# Introduction à la théorie des sondages
# Cours 6 : Révisions

Thomas Merly-Alpa
thomas.merly-alpa@insee.fr

http://nc233.com/teaching

INSEE, Département des méthodes statistiques

12 mars 2018

Chapitre 1

# Retour sur l'ensemble du cours

Principe de l'estimation
Sondage aléatoire simple
Stratification : application au sondage aléatoire simple
Stratification : choix des allocations
La non-réponse et sa correction
Les méthodes de redressement

# Sommaire

- Principe de l'estimation
- Sondage aléatoire simple
- Stratification : application au sondage aléatoire simple
- Stratification : choix des allocations
- La non-réponse et sa correction
- Les méthodes de redressement

Partie 1

# Principe de l'estimation

# Objectifs et méthode

**Objectif** Estimer la valeur d'une statistique $\theta$ sur une population $\mathcal{U}$.

En pratique, il est inenvisageable d'interroger les $N$ individus de $\mathcal{U}$ (**recensement**) :

- très cher ;
- nombre limité de questions ;
- diminution rapide des taux de réponse (biais de non-réponse).

**Solution** **Sondage probabiliste** : seul un échantillon $s$ de taille $n$ **tiré aléatoirement** dans une base de sondage est enquêté.

**Base de sondage** Fichier comportant les **informations de contact** de toutes les unités de la population $\mathcal{U}$ et éventuellement des **informations auxiliaires**.

# Plan de sondage

On note $\mathcal{S}$ l'ensemble des parties de $\mathcal{U}$. On appelle alors **plan de sondage** la loi de probabilité $p$ définie sur $\mathcal{S}$ telle que :

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

**Exemple** $\mathcal{U} = \{a, b, c\}$ hence

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

$p(\{a\}) = 0$ $\quad p(\{a, b\}) = 0{,}5$ $\quad p(\{a, b, c\}) = 0$
$p(\{b\}) = 0$ $\quad p(\{a, c\}) = 0{,}25$ $\quad p(\emptyset) = 0$
$p(\{c\}) = 0$ $\quad p(\{b, c\}) = 0{,}25$

# Estimation sans biais : probabilités d'inclusion

**Biais de l'estimateur « naturel »** L'estimateur « naturel » (ou *plugin*) n'est pas nécessairement sans biais.

**Exemple** La moyenne arithmétique simple dans l'échantillon comme estimateur de la moyenne dans la population.

Pour construire un **estimateur sans biais sous le plan de sondage**, on introduit les quantités suivantes :

- **probabilités d'inclusion simple** :

$$\forall k \in \mathcal{U}, \quad \pi_k = \sum_{s \in \mathcal{S}} p(s) \mathbb{1}_{k \in s}$$

- **probabilités d'inclusion double** :

$$\forall k, l \in \mathcal{U}, \quad \pi_{k,l} = \sum_{s \in \mathcal{S}} p(s) \mathbb{1}_{k \in s} \mathbb{1}_{l \in s}$$

# Estimation sans biais : estimateur d'Horvitz-Thompson

**Définition** On appelle **estimateur d'Horvitz-Thompson** :

- du total de la variable $Y$ : $\hat{T}_{HT}(Y) = \sum_{k \in s} \frac{y_k}{\pi_k}$
- de la moyenne de la variable $Y$ : $\hat{\bar{Y}}_{HT} = \frac{1}{N} \sum_{k \in s} \frac{y_k}{\pi_k}$

**Propriété** Si $\forall k \in \mathcal{U},\ \pi_k > 0$ alors l'estimateur d'Horvitz-Thompson est **sans biais** pour le total et la moyenne.

# Variance de l'estimateur d'Horvitz-Thompson

**Définition** La variance de l'estimateur $\hat{\theta}$ (sans biais) est définie par :

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

Dans le cas de l'estimateur d'Horvitz-Thompson du total, cette quantité se réécrit :

$$V\left(\hat{T}_{HT}(Y)\right) = \sum_{k \in \mathcal{U}} \sum_{l \in \mathcal{U}} (\pi_{kl} - \pi_k \pi_l) \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}$$

Et quand le plan de sondage est de taille fixe (Sen-Yates-Grundy) :

$$V_{SYG}\left(\hat{T}_{HT}(Y)\right) = -\frac{1}{2} \sum_{k \in \mathcal{U}} \sum_{\substack{l \in \mathcal{U} \\ l \neq k}} (\pi_{kl} - \pi_k \pi_l) \left( \frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2$$

Partie 2

# Sondage aléatoire simple

# Définition et probabilités d'inclusion

Un sondage aléatoire simple (SAS) sans remise de taille fixe $n$ est un plan de sondage tel que :

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

Le **taux de sondage** $f$ est défini par $f = \dfrac{n}{N}$.

