Introduction à la théorie des sondages
Cours 4

Antoine Rebecq et Martin Chevalier
antoine.rebecq@insee.fr et martin.chevalier@insee.fr

INSEE, direction de la méthodologie

15 février 2016

Sommaire

Chapitre 1

Sondage à probabilités inégales

Partie 1

Principe du sondage à probabilités inégales

Sondage à probabilités inégales

Principe du sondage à probabilités inégales

Dans le cas du sondage stratifié avec allocation de Neyman, on a vu qu'il est avantageux de donner une **probabilité de sélection supérieure aux unités des strates dans lesquelles la variable d'intérêt est la plus dipersée**.

Il est possible d'**appliquer ce principe en dehors du cadre des plans de sondages stratifiés** pour espérer là encore des gains en termes de précision.

On s'intéresse alors à la classe des plans de sondage à probabilités inégales, c'est-à-dire des plans de sondages $p(s)$ pour lesquels les probabilités d'inclusion simples ne sont pas égales.

Sondage à probabilités inégales

Exemples de plans de sondages à probabilités inégales

Dans le cadre d'une enquête agricole, on cherche à estimer le **volume de la production totale Y d'une céréale**.

Avec un sondage aléatoire simple, les exploitations agricoles ont toutes la même probabilité d'être sélectionnées.

Néanmoins, la présence ou non dans l'échantillon des exploitations agricoles les plus étendues (qui contribuent donc le plus à la production de Y) induit une fore variance de l'estimateur du total de Y.

En faisant dépendre la probabilité de sélection de la taille des exploitations agricoles, il est possible de **limiter fortement la variance de l'estimateur**.

Partie 2

Sondage à probabilités proportionnelles

Sondage à probabilités inégales

Rappel : Estimateur d'Horvitz-Thompson

Soit s un échantillon dont les unités sont tirées selon des probabilités d'inclusion simple π_k.

On appelle estimateur d'Horvitz-Thompson (ou $\pi - estimateur$) du total de la variable Y l'estimateur :

$$\hat{T}_{HT}(Y) = \sum_{k \in s} \frac{y_k}{\pi_k}$$

Cet estimateur est sans biais. Sa variance peut-être calculée à partir des probabilités d'inclusion double π_{kl} :

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

Sondage à probabilités inégales

Sondage à probabilités proportionnelles à une variable Y

Pour une variable d'intérêt Y donnée, on appelle plan de sondage à probabilités proportionnelles à Y tout plan de sondage de taille n tel que :

$$\forall k \in \mathcal{U} \quad \pi_k = c \times Y_k, \quad c \in \mathbb{R}$$

Dans ce cadre, l'estimateur d'Horvitz-Thompson du total de Y devient :

$$\hat{T}_{HT}(Y) = \sum_{k \in s} \frac{Y_k}{\pi_k} = \sum_{k \in s} \frac{Y_k}{c \times Y_k} = \sum_{k \in s} \frac{1}{c} = \frac{n}{c}$$

Sondage à probabilités inégales

Efficacité pour l'estimation de Y

Cet estimateur estime parfaitement $T(Y)$:

$$\hat{T}_{HT}(Y) = \frac{n}{c} = \frac{\sum_{k \in \mathcal{U}} \pi_k}{c} = \frac{c \times T(Y)}{c} = T(Y)$$

On est au-delà du caractère sans biais : **quel que soit l'échantillon on retombe toujours sur le vrai total de Y.**

La variance de cet estimateur est donc nulle.

En d'autres termes, **le sondage à probabilités proportionnelles à Y est efficace pour estimer Y.**

Sondage à probabilités inégales

Mise en œuvre pratique

En pratique, on ne connaît pas les Y_k (le sondage serait inutile si on les avait !).

Solution : On choisit les π_k proportionnelles aux valeurs X_k d'une **variable auxiliaire** X connue sur la population et **supposée bien corrélée à** Y.

$$\pi_k = n \times \frac{X_k}{\sum_{k \in \mathcal{U}} X_k}$$

En règle générale, X correspond à une taille : on parle de **sondage proportionnel à la taille**.

Remarque : avec ce plan de sondage on estime parfaitement le total de X.

