Main sampling techniques

ELSTAT Training Course

January 23-24 2017

Martin Chevalier
Department of Statistical Methods
Insee

Outline

Sampling theory

Sampling theory

Context

Goal Obtain information about a given population U (of size N).

Census is an option, but:

- expensive;
- can contain only a limited number of questions;
- one can expect some strong decrease in response rate.

Hence **sample surveys**: only a sample s (of size n) takes part in the survey, but **the inference is made** on the whole population U.

Remark **Probabilistic sampling**

- $\neq$ **deterministic sampling**: the sample is drawn at random ($\neq$ volunteers, $\neq$ chosen on purpose);
- $\neq$ **quota sampling**: once a unit is sampled, it should answer the survey.

Sampling frame

The **sampling frame** is a database containing **minimal information** about the **inference population** U:

- identifiers (must be unique !);
- contact information;
- possibly auxiliary information:
  - on individuals: sex, age, income, profession;
  - on firms: number of employees, turnover.

Remark In some contexts one sampling frame does not cover the whole **inference population**:

- multiple sampling frames with coverage issues;
- panel survey.

A complex methodology allows for an unbiased estimation $\rightarrow$ detailed by **Pascal Ardilly** on Thursday and Friday.

Sampling design

Let $\mathcal{U}$ denote the **power set** of U, *i.e.* the **set of all subsets of** U.

The sampling design p is a probability distribution defined on $\mathcal{U}$ such as:

1. $\forall s \in \mathcal{U}, \quad p(s) \geq 0$
2. $\sum_{s \in \mathcal{U}} p(s) = 1$

Example $U = \{a, b, c\}$ hence

$$\mathcal{U} = \{\emptyset, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, \{a, b, c\}\}$$

Sampling design: example

It is then possible to define the sampling design p_1...:

$$\begin{array}{lll} p_1(\{a\}) = 0.1 & p_1(\{a,b\}) = 0.1 & p_1(\{a,b,c\}) = 0.2 \\ p_1(\{b\}) = 0.1 & p_1(\{a,c\}) = 0.1 & p_1(\emptyset) = 0.2 \\ p_1(\{c\}) = 0.1 & p_1(\{b,c\}) = 0.1 & \end{array}$$

... or the sampling design p_2:

$$\begin{array}{lll} p_2(\{a\}) = 0 & p_2(\{a,b\}) = 0.5 & p_2(\{a,b,c\}) = 0 \\ p_2(\{b\}) = 0 & p_2(\{a,c\}) = 0.25 & p_2(\emptyset) = 0 \\ p_2(\{c\}) = 0 & p_2(\{b,c\}) = 0.25 & \end{array}$$

Remark p_2 is a sampling design of **fixed size**.

Estimation mechanism

Let Y denote the variable and $\theta(Y)$ the parameter of interest defined on the whole population U. For example:

- $\theta(Y) = T(Y)$ for the total of Y;
- $\theta(Y) = \bar{Y}$ for the mean of Y.

Using only the information from a drawn sample s, one can compute $\hat{\theta}_s(Y)$, an estimator of $\theta(Y)$.

Question On average, given the sampling design $p(s)$, how far is $\hat{\theta}_s(Y)$ from the true value in the population $\theta(Y)$?

Sampling error

Bias

$$B(\hat{\theta}(Y)) = \mathbb{E}(\hat{\theta}(Y)) - \theta(Y) = \sum_{s \in \mathcal{U}} p(s)\hat{\theta}_s(Y) - \theta(Y)$$

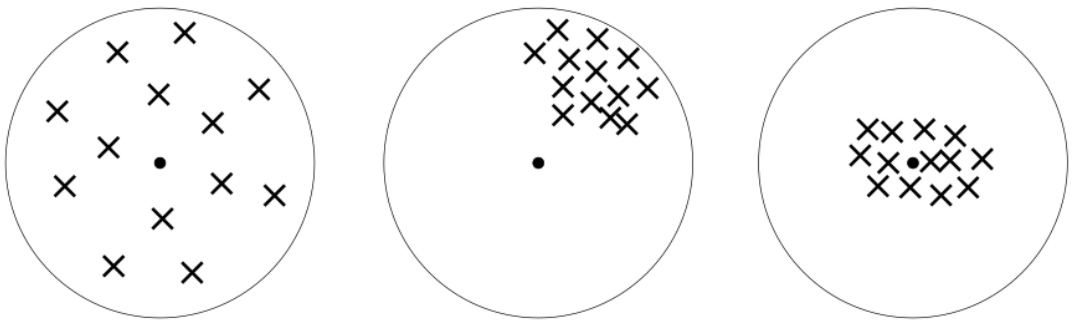
If $B(\hat{\theta}(Y)) = 0$ then $\hat{\theta}(Y)$ is said to be **unbiased**.

Variance

$$V(\hat{\theta}(Y)) = \sum_{s \in \mathcal{U}} p(s) \left[\hat{\theta}_s(Y) - \mathbb{E}(\hat{\theta}(Y))\right]^2$$

The smaller the variance, the more accurate the estimator.

Sampling error



1. No bias, big variance
2. Bias, small variance
3. No bias, small variance

Sampling error: example

Let's go back to $U = \{a, b, c\}$ and p_2 defined by:

$$p_2(\{a, b\}) = 0.5 \quad p_2(\{a, c\}) = 0.25 \quad p_2(\{b, c\}) = 0.25$$

The parameter of interest is the mean of a given variable Y:

	a	b	c
Y	20	10	3

$\theta(Y) = \bar{Y} = 11$ in the whole population.

Estimator Let the natural mean $\hat{\theta}(Y) = \bar{y}$ be the estimator of $\theta(Y)$:

$$\bar{y}_{\{a,b\}} = 15 \quad \bar{y}_{\{a,c\}} = 11.5 \quad \bar{y}_{\{b,c\}} = 6.5$$

Sampling error: example

$$\begin{aligned} B(\bar{y}) &= \sum_{s \in \mathcal{U}} p(s)\bar{y}_s - \bar{Y} \\ &= (0.50 \times 15 + 0.25 \times 11.5 + 0.25 \times 6.5) - 11 \\ &= 12 - 11 = 1 \neq 0 \end{aligned}$$

Given the sampling design p_2 the natural mean is a **biased estimator**.

$$\begin{aligned} V(\bar{y}) &= \sum_{s \in \mathcal{U}} p(s) \left[\bar{y}_s - \mathbb{E}(\bar{y}_s)\right]^2 \\ &= 0.50 \times (15 - 12)^2 + 0.25 \times (11.5 - 12)^2 \\ &\qquad + 0.25 \times (6.5 - 12)^2 \\ &= 12.125 \end{aligned}$$

Inclusion probabilities

In order to define an **unbiased estimator** under a given sampling design p, let's introduce the **first- and second-order inclusion probabilities**.

- first-order inclusion probabilities:

$$\pi_i = \sum_{s \in \mathcal{U}^{(i)}} p(s)$$

  where $\mathcal{U}^{(i)}$ is the set of subsets of U contaning unit i
- second-order inclusion probabilities:

$$\pi_{ij} = \sum_{s \in \mathcal{U}^{(ij)}} p(s)$$

  where $\mathcal{U}^{(ij)}$ is the set of subsets of U contaning both units i and j

Inclusion probabilities: example

Let's go back to $U = \{a, b, c\}$ and p_2 defined by:

$$p_2(\{a, b\}) = 0.5 \quad p_2(\{a, c\}) = 0.25 \quad p_2(\{b, c\}) = 0.25$$

Hence

- $\pi_a = p_2(\{a, b\}) + p_2(\{a, c\}) = 0.75$
- $\pi_b = p_2(\{a, b\}) + p_2(\{b, c\}) = 0.75$
- $\pi_c = p_2(\{a, c\}) + p_2(\{b, c\}) = 0.50$

And:

- $\pi_{a,b} = p_2(\{a, b\}) = 0.50$
- $\pi_{a,c} = p_2(\{a, c\}) = 0.25$
- $\pi_{b,c} = p_2(\{b, c\}) = 0.25$

Horvitz-Thompson estimator of a total

In this context, the **Horvitz-Thompson estimator of a total** is defined as:

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

One can demonstrate that **the Horvitz-Thompson estimator of a total is unbiased** under the sampling design p:

$$B(\hat{T}^{HT}(Y)) = \mathbb{E}(\hat{T}^{HT}(Y)) - T(Y) = 0$$

The Horvitz-Thompson estimator is a **weighted estimator**, where the sampling weights are defined as

$$\forall i \in s \quad d_i = \frac{1}{\pi_i}$$

.

Unbiasedness of the HT estimator: proof

$$\begin{aligned}
\mathbb{E}\left[\hat{T}^{HT}(Y)\right] &= \mathbb{E}\left[\sum_{i \in s} \frac{y_i}{\pi_i}\right] = \mathbb{E}\left[\sum_{i \in U} \frac{y_i \times \delta_i}{\pi_i}\right] \\
&= \sum_{i \in U} \frac{y_i \times \mathbb{E}\left[\delta_i\right]}{\pi_i} = \sum_{i \in U} \frac{y_i \times \pi_i}{\pi_i} \\
&= \sum_{i \in U} y_i = T(Y)
\end{aligned}$$

where δ_i is a so-called **Cornfiled variable**:

$$\delta_i = \begin{cases} 1 & \text{if } i \in s \\ 0 & \text{if } i \notin s \end{cases}$$

Estimators of a mean

The **Horvitz-Thompson estimator of a mean** directly derives from the Horvitz-Thompson estimator of a total:

$$\hat{\bar{Y}}^{HT} = \frac{\hat{T}^{HT}(Y)}{N} = \frac{1}{N}\sum_{i\in s}\frac{y_i}{\pi_i}$$

This estimator **requires the total size of the population N to be known**.

When it is not the case, one can compute the (slightly biased) **Hájek estimator of the mean**:

$$\hat{\bar{Y}}^{Hájek} = \frac{1}{\hat{N}}\sum_{i\in s}\frac{y_i}{\pi_i}$$

where $\hat{N} = \sum_{i\in s}\frac{1}{\pi_i}$ is an estimator of the population size.

