Données et statistiques descriptives en sciences sociales

Ahmed Fouad EL HADDAD

UPEC - IEP de Fontainebleau

October 6, 2025

Plan de la séance

1. La démarche méthodologique en science politique
2. Méthodes qualitatives et quantitatives
3. Pourquoi enquêter et mesurer ?
4. Qu'est-ce qu'une donnée ?
5. Données individuelles vs. agrégées
6. Erreur écologique
7. Statistiques et types de données
8. Recueil des données
9. Sondages et échantillonnage
10. Biais et incertitudes
11. Statistiques descriptives
    - Tendance centrale
    - Mesures de variabilité

Les étapes de la recherche

Processus général de la recherche empirique en sciences sociales :

1. Définir et délimiter lobjet de recherche.
2. Identifier les objectifs à atteindre.
3. Poser une question de recherche claire et précise.
4. Effectuer une revue de la littérature existante.
5. Formuler des hypothèses à tester.
6. Observer et recueillir des données empiriques.
7. Choisir le type de données le plus pertinent.

La démarche scientifique appliquée

Trois grandes étapes :

- **Classer et ordonner :** établir des catégories, créer des tableaux croisés, typologies.
- **Établir des relations empiriques :** corrélations, causalité, modèles explicatifs.
- **Comparer :** logique small-N (études de cas) vs large-N (analyses statistiques).

But final : interpréter les données, proposer des conclusions, contribuer à la connaissance scientifique.

Méthodes qualitatives vs quantitatives

Tableau comparatif :

Qualitatives	**Quantitatives**
Entretiens, archives, observations	Statistiques, indicateurs numériques
Nombre de cas : petit	Nombre de cas : grand
Analyse contextualisée, par cas	Analyse par variables, abstraction hors contexte
Niveau dabstraction : faible	Niveau dabstraction : élevé
Singularité, complexité	Généralisation et comparabilité
Limites : faible montée en généralité	Limites : absence de contexte, interprétation discutable

Objectifs des deux approches

Qualitatives :

- Tester et nuancer des théories.
- Identifier des contre-exemples ou cas déviants.
- Appréhender la singularité des phénomènes sociaux.

Quantitatives :

- Confirmer des théories par des tests robustes.
- Produire des résultats synthétiques et généralisables.
- Mesurer la force et la direction des relations entre variables.

Pourquoi mesurer en sciences sociales ?

- Pour appréhender la complexité des phénomènes sociaux.
- Pour découvrir des causes cachées.
- Pour démêler les liens entre individus et groupes.
- Pour donner une assise empirique aux débats théoriques.

Exemples de questions :

- Comment définir la participation politique ?
- Est-ce la fin du clivage gauche/droite ?
- Comment mesurer le bien-être objectif et subjectif ?

Qu'est-ce qu'une donnée ?

- **Définition :** une donnée désigne des faits, mesures ou observations collectés.
- **Finalité :** utilisée pour décrire, analyser et interpréter les phénomènes sociaux.

Exemples

- Nombre de propositions de loi déposées par député.
- Taux de participation électorale par circonscription.
- Part des femmes au sein du parlement (%).

Représentation des données :

- Elles sont organisées en tableaux : les lignes représentent des **observations**, les colonnes représentent des **variables**.

Exemple : Données individuelles (aperçu)

- Les données sont collectées auprès de 10 individus, mesurant :
  - Niveau de revenu (faible, moyen, élevé)
  - Années de scolarité
- Il s'agit de données brutes au niveau individuel.

Pourquoi les données individuelles sont-elles importantes ?

- Elles capturent les différences à l'intérieur d'un groupe.
- Elles évitent les conclusions trompeuses qui apparaissent lorsqu'on n'observe que les données agrégées.

Exemple : Données individuelles (tableau)

Individu	Niveau de revenu	Années de scolarité
Personne 1	Faible	16
Personne 2	Faible	12
Personne 3	Moyen	14
Personne 4	Moyen	10
Personne 5	Moyen	8
Personne 6	Élevé	18
Personne 7	Élevé	17
Personne 8	Élevé	9
Personne 9	Élevé	20
Personne 10	Faible	5

Table: Données individuelles : revenu et éducation

Exemple : Données agrégées

- Si nous agrégeons les données par **groupe de revenu**, nous obtenons :

Groupe de revenu	**Années moyennes de scolarité**
Faible	11
Moyen	10,67
Élevé	16

Table: Données agrégées : revenu et éducation

Problème principal :

- La **variation interne aux groupes est perdue**.
- On a l'impression que **tous les individus à haut revenu ont 16 ans de scolarité**, alors qu'en réalité certains ont un niveau beaucoup plus faible.
- Cela conduit à des **conclusions trompeuses** si on applique ces résultats à des individus.

Qu'est-ce que l'erreur écologique ?

