# Echantillonnage aléatoire simple sur des données réelles, issues de ressources numériques

Christelle Breuils Degryse

I.U.T. de Metz, Département STID

24 novembre 2022

# Cadre

- Etudiants en première année de BUT STID
- Module de 10h TP en fin d'année scolaire
- Etude majoritairement en autonomie, par groupe (en salle TP avec le professeur)
- Habitudes : plutôt un contexte d'enseignement théorique, avec des TP de type exercices (données nettoyées avant le TP, situations d'études déjà décrites)

# Définition du problème posé

- Suggestion de Franck Gaüzère (MCF à L'IUT) : étude sur les données des élections présidentielles
- Contour choisi : première étude sur 2017
- Choix de la variable étudiée à déterminer par le groupe
- Seconde étude en 2022
- Comparaison 2017/2022 (si le temps le permet)

Motivations pédagogiques

- Premières élections (2022) auxquelles la plupart des étudiants ont eu le choix de voter
- Choix de la variable qui suscite l'intérêt des étudiants

1. Calcul de la moyenne pour taux d'abstention au premier tour, pour toutes les données de 2017,
2. Echantillonnage et calcul du taux d'abstention national, que l'on compare au taux du 1. Modification de la taille des échantillons.
3. Pour les données de 2022, effectuer la même chose qu'aux 1. et au 2. pour 2022
4. Calcul d'intervalles de confiance pour 2017 et 2022
5. Test de comparaison, avec comme valeur de référence la valeur de 2017, si l'abstention estimée en 2022 est statistiquement supérieure ou non.
6. Si temps : cartographie (par exemple abstention par département) ; changement de variable d'échantillonnage (villes...)
7. Rapport complet

- Année de référence 2017
- Exemple de variable : abstention
- Source des documents : data.gouv.fr, éventuellement fichiers modifiés par utilisateurs et validés sur le site
- Aspects pédagogiques (en dehors de la statistique) : décompte des résultats des élections (différence entre votes blancs, nuls et abstention)
- Aspects techniques : importation des mauvais fichiers en fonction du thème choisi (pourcentages ordonnés par score des candidats, ...) + procédure d'importation sous R (difficulté informatique par exemple pour codage département)

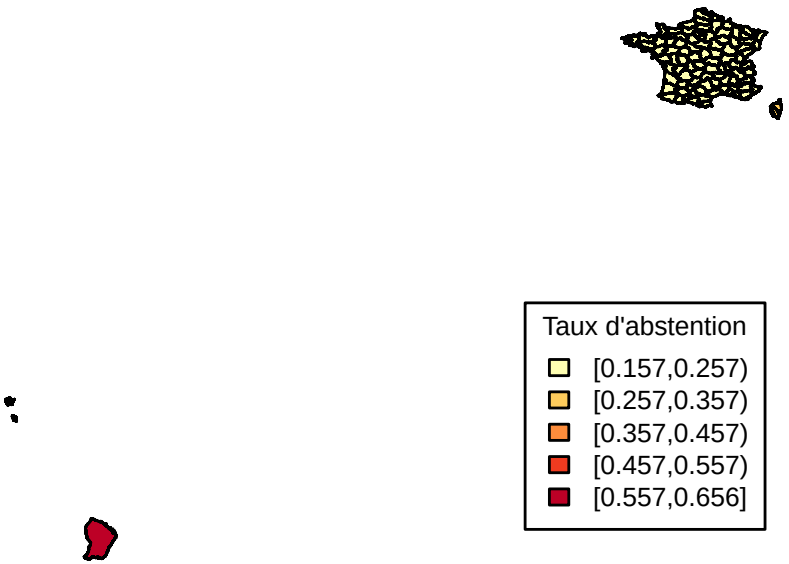
```
donnees=read_excel('V2017.xlsx',guess_max = Inf)
```

- Fichier un individu = un bureau de vote
- 69242 bureaux de vote et 98 variables