Variance of the HT estimator

One can demonstrate that the variance of the Horvitz-Thompson estimator of a total is:

$$V\left(\hat{T}^{HT}(Y)\right) = \sum_{i\in U}\sum_{j\in U}(\pi_{ij} - \pi_i\pi_j)\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}$$

If the sampling design if of **fixed size**, it can be rewritten (Sen-Yates-Grundy):

$$V^{SYG}\left(\hat{T}^{HT}(Y)\right) = -\frac{1}{2}\sum_{i\in U}\sum_{\substack{j\in U\\ j\neq i}}(\pi_{ij} - \pi_i\pi_j)\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$$

Remark Both formulae rely on **summations on the whole population**, which are by definition unknown in practice: **variance has to be estimated**.

Variance estimator of the HT estimator

Each of the previously defined variance formula has its unbiased estimator:

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

$$\hat{V}^{SYG}\left(\hat{T}^{HT}(Y)\right) = -\frac{1}{2}\sum_{i\in s}\sum_{\substack{j\in s\\ j\neq i}}\frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_{ij}}\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2$$

Remark Sen-Yates-Grundy condition
If $\forall i, j \neq i \quad \pi_{ij} - \pi_i\pi_j \leq 0$ then $\hat{V}^{SYG}\left(\hat{T}^{HT}(Y)\right)$ can only take positive values.

Horvitz-Thompson estimaton: example

Let's go back to $U = \{a, b, c\}$. Recall that under the sampling design p_2:

$$\pi_a = 0.75 \quad \pi_b = 0.75 \quad \pi_c = 0.50$$

and that the interest variable is Y:

	a	b	c
Y	20	10	3

Estimator Hence

$$\hat{\hat{Y}}^{HT}_{\{a,b\}} = \frac{1}{3}\left(\frac{20}{0.75}+\frac{10}{0.75}\right) = 13.3 \quad \hat{\hat{Y}}^{HT}_{\{a,c\}} = \frac{1}{3}\left(\frac{20}{0.75}+\frac{3}{0.50}\right) = 10.9$$

$$\text{and } \hat{\hat{Y}}^{HT}_{\{b,c\}} = \frac{1}{3}\left(\frac{10}{0.75}+\frac{3}{0.50}\right) = 6.4$$

Horvitz-Thompson estimaton: example

$$
\begin{aligned}
B(\hat{\bar{Y}}^{HT}) &= \sum_{s \in \mathcal{U}} p(s) \hat{\bar{Y}}_s^{HT} - \bar{Y} \\
&= (0.50 \times 13.3 + 0.25 \times 10.9 + 0.25 \times 6.4) - 11 \\
&= 11 - 11 = 0
\end{aligned}
$$

Given the sampling design p_2 the Horvitz-Thompson estimator is **unbiased** (this is always true!).

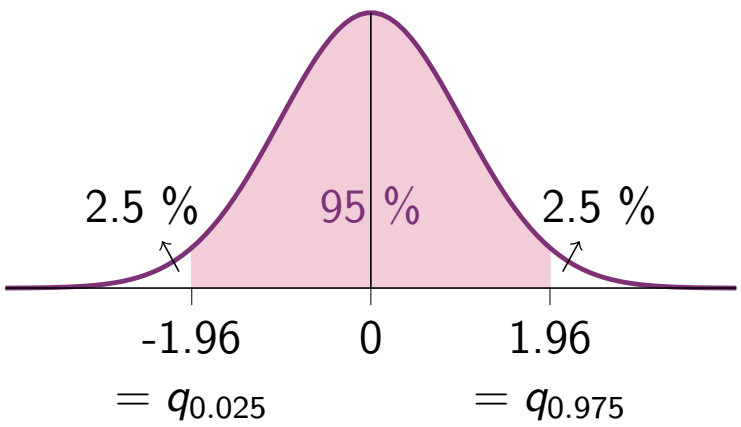
$$
\begin{aligned}
V(\hat{\bar{Y}}^{HT}) &= \sum_{s \in \mathcal{U}} p(s) \left[\bar{y}_s - \mathbb{E}(\bar{y}_s)\right]^2 \\
&= 0.50 \times (13.3 - 11)^2 + 0.25 \times (10.9 - 11)^2 \\
&\qquad + 0.25 \times (6.4 - 11)^2 \\
&= 7.9
\end{aligned}
$$

Confidence interval

Under some regularity conditions, one can establish

$$
\frac{\hat{T}^{HT}(Y) - T(Y)}{\sqrt{\hat{V}\left(\hat{T}^{HT}(Y)\right)}} \xrightarrow{d} \mathcal{N}(0,1)
$$

In other words: the Horvitz-Thompson estimator of a total is **asymptotically normally distributed**:



Confidence interval

For n "big enough", this result allows for hypothesis testing or confidence intervals. One can demonstrate that:

$$\left[\hat{T}^{HT}(Y) - q_{1-\alpha/2}\hat{\sigma}; \hat{T}^{HT}(Y) + q_{1-\alpha/2}\hat{\sigma}\right]$$

is a **confidence interval for a type 1 error** α, where:

- $q_{1-\alpha/2}$ is the quantile at level $1-\alpha/2$ of a normal distribution $\mathcal{N}(0,1)$;
- $\hat{\sigma} = \sqrt{\hat{V}\left(\hat{T}^{HT}(Y)\right)}$ is the estimated standard deviation of the HT estimator of the total.

In particular:

$$\left[\hat{T}^{HT}(Y) - 1.96\hat{\sigma}; \hat{T}^{HT}(Y) + 1.96\hat{\sigma}\right]$$

is the **confidence interval at a 95 % level** of the total of Y based on the HT estimator.

Simple random sampling

Definition

A simple random sampling without replacement is a sampling design of fixed size n such as:

$$\forall s \in \mathcal{U}, \quad p(s) = \begin{cases} \dfrac{1}{\binom{N}{n}} = \dfrac{n!(N-n)!}{N!} & \text{if } |s| = n \\ 0 & \text{if } |s| \neq n \end{cases}$$

The **sampling rate** is defined as $f = \dfrac{n}{N}$.

Inclusion probabilities

First-order inclusion propabilities

$$\forall i \in U \quad \pi_i = \sum_{s \in \mathcal{U}^{(i)}} \frac{1}{\binom{N}{n}} = \frac{1}{\binom{N}{n}} \times \binom{N-1}{n-1}$$
$$= \frac{n!(N-n)!}{N!} \frac{(N-1)!}{(n-1)!(N-n)!} = \mathbf{\frac{n}{N}}$$

Second-order inclusion propabilities

$$\forall i,j \quad i \neq j \in U \quad \pi_{ij} = \sum_{s \in \mathcal{U}^{(ij)}} \frac{1}{\binom{N}{n}} = \frac{1}{\binom{N}{n}} \times \binom{N-2}{n-2}$$
$$= \frac{n!(N-n)!}{N!} \frac{(N-2)!}{(n-2)!(N-n)!}$$
$$= \mathbf{\frac{n(n-1)}{N(N-1)}}$$

Horvitz-Thompson estimator

When $p(s)$ is a simple random sampling the corresponding Horvitz-Thompson estimator of the total rewrites:

$$\hat{T}^{HT}(Y) = \sum_{i \in s} \frac{y_i}{\pi_i} = \sum_{i \in s} \frac{y_i}{n/N} = \frac{N}{n} \sum_{i \in s} y_i$$

The same applies to the Horvitz-Thompson estimator of the mean:

$$\hat{\bar{Y}}^{HT} = \frac{1}{N} \hat{T}^{HT}(Y) = \frac{1}{N} \frac{N}{n} \sum_{i \in s} y_i = \frac{1}{n} \sum_{i \in s} y_i = \bar{y}$$

To sum up In case of simple random sampling, the Horvitz-Thompson estimator of the mean is the **natural sample mean** $\bar{y} = \frac{1}{n} \sum_{i \in s} y_i$.

Variance of the HT estimator

Going from the general Sen-Yates-Grundy formula (slide 18), one can demonstrate that in the specific case of simple random sampling $V^{SYG}\left(\hat{T}^{HT}(Y)\right)$ **rewrites**:

$$V\left(\hat{T}^{HT}(Y)\right) = N^2(1-f)\frac{S^2}{n}$$

where S^2 is the **empirical variance**:

$$S^2 = \frac{1}{N-1} \sum_{i \in U} (y_i - \bar{y})^2$$

The same applies to the Horvitz-Thompson estimator of the mean:

$$V\left(\hat{\bar{Y}}^{HT}\right) = V\left(\frac{\hat{T}^{HT}(Y)}{N}\right) = \frac{V\left(\hat{T}^{HT}(Y)\right)}{N^2} = (1-f)\frac{S^2}{n}$$

Variance of the HT estimator: proof

$$
\begin{aligned}
V\left(\hat{T}^{HT}(Y)\right) &= -\frac{1}{2}\sum_{i\in U}\sum_{\substack{j\in U\\ j\neq i}}(\pi_{ij}-\pi_i\pi_j)\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2 \\
&= \frac{1}{2}\sum_{\substack{j\in U\\ j\neq i}}\frac{n(N-n)}{N^2(N-1)}\left(\frac{y_i N}{n}-\frac{y_l N}{n}\right)^2 \\
&= \frac{N-n}{n}\frac{1}{2(N-1)}\sum_{i\in U}\sum_{\substack{j\in U\\ j\neq i}}(y_i-y_j)^2 \\
&= N^2\frac{N-n}{nN}\underbrace{\frac{1}{2N(N-1)}\sum_{i\in U}\sum_{\substack{j\in U\\ j\neq i}}(y_i-y_j)^2}_{S^2} \\
&= N^2\frac{N-n}{nN}S^2 = \mathbf{N^2(1-f)\frac{S^2}{n}}
\end{aligned}
$$

Variance estimator of the HT estimator

Both variances are estimated without bias by:

- $\hat{V}\left(\hat{T}^{HT}(Y)\right) = N^2(1-f)\frac{s^2}{n}$
- $\hat{V}\left(\hat{\bar{Y}}^{HT}\right) = (1-f)\frac{s^2}{n}$

where s^2 is the empirical variance **in the sample**:

$$s^2 = \frac{1}{n-1}\sum_{i\in s}(y_i-\bar{y})^2$$

Example: SRS of 20,000 households

Typical household survey in France: 20,000 households out of 27,000,000. The variable of interest Y has a sample mean $\bar{y} = 28{,}000$ and an empirical variance 1.3×10^9.

Question What is the 95 % confidence interval of the mean based on the Horvitz-Thompson estimator?

