

Les grands types de données statistiques

Méthodes Quantitatives - Séance 5 (Partie 1)

Sciences Po

2024-2025

Plan de la séance

- 1 Introduction : La mesure du crime
- 2 Les trois grands types de données
- 3 Les données d'enquête
- 4 Les données administratives
- 5 Les données issues de traces
- 6 Travail pratique
- 7 Représentativité et biais
- 8 L'échantillonnage aléatoire
- 9 Taille d'échantillon vs taux de sondage
- 10 Alternatives à l'échantillonnage aléatoire
- 11 Conclusion
- 12 Pratique avec Google Sheets

Un débat révélateur : la mesure du crime et de l'insécurité

"J'aime beaucoup les enquêtes de victimation et les experts médiatiques, mais je préfère le bon sens du boucher-charcutier de Tourcoing"



Un débat révélateur : la mesure du crime et de l'insécurité

"La réalité dément cette étude"



Les enjeux de la mesure

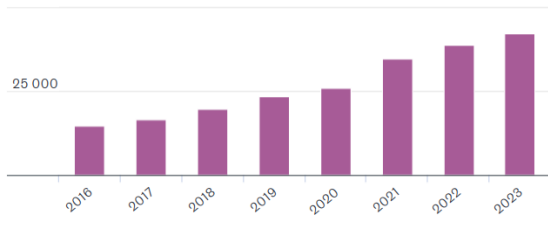
Différence entre la criminalité réelle et la criminalité connue
Par définition, il n'existe pas de données exhaustive sur la criminalité.

La Mission interministérielle pour la protection des femmes contre les violences et la lutte contre la traite des êtres humains :

- En moyenne, le nombre de femmes âgées de 18 ans et plus qui, en 2022, ont été victimes de viols, tentatives de viol et/ou agressions sexuelles est estimé à 230 000 femmes.
- Parmi l'ensemble des victimes, femmes et hommes :
 - 61 % déclarent que ces agressions ont été perpétrées par une personne connue de la victime,
 - 28 % déclarent que le conjoint ou ex-conjoint était l'auteur des faits.
- Suite aux viols, tentatives de viol et/ou agressions sexuelles qu'elles ont subi, seules 6 % des femmes victimes déclarent avoir porté plainte.

Cette estimation est issue des résultats de l'enquête de victimation "Vécu et ressenti en matière de sécurité" du SSMSI, réalisée en 2023.

Bases statistiques des victimes de violences sexuelles enregistrées par la police et la gendarmerie entre 2016 et 2023, viols et tentatives de viol



Source : [SSMSI](#)

Les deux sources vous semblent-elles cohérentes ?

- Toutes les formes de criminalité ne sont pas reportées de la même manière aux autorités, et n'ont pas le même écho dans la presse
- Il existe même des études sur la corrélation sur la perception d'insécurité et l'insécurité effective.

- Toutes les formes de criminalité ne sont pas reportées de la même manière aux autorités, et n'ont pas le même écho dans la presse
- Il existe même des études sur la corrélation sur la perception d'insécurité et l'insécurité effective.

Facteurs influençant le signalement :

- Gravité perçue
- Confiance dans les institutions
- Accessibilité des services de police
- Normes sociales et culturelles

- Les enquêtes de victimation permettent de contourner certaines limites mais ne règlent pas tout.

Ce problème est-il propre aux données sur la criminalité ?

Ce problème est-il propre aux données sur la criminalité ?

Non, il est lié à un problème "d'observabilité" inhérent à bien des comportements.

Par exemple :

- orientation sexuelle
- répartition du travail domestique
- données de vote individuelles

Les individus eux-mêmes ne sont pas très fiables (intentionnellement ou non) pour rapporter leur propre comportement.

Ce problème est-il propre aux données sur la criminalité ?

Non, il est lié à un problème "d'observabilité" inhérent à bien des comportements.

Par exemple :

- orientation sexuelle
- répartition du travail domestique
- données de vote individuelles

Les individus eux-mêmes ne sont pas très fiables (intentionnellement ou non) pour rapporter leur propre comportement.

Le problème ne concerne les données d'enquêtes et les données administratives.