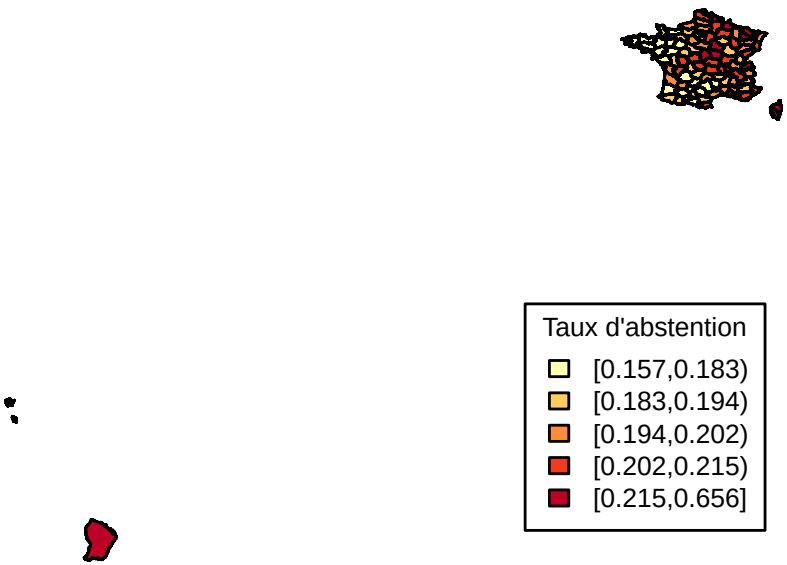
Voici un extrait du fichier

| | C-dpt | Dpt | Commune | Inscrits | Abstentions |
|---|---|---|---|---|---|
| 1 | 1 | Ain | L'Abergement-Clémenciat | 598 | 92 |
| 2 | 1 | Ain | L'Abergement-de-Varey | 209 | 25 |
| 3 | 1 | Ain | Ambérieu-en-Bugey | 1116 | 233 |
| 4 | 1 | Ain | Ambérieu-en-Bugey | 1128 | 256 |

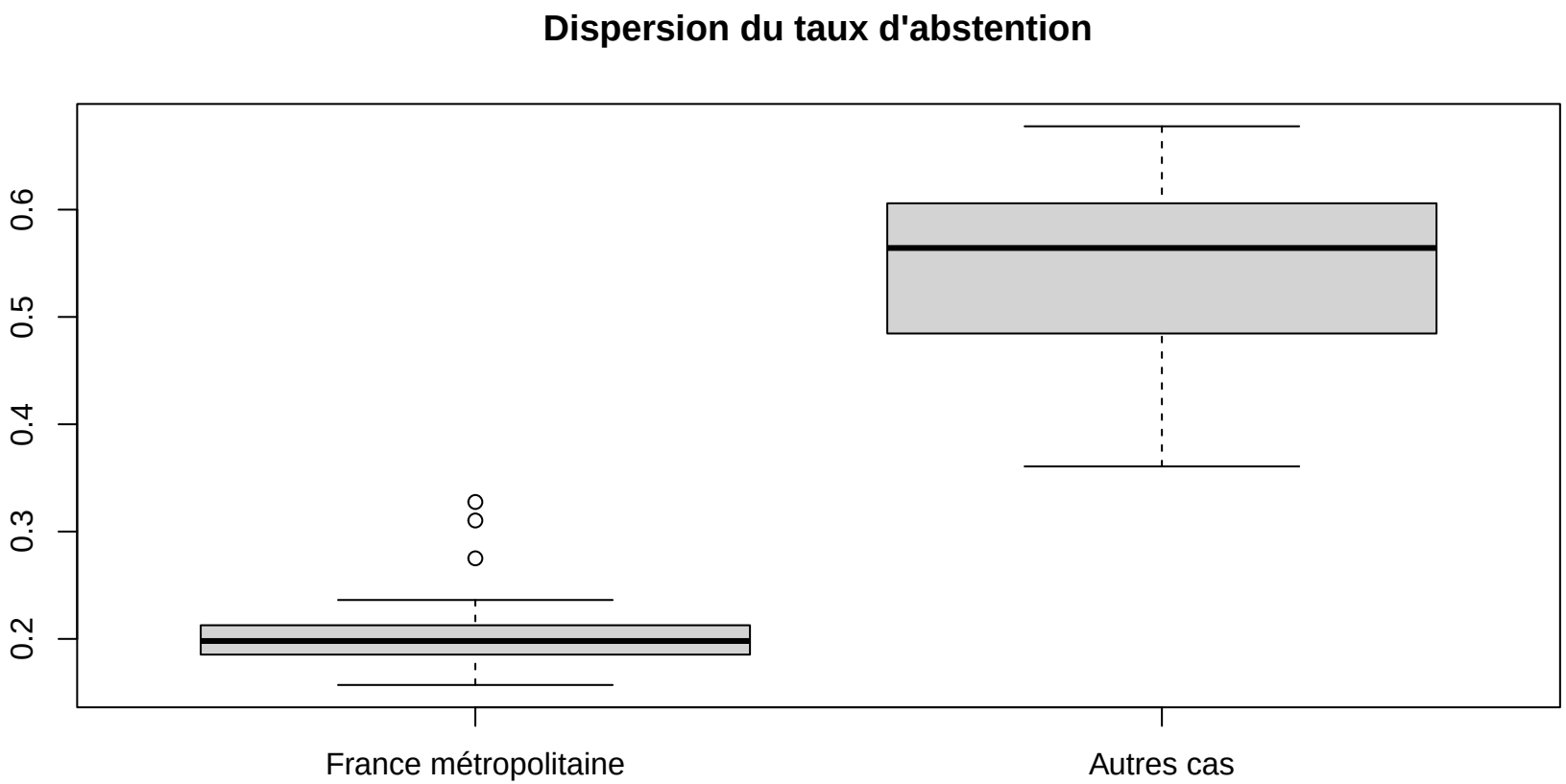
# Taux abstention (premier tour) 0.2223197 (population)



