Séance 7 : Construire un échantillon

Représentativité, Quotas et Enquêtes par traces

Mattéo Lanoë

Printemps 2026

Où en sommes-nous ?

Le cycle de l'enquête quantitative :

- Nous avons vu comment **formuler** les questions (Séance 6).
- Aujourd'hui : **À qui** allons-nous poser ces questions ?

Où en sommes-nous ?

Le cycle de l'enquête quantitative :

- Nous avons vu comment **formuler** les questions (Séance 6).
- Aujourd'hui : **À qui** allons-nous poser ces questions ?

L'objectif du jour : L'échantillonnage

Comprendre comment on sélectionne les individus à interroger pour s'assurer que leurs réponses reflètent bien la réalité de l'ensemble de la population.

Un cas historique : L'élection américaine de 1936

Le contexte : États-Unis, 1936. L'élection présidentielle oppose Franklin D. Roosevelt à Alfred Landon.

Le défi

Prédire le résultat de l'élection. Deux acteurs s'affrontent sur la méthode pour construire leur échantillon :

- Le célèbre magazine ***Literary Digest*** (qui avait réussi ses prédictions aux élections précédentes).
- Un pionnier des sondages : **George Gallup**.

Deux méthodes d'échantillonnage opposées

1. Le *Literary Digest*

La force du nombre :

- Envoi de **10 millions** de bulletins de vote factices.
- **Cibles :** Leurs abonnés, les propriétaires de voitures et les personnes dans l'annuaire téléphonique.
- **Recrutement volontaire** (les gens choisissent de renvoyer le bulletin).

Deux méthodes d'échantillonnage opposées

1. Le *Literary Digest*

La force du nombre :

- Envoi de **10 millions** de bulletins de vote factices.
- **Cibles :** Leurs abonnés, les propriétaires de voitures et les personnes dans l'annuaire téléphonique.
- **Recrutement volontaire** (les gens choisissent de renvoyer le bulletin).

2. George Gallup

La force de la méthode :

- Un échantillon **beaucoup plus restreint** (quelques milliers de personnes).
- **Cibles :** Des individus choisis de manière **aléatoire**.
- Le recrutement ne repose pas sur le seul volontariat spontané.

Qui a eu raison ?

D'après vous, quelle méthode a été la plus efficace ?

Qui a eu raison ?

D'après vous, quelle méthode a été la plus efficace ?

Le double succès de Gallup

Malgré un nombre gigantesque de réponses, le *Literary Digest* a annoncé la victoire de Landon.
Gallup a eu raison sur toute la ligne : il a **prédit la victoire de Roosevelt**, et a même prédit l'erreur du magazine !

Qui a eu raison ?

D'après vous, quelle méthode a été la plus efficace ?

Le double succès de Gallup

Malgré un nombre gigantesque de réponses, le *Literary Digest* a annoncé la victoire de Landon.
Gallup a eu raison sur toute la ligne : il a **prédit la victoire de Roosevelt**, et a même prédit l'erreur du magazine !

Pourquoi une telle erreur du magazine ?
Contacter ses abonnés et les propriétaires de voitures en 1936 (en pleine Grande Dépression), c'est interroger une population très spécifique (plus aisée) qui ne vote pas comme le reste du pays.

La leçon de 1936 : La représentativité

Cet échec historique soulève un point crucial pour évaluer la qualité d'une enquête.

La leçon de 1936 : La représentativité

Cet échec historique soulève un point crucial pour évaluer la qualité d'une enquête.

Le Volume (La taille)
Avoir 10 millions de réponses ne garantit pas la vérité.

La leçon de 1936 : La représentativité

Cet échec historique soulève un point crucial pour évaluer la qualité d'une enquête.

Le Volume (La taille)
Avoir 10 millions de réponses ne garantit pas la vérité.

$\neq$

La Représentativité
L'échantillon doit être un miroir fidèle de la population étudiée.

La leçon de 1936 : La représentativité

Cet échec historique soulève un point crucial pour évaluer la qualité d'une enquête.