## Probabilités d'inclusion

- probabilités d'inclusion simples : $\forall k \in \mathcal{U},\ \pi_k = \dfrac{n}{N}$ ;
- probabilités d'inclusion doubles : $\forall k, l \in \mathcal{U},\ \pi_{k,l} = \dfrac{n(n-1)}{N(N-1)}$.

# Estimateur d'Horvitz-Thompson

On spécifie dans le cas particulier du sondage aléatoire simple la valeur de l'estimateur d'Horvitz-Thompson :

- de la moyenne de la variable $Y$ :

$$\hat{\bar{Y}}_{SAS} = \frac{1}{N}\sum_{k\in s}\frac{y_k}{\pi_k} = \frac{1}{N}\sum_{k\in s}\frac{y_k}{n/N} = \frac{1}{n}\sum_{k\in s} y_k = \bar{y}$$

- du total de la variable $Y$ : $\hat{T}_{SAS}(Y) = N\bar{y}$

**À retenir** Dans le cas du sondage aléatoire simple, l'estimateur d'Horvitz-Thompson de la moyenne **coïncide avec la moyenne arithmétique simple**.

# Variance de l'estimateur d'Horvitz-Thompson

**Population** On spécifie également les formules de variance des estimateurs dans le cas du sondage aléatoire simple :

$$V(\hat{\bar{Y}}_{SAS}) = \left(1 - \frac{n}{N}\right) \frac{S_Y^2}{n} \quad \text{et} \quad V(\hat{T}_{SAS}(Y)) = N^2 \left(1 - \frac{n}{N}\right) \frac{S_Y^2}{n}$$

avec $S_Y^2$ la variance empirique de $Y$ dans la population $\mathcal{U}$ :

$$S_Y^2 = \frac{1}{N-1} \sum_{k \in \mathcal{U}} (y_k - \bar{Y})^2$$

**Échantillon** On estime ces quantités à partir de l'échantillon $s$ par :

$$\hat{V}(\hat{\bar{Y}}_{SAS}) = \left(1 - \frac{n}{N}\right) \frac{s_Y^2}{n} \quad \text{et} \quad \hat{V}(\hat{T}_{SAS}(Y)) = N^2 \left(1 - \frac{n}{N}\right) \frac{s_Y^2}{n}$$

avec $s_Y^2$ la variance empirique de $Y$ dans l'échantillon $s$ :

$$s_Y^2 = \frac{1}{n-1} \sum_{k \in s} (y_k - \bar{y})^2$$

# Cas particulier d'une proportion

Une proportion $P$ est un **cas particulier de moyenne** (où la variable $Y$ est une indicatrice), aussi :

$$\hat{P}_{SAS} = p$$

où $p$ est la proportion calculée dans l'échantillon $s$.

Par ailleurs, on peut montrer que la **variance empirique** dans l'échantillon associée à une proportion $p$ est :

$$s_p^2 = \frac{n}{n-1} p(1-p) \quad \text{et ainsi} \quad \hat{V}(\hat{P}_{SAS}) = \left(1 - \frac{n}{N}\right) \frac{p(1-p)}{n-1}$$

Sous l'hypothèse que le taux de sondage $f = \frac{n}{N}$ est négligeable, la taille d'échantillon $n^*$ **pour obtenir le coefficient de variation** $CV_0$ est alors :

$$n^* \approx \frac{1 - p_0}{p_0 \times CV_0^2}$$

Partie 3

# Stratification : application au sondage aléatoire simple

# Principe de la stratification

**Objectif** **Améliorer la précision** de l'estimateur d'Horvitz-Thompson obtenu par sondage aléatoire simple.

**Méthode** Exploiter l'information auxiliaire présente dans la base de sondage **au moment du tirage**.

En pratique :

- à l'aide de **variables auxiliaires**, on définit des **strates** au sein desquelles la variable $Y$ est **homogène** ;
- on mène un **tirage indépendant** au sein de chaque strate.

**Résultat** L'échantillon $s$ obtenu contient nécessairement des **observations provenant de chaque strate**, ce qui **stabilise l'estimateur** (*i.e* diminue sa variance).

# Sondage aléatoire simple stratifié

On définit ainsi le sondage aléatoire simple stratifié de taille $n$ :

1. Découpage de la population $\mathcal{U}$ en $H$ strates (la strate $h$ est de taille $N_h$) à l'aide des variables auxiliaires de la base de sondage.
2. Sondage aléatoire simple de $n_h$ unités au sein de chaque strate $h$ de telle sorte que $\sum_{h=1}^{H} n_h = n$.