- L'**erreur écologique** survient lorsque l'on tire des **conclusions individuelles** à partir de **données agrégées**.
- Il s'agit d'une **erreur logique** d'inférence statistique, qui peut mener à des conclusions erronées.

Exemple :

- Supposons que l'on observe que les quartiers à **haut revenu** présentent un **taux de scolarité plus élevé**.
- Cela signifie-t-il que tous les **individus riches** sont hautement éduqués ? **Pas nécessairement !**

Pourquoi est-ce un problème ?

- Les données agrégées masquent la **variabilité individuelle**.
- Les corrélations observées au niveau des groupes ne reflètent pas nécessairement la réalité au niveau des individus.
- Des politiques publiques fondées sur une erreur écologique risquent d'être mal conçues.

Exemple :

- Un urbaniste constate que les **quartiers aisés présentent un niveau d'éducation plus élevé**.
- Il en déduit que **tous les individus aisés sont bien éduqués**.
- En réalité, certains individus à haut revenu ont un faible niveau d'éducation, et certains individus à faible revenu sont très éduqués.

Exemples classiques d'erreur écologique

1. **Robinson (1950)**
   - Article : *Ecological Correlations and the Behavior of Individuals*, ASR 15(3).
   - Corrélation agrégée : États avec plus dimmigrés = meilleur taux moyen dalphabétisation.
   - Mauvaise conclusion : les immigrés sont plus alphabétisés.
   - En réalité : au niveau individuel, taux plus faible.

2. **Vote nazi (années 1930)**
   - Analyses régionales : zones protestantes = fort vote NSDAP ; zones catholiques = faible vote.
   - Mauvaise conclusion : tous les protestants votent nazi, les catholiques jamais.
   - Travaux : Hamilton (1982, *Who Voted for Hitler?*), Falter (1991, *Hitlers Voters*).
   - Résultats : variation individuelle forte, traversant classes sociales et religions.

Exemples contemporains d'erreur écologique

3. Emmanuel Todd (2015)

- Essai : *Qui est Charlie ? Sociologie dune crise religieuse*, Seuil.
- Thèse : mobilisation du 11 janvier = France périphérique, blanche, catholique, islamophobe.
- Inférence erronée : profils individuels déduits de cartes régionales.

4. Mayer et Tiberj (2015)

- Tribune : *Le simplisme dEmmanuel Todd démonté par la sociologie des Je suis Charlie*, Le Monde, 18 mai 2015.
- Données CNCDH (sondage représentatif).
- Résultats : mobilisation diverse (jeunes, diplômés, enfants dimmigrés, catégories moyennes et populaires).
- Les manifestants sont moins islamophobes et moins antisémites que les non-manifestants.

À retenir

- L'**erreur écologique** survient lorsqu'on suppose que les tendances observées au niveau des groupes s'appliquent aux individus.
- Les données agrégées **peuvent induire en erreur**, notamment lorsqu'elles servent de base à des décisions publiques.
- Il est essentiel d'analyser la **variabilité individuelle** avant de tirer des conclusions.

Idée finale :

- **Tous les riches ne sont pas hautement éduqués.**
- **Tous les pauvres ne sont pas dépourvus d'éducation.**
- **L'agrégation statistique ne reflète pas la réalité individuelle !**

Types de variables

Variables qualitatives (catégorielles) :

- **Nominales :** catégories sans ordre hiérarchique.
  - Exemple : Réseau social préféré (TikTok, Instagram, Reddit).
  - Exemple : Groupe sanguin (A, B, AB, O).
- **Ordinales :** catégories ordonnées, sans intervalle numérique constant.
  - Exemple : Degré d'accord (Pas du tout d'accord, Plutôt pas d'accord, Neutre, Plutôt d'accord, Tout à fait d'accord).
  - Exemple : Niveau d'éducation (Lycée, Licence, Master, Doctorat).
- **Dichotomiques (binaires) :** cas particulier avec deux modalités seulement.
  - Exemple : Réussite/Échec à un examen.
  - Exemple : Possède un animal domestique ? (Oui/Non).

Types de variables :(suite)

Variables quantitatives (numériques) :

- **Discrètes :** valeurs dénombrables, souvent des entiers.
  - Exemple : Nombre de livres possédés.
  - Exemple : Nombre de fois où vous avez changé de sujet de thèse.
- **Continues :** valeurs pouvant prendre n'importe quelle valeur dans un intervalle ; généralement mesurées.
  - Exemple : Temps passé sur une tâche (en minutes).
  - Exemple : Taille en centimètres.