Sondage à probabilités inégales

Exemple : Sondage agricole

Ferme (k)	Taille (X_k)	Production (Y_k)	π_k		Poids	
			SAS	INEG	SAS	INEG
A	100	26	0,33	0,1	3	10
B	1000	470	0,33	1	3	1
C	125	66	0,33	0,125	3	8
D	250	145	0,33	0,25	3	4
E	500	280	0,33	0,5	3	2
F	25	13	0,33	0,025	3	40

Sondage à probabilités inégales

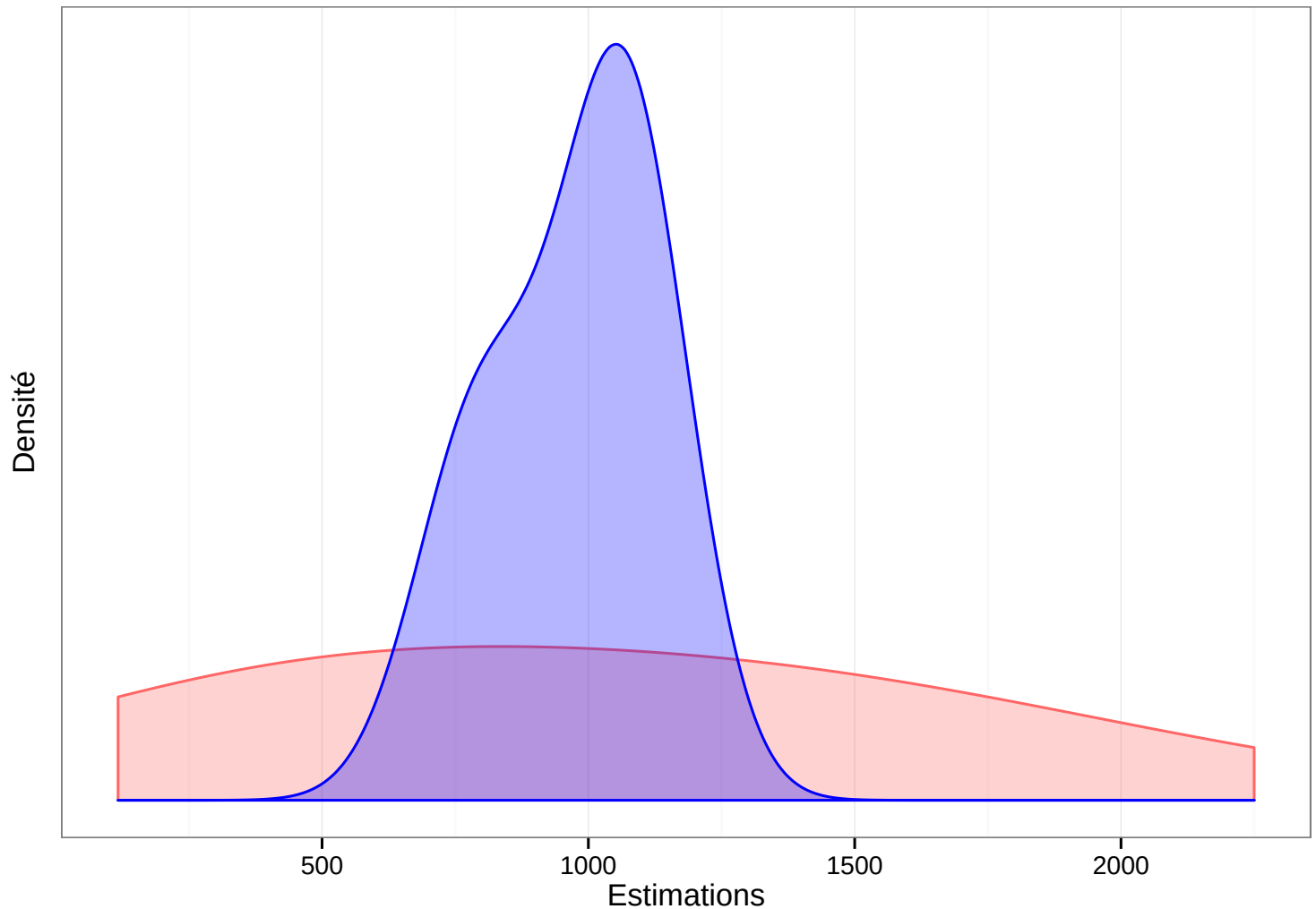
Exemple : Sondage agricole

Éléments	$\hat{T}_{SAS}(Y)$	$\hat{T}_{INEG}(Y)$
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

Éléments	$\hat{T}_{SAS}(Y)$	$\hat{T}_{INEG}(Y)$
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

Sondage à probabilités inégales

Exemple : Sondage agricole



Sondage à probabilités inégales

Traitement des unités exhaustives

Comme dans le cas d'un plan de sondage stratifié avec allocation de Neyman, le sondage à probabilités proportionnelles peut parfois conduire à des **probabilités d'inclusion supérieures à 1**.

Les unités concernées sont donc sélectionnées à coup sûr dans l'échantillon ($\pi_k = 1$) et **les probabilités des autres unités sont réévaluées** pour atteindre la taille d'échantillon souhaitée.

