

Méthodes Quantitatives

Scéance 3: Relier des variables

2024-2025

Introduction : La statistique, c'est comparer

Pourquoi comparer ?

- Limite de l'analyse univariée
 - La moyenne seule ne dit pas tout
 - La dispersion seule ne suffit pas
 - Besoin de mettre en relation
- Comprendre les liens entre phénomènes
 - Relations causales potentielles
 - Corrélations et associations
 - Structures sous-jacentes
- Applications pratiques
 - Politique publique
 - Marketing
 - Recherche sociale

Types de comparaisons

- Comparaison dans le temps
 - Évolution d'une variable
 - Séries temporelles
 - Exemple : évolution des salaires
- Comparaison entre groupes
 - Différences démographiques
 - Variations géographiques
 - Exemple : différences hommes-femmes
- Comparaison entre variables
 - Relations entre phénomènes
 - Interdépendances
 - Exemple : éducation et revenu

Liaison entre variables catégorielles

Le tableau croisé : principe

- Définition
 - Représentation bidimensionnelle
 - Croisement de deux variables qualitatives
 - Base de l'analyse bivariée catégorielle
- Construction
 - Variables en lignes et colonnes
 - Cases : intersections des modalités
 - Effectifs ou pourcentages
- Objectifs
 - Visualiser les relations
 - Identifier les associations
 - Tester l'indépendance

Exemple concret : Diplôme et possession d'animal

Diplôme	Chien	Chat	Total
Sans diplôme	39.4%	24.2%	100%
CAP/BEP	42.2%	24.5%	100%
Bac	26.7%	21.0%	100%
Supérieur	24.3%	20.1%	100%

Table – Source : Données Hérans (1987)

- Lecture en ligne : comparaison des taux par niveau d'études
- Gradient social visible
- Question d'interprétation : causalité ?

Les différentes lectures du tableau croisé

- Lecture en effectifs
 - Nombres bruts
 - Importance numérique
 - Taille des sous-groupes
- Lecture en ligne (%)
 - Structure horizontale
 - Distribution d'une modalité
 - " Parmi les bacheliers..."
- Lecture en colonne (%)
 - Structure verticale
 - Composition d'une catégorie
 - " Parmi les possesseurs de chiens..."

Conventions et bonnes pratiques

1 Présentation

- Variable explicative en ligne
- Variable expliquée en colonne
- Titres clairs et précis

2 Contenu

- Effectifs ET pourcentages
- Totaux en marges
- Note de lecture

3 Format

- Tableau détachable
- Source indiquée
- Unités précisées

Liaison entre variable catégorielle et quantitative

Deux approches complémentaires

Approche 1 : Catégorisation

- Transformation de la variable quantitative
- Création de classes
- Tableau croisé classique

Approche 2 : Statistiques par groupe

- Conservation des valeurs
- Comparaison de moyennes
- Analyse de la dispersion

Exemple : Salaire selon le genre

Genre	Moyenne	Écart-type	Médiane	N
Homme	2850€	1250€	2600€	476
Femme	2450€	980€	2300€	175

- Différence de moyennes : 400€
- Dispersions différentes
- Question d'interprétation :
 - Autres facteurs explicatifs
 - Tests statistiques nécessaires

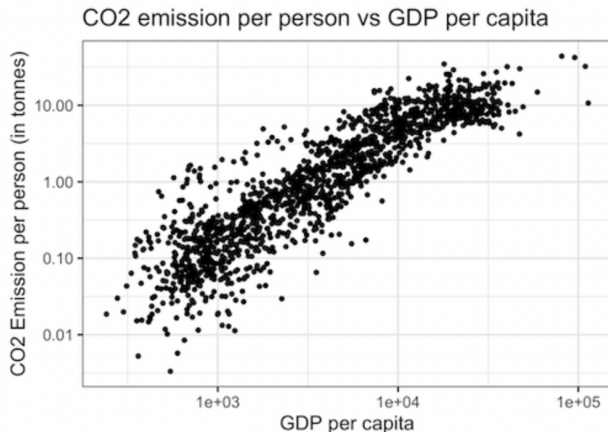
Liaison entre variables quantitatives

Analyse de la covariation

- Question fondamentale :
 - Variation conjointe ?
 - Dans quel sens ?
 - Quelle intensité ?
- Outils principaux :
 - Nuage de points
 - Covariance
 - Coefficient de corrélation
- Limites :
 - Relations non linéaires
 - Points influents
 - Causalité vs corrélation