Exercice : Classifier les variables (science politique)

Classez les variables suivantes comme qualitatives, discrètes ou continues :

1. *Orientation partisane déclarée* (Gauche, Droite, Centre, Abstention)
2. *Niveau de confiance dans le parlement* (Aucune, Faible, Moyenne, Forte)
3. *Nombre de mandats parlementaires exercés par un député*
4. *Taux de participation électorale dans une circonscription* (en
5. *Type de régime politique du pays* (Démocratie, Autoritarisme, Hybride)
6. *Nombre de manifestations collectives auxquelles un individu a participé en un an*

Réponse : Classification des variables (science politique)

Variable	**Type**
Orientation partisane déclarée	Qualitative (nominale)
Niveau de confiance dans le parlement	Qualitative (ordinale)
Nombre de mandats parlementaires	Discrète
Taux de participation électorale	Continue
Type de régime politique	Qualitative (nominale)
Nombre de manifestations en un an	Discrète

Exercice : Classification des variables (politiques publiques)

Classez les variables suivantes :

1. *Mode principal de financement dun parti politique* (Dons privés, Subventions publiques, Cotisations)
2. *Niveau de satisfaction à légard des services de santé* (Très insatisfait, Insatisfait, Satisfait, Très satisfait)
3. *Nombre de lois adoptées par le parlement en une année*
4. *Dépenses publiques déducation par habitant* (en euros)
5. *Type de scrutin utilisé pour les élections législatives* (Proportionnel, Majoritaire, Mixte)
6. *Nombre de questions parlementaires posées par un député en une session*

Réponse : Classification des variables (politiques publiques)

Variable	**Type**
Mode de financement dun parti	Qualitative (nominale)
Satisfaction vis-à-vis des services de santé	Qualitative (ordinale)
Nombre de lois adoptées	Discrète
Dépenses publiques déducation par habitant	Continue
Type de scrutin	Qualitative (nominale)
Nombre de questions parlementaires	Discrète

Statistique et statistiques

Deux sens différents :

- **La statistique** : discipline des mathématiques appliquées à la collecte, analyse et interprétation des données.
- **Une statistique** : indicateur numérique particulier (ex. moyenne, médiane).

Processus statistique :

1. Collecte des données.
2. Analyse et traitement (résumer, transformer, modéliser).
3. Interprétation.
4. Présentation des résultats.

Modes de recueil des données

- **Recensement** : observation exhaustive de tous les éléments dune population.
  - Avantage : couverture totale.
  - Inconvénient : dispositif lourd et coûteux.
- **Sondage** : échantillon représentatif de la population.
  - Avantage : plus léger et moins coûteux.
  - Inconvénient : repose sur la représentativité.

Sources de données

Individuelles :

- CDSP (Sciences Po), Réseau Quetelet, ICPSR.
- Eurobaromètres, ESS, ISSP, EVS, WVS.
- Baromètres régionaux : Afrobarometer, Arab Barometer, etc.

Agrégées :

- Insee, Ined, Crédoc.
- Eurostat, OCDE, ONU, Banque mondiale, V-Dem.

Historique des sondages

- Développement aux États-Unis au début du XXe siècle (Gallup 1936).
- En France : après 1945, Jean Stoetzel (IFOP).
- Controverse : ń Lopinion publique nexiste pas ż (Bourdieu).

Principe de léchantillon

- Exemple de la soupe : une cuillère suffit, à condition quelle soit bien mélangée.
- Exemple de la bière : une gorgée suffit, à condition de ne pas se limiter à la mousse.
- La représentativité est lenjeu central.

Échantillonnage aléatoire (probabiliste)

- Chaque individu de la population a une probabilité connue et non nulle dêtre sélectionné.
- Fondé sur la théorie des probabilités et la loi des grands nombres.
- Permet destimer des paramètres avec un intervalle de confiance.
- Exemple : marge derreur ś2% avec un seuil de confiance de 95%.

Échantillonnage non-aléatoire (empirique)

- **Quotas** : respect de proportions selon critères connus (sexe, âge, profession, etc.).
- **Boule de neige** : propagation par réseau relationnel (souvent en recherche qualitative).
- **Volontaire** : panels web.
- **Accidentel** : cas fortuit, souvent utilisé par les médias.

Les biais des sondages

- **Erreur aléatoire** : variation due au hasard de léchantillonnage.
- **Erreur systématique** : biais lié au dispositif de collecte.
- Non-réponse totale ou partielle (répondant fantôme).
- Désirabilité sociale : réponses biaisées par pression normative.
- Mode de collecte (face-à-face, téléphone, internet).

Corriger les erreurs

- **Pondération** : ajustement selon caractéristiques connues des non-répondants.
- **Intervalle de confiance** : estimation de lincertitude statistique.
- Limites : aucune correction ne supprime totalement les biais.

Mesures de tendance centrale

Définition : les mesures de tendance centrale résument un ensemble de données en identifiant une valeur centrale ou typique.