Il s'agit d'un mécanisme itératif analogue à celui présenté au cours 3.

Partie 3

Introduction au sondage équilibré

Sondage à probabilités inégales

Principe du sondage équilibré

L'échantillonnage dit « équilibré » permet de généraliser la propriété d'estimation parfaite à plusieurs variables auxiliaires.

Il est ainsi possible de renseigner plusieurs variables X telles que **quel que soit l'échantillon tiré** :

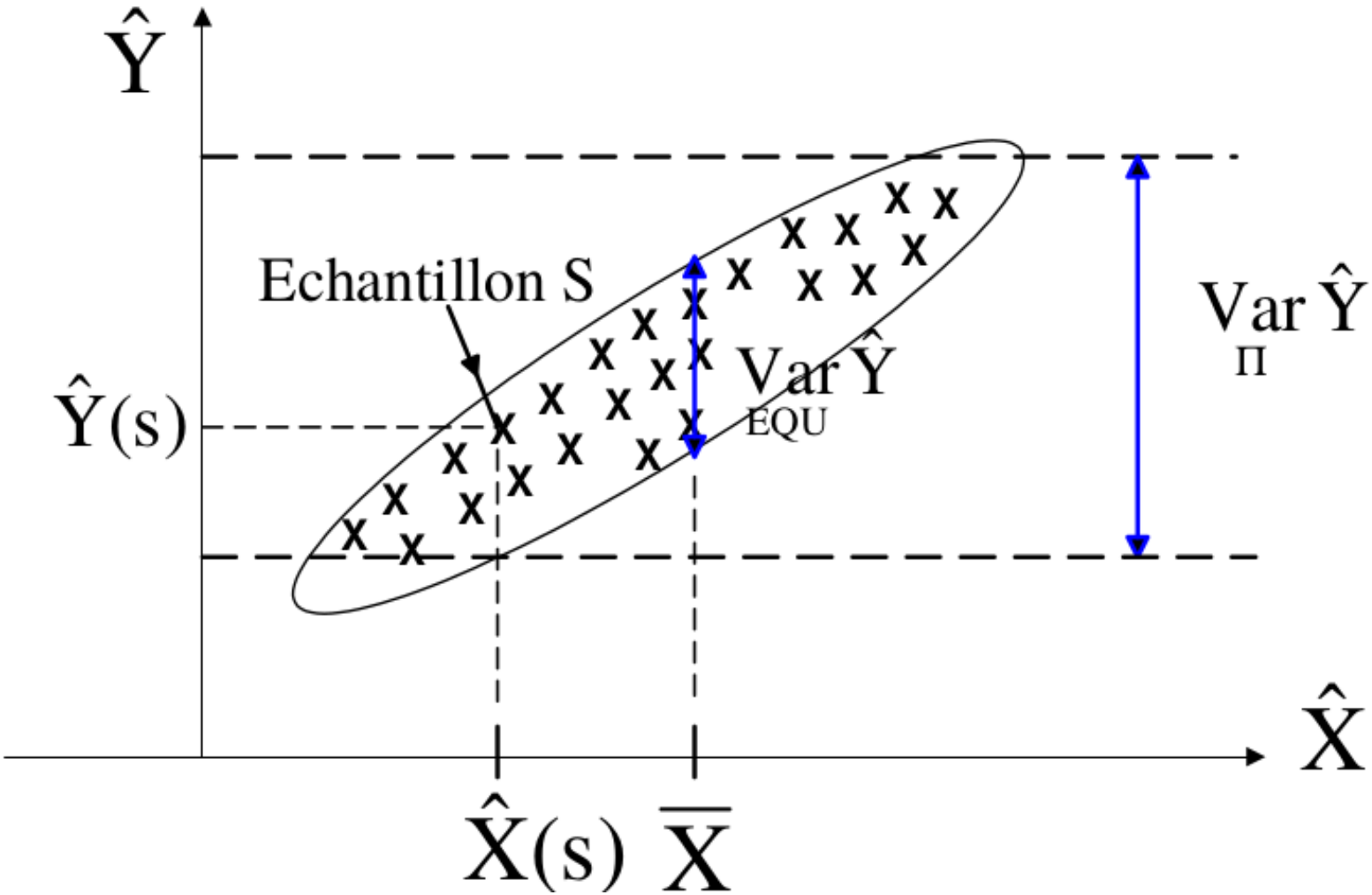
$$\hat{T}_{HT}(X) = T(X)$$

Les avantages de ce plan de sondage sont multiples :

1. les variables auxiliaires sont toujours parfaitement estimées ;
2. la variance des estimateurs de variables bien corrélées aux variables auxiliaires est bien plus faible qu'avec un SAS.