1. Horvitz-Thompson estimator of the mean:
$$\hat{\bar{Y}}^{HT} = \bar{y} = 28,000$$
2. Variance estimator :
$$\hat{V}\left(\hat{\bar{Y}}^{HT}\right) = \left(1 - \frac{20{,}000}{27{,}000{,}000}\right) \frac{1.3 \times 10^9}{20{,}000} = 64{,}952$$
3. Confidence interval:

$$\begin{aligned} CI_{95\ \%} &= [28{,}000 - 1.96 \times \sqrt{64{,}952}; 28{,}000 + 1.96 \times \sqrt{64{,}952}] \\ &= [27{,}500; 28{,}500] \end{aligned}$$

Context and example

The parameter of interest is quite commonly a **proportion** rather than a total.

Examples Unemployment rate, percentage of part-time jobs, etc.

Any proportion P can be rewritten as the mean of a dichotomous variable.

Example If P is the proportion of unemployed persons, $P = \bar{Y}$ with

$$Y_i = \begin{cases} 1 & \text{if } i \text{ is unemployed} \\ 0 & \text{if } i \text{ is not unemployed} \end{cases}$$

Estimation and precision

Under simple random sampling, the Horvitz-Thompson estimator of the proportion P in the population U is its counterpart p in the sample s:

$$\hat{P}^{HT} = \hat{\bar{Y}}^{HT} = \bar{y} = p$$

The variance of this estimator is given by:

$$V(\hat{P}^{HT}) = (1 - f)\frac{S^2}{n}$$

But note that the empirical variance of the dichotomous variable Y can be rewritten in terms of P:

$$S^2 = \frac{N}{N-1}P(1-P)$$

Estimation and precision

The same applies for the variance estimator of $\hat{P}^{HT}$:

$$\hat{V}(\hat{P}^{HT}) = (1 - f)\frac{s^2}{n}$$

with

$$s^2 = \frac{n}{n-1}p(1-p)$$

Hence

$$\hat{V}(\hat{P}^{HT}) = (1 - f)\frac{1}{n}\frac{n}{n-1}p(1-p) = (1 - f)\frac{p(1-p)}{n-1}$$

Idea This closed-form formula allows for an **easy determination of a sample size for a given precision threshold**.

Example: Determining the sample size

Given the *a priori* knowledge that the share of retired farmers is about 2 %, determine the sample size in order to obtain a coefficient of variation of 5 %.

1. **Assumption** Negligible sampling rate ($f << 1$) hence
$$\hat{V}(\hat{P}^{HT}) \approx \frac{p(1-p)}{n-1}$$
2. $\hat{CV} = \frac{\sqrt{\hat{V}(\hat{P}^{HT})}}{p}$ hence $\hat{V}(\hat{P}^{HT}) = \hat{CV}^2 \times p^2$.
3. Thus the minimal sample size n^* for a given coefficient of variation CV_0 is
$$n^* \approx \frac{p(1-p)}{CV_0^2 \times p^2} = \frac{1-p}{CV_0^2 \times p}$$
4. **Result** $n^* \approx \frac{1-0.02}{0.05^2 \times 0.02} = 19{,}600$

Table: Sample sizes for given proportions

P \ CV_0	**1 %**	**2 %**	**5 %**
0,05	190,000	47,500	7,600
0,10	90,000	22,500	3,600
0,20	40,000	10,000	1,600
0,30	23,333	5,833	933
0,40	15,000	3,750	600
0,50	10,000	2,500	400

Nature of the budget constraint

- If the budget constraint is **strong**, the sample size is entirely determined by the unit cost of a survey:

$$n^* = \frac{C}{c} = \frac{\text{Total cost}}{\text{Unit cost}}$$

Note Other sampling designs allow for reducing the unit cost: cluster sampling, two-stages sampling.

- If the budget constraint is **weaker** (or **if the precision constraint is stronger**), the previous results offer some guidance in the determination of an optimal sample size n^*.

Steps for determining the sample size

1. Fix a precision target (in terms of V_0 or CV_0) and express the desired sample size n^* as a fonction of this target:

$$n^* = \left(\frac{V_0}{S^2} - \frac{1}{N} \right)^{-1}$$

2. Assess the value of S^2:
   - other survey related to the same topic;
   - use of a "proxy" variable;
   - preliminary lightweight survey.

Notes

- For proportions, $S^2 \leq \max_{0 \leq p \leq 1} p(1-p) = 0,25$: a conservative sample size is always computable.
- Be careful when the survey has several variables of interest.

Algorithm based on file ordering

Several algorithms are available in order to draw a sample according to a simple random sampling.

The simplest one relies on the possibility to order the whole sampling frame:

1. For each unit i in the frame, draw a value a_i from a uniform distribution on [0;1].
2. Sort the sampling frame by a_i value.
3. Select the n first units in the sample.

Assets Exact, easy to understand and to implement.

Systematic sampling

The systematic sampling is a **general drawing algorithm** which can yield simple random sampling under **certain conditions**.

1. Randomly order the sampling frame.

Let's define a_i the cumulated probabilities of inclusion of the i first units in the sampling frame:

$$a_i = \sum_{k=1}^{i} \pi_k = \frac{n}{N} \times i$$

2. Draw a value η from a uniform distribution on [0;1].
3. Select all units i such as

$$a_{i-1} \leq \eta + k - 1 < a_i$$

for $k = 1, \ldots, n$.

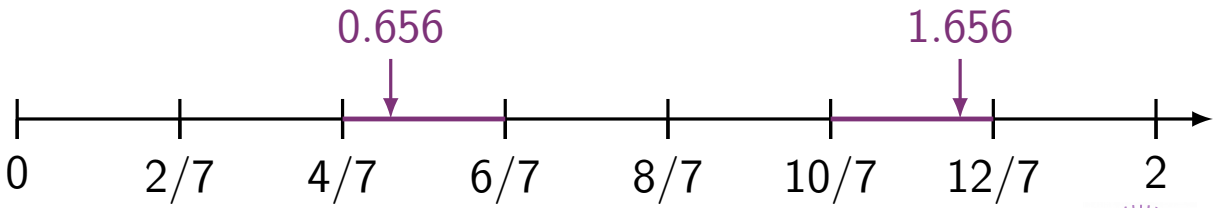
Systematic sampling: example

$N = 7$ and $n = 2$ hence $\forall i \in U\ \pi_i = 2/7$

1. Randomly order the sampling frame

i	G	A	C	E	F	B	D
π_i	2/7	2/7	2/7	2/7	2/7	2/7	2/7
a_i	2/7	4/7	6/7	8/7	10/7	12/7	2

2. Draw a value η from a uniform distribution on [0;1]: $\eta = 0.656$



3. **Result** The drawn sample is $s = \{C, B\}$.

Drawing a sample with SAS

1. Algorithm based on file ordering

```
/*Step 1*/
DATA frame;
  SET frame;
  CALL STREAMINIT(1234);
  a = RAND("UNIFORM");
RUN;

/*Step 2*/
PROC SORT DATA = frame;
  BY a;
RUN;

/*Step 3*/
DATA sample;
  SET frame(OBS = 100);
RUN;
```

Drawing a sample with SAS

2. Systematic sampling

```
/*Step 1*/
DATA frame;
  SET frame;
  CALL STREAMINIT(1234);
  a = RAND("UNIFORM");
RUN;
PROC SORT DATA = frame;
  BY a;
RUN;

/*Step 2 and 3*/
PROC SURVEYSELECT DATA = frame METHOD = SYS N
    = 100 SEED = 1234 OUT = sample;
RUN;
```

Drawing a sample with **R**

1. Algorithm based on file ordering

```
N <- nrow(frame)
n <- 100
set.seed(1234)
a <- runif(N)
sample <- frame[order(a), , drop = FALSE][1:n, ]
```

2. Systematic sampling

```
library(sampling)
set.seed(1234)
pik <- rep(n/N, N)
sample <- frame[
  as.logical(UPrandomsystematic(pik))
  , , drop = FALSE
]
```

Going beyond simple random sampling

Simple random sampling is a simple and robust sampling mechanism which **does not rely on auxiliary information from the sampling frame**.

In most cases some information is **nonetheless available**:

- through contact information: region, city, etc.
- through data source: sex, age, education (census), income, turnover (tax files).

More **advanced sampling designs take advantage of this auxiliary information** in order to improve the sampling mechanism:

- **lower variance at given sample size**: probability proportional to size sampling, stratified sampling;
- **lower unit cost**: cluster sampling, two-stages sampling.

Probability proportional to size sampling

Context and definition

Probability proportional to size sampling is one of the most commonly used unequal probabiliy sampling design.

In this sampling design the **first-order probabilities of inclusion** are defined as **proportional to an auxiliary variable** X:

$$\forall i \in U \quad \pi_i = c \times x_i, \quad c \in \mathbb{R}$$

From

$$\sum_{i \in U} \pi_i = n = \sum_{i \in U} c \times x_i = c \times T(X)$$

follows that

$$c = \frac{n}{T(X)}$$

Remark The auxiliary variable X must be **available for the whole population** U, not only for the selected sample s.

Estimation and properties

The Horvitz-Thompson estimator of the total of the auxiliary variable X is:

$$\hat{T}^{HT}(X) = \sum_{i \in s} \frac{x_i}{\pi_i} = \sum_{i \in s} \frac{x_i}{c \times x_i} = \sum_{i \in s} \frac{1}{c} = \frac{n}{c}$$

As $c = \frac{n}{T(X)}$ one can conclude that

$$\hat{T}^{HT}(X) = n \times \frac{T(X)}{n} = T(X)$$

To sum up Probability proportional to size sampling ensures that **the auxiliary variable is perfectly estimated**.

Assets

This property might seem a little futile since the auxiliary variable is by definition known on the whole population: **it doesn't need to be estimated!**

But **if the interest variable Y is positively correlated with the auxiliary variable X**, probability proportional to size sampling **decreases the sampling variance** compared to simple random sampling.

However, **if X and Y are negatively correlated**, probability proportional to size sampling **increases sampling variance compared to simple random sampling**.

Example: Farm survey

One aims to estimate the total production $T(Y)$ of a population of $N = 6$ farms through a sample of size $n = 2$.