1. Moyenne (arithmétique)

- Représente la valeur moyenne de l'ensemble des observations.
- Sensible aux valeurs extrêmes (outliers).
- Formule :

$$\bar{x} = \frac{\sum x_i}{n}$$

où :

- x_i = observations individuelles,
- n = nombre total d'observations.

- Exemple : densités de quartiers : $3000, 4000, 5000$

$$\bar{x} = \frac{3000 + 4000 + 5000}{3} = 4000$$

Mesures de tendance centrale

2. Médiane (valeur centrale)

- La valeur qui divise l'échantillon en deux parties égales.
- Moins affectée par les valeurs extrêmes que la moyenne.
- Calcul :
  - Si n est impair : la médiane est la valeur centrale.
  - Si n est pair : la médiane est la moyenne des deux valeurs centrales.
- Exemple :
  - Données ordonnées : $3000, 4000, 5000 \Rightarrow$ Médiane $= 4000$
  - Données ordonnées : $2500, 3000, 4000, 5000 \Rightarrow$ Médiane $= \frac{3000+4000}{2} = 3500$

Mesures de tendance centrale

3. Mode (valeur la plus fréquente)

- La valeur qui apparaît le plus souvent dans l'échantillon.
- Peut être utilisée pour des données qualitatives comme quantitatives.
- Exemple :
  - Mode de transport le plus courant : Bus (si la majorité répond ń Bus ż).
  - Densités de quartiers : $2000, 3000, 4000, 4000, 5000 \Rightarrow$ Mode $= 4000$ (car apparaît deux fois).

Au-delà de la tendance centrale : pourquoi la variabilité compte

- Les mesures de tendance centrale résument l'ensemble, mais ne disent rien sur la dispersion des données.
- Exemple : deux villes peuvent avoir la même moyenne de pollution, mais l'une connaît des fluctuations extrêmes.
- Les mesures de variabilité permettent de comprendre la dispersion et la stabilité des données.

Illustration : deux villes avec la même moyenne mais des variabilités différentes

Scénario : Deux villes ont le même niveau moyen de pollution quotidienne (PM2.5 en ţg/mş), mais une variabilité différente.

Jour	**Ville A**	**Ville B**
Lundi	20	10
Mardi	22	40
Mercredi	18	15
Jeudi	21	35
Vendredi	19	20

Table: Niveaux quotidiens de PM2.5 dans deux villes

Illustration : deux villes avec la même moyenne mais des variabilités différentes (suite)

Observations principales :

- Les deux villes ont la même moyenne :

$$\bar{x} = \frac{20 + 22 + 18 + 21 + 19}{5} = 20$$

- **Ville A :** faible variabilité, qualité de l'air stable.
- **Ville B :** forte variabilité, fluctuations importantes, pollution imprévisible.

Mesures de variabilité

Définition : elles indiquent à quel point les données sont dispersées autour de la tendance centrale.

1. Étendue (ou amplitude)

- Mesure la différence entre la valeur maximale et la valeur minimale.
- Formule :

$$\text{Étendue} = \max(x) - \min(x)$$

- Exemple :

$$\text{Étendue} = 5000 - 3000 = 2000$$

Mesures de variabilité

2. Variance (σ^2) (écart quadratique moyen)

- Indique l'écart moyen au carré par rapport à la moyenne.
- Formule (population) :

$$\sigma^2 = \frac{\sum(x_i - \bar{x})^2}{n}$$

- Exemple : données : $3000, 4000, 5000$, moyenne $= 4000$

$$\sigma^2 = \frac{(3000 - 4000)^2 + (4000 - 4000)^2 + (5000 - 4000)^2}{3} = 666667$$

Mesures de variabilité

3. Écart-type (σ) (racine carrée de la variance)

- Exprime la dispersion dans les mêmes unités que les données initiales.
- Formule :

$$\sigma = \sqrt{\sigma^2}$$

- Exemple :

$$\sigma = \sqrt{666667} \approx 816.5$$

Quelles mesures selon le type de variables ?

Mesure	Nominales	Ordinales	Continues
Mode	✓	✓	✓
Médiane	×	✓	✓
Moyenne	×	×	✓
Étendue	×	×	✓
Variance	×	×	✓
Écart-type	×	×	✓

Table: Pertinence des mesures statistiques selon le type de variables

Quelles mesures selon le type de variables ? (suite)

Points clés :

- **Variables nominales :** seul le mode est applicable (ex. : source d'énergie dominante dans un quartier).
- **Variables ordinales :** la médiane est appropriée, mais les mesures de dispersion (variance, écart-type) n'ont pas de sens (ex. : perception de la qualité de l'air).
- **Variables continues :** toutes les mesures s'appliquent, permettant d'étudier à la fois la tendance centrale et la dispersion (ex. : fluctuations quotidiennes de température).