Sondage à probabilités inégales

Gains de variance associés au sondage équlibré



Sondage à probabilités inégales

Le sondage équilibré en pratique

En pratique, toutes les configurations ne permettent pas d'aboutir à des échantillons parfaitement équilibrés.

Exemple : dans une base de 100 personnes (50 hommes et 50 femmes), on ne peut pas tirer un échantillon de 11 personnes équilibré selon le sexe.

L'algorithme qui implémente cette méthode, dit algorithme du « Cube », comporte ainsi deux phases :

1. une phase de « vol » dans laquelle l'objectif est de constituer un échantillon qui respecte exactement les contraintes d'équilibrage ;
2. une phase d'« atterrissage » (optionnelle mais très fréquente) dans laquelle l'échantillon précédent est complété de façon à dégrader le moins possible ses bonnes propriétés.

Sondage à probabilités inégales

Le sondage équilibré en pratique

La méthode du sondage équilibré a été développée notamment en France au sein de l'Insee (travaux de Jean-Claude Deville).

Elle est implémentée sous SAS (macro *%cube*) et sous **R** (package *sampling* d'Yves Tillé, package *BalancedSampling*).

Elle est **très utilisée à l'Insee dans les enquêtes auprès des ménages** : c'est par cette méthode que sont sélectionnés les communes de l'≪ Échantillon-maître ≫.

Chapitre 2

Introduction au sondage à deux degrés

Partie 1

Principe du sondage à deux degrés

Introduction au 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 » :

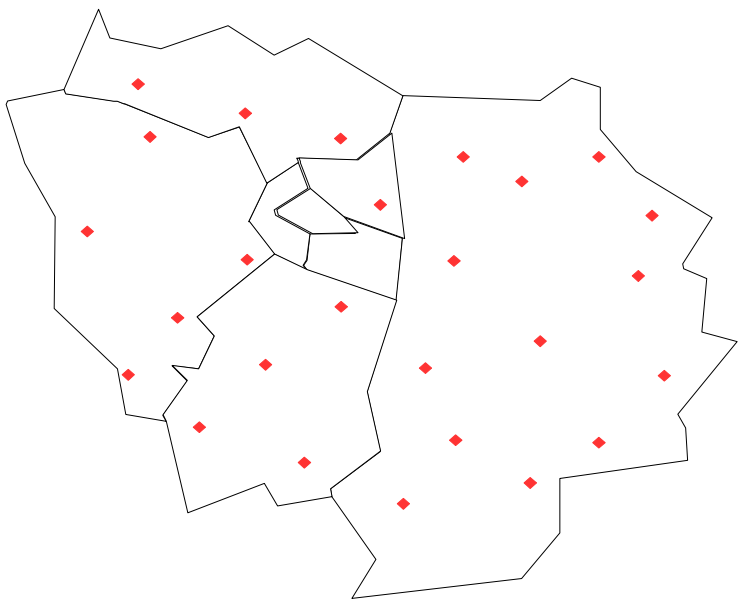
1. qualité de l'information recueillie (relances, vérifications sur documents, aide à la compréhension du questionnaire, etc.) ;
2. temps de passation parfois longs (inenvisageables au téléphone ou par internet) ;
3. publics visé (enquête sur les personnes âgées, etc.).

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

Introduction au sondage à deux degrés

Coût de déplacement et tirage aléatoire

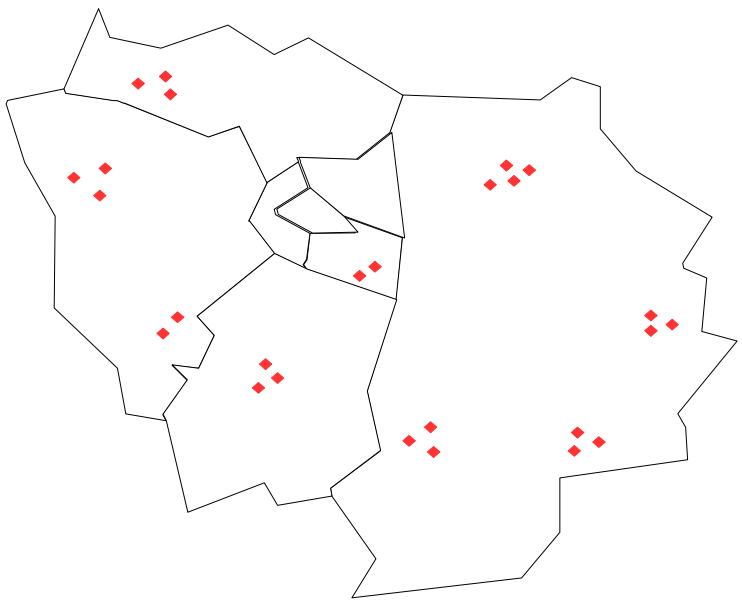
Mais le principe-même du sondage conduit à considérablement augmenter les coûts de collecte : les logements tirés au sort peuvent être très éloignés les uns des autres.



