La disjonction des échantillons des enquêtes auprès des ménages : de la théorie à la pratique

Journées de la méthodologie statistique 2022

Nicolas Paliod

Département des méthodes statistiques - Insee

30 mars 2022

Sommaire

Plan

Limiter la charge d'enquête

- L'article 9 du code des bonnes pratiques de la statistique européenne porte sur la charge non excessive pour les déclarants.
  Article 9.2 : « La charge de réponse est répartie aussi largement que possible entre les populations sondées et contrôlée par l'autorité statistique »

- Site internet du comité du label : « Selon les termes du décret de 2013, le Comité du label couvre tous les aspects de l'enquête : il évalue (...) la charge qu'implique l'enquête pour les personnes physiques ou morales qui en font l'objet »

Favoriser des taux de réponse élevés

- Enquêtes longues, parfois plusieurs heures d'interrogation pour certaines sous-populations d'intérêt dans quelques enquêtes
- Enquêtes récurrentes sous forme de panel, jusqu'à 6 interrogations pour l'enquête Emploi
- Besoin d'avoir des taux de réponse élevés pour limiter les biais de non-réponse et les hypothèses lors des redressements
- Important de limiter la charge de collecte pour ne pas lasser les enquêtés : limiter le recouvrement entre les échantillons d'enquêtes différentes est une solution

Plan

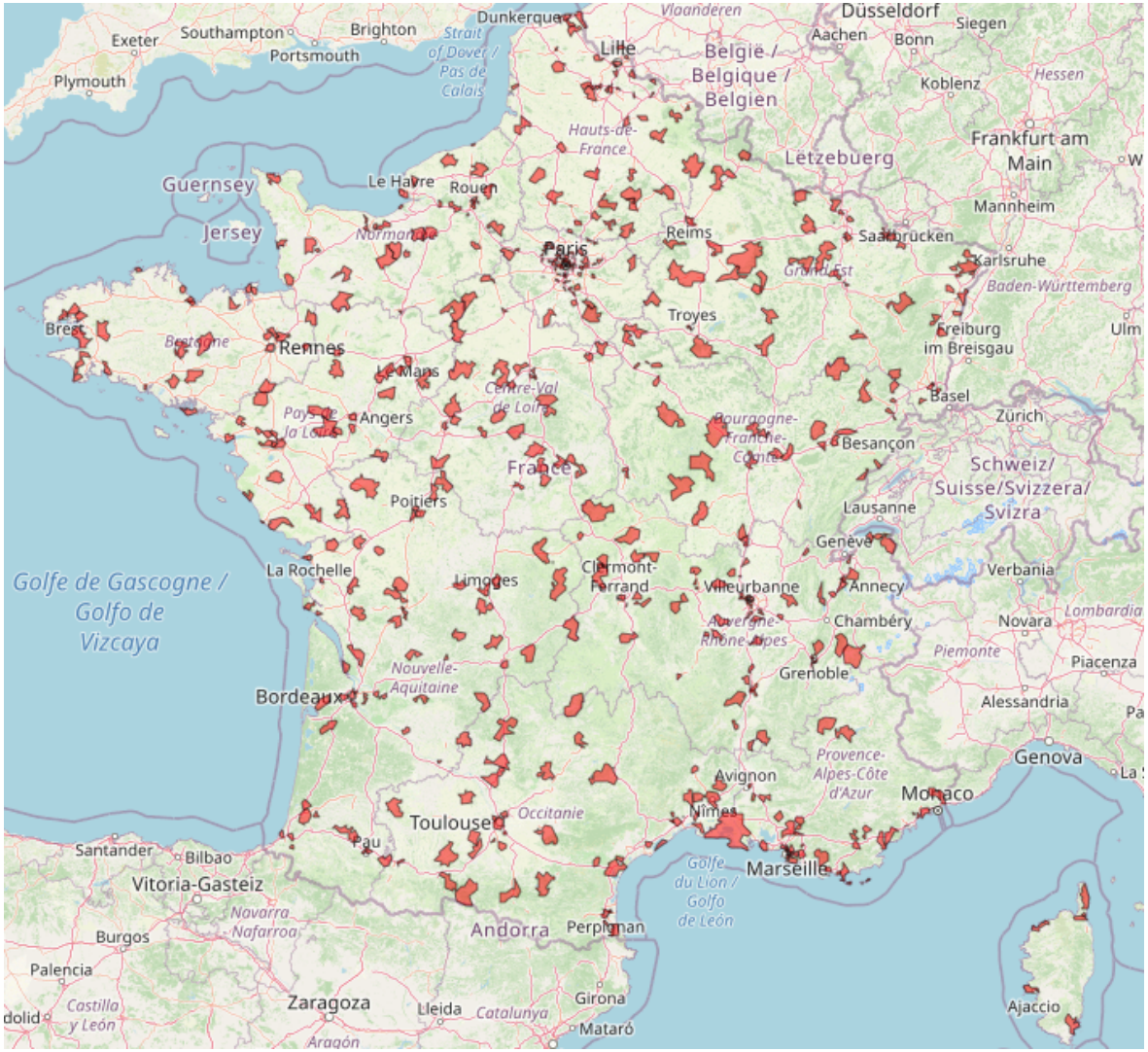
Tirage d'échantillons pour les enquêtes ménages

