampling**.

Ensure a minimum number of sampled units with certain characteristics and increase precision: **stratified sampling and systematic sampling on a sorted file**.

Reduce the unit cost per interview in order to increase sample size or decrease the global cost of the survey: **cluster and two-stages samplings**.

Conduct a survey without an exhaustive list of the population of interest or combine several sampling frames: **indirect sampling**.

February-March 2015 – ELSTAT

Main sampling techniques
Stratified sampling

Martin Chevalier (INSEE)

Stratified sampling

Principles and notations

Let Y be a quantitative variable defined on U.

When using a simple random sampling: when dispersion S^2 of Y increases, the precision of the estimation decreases.

Hence the **core principle** of stratification:

- Let's partition the population U into H parts called "strata" and denoted $U_1, U_2, \ldots, U_h, \ldots, U_H$ so that, in each stratum h, the dispersion S_h^2 of Y is low.
- In each stratum h, draw independently a sample according to a sampling design p_h.

Principles and notations

Justification Because of the low dispersion in each stratum, estimators might be more accurate, which should lead to more precision in the whole sample.

Other goal Stratification allows to set a lower bound for precision in each stratum by controlling the number of units per stratum in the sample.

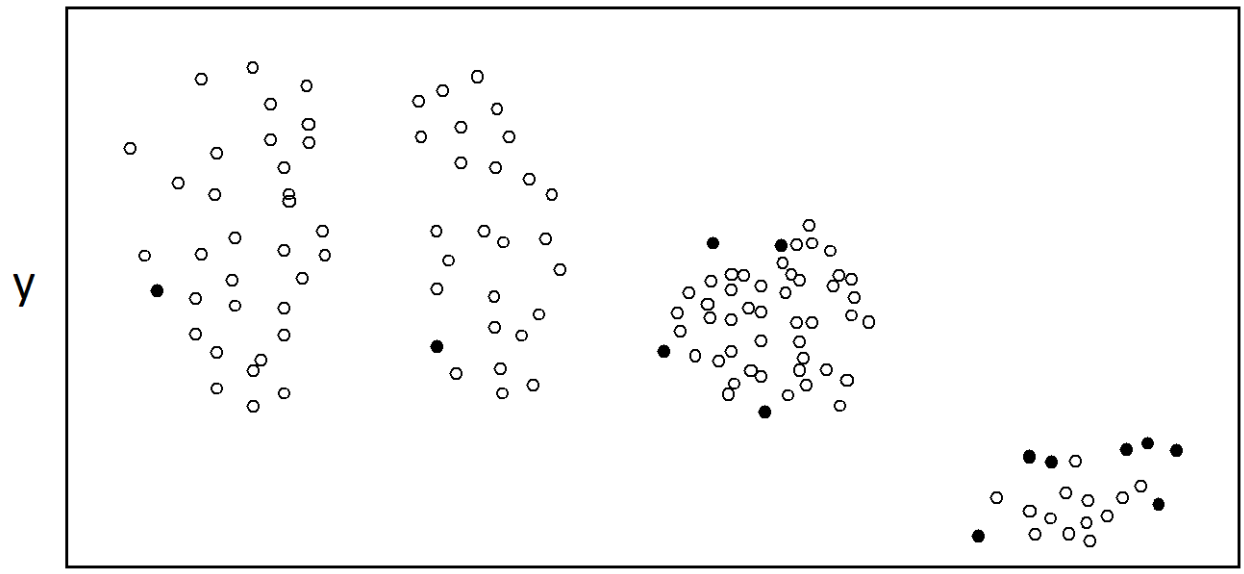
Remark Contrary to the simple random sampling, this method requires **auxiliary information** in the sampling frame, i.e. one or more variables in order to determine the strata.

It is assumed that the sizes of the strata N_h are known (usually from the sampling frame).

Principles and notations

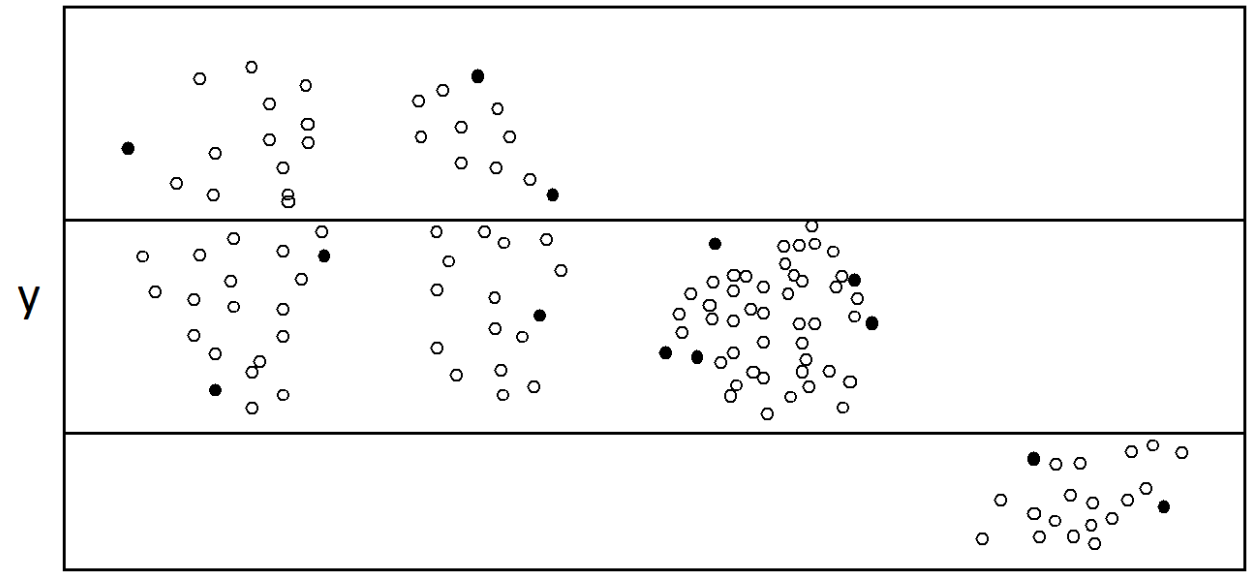
Representation of a SRS

SRS of $n = 13$ units in a population of size $N = 130$ units.



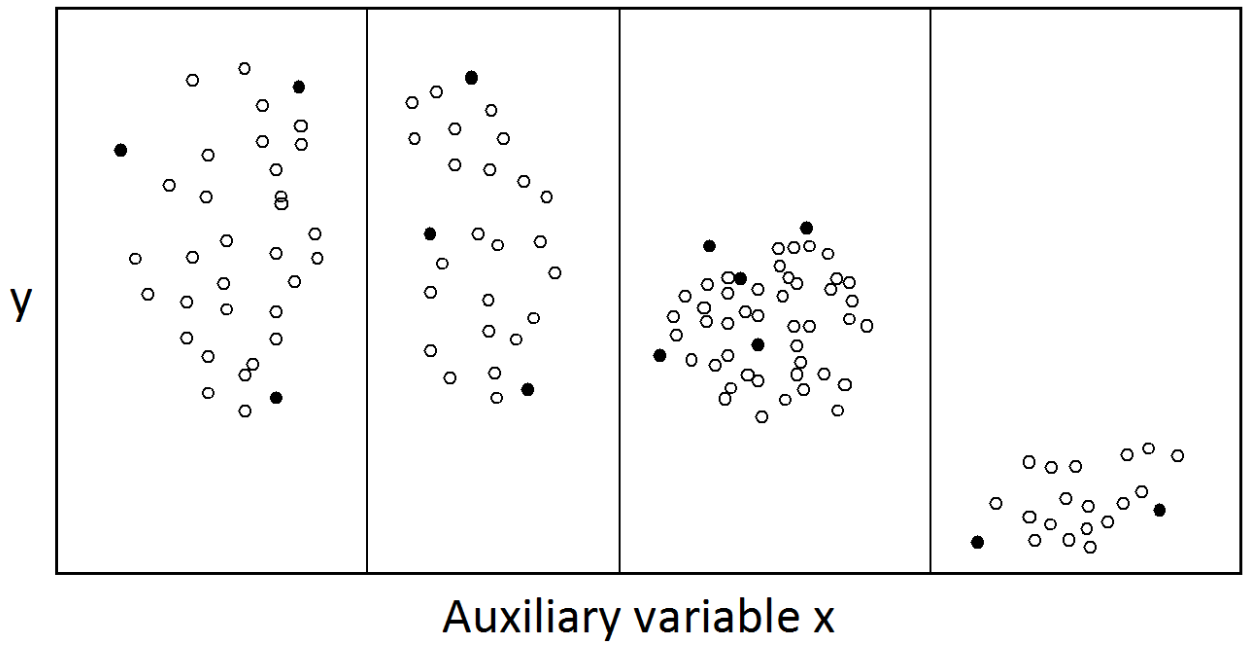
Principles and notations

Representation of a stratified sampling: strata constitution



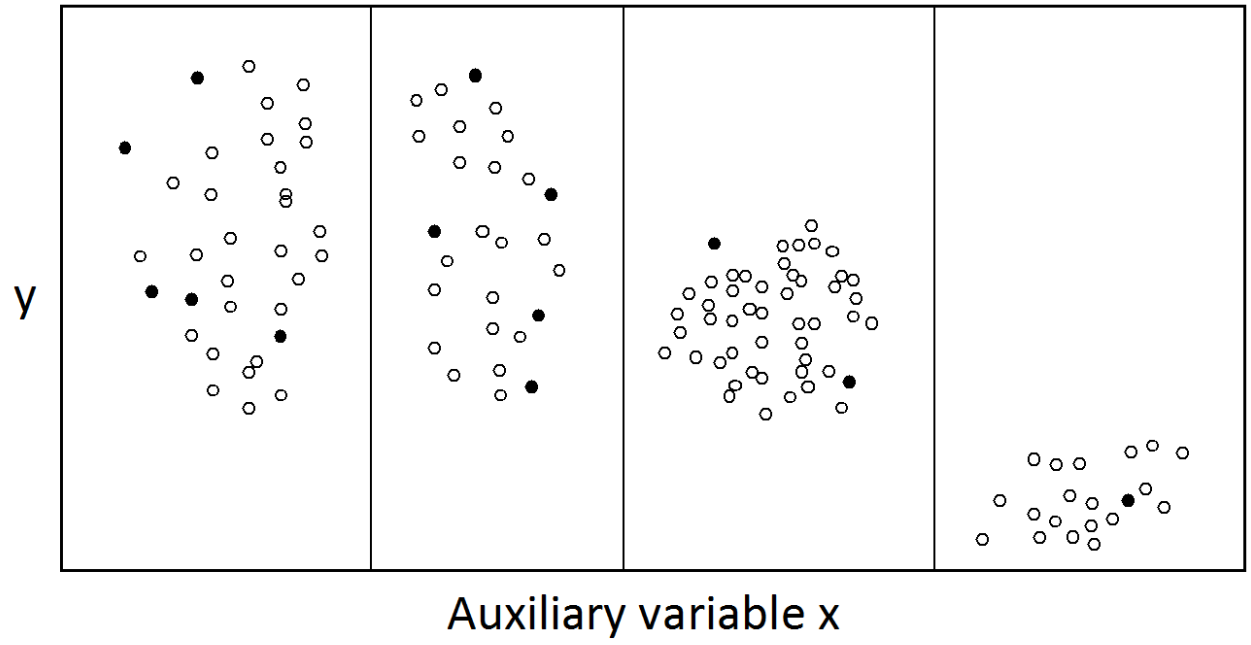
Principles and notations

Representation of a stratified sampling: strata constitution



Principles and notations

Representation of a stratified sampling: sample allocation



Principles and notations

Steps in order to achieve a stratified sampling design of size n

1. Partition the population U into H strata. Every unit of the sampling frame must be associated with one and only one stratum.
2. Determine the allocation in each stratum under the following constraint:

$$\sum_{h=1}^{H} n_h = n$$

   n is assumed to be known (depends on the goals and budget allocated to the survey).
3. In each stratum U_h, draw a sample s_h of size n_h using a sampling design p_h.

The final sample s is the union of all samples s_h:

$$s = s_1 \cup s_2 \cup \ldots \cup s_H$$

Principles and notations

	Population U	Sample s
Size of stratum h	N_h	n_h
Number of observations	$N = \sum_h N_h$	$n = \sum_h n_h$
Total of Y in stratum h	$T_h(Y) = \sum_{i \in U_h} Y_i$	$t_h(Y) = \sum_{i \in s_h} Y_i$
Total if Y in U	$T(Y) = \sum_h T_h(Y)$	$t(Y) = \sum_h t_h(Y)$
Mean of Y in stratum h	$\bar{Y}_h = \frac{T_h(Y)}{N_h}$	$\bar{y}_h = \frac{t_h(Y)}{n_h}$
Mean of Y in U	$\bar{Y} = \sum_h \frac{N_h}{N} \bar{Y}_h$	$\bar{y} = \sum_h \frac{n_h}{n} \bar{y}_h$

Stratified sampling

Estimation and precision

Estimator The total of Y is estimated without bias by:

$$\hat{T}_{str}(Y) = \sum_{h=1}^{H} \hat{T}_h(Y)$$

where $\hat{T}_h(Y)$ is the Horvitz-Thompson estimator of $T_h(Y)$:

$$\hat{T}_h(Y) = \sum_{i \in s_h} \frac{y_i}{\pi_i}$$

Estimation and precision

Precision The $\hat{T}_h(Y)$ are independent from one another, hence:

$$V(\hat{T}_{str}(Y)) = \sum_{h=1}^{H} V(\hat{T}_h(Y)) \quad \text{and} \quad \hat{V}(\hat{T}_{str}(Y)) = \sum_{h=1}^{H} \hat{V}(\hat{T}_h(Y))$$

with $V(\hat{T}_h(Y)) = \sum_{i \in U_h} \sum_{j \in U_h} (\pi_{ij} - \pi_i \pi_j) \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}$

and $\hat{V}(\hat{T}_{str}(Y))$ its unbiased estimator (Horvitz-Thompson or Yates-Grundy).

$V(\hat{T}_h(Y))$ can also be computed using the classical Horvitz-Thompson variance estimator, once one noticed that

$$\pi_{ij} - \pi_i \pi_j = 0 \quad \text{if} \quad i \in U_h \quad \text{and} \quad j \in U_{h'}, \quad h \neq h'$$

Estimation and precision

Stratified sampling with a SRS in each stratum

Now suppose that in each stratum, the sample is drawn by simple random sampling without replacement with a sampling rate

$$f_h = \frac{n_h}{N_h}$$

Estimators The total $T(Y)$ and the mean $\bar{Y}$ are estimated without bias by

$$\hat{T}_{str}(Y) = \sum_{h=1}^{H} N_h \bar{y}_h \quad \text{and} \quad \hat{\bar{Y}}_{str} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h$$

Estimation and precision

Stratified sampling with a SRS in each stratum

Remarks

1. $\hat{\bar{Y}}_{str} \neq \bar{y}$ The stratified estimator differs from the arithmetic mean.