On note $f_h = \frac{n_h}{N_h}$ le taux de sondage dans la strate $h$ et on a les probabilités d'inclusion :

- $\forall k \in s_h, \quad \pi_k = \dfrac{n_h}{N_h}$
- $\forall k, l \in s_h, \quad \pi_{kl} = \dfrac{n_h(n_h-1)}{N_h(N_h-1)}$

# Estimateur d'Horvitz-Thompson

Les tirages étant indépendants d'une strate à une autre, les estimateurs d'Horvitz-Thompson se réécrivent simplement en fonction des estimateurs au sein de chaque strate :

- pour le total de $Y$ : $\hat{T}_{SAS-str}(Y) = \sum_{h=1}^{H} N_h \bar{y}_h$
- pour la moyenne de $Y$ : $\hat{\bar{Y}}_{SAS-str} = \frac{1}{N} \sum_{h=1}^{H} N_h \bar{y}_h$

# Variance de l'estimateur d'Horvitz-Thompson

Les estimateurs de variance des estimateurs d'Horvitz-Thompson font eux aussi intervenir uniquement des quantités calculées au sein de chaque strate :

- pour le total de $Y$ : $\hat{V}(\hat{T}_{SAS-str}(Y)) = \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s_h^2}{n_h}$
- pour la moyenne de $Y$ :
$$\hat{V}(\hat{\bar{Y}}_{SAS-str}) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s_h^2}{n_h}$$

avec $s_h^2$ la variance empirique de $Y$ au sein de la strate $h$ dans l'échantillon $s$ :

$$s_h^2 = \frac{1}{n_h - 1} \sum_{k \in s_h} (y_k - \bar{y}_h)^2$$

Partie 4

# Stratification : choix des allocations

# Enjeu du choix des allocations

Dans un sondage stratifié, le nombre d'unités à allouer pour chaque strate est déterminé par le concepteur de l'enquête.

**Question** Quel est l'impact de l'**allocation de l'échantillon** entre les strates (*i.e.* des valeurs $n_h$) sur la **précision de l'estimateur** ?

Deux situations :

- on souhaite une amélioration de la précision de toutes les variables par rapport au sondage aléatoire simple $\rightarrow$ **allocation proportionnelle**
- on souhaite la meilleure précision possible pour une variable, quitte à détériorer la précision pour d'autres $\rightarrow$ **allocation de Neyman**

# Allocation proportionnelle

**Définition** L'allocation proportionnelle est définie par :

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

**Intuition** La proportion d'unités de la strate $h$ dans l'échantillon correspond à la proportion d'unités de la strate $h$ dans la population :

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

**Propriété**

$$V(\hat{T}_{SAS-str,prop}(Y)) \leq V(\hat{T}_{SAS}(Y))$$

# Allocation de Neyman

**Définition** L'allocation de Neyman est définie par :

$$\forall h = 1, \ldots, H, \quad n_h = n \times \frac{N_h S_{Y,h}}{\sum_{h'=1}^{H} N_{h'} S_{Y,h'}}$$

**Intuition** L'allocation est d'autant plus forte que la variance de la variable d'intérêt $Y$ dans la strate $h$ $S^2_{Y,h}$ est importante : **on va chercher l'information là où elle est**.

**Propriété** L'allocation de Neyman est **optimale** au sens où elle conduit à des estimateurs d'Horvitz-Thompson de $Y$ dont la **variance est minimale**.

# Strate exhaustive

Une allocation de Neyman peut conduire à devoir sélectionner **davantage d'unités dans l'échantillon que n'en compte la strate dans la population** :

$$n_h^{Neyman} > N_h$$

Dans ce cas :

1. On intègre à l'échantillon les $N_h$ observations de la strate $h$ (pas de tirage au sort) : $h$ est une **strate exhaustive**.
2. On poursuit le mécanisme d'allocation en excluant totalement la strate $h$ du problème : on se ramène à un sondage de $n - N_h$ unités parmi $N - N_h$ stratifié en $H - 1$ strates.

Partie 5

# La non-réponse et sa correction

# Deux types de non-réponse

On distingue classiquement deux types de non-réponse :

- **non-réponse totale** : l'unité n'a répondu à aucune question de l'enquête.

  **Exemple** Impossible à joindre, refus lors de la prise de contact.

- **non-réponse partielle** : l'unité a globalement répondu à l'enquête mais pas à certaines questions spécifiques.

  **Exemple** Questions difficiles à comprendre ou gênantes.

**Conséquences de la non-réponse** En règle générale **en présence de non-réponse les estimateurs d'Horvitz-Thompson sont biaisés**.

# Trois mécanismes de non-réponse

On distingue classiquement trois mécanismes de non-réponse :

- **non-réponse MCAR** (*Missing completely at random*) : non-réponse et variable d'intérêt $Y$ sont complètement indépendantes ;
- **non-réponse MAR** (*Missing at random*) : non-réponse et variables d'intérêt $Y$ sont indépendantes conditionnellement à certaines variables auxiliaires disponibles ;
- **non-réponse MNAR** (*Missing not at random*) : même en conditionnant par toutes les variables auxiliaires disponibles, non-réponse et variable d'intérêt $Y$ ne sont pas indépendantes.