Le Volume (La taille)
Avoir 10 millions de réponses ne garantit pas la vérité.

$\neq$

La Représentativité
L'échantillon doit être un miroir fidèle de la population étudiée.

$\rightarrow$ *C'est la différence entre un "bon" et un "mauvais" échantillon, et c'est ce que nous allons voir avec les concepts fondamentaux.*

Les concepts fondamentaux

Avant de commencer, il faut distinguer deux niveaux :

Les concepts fondamentaux

Avant de commencer, il faut distinguer deux niveaux :

La population de référence (ou population cible)

C'est l'ensemble total des individus que l'on souhaite étudier (ex : Tous les étudiants de Sciences Po, tous les électeurs français).

Les concepts fondamentaux

Avant de commencer, il faut distinguer deux niveaux :

La population de référence (ou population cible)

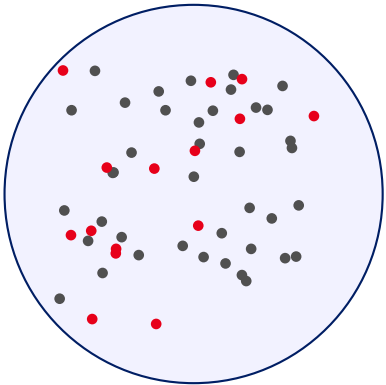
C'est l'ensemble total des individus que l'on souhaite étudier (ex : Tous les étudiants de Sciences Po, tous les électeurs français).

L'échantillon

C'est le sous-ensemble de cette population qui sera **effectivement interrogé**.
L'objectif est qu'il soit une "miniature fidèle" de la population de référence.

De la population à l'échantillon

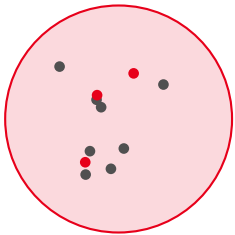
Population de référence



Ex : 50 millions d'électeurs

Échantillonnage

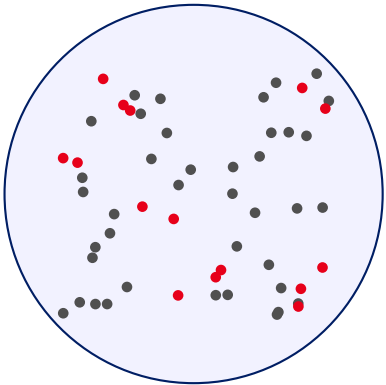
Échantillon



Ex : 1 000 personnes

De la population à l'échantillon

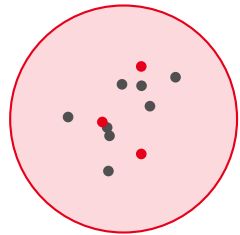
Population de référence



Ex : 50 millions d'électeurs

Échantillonnage

Échantillon



Ex : 1 000 personnes

Le défi : Comment s'assurer que les points rouges et gris sont dans les mêmes proportions à gauche et à droite ? C'est le problème de la **représentativité**.

La condition indispensable : La base de sondage

Pour tirer au sort de manière rigoureuse, il faut disposer d'une **base de sondage**.

La condition indispensable : La base de sondage

Pour tirer au sort de manière rigoureuse, il faut disposer d'une **base de sondage**.

Qu'est-ce que c'est ?

C'est la liste exhaustive et nominative de tous les membres de la population de référence (ex : Listes électorales, registres d'inscription de l'université, annuaire).

La condition indispensable : La base de sondage

Pour tirer au sort de manière rigoureuse, il faut disposer d'une **base de sondage**.

Qu'est-ce que c'est ?

C'est la liste exhaustive et nominative de tous les membres de la population de référence (ex : Listes électorales, registres d'inscription de l'université, annuaire).

Le problème en sciences sociales :

- Cette liste existe rarement de manière parfaite.
- Elle est souvent obsolète, incomplète, ou inaccessible pour des raisons de confidentialité.