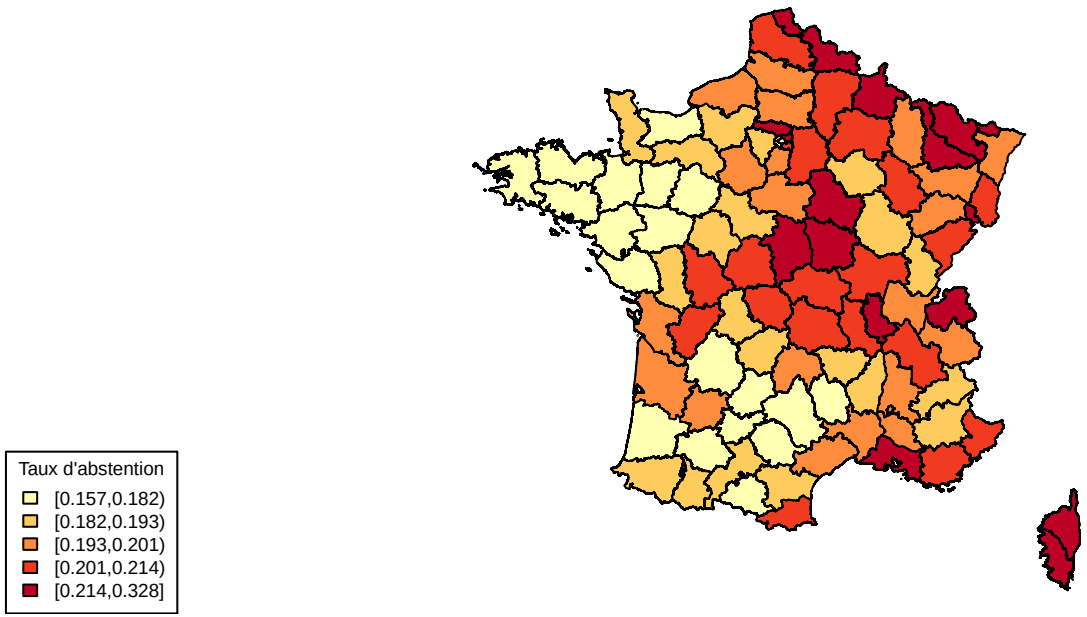
# Carte avec la méthode des quantiles pour séparer les classes