Introduction au sondage à deux degrés

Coût de déplacement et tirage aléatoire

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



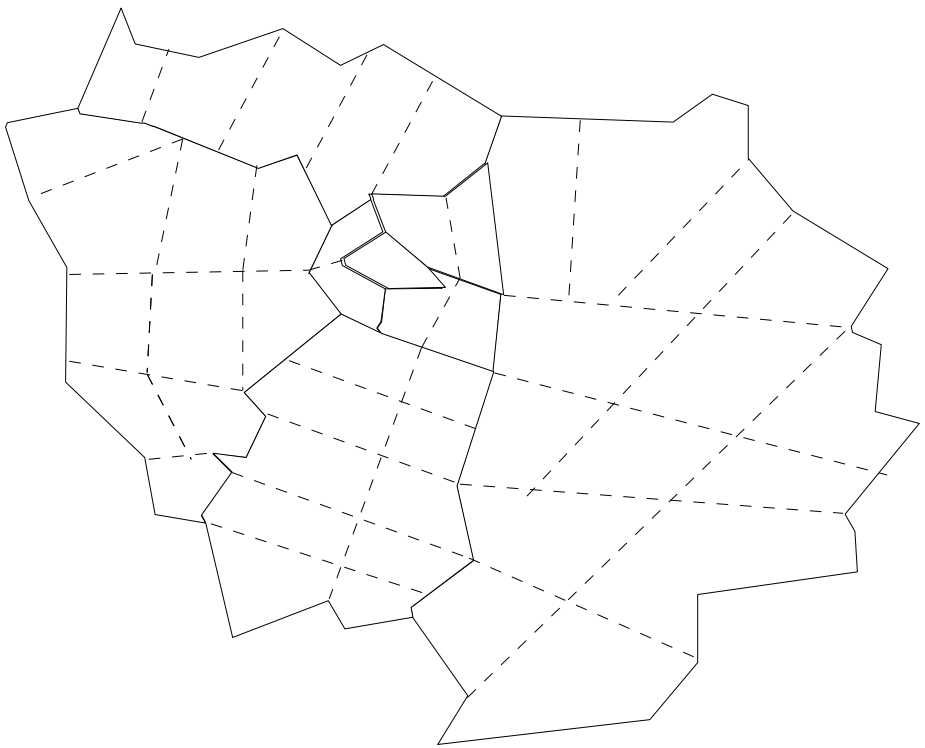
Mais on ne peut pas le décider directement, sinon ce ne serait plus un tirage aléatoire !