Les deux grandes familles d'échantillonnage

Selon que l'on possède ou non une base de sondage, on utilise des méthodes différentes :

Les deux grandes familles d'échantillonnage

Selon que l'on possède ou non une base de sondage, on utilise des méthodes différentes :

1. Méthodes Probabilistes (Aléatoires)

- Nécessitent une **base de sondage**.
- Chaque individu a une probabilité **connue** (souvent égale) d'être tiré au sort.
- C'est le "pile ou face" statistique, la méthode la plus pure scientifiquement.

Les deux grandes familles d'échantillonnage

Selon que l'on possède ou non une base de sondage, on utilise des méthodes différentes :

1. Méthodes Probabilistes (Aléatoires)

- Nécessitent une **base de sondage**.
- Chaque individu a une probabilité **connue** (souvent égale) d'être tiré au sort.
- C'est le "pile ou face" statistique, la méthode la plus pure scientifiquement.

2. Méthodes Empiriques

- Utilisées quand il n'**y a pas de base de sondage**.
- Le chercheur (ou l'enquêteur) choisit les individus selon des règles précises pour "imiter" le hasard.
- Ex : La méthode des quotas.

La méthode empirique reine : Les quotas

C'est la méthode utilisée par **tous les instituts de sondage** en France.

Le principe : Puisqu'on ne peut pas tirer au sort, on force l'échantillon à avoir **exactement la même structure** que la population totale sur des critères sociodémographiques connus (grâce au recensement de l'INSEE).

La méthode empirique reine : Les quotas

C'est la méthode utilisée par **tous les instituts de sondage** en France.

Le principe : Puisqu'on ne peut pas tirer au sort, on force l'échantillon à avoir **exactement la même structure** que la population totale sur des critères sociodémographiques connus (grâce au recensement de l'INSEE).

Critère (ex : Sexe)	**Population Française**	**Notre Échantillon (N=1000)**
Femmes	51,5 %	**515** femmes à interroger
Hommes	48,5 %	**485** hommes à interroger

On croise souvent ces quotas : Sexe $\times$ Âge $\times$ Profession (ex : il me faut "3 femmes, ouvrières, de moins de 30 ans").

Application 1 : Enquêter sans base de sondage

Texte de référence : J.-M. Firdion, *Construire un échantillon* (L'enquête sur les jeunes sans-abri).

Le problème méthodologique

Comment faire un échantillon représentatif de "jeunes en errance" ?

- Il n'y a **pas de liste** (pas de base de sondage).
- Il n'y a **pas de statistiques INSEE** sur eux (impossible de faire des quotas stricts).

Application 1 : Enquêter sans base de sondage

Texte de référence : J.-M. Firdion, *Construire un échantillon* (L'enquête sur les jeunes sans-abri).

Le problème méthodologique

Comment faire un échantillon représentatif de "jeunes en errance" ?

- Il n'y a **pas de liste** (pas de base de sondage).
- Il n'y a **pas de statistiques INSEE** sur eux (impossible de faire des quotas stricts).

La solution trouvée (Échantillonnage sur les lieux de passage) : Tirer au sort des lieux d'aide (foyers, centres de distribution de repas, centres d'hébergement) et y interroger les usagers.

Les limites de l'échantillonnage de Firdion

Est-ce que cet échantillon est parfaitement représentatif ? **Non.**

Les limites de l'échantillonnage de Firdion

Est-ce que cet échantillon est parfaitement représentatif ? **Non.**

Le biais de sélection par l'institution

En passant par les foyers et les centres d'aide, les chercheurs écartent d'office la frange la plus désocialisée des sans-abri : ceux qui ne fréquentent plus **aucune** institution et vivent de manière totalement isolée dans la rue ou les squats.

Les limites de l'échantillonnage de Firdion

Est-ce que cet échantillon est parfaitement représentatif ? **Non.**

Le biais de sélection par l'institution

En passant par les foyers et les centres d'aide, les chercheurs écartent d'office la frange la plus désocialisée des sans-abri : ceux qui ne fréquentent plus **aucune** institution et vivent de manière totalement isolée dans la rue ou les squats.