# Etude de la dispersion des données

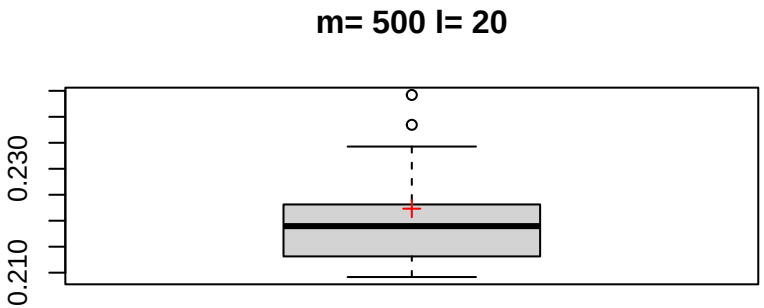


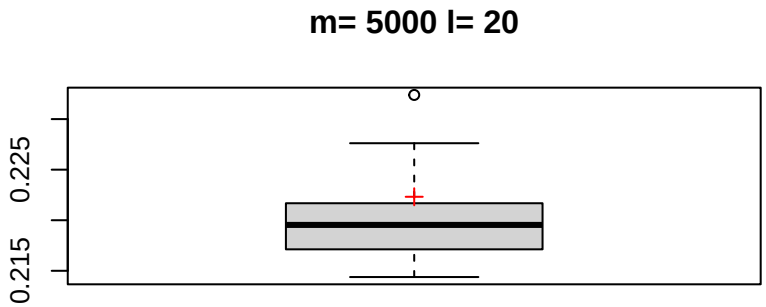
# Zoom sur la France métropolitaine

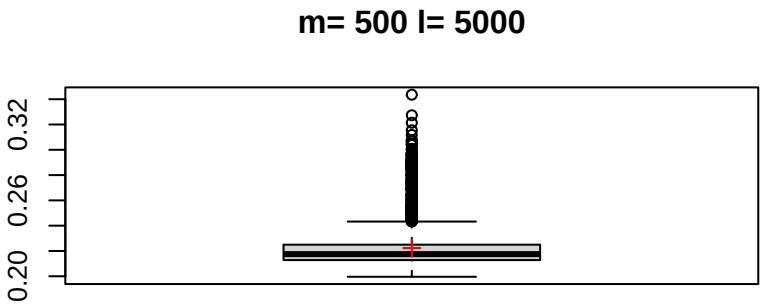


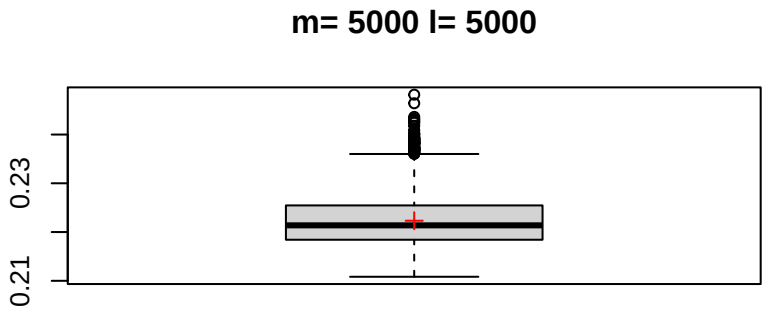

Notion de paramètre à estimer (difficulté des étudiants à comprendre qu'on est omniscient dans cette étude)

Echantillonnage "naïf" (toutes les données) : on tire au sort $m$ bureaux de vote et on étudie la variable abstention sur ces bureaux de vote, pour $l$ échantillons.