Le nuage de points

- Axe X : première variable
- Axe Y : seconde variable
- Un point = une observation



La covariance

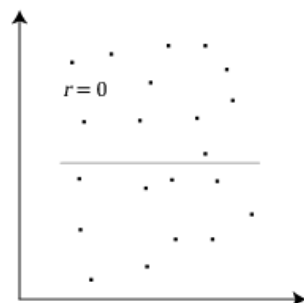
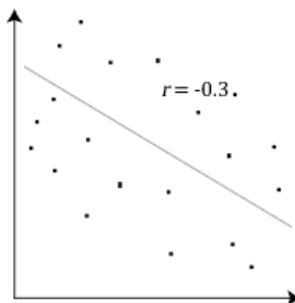
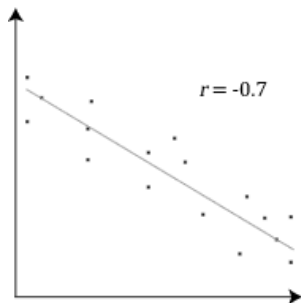
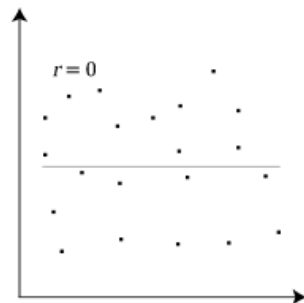
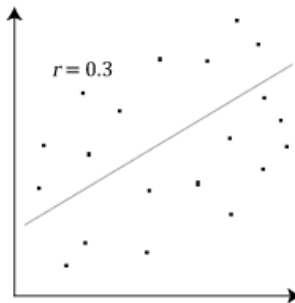
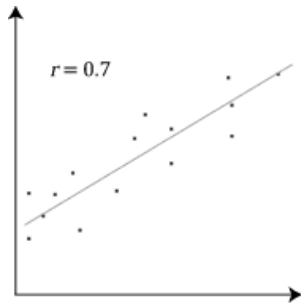
- Définition mathématique :

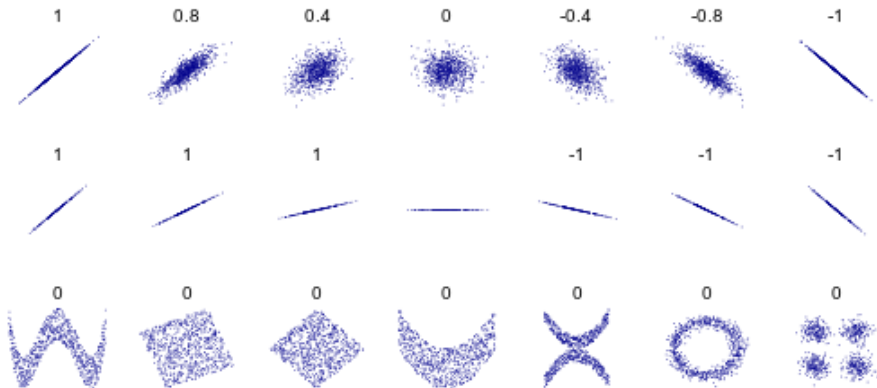
$$COV(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n}$$

- Interprétation
 - Signe : direction de la relation
 - Amplitude : force (avec précaution)
 - Unité : produit des unités
- Limites
 - Dépend des échelles
 - Non borné
 - Peu comparable