Introduction au sondage à deux degrés

Principe du sondage à deux degrés

Solution : le tirage à deux degrés

Étape 1 : Découper l'espace en zones

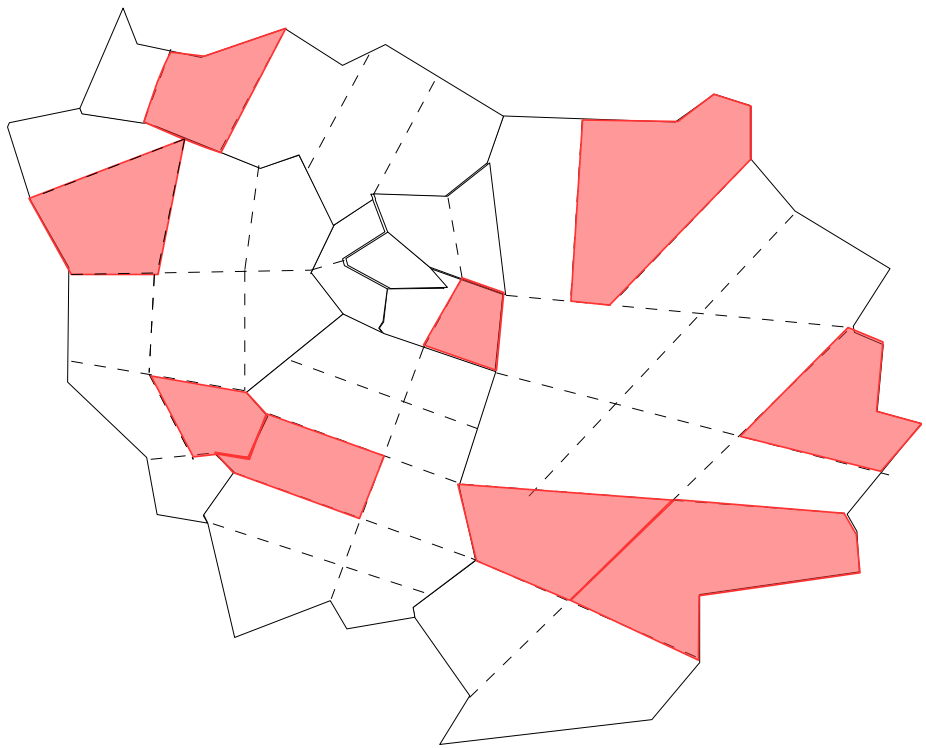


Introduction au sondage à deux degrés

Principe du sondage à deux degrés

Solution : le tirage à deux degrés

Étape 2 : Tirer un échantillon de zones

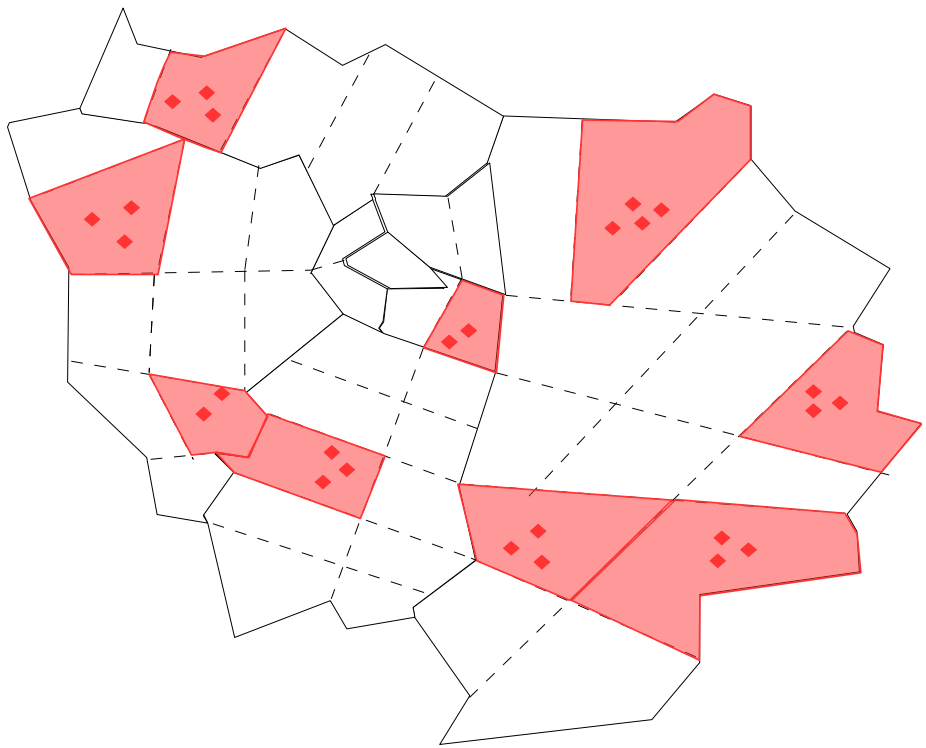


Introduction au sondage à deux degrés

Principe du sondage à deux degrés

Solution : le tirage à deux degrés

Étape 3 : Tirer un échantillon d'individus dans les zones tirées



Introduction au 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.

Introduction au sondage à deux degrés

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

Introduction au sondage à deux degrés

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.

Introduction au sondage à deux degrés

Sondage aléatoire simple à chaque degré

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).

Introduction au sondage à deux degrés

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 ρ 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 ρ 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.

Introduction au sondage à deux degrés

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 $Deff^{prop}_{SAS-str}(Y) \leq 1$.

Introduction au sondage à deux degrés

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

Le sondage à deux degrés en pratique

Introduction au sondage à deux degrés

De l'importance du coût relatif

Un sondage à deux degrés est presque toujours moins efficace qu'un sondage aléatoire simple de même taille.

En pratique cependant, le sondage à deux degrés permet dans la plupart des cas d'obtenir la même précision que le SAS à un coût moindre.

Pour comprendre pourquoi, on peut décomposer le coût d'une enquête en deux composantes :

- le **coût variable** est intrinsèquement lié au nombre d'entretiens réalisés (durée de passation) ;
- le **coût fixe** à l'inverse est susceptible d'être mutualisé entre plusieurs entretiens (repérage, déplacement,etc.).

Introduction au sondage à deux degrés

De l'importance du coût relatif

En rapprochant géographiquement les entretiens à réaliser, le sondage à deux degrés permet de **faire des économies sur le coût fixe**, ce qui permet d'**augmenter la taille de l'échantillon**.

