Echantillonnage

Professeur **Francis GUILLEMIN**
> Ecole de santé publique - Faculté de Médecine

Nancy-Université
Université Henri Poincaré
ministère éducation nationale enseignement supérieur recherche

Plan

- Terminologie
- Méthodes de sondage
- Qualité des estimateurs

Comment dénombrer ?

- ***Question*** : combien y a-t-il de personnes atteintes de troubles de la vue parmi les conducteurs automobiles en France ?
- ***Réponse*** : 10% ? 40 % ? 75 % ?

- Il est impossible de les compter toutes en examinant toute la population des conducteurs français
- Il va être nécessaire d'utiliser une procédure particulière (l'échantillonnage) et des méthodes statistiques pour **estimer** la précision du résultat (incertitude)

Un peu de terminologie

- **Population** : Toutes les personnes à qui les résultats doivent s'appliquer

- **Echantillon** : Dans la plupart des cas, la taille de la population est trop importante pour que l'on puisse étudier tous les individus qui la compose. On étudie un sous-groupe appelé échantillon.

- **Unités** : il peut s'agir d'unité individuelle (sujet) ou collective (foyer, hôpitaux)

Un peu de terminologie

- **Phénomène d'intérêt** : c'est la caractéristique de santé qui fait l'objet de l'étude

- **Sondage** : toute forme d'échantillonnage qui permet de constituer un échantillon à partir de la population

- **Estimateur** : résultat estimé à partir des données observées dans l'échantillon qui représente la valeur vraie du phénomène dans la population, avec un certain degré d'incertitude

Différentes méthodes

- Sondage empirique
- Sondage aléatoire simple
- Sondage stratifié
- Sondage en grappe
- Sondage pseudo-aléatoire

Sondage empirique

- Constituer un échantillon de telle façon qu'un nombre fixe de personnes à enquêter soit atteint.
- On utilisera volontiers la méthode des quotas, indiquant à l'enquêteur de s'arrêter lorsqu'il a atteint le quota voulu dans chaque catégorie:
  - X hommes, Y femmes
  - Z_1 [18– 25 ans[, Z_2 [25 – 60 ans[, Z_3 [60 ans et +]
  - etc…

Sondages probabilistes

- Ensemble de méthodes appelées sondages probabilistes, parce que chaque unité échantillonnée a une probabilité connue à l'avance de figurer dans l'échantillon

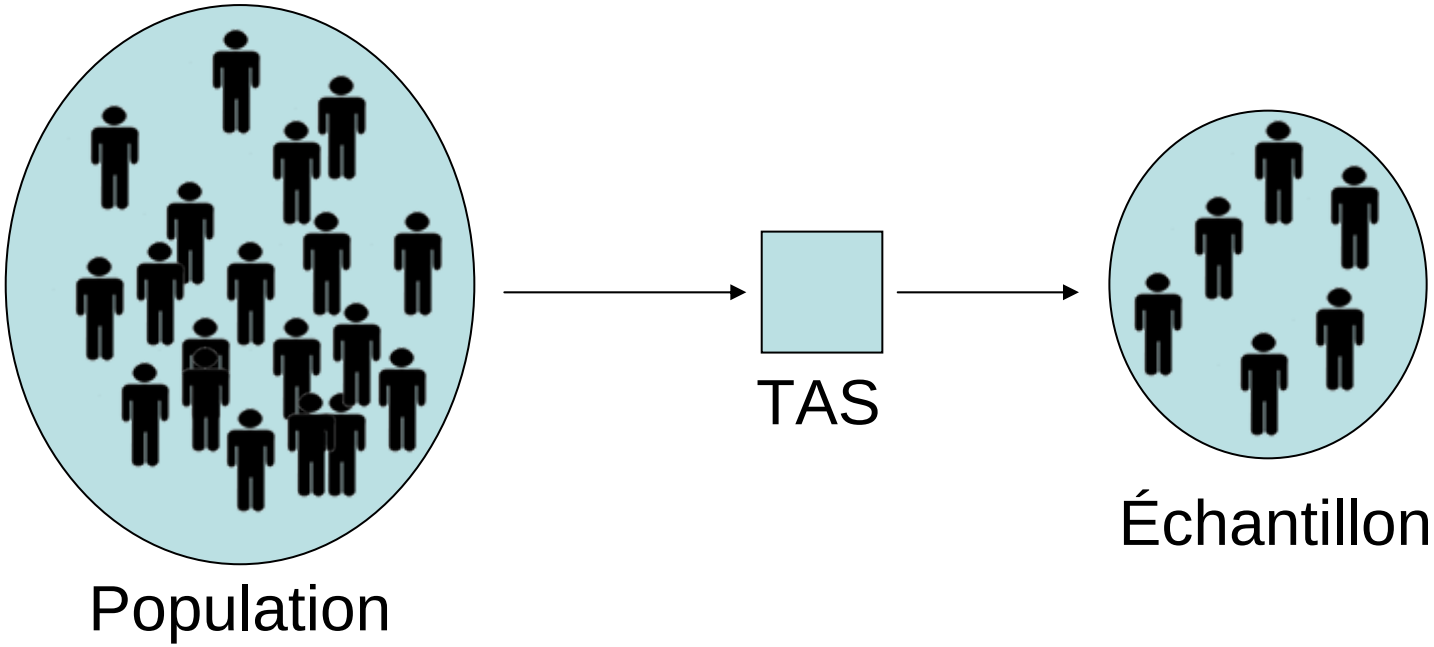
- Ceci permet
  - de généraliser l'estimation du phénomène à la population dont est issu l'échantillon
  - d'apprécier la marge d'erreur, le degré d'incertitude de l'estimateur