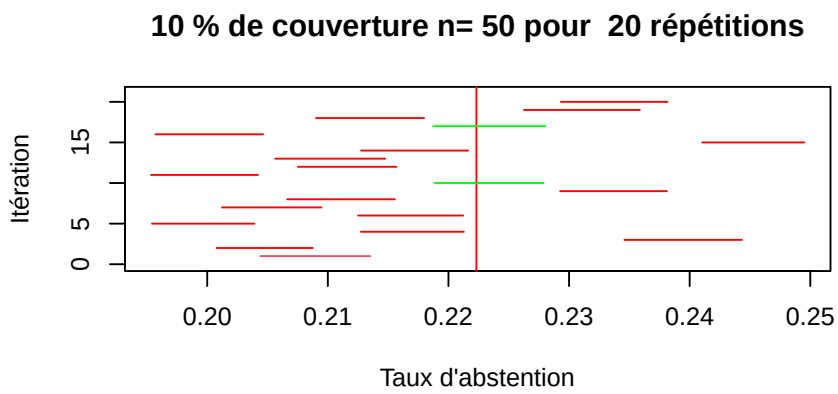



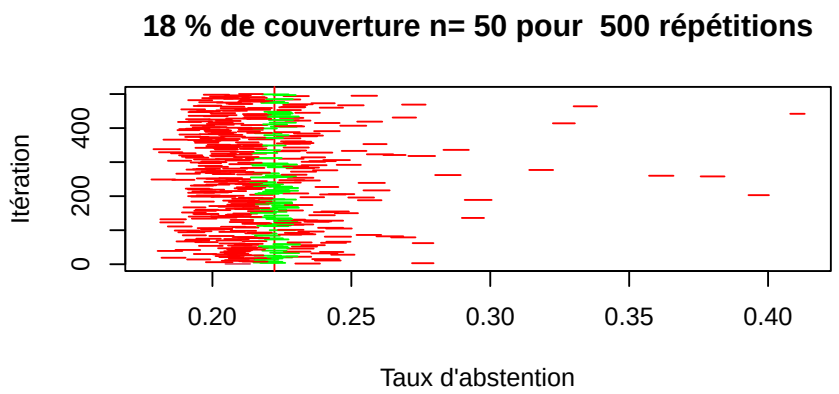



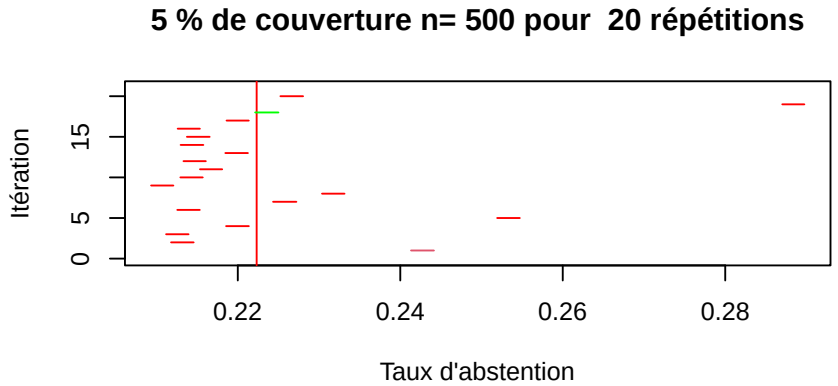
Echantillonnage "naïf" : question de la taille de l'échantillon

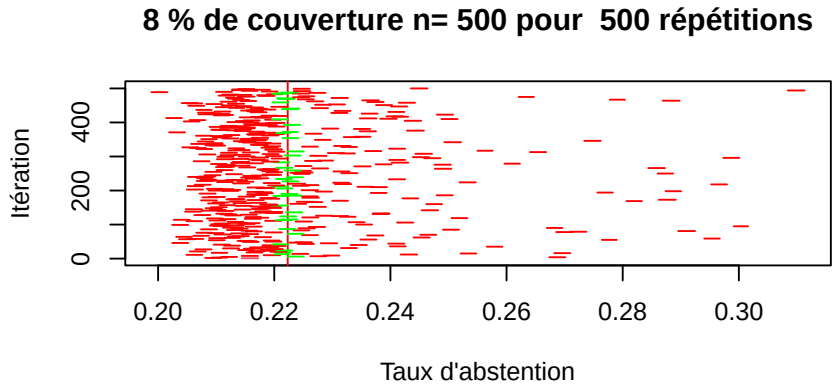
Problème posé avec étudiants : on effectue $l$ tirages de $m$ bureaux de vote. On ne cherche pas le taux d'abstention moyen par bureau de vote mais total.

Si on prend pour taille d'échantillon, le nb total d'inscrits du tirage, l'intervalle de confiance "classique" est mauvais (au passage, bien vérifier les conditions d'application)