Two sampling designs are considered:

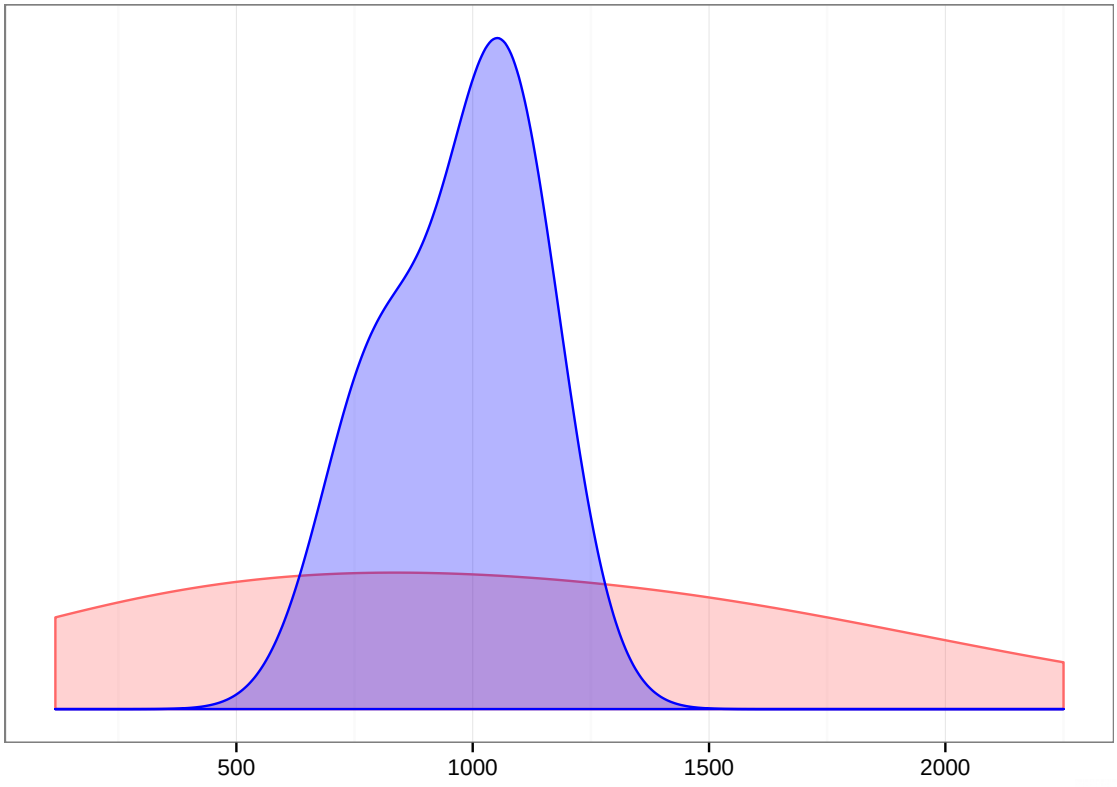
- simple random sampling (SRS);
- probability proportional to size sampling (PPS) with the size of the farms as auxiliary variable.

Farm (i)	Size (X_i)	Production (Y_i)	π_i	
			SRS	PPS
A	100	26	0,33	0,1
B	1000	470	0,33	1
C	125	66	0,33	0,125
D	250	145	0,33	0,25
E	500	280	0,33	0,5
F	25	13	0,33	0,025

Example: Farm survey

Sample	$\widehat{\mathbf{T}}_{\mathbf{SRS}}(\mathbf{Y})$	$\widehat{\mathbf{T}}_{\mathbf{PPS}}(\mathbf{Y})$
$\{A, B\}$	1,488	730
$\{A, C\}$	276	788
$\{A, D\}$	513	840
$\{A, E\}$	918	820
$\{A, F\}$	117	780
$\{B, C\}$	1,608	998
$\{B, D\}$	1,845	1,050
$\{B, E\}$	2,250	1,030
$\{B, F\}$	1,449	990
$\{C, D\}$	633	1,108
$\{C, E\}$	1,038	1,088
$\{C, F\}$	237	1,048
$\{D, E\}$	1,275	1,140
$\{D, F\}$	474	1,100
$\{E, F\}$	879	1,080

Example: Farm survey



Red: SRS – Blue: PPS

Exhaustive units

PPS sampling may yield so-called **exhaustive units**, that is units whose **inclusion probability might be greater than 1**.

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

1. Compute the inclusion probabilities using all units.
2. Until all computed inclusion probabilities are smaller than 1:
   - 2.1 Select the units whose inclusion probability is equal or greater than 1.
   - 2.2 Calculate a new set of inclusion probabilities for all remaining units after removing the exhaustive units.
3. Sample the non-exhaustive units using their calculated inclusion probability.

Exhaustive units: Example

Exhaustive units arise when there is a **significant gap** between one unit and the others in terms of the auxiliary variable X.

Let's imagine one wants to sample 2 units out of a population of size 3 using PPS sampling.

	X_i	Y_i	$\text{PPS}^{(1)}$	$\text{PPS}^{(2)}$
A	300	180	0,2	0,33
B	600	240	0.4	0.67
C	2100	760	1.4	1

Variance estimator for the HT estimator

As the estimator derives from Horvitz-Thompson theoretical frame, the usual variance estimators can be used:

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

$$\hat{V}^{SYG}\left(\hat{T}^{HT}(Y)\right) = -\frac{1}{2}\sum_{i\in s}\sum_{\substack{j\in s\\ j\neq i}}\frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_{ij}}\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2$$

Some issues are nonetheless of **specific relevance** for PPS sampling:

- the Sen-Yates-Grundy condition $\forall i, j \neq i \quad \pi_{ij} - \pi_i\pi_j \leq 0$ is rarely met;
- some π_{ij} might be equal to 0, which can introduce bias in $\hat{V}\left(\hat{T}^{HT}(Y)\right)$.

Variance estimator for the HT estimator

Moreover, the second-order inclusion probability are not always easy to compute (*e.g.* balanced sampling).

In practice, **first-order approximations are commonly used** such as:

$$\hat{V}^{Deville}(\hat{T}^{HT}(Y)) = \frac{n}{n-1}\sum_{i\in s}(1-\pi_i)\left(\frac{y_i}{\pi_i}-\frac{\sum_{j\in s}(1-\pi_j)\frac{y_j}{\pi_j}}{\sum_{j\in s}(1-\pi_j)}\right)^2$$

$\rightarrow$ detailed by **Pascal Ardilly** on wednesday.

Systematic sampling

Systematic sampling may be used in order to draw a sample with probabilities proportional to size.

1. Randomly order the sampling frame.

Given the desired first-order inclusion probabilities π_i (once the exhaustive units have been taken care of), let's define $a_i = \sum_{k=1}^{i} \pi_k$

2. Draw a value η from a uniform distribution on [0;1].
3. Select all units i such as

$$a_{i-1} \leq \eta + k - 1 < a_i$$

for $k = 1, \ldots, n$.

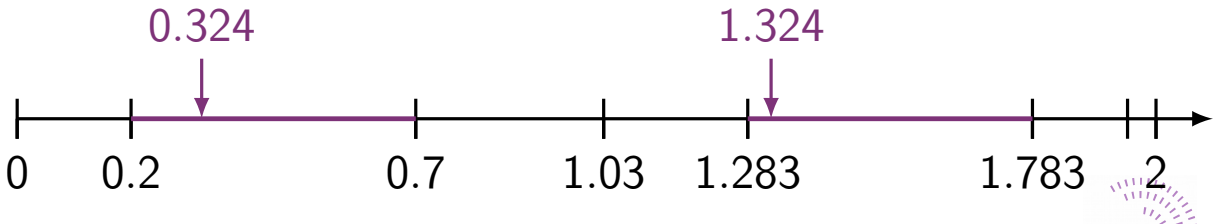
Systematic sampling: example

$N = 7$ and $n = 2$.

1. Randomly order the sampling frame

i	F	A	D	E	G	C	B
π_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

2. Draw a value η from a uniform distribution on [0;1]: $\eta = 0.324$



3. **Result** The drawn sample is $s = \{A, G\}$.

Properties

1. Systematic sampling yields the desired sample size and first-order inclusion probabilities.

2. Very easy and efficient to implement.

3. It may lead to $\pi_{ij} = 0$ for some i and j even after random reordering of the sampling frame.
   $\rightarrow$ Variance estimators of the HT estimator might be biased.

Example $X = \{1, 2, 4, 5, 6\}$ and $n = 2$: even with reordering $\pi_{A,B} = 0$.

To go further Tillé Y. (2006), *Sampling algorithms*, Springer

Drawing a sample with SAS

1. Systematic sampling

```
/*Step 1*/
DATA frame;
  SET frame;
  CALL STREAMINIT(1234);
  a = RAND("UNIFORM");
RUN;
PROC SORT DATA = frame;
  BY a;
RUN;

/*Step 2 and 3*/
PROC SURVEYSELECT DATA = frame METHOD =
    PPS_SYS N = 100 SEED = 1234 OUT = sample;
  SIZE size;
RUN;
```

Drawing a sample with SAS

2. Hanurav-Vijayan (SAS default)

```
PROC SURVEYSELECT DATA = frame METHOD = PPS N
    = 100 SEED = 1234 OUT = sample;
  SIZE size;
RUN;
```

To go further SAS help

Drawing a sample with **R**

1. Systematic sampling

```
library(sampling)
set.seed(1234)
sample <- frame[
  as.logical(UPrandomsystematic(pik))
  , , drop = FALSE
]
```

2. Sampford algorithm (among others)

```
library(sampling)
set.seed(1234)
sample <- frame[
  as.logical(UPsampford(pik))
  , , drop = FALSE
]
```

To go further `sampling` package documentation

Stratified sampling

Principles of a statified sampling design

Notations

Let Y be a quantitative variable defined on U.

With a simple random sampling: when dispersion S^2 of Y increases, the precision of the estimator 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

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

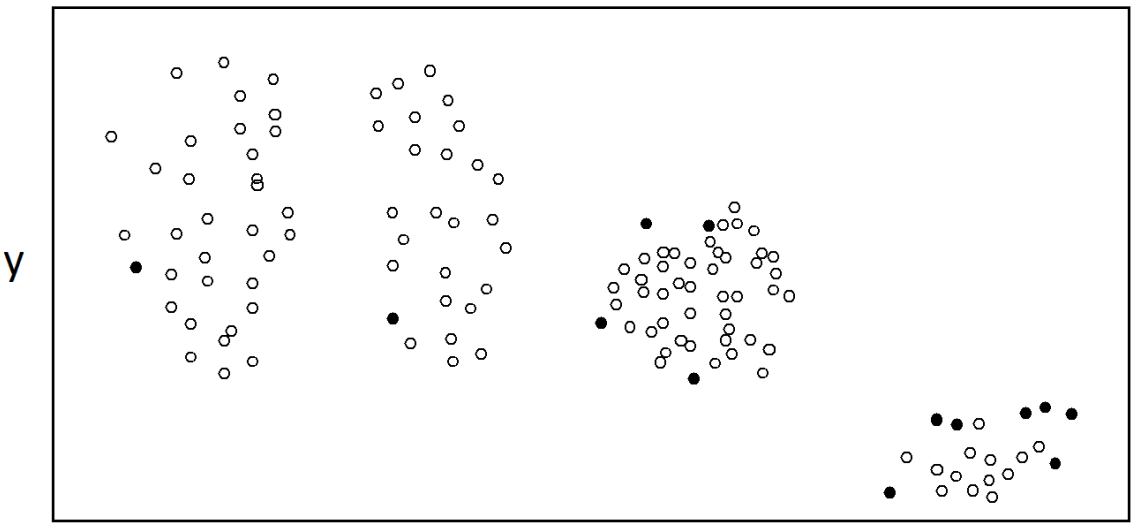
Other goal Stratification allows to set a lower bound for precision in each stratum by controlling the number of units per stratum in the sample.