2. $\hat{T}_{str}(Y) = \sum_{h=1}^{H} N_h \bar{y}_h = \sum_{h=1}^{H} N_h \left(\frac{1}{n_h} \sum_{i \in s_h} y_i \right) = \sum_{h=1}^{H} \sum_{i \in s_h} \frac{N_h}{n_h} y_i$
   For each observation of stratum h, the sampling weight is $\frac{N_h}{n_h}$. Stratification can also yield unequal probability sampling (but not necessarily).

Estimation and precision

Stratified sampling with a SRS in each stratum

Precision The variance of the stratified estimator of a total is

$$V(\hat{T}_{str}(Y)) = \sum_{h=1}^{H} N_h^2 V(\bar{y}_h) = \sum_{h=1}^{H} N_h^2 (1 - f_h) \frac{S_h^2}{n_h}$$

Remark The precision of the stratified estimator only depends on the dispersion of Y within the strata: the more the variance within the strata is low, the more the stratification is efficient.

The estimated variance of the stratified estimator is

$$\hat{V}(\hat{T}_{str}(Y)) = \sum_{h=1}^{H} N_h^2 (1 - f_h) \frac{s_h^2}{n_h}$$

Remark In order to be computed, this estimator requires at least 2 observations per stratum.

Estimation and precision

Stratified sampling with a SRS in each stratum

The variance of the stratified estimator of a mean is

$$V(\hat{\bar{Y}}_{str}) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 (1 - f_h)\frac{S_h^2}{n_h}$$

This variance is estimated without bias by

$$\hat{V}(\hat{\bar{Y}}_{str}) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 (1 - f_h)\frac{s_h^2}{n_h}$$

Estimation and precision

Example: 2 units per stratum

Population U	A	B	C	D	E	F
Values	2	6	8	10	10	12
Stratification A	I	I	II	I	II	II

Sample	1	2	3	4	5	6	7	8	9
Stratum I	2 6	2 6	2 6	2 10	2 10	2 10	6 10	6 10	6 10
Mean	4	4	4	6	6	6	8	8	8
Stratum II	8 10	8 12	10 12	8 10	8 12	10 12	8 10	8 12	10 12
Mean	9	10	11	9	10	11	9	10	11
Estimator	6.5	7	7.5	7.5	8	8.5	8.5	9	9.5

Sampling variance 0.83 (1.07 for a SRS)

Estimation and precision

Example: 2 units per stratum

Variance estimator

Sample	1	2	3	4	5	6	7	8	9
Stratum I	2 6	2 6	2 6	2 10	2 10	2 10	6 10	6 10	6 10
Variance	8	8	8	32	32	32	8	8	8
Stratum II	8 10	8 12	10 12	8 10	8 12	10 12	8 10	8 12	10 12
Variance	2	8	2	2	8	2	2	8	2
Estimator	0.4	0.7	0.4	1.4	1.7	1.4	0.4	0.7	0.4

Mean of variance estimator 0.83 (unbiased)

Variance of variance estimator 0.236 (0.251 for a SRS)

Stratified sampling

Strata constitution

These results provide some guidance in the problem of **strata constitution** and **sample allocation between strata**.

As the variance of the estimation of Y is directly related to the variance of Y within the strata, a "good" stratification should aim to minimize this within-variance.

In order to obtain the most efficient stratification, **the values of Y must be as close as possible within each stratum**.

Strata constitution

Example: 2 units per stratum

Population U	A	B	C	D	E	F
Values	2	6	8	10	10	12
Stratification B	I	II	II	I	II	I

Sample	1	2	3	4	5	6	7	8	9
Stratum I	2 10	2 10	2 10	2 12	2 12	2 12	10 12	10 12	10 12
Mean	6	6	6	7	7	7	11	11	11
Stratum II	6 8	6 10	8 10	6 8	6 10	8 10	6 8	6 10	8 10
Mean	7	8	9	7	8	9	7	8	9
Estimator	6.5	7	7.5	7	7.5	8	9	9.5	10

Sampling variance 1.33 (1.07 for a SRS)

Strata constitution

Example: 2 units per stratum

Variance estimation

Sample	1	2	3	4	5	6	7	8	9
Stratum I	2 10	2 10	2 10	2 12	2 12	2 12	10 12	10 12	10 12
Variance	32	32	32	50	50	50	2	2	2
Stratum II	6 8	6 10	8 10	6 8	6 10	8 10	6 8	6 10	8 10
Variance	2	8	2	2	8	2	2	8	2
Estimator	1.4	1.7	1.4	2.2	2.4	2.2	0.2	0.4	0.2

Mean of variance estimator 1.33 (unbiased)

Variance of variance estimator 0.944 (0.251 for a SRS)

Strata constitution

Example: 2 units per stratum

Population U	A	B	C	D	E	F
Values	2	6	8	10	10	12
Stratification C	I	I	I	II	II	II

Sample	1	2	3	4	5	6	7	8	9
Stratum I	2 6	2 6	2 6	2 8	2 8	2 8	6 8	6 8	6 8
Mean	4	4	4	5	5	5	7	7	7
Stratum II	10 10	10 12	10 12	10 10	10 12	10 12	10 10	10 12	10 12
Mean	10	11	11	10	11	11	10	11	11
Estimator	7	7.5	7.5	7.5	8	8	8.5	9	9

Sampling variance 0.44 (1.07 for a SRS)

Strata constitution

How to approximate S_h^2?

As Y is the variable that shall be estimated with the survey, it does not appear in the sampling frame: the S_h^2 are therefore unknown.

The basic idea is so to use some **auxiliary information** from the sampling frame which **might be correlated with** Y.

Depending on the auxiliary variables available in the sampling frame, the stratification might rely on **one or more variables**, in order to :

- maximize homogeneity within each stratum
- maximize heterogeneity between the strata

Remark One stratification can be efficient for one variable Y, but not for another one.

Strata constitution

How many strata?

In theory, the higher the number of strata, the better. Indeed, if one split a stratum, the within-variance can only decrease...

In practice, there is a "critical threshold" :

- a more complex data collection and estimation may cancel out the gains in terms of precision when adding one more stratum.
- at least one surveyed unit per stratum is required in order to obtain unbiased estimators and two to estimate precision. In order to anticipate non-response, one should sample more than two units in each stratum.

Strata constitution

Usual criteria for stratification at INSEE

Household surveys

- Region (NUTS2)
- Habitat: urban, semi-urban, rural
- Diploma

Business surveys

- Industry sector (NACE sections)
- Firm size: number of employees or turnover
- Region (NUTS2)

Strata constitution

Example: Strata boundaries for the number of employees

The variable "number of employees" is in general available as a number is the sampling frame (not interval coded).

In order to use it in as a stratification variable, one must set some boundaries to define the strata.

The usual boundaries in French business surveys are the following: 10-19, 20-49, 50-99, 100-249, 250-499, 500-999, 1,000-4,999, 5,000 and above.

A study has been conducted by one of our colleagues about the optimality of theses boundaries in terms of sampling variance.

Strata constitution

Example: Strata boundaries for the number of employees

There are several methods which determine "optimal" boundaries $b_0, b_1, \dots b_H$ in some sense for variable y.

One of the most straightforward is the **geometric method**. It is based on the idea that with boundaries near the optimum, the coefficients of variation should be equal across strata.

$$\forall h \in \{1, \dots H\}, \quad \frac{s_h}{\bar{y}_h} = constant$$

As the coefficients of variation cannot always be computed, let assume that the y are distributed roughly following a **uniform probability distribution** in each stratum h. Then:

$$\bar{y}_h \approx \frac{b_h + b_{h+1}}{2} \quad \text{and} \quad s_h \approx \frac{b_h - b_{h-1}}{\sqrt{12}}$$

Strata constitution

Example: Strata boundaries for the number of employees

For any given $h < H$:

$$\frac{s_h}{\bar{y}_h} = \frac{s_{h+1}}{\bar{y}_{h+1}} \quad \Rightarrow \quad \frac{b_h - b_{h-1}}{b_h + b_{h-1}} = \frac{b_{h+1} - b_h}{b_{h+1} + b_h}$$

$$\Rightarrow \quad b_h^2 = b_{h+1} b_{h-1}$$

With $b_0 > 0$, it implies:

$$\forall h \in \{1, \dots H\}, \quad b_h = b_0 \left(\frac{b_H}{b_0} \right)^{\frac{h}{H}}$$

where b_0 and b_H are respectively the minimum and maximum values of y.

Strata constitution

Example: Strata boundaries for the number of employees

Application on French data

The boundaries yielded by this method on French data are: 10-24, 25-59, 60-143, 144-348, 349-846, 847-2,055, 2,056-4,999,5,000 and above (the last stratum is defined *ex ante*).

For a given precision of the estimation of the number of employees, one can **compare** the number of units needed by a SRS, a stratified sampling with usual boundaries and stratified sampling with boundaries determined by the geometric method.

CV	SRS	Usual boundaries	Geometric method
1 %	57,922	666	611
5 %	3,276	156	151
10 %	925	138	136

Strata constitution

Example: Strata boundaries for the number of employees

In this situation, the variable to be estimated is known on the whole population (available in the sampling frame): hence the **magnitude of the gains associated with stratification**.

In general, if the variable of interest is correlated with the stratification variable, **the position of the boundaries might influence the efficiency of the stratification**.

The **R** package `stratification` implements **several methods for optimizing strata boundaries** (including the geometric method) in this context.

See Baillargeon S., Rivest L.-P. (2011), "The construction of stratified designs in R with the package stratification", *Survey methodology*, Vol. 37, No. 1, pp. 53-65

Stratified sampling

Sample allocation between strata

Once the strata are defined and assuming that the size n of the sample is known, **is there a best way to allocate the sampled units between the strata?**

The answer to that question differs with the goal of the survey:

- To obtain the best precision for one variable
- To obtain the best precision for several variables simultaneously
- To obtain a good precision in each stratum in order to compare the estimators between strata

Sample allocation between strata

Optimal allocation

Let's assume that the cost of a survey can be written as:

$$C = \sum_{h=1}^{H} n_h c_h \quad (+c_0)$$

where c_h is the cost of one interview in the stratum h.

Problems

- Determine n_h which minimize $V(\hat{T}_{str}(Y))$ for a given cost C.
- Determine n_h which minimize the cost C for a given $V(\hat{T}_{str}(Y))$.

Sample allocation between strata

Optimal allocation

Optimal precision at a given cost

The n_h which minimize the variance $V(\hat{T}_{str}(Y))$ for a given cost C are

$$n_h = \frac{N_h S_h}{\sqrt{c_h}} \frac{C}{\sum_{k=1}^{H} \sqrt{c_k} N_k S_k}$$

and the minimal variance is

$$V_{opt}(\hat{T}_{str}(Y)) = \frac{1}{C} \left(\sum_{h=1}^{H} \sqrt{c_h} N_h S_h \right)^2 - \sum_{h=1}^{H} N_h S_h^2$$

Sample allocation between strata

Optimal allocation

Proof

$$\begin{cases} \min_{n_h} \sum_h N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right) S_h^2 \\ \text{with constraint} \quad C = \sum_h n_h c_h \end{cases}$$

Keeping only terms which include n_h, let's write the Lagrangian of this minimization problem:

$$L(n_1, n_2, \ldots, n_H, \lambda) = \sum_h \frac{N_h^2 S_h^2}{n_h} - \lambda \left(C - \sum_h n_h c_h \right)$$

The first-order conditions yield:

$$\begin{cases} \frac{\delta L}{\delta n_h} = 0 \Rightarrow \frac{N_h^2 S_h^2}{n_h^2} = \lambda c_h \Rightarrow n_h = \frac{N_h S_h}{\sqrt{\lambda c_h}} \\ \frac{\delta L}{\delta \lambda} = 0 \Rightarrow C = \sum_h n_h c_h = \sum_h \frac{N_h S_h \sqrt{c_h}}{\sqrt{\lambda}} \Rightarrow \frac{1}{\sqrt{\lambda}} = \frac{C}{\sum_h N_h S_h \sqrt{c_h}} \end{cases}$$

Hence $n_h = \frac{N_h S_h}{\sqrt{c_h}} \frac{C}{\sum_{k=1}^H \sqrt{c_k} N_k S_k}$

Sample allocation between strata

Optimal allocation

Optimal cost at a given precision

The n_h which minimize the cost C for a given precision $V(\hat{T}_{str}(Y))$ are

$$n_h = \frac{N_h S_h}{\sqrt{c_h}} \frac{\sum_{k=1}^H \sqrt{c_k} N_k S_k}{V(\hat{T}_{str}(Y)) + \sum_{k=1}^H N_k S_k^2}$$

and the minimal cost is

$$C_{opt} = \frac{\left(\sum_{h=1}^H \sqrt{c_h} N_h S_h \right)^2}{V(\hat{T}_{str}(Y)) + \sum_{h=1}^H N_h S_h^2}$$

Sample allocation between strata

Optimal allocation

Interpretation

In both cases

$$\frac{n_h}{N_h} \propto \frac{S_h}{\sqrt{c_h}}$$

- One should over-represent the strata where the dispersion of Y is the highest: in other terms, the survey should go get information where it is.
- One should over-represent the strata where the unit cost c_h is the lowest.

Sample allocation between strata

Optimal allocation

Neyman allocation If we assume that the cost of an interview c_h does not vary accross strata, the optimal allocation is also called Neyman allocation:

$$n_h = n \times \frac{N_h S_h}{\sum_{k=1}^{H} N_k S_k}$$

Dalenius rule When using Neyman allocation, it can be useful to define the strata so that $N_h S_h$ is constant across strata (Dalenius rule). It yields the same sample size in every stratum:

$$n_h = \frac{n}{H}$$

Sample allocation between strata

Example: 3 units in stratum I, 1 unit in stratum II

Population U	A	B	C	D	E	F
Values	2	6	8	10	10	12
Stratification A	I	I	II	I	II	II

Sample	1	2	3
Stratum I	2 6 10	2 6 10	2 6 10
Mean	6	6	6
Stratum II	8	10	12
Mean	8	10	12
Estimator	7	8	9

Sampling variance 0.67 (1.07 for a SRS)

Sample allocation between strata

Example: 1 unit in stratum I, 3 units in stratum II

Population U	A	B	C	D	E	F
Values	2	6	8	10	10	12
Stratification A	I	I	II	I	II	II

Sample	1	2	3
Stratum I	2	6	10
Mean	2	6	10
Stratum II	8 10 12	8 10 12	8 10 12
Mean	10	10	10
Estimator	6	8	10

Sampling variance $8/3 = 2.67$ (1.07 for a SRS)

Sample allocation between strata

Example: Neyman allocation

Population U	A	B	C	D	E	F
Values	2	6	8	10	10	12
Stratification A	I	I	II	I	II	II

In this example, the data are:

$$n = 4, \quad N_I = N_{II} = 3, \quad S_I = 4, \quad \text{and} \quad S_{II} = 2$$

Then it follows:

$$\begin{cases} n_I = 4 \times \dfrac{3 \times 4}{3 \times 4 + 3 \times 2} = \dfrac{48}{18} = 2.7 \\ n_{II} = 4 \times \dfrac{3 \times 2}{3 \times 4 + 3 \times 2} = \dfrac{24}{18} = 1.3 \end{cases}$$

So it explains the previous results.