- Plusieurs sources de données pour les tirages d'échantillons d'enquêtes auprès des ménages réalisés par l'Insee :
  - Sources fiscales (tirages en France métropolitaine et dans les DROM)
  - Enquêtes annuelles de recensement (tirages dans les DROM hors Mayotte)
  - Bases cartographiques (tirages à Mayotte)
- Cadre de la présentation : tirages dans les sources fiscales
- Deux types d'unités possibles pour le tirage : logements ou individus
- Des strates qui changent d'une enquête à l'autre

Des tirages à 1 ou 2 degrés

- Les modes de collecte
  - Face-à-face (mode de collecte historique)
  - Téléphone
  - Internet
  - Papier
- Les spécificités du face-à-face
  - Nécessité de se restreindre à des zones de collecte (des « unités primaires ») pour limiter le déplacement des enquêteurs
  - Tirage d'un échantillon d'unités primaires
  - Tirage à 2 degrés : les logements ou les individus sont ensuite tirés dans l'échantillon d'unités primaires

Exemple d'échantillon d'unités primaires



Plan

Mécansime de disjonction

- Disjonction d'échantillons : suppression des logements (resp. des individus) tirés lors d'enquêtes passées de la base de sondage utilisée lors d'un tirage de logements (resp. d'individus)

- Base de tirage : base de sondage privée des unités tirées par le passé

Pondérations

- Pour un tirage dans une strate s, on souhaite pouvoir utiliser les poids de sondage en l'absence de disjonction pour une unité l sélectionnée :
  - $w_l = \frac{N_s}{n_s}$ pour les tirages à 1 degré où N_s est la taille de la strate s dans la base de sondage avant disjonction et n_s est l'allocation dans la strate s
  - $w_l = \frac{1}{\pi_{up}} \frac{N_{s,up}}{n_{s,up}}$ pour les tirages à 2 degrés où $N_{s,up}$ est la taille de la strate s dans l'unité primaire up dans la base de sondage avant disjonction et $n_{s,up}$ est l'allocation dans le croisement de la strate s avec l'unité primaire up

Des estimations sans biais sous conditions (1/2)

- Pondérations qui sont calculées par post-stratification « comme si » on avait fait le tirage dans la base de sondage
- Les estimations sont-elles sans biais ?
  - Impose que toutes les unités aient la même probabilité d'avoir été tirées dans le passé au sein d'une unité primaire donnée, conditionnellement aux tirages d'échantillons d'unités primaires
  - Impose que toutes les unités aient la même probabilité d'avoir été tirées dans le passé quelle que soit leur unité primaire, en considérant les tirages d'unités primaires comme aléatoires
  - Pas évident à démontrer même en tenant compte de ces éléments (voir l'article)

Des estimations sans biais sous conditions (2/2)

- Les tirages réalisés à l'Insee doivent nécessairement garantir :
  - que toutes les unités d'une région ont la même probabilité d'être tirées à 1 degré
  - que toutes les unités d'une unité primaire ont la même probabilité d'être tirées à 2 degrés
  - que toutes les unités d'une région ont la même probabilité d'être tirées à 2 degrés, quand on considère que l'échantillon d'unités primaires est aléatoire