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

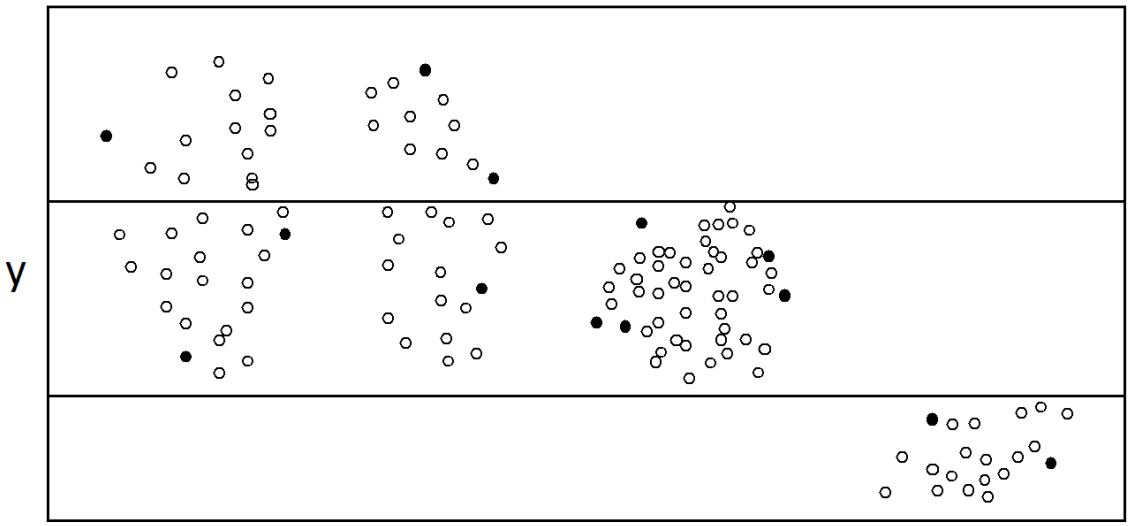
It is assumed that the sizes of the strata N_h are known (usually from the sampling frame).

Representation of a simple random sampling

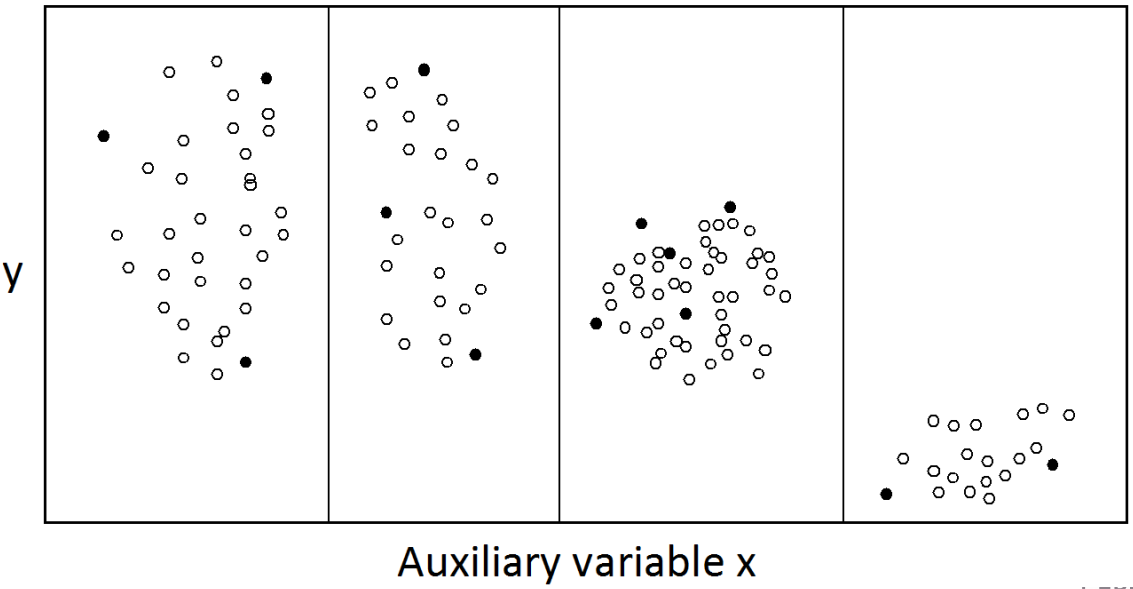
SRS of $n = 13$ units in a population of size $N = 130$ units.



Ideal stratification Using the values of the variable of interest Y.

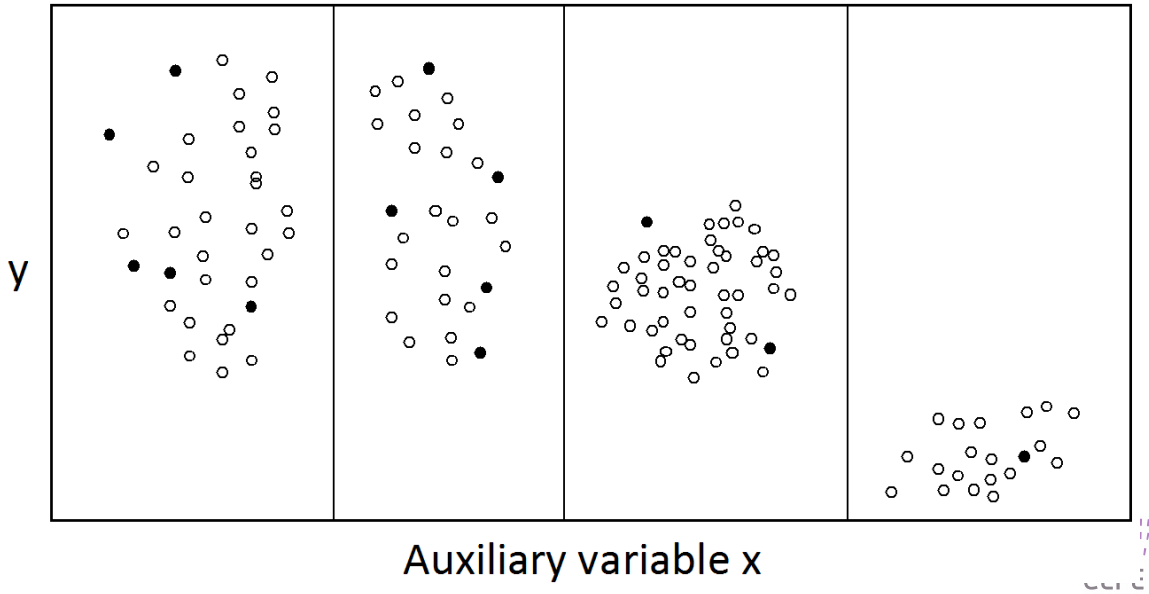


Feasible stratification Using the values of an auxiliary variable X correlated with the variable of interest Y.



Representation of a stratified sampling

Sample allocation How many units should be sampled in each stratum in order to minimize the sampling variance?



Steps to obtain a stratified sample 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:

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

General case

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 $T_h(Y)$:

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

General case

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 notices that

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

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

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

Remarks

1. $\hat{\bar{Y}}_{str} \neq \bar{y}$ The stratified estimator may differ 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

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

   Stratification can yield unequal probability sampling.

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.

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

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

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)

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 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

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

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 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.

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)

Additional material Optimization of strata boundaries based on the number of employees.

Context and strategies

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 depending on 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.

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

Optimal precision at a given cost

The n_h which minimizes 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

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

Optimal precision at a given cost: 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 $n_h = \frac{N_h S_h}{\sqrt{c_h}} \frac{C}{\sum_{k=1}^H \sqrt{c_k} N_k S_k}$

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

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

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 words, the survey should go get information where it is.

- One should **over-represent the strata where the unit cost c_h is the lowest**.

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 2.67 (1.07 for a SRS)

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

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

This explains the previous results.

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.

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**.

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!

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

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):

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

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

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^2_h differ a lot from one stratum to another.

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

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

$$n_h^{mixed} = \alpha n_h^{Neyman} + (1-\alpha) n_h^{prop}$$

with $0 < \alpha < 1$

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):

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

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

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

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. Hence

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

Problem How to choose the α_j...

Exhaustive strata

Using other allocations than the proportional (*e.g.* 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.

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.

Example: Sampling of a business survey

Goals Sample $n = 300$ firms out of a population U of size $N = 1,060$ (e.g. a specific sector).

Auxiliary variable The size of the firm in terms of employees is known. 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 and in each case compute the variance of the estimator.

Example: Sampling of a business survey

Proportional allocation

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

For example $n_5 = 300 \times \dfrac{10}{1,060} \approx 3$

Optimal allocation

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

For example

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

$$\approx 34 > 10$$

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

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

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

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

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

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:

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

Example: Sampling of a business survey

Variance computation: proportional allocation

For example

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

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

Example: Sampling of a business survey

Variance computation: optimal allocation

For example

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

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

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).

Additional material The sampling of the PRODCOM survey.

Stratified and systematic sampling designs

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**.

Stratified and systematic sampling designs

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

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.

SRS and systematic sampling: a trade-off

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**.

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**.

Additional material

Strata optimization: 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 about the optimality of theses boundaries in terms of sampling variance.

Additional material

Strata optimization: Number of employees

There are several methods which determine "optimal" boundaries $b_0, b_1, \ldots b_H$ 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.

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

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.

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

Additional material

Strata optimization: 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.

Additional material

Strata optimization: Number of employees

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.