Sample allocation between strata

Optimal allocation

Estimation of the S_h

The variance of Y within each stratum is unknown. In order to apply optimal allocation, it can be estimated using various methods:

- Expert opinions
- Auxiliary information from the sampling frame
- Previous surveys
- A lightweight preliminary survey (as long as the additional cost allows a far better quality of estimation in the main survey)

Sample allocation between strata

Proportional allocation

Definition The allocation of the sample between strata is identical to the allocation of the population between strata:

$$\forall h \in \{1, \dots H\} \quad \frac{n_h}{n} = \frac{N_h}{N}$$

It yields the **same sampling rate** in each stratum

$$f_h = \frac{n_h}{N_h} = \frac{n}{N} = f$$

This sampling is so-called "representative" or proportional. It is an equal probability sampling.

Sample allocation between strata

Proportional allocation

Estimator The estimator is identical to the one used in simple random sampling...

$$\hat{\bar{Y}}_{prop} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h = \sum_{h=1}^{H} \frac{n_h}{n} \bar{y}_h = \bar{y}$$

Variance ... but its variance differs!

$$V(\hat{\bar{Y}}_{prop}) = \frac{1-f}{n} \sum_{h=1}^{H} \frac{N_h}{N} S_h^2 \simeq (1-f) \frac{S_{within}^2}{n}$$

Sample allocation between strata

Proportional allocation

Comparison with simple random sampling

As $V(\hat{\bar{Y}}_{SRS}) = (1-f)\frac{S^2}{n}$ and $S^2_{within} \leq S^2$ (variance decomposition formula):

$$V(\hat{\bar{Y}}_{prop}) \leq V(\hat{\bar{Y}}_{SRS})$$

The stratified sampling with proportional allocation **always outperforms the simple random sampling in terms of precision**.

The larger the dispersion between the strata, the bigger the gain associated with the stratification.

Sample allocation between strata

Proportional allocation

Comparison with Neyman allocation

For a given variable of interest Y, Neyman allocation yields significant gains compared to proportional allocation if the dispersions S_h differ a lot from one stratum to another.

However, Neyman allocation is optimal with respect to variable Y, and may be harmful for the estimation of another variable.

If one uses an allocation "not too far" from Neyman allocation but closer to proportional allocation, the precision is nearly "optimal".

Sample allocation between strata

Other allocations

Same precision in each stratum

The variance of $\bar{Y}$ in each stratum is a function of S_h^2 and n_h (assuming a negligible sampling rate) :

$$V(\bar{Y}) \approx \frac{S_h^2}{n_h}$$

If one aims to achieve the same precision in each stratum, the allocation should be proportionate to the variance of Y within each stratum:

$$n_h = n \times \frac{S_h^2}{\sum_{k=1}^{H} S_k^2}$$

Sample allocation between strata

Other allocations

Efficient allocation for several variables

The optimal allocation for a variable Y may yield a worse precision regarding other variables than simple random sampling.

It is possible to weight the J different variables of interest through their variance:

$$V = \sum_{j=1}^{J} \alpha_j V(\hat{T}_{str}(Y^j))$$

in order to minimize V given a total cost C, and conversely:

$$n_h \propto \frac{N_h \sqrt{\sum_{j=1}^{J} \alpha_j S_{Y_h^j}^2}}{\sqrt{c_h}}$$

Problem How to choose the α_j...

Sample allocation between strata

Exhaustive strata

Using other allocations than the proportional (in particular Neyman allocation), the calculated allocation for a stratum may be **larger than its actual size in the population**.

All units belonging to this stratum should then be sampled: this is a so-called **exhaustive stratum**.

This configuration may yield a **sample size n smaller than the expected one**: too few units are sampled from the exhaustive strata.

Sample allocation between strata

Exhaustive strata

In order to achieve the desired sample size n, these strata should be treated in an **iterative process**.

1. Calculate allocations using all strata.
2. Until all calculated allocations are smaller than the actual size of the strata in the population:
   - 2.1 Saturate the exhaustive strata.
   - 2.2 Calculate a new allocation for all remaining strata after removing the units from the exhaustive strata.
3. Sample the non-exhaustive strata using their calculated allocation.

Sample allocation between strata

Example: Sampling of a business survey

Goals Sample $n = 300$ firms out of a population U of size $N = 1060$ (e.g. a specific sector).

Auxiliary variable The size of the firm in terms of employees is known on U and coded in intervals. For each firm size, the mean ($\bar{y}$) and the variance (s_h^2) of the turnover are known.

Size of the firm	N_h	$\bar{y}_h$	S_h^2	Prop.	Opti.
0-9	500	10	2		
10-19	300	50	15		
20-49	150	200	50		
50-499	100	500	100		
500 and more	10	1,000	2,500		

To do Determine the proportional and optimal allocations (where the turnover is the auxiliary variable). In each case, compute the variance in the estimation of the turnover.

Sample allocation between strata

Example: Sampling of a business survey

Proportional allocation

$$n_h = n \times \frac{N_h}{N}$$

For example $n_5 = 300 \times \frac{10}{1060} \approx 3$

Optimal allocation

$$n_h = n \times \frac{N_h S_h}{\sum_k N_k S_k}$$

For example

$$n_5 = 300 \times \frac{10 \times \sqrt{2500}}{500\sqrt{2} + 300\sqrt{15} + 150\sqrt{50} + 100\sqrt{100} + 10\sqrt{2500}}$$

$$\approx 34 > 10$$

Sample allocation between strata

Example: Sampling of a business survey

Exhaustive stratum

As $34 > 10$, the last stratum is to be considered as exhaustive.

In order to determine the allocations for the four remaining strata, from now on one must act as if the question was to sample $300 - 10 = 290$ units out of the population formed by the four first strata.

Then

$$n_4 = 290 \times \frac{100\sqrt{100}}{500\sqrt{2} + 300\sqrt{15} + 150\sqrt{50} + 100\sqrt{100}}$$

$$\approx 74 < 100 \quad \text{(non-exhaustive stratum)}$$

Sample allocation between strata

Example: Sampling of a business survey

Sample allocations

Size of the firm	N_h	$\bar{y}_h$	S_h^2	Prop.	Opti.
0-9	500	10	2	142	52
10-19	300	50	15	85	86
20-49	150	200	50	42	78
50-499	100	500	100	28	74
500 and more	10	1,000	2,500	3	10

As the variance is very different from one stratum to another, the two sampling allocations are themselves very different.

Sample allocation between strata

Example: Sampling of a business survey

Variance computation

In both cases, the true values of N_h and S_h are known from the sampling frame. The calculation uses the formula:

$$V(\hat{\bar{Y}}_{str}) = \sum_{h=1}^{H} V_h = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 (1 - f_h)\frac{S_h^2}{n_h}$$

Sample allocation between strata

Example: Sampling of a business survey

Proportional allocation

For example $V_1 = \left(\frac{500}{1060}\right)^2 \times (1 - \frac{142}{500}) \times \frac{2}{142} = 2.24 \times 10^{-3}$

Size of the firm	N_h	S_h^2	n_h	V_h
0-9	500	2	142	2.24×10^{-3}
10-19	300	15	85	10.13×10^{-3}
20-49	150	50	42	17.16×10^{-3}
50-499	100	100	28	22.89×10^{-3}
500 and more	10	2,500	3	51.92×10^{-3}

Then $V(\hat{\bar{Y}}_{str-prop}) = 104.34 \times 10^{-3}$.

Sample allocation between strata

Example: Sampling of a business survey

Optimal allocation

For example $V_1 = \left(\frac{500}{1060}\right)^2 \times (1 - \frac{52}{500}) \times \frac{2}{52} = 7.67 \times 10^{-3}$

Size of the firm	N_h	S_h^2	n_h	V_h
0-9	500	2	52	7.67×10^{-3}
10-19	300	15	86	9.97×10^{-3}
20-49	150	50	78	6.16×10^{-3}
50-499	100	100	74	3.13×10^{-3}
500 and more	10	2,500	10	0

Then $V(\hat{\bar{Y}}_{str-opti}) = 26.92 \times 10^{-3}$.

Sample allocation between strata

Example: Sampling of a business survey

Conclusion

In this context, optimal allocation yields far better precision than proportional allocation.

This can be explained by the fact that the within stratum variance strongly differs from one stratum to another.

Note that in general, one does not have the true value of the variance of the variable of interest in the strata (here S_h^2).

Stratified sampling

Systematic sampling on a sorted file

Principle (reminder from unequal probabilities sampling)

1. Given a population of size N, a desired size n and first-order inclusion probabilities p_i, let's define $a_i = \sum_{j=1}^{i} p_j$
2. Draw one value X in $U_{[0;1]}$.
3. Select all the units i such that:

$$a_{i-1} \leq X + j - 1 < a_i$$

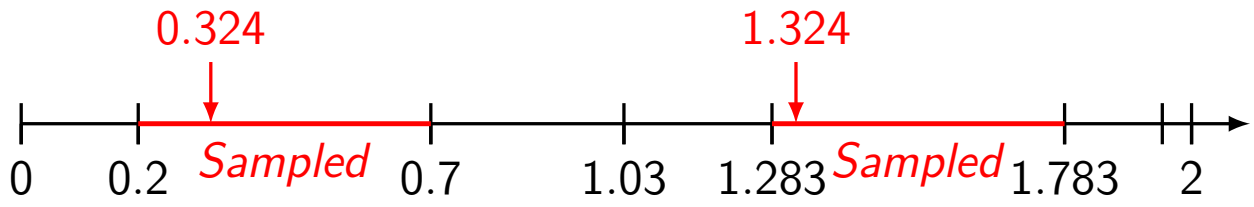
where j is an index varying from 1 to n.

Systematic sampling on a sorted file

Example

$N = 7 \quad n = 2 \quad \sum_{i=1}^{7} p_i = 2 \quad X = 0.324$

i	1	2	3	4	5	6	7
P_i	0.2	0.5	0.33	0.25	0.5	0.166	0.05
A_i	0.2	0.7	1.03	1.283	1.783	1.950	2.00



The sample drawn is $s = \{2, 5\}$.

Systematic sampling on a sorted file

Properties

- This sampling algorithm yields the desired sample size and first-order inclusion probabilities.
- Very easy and efficient to implement (it only needs to read the sampling frame one time).
- It may lead to $p_{ij} = 0$ for some i and j even after reordering of the sampling frame: estimators of the variance of the Horvitz-Thompson estimator might be biased.

Systematic sampling on a sorted file

Stratified sampling and systematic sampling

When the sampling frame is sorted by the stratification variables, the systematic sampling with equal probabilities is roughly **equivalent in terms of precision** to a stratified sampling:

- with allocation proportionate to size
- and a SRS in each stratum.

BUT its second-order inclusion probabilities differ: with such a particular ordering, **a lot of second-order inclusion probabilities equal 0**.

Systematic sampling on a sorted file

Justifications

- Systematic sampling on a sorted file yields some implicit stratification **which can only increase precision compared to SRS**.
- It allows stratification at a low level (with only a few units in each stratum), whereas explicit stratification would yield empty strata and therefore **coverage issues**.

Examples at INSEE

- In business surveys, the region (NUTS2) is often introduced implicitly as stratification variable by sorting within each stratum by region.
- In household surveys, the stratification related to the topic of the survey is introduced through systematic sampling.

Systematic sampling on a sorted file

Trade-off between variance of the Horvitz-Thompson estimator and bias of its estimator of variance

The properties of the systematic sampling on a sorted file can be summarized as a trade-off:

- On the one hand, using systematic sampling on a sorted file **always decrease the variance of the Horvitz-Thompson estimator**.
- On the second hand, the large number of null second-order inclusion probabilities yields a **biased estimator of the variance of the Horvitz-Thompson estimator**.

In practice, **a smaller variance is often preferred even if it implies that it can't be estimated without bias**.

Stratified sampling

Focus: The sampling of the PRODCOM survey

The PRODCOM survey is a European Union statistical survey on the volume of industrial output sold by product.

It is conducted each year in France in order to meet European regulation.

The firms covered by PRODCOM are those who belong to the sections B to E of the Statistical Classification of Economic Activities in the European Community (NACE) excluding agro-food industry and sawmilling and planing of wood.

In France in 2014, the sampling frame contains 146,249 units (legal units or firms) and the sample 35,003 units.

Focus: The sampling of the PRODCOM survey

Stratification

The strata are defined as the **intersection** of the following variables:

- **Economic activity**: NACE 5-digits.
- **Number of employees** coded in intervals: 0, 1-5, 6-9, 10-19, 20 and more.
- **Turnover**.

The introduction of turnover as stratification variable depends on the size of the stratum economic activity $\times$ number of employees:

- Less than 20 units: no stratification by turnover.
- Between 20 and 50 units: the median is used as stratification threshold.
- Above 50 units: the quartiles are used as stratification thresholds.

Focus: The sampling of the PRODCOM survey

Exhaustive stratum

The exhaustive stratum is defined in order to meet a Eurostat constraint: the **surveyed firms must represent 85% of the turnover in each economic activity** (NACE 5 digits).

Hence a "cut-off" rule:

- In each activity, the firms are **sorted by decreasing turnover**.
- The first firms are selected in order to ensure a **coverage rate of 85%** of the sector.

Moreover, the strata containing less than 10 units are automatically considered as exhaustive.

As a consequence, in this particular survey **the exhaustive stratum** is particularly large: 27,123 units in 2014.

Focus: The sampling of the PRODCOM survey

Allocation The remaining sample is allocated between the non-exhaustive strata according to the following rules:

- **Neyman allocation** on the turnover in each stratum...
- ...but **adapted in order to ensure at least 10 units per stratum and reliable estimations of precision**.

The special case of 3511Z: Production of electricity

- The sector 3511Z represents 18,210 units including 17,546 without any employee: domestic production.
- Neyman allocation: exhaustive stratum.
- Proportionate allocation: 2,000 units.

Solution The units with a turnover of less than 100,000€ and some legal categories (households) are excluded.

A brief conclusion

Stratification is an **efficient way to improve the precision of the estimations** when auxiliary information is available.

It requires some **methodological expertise** in the building of the strata in order to optimize the gains in accuracy and to avoid coverage issues.

The various allocation methods enable to adapt the sampling design to the objectives of each survey.

When applied to a sorted file, **the systematic sampling algorithm yields an implicit yet efficient stratification with allocation proportionate to size**.

February-March 2015 – ELSTAT

Main sampling techniques
Cluster and two-stages samplings

Martin Chevalier (INSEE)

Cluster and two-stages samplings

Cluster sampling: Principles and notations

In the context of household surveys with face-to-face interviews, the unit cost per interview may be high.

Spatial dispersion of the sampled dwellings in the case of SRS yields indeed significant travel costs (including travel time).

Hence the **core principle** of cluster sampling:

- Let's partition the population U into M parts called "clusters" and denoted $U_1, U_2, \ldots, U_g, \ldots, U_M$ so that, in each cluster g, the spatial dispersion of the units is low.
- Using a sampling design p_{CLUST}, sample m clusters and form the sample of clusters s_{CLUST}.