Le coefficient de corrélation

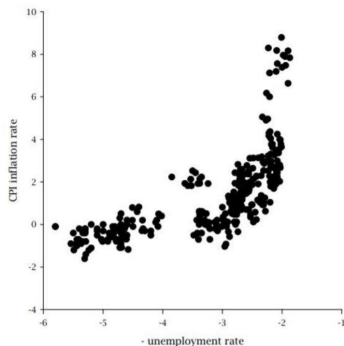
- Définition : $r_{X,Y} = \frac{COV(X,Y)}{\sigma_X \cdot \sigma_Y}$
- Propriétés
 - Borné entre -1 et 1
 - Sans unité
 - Symétrique





"Japan's phillip's curve looks like Japan"

Figure 2: Japan's Inflation Rate and (Minus) Unemployment Rate
January 1980 to August 2005



Points d'attention méthodologique

- Valeurs manquantes
 - Traitement explicite
 - Impact sur les résultats
 - Documentation des choix
- Points influents
 - Détection
 - Impact sur la corrélation
 - Analyse de sensibilité
- Non-linéarité
 - Détection graphique
 - Corrélation trompeuse
 - Méthodes alternatives

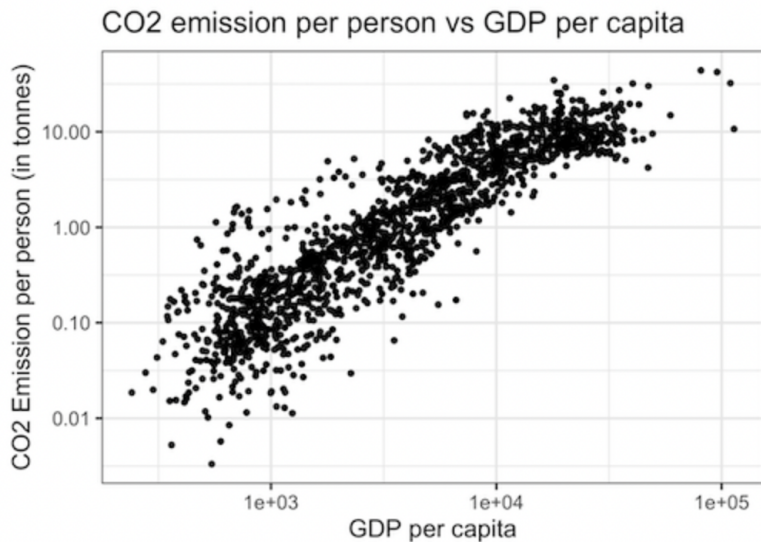
Synthèse des méthodes

Type de liaison	Méthode	Visualisation
Cat * Cat	Tableau croisé	Barres groupées
Cat * Quant	Moyennes groupées	Boîtes à moustaches
Quant * Quant	Corrélation	Nuage de points

- Choix en fonction des variables
- Complémentarité des approches
- Importance de la visualisation

Mesurer la causalité

Les limites de la corrélation



Mesurer la causalité

Ces slides sont basés sur les excellents slides de Clément de Chaisemartin

La causalité en sciences sociales

- Le but d'une recherche est souvent de mesurer une relation de cause à effet
- Exemples :
 - Effet du vaccin Covid sur la mortalité
 - Impact d'une classe de CP à 12 élèves sur l'apprentissage
 - Effet d'une bourse d'études sur la scolarité au Ghana

Traitement et Résultat

- La littérature sur l'inférence causale vient de la médecine
- Vocabulaire :
 - **Traitement** : l'intervention étudiée (vaccin, classe à 12, bourse...)
 - **Résultat** : la variable d'intérêt (mortalité, niveau de lecture...)

Le problème fondamental de l'inférence causale

- Peut-on **mesurer** l'effet causal du traitement sur un individu ?

Le problème fondamental de l'inférence causale

- Peut-on **mesurer** l'effet causal du traitement sur un individu ?
- Non.