Si le rapport coût fixe / coût variable est suffisamment élevé, alors cette augmentation de la taille de l'échantillon peut **compenser la perte de précision** associée au plan de sondage à deux degrés.

Selon l'organisation de la collecte, le coût relatif peut conduire à privilégier ou non le sondage à deux degrés sur le sondage aléatoire simple.

Introduction au sondage à deux degrés

Un exemple de sondage à deux degrés atypique : l'enquête AES 2016

L'enquête AES (*Adult education survey*) est une enquête portant sur la formation des adultes de 18 à 64 ans.

L'unité d'intérêt est l'individu, qui est interrogé en face-à-face lors d'entretiens d'environ 30 min (coût variable).

Le temps moyen de repérage et de déplacement est de 2 heures (coût fixe).

On est donc dans une situation où le coût fixe (repérage et déplacement) excède largement le coût variable (durée d'entretien).

Introduction au sondage à deux degrés

Un exemple de sondage à deux degrés atypique : l'enquête AES 2016

Dans ce contexte, on a proposé l'ajout d'un degré de sondage supplémentaire, le ménage :

- l'idée est d'interroger non plus un mais deux individus du champ de l'enquête par ménage ;
- pour les ménages concernés, on passe donc de 2h30 pour 1 entretien à 3h pour 2 entretiens.

Cependant, la qualité de l'information statistique recueillie par ce biais est moindre qu'avec un individu par logement :

- les individus d'un même ménage « se ressemblent » : proximité d'âge et homogamie ;
- la corrélation intra-grappe (ici au sein du ménage) est donc non-nulle : à taille d'échantillon identique, on perd en précision.

Introduction au sondage à deux degrés

Un exemple de sondage à deux degrés atypique : l'enquête AES 2016

Un travail méthodologique a permis de déterminer le nombre d'individus supplémentaires à interroger pour maintenir le niveau de précision.

En pratique, il a fallu estimer la valeur de la corrélation intra-grappe pour les variables d'intérêt de l'enquête (nombre de formations suivies).

La principale source en ce domaine est l'Enquête emploi en continu (EEC) réalisée sur l'ensemble des individus de 15 ans et plus.

En définitive, l'échantillon a dû être gonflé de 20 %, mais comme les « seconds interrogés » coûtent beaucoup moins chers, le coût total de l'enquête a pu être diminué de 10 % à précision constante.

Partie 4

Principe de l'Échantillon-maître

Introduction au sondage à deux degrés

Limites du plan de sondage à deux degrés

Au-delà de l'exemple de l'enquête AES, 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.

Introduction au sondage à deux degrés

Principe de l'Échantillon-maître

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.

Introduction au sondage à deux degrés

Principe de l'Échantillon-maître

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 :

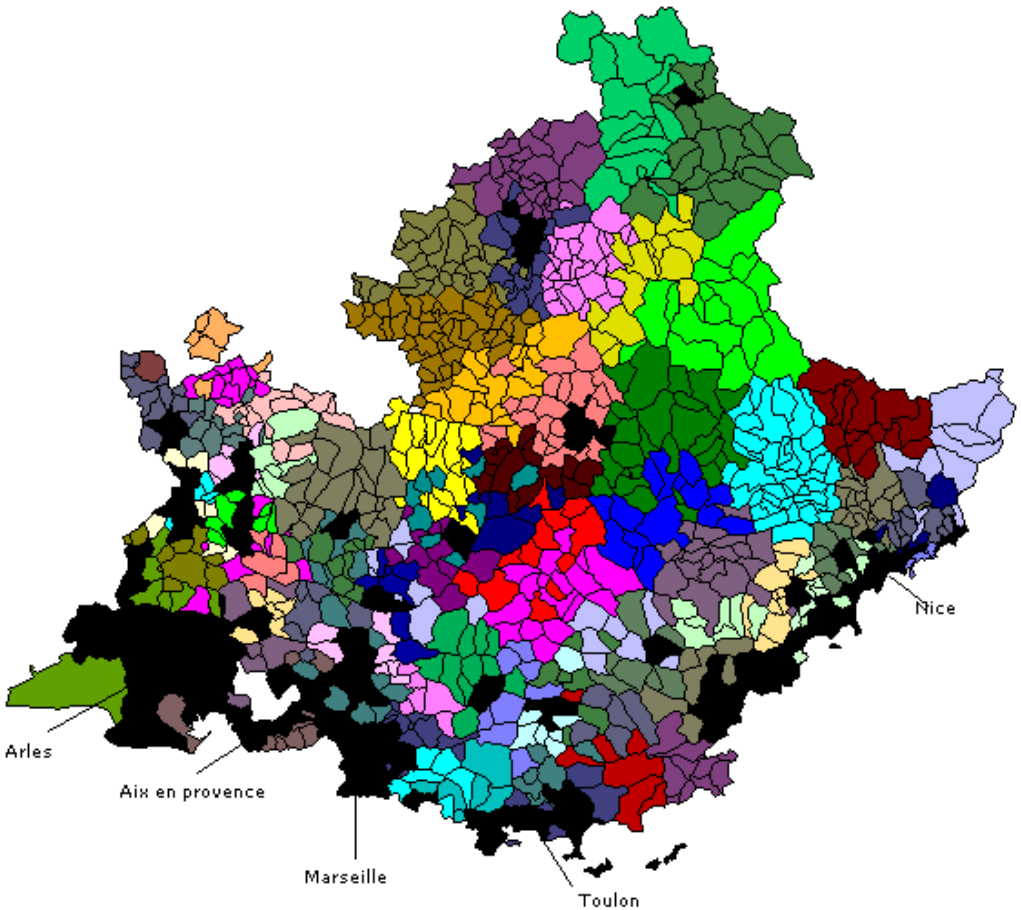
- diminution des coûts de gestion du réseau ;
- professionnalisation des enquêteurs.