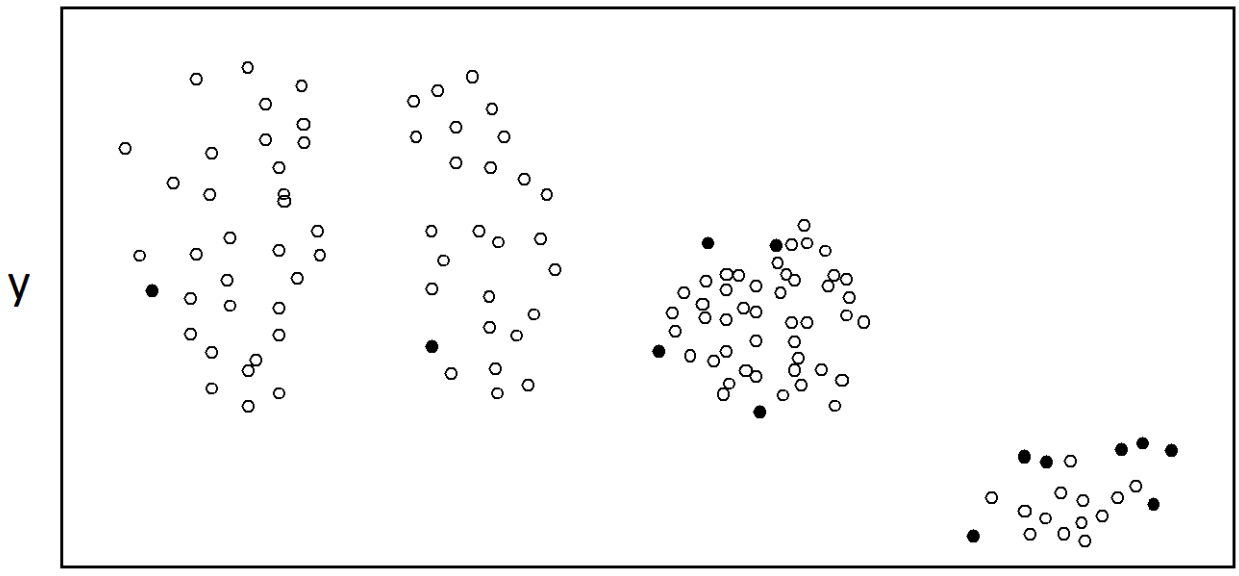
The final sample s is the union of all the units in the sampled clusters forming s_{CLUST}:

$$s = \bigcup_{g \in s_{CLUST}} U_g$$

Cluster sampling: Principles and notations

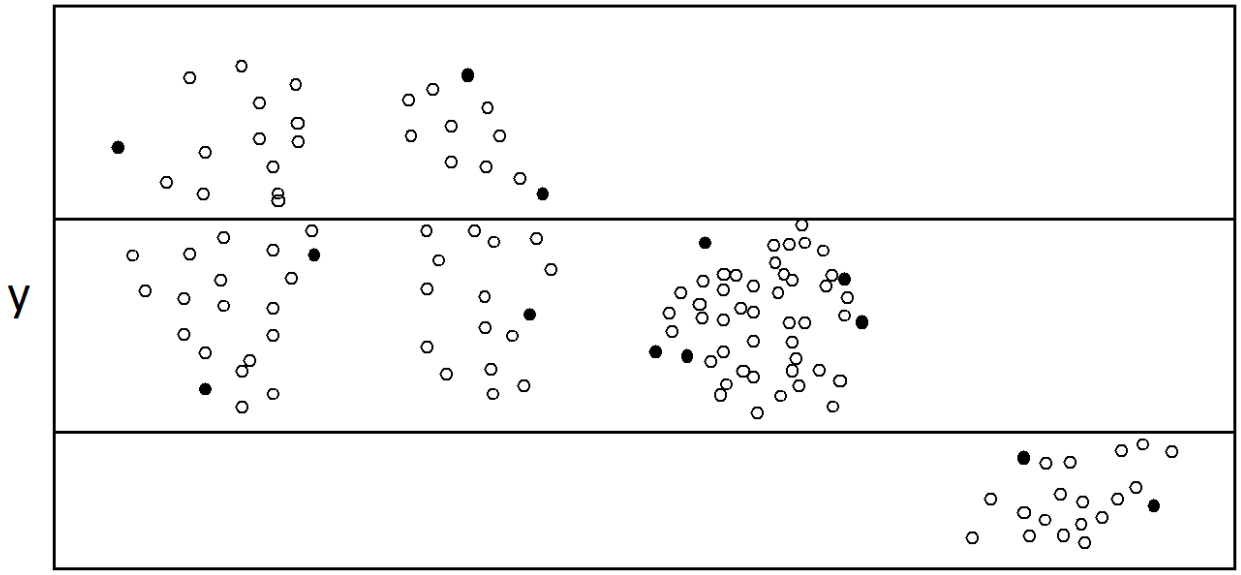
Representation of a SRS

SRS of $n = 13$ units in a population of size $N = 130$ units.



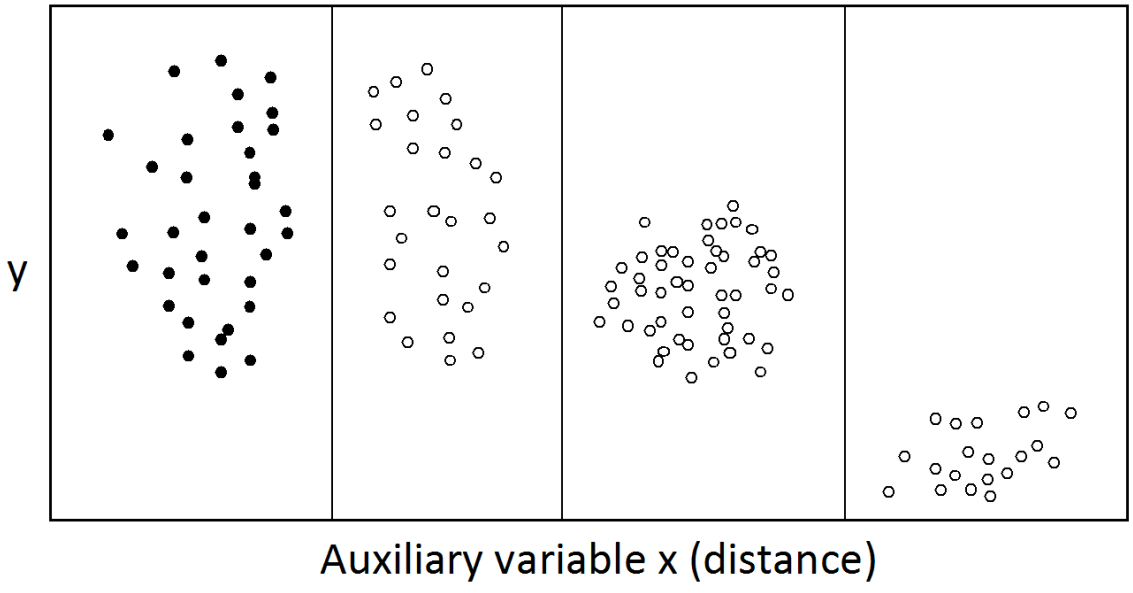
Cluster sampling: Principles and notations

Representation of a stratified sampling



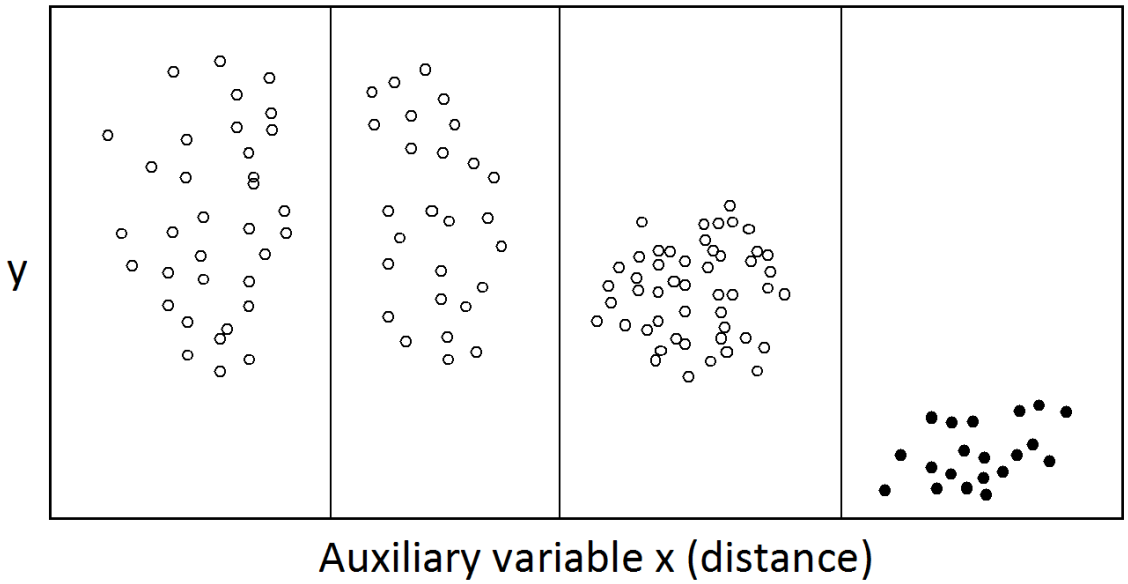
Cluster sampling: Principles and notations

Representation of a cluster sampling



Cluster sampling: Principles and notations

Representation of a cluster sampling



Cluster sampling: Principles and notations

Interpretation

In the selected clusters, the cluster sampling is a census.

Justification

If the data collection costs are strongly related to the sample drawn (e.g. face-to-face interviews *versus* telephone), cluster sampling may significantly reduce global survey costs.

If there is no sampling frame for the unit surveyed (e.g. dwellings) but a list of the clusters (e.g. neighbourhoods), cluster sampling may yield estimations with sufficient precision for a reasonable cost (compared to a census).

Cluster sampling: Principles and notations

Inclusion probabilities

The sampling design p_{CLUST} yields the following inclusion probabilities for the clusters:

$$\pi_g = P(g \in s_{CLUST})$$

$$\pi_{gh} = P(g \in s_{CLUST} \quad \text{AND} \quad h \in s_{CLUST})$$

As long as its clusters is selected, a unit is selected. Hence the first- and second-order inclusion probabilities of the units:

$$\pi_i = \pi_g \quad \text{if} \quad i \in U_g$$

$$\pi_{ij} = \begin{cases} \pi_g & \text{if} \quad i \neq j \in U_g \\ \pi_{gh} & \text{if} \quad i \in U_g, j \in U_h \end{cases}$$

Cluster sampling: Principles and notations

Horvitz-Thompson estimator

In a cluster sampling the total $T_g(Y) = \sum_{i \in U_g} y_i$ of Y in each sampled cluster g is known.

The Horvitz-Thomson estimator of the total in the population U is then

$$\hat{T}_{CLUST}(Y) = \sum_{g \in s_{CLUST}} \frac{T_g(Y)}{\pi_g}$$

with variance

$$V(\hat{T}_{CLUST}(Y)) = \sum_{g \in s_{CLUST}} \sum_{h \in s_{CLUST}} (\pi_{gh} - \pi_g \pi_h) \frac{T_g(Y)}{\pi_g} \frac{T_h(Y)}{\pi_h}$$

and $\hat{V}(\hat{T}_{CLUST}(Y))$ its unbiased estimator (Horvitz-Thompson or Yates-Grundy).

Cluster sampling: Principles and notations

Remark $V(\hat{T}_{CLUST}(Y))$ can also directly be derived from the general formula

$$V(\hat{T}(Y)) = \sum_{i \in s} \sum_{j \in s} (\pi_{ij} - \pi_i \pi_j) \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}$$

once one noticed that:

- if $(i,j) \in (U_g)^2$: $\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} = \frac{\pi_g - \pi_g^2}{\pi_g^2} = \frac{1}{\pi_g} - 1$
- if $i \in U_g$ and $j \in U_h$, $g \neq h$: $\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} = \frac{\pi_{gh} - \pi_g \pi_h}{\pi_g \pi_h}$

and uses these terms as common factors in order to form $T_g(Y)$ and $T_h(Y)$.

Cluster sampling: Principles and notations

Cluster effect

Cluster sampling may decrease survey cost for a given sample size, but it might also decrease the quality of the information collected.

Socio-economical phenomena are indeed often spatially correlated: sampling units from the same spatial area may decrease the variability of the sample with respect to Y.

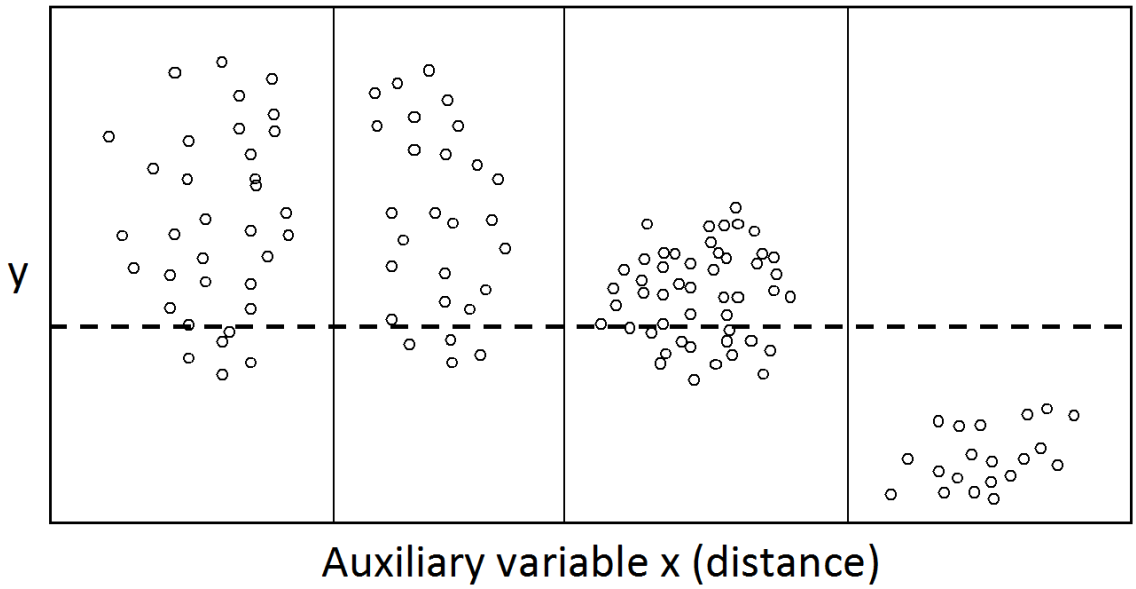
The within-cluster correlation coefficient ρ accounts for this so-called "cluster effect":

$$\rho = \frac{1}{\bar{N} - 1} \frac{\sum_{g \in s_{CLUST}} \sum_{i \in U_g} \sum_{j \in U_g, i \neq j} (y_i - \bar{y})(y_j - \bar{y})}{\sum_{g \in s_{CLUST}} \sum_{i \in U_g} (y_i - \bar{y})^2}$$

With $\bar{N}$ the mean size of the clusters. If the units within the clusters are close with respect to variable Y then $\rho > 0$.