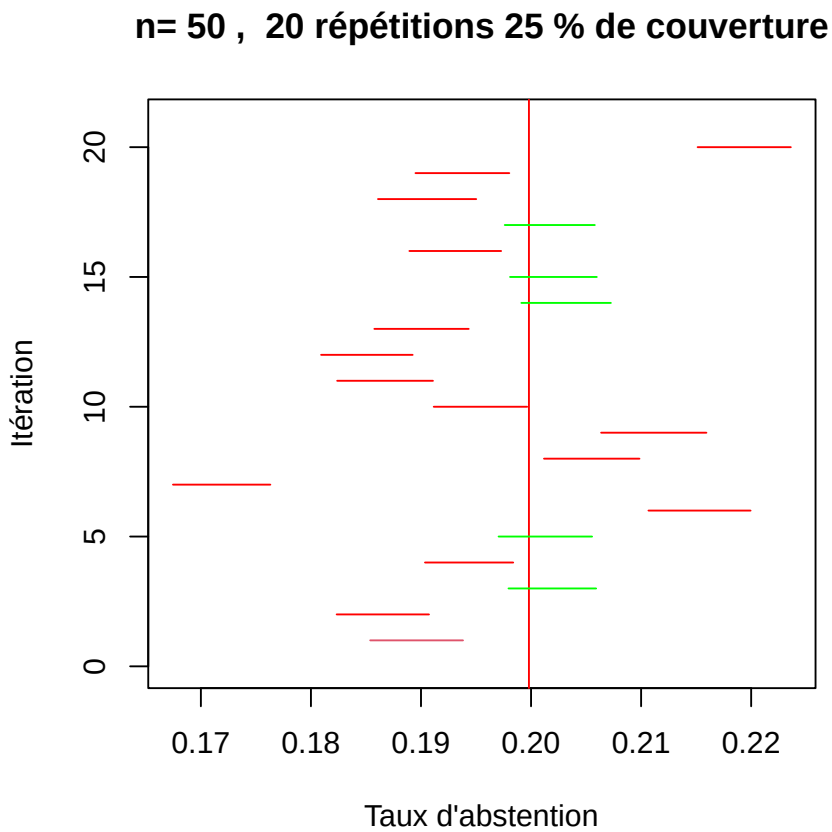



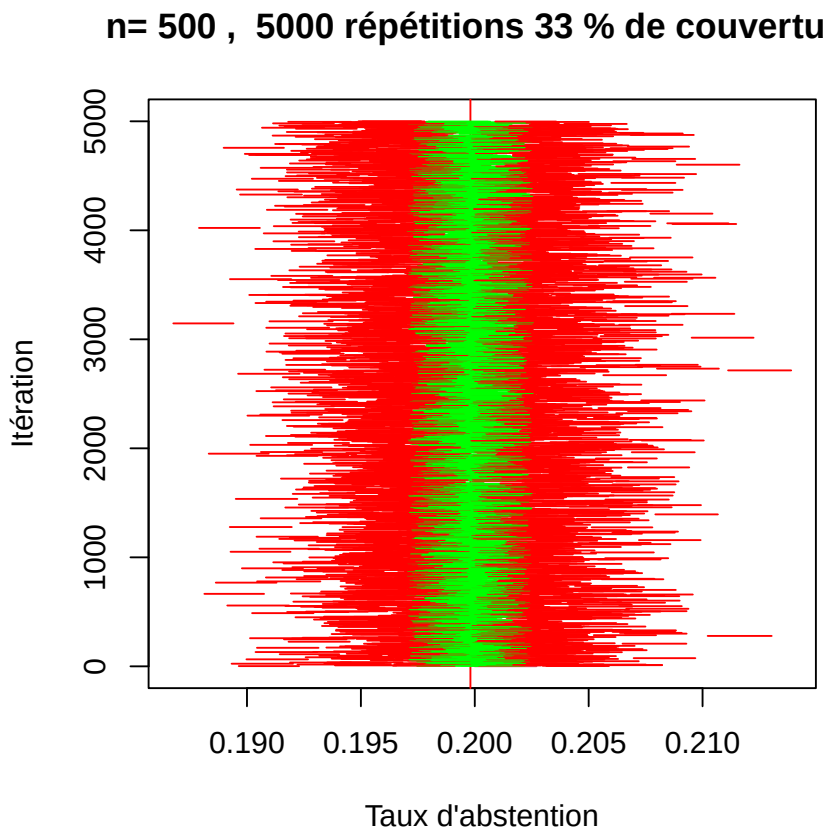




Echantillonnage ”naïf” (toutes les données) : intervalles de confiance, taux de couverture mauvais => échantillons peu représentatifs

# Echantillonnage ”naïf” : France métropolitaine uniquement



Données obtenues pour 50000 répétitions

| n | Population | Taux |
|---|---|---|
| 50 | Toute | 0.16414 |
| | France métropolitaine | 0.3235 |
| | Sans la Corse | 0.3226 |
| 500 | Toute | 0.07906 |
| | France métropolitaine | 0.32766 |
| | Sans la Corse | 0.33018 |
| 5000 | Toute | 0.05972 |
| | France métropolitaine | 0.33726 |
| | Sans la Corse | 0.34204 |

Problème : on va finir par ne tirer que les échantillons des bureaux de vote égaux au taux d'abstention demandés et outrepasser l'intérêt géographique pour un calcul pratique...

Pourquoi ça ne fonctionne pas ?

- Le tirage se fait par grappe (une grappe = un bureau de vote)
- Pour chaque grappe tirée au sort, on interroge tous les individus.
- Conséquence : les individus n'ont pas la même probabilité d'être dans l'échantillon (un individu issu d'un "petit" bureau de vote a plus de chance d'être tiré au sort)

Les données sont trop dispersées (beaucoup de petits bureaux de vote + disparités géographiques) pour obtenir une bonne estimation du taux d'abstention !...