Il est possible de **corriger le biais associé à des non-réponses régies par des mécanismes MCAR et MAR**.

# Deux méthodes de correction de la non-réponse

On distingue classiquement deux ensembles de méthodes de correction de la non-réponse :

- **repondération** : répartir le poids de sondage des non-répondants (au sens de la non-réponse totale) sur les répondants.

  $\rightarrow$ convient à la correction de la **non-réponse totale**.

- **imputation** : attribuer aux non-répondants une valeur « crédible », de façon déterministe (*cold-deck*, moyenne) ou aléatoire (*hot-deck*).

  $\rightarrow$ convient surtout à la correction de la **non-réponse partielle**.

# Principe de la correction de la non-réponse

**Pour une non-réponse MCAR** Appliquer la méthode de correction de la non-réponse (repondération ou imputation) globalement.

**Pour une non-réponse MAR** Appliquer la méthode de correction de la non-réponse **au sein de classes de correction de la non-réponse** définies à l'aide des **variables auxiliaires**.

On parle de **groupes de réponse homogènes** pour les méthodes de repondération et de **classes d'imputation** pour les méthodes d'imputation.

Partie 6

# Les méthodes de redressement

# Principe des redressements

**Objectif** **Améliorer la précision** de l'estimateur d'Horvitz-Thompson.

**Méthode** Exploiter l'information auxiliaire dont on dispose **au moment de l'estimation**.

En pratique :

- à partir de l'échantillon $s$ tiré, calculer l'estimateur d'Horvitz-Thompson d'une ou plusieurs variables auxiliaires ;
- redresser l'estimateur d'Horvitz-Thompson en comparant ces estimations aux vraies valeurs dans la population.

**Résultat** L'estimateur redressé garantit une **estimation parfaite des variables auxiliaires** et est donc de ce fait **plus stable**.

# Redressement par le ratio

**Définition** Pour une variable auxiliaire $X$ quantitative :

$$\hat{T}_{ratio}(Y) = \hat{T}_{HT}(Y) \times \frac{T(X)}{\hat{T}_{HT}(X)}$$

**Variance dans le cas du SAS**

$$\hat{V}(\hat{T}_{ratio}(Y)) \approx N^2 \left(1 - \frac{n}{N}\right) \frac{s_Y^2 + \hat{R}^2 s_X^2 - 2\hat{R} s_{X,Y}}{n}$$

avec $\hat{R} = \dfrac{\hat{T}_{HT}(Y)}{\hat{T}_{HT}(X)}$ et $s_{X,Y}$ la covariance empirique de $X$ et $Y$ dans l'échantillon.

**Intuition** Plus la corrélation entre $X$ et $Y$ est forte, plus la variance de l'estimateur redressé est faible.

# Redressement par post-stratification

**Définition** Pour une variable auxiliaire $X$ qualitative :

$$\hat{T}_{post}(Y) = \sum_{h=1}^{H} \hat{T}_{h,HT}(Y) \frac{N_h}{\hat{N}_{h,HT}}$$

**Variance dans le cas du SAS**

$$V(\hat{T}_{post}(Y)) \approx \underbrace{N^2 \frac{1-f}{n} \sum_{h=1}^{H} \frac{N_h}{N} S_h^2}_{\text{Variance d'un SAS stratifié avec alloc. proportionnelle}} + \underbrace{N^2 \frac{1-f}{n^2} \sum_{h=1}^{H} \frac{N-N_h}{N} S_h^2}_{\text{Variance supplémentaire due à la post-stratification}}$$

**Intuition** La variance de l'estimateur post-stratifié est toujours supérieure à celle d'un SAS stratifié avec allocation proportionnelle, mais la différence est négligeable si $n$ est grand.

# Généralisation : calage sur marges

Le calage sur marges est une **généralisation des autres méthodes de redressement** qui permet notamment de tenir compte de **plusieurs variables auxiliaires** simultanément.

L'objectif de l'algorithme de calage sur marge est de trouver le vecteur de poids $w_k$ **le plus proche possible des poids de sondage** $d_k$ et tel que les totaux des variables de calage soient **parfaitement estimés**.

C'est en pratique cette méthode de redressement qui est **la plus employée**, qui est mise en œuvre implicitement dès lors qu'on utilise le **poids calé à la place du poids de sondage** pour mener à bien des estimations.

Chapitre 2

# Introduction au sondage à deux degrés

Partie 1

# Principe du sondage à deux degrés

# Le contexte : l'importance des enquêtes en face-à-face

À l'Insee, la plupart des enquêtes auprès des ménages sont effectuées en face-à-face.