Cluster sampling: Principles and notations

Cluster effect



Cluster and two-stages samplings

Cluster sampling: SRS of clusters

Let's use simple random sampling without replacement as sampling design p_{CLUST}. The previous results yield:

$$\hat{T}_{CLUST-SRS}(Y) = \sum_{g \in s_{CLUST}} \frac{T_g(Y)}{m/M} = M\bar{y}_{CLUST}$$

where $\bar{y}_{CLUST} = \frac{1}{m}\sum_{g \in s_{CLUST}} T_g(Y)$ is the between-cluster mean of the total of Y in each cluster and

$$\hat{V}(\hat{T}_{CLUST-SRS}(Y)) = M^2\left(1 - \frac{m}{M}\right)\frac{s^2_{CLUST}}{m}$$

where $s^2_{CLUST} = \frac{1}{m-1}\sum_{g \in s_{CLUST}}(T_g(Y) - \bar{y}_{CLUST})^2$ is the between-cluster variance of the total of Y.

Cluster sampling: SRS of clusters

Variance as a function of ρ

When the clusters are sampled using SRS, the variance of $\hat{T}_{CLUST-SRS}(Y)$ can be rewritten as

$$V(\hat{T}_{CLUST-SRS}(Y)) \approx N^2\frac{S^2_Y}{n}(1 + \rho(\bar{N} - 1) + \Delta)$$

with $\Delta = \bar{N}\frac{CV(N)}{CV(Y)}$

- As long as $\rho > 0$, $\bar{N}$ should be as little as possible : it should reach 1 and so $m = n/\bar{N} = n$.
- The clusters should have the same size.

Cluster sampling: SRS of clusters

Design effect

The design effect of a sampling for a variable Y is defined as the ratio between the variance yielded by this sampling design and the variance of a SRS of same size:

$$Deff_{CLUST-SRS}(Y) = \frac{V(\hat{T}_{CLUST-SRS}(Y))}{V(\hat{T}_{SRS}(Y))} = 1 + \rho(\bar{N} - 1) + \Delta$$

As long as $\rho > 0$ (probable due to spatial correlation) **cluster sampling is always outperformed by a SRS of same size**.

Cluster sampling: SRS of clusters

Sampling size gain (1)

But the essential goal of cluster sampling is to reduce the unit cost of an interview compared to SRS.

In order to compare the two sampling designs, one should take into account the various costs related to the organization of an interview and the different related sample sizes for a given global cost C.

Let's assume that in a cluster sampling, the global cost can be separated into two components:

$$C = mc_1 + n_{CLUST-SRS}c_2$$

The first component c_1 refers to the fixed cost of a cluster (e.g. travel cost) while the second refers to the variable cost per interview c_2.

Cluster sampling: SRS of clusters

Sampling size gain (2)

Let's assume than in the corresponding SRS, each interview implies the two components of the cost:

$$C = n_{SRS}(c_1 + c_2)$$

Then a same global cost C yields:

$$n_{CLUST-SRS} = n_{SRS} + (n_{SRS} - m)\frac{c_1}{c_2} \geq n_{SRS}$$

- The cluster sampling always yields a larger sample size than SRS.
- The sampling size gain is directly related to the ratio between fixed and variable costs.

Cluster sampling: SRS of clusters

Practical recommendations regarding cluster sampling

The dispersion of Y should be as large as possible within the clusters and as small as possible between the clusters:

$$\hat{V}(\hat{T}_{CLUST-SRS}(Y)) = M^2\left(1 - \frac{m}{M}\right)\frac{s^2_{CLUST}}{m}$$

As long as the variable Y is spatially correlated, the number of clusters should be as large as possible.

Clusters should have the same size.

Cluster sampling: SRS of clusters

Example: 1 cluster out of 3

Population U	A	B	C	D	E	F
Values	2	6	8	10	10	12

	Cluster 1	**Cluster 2**	**Cluster 3**
Units	A, B	C, D	E, F
Values	2, 6	8, 10	10, 12
Mean	4	9	11

	Cluster 1	**Cluster 2**	**Cluster 3**
Units	A, D	B, E	C, F
Values	2, 10	6, 10	8,12
Mean	6	8	10

Sampling variance (1.07 for SRS)

- First situation: $26/3 = 8.67$
- Second situation: $8/3 = 2.67$

Cluster sampling: SRS of clusters

Focus: Some elements about the sampling design of the French LFS

The Labour force survey (LFS) is one of the most important household surveys conducted in France.

It enables INSEE to compute the **unemployment rate as defined by the International Labour Organization** (ILO) on a quarterly basis, together with other labour markets statistics (e.g. employment-to-population ratio).

Since 2003 it is **conducted continuously** (each week about 4,000 dwellings are surveyed) using a **complex rotating survey design**.

The methodology is **described in details** and in English in the document: http://www.insee.fr/en/methodes/sources/pdf/methodologie_eeencontinu_anglais.pdf

Cluster sampling: SRS of clusters

Focus: Some elements about the sampling design of the French LFS

This survey must meet several constraints at a time:

- **large sample size**: to produce estimations of unemployment rate with small variance in level and in evolution, both at national and regional level, the sample size must be quite large.
- **speed of the data gathering process**: the survey must take place less than two weeks and two days after the reference week.

In order to satisfy these two constraints simultaneously while keeping the survey costs as low as possible, a **cluster sampling is used at the last sampling stage**.

Cluster sampling: SRS of clusters

Focus: Some elements about the sampling design of the French LFS

Each quarter, the dwellings surveyed by an interviewer belong to a **cluster of about 20 main dwellings**.

These clusters have been built based on **geographical proximity** and in order to yield the same sample size (controlling for main/secondary dwellings).

In collective housing, **the dwellings located on the same floor** belong to the same cluster.

The building of the clusters used informations from **land register** and **dwelling taxation** where every building is located.

Cluster sampling: SRS of clusters

Focus: Some elements about the sampling design of the French LFS

Cluster sampling: SRS of clusters

Focus: Some elements about the sampling design of the French LFS

Cluster sampling: SRS of clusters

Focus: Some elements about the sampling design of the French LFS

Cluster sampling: SRS of clusters

Focus: Some elements about the sampling design of the French LFS

In the context of a rotating sampling, **a cluster is surveyed 6 quarters in a row before being replaced**.

In order to minimize the distance between two clusters successively surveyed by the same interviewer, **clusters are grouped in so-called "sectors"** on a geographical basis.

If the sector contains more than 6 clusters, **6 clusters are sampled using simple random sampling**.

The sector are themselves sampled within primary units, themselves sampled using a stratified sampling design per region (NUTS2)... The whole design is quite complex!

Cluster and two-stages samplings

Two-stages sampling: Principles and notations

Stratified and cluster samplings both rely on a partition of the population of interest U:

- In the stratified sampling, a sampling is conducted within each stratum.
- In the cluster sampling, a census is conducted within a selection of clusters.

It is possible to encompass these two sampling techniques by distinguishing **two stages of sampling units**:

1. The M primary sampling units (PSUs) correspond to strata and clusters and form a partition of a U.
2. The N secondary sampling units (SSUs) correspond to the units of interest in the population (e.g. dwellings) and are associated with exactly one PSU.

Two-stages sampling: Principles and notations

Given a partition of PSUs, a two-stages sampling is defined by the sampling designs applied at each stage:

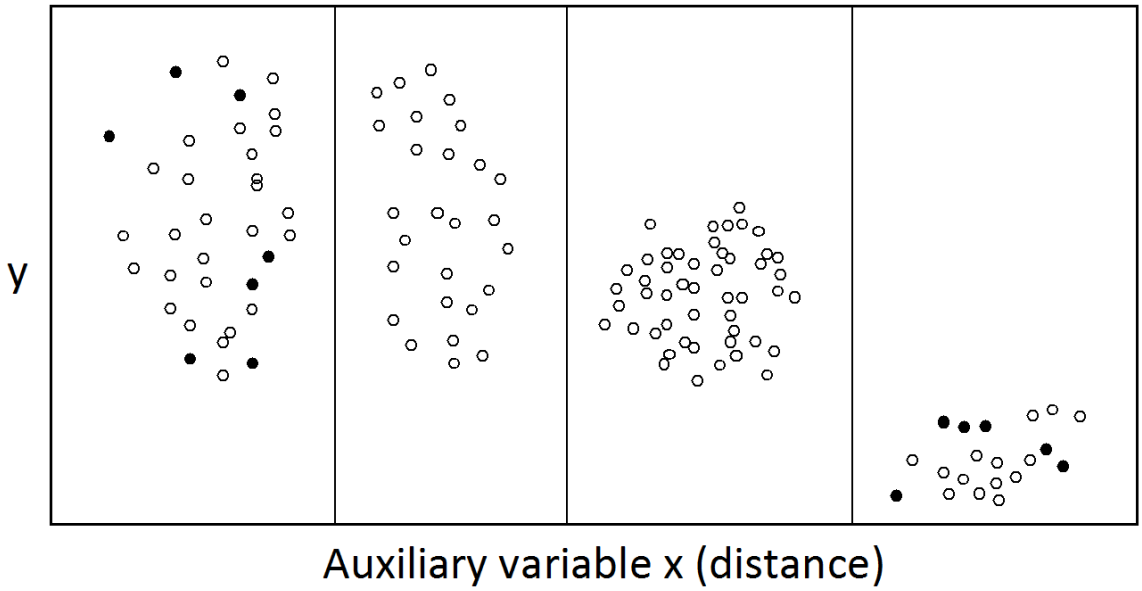
- First m PSUs are sampled out of M using a sampling design p_{PSU} and form the sample of PSUs s_{PSU}.
- Then in each **sampled** PSU g, n_g SSUs are sampled out of N_g using a sampling design p_g and form the sample of SSUs s_g.

The final sample s in the union of the m samples of SSUs:

$$s = \bigcup_{g \in s_{PSU}} s_g$$

Two-stages sampling: Principles and notations

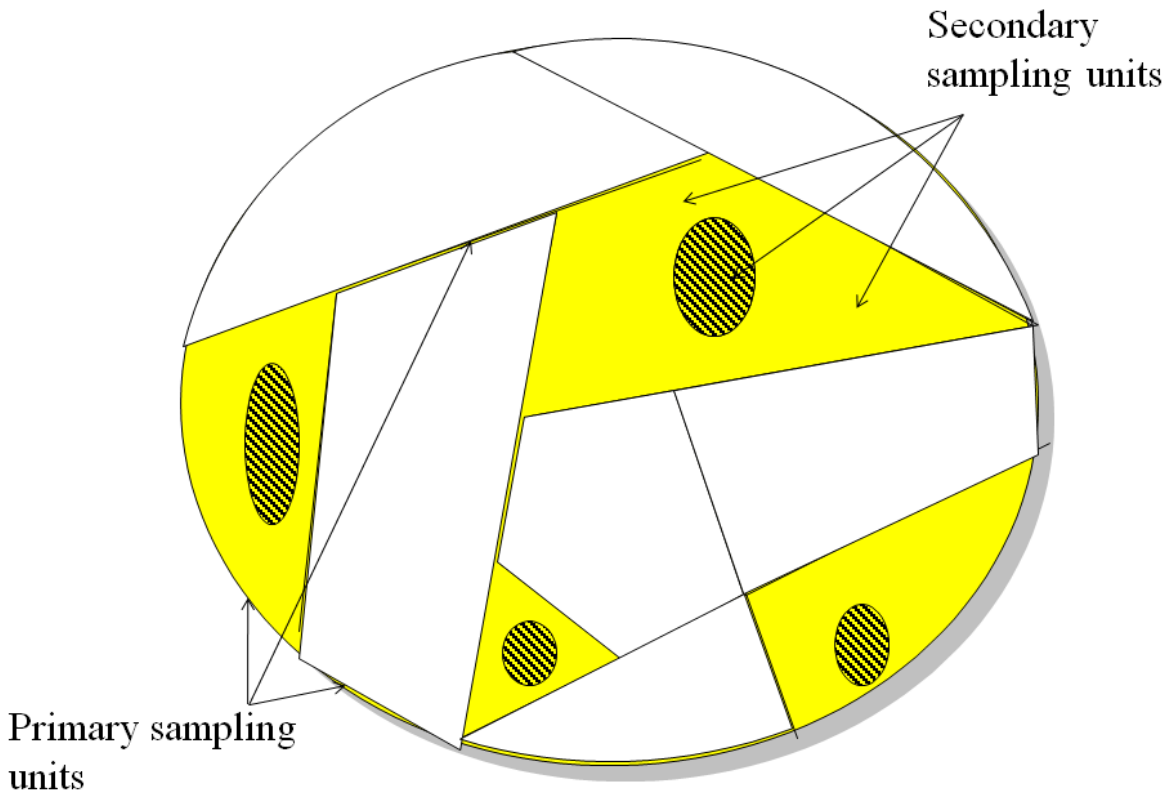
Two-stages sampling with 2 PSUs sampled out of 4, 7 SSUs sampled out of 30 in the first PSU and 6 out of 20 in the second PSU.



18 / 44

Two-stages sampling: Principles and notations

Focus: Some elements about the sampling design of the French LFS



Two-stages sampling: Principles and notations

Focus: Some elements about the sampling design of the French LFS

Inclusion probabilities

PSU The first- and second-order probabilities of the PSUs π_g and π_{gh} are determined by the sampling design of the PSUs p_{PSU}.

SSU of a sampled PSU Within a sampled PSU g, the $\pi_{i|g}$ and $\pi_{ij|g}$ are determined by the sampling design within the PSU p_g.

SSU in the population The first-order probability inclusion of a SSU i belonging to a PSU g (sampled or not) can be computed as

$$\pi_i = \pi_g \times \pi_{i|g}$$

Two-stages sampling: Principles and notations

Two-stages sampling and stratified sampling

Stratified sampling can be seen as a special case of two-stages sampling, where:

- The PSUs are the strata.
- There is no sampling at the first degree, that is

$$\forall (g,h) \in \{1,\ldots,M\}^2 \quad \pi_g = 1 \quad \text{and} \quad \pi_{gh} = 1$$

  In other terms the first degree is a census.
- The SSUs are sampled in each PSU g according to sampling design p_g. Then

$$\forall (i,j) \in U_g^2 \quad \pi_i = \pi_{i|g} \quad \text{and} \quad \pi_{ij} = \pi_{ij|g}$$

Two-stages sampling: Principles and notations

Two-stages sampling and cluster sampling

Cluster sampling can be seen as a special case of two-stages sampling, where:

- The PSUs are the clusters.
- The PSUs are sampled according to sampling design p_{CLUST} which defines π_g and π_{gh}.
- There is no sampling at the second degree, that is

$$\forall i \in U_g \quad j \in U_h \quad \pi_i = \pi_g \quad \text{and} \quad \pi_{ij} = \begin{cases} \pi_g & \text{if} \quad g = h \\ \pi_{gh} & \text{if} \quad g \neq h \end{cases}$$

  In other terms the second degree is a census.

Two-stages sampling: Principles and notations

Justification

In the context of high fixed costs (face-to-face interview with travel costs), SRS can lead to a high unit cost per interview and then smaller samples.

On the other hand, cluster sampling might affect the precision of the results when within-cluster correlation is high.

Two-stage sampling appears as a **potential compromise** between SRS and cluster sampling:

- Through sampling at the first stage, it allows to **concentrate the interviews in rather small areas**.
- Through sampling at the second stage, it allows to increase the number of PSUs and then to **decrease cluster effect**.

Two-stages sampling: Principles and notations

Generalization

It is possible to define multi-stage samples with three, four or more stages.