- $N$ unités (=inscrits) répartis en $M$ grappes (= bureaux de vote, unité primaire)
- On tire au sort $m$ grappes que l'on sonde intégralement.
- Une grappe $i$ contient $N_i$ inscrits (unités secondaires, $N = \sum N_i$)
- Pour estimer $Y/N$ (taux d'abstention national), on considère

$$\hat{Y} = \frac{1}{m} \sum_{i=1}^{m} M \frac{Y_i}{N},$$

qui est le taux moyen d'abstentions par grappe multiplié par $M/N$ ($Y_i$ est le nombre d'abstentions de la grappe $i$), dont l'estimateur de la variance est

$$\hat{V}(\hat{Y}) = \frac{M-m}{Mm} \frac{1}{m-1} \sum_{i=1}^{m} \left( \frac{Y_i N}{M} - \hat{Y} \right)^2$$

Résultats avec estimation par grappe
Données obtenues pour 50000 répétitions

| $m$ | Taux |
|---|---|
| 50 | 0.87148 |
| 500 | 0.7507 |
| 5000 | 0.81918 |

Une réponse : vrai échantillonnage aléatoire simple
Algorithme naïf :

1. Je tire au sort $n$ valeurs entre 1 et $nb\_inscrits$
2. Je calcule les effectifs cumulés des inscrits
3. Je cherche à quel bureau de vote les tirages appartiennent et pour chaque bureau de vote, je détermine si l'individu s'abstient

Problème : Défilement du fichier de bureaux (trop long) et recherche "manuelle"
Une réponse possible : astuces de programmation

Algorithme amélioré (pas parfait !) :

1. Je tire au sort $n$ valeurs entre 1 et $nb\_inscrits$
2. Je calcule les effectifs cumulés des inscrits $inscrits\_cum$
3. J'intègre les valeurs tirées à $inscrits\_cum$ et à l'aide de match, je récupère les indices de leur emplacements, auxquels je retranche $0, 1, 2, ..., (n-1)$, ce qui me donne leurs bureaux de vote
4. Je détermine si l'individu s'abstient

Les résultats : enfin les bons !
50000 répétitions

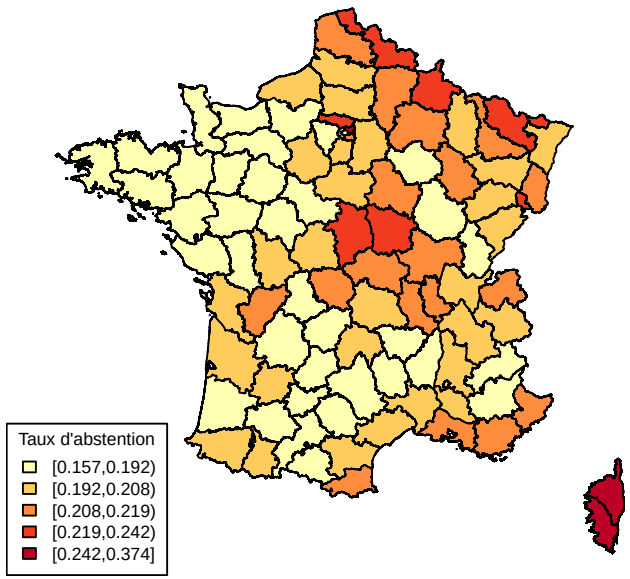
| $n$ | Taux |
|---|---|
| 50 | 0.9302 |
| 500 | 0.94588 |
| 5000 | 0.94864 |

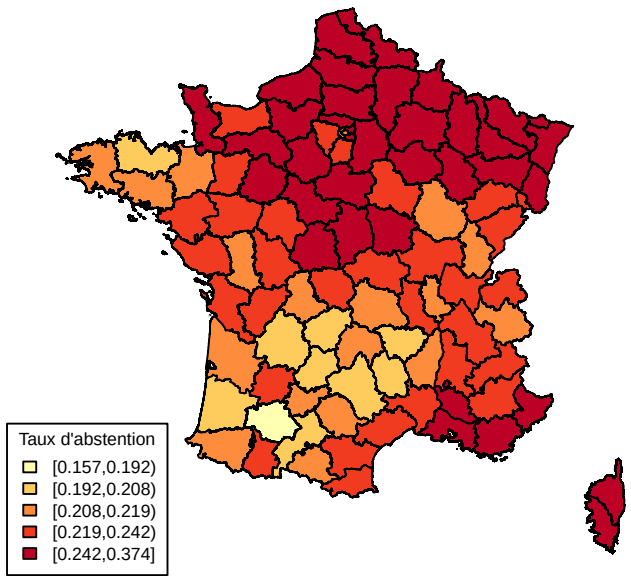
- Apport pédagogique : des erreurs, on tire de l'info (difficile à comprendre pour les étudiants)