Le problème fondamental de l'inférence causale

- Peut-on **mesurer** l'effet causal du traitement sur un individu ?
- Non.
- Pourquoi ?
 - On ne peut jamais observer le même individu avec ET sans traitement
 - On n'observe pas le résultat contrefactuel

Le problème fondamental de l'inférence causale

- Peut-on **mesurer** l'effet causal du traitement sur un individu ?
- Non.
- Pourquoi ?
 - On ne peut jamais observer le même individu avec ET sans traitement
 - On n'observe pas le résultat contrefactuel
- On ne peut que l'estimer.

Le contrefactuel

- Exemple : un élève en CP à 12 élèves
- On observe son niveau de lecture en fin d'année
- Mais on ne peut pas observer quel aurait été son niveau s'il avait été dans une classe à 24
- C'est ce qu'on appelle le contrefactuel
- Il est par définition inobservable

La solution : comparer des groupes

Au lieu de comparer le même individu avec/sans traitement :

- On compare deux groupes :
 - **Groupe test** : reçoit le traitement
 - **Groupe contrôle** : ne reçoit pas le traitement
- On compare leurs résultats moyens :

$$\text{Effet moyen} = \text{Moyenne}(\text{groupe test}) - \text{Moyenne}(\text{groupe contrôle})$$

Le problème des variables omises

- La simple comparaison peut être trompeuse
- Exemple : CP à 12 vs CP à 24
 - CP à 12 = écoles REP (milieux défavorisés)
 - CP à 24 = autres écoles
 - Différence de résultats due à la taille de classe ou au milieu social ?
- C'est le biais de variable omise

$$\begin{aligned}
 & \underbrace{\text{Moyenne}(\text{groupe test}) - \text{Moyenne}(\text{groupe contrôle})}_{\text{Différence observée}} \\
 &= \underbrace{\text{Effet du traitement}}_{\text{Ce qu'on veut mesurer}} \\
 &+ \underbrace{\text{Moyenne}(\text{contrefact. groupe test}) - \text{Moyenne}(\text{groupe contrôle})}_{\text{Biais de variable omise}}
 \end{aligned}$$

Interprétation des composantes :

- **Différence observée** : C'est ce qu'on peut calculer directement à partir des données
- **Effet du traitement** : C'est l'effet causal qu'on cherche à mesurer
- **Biais de variable omise** : C'est la différence qui existe entre les groupes

La randomisation : principe

- Solution : tirer au sort les groupes
- Exemple :
 - Pile : l'école aura des CP à 12
 - Face : l'école gardera des CP à 24
- Conséquence mathématique :
 - Les groupes sont similaires en moyenne
 - Plus de biais de variable omise

La marge d'erreur statistique

Formule :

$$\text{MES} = 1.96 \times \text{écart-type} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_0}}$$

où :

- n_1 = taille du groupe test
- n_0 = taille du groupe contrôle

Interpréter la significativité

- Si la différence entre les moyennes des deux groupes est plus grande que la MES ou plus petite que $-MES$:
 - 95% de chances que ce ne soit pas le hasard
 - On parle d'effet statistiquement significatif
- La p-value :
 - Probabilité que la différence soit due au hasard
 - Significatif si $p\text{-value} < 0.05$ (ou 0.01 ou 0.001 selon le niveau d'exigence)

Conclusion

- La randomisation est l'outil le plus puissant
- Elle permet de mesurer la causalité
- Mais attention :
 - Pas toujours possible
 - Coûteux et long
 - Questions éthiques

Les études observationnelles

Le cadre observationnel : définition

- Situation où la randomisation n'est pas possible
- Raisons courantes :
 - Contraintes éthiques
 - Contraintes politiques
 - Contraintes pratiques
 - Politique déjà mise en place
- Défi : trouver un groupe témoin comparable

Exemple : Le dédoublement des CP en REP+

Contexte :

- Politique mise en place en REP+
- Impossible de randomiser
- Question : quel effet sur l'apprentissage ?