Stratification can for example be introduced as each stage of a two-stages sampling, in order to ensure the presence of some profiles of PSUs and SSUs in the final sample.

For household surveys at INSEE, the sampling of PSUs is stratified by region (NUTS2) while the sampling of SSUs can be stratified by several variables, depending on the subject of the survey (systematic sampling on a sorted file).

Cluster and two-stages samplings

Two-stages sampling: Estimation and precision

In the context of cluster sampling, the total $T(Y)$ of a variable Y is estimated without bias by

$$\hat{T}_{CLUST}(Y) = \sum_{g \in s_{CLUST}} \frac{T_g(Y)}{\pi_g}$$

In the context of stratified sampling, the total $T_g(Y)$ of a variable Y in stratum g is estimated without bias by

$$\hat{T}_g(Y) = \sum_{i \in s_g} \frac{y_i}{\pi_{i|g}}$$

It follows that in the context of two-stages sampling the Horvitz-Thompson estimator:

$$\hat{T}_{TS}(Y) = \sum_{g \in s_{PSU}} \sum_{i \in s_g} \frac{y_i}{\pi_g \times \pi_{i|g}}$$

estimates the total of Y in the population U without bias.

Two-stages sampling: Estimation and precision

Proof

P denotes the alea associated with the sampling of the PSUs and S the alea associated with the sampling of the SSUs.

$$\begin{aligned} E\left(\hat{T}_{TS}(Y)\right) &= E_P\left[E_S\left(\hat{T}_{TS}(Y)|P\right)\right] \\ &= E_P\left[E_S\left(\sum_{g\in s_P}\frac{\hat{T}_g(Y)}{\pi_g}|P\right)\right] \\ &= E_P\left[\sum_{g\in s_P}\frac{E_S\left(\hat{T}_g(Y)|P\right)}{\pi_g}\right] \\ &= E_P\left[\sum_{g\in s_P}\frac{T_g(Y)}{\pi_g}\right] = T(Y) \end{aligned}$$

Two-stages sampling: Estimation and precision

Variance of the Horvitz-Thompson estimator

It is possible to show that the variance of $\hat{T}_{TS}(Y)$ can be rewritten:

$$V\left(\hat{T}_{TS}(Y)\right) = V_{PSU} + V_{SSU} = V_{BETWEEN} + V_{WITHIN}$$

where

$$V_{PSU} = \sum_{g\in s_{PSU}}\sum_{h\in s_{PSU}}(\pi_{gh} - \pi_g\pi_h)\frac{T_g}{\pi_g}\frac{T_h}{\pi_h}$$

and

$$V_{SSU} = \sum_{g\in s_{PSU}}\frac{V_g}{\pi_g} \quad \text{with} \quad V_g = \sum_{i\in s_g}\sum_{j\in s_g}(\pi_{ij|g} - \pi_{i|g}\pi_{j|g})\frac{y_i}{\pi_{i|g}}\frac{y_j}{\pi_{j|g}}$$

Two-stages sampling: Estimation and precision

Proof

$$V\left(\hat{T}_{TS}(Y)\right) = V_P\left[E_S\left(\hat{T}_{TS}(Y)|P\right)\right] + E_P\left[V_S\left(\hat{T}_{TS}(Y)|P\right)\right]$$

$$E_S\left(\hat{T}_{TS}(Y)|P\right) = \sum_{g\in s_{PSU}} \frac{E_S\left(\hat{T}_g(Y)|P\right)}{\pi_g} = \sum_{g\in s_{PSU}} \frac{T_g(Y)}{\pi_g}$$

$$V_P\left[E_S\left(\hat{T}_{TS}(Y)|P\right)\right] = V_P\left[\sum_{g\in s_{PSU}} \frac{T_g(Y)}{\pi_g}\right] = V_{PSU}$$

$$V_S\left(\hat{T}_{TS}(Y)|P\right) = \sum_{g\in s_{PSU}} \frac{V_S\left(\hat{T}_g(Y)|P\right)}{\pi_g^2} = \sum_{g\in s_{PSU}} \frac{V_g}{\pi_g^2}$$

$$E_P\left[V_S\left(\hat{T}_{TS}(Y)|P\right)\right] = E_P\left[\sum_{g\in U_{PSU}} \frac{V_g}{\pi_g^2}\delta_g\right] = \sum_{g\in U_{PSU}} \frac{V_g}{\pi_g^2} E_P\left[\delta_g\right] = V_{SSU}$$

Two-stages sampling: Estimation and precision

Estimated variance of the Horvitz-Thompson estimator

If the sampling designs at the second stage **do not depend on the sample produced at the first stage**, this variance can be estimated without bias by

$$\hat{V}\left(\hat{T}_{TS}(Y)\right) = \underbrace{\sum_{g\in s_{PSU}}\sum_{h\in s_{PSU}} \frac{\pi_{gh}-\pi_g\pi_h}{\pi_{gh}}\frac{\hat{T}_g}{\pi_g}\frac{\hat{T}_h}{\pi_h}}_{(a)} + \underbrace{\sum_{g\in s_{PSU}} \frac{\hat{V}_g}{\pi_g}}_{(b)}$$

where $\hat{V}_g = \sum_{i\in s_g}\sum_{j\in s_g} \frac{\pi_{ij|g}-\pi_{i|g}\pi_{j|g}}{\pi_{ij|g}}\frac{y_i}{\pi_{i|g}}\frac{y_j}{\pi_{j|g}}$

Remark $(a)+(b)$ estimates $V_{PSU}+V_{SSU}$ without bias, however:

- (a) is an upward biased estimator of V_{PSU}
- (b) is a downward biased estimator of V_{SSU}

Two-stages sampling: Estimation and precision

Back to stratified and cluster samplings

This decomposition between $V_{BETWEEN}$ and V_{WITHIN} gives a new understanding of variance formulae for stratified and clusted samplings:

- In the stratified sampling, first- and second order inclusion probability at the first stage yield $V_{BETWEEN} = 0$. In order to reduce the variance of the estimation, the stratification should ensure a **small within-stratum variance**.
- In the cluster sampling, first- and second order inclusion probability at the second stage yield $V_{WITHIN} = 0$. In order to reduce the variance of the estimation, **the cluster means should be near from one another** i.e **small between-clusters variance**.

This opposite situation explains the **opposite recommendations** regarding strata and clusters constitution.

Cluster and two-stages samplings

Two-stages sampling: SRS at each stage

First stage m PSUs are sampled among M by SRS.

Second stage Within each sampled PSU U_g, n_g SSUs are sampled among N_g by SRS.

Horvitz-Thompson estimator

$$\hat{T}_{TS-SRS}(Y) = \frac{M}{m} \sum_{g \in s_{PSU}} \left[\frac{N_g}{n_g} \sum_{i \in s_g} yi \right] = \frac{M}{m} \sum_{g \in s_{PSU}} N_g \bar{y}_g$$

Two-stages sampling: SRS at each stage

Variance of the Horvitz-Thompson estimator

$$V\left(\hat{T}_{TS-SRS}(Y)\right) = M^2 \left(1 - \frac{m}{M}\right) \frac{S^2_{PSU}}{m} + \frac{M}{m} \sum_{g \in s_{PSU}} N_g^2 \left(1 - \frac{n_g}{N_g}\right) \frac{S_g^2}{n_g}$$

where S^2_{PSU} is the variance of the total of Y between the PSUs and S_g^2 is the variance of the total of Y within the PSUs.

Omitting the sampling rates:

$$V\left(\hat{T}_{TS-SRS}(Y)\right) \approx M^2 \frac{S^2_{PSU}}{m} + \frac{M}{m} \sum_{g \in s_{PSU}} N_g^2 \frac{S_g^2}{n_g}$$

- The size m of the sample of PSUs appears in both terms, while the size n of the sample of SSUs appears only in the second (through n_g)
- Empirically V_{PSU} is greater than V_{SSU}

Two-stages sampling: SRS at each stage

Practical recommendations

Similar recommendations than concerning cluster sampling:

- Sample more PSU and consecutively less SSU per PSU.
- Constitute the PSUs so that S^2_{PSU} is low: have PSUs with roughly the same size and the same mean for Y

$$\forall g \in \{1, \ldots, M\} \quad T_g = N_g \bar{Y}_g = constant$$

"Good" PSUs should therefore be quite numerous, with a large heterogeneity of within and a small dispersion of their mean for Y.

Two-stages sampling: SRS at each stage

Cluster and design effects

Under the assumptions that the PSUs are of same size $\bar{N}$ which leads to a sample size n/m in each PSU, it can be shown that:

$$V\left(\hat{T}_{TS-SRS}(Y)\right) \approx N^2 \frac{S^2_{PSU}}{n}(1 + \rho(n/m - 1))$$

where ρ is the cluster effect defined for the partition formed by the PSUs.

Thus

$$Deff_{TS-SRS} \approx 1 + \rho(n/m - 1) > 1$$

A two-stages sampling is in general less efficient than a SRS.

Two-stages sampling: SRS at each stage

Estimated variance of the Horvitz-Thompson estimator

$$\hat{V}\left(\hat{T}_{TS-SRS}(Y)\right) = M^2\left(1-\frac{m}{M}\right)\frac{s^2_{PSU}}{m} + \frac{M}{m}\sum_{g\in s_{PSU}} N_g^2\left(1-\frac{n_g}{N_g}\right)\frac{s_g^2}{n_g}$$

where

$$s^2_{PSU} = \frac{1}{m-1}\sum_{g\in s_{PSU}}\left(N_g\bar{y}_g - \frac{1}{m}\sum_{h\in s_{PSU}} N_h\bar{y}_h\right)^2$$

$$s_g^2 = \frac{1}{n_g-1}\sum_{i\in s_g}(y_i-\bar{y}_g)^2$$

$$\bar{y}_g = \frac{1}{n_g}\sum_{i\in s_g} y_i$$

Two-stages sampling: SRS at each stage

Remarks (1)

The size of the population is not always estimated with a null variance:

$$V(\hat{N}) = V\,\hat{T}_{TS-SRS}(1)) = V_{PSU}(1)$$

The variance of $\hat{N}$ is null only if all PSUs have the same size.

The size of the sample is not fixed: $n = \sum_{g\in s_{PSU}} n_g$ (it depends on the size if the sampled PSU), except if a constant number of SSUs are sampled in each PSU.

The first-order inclusion probability of a SSU i in PSU g $\pi_i = \frac{m}{M}\times\frac{n_g}{N_g}$ varies accoss units, unless n_g is proportionate to N_g for all g.

Two-stages sampling: SRS at each stage

Remarks (2)

The **variability in the size of the PSUs** is a source of problems in two-stages sampling with a SRS at each stage. It yields indeed **variable inclusion probabilities**, **variable size of the sample** and **variable estimations of the size of the population**.

For these reasons, one often prefers a sampling design where the **PSUs are sampled in proportion to their size** and where **the number of SSUs in each PSU is constant**.

Two-stages sampling: SRS at each stage

Example: Two-stages sampling *versus* cluster sampling

1. Cluster sampling: SRS of 1 cluster among 3
   Sampling variance 6

	Cluster 1	**Cluster 2**	**Cluster 3**
Units	A, C	B, D	E, F
Values	2, 8	6, 10	10, 12
Mean	5	8	11

2. Two-stages sampling: 2 PSUs among 3 (SRS), 1 SSU per PSU (SRS)
   Sampling variance 3.83

Selected PSUs	I,II				I,III				II,III			
SSU from PSU 1	2	2	8	8	2	2	8	8	6	6	10	10
SSU from PSU 2	6	10	6	10	10	12	10	12	10	12	10	12
Mean	4	6	7	9	6	7	9	10	8	9	10	11

Cluster and two-stages samplings

Two-stages sampling: Equally weighted sampling

First stage m PSUs are sampled among M according to a sampling with probability proportionate to their size.

Second stage Within each sampled PSU U_g, $\bar{n}$ SSU are sampled among N_g by SRS. $\bar{n}$ is constant across PSUs.

First-order inclusion probability For SSU i of PSU g:

$$\pi_i = \pi_g \times \pi_{i|g} = \frac{mN_g}{N} \times \frac{\bar{n}}{N_g} = \frac{m\bar{n}}{N} = constant$$

Size of the sample $n = m\bar{n}$ and is fixed.

This configuration thus yields an **equally weighted sampling of fixed size**.

Two-stages sampling: Equally weighted sampling

Horvitz-Thompson estimator

$$\hat{T}_{TS-EQS} = \frac{N}{m\bar{n}} \sum_{i \in s} y_i = N \times \frac{1}{n} \sum_{i \in s} y_i = N\bar{y}$$

As the sample is equally weighted, the Horvitz-Thompson estimators are the same as in SRS.

Estimated variance of the Horvitz-Thompson estimator

$$\begin{aligned} \hat{V}(\hat{T}_{TS-EQS}) = & - \frac{1}{2}\frac{N^2}{m^2} \sum_{g,h \in s_{PSU}} \frac{\pi_{gh} - \pi_g \pi_h}{\pi_{gh}} \left(\frac{\hat{T}_g}{\pi_g} - \frac{\hat{T}_h}{\pi_h} \right)^2 \\ & + \frac{N}{m\bar{n}} \sum_{g \in s_{PSU}} N_g \left(1 - \frac{\bar{n}}{N_g} \right) s_g^2 \end{aligned}$$

A brief conclusion

Cluster and two-stages sampling are to be used when one aims to **reduce the mean cost of an interview** in the context of face-to-face interviews (with significant fixed costs).

Less efficient than SRS owing to cluster effect, **they can lead to larger samples without increasing the global cost of a survey**.

When the first-stage is a sampling proportionate to size and the second a SRS with constant allocation across primary units, **two-stages sampling yields an equally weighted sample**.

February-March 2015 – ELSTAT

Main sampling techniques
Case study: The French master sample

Martin Chevalier (INSEE)

Case study: The French master sample

Principle of a master sample

Cluster and two-stages sampling are efficient methods in order to lower unit mean cost in the case of face-to-face interviews.

However, as the selected primary sampling units (PSUs) might change from one survey to another, **two-stages sampling requires a high flexibility from the network of interviewers**:

- Interviewers would eventually have to travel a long distance between the PSUs of one survey and the PSUs of another.
- Several surveys could not be be conducted at the same time.
- A significant number of interviewers would be hired specifically for one survey, which would raise training costs and lower the quality of the information collected.

Principle of a master sample

Owing to this possible change in PSUs from one survey to another, repeated two-stages samplings seem difficult and quite costly to implement.