Vue d'ensemble

① Données d'enquête

- Collecte active d'informations
- Protocole standardisé

② Données administratives

- Sous-produit d'activités
- Finalité pratique

③ Données issues de traces

- Construction a posteriori
- Sources variées

Les données d'enquête : définition approfondie

● **Producteurs :**

- Instituts nationaux (INSEE, INED...)
- Organismes internationaux (Eurostat, OCDE...)
- Centres de recherche
- Instituts de sondage

● **Finalités :**

- Recherche académique
- Statistique publique
- Études de marché
- Sondages d'opinion

Focus : L'enquête Emploi de l'INSEE

- **Objectif** : mesurer le chômage selon les critères du BIT
- **Méthode** :
 - Échantillon de 100 000 logements
 - Interrogation en continu
 - Suivi pendant 6 trimestres
- **Particularités** :
 - Questions très précises sur la recherche d'emploi
 - Différent des chiffres de Pôle Emploi (donnée administrative)
 - Comparabilité internationale

Les grandes enquêtes internationales

- **European Social Survey (ESS)**
 - Valeurs, attitudes, comportements
 - Tous les 2 ans depuis 2002
- **Programme for International Student Assessment (PISA)**
 - Compétences des élèves de 15 ans
 - Organisation : OCDE
- **World Values Survey**
 - Évolution des valeurs dans le monde
 - Depuis 1981

Les modes de collecte

En face-à-face

- Qualité des réponses
- Coût élevé
- Biais de désirabilité

Par téléphone

- Moins coûteux
- Problème des mobiles
- Taux de refus

Auto-administré

- Papier ou web
- Questions sensibles
- Risque d'incompréhension

Mixte

- Combinaison de modes
- Optimisation des coûts
- Complexité accrue

Les enjeux actuels des enquêtes

- **Baisse des taux de réponse**
 - Multiplication des sollicitations
 - Méfiance croissante
 - Impact sur la qualité
- **Nouveaux modes de collecte**
 - Enquêtes en ligne
 - Panels d'accès
 - Enquêtes par smartphone
- **Questions méthodologiques**
 - Représentativité
 - Comparabilité
 - Protection des données

Les données administratives : diversité des sources

Types d'organisations productrices

- Administrations publiques
- Entreprises
- Associations
- Organisations internationales

- **Exemples concrets :**

- État civil
- Données fiscales
- Dossiers médicaux
- Registres scolaires
- Données de consommation

Exemple : Le système d'information scolaire

- **Base élèves**

- Gestion administrative
- Suivi des effectifs
- Attribution des moyens

- **Usages secondaires**

- Études sur les parcours
- Évaluation des politiques
- Recherche en éducation

- **Limites**

- Enseignement public uniquement
- Variables disponibles limitées
- Changements de nomenclature

Les nouvelles sources administratives

- **Big data administratives**

- Données de géolocalisation
- Transactions bancaires
- Consommation d'énergie

- **Enjeux**

- Volume des données
- Temps réel vs différé
- Protection de la vie privée
- Accès aux chercheurs

Les traces : une typologie

Traces historiques

- Archives
- Correspondances
- Presse ancienne
- Documents comptables

Traces contemporaines

- Réseaux sociaux
- Blogs et forums
- Sites web

Traces involontaires

- Logs de connexion
- Géolocalisation
- Capteurs

Traces volontaires

- Publications
- Commentaires
- Photos partagées

Exemple : étudier le 19e siècle

- **Sources possibles**

- Registres paroissiaux
- Actes notariés
- Journaux d'époque
- Correspondances privées

- **Défis méthodologiques**

- Conservation sélective
- Lisibilité des documents
- Contexte historique
- Représentativité

Les traces numériques

Définition

Données générées par l'utilisation des technologies numériques

- **Types de traces**

- Navigation web
- Interactions sur réseaux sociaux
- Achats en ligne
- Géolocalisation

- **Enjeux éthiques**

- Consentement
- Anonymisation
- Droit à l'oubli

Exercice en groupes

Consignes du travail en groupe

Par groupes de 2-4 étudiants :

① Analyser le document fourni

- Identifier le type de données
- Repérer la source précise
- Noter le contexte de production

② Réfléchir aux biais

- Population manquante
- Impact des catégories
- Objectifs du producteur

Grille d'analyse

Pour chaque document, se poser systématiquement ces questions :

- **Sur la source**

- Qui a produit ces données ?
- Dans quel but initial ?
- À quelle date ?

- **Sur les biais**

- Qui est inclus/exclu ?
- Quelles sont les catégories utilisées ?
- Quels comportements sont visibles/invisibles ?

Synthèse

Points clés à retenir

- Pas de source parfaite
- Importance de croiser les sources
- Conscience des biais
- Documentation des choix

Le défi de la représentativité

- **Question initiale :**

- " Pourquoi le résultat dépend-il de la population interrogée ?"
- " Comment résoudre les problèmes de non-représentativité ?"

