Tests d'association : Fondements et applications

Ahmed Fouad EL HADDAD

Sciences Po Paris

October 10, 2025

De l'univarié au bivarié

- Jusqu'ici, nous avons étudié les **statistiques univariées**, c'est-à-dire l'analyse de variables isolées.
- Or, les sciences sociales s'intéressent rarement à des variables isolées – il s'agit de comprendre les relations entre elles.
- Trois manières principales de mesurer les associations entre variables :
  - **Covariance** – mesure de l'association linéaire.
  - **Corrélation** – mesure normalisée de l'association.
  - **Tableaux de contingence** – utilisés pour les variables catégorielles.

Pourquoi mesurer l'association ?

- En sciences sociales, la plupart des questions de recherche portent sur des relations :
  - Le **revenu** influence-t-il la **participation politique** ?
  - Existe-t-il une relation entre **niveau d'éducation** et **soutien à la démocratie** ?
  - Comment **classe sociale** et **comportement électoral** interagissent-ils ?
- Mesurer les associations permet de :
  - Identifier des tendances et des régularités sociales.
  - Tester des hypothèses théoriques avec des données empiriques.
  - Distinguer entre corrélations **réelles** et **fallacieuses**.
- Sans tests statistiques formels, nous risquons de nous appuyer sur l'intuition et l'observation brute, ce qui peut induire en erreur.

Au-delà de la distinction quali/quanti

- Jusqu'à présent, nous avons distingué les variables **quantitatives** (numériques) et **qualitatives** (catégorielles).
- Mais pour construire une analyse pertinente, il faut se concentrer sur le **rôle** des variables dans la question de recherche :
  - **Variable dépendante (VD)** : le phénomène à expliquer.
  - **Variable indépendante (VI)** : le ou les facteurs supposés influencer la VD.
- La question centrale devient : **Que veux-je expliquer, et à partir de quoi ?**

Formuler des hypothèses d'association

- Une fois la VD et la VI définies, on peut formuler des hypothèses sur :
  - **La direction de la relation** : la VI augmente-t-elle ou diminue-t-elle la VD ?
  - **La force de la relation** : effet faible, modéré ou fort ?
- Exemple (études urbaines) :
  - Hypothèse : **Plus l'accès aux transports publics est élevé, plus le taux de possession automobile est faible**.
  - Ici, **possession automobile** = VD ; **accessibilité des transports publics** = VI.
  - Relation attendue : **association négative**.
- Ces hypothèses guident le choix des tests statistiques.

Exercice : Identifier VD et VI

Consignes : Pour chacune des questions suivantes, identifiez :

- **Variable dépendante (VD)** – ce que l'on cherche à expliquer.
- **Variable indépendante (VI)** – ce que l'on suppose influencer la VD.

Questions :

1. Les quartiers avec plus d'espaces verts ont-ils un taux de criminalité plus faible ?
2. Comment le pourcentage de logements sociaux influence-t-il les prix immobiliers ?
3. Le niveau de pollution de l'air affecte-t-il le nombre d'admissions hospitalières pour maladies respiratoires ?
4. Quel est l'effet du télétravail sur l'usage quotidien des transports publics ?
5. L'accès aux pistes cyclables influence-t-il l'usage du vélo en ville ?

Correction : Variables dépendantes vs indépendantes

Réponses :