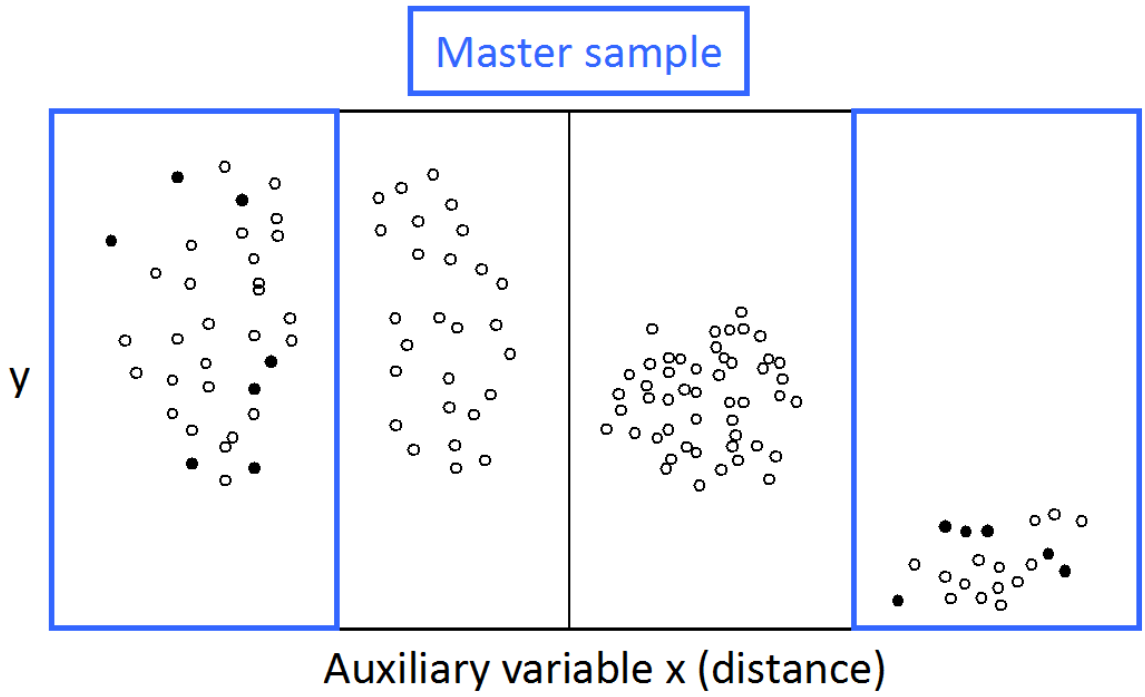
Yet it is the most efficient way to organize face-to-face interviews (household surveys) compared to simple random sampling.

Hence the **core principle** of a master sample:

- **After each census, define a partition of PSUs and draw a sample out of it**.
- **Until the next census**, **draw every sample of secondary sampling units (SSUs) in these once and for all selected PSUs**.

Principle of a master sample

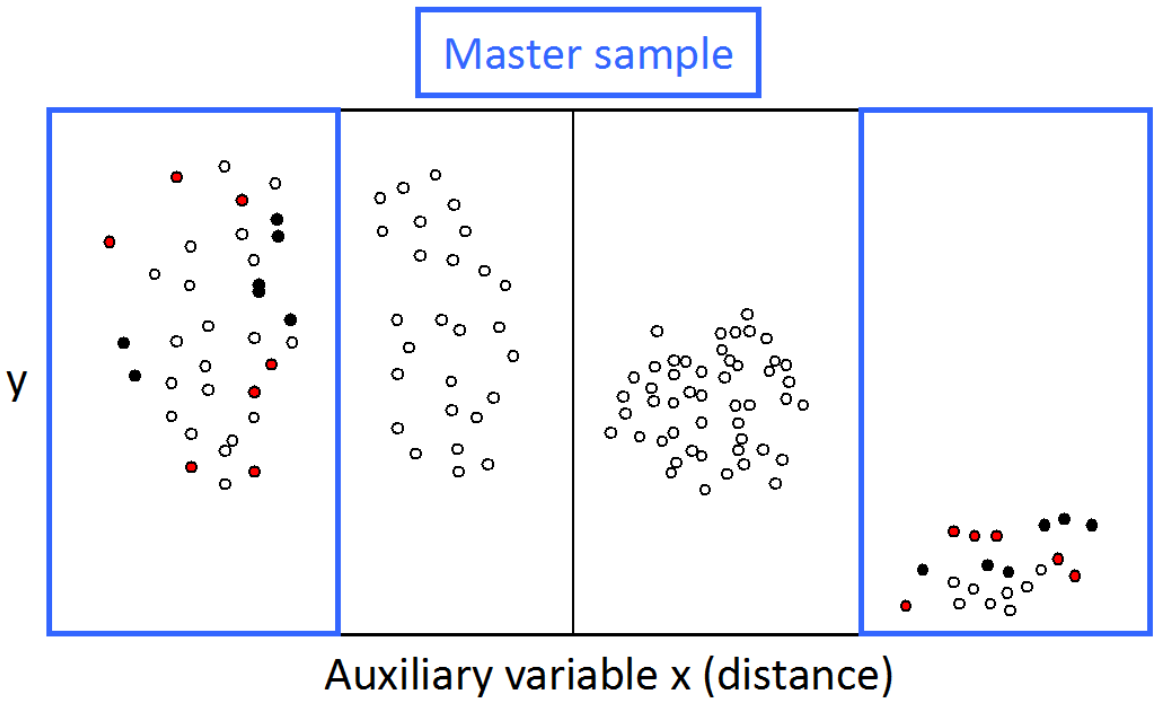
First sample after the census



Principle of a master sample

Second sample after the census

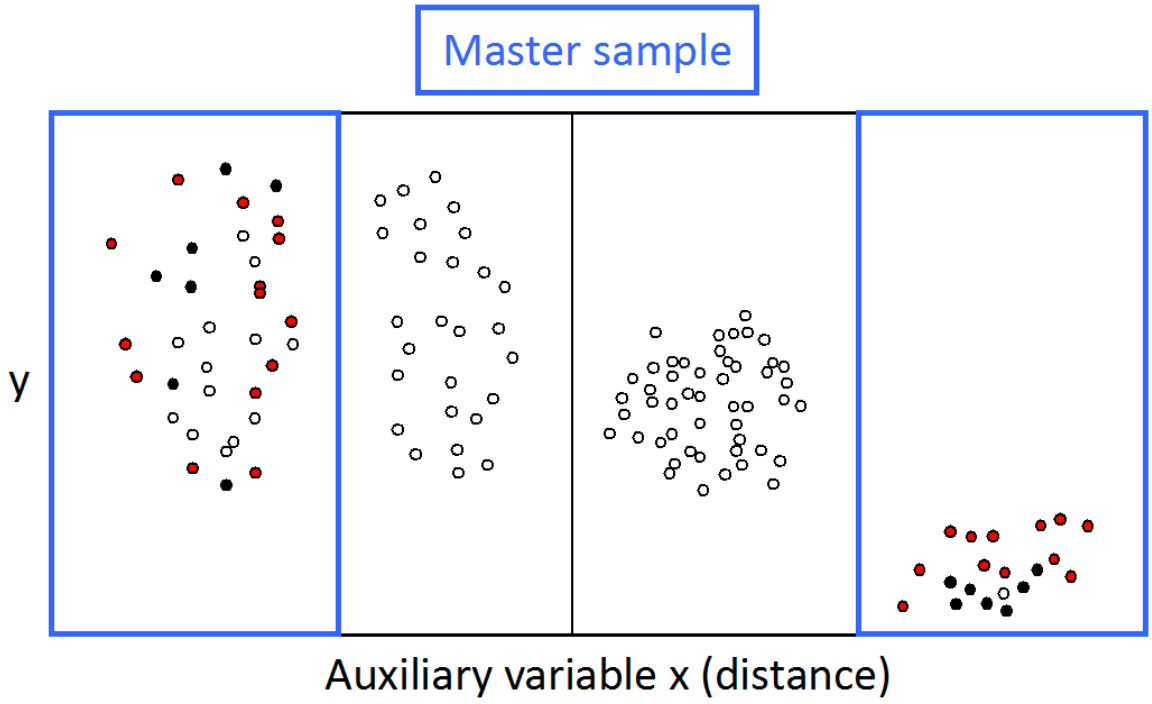
The units in red took part of a previous survey: they are "flagged" and do not participate in the current sampling.



Principle of a master sample

Third sample after the census

The units in red have been sampled by a previous survey: they are "flagged" and do not participate in the current sampling.



Principle of a master sample

Justification A master sample enables to **stabilize the network of interviewers**. This has a positive impact on the data collection process:

- Significant **reduction of the travel costs**: the interviewer lives near the PSU he or she is in charge of.
- **Flexibility in data collection organization**: several surveys can be conducted at the same time, household surveys interviewers can also participate in price index surveys.
- **Better preparation of the interviewers**: the interviewers can be hired for several years and trained accordingly which yield better response rates, better quality of the collected data.
- **Knowledge about local context and geography**: the interviewers know better how to reach the dwellings in order to reduce unit non-response.

Principle of a master sample

Challenges

Define the optimal size of PSUs in order to have enough dwellings to survey during the inter-census period.

Define the **optimal partition of PSUs regarding travel costs and precision**.

Ensure representativeness at different geographical levels (national and regional).

Draw a sample of PSUs which can be **used by any household survey**.

Case study: The French master sample

Building of the PSUs

The sampling frame: The French census

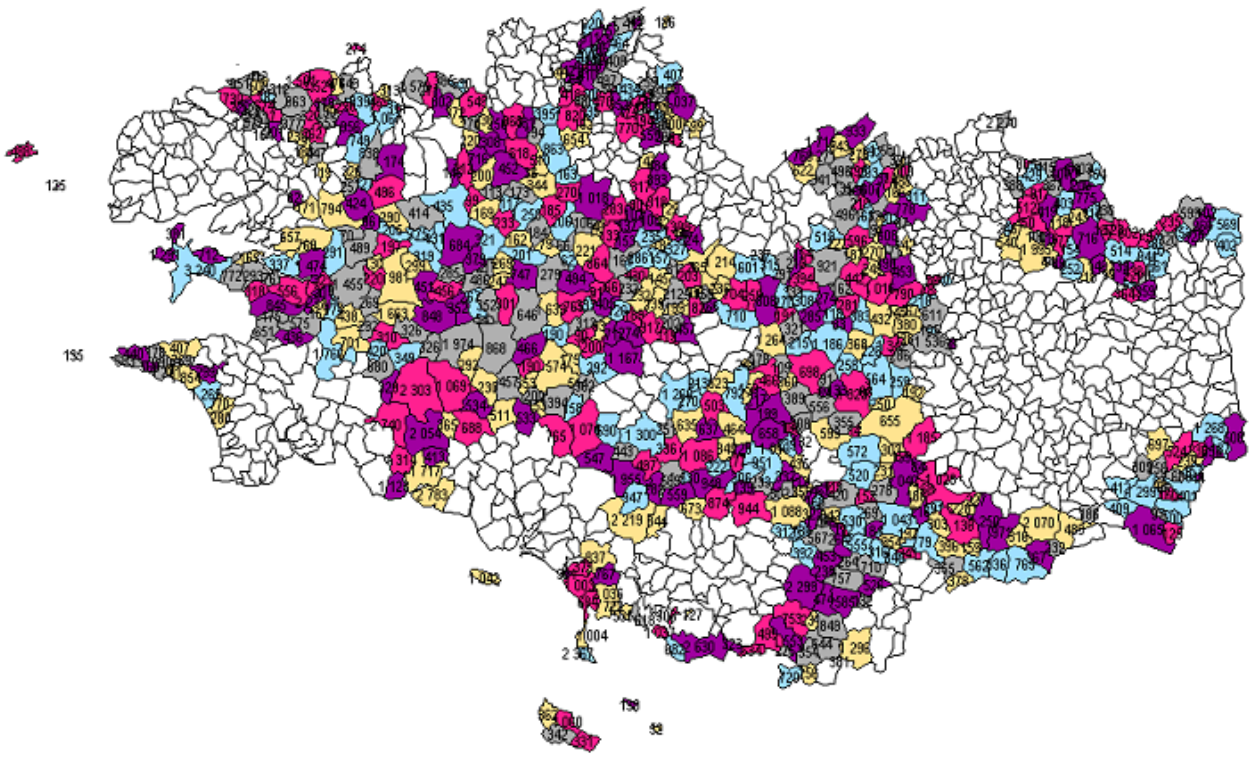
Since 2004, The French census is a rotating census:

- Small municipalities (less than 10,000 inhabitants):
  - Building 5 random samples of municipalities ("rotation groups"), with equal probabilities
  - Whole census each year of all municipalities belonging to one of the rotation groups.
- Big municipalities (over 10,000 inhabitants):
  - Building in each of them 5 samples of addresses ("rotation groups") from a file updated each year.
  - Drawing each year a sample of dwellings: the average sample rate is about 40% of all dwellings belonging to the current rotation group.
  - Census of this sample of dwellings.

Building of the PSUs

The sampling frame: The French census

Example: Brittany



Building of the PSUs

The sampling frame: The French census

Impact on the master sample

To take advantage of the "update" brought by the new census: using as a frame of a given year $n+1$ all the dwellings covered by the census at year n.

- To draw in a more efficient way samples on particular sub-populations (whose recent characteristics are known).
- To get rid of a specific system to cover new dwellings.
- To ensure that dwellings surveyed one given year will not be surveyed again before 5 years.

Building of the PSUs

The sampling frame: The French census

Impact on the building of the PSUs

Issue How to conciliate the principle of drawing **"rotating" samples** from the most recent census and building **fixed PSUs**?

Constraints and objectives Build primary units within each region in order to create a division of the territory:

- ...composed with municipalities belonging to the 5 rotation groups
- ...with a minimum number of dwellings (300) in each of them.

Building of the PSUs

PSU building process

Big municipalities Each of them constitutes one single PSU (they contain the 5 rotation groups of addresses).

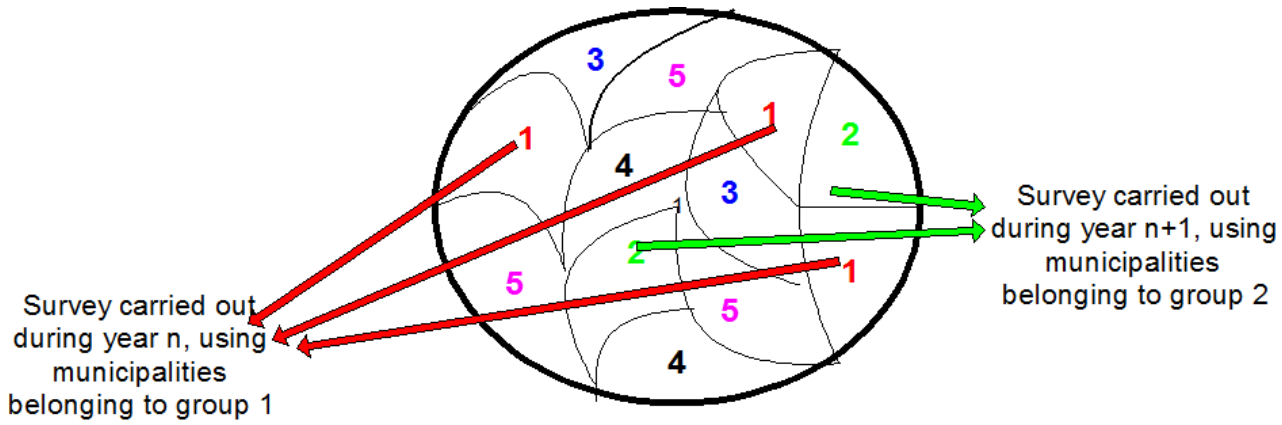
Small municipalities The aim is to build an optimal partition from the territory:

- Under constraints of minimum size (number of dwellings in each group) and with respect to regional boundaries.
- With PSUs being as spatially concentrated as possible.