For a given precision, 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

Additional material

Strata optimization: Number of employees

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

Additional material

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.

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.

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.

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 euros and some legal categories are excluded.

Cluster sampling

Motivation for cluster sampling

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**.

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

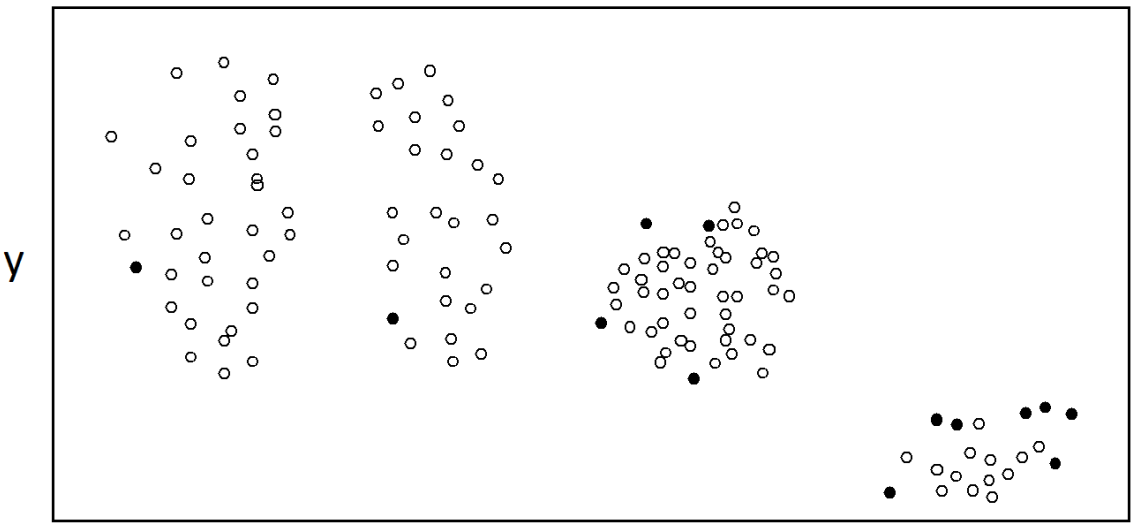
- Let's partition the population U into M parts called "clusters" and denoted $U_1, U_2, \ldots, U_g, \ldots, U_M$ so that, in each cluster g, the spatial dispersion of the units is low.
- Using a sampling design p_{CLUST}, sample m clusters and form the sample of clusters s_{CLUST}.

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

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

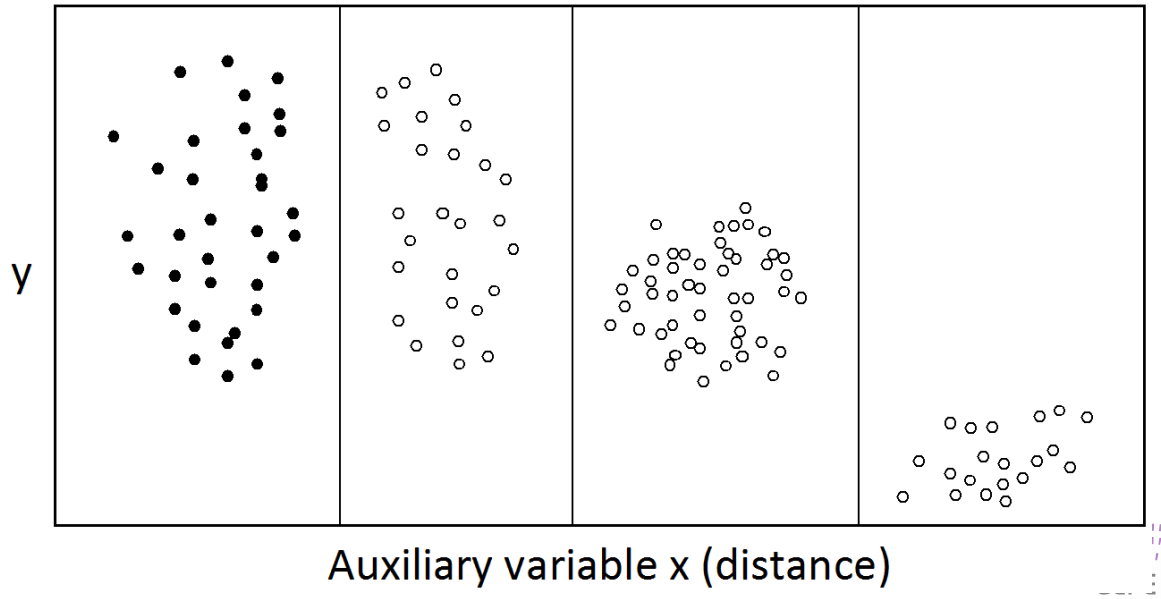
Representation of a simple random sampling

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



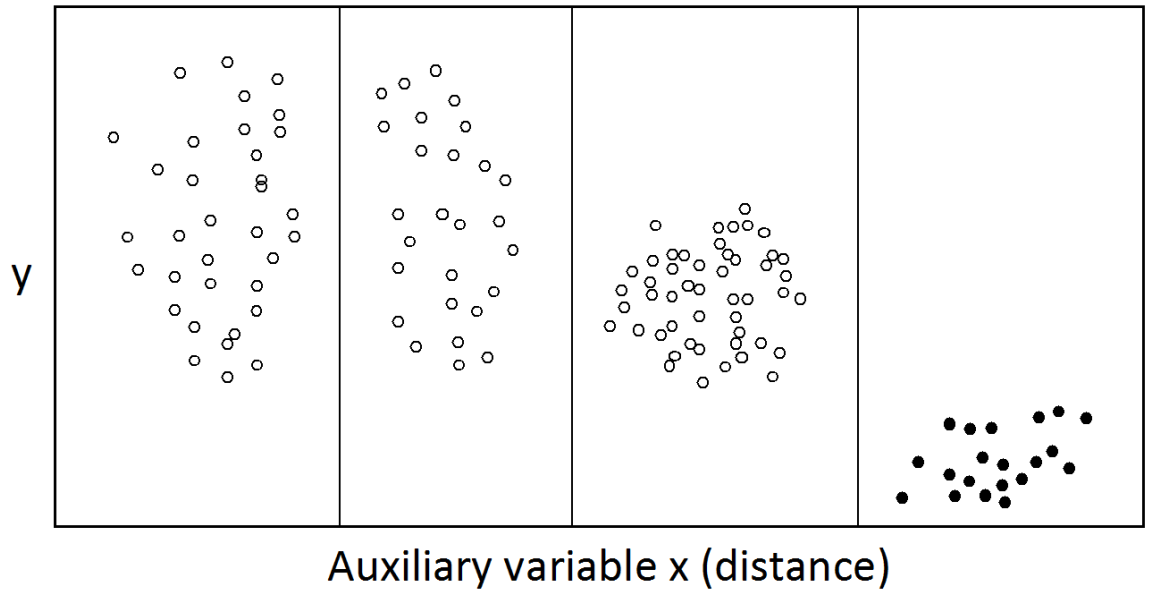
Representation of a cluster sampling

Note In the context of cluster sampling, the auxiliary variable X often represents a distance.



Representation of a cluster sampling

In cluster sampling, all units of the sampled cluster(s) takes part in the survey.



Justification

If the data collection costs are strongly related to the sample drawn (e.g. face-to-face interviews), 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.

Additional material The sampling of the French Labour force survey (LFS).

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:

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

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

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).

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 notices 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 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.

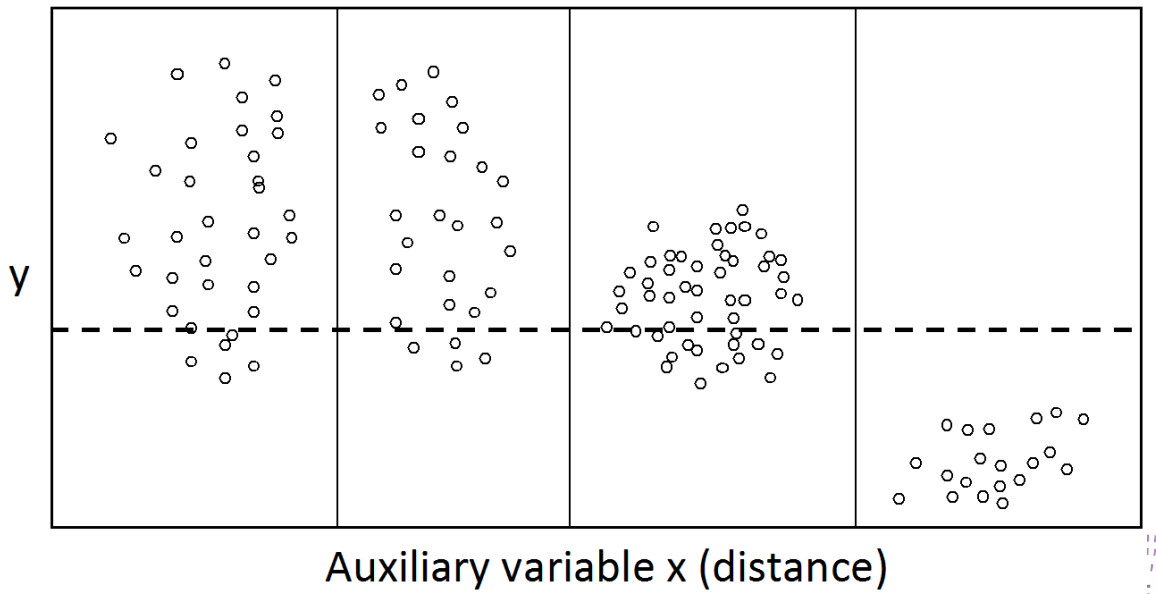
The **within-cluster correlation coefficient** ρ accounts for this so-called "cluster effect":

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

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 effect

Note The values of the interest variable Y are much more concentrated in the fourth cluster than in the two firsts.



Definition and notations

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.

Example: 1 cluster out of 3

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

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

Case 2	**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)

- Case 1: 8.67
- Case 2: 2.67

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

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

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

Remarks

- 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 (in order to have $CV(N) = 0$).

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**.

Remark In practice the estimation of $V(\hat{T}_{SRS}(Y))$ is not straightforward, since the real sampling design is not a simple random sampling.

Sampling size gain

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 related sample sizes for a given global cost C.

Let's assume that in a cluster sampling (simple random sampling of clusters), 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.