Sondage aléatoire simple

- Chaque sujet de la population a la même probabilité d'être inclus dans l'échantillon
- Maximise la possibilité de conclure pour toute la population
- Base de sondage : liste pré-établie des sujets
  - Liste des conducteurs
  - Liste des foyers
  - Liste des abonnés au téléphone
  - …

Sondage aléatoire simple

- Procéder à un tirage au sort des sujets dans la base :
  - Programme informatique
  - Tables de nombre au hasard



Sondage aléatoire simple

- Le sondage permet de limiter la taille de l'investigation
- Avantages :
  - Réduction des coûts d'investigation
  - Meilleure qualité de l'observation chez chaque sujet (enquête, questionnaire, investigation clinique)
  - Délai d'obtention des résultats plus rapide
- Limite :
  - il est nécessaire d'avoir une base de sondage fiable

Sondage stratifié

- Dans certains cas, on peut craindre d'obtenir trop peu de sujets d'un sous-groupe particulier (*p.ex. les conducteurs occasionnels*), alors qu'on peut supposer une fréquence particulière du phénomène dans ce sous-groupe.

- On risque que l'échantillon de ce sous-groupe de la population ne permette pas de calculer un estimateur suffisamment précis

- Par le simple fait du hasard, on peut sous-estimer ou sur-estimer la fréquence du phénomène dans ce sous-groupe

Sondage stratifié

- La méthode consiste à identifier les niveaux / catégories de la variable qui caractérise cet aspect de la population

- *exemple 1* : fréquence de la conduite
  - Quotidienne longs trajets
  - Quotidienne courts trajets
  - Occasionnelle

- Chaque catégorie définit une strate de la population

Sondage stratifié

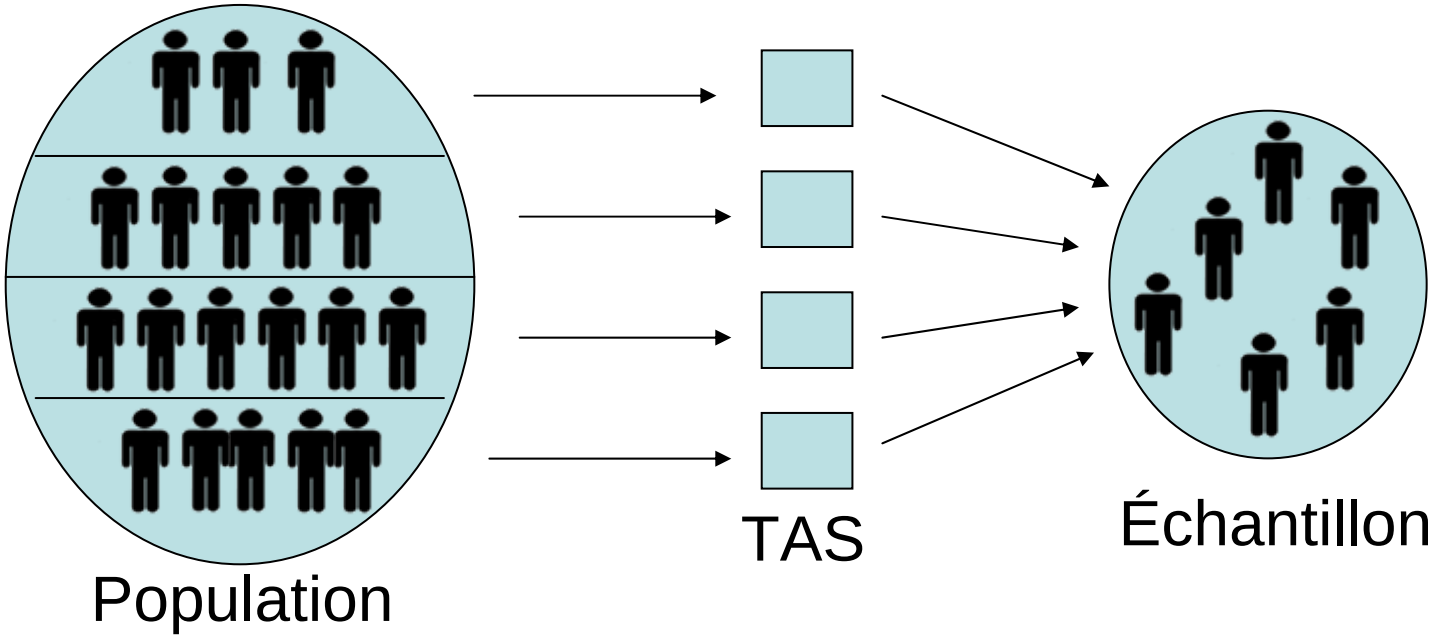
- La méthode consiste à identifier les niveaux / catégories de la variable qui caractérise cet aspect de la population