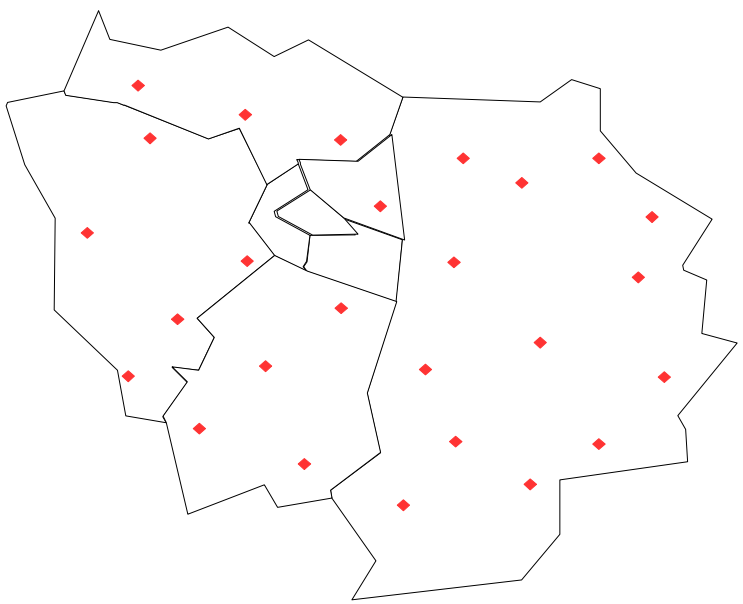
Plusieurs raisons justifient ce choix de « mode de collecte » :

1. qualité de l'information recueillie (relances, vérifications sur documents, aide à la compréhension du questionnaire, etc.) ;
2. temps de passation parfois longs (inenvisageables au téléphone ou par internet) ;
3. publics visé (enquête sur les personnes âgées, etc.).

Dans ce contexte, le **repérage des logements** et le **déplacement des enquêteurs** représentent une grande partie du coût de l'enquête.

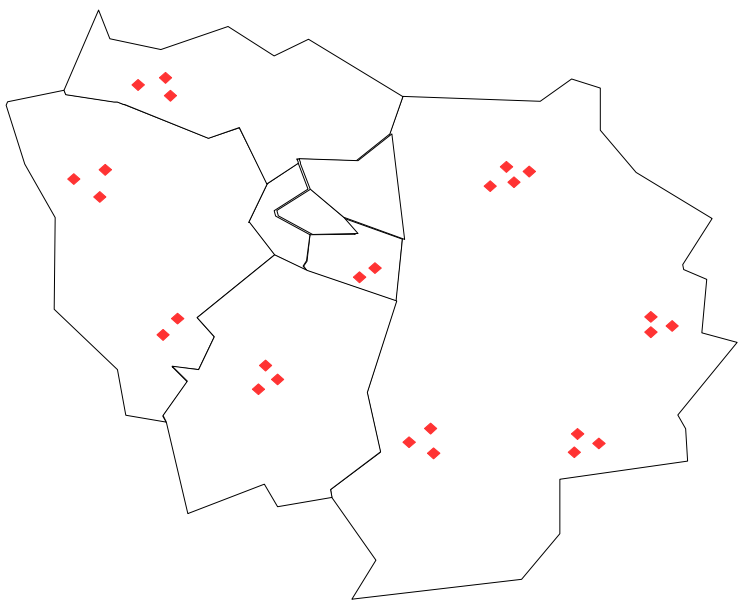
# Coût de déplacement et tirage aléatoire

Mais le principe-même du sondage conduit à considérablement augmenter les coûts de collecte : les logements tirés au sort peuvent être très éloignés les uns des autres.



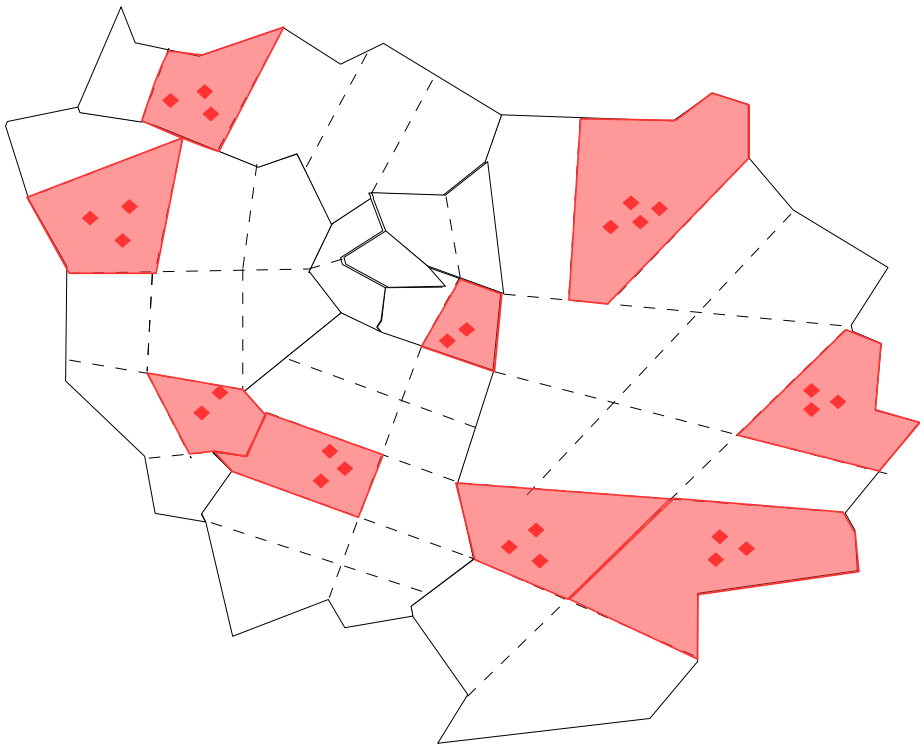
# Coût de déplacement et tirage aléatoire