Sampling size gain

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:

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

Remarks

- 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**.

Practical recommendations

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

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

2. As long as the variable Y is spatially correlated, **the number of clusters should be as large as possible**.
3. Clusters should have the **same size**.

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**.

Constraints

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**.

Definition of the clusters

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).

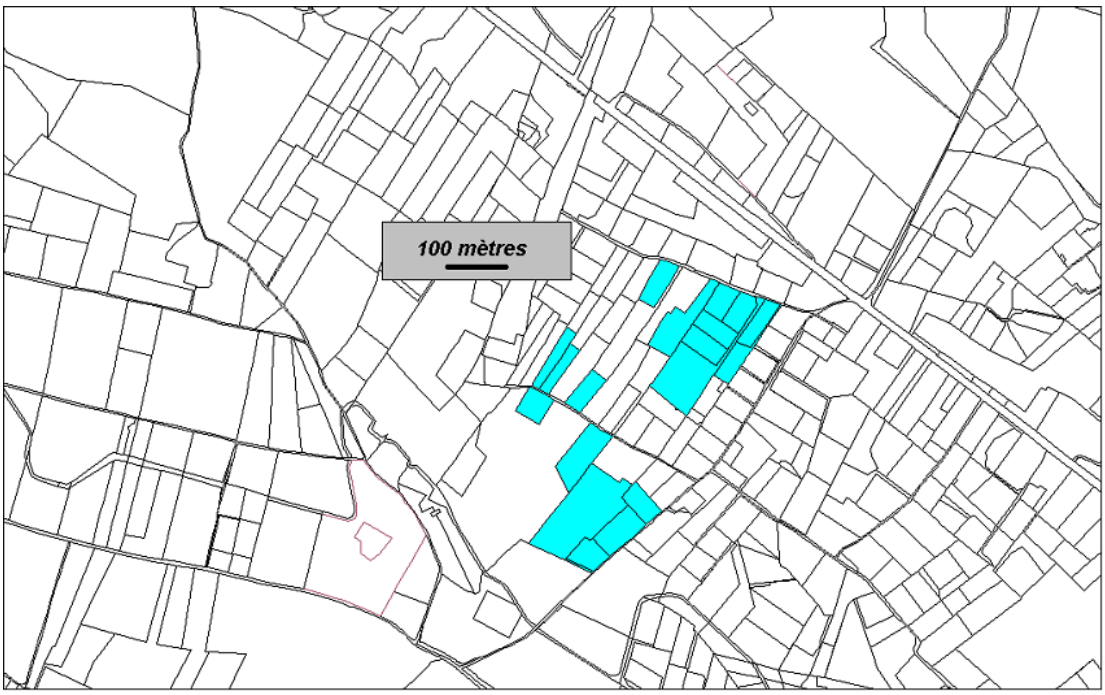
In collective housing, **the dwellings located on the same floor** belong to the same cluster.

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

Definition of the clusters

Definition of the clusters

Definition of the clusters



Sampling design

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!

Two-stages sampling

Principles and notations

Looking back at stratified and cluster sampling

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.

Definition

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

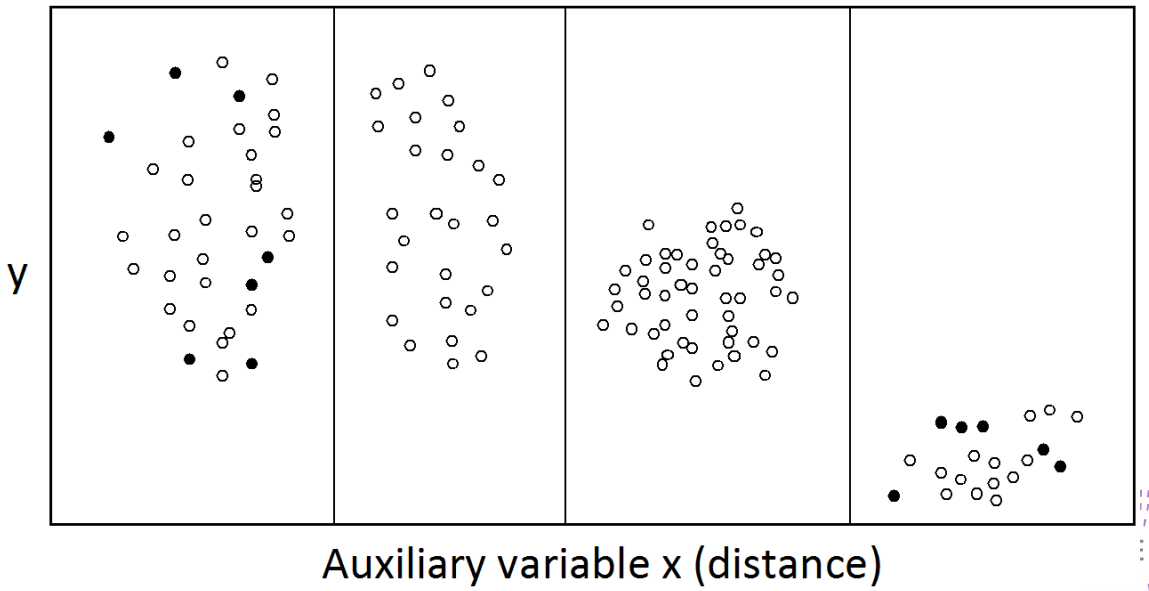
- First m PSUs are sampled out of M using a sampling design p_{PSU} and form the sample of PSUs s_{PSU}.
- Then in each **sampled** PSU g, n_g SSUs are sampled out of N_g using a sampling design p_g and form the sample of SSUs s_g.

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

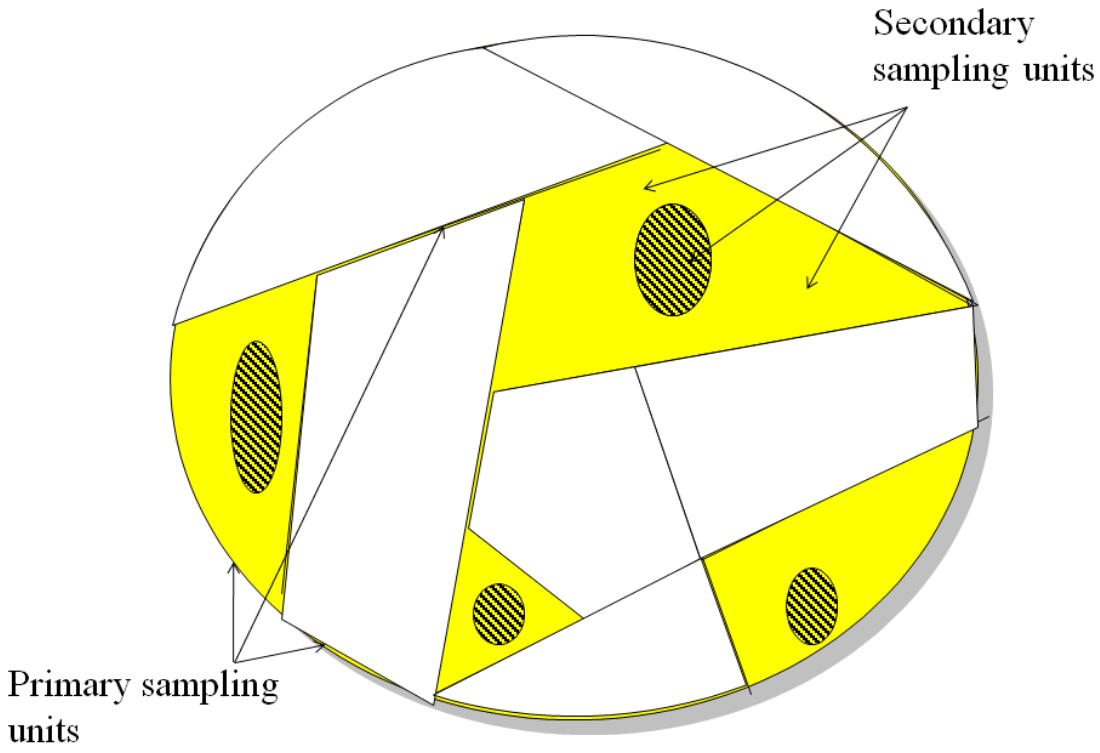
$$s = \bigcup_{g \in s_{PSU}} s_g$$

Representation of a two-stages sampling

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.



Representation of a two-stages sampling



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 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 stage, 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 stage is a census**.

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 stage, 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 stage is a census**.
- The SSUs are sampled in each PSU g according to sampling design p_g. Then

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

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 **trade-off** 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**.

Generalization

It is possible to define **multi-stages 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 topic of the survey (systematic sampling on a sorted file).

Horvitz-Thompson estimator

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.

Unbiasedness of the HT estimator: proof

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

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

Variance of the HT 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

$$V_{SSU} = \sum_{g \in s_{PSU}} \frac{V_g}{\pi_g^2} \quad \text{with} \quad V_g = \sum_{i \in s_g} \sum_{j \in s_g} (\pi_{ij|g} - \pi_{i|g}\pi_{j|g}) \frac{y_i}{\pi_{i|g}} \frac{y_j}{\pi_{j|g}}$$

Variance of the HT estimator: proof

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

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

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

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

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

Variance estimator of the HT 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

$$\hat{V}\left(\hat{T}_{TS}(Y)\right) = \underbrace{\sum_{g\in s_{PSU}} \sum_{h\in s_{PSU}} \frac{\pi_{gh}-\pi_g\pi_h}{\pi_{gh}} \frac{\hat{T}_g}{\pi_g} \frac{\hat{T}_h}{\pi_h}}_{(a)} + \underbrace{\sum_{g\in s_{PSU}} \frac{\hat{V}_g}{\pi_g^2}}_{(b)}$$

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}

Definition and Horvitz-Thompson estimator

First stage m PSUs are sampled among M by simple random sampling.

Second stage Within each sampled PSU U_g, n_g SSUs are sampled among N_g by simple random sampling

Horvitz-Thompson estimator

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

Variance of the HT 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}

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

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

To sum up "Good" PSUs should be quite numerous, with a large heterogeneity within for Y.

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:

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

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

Thus

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

To sum up A two-stages sampling is in general **less efficient than a simple random sampling**.

Variance estimator for the HT 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

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

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

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

Example: Two-stages *versus* cluster sampling

1. Cluster sampling: SRS of 1 cluster among 3

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

Sampling variance 6

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

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

Sampling variance 3.83

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.