- *exemple 2* : on peut supposer que les personnes d'un même groupe partagent des caractéristiques qui déterminent plus particulièrement le phénomène
  - Les troubles de la vue peuvent comporter une composante d'origine génétique : daltonisme, myopie
  - Les personnes d'une même famille ont donc une probabilité différente d'une autre famille

- Chaque famille définit une strate de la population

Sondage stratifié

- L'échantillon est constitué par un sondage aléatoire simple par strate :
- Tirage au sort des unités dans chaque strate



Sondage stratifié

- Ainsi, connaissant le poids (la proportion) de chaque strate dans la population, on peut en tenir compte au moment du calcul des estimateurs

- Avantage : cette méthode permet d'améliorer la précision du sondage

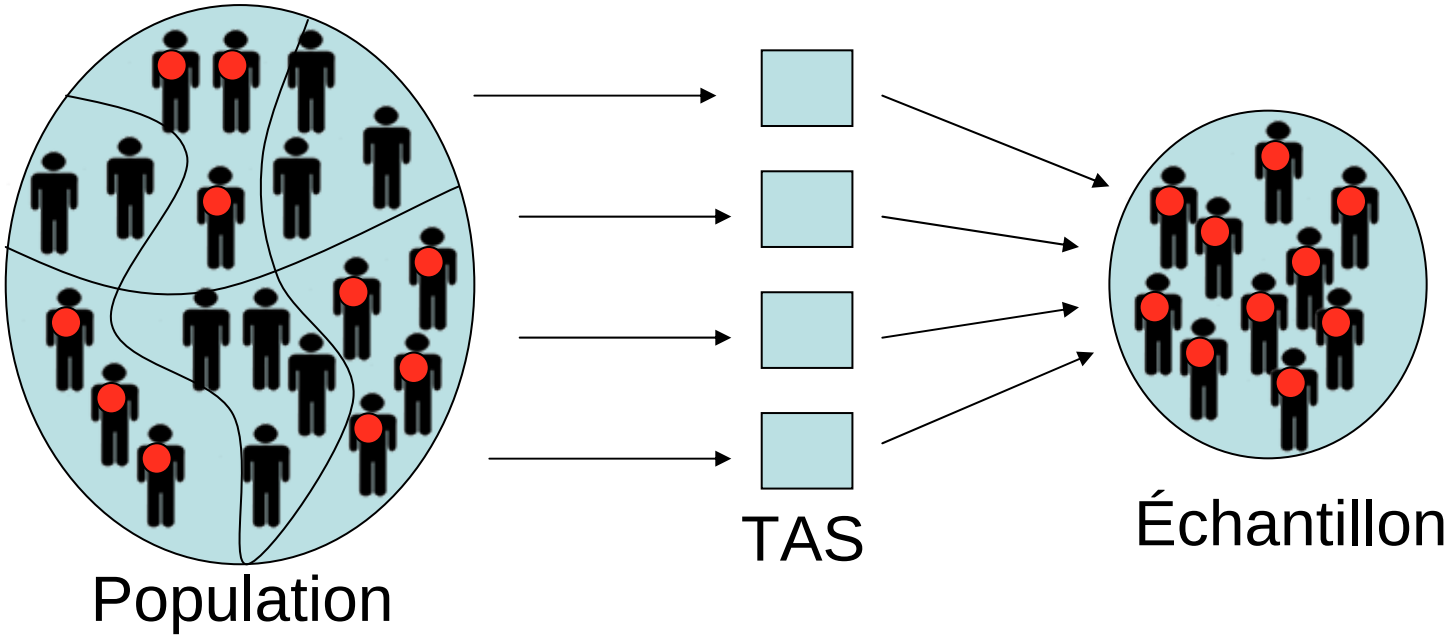
- Inconvénient : le calcul de l'estimateur est plus complexe