→ *Leçon méthodologique :* La manière dont on recrute l'échantillon définit les limites des conclusions que l'on peut en tirer. Les résultats de Firdion sont valables pour "les jeunes sans-abri **usagers des services d'aide**".

Une nouvelle ère : Les enquêtes par traces (Big Data)

Aujourd'hui, avec le web, les chercheurs disposent de jeux de données massifs produits automatiquement : **les données par traces**.

Une nouvelle ère : Les enquêtes par traces (Big Data)

Aujourd'hui, avec le web, les chercheurs disposent de jeux de données massifs produits automatiquement : **les données par traces**.

- **Qu'est-ce que c'est ?** Des données laissées par les utilisateurs lors de leurs activités numériques (achats, réseaux sociaux, clics, sites de rencontres).

Une nouvelle ère : Les enquêtes par traces (Big Data)

Aujourd'hui, avec le web, les chercheurs disposent de jeux de données massifs produits automatiquement : **les données par traces**.

- **Qu'est-ce que c'est ?** Des données laissées par les utilisateurs lors de leurs activités numériques (achats, réseaux sociaux, clics, sites de rencontres).
- **L'avantage :** Des échantillons gigantesques (le "N" est énorme, souvent plusieurs centaines de milliers).

Une nouvelle ère : Les enquêtes par traces (Big Data)

Aujourd'hui, avec le web, les chercheurs disposent de jeux de données massifs produits automatiquement : **les données par traces**.

- **Qu'est-ce que c'est ?** Des données laissées par les utilisateurs lors de leurs activités numériques (achats, réseaux sociaux, clics, sites de rencontres).
- **L'avantage :** Des échantillons gigantesques (le "N" est énorme, souvent plusieurs centaines de milliers).
- **L'illusion :** Penser que "beaucoup de données" (Big Data) équivaut automatiquement à "données représentatives".

Application 2 : Le cas de Meetic

Texte de référence : M. Bergström (2018), *De quoi l'écart d'âge est-il le nombre ?*

La base de données étudiée

- Utilisateurs actifs du site Meetic.
- **N = 119 501 individus.**
- Suivi de leurs interactions (qui contacte qui ?).

Avec près de 120 000 personnes, peut-on dire que l'étude de Bergström reflète les pratiques de couple de **tous** les Français ?

Le mythe du "N = Tout" (Big Data vs Représentativité)

Bergström montre que **la taille de l'échantillon ne garantit pas la représentativité**. La population de Meetic a des caractéristiques très spécifiques par rapport à la population générale (INSEE) :

Le mythe du "N = Tout" (Big Data vs Représentativité)

Bergström montre que **la taille de l'échantillon ne garantit pas la représentativité**. La population de Meetic a des caractéristiques très spécifiques par rapport à la population générale (INSEE) :

- **Géographie :** Surreprésentation massive des urbains et des habitants de l'Île-de-France (au détriment des zones rurales).

Le mythe du "N = Tout" (Big Data vs Représentativité)

Bergström montre que **la taille de l'échantillon ne garantit pas la représentativité**. La population de Meetic a des caractéristiques très spécifiques par rapport à la population générale (INSEE) :

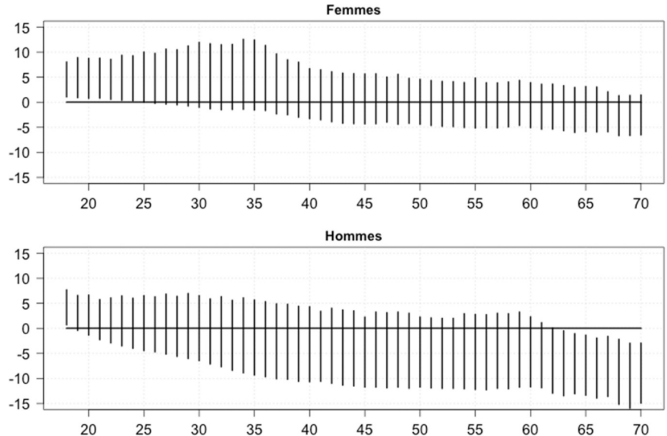
- **Géographie :** Surreprésentation massive des urbains et des habitants de l'Île-de-France (au détriment des zones rurales).
- **Âge :** Sous-représentation des plus jeunes et des plus âgés. C'est le site des trentenaires/quadragénaires en "seconde formation de couple" (après une première séparation).