Dans l'idéal, on souhaiterait que les logements tirés dans l'échantillon soient proches les uns des autres.



Mais on ne peut pas le décider directement, sinon ce ne serait plus un tirage aléatoire !

# Principe du sondage à deux degrés



# Principe du sondage à deux degrés

Le sondage à deux degrés peut être appliqué dans des contextes plus larges.

Mais dans tous les cas, on distingue trois étapes :

1. **partitionner la population en unités primaires** (les zones géographiques dans l'exemple) ;
2. **sélectionner un échantillon d'unités primaires** selon un certain plan de sondage ;
3. **sélectionner**, au sein de chaque unité primaire, **des unités secondaires** (les logements dans l'exemple) selon un certain plan de sondage.

# Avantages et inconvénients du sondage à deux degrés

## Avantages

- réduction du coût de collecte unitaire ;
- à budget constant, plus d'enquêtes peuvent être réalisées.

## Inconvénients

- un peu plus complexe que le SAS (mais pas tellement !) ;
- perte de précision si les zones sont très différentes les unes des autres pour la variable d'intérêt $Y$.

Autrement dit, le sondage à deux degrés est **peu efficace quand la variable $Y$ à mesurer présente une forte corrélation spatiale**.

Partie 2

# Sondage aléatoire simple à chaque degré

Comme dans les chapitres précédents, on applique le cadre général de Horvitz-Thompson (probabilités d'inclusion $\rightarrow$ estimateur sans biais $\rightarrow$ variance de l'estimateur).

Dans le cadre de cette introduction, on n'évoque que le cas du sondage à deux degrés **avec un SAS à chaque degré** :

1. tirage de $m$ unités primaires parmi $M$ par sondage aléatoire simple ;
2. au sein de chaque unité primaire $h$ ($1 \leq h \leq m$), tirage de $n_h$ unités secondaires parmi $N_h$ par sondage aléatoire simple.

Dans ce contexte, on peut montrer que l'estimateur d'Horvitz-Thompson du total d'une variable $Y$ s'écrit :

$$\hat{T}_{SAS-2D}(Y) = \frac{M}{m} \sum_{h=1}^{m} \left[ \frac{N_h}{n_h} \sum_{k \in s_h} y_k \right] = \frac{M}{m} \sum_{h=1}^{m} N_h \bar{y}_h$$

Cet estimateur est sans biais et sa variance s'écrit :

$$V\left(\hat{T}_{SAS-2D}(Y)\right) = \underbrace{M^2 \left(1 - \frac{m}{M}\right) \frac{S_{UP}^2}{m}}_{\text{1er degré}} + \underbrace{\frac{M}{m} \sum_{h=1}^{m} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_h^2}{n_h}}_{\text{2nd degré}}$$

où $S_{UP}^2$ est la variance inter unités primaires et $S_h^2$ la variance intra unités primaires (*cf.* décomposition de la variance).

# Mise en évidence d'un effet de grappe

Sous l'hypothèse que toutes les unités sont de même taille, on peut montrer que :

$$V\left(\hat{T}_{SAS-2D}(Y)\right) \approx N^2 \frac{S_{UP}^2}{n}\left(1+\rho\left(\frac{n}{m}-1\right)\right)$$

où $\rho$ est l'effet de grappe (ou corrélation intra-grappe) associé à la partition formée par les $M$ unités primaires.

**Plus les unités secondaires d'une même unité primaires sont homogènes, plus $\rho$ est élevé**.

Plus les unités primaires sont homogènes pour la variable $Y$, plus la variance de l'estimateur du total de la variable $Y$ est élevée.

# Le concept d'effet de sondage

Pour un plan de sondage $P$ et une variable $Y$ donnés, on appelle **effet de sondage** (ou *design effect*, *Deff*) le rapport :

$$Deff_P(Y) = \frac{V_P(Y)}{V_{SAS}(Y)}$$

L'effet de sondage est une mesure d'**efficacité relative du plan de sondage** $P$ pour la variable $Y$.