- **Exemple concret :**

- Sondage Literary Digest (1936)
- Prédiction élection présidentielle US
- Échantillon : 2.4 millions de personnes
- Erreur majeure : biais de sélection

Les idées reçues sur l'échantillonnage

- **Fausse croyances courantes :**

- "Il faut interroger beaucoup de monde"
- "Il faut un pourcentage minimum de la population"
- "Un échantillon sur dix est représentatif"
- "Plus on interroge, plus c'est précis"

- **Point clé :**

- La qualité prime sur la quantité
- L'aléatoire est crucial

Qu'est-ce que l'aléatoire ?

Expérience

Pensez à un nombre "au hasard" entre 1 et 10

- **Constat habituel :**
 - Surreprésentation de certains nombres (3, 7)
 - Sous-représentation d'autres (1, 10)
- **Principe de l'aléatoire :**
 - Équiprobabilité
 - Indépendance des tirages

Ce qui n'est PAS aléatoire

- **"Ce qui me tombe sous la main"**
 - Biais de disponibilité
 - Biais de sélection
- **"Un sur dix"**
 - Risque de périodicité
 - Dépendance à l'ordre initial
- **"Au feeling"**
 - Biais cognitifs
 - Préférences inconscientes

Une découverte contre-intuitive

Principe fondamental

La précision dépend de la taille de l'échantillon, pas du taux de sondage

- **Exemples :**

- 1000 personnes sur 67 millions (France)
- 1000 personnes sur 1.4 milliard (Chine)
- Même précision !

Pourquoi est-ce vrai ?

- **Explication intuitive**

- Variété des profils capturés
- Combinaisons de caractéristiques
- Loi des grands nombres

- **Histoire**

- Démonstré depuis plus d'un siècle
- Base des sondages modernes
- Utilisé dans tous les pays

Quand l'aléatoire est impossible

- **Situations courantes**

- Absence de base de sondage
- Population difficile à joindre
- Contraintes pratiques
- Coûts trop élevés

- **Populations spécifiques**

- Populations rares
- Populations cachées
- Populations mobiles

Méthodes alternatives

Échantillonnage en boule de neige

- Populations difficiles d'accès
- Réseaux sociaux
- Effet de grappe

Échantillonnage par quotas

- Critères prédéfinis
- Rapidité
- Coût modéré

Exemples concrets

- **Études sur les sans-abri**

- Pas de base de sondage
- Points de contact
- Réseaux d'aide

- **Études de marchés**

- Quotas socio-démographiques
- Lieux de passage
- Panels en ligne

Points clés à retenir

- L'aléatoire est crucial mais pas toujours possible
- La taille d'échantillon prime sur le taux de sondage
- Différentes méthodes pour différents objectifs
- Importance de documenter ses choix

La fonction ALEA()

Principe

Génère un nombre aléatoire entre 0 et 1

- **Points importants :**
 - Se recalcule à chaque modification
 - Nécessite une stabilisation
 - Permet un tri aléatoire

Exercice pratique

Base de données des résultats des élections présidentielles :

- **Objectif** : tirer 1000 communes au hasard
- **Étapes** :
 - 1 Générer nombres aléatoires
 - 2 Stabiliser le tirage
 - 3 Trier
 - 4 Sélectionner l'échantillon
- **Vérification** :
 - Calculer moyennes sur quelques variables
 - Comparer avec population totale

Lectures d'approfondissement suggérées :

- Olivier Martin, L'enquête et ses méthodes. L'analyse de données quantitatives, Paris, Armand Colin, 2020 (1re éd 2005), chapitre 1, [Lien](#)
- Jean-Marie Firdion, "Construire un échantillon", in Serge Paugam (dir.), L'enquête sociologique, Paris, Presses Universitaires de France, 2012, p. 69-92. [Lien](#)
- Claude Dargent, "Construire un échantillon", Sociologie des opinions, Paris, Armand Colin, 2011, p. 93-114, [Lien](#)
- optionnel : exemples d'échantillonnage stratifié et d'adaptation aux contraintes pratiques : encadré 2 dans Anton Perdoncin, "Tous au charbon ? Inégalités de carrières et mobilités ouvrières dans la récession (Nord-Pas-de-Calais, 1945-1990)", Genèses, 122-1, 2021, p. 9-33 [Lien](#) et/ou partie "1.1 La méthode de suivi de carrières : l'échantillonnage" dans Michel Grignon, "Reconstitution de carrières démographiques et professionnelles sur une population d'ouvriers migrants du XIXe siècle", Histoire & Mesure, 8-1, 1993, p. 145-162. [Lien](#)