Le mythe du "N = Tout" (Big Data vs Représentativité)

Bergström montre que **la taille de l'échantillon ne garantit pas la représentativité**. La population de Meetic a des caractéristiques très spécifiques par rapport à la population générale (INSEE) :

- **Géographie :** Surreprésentation massive des urbains et des habitants de l'Île-de-France (au détriment des zones rurales).
- **Âge :** Sous-représentation des plus jeunes et des plus âgés. C'est le site des trentenaires/quadragénaires en "seconde formation de couple" (après une première séparation).
- **Diplôme / PCS :** Probable surreprésentation des classes moyennes et supérieures urbaines.

FIGURE 3. – ***Préférences d'âge déclarées dans le profil par sexe et âge : intervalles entre âge minimum moyen et âge maximum moyen, rapportés à l'âge d'ego***



Note : Les barres indiquent l'intervalle entre l'*âge minimum moyen* et l'*âge maximum moyen*, par sexe et à chaque âge. Afin de faciliter la comparaison, les intervalles sont indiqués sous forme de rapport à l'âge propre (0), indiquant une préférence pour un partenaire plus âgé (+) ou plus jeune (–).
Lecture : Les profils présentant une femme de 20 ans affichent comme préférence un âge minimum de 20,8 ans en moyenne, et un âge maximum de 29,5 en moyenne.
Champ : Comptes d'utilisateurs actifs inscrits sur Meetic en 2014.
Source : Base d'utilisateurs de *Meetic.fr*, Meetic Group, 2014.

Que retenir de l'étude par traces ?

Le biais de la plateforme

Chaque site internet attire une sociologie particulière. Les utilisateurs de Meetic ne sont pas ceux de Tinder, qui ne sont pas ceux de Facebook, etc.

Que retenir de l'étude par traces ?

Le biais de la plateforme

Chaque site internet attire une sociologie particulière. Les utilisateurs de Meetic ne sont pas ceux de Tinder, qui ne sont pas ceux de Facebook, etc.

Comment le chercheur doit-il réagir ? Il ne s'agit pas de jeter ces données (elles sont très riches !), mais de **circonscrire ses conclusions**. Bergström ne prétend pas analyser la formation des couples en France, mais la formation des couples *chez les urbains connectés d'âge intermédiaire*.

Bilan de la séance

Les règles d'or de l'échantillonnage :

1. **Définir** précisément sa population de référence.

Bilan de la séance

Les règles d'or de l'échantillonnage :

1. **Définir** précisément sa population de référence.
2. **Vérifier** l'existence d'une base de sondage (si oui $\rightarrow$ méthode aléatoire ; si non $\rightarrow$ méthode des quotas ou autres empiriques).

Bilan de la séance

Les règles d'or de l'échantillonnage :

1. **Définir** précisément sa population de référence.
2. **Vérifier** l'existence d'une base de sondage (si oui $\rightarrow$ méthode aléatoire ; si non $\rightarrow$ méthode des quotas ou autres empiriques).
3. **Ne jamais confondre "Taille" et "Représentativité"** : Un mauvais échantillon de 100 000 personnes est moins fiable qu'un bon échantillon de 1 000 personnes tiré rigoureusement.

Bilan de la séance

Les règles d'or de l'échantillonnage :

1. **Définir** précisément sa population de référence.
2. **Vérifier** l'existence d'une base de sondage (si oui $\rightarrow$ méthode aléatoire ; si non $\rightarrow$ méthode des quotas ou autres empiriques).
3. **Ne jamais confondre "Taille" et "Représentativité"** : Un mauvais échantillon de 100 000 personnes est moins fiable qu'un bon échantillon de 1 000 personnes tiré rigoureusement.
4. **Toujours identifier le biais de sélection** pour ne pas faire dire à ses données ce qu'elles ne peuvent pas dire.