- Or, il y a des limitations de champ ou des surreprésentations dans les enquêtes : nécessité de tirer des échantillons complémentaires pour rééquilibrer la base pour de futurs tirages

Illustration tirages à 1 degré - Base de sondage avant tirage

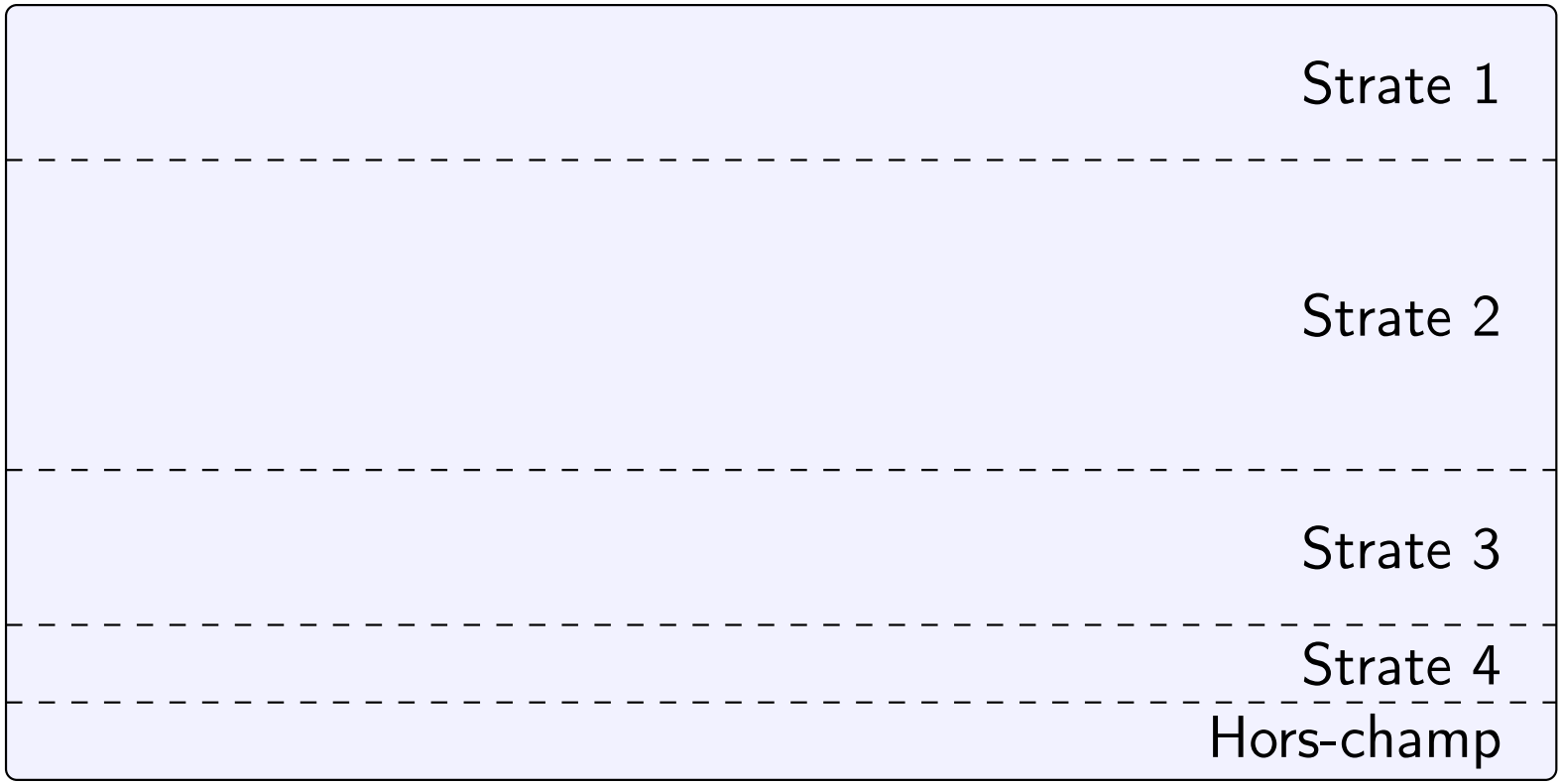


Illustration tirages à 1 degré - échantillon tiré dans l'enquête

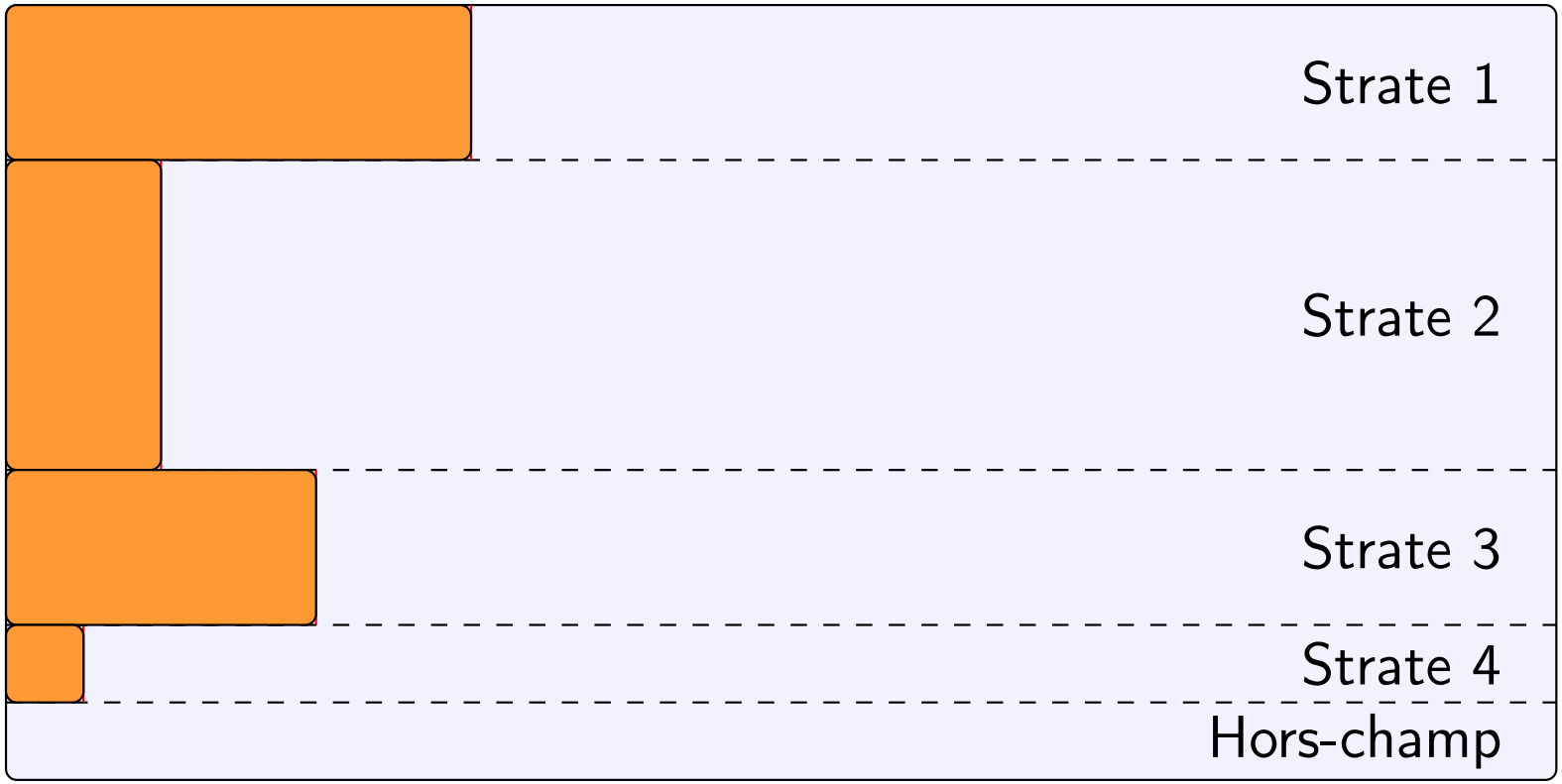


Illustration tirages à 1 degré - échantillon complémentaire pour rééquilibrer la base