Sondage en grappe

- Dans certains cas, il est difficile d'obtenir un échantillon d'individus indépendants les uns des autres. Il peut être plus facile d'enquêter dans un lieu où ils sont rassemblés
- *Exemple* : les sujets d'un même foyer (résidence)
- Le sous-groupe de la population définit une grappe

Sondage en grappe

- Ce sont les grappes qui sont tirées au sort dans la population
- L'ensemble des sujets d'une grappe tirée au sort sera enquêté



Sondage en grappe

- Avantages :
  - il n’est pas nécessaire de disposer d’une base de sondage des individus, une liste des grappes suffit
- Inconvénients :
  - le sondage est moins précis que le sondage aléatoire simple
  - L’analyse doit prendre en compte l’effet grappe, ce qui est plus complexe

Sondage pseudo-aléatoire

- En l'absence de base de sondage, on peut prendre des méthodes d'allure organisée, sur un caractère supposé indépendant du phénomène étudié, mais qui ne garantissent pas un vrai tirage au sort

- Ces méthodes ne garantissent pas la représentativité comme le ferait un véritable tirage au sort

Sondage pseudo-aléatoire

- Méthode systématique : les conducteurs qui franchissent un carrefour, qui se garent sur un parking

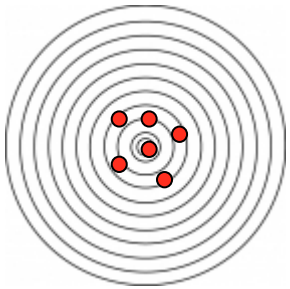
- Méthode dite « des itinéraires » : les conducteurs de telle maison, puis telle autre plus loin.

La qualité de l’estimation

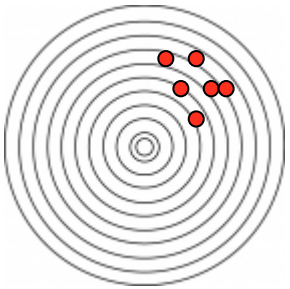
- La qualité d'une estimation repose sur sa précision et sur l'absence de biais.
- La représentativité de l’échantillon est la qualité garantie par une estimation sans biais.
- La précision n’est jamais parfaite et se traduit par une incertitude sur la valeur de l’estimateur

La qualité de l'estimation

- une estimation sans biais est obtenue au mieux par les méthodes de sondage aléatoire



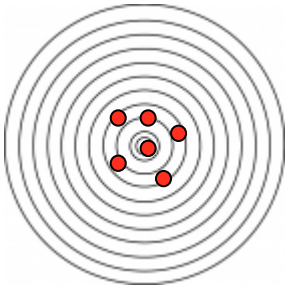
estimation
non biaisée

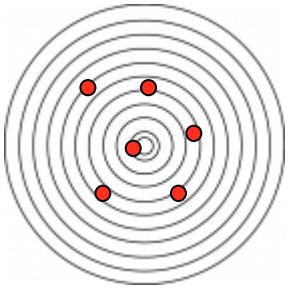


estimation
biaisée

La qualité de l'estimation

- La précision d'une estimation dépend du degré d'erreur de la méthode de mesure

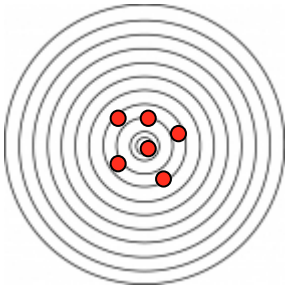
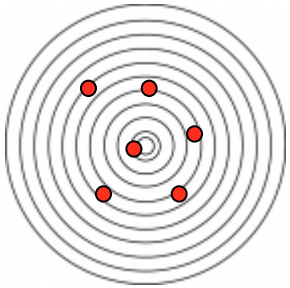
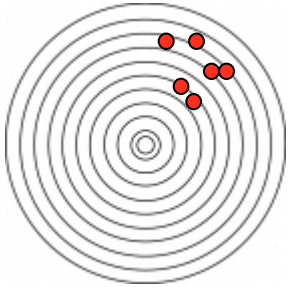
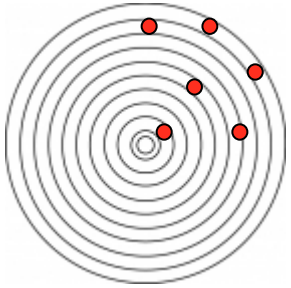




- L'incertitude sur la valeur de l'estimateur est exprimée par son intervalle de confiance
- L'incertitude diminue lorsque la taille de l'échantillon augmente

La qualité de l’estimation

- Elle dépend donc de la méthode d’échantillonnage choisie et de la taille de l’échantillon

Nancy-Université
Université Henri Poincaré

ministère éducation nationale enseignement supérieur recherche