Challenge :

- Comment évaluer l'effet sans randomisation ?
- Comment construire un groupe témoin ?

Exemple : Résultats sur les CP dédoublés

Étude DEPP :

- Résumé de l'étude
- Présentez l'approche de la DEPP
- Présentez les résultats

Construction du groupe témoin

Approche de la DEPP :

- **Groupe test** : élèves en REP+ avec CP dédoublés
- **Groupe témoin** : élèves hors REP+ mais...
 - Profil social proche
 - Niveau scolaire initial comparable
 - Conditions d'enseignement similaires

Le problème du biais de sélection

Risques potentiels :

- Différences systématiques entre groupes
- Variables inobservables
- Effet de sélection

Questions à se poser :

- Les groupes sont-ils vraiment comparables ?
- Y a-t-il des différences non mesurées ?
- Les résultats sont-ils généralisables ?

Solutions méthodologiques

1. Matching (appariement) :

- Associer chaque individu traité à un "jumeau" non traité
- Sur la base de caractéristiques observables
- Limites : variables non observées

2. Différence de différences :

- Comparer l'évolution dans le temps
- Entre groupe traité et non traité
- Hypothèse : tendances parallèles

Vérifier la comparabilité

Tests préliminaires :

- Comparer les caractéristiques observables
- Tester les différences avant traitement
- Analyser les tendances pré-traitement

Variables à comparer :

- Caractéristiques socio-démographiques
- Résultats scolaires antérieurs
- Caractéristiques des écoles

Interpréter les résultats

Points de vigilance :

- Hypothèses sous-jacentes
- Robustesse des résultats
- Tests de sensibilité
- Généralisation possible

Questions critiques :

- Les résultats résistent-ils aux tests ?
- Les hypothèses sont-elles plausibles ?
- Quelle est la portée des conclusions ?

Limites des études observationnelles

- Impossibilité d'éliminer tous les biais
- Hypothèses non testables
- Résultats plus fragiles
- Généralisabilité limitée
- Nécessité de réplication

Bonnes pratiques

Pour renforcer la validité :

- Transparence sur les méthodes
- Tests de robustesse multiples
- Documentation des hypothèses
- Réplication avec différentes méthodes
- Publication des analyses de sensibilité

Conclusion : Études observationnelles

Points clés :

- Utiles quand la randomisation est impossible
- Nécessitent des méthodes rigoureuses
- Importance de la construction du groupe témoin
- Résultats à interpréter avec prudence
- Complémentaires aux études randomisées

Un regard critique

Le problème des tests multiples

Problème :

- Quand on teste plusieurs variables, on augmente le risque de trouver des résultats "significatifs" par hasard

Exemple :

- Si on fait 20 tests avec $\alpha = 5\%$
- Probabilité de trouver au moins un faux positif :

$$P(\text{au moins un faux positif}) = 1 - (0.95)^{20} \approx 64\%$$

Solutions :

- Correction de Bonferroni : $\alpha' = \alpha/n$ pour n tests
- Pré-enregistrement des hypothèses à tester
- Réplication des résultats

Le P-Hacking

Définition :

- Manipulation des données ou des analyses pour obtenir des résultats "significatifs"

Exemples de pratiques :

- Tester plusieurs variables jusqu'à en trouver une "significative"
- Exclure des observations "gênantes"
- Arrêter la collecte de données quand le résultat devient significatif
- Choisir la spécification qui donne le résultat voulu

Solutions :

- Pré-enregistrement des études
- Transparence des données et du code
- Publication des résultats nuls
- Réplication indépendante

Pour développer son esprit critique

Face à une affirmation causale :

- Quel est le contrefactuel implicite ?
- Le groupe contrôle est-il vraiment comparable ?
- Y a-t-il des variables omises ?
- Les différences sont-elles significatives ?
- Les résultats sont-ils généralisables ?