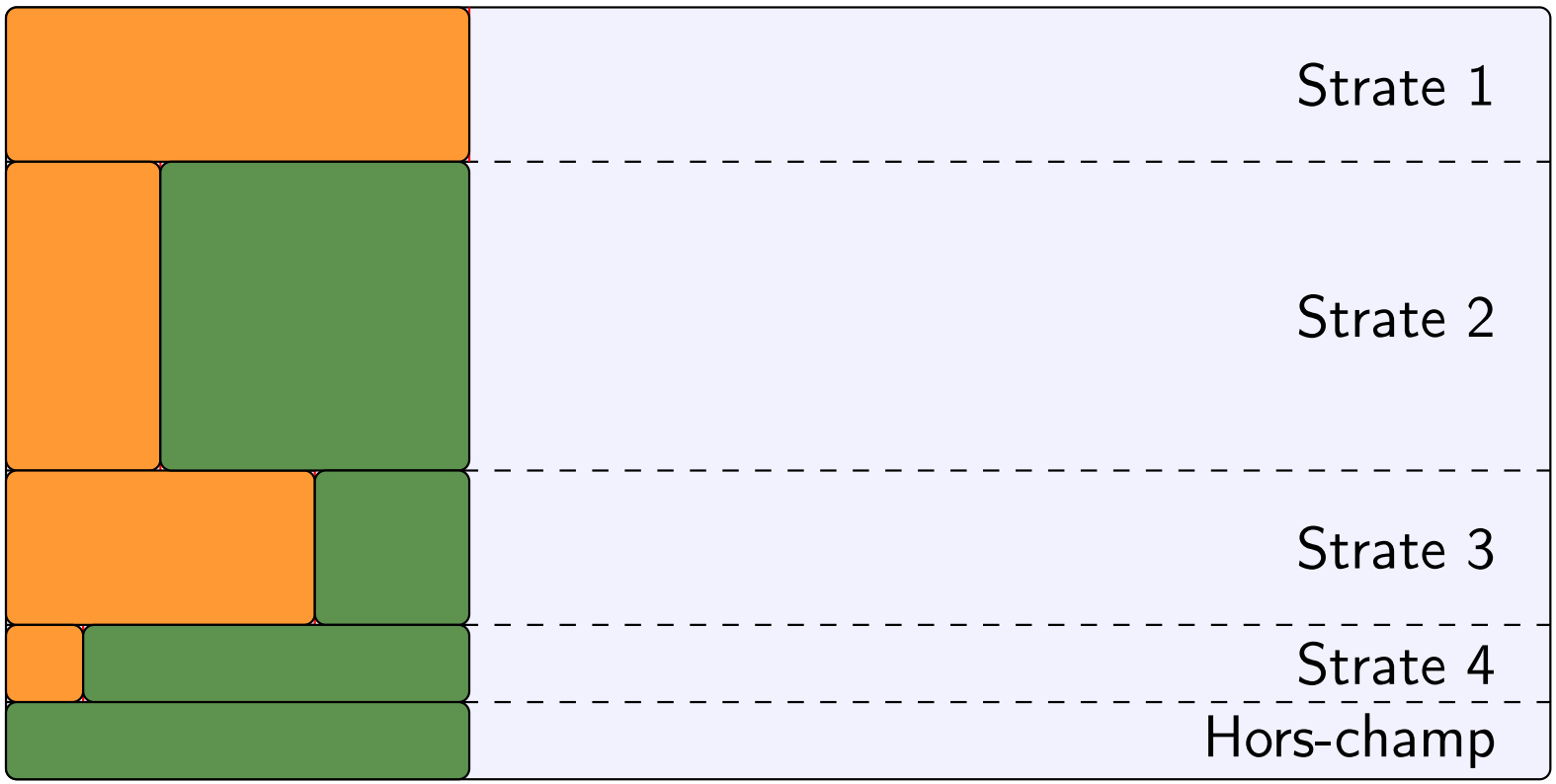


Illustration tirages à 1 degré - Base de tirage pour les tirages ultérieurs

Plan

Quels échantillons disjoint-on lors d'un tirage ?