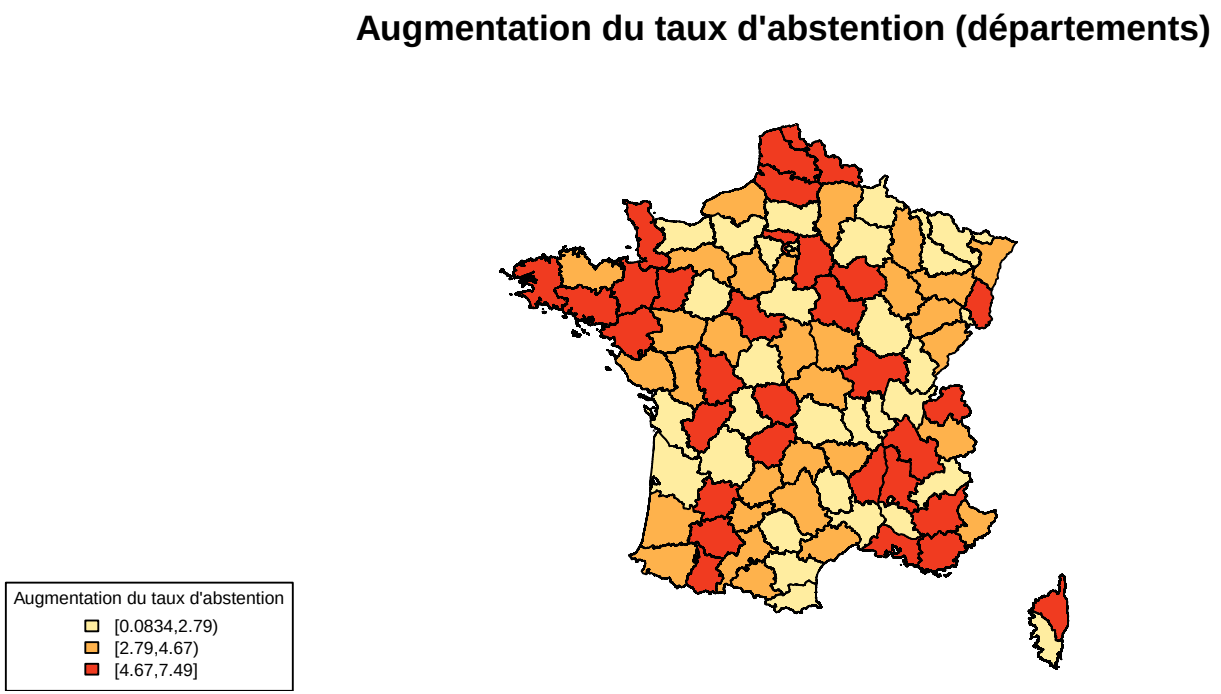
- Perspectives d'études :
  - Approche du sondage stratifiés par taille de commune ou avec une autre variable auxiliaire
  - Etude propre aux français de l'étranger
  - Etude des régions etc

# Etude en France métropolitaine uniquement

Vrai taux abstention : 0.2630713



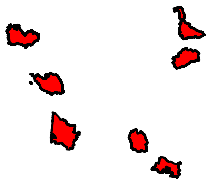






Etude sur toute la population

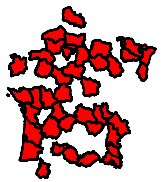
# Test de comparaison de moyenne à une valeur de référence pour chaque département (risque 5%)

**Augmentation du taux d'abstention, n= 500**

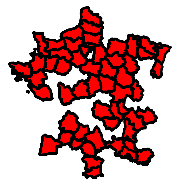


**Augmentation du taux d'abstention, n= 5000**

**Augmentation du taux d'abstention, n= 20000**



**Augmentation du taux d'abstention, n= 50000**



Conclusions personnelles et pédagogiques

- Principales difficultés pour l'étudiant : techniques
- Pour l'enseignant :
  Intérêt accru pour certains groupes : libres dans leur choix, discussions entre étudiants intéressantes
  D'un problème, on tire des informations sur les données : difficile à comprendre
- L'an prochain : même sujet, plus abouti (RShiny ?), mais approche de départ finalement identique.

Merci de votre attention !