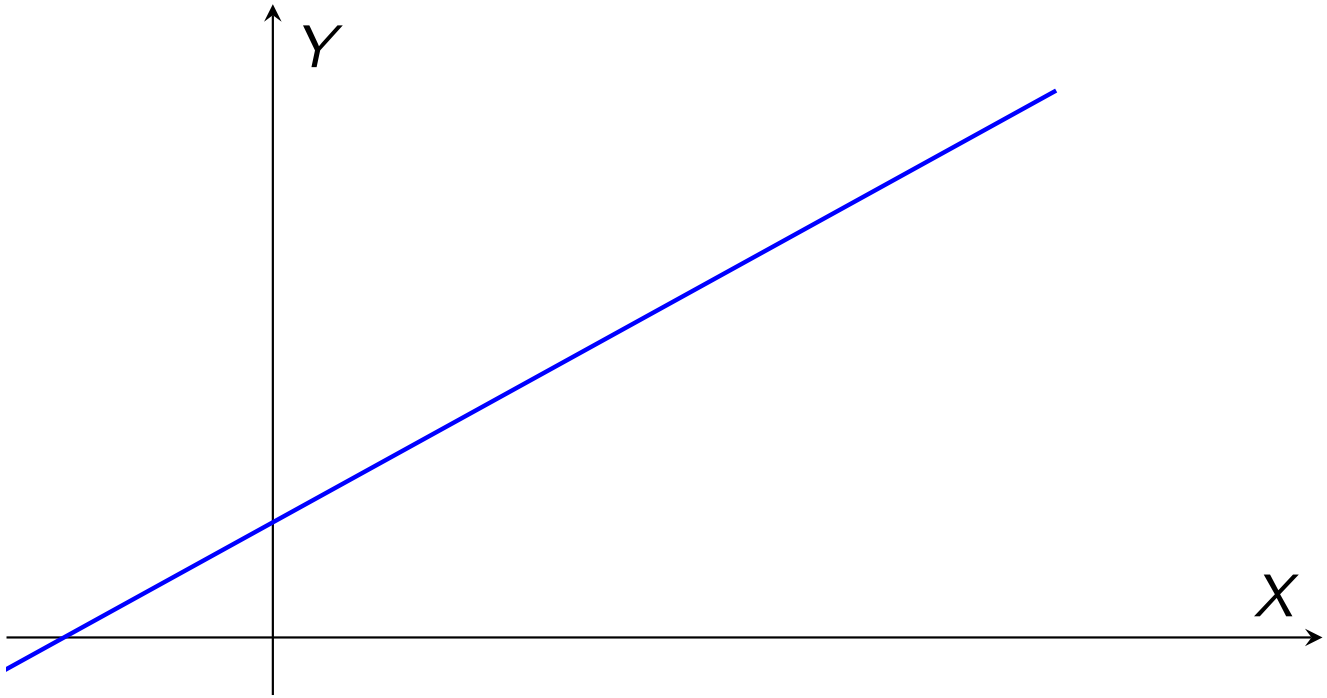
1. **VD :** Taux de criminalité
   VI : Espaces verts.
2. **VD :** Prix immobiliers
   VI : Pourcentage de logements sociaux.
3. **VD :** Admissions hospitalières (respiratoires)
   VI : Niveau de pollution.
4. **VD :** Usage des transports publics
   VI : Télétravail.
5. **VD :** Usage du vélo
   VI : Accès aux pistes cyclables.

Toutes les relations sont-elles mesurables ?

- Toutes les relations ne sont pas **mesurables**. Certaines existent conceptuellement ou socialement sans représentation numérique claire.
- Mais lorsque c'est possible, une relation doit être exprimée sous forme de **fonction mathématique**.
- Il n'existe pas une unique fonction – différents modèles décrivent différents types de relations.
- Nous allons nous concentrer sur une fonction particulière : la **fonction linéaire**.

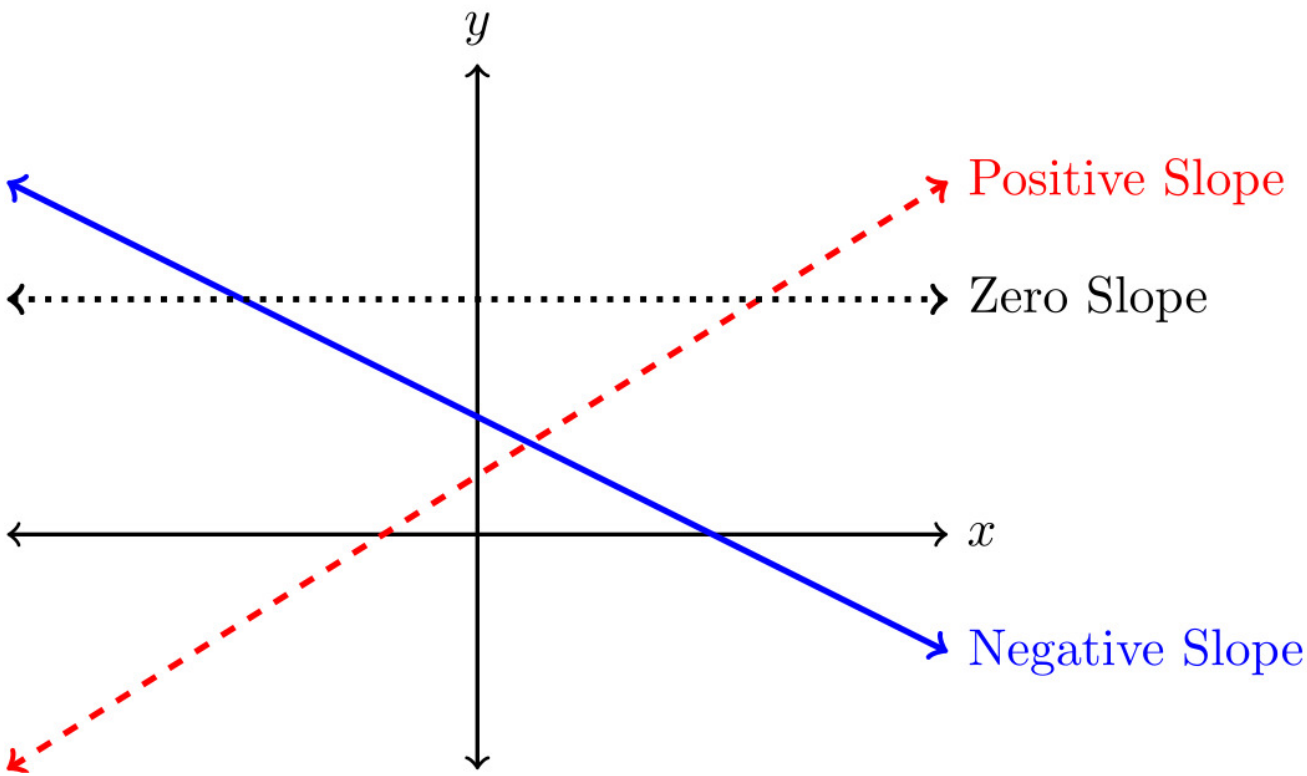
Pourquoi se concentrer sur la fonction linéaire ?