## Remarques

- par définition, $Deff_{SAS}(Y) = 1$ ;
- pour un SAS stratifié avec allocation proportionnelle, on a vu que $V^{prop}_{SAS-str}(Y) \leq V_{SAS}(Y)$ donc $Deff^{prop}_{SAS-str}(Y) \leq 1$.

# Effet de grappe et efficacité du plan de sondage à deux degrés

En raison de l'effet de grappe, **le sondage à deux degrés est toujours moins efficace que le sondage aléatoire simple**.

En effet, on peut montrer que :

$$Deff_{SAS-2D}(Y) \approx 1 + \rho\left(\frac{n}{m} - 1\right) > 1$$

Pour minimiser l'ampleur de la « pénalité » associée au plan de sondage à deux degrés, il faut :

- privilégier un échantillon d'unités primaires important ;
- faire en sorte que les unités primaires soient les plus hétérogènes possibles pour la variable d'intérêt $Y$.

Partie 3

# Principe de l'Échantillon-maître

# Limites du plan de sondage à deux degrés

Les plans de sondage à deux degrés sont centraux à l'Insee dans l'organisation des enquêtes auprès des ménages.

Le plan de sondage à deux degrés permet en effet de réaliser d'importantes économies tout en garantissant un certain niveau de précision.

Cependant, l'Insee réalise chaque année une dizaine d'enquête auprès des ménages avec le **même réseau d'enquêteurs**.

**Problème** Si d'une enquête à l'autre les unités primaires tirées changent du tout au tout, les enquêteurs doivent se déplacer très loin de leur domicile pour réaliser les enquêtes.

D'où l'idée de **mettre en commun le premier degré de toutes les enquêtes auprès des ménages** : c'est le principe de l'Échantillon-maître.

En pratique (jusqu'à 2009) :

- les unités primaires de toutes les enquêtes sont tirées une seule fois (après le recensement) ;
- pour chaque enquête, on tire dans le « stock » de logements des unités primaires tirées (puis les logements interrogés sont mis de côté) ;
- on met à jour chaque année la liste des logements avec les constructions et les destructions.

**Note** Depuis 2009, les enquêtes sont tirées dans le recensement rénové de la population (rotatif). Le nouvel échantillon-maître est plus complexe.

## Enjeux de la constitution de l'Échantillon-maître

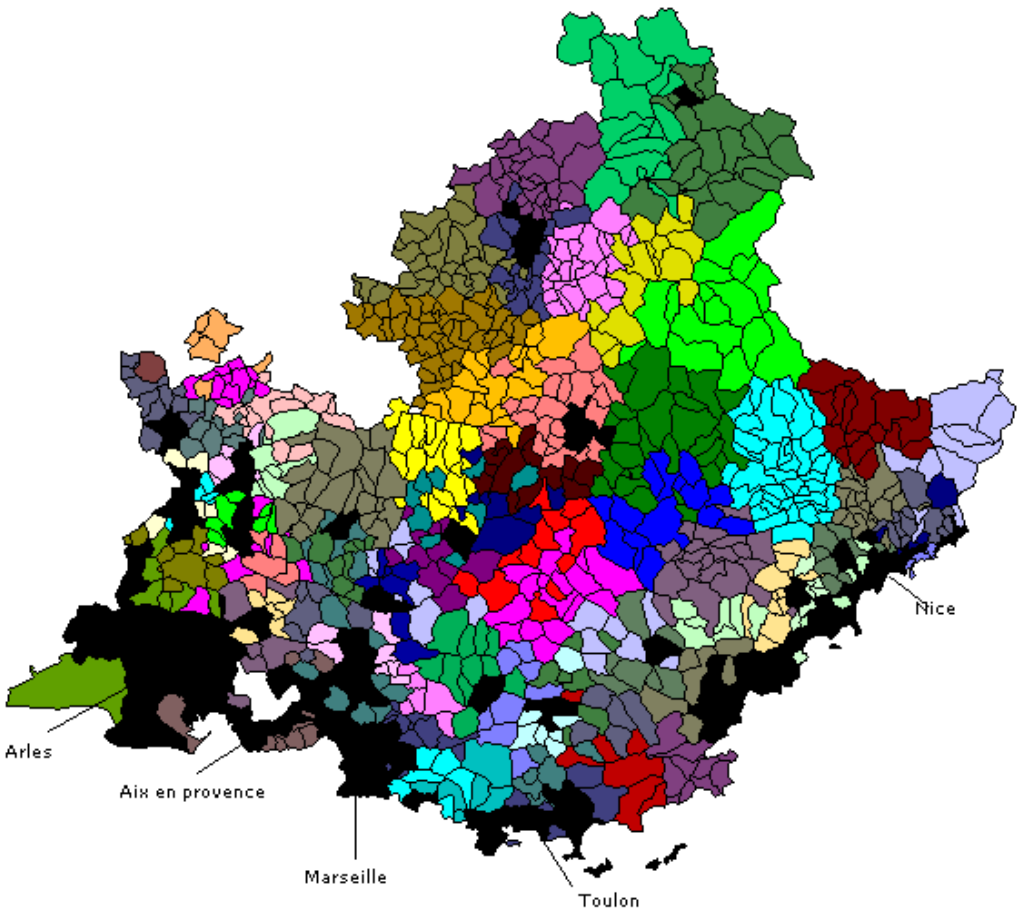
- bien construire les unités primaires pour permettre une variance faible pour beaucoup d'enquêtes différentes ;
- bien anticiper le nombre de logements nécessaires pour toute la durée de vie de l'échantillon.