For that purpose, considering the great number of constraints and the complexity of the problem, a **specific algorithm** has been implemented.

Building of the PSUs

Theoretical scheme



Building of the PSUs

Algorithm to build PSUs with small municipalities

In each region, it begins with the **largest municipality** (number of main dwellings) among the small ones: there is an attempt to build a PSU around this municipality (that will be the "center" or "**pivot**" of the PSU).

A PSU is achieved if, among municipalities of the same region (not yet allocated) and whose distance to the pivot is less than a given threshold, **it is possible to find enough municipalities in order to reach 300 main dwellings in each rotation group**. If not, the PSU is not validated.

At each step, the biggest municipality not yet allocated to one PSU is tested as a possible pivot. At the end, all remaining communities are allocated to the closest PSU (if the distance to the "center municipality" does not exceed the fixed threshold).

Building of the PSUs

Simulations carried out in order to find the "optimal" partition

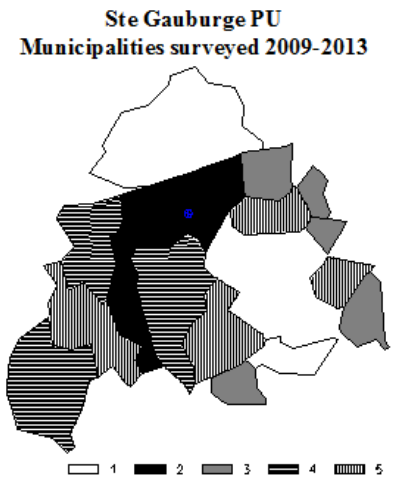
Automatic process of building of PSU developped, several values of the threshold tested.

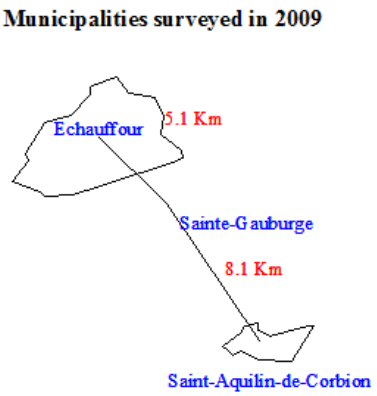
Maximal extent	Number of PSUs built	Unaffected municipalities	Mean extent
10	1,788	10,996	7.8
15	2,565	1,746	10
18	2,779	645	10.9
19	2,848	465	11.2
20	2,886	363	11.4
21	2,944	247	11.7
22	2,969	175	11.9
...	...	...	...
27	3,115	32	12.9
28	3,144	15	13.2

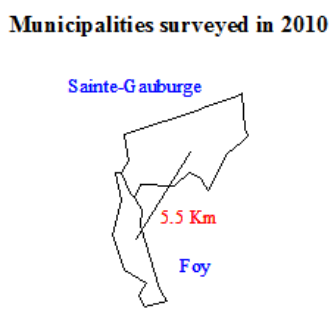
Criteria Number of unaffected municipalities and the extent of the PSUs. **Chosen threshold value: 20 km**. All 363 remaining municipalities have been affected to a PSU.

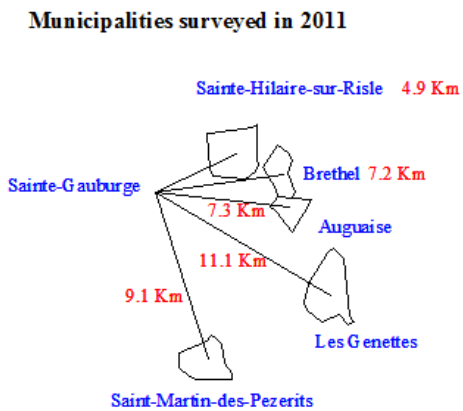
Building of the PSUs

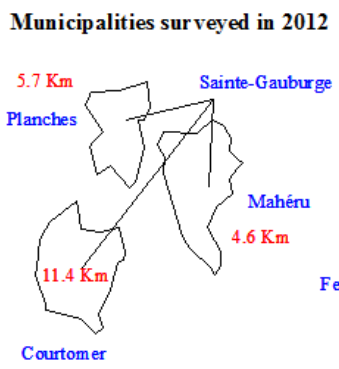
Example: Sainte-Gauburge

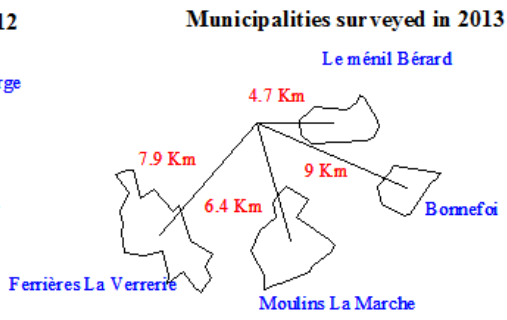












Building of the PSUs

Result of PSUs building

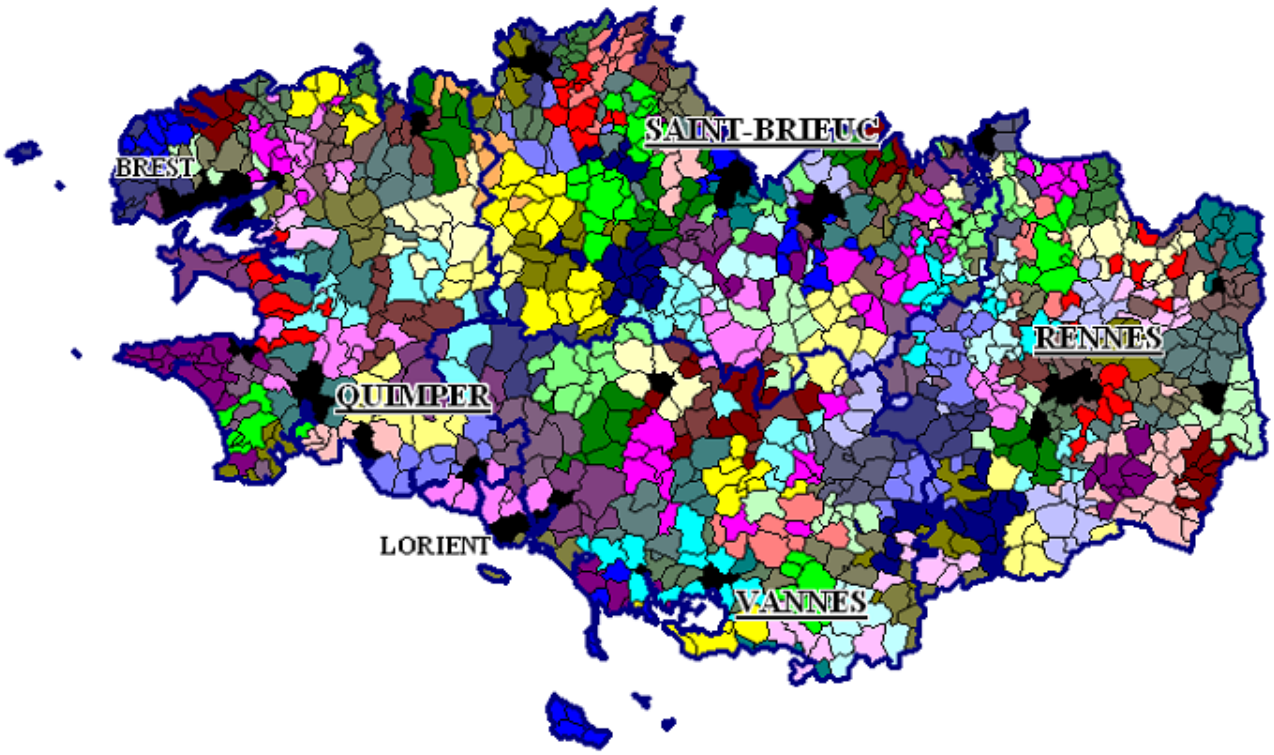
3,785 PSUs

- 892 PSUs for big municipalities (over 10,000 inhabitants): the three major cities Paris, Lyon and Marseille are divided into several districts.
- 2,893 PSUs for small municipalities (less than 10,000 inhabitants).

Remark The algorithm for building PSUs is deterministic but the initial assignment of municipalities to different rotation groups is random.

Building of the PSUs

Example: Back to Brittany



Case study: The French master sample

Sampling of the primary units

Number of PSUs to be sampled and exhaustive PSUs

Basic hypotheses

- PSUs are drawn proportionally to their size (number of main dwellings)
- Some of them are systematically kept (exhaustive or "take-all PSUs").

Parameters

- A regular household survey includes about 12,000 dwellings, which means a sampling rate $\tau = \frac{1}{2000}$.
- Except for the take-all PSUs, there are 1 interviewer and $e = 20$ units sampled per PSU and per survey.

Note With 300 dwellings per rotation group, this implies a maximum of 15 surveys per year.

Sampling of the primary units

Number of PSUs to be sampled and exhaustive PSUs

These parameters yield a threshold T for take-all PSUs. One can indeed show that:

$$T = \frac{e}{\tau}$$

Proof Let denote k the number of non-exhaustive PSUs to sample and N^{exh} the size of the population in the exhaustive PSUs. Then

$$k = \frac{\tau(N - N^{exh})}{e}$$

Moreover, denoting N_g the size of the non-exhaustive PSU g, the sampling probability of g is

$$\pi_g = k\frac{N_g}{N - N^{exh}} = \frac{\tau N_g}{e}$$

As $\pi_g < 1$, one can define T as the value of N_g such as

$$\pi_g = 1 \Leftrightarrow \frac{\tau T}{e} = 1 \Leftrightarrow T = \frac{e}{\tau}$$

Sampling of the primary units

Number of PSUs to be sampled and exhaustive PSUs

Results

Given the parameters, the value of the threshold is **40,000 inhabitants**. All exhaustive PSUs are then big municipalities ($>$ 10,000 inhabitants).

37 big municipalities are exhaustive PSUs assigned to several interviewers.

488 non-exhaustive PSUs are to be sampled in order to have the desired number of dwellings.

Sampling of the primary units

Drawing of the primary units

Sampling strategy

Stratified according to the regions (NUTS2). Particular case: Paris area splitted in two "crowns".

Balanced on regional totals:

- It is necessary to balance not only on the level of whole PSU but also for **each rotation group**...
- ...in order to benefit each year from a "representative" sampling frame.

It increases the number of balancing constraints and reduces the number of allowed balancing variables.

Sampling of the primary units

Drawing of the primary units

Balancing variables

Number of main dwellings of municipalities belonging to the PSU, for each of the five rotation groups.

Total income (from tax sources) of municipalities belonging to the PSU, for each of the five rotation groups.

Total number of dwellings in the whole PSU in peri-urban areas, rural areas and urban areas.

Additional balancing variables were used in Paris area (e.g. age, household structure).

Sampling of the primary units

Drawing of the primary units

The master sample and regional extensions

The master sample yields results with a good enough precision only at a national level.

In order to address the question of regional extensions, a broader master sample is defined (called "EMEX").

- for regular regional extensions (roughly twice as many dwellings as in the master sample), a first EMEX is defined (restricted EMEX).
- for broader regional extensions (roughly thrice as many dwellings as in the master sample), a second EMEX is defined (enlarged EMEX).

The master sample, restricted EMEX and enlarged EMEX are **drawn simultaneously using multi-phases sampling**, in order to control inclusion probabilities in each sample.

Sampling of the primary units

Drawing of the primary units

Quality of the sample

One looks at the quality of the sample of PSUs per rotating group, comparing:

- the estimate (from the sample of PSUs) of totals of the number of main dwellings in rural space (for example)
- with the true total known through census 1999.

Rotating group	Relative error
1	+3.4%
2	-3.3%
3	-7.9%
4	-8.1%
5	-9.4%

Sampling of the primary units

Drawing of the primary units

Calibration

In order to obtain a yearly "representative" sampling frame, the weights of the PSUs are each year **calibrated on the last census**.

Calibration is a technique that ensures that **the value of a total in a sample** (here the sampled PSUs) **corresponds to its counterpart in a population** (here the census).

Moreover, calibration asymptotically yields **unbiased estimators and better precision**:

- Relative error equals zero for all calibration variables...
- ...and does not increase for other variables of interest.

More about calibration will be discussed in session 2 with Emmanuel Gros.

Case study: The French master sample

Principle of a master sample

Building of the PSUs

Sampling of the primary units

Sampling of the secondary units

Sampling of the secondary units

Allocation per PSU

In the **exhaustive PSUs**, the sampling allocation is proportionate to their size in number of main dwellings.

In the **non-exhaustive PSUs**, two contradictory objectives:

- **Equal final weights** of the drawn dwellings.
- **Equal number of dwellings drawn in each PSU** (in order to have the same workload for each interviewer).

The allocation algorithm tries to **minimize the dispersion of final dwelling weights** with the following constraints:

- **The total size of the sample is fixed**.
- A **lower and an upper bound** of workload per interviewer is defined (between 20 and 40 interviews).

Sampling of the secondary units

Sampling frame in each PSU

In each PSU, the sampling frame for the surveys of a given year t is constituted by the dwellings covered by the census in year $t-1$.

The rotating census allows the auxiliary information used in the sampling to be quite "fresh": the probability that the characteristics of a dwelling change between the census and the survey is lower than with a traditional census.

In order to avoid the situation where a dwelling is surveyed by two different surveys within a short time, the dwellings are "flagged" once they enter a sample.

A "flagged" dwelling does not participate in the following samplings.

Sampling of the secondary units

Sampling design of the SSUs

In each PSU, the SSUs are drawn using a **systematic sampling algorithm**.

In most cases, the sampling frame is **sorted by some variable of interest for the survey**, yielding an implicit yet efficient stratification with proportional allocation.

Example In the 2013 survey about accommodation, the sampling frame was sorted by:

- housing occupation status (tenant rather than owner),
- period of construction of the dwelling,
- housing type (building or house).

Sampling of the secondary units

Special case: Survey with over-representation

Some surveys requires an **over-representation of some dwellings** on the basis of information from the sampling frame.

This is implemented in the sampling program through **stratification with different sampling rates per stratum**.

Yet a problem remains: as the dwellings are "flagged" once they enter a sample, **such over-representation might change the statistical properties of the sample of the remaining dwellings**.

Example If a survey dramatically over-represent the persons without qualification, the following surveys might be biased towards qualified persons.

Sampling of the secondary units

Special case: Survey with over-representation

In order to avoid this phenomenon, such a sample is drawn through **two-phases sampling**:

- the first phase is a **simple random sampling**,
- within the sample of first phase, a stratified sampling with specific allocations (depending on the desired over-representation) is drawn.

The key idea here is that every dwelling sampled **during the first phase** is flagged: the remaining stock of dwellings is therefore not biased by the over-representation of the survey.

A brief conclusion

Introducing a master sample can improve the **reliability**, the **efficiency** and the **methodological quality** of a sampling design.

Its numerous advantages come with a cost in terms of **organization** and **methodological complexity**.

In the French case, the rotating census allows the use of quite **"fresh" auxiliary information** in order to optimize representativeness at the first stage (calibration) and allocation across PSUs at the second stage.

Such a complex sampling design raises however the issue of **variance estimation**.