- On retire de la base de tirage d'individus :
  - les échantillons d'individus tirés par le passé
  - les individus vivant dans des logements tirés par le passé (actuellement disjonction uniquement avec les occupants des logements de l'Enquête Emploi en Continu, à termes avec tous les occupants des logements tirés par le passé)
- On retire de la base de tirage de logements :
  - les échantillons de logements tirés par le passé

Efficacité de la disjonction pour la collecte (1/2)

- Essentiel des cas de réinterrogation :
  - déménagement d'un ménage ou d'un individu
  - appariement imparfait entre millésimes d'une même source administrative
  - tirage d'un individu, puis de son logement

Efficacité de la disjonction pour la collecte (2/2)

- Des réinterrogations impossibles à éviter
  - Hypothèse-clé du marquage : les unités (logements ou individus) au sein d'une unité primaire donnée doivent avoir la même probabilité d'avoir été tiré dans le passé
  - Les déménagements nécessitent des opérations spécifiques pour continuer à satisfaire cette hypothèse et ne pas s'autoriser à réinterroger des individus qui ont déménagé ne permet pas de respecter cette hypothèse (voir présentation du colloque francophone sur les sondages 2020)
  - Impossibilité de supprimer de la base d'un tirage de logements les logements dont des occupants ont été tirés par le passé lors d'un tirage d'individus (les logements avec plusieurs individus auraient plus de probabilités d'être supprimés de la base que ceux occupés par un seul individu)

La disjonction en chiffres en France métropolitaine

- 2,2 millions de logements supprimés de la base de tirage de logements (en comptant les résidences secondaires, logements vacants, etc.), soit 5,5 % de la base de sondage
- 6,2 millions d'individus supprimés des tirages d'individus, soit 9,1 % de la base de sondage

Efficacité statistique de la disjonction (1/4)

- Sur le plan théorique, les démonstrations présentes dans l'article donnent des intuitions, permettent des démonstrations dans des cas simples
- Mais les démonstrations dans le cas général n'ont pas été réalisées
- Nécessité de mettre la pratique à l'épreuve des faits
- On regarde si la base de sondage et la base de tirage (base de sondage dont sont supprimées les unités disjointes) ont la même structure

Efficacité statistique de la disjonction (2/4)

- A un degré, on suppose que le poids d'une unité dans la base de tirage est $\frac{N_{reg}}{\tilde{N}_{reg}}$ où N_{reg} est la taille de la région dans la base de sondage et $\tilde{N}_{reg}$ la taille de la région dans la base de tirage
- A deux degrés, on suppose que le poids d'une unité dans la base de tirage est $\frac{1}{\pi_{up}} \frac{N_{up}}{\tilde{N}_{up}}$ où N_{up} est la taille de l'unité primaire dans la base de sondage et $\tilde{N}_{up}$ la taille de l'unité primaire dans la base de tirage

Efficacité statistique de la disjonction (3/4)

Base de tirage de logements à 2 degrés, comparaison avec les totaux estimés sur la base de sondage restreinte à l'échantillon-maître

Variable	Totaux base de sondage	Totaux base de tirage
Total	28 655 000	28 655 000
Appartements	12 850 000	12 849 000
Logements sociaux	4 355 000	4 355 000
Sous seuil pauvreté	3 565 000	3 569 000
QPV	1 813 000	1 815 000
1^{er} décile	2 574 000	2 578 000
Moins de 30 m^2	1 245 000	1 253 000
Construit après 2005	4 352 000	4 354 000

Efficacité statistique de la disjonction (4/4)

Base de tirage d'individus à 1 degré

Variable	Totaux base de sondage	Totaux base de tirage
Total	66 491 000	66 490 000
10-14 ans	4 085 000	4 091 000
40-44 ans	4 107 000	4 105 000
Femmes	27 808 000	27 779 000
Perception salaire	28 959 000	28 948 000
Perception chômage	5 578 000	5 579 000
Perception bén agri	539 000	538 000
QPV	4 974 000	4 977 000
1^{er} décile	6 024 000	6 032 000

Plan

Conclusion

- La méthode de disjonction fonctionne si :
  - pour des tirages à 1 degré, toutes les unités du champ et hors-champ de l'enquête ont la même probabilité d'être tirées
  - pour des tirages à 2 degrés, des allocations auto-pondérées sont retenues et si toutes les unités du champ et hors-champ d'une unité primaire ont la même probabilité d'être tirées

- Même en disjoignant 10 % de la base de sondage dans le cas de tirage d'individus, la qualité de la base utilisée pour le tirage reste très satisfaisante

- Pas de preuve théorique à ce jour sur le marquage mais fonctionne empiriquement, c'est une opération de post-stratification

Conclusion

Merci de votre attention !