2. **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.

3. **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.

Remarks

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 with probabilities proportional to size** and where **the number of SSUs in each PSU is constant**: **self-weighted sample**.

Definition

First stage m PSUs are sampled among M according to a sampling with **probability proportional 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 a **self-weighted sample of fixed size**.

Horvitz-Thompson estimator and its variance

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

As the sample is equally weighted, **the Horvitz-Thompson estimators are the same as in simple random sampling**.

Variance of the Horvitz-Thompson estimator

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

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.

Less efficient than simple random sampling for a given sample size 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 a self-weighted sample**.

Additional material

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 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.

Additional material

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.

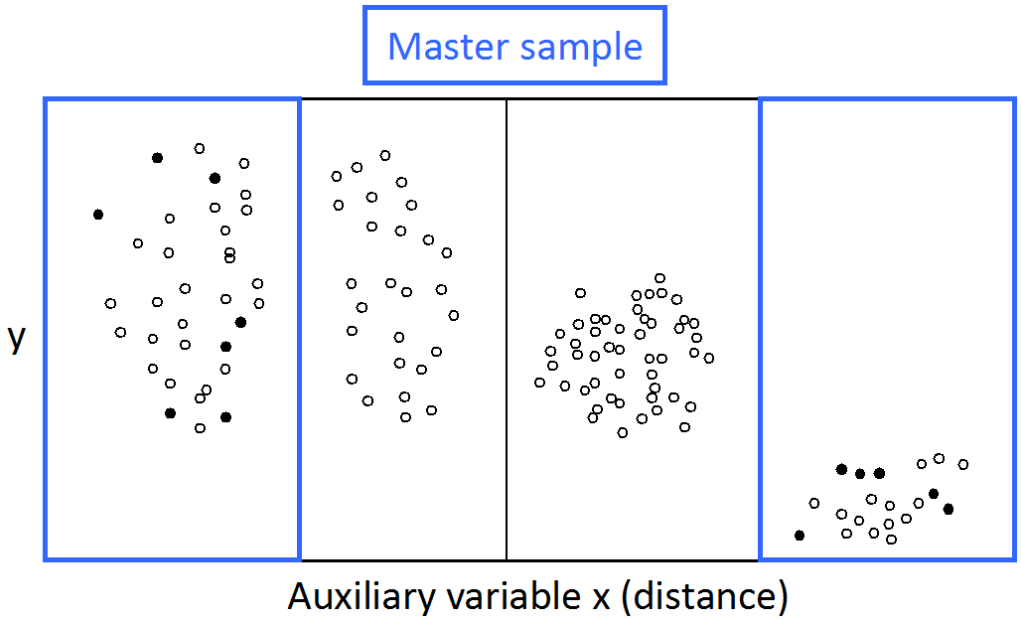
Yet it is the most efficient way to organize face-to-face interviews (household surveys) compared to simple random sampling.

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

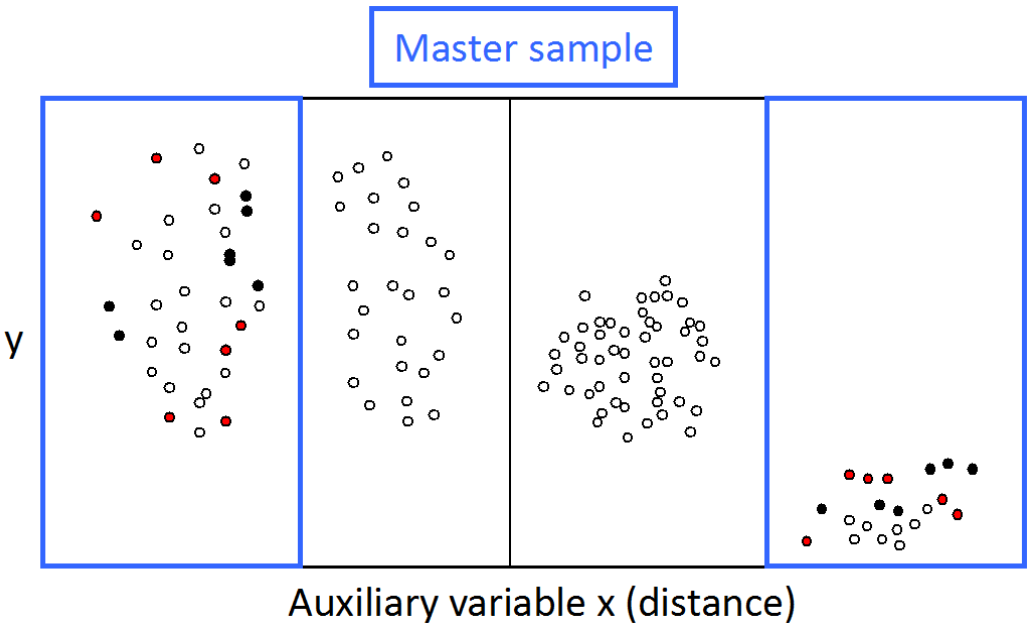
First sample after the census



Principle of a master sample

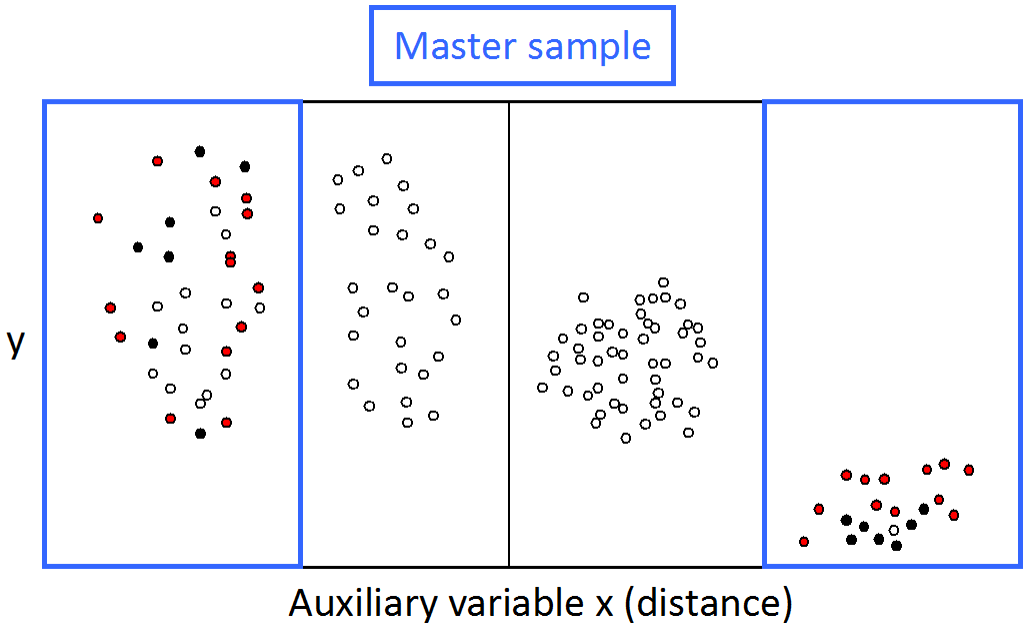
Second sample after the census

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



Third sample after the census

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



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 training 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.

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**.

The sampling of the Adult education survey

The Adult education survey (AES) was conducted in France in 2012 and 2016 in application of european regulations. It aims at measuring the **participation in training of adults aged 25-64**.

In 2012, its sampling design was typical of a French individual survey:

1. Sampling of households in the master sample:
   - 1.1 Sampling of primary sampling units in 2009: balanced sampling.
   - 1.2 Sampling of households within the primary sampling units: systematic sampling on a sorted file.
2. Sampling of one individual among the household: simple random sampling.

Additional material
The sampling of the Adult education survey

As this survey relies on face-to-face interviews, the **travel** represents the most part of its cost:

- 1 hour for locating the dwelling and getting in touch
- 1 hour for the travel itself
- 1/2 hour for the interview itself

In this context, one can distinguish **two types of costs**:

- c_1: the fixed cost per household, 2 hours
- c_2: the variable cost per interview, 1/2 hour

In 2016 we implemented a **new methodology**, where **two individuals are sampled from the same household**.

Additional material
The sampling of the Adult education survey

The gains associated with this strategy should not be **overstated**: as the individual belonging to the same household are often alike, it introduces **cluster effect**.

Intuition In order to maintain the precision of a one-individual-per-household strategy, one should **increase the overall sample size**.

Technically, **two strata of households** are defined:

- the households with only one individual 25-64
- the households with two or more individuals 25-64

Additional material
The sampling of the Adult education survey

Questions

- what should be the total sample size n in order to achieve the same precision than in 2012?
- how to allocate between the two strata in order to maximize the gains associated with this strategy?

Assumptions

- negligible sampling rates
- same design effect in 2012 and in 2016
- same empirical variance for the variable of interest in the two strata

Additional material
The sampling of the Adult education survey

This yields the minimization problem:

$$\begin{cases} \min_{n_1,n_2}(c_1+c_2)n_1 + \dfrac{(c_1+2c_2)}{2}n_2 \\ \text{s.c.} \left(\dfrac{N_1}{N_1+N_2}\right)^2 \dfrac{1}{n_1} + \left(\dfrac{N_2}{N_1+N_2}\right)^2 \dfrac{1+\rho}{n_2} = \dfrac{1}{n_{2012}} \end{cases}$$

where ρ is the intra-cluster correlation coefficient.

The solution of this problem is:

$$\begin{cases} n_1 = n_{2012} \times \left[(1-q_2)^2 + q_2(1-q_2)\sqrt{\dfrac{1+\rho}{2\tau}}\right] \\ n_2 = n_{2012} \times \left[q_2(1-q_2)\sqrt{2\tau(1+q_2)} + q_2^2(1+\rho)\right] \end{cases}$$

with $q_2 = \dfrac{N_2}{N_1+N_2}$ and $\tau = \dfrac{c_1+c_2}{c_1+2c_2}$

The sampling of the Adult education survey

In order to calculate n_1 and n_2, ρ had to be estimated.

The **Labour force s**urvey was used for that purpose:

- all persons 15 + of the household are surveyed
- questions are asked about training during the last three months (proxy)

The final value for ρ was set to 0.20.

In order to keep the sampling variance at the level of 2012, about 14,000 individuals were interviewed instead of 11,000, but for a **lower global cost**.

Other advantage It allows some study about the proximity in terms of training of the members of the same household.