## Avantages

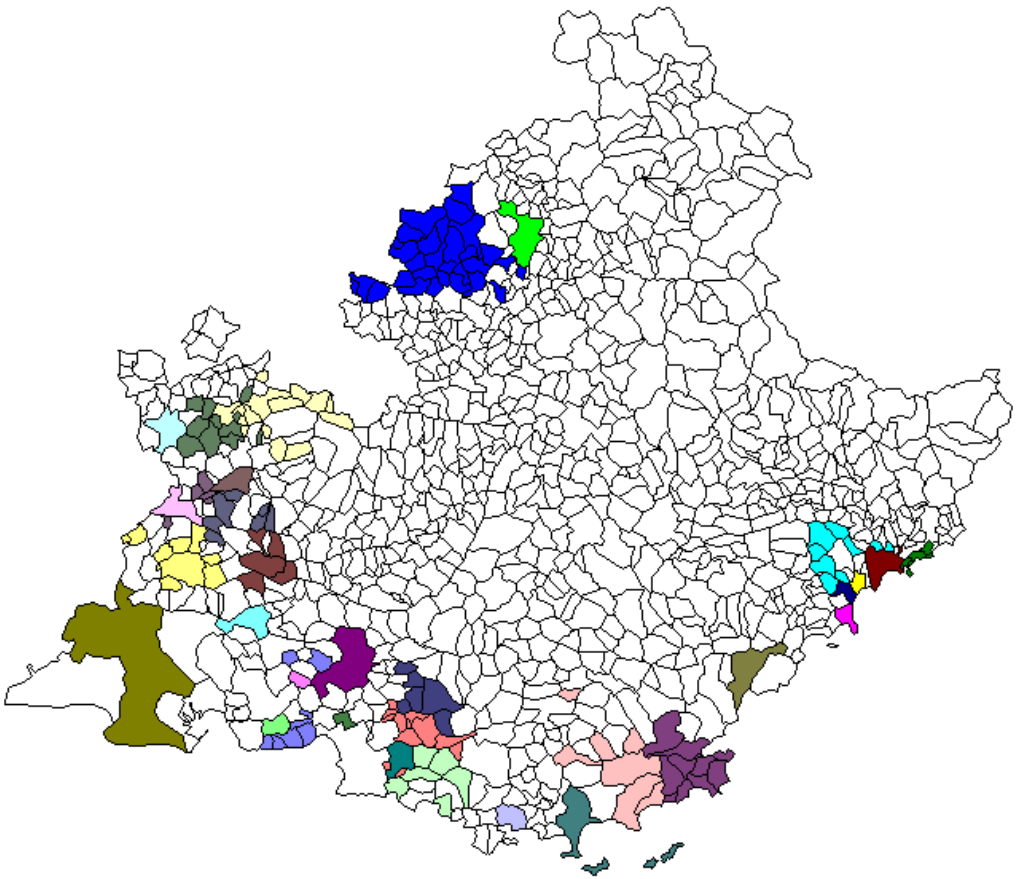
- diminution des coûts de gestion du réseau ;
- professionnalisation des enquêteurs.

**Inconvénient** Plus complexe qu'un sondage aléatoire simple (surtout avec un tirage équilibré au premier degré).

# Exemple : les unités primaires de la région PACA



# Exemple : les unités primaires de la région PACA



# Conclusion

Le sondage à plusieurs degrés est un moyen particulièrement efficace de réduire le coût d'une enquête sur le terrain.

À taille d'échantillon constante, l'effet de grappe inhérent à ce type de plan de sondage conduit à des estimateurs moins précis que ceux obtenus par un SAS.

Mais si le rapport coût fixe / coût variable est suffisamment élevé, une augmentation de la taille d'échantillon permet de compenser cette perte tout en diminuant le coût total.

Ce type de plan de sondage est central dans les enquêtes auprès des ménages de l'Insee *via* un Échantillon-maître.