Inconvénients : plus complexe qu'un sondage aléatoire simple (surtout avec un tirage équilibré au premier degré).

Introduction au sondage à deux degrés

Principe de l'Échantillon-maître

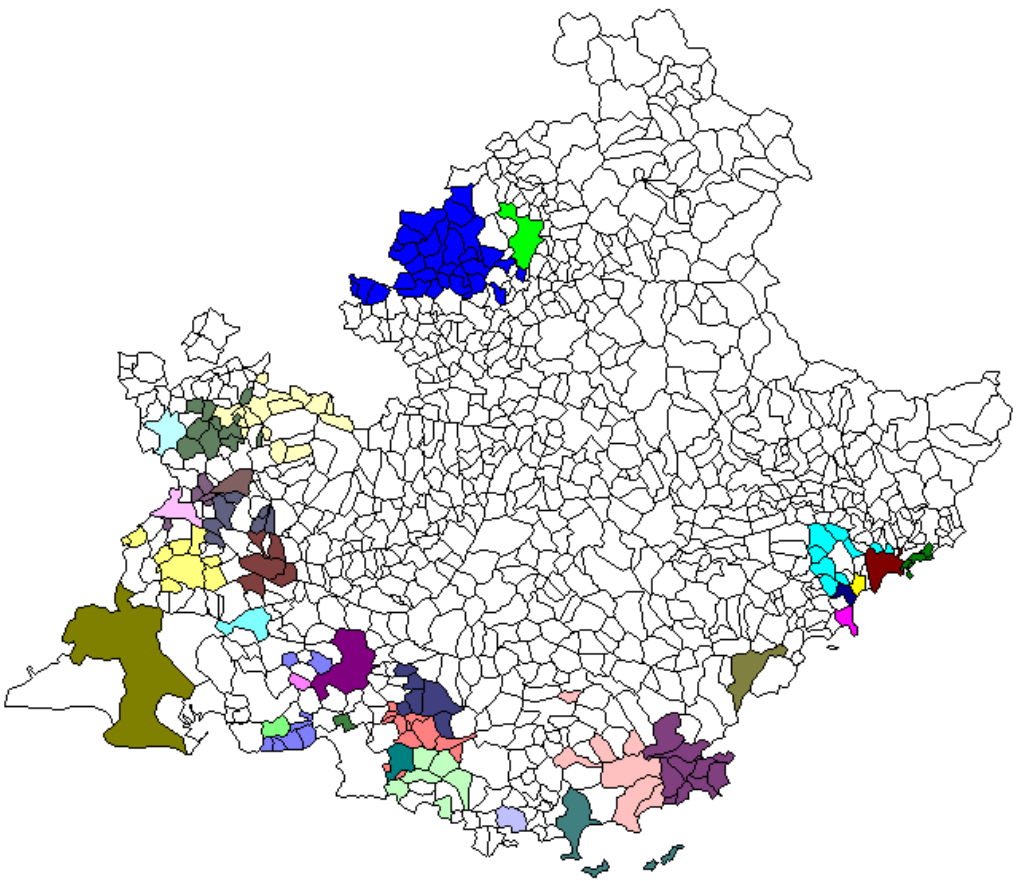
Exemple : les unités primaires de la région PACA



Introduction au sondage à deux degrés

Principe de l'Échantillon-maître

Exemple : les unités primaires de la région PACA sélectionnées dans l'échantillon-maître



Introduction au sondage à deux degrés

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.