- La **fonction linéaire** est un outil fondamental pour modéliser les relations entre variables.
- Elle suppose un **effet proportionnel** : une augmentation d'une unité de X entraîne un changement constant de Y.



Deux propriétés clés de la droite ajustée

- **Pente (b)** :
  - Si la pente est positive, Y augmente lorsque X augmente.
  - Si la pente est négative, Y diminue lorsque X augmente.
  - Une pente plus raide indique une relation plus forte.



Deux propriétés clés de la droite ajustée

- **Dispersion** :
- Si les points sont fortement regroupés autour de la droite, la relation est forte.
- S'ils sont très dispersés, la relation est faible.
- Cette dispersion permet de mesurer la **force de la relation**.

Voir le graphique

Le cœur de la modélisation statistique : l'ajustement de fonctions

La statistique, c'est l'ajustement

- Au fond, la statistique consiste à ajuster des fonctions mathématiques aux données.
- On prend une forme mathématique prédéfinie — ici, une droite — et on l'ajuste le mieux possible aux observations.
- La plupart des méthodes statistiques appartiennent à deux grandes catégories :
  - **Fonctions mathématiques** : modèles linéaires, polynomiaux, etc.
  - **Distributions probabilistes** : normale, exponentielle, etc., pour modéliser l'aléa.
- Si l'on retient cette idée simple — la statistique comme ajustement de fonctions — le reste s'éclaire naturellement.

Covariance : mesurer l'association linéaire

Définition : La covariance mesure la tendance de deux variables X et Y à varier ensemble.

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

Annotations :

- $\text{Cov}(X, Y)$: covariance entre X et Y.
- n : nombre total d'observations.
- X_i, Y_i : valeurs individuelles.
- $\bar{X}, \bar{Y}$: moyennes respectives.
- $(X_i - \bar{X}), (Y_i - \bar{Y})$: écarts à la moyenne.

Covariance : interprétation

- Si $\text{Cov}(X, Y) > 0$: valeurs élevées de X associées à valeurs élevées de Y (**association positive**).
- Si $\text{Cov}(X, Y) < 0$: valeurs élevées de X associées à valeurs faibles de Y (**association négative**).
- Si $\text{Cov}(X, Y) = 0$: pas de relation linéaire entre les deux variables.

Corrélation : une covariance standardisée

- La covariance dépend des unités de mesure, difficile à interpréter.
- Le coefficient de corrélation de Pearson la standardise :

$$r_{XY} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

- Propriétés :
  - r_{XY} est toujours entre -1 et 1.
  - $r_{XY} > 0$: relation **positive**.
  - $r_{XY} < 0$: relation **négative**.
  - $r_{XY} = 0$: **pas de relation linéaire**.

Association pour variables catégorielles

Covariance et corrélation s'appliquent aux variables continues.

- Pour les variables catégorielles, on utilise les **tableaux de contingence**.
- Exemple : survie sur le Titanic.

	Décès	**Survivants**
Femmes	154	308
Hommes	709	142

Table: Taux de survie selon le genre sur le Titanic.

Choisir le bon test : selon le type de variable

- Comme en univarié, **tous les tests ne s'appliquent pas à toutes les variables**.
- Le type de variable — **qualitative** ou **quantitative** — détermine le test possible.
- Certains tests exigent des données **continues**, d'autres des données **catégorielles**.

Quand la corrélation échoue : relations non linéaires

- La corrélation de Pearson ne capte que les relations **linéaires**.
- Exemple : si $Y = X^2$, la corrélation peut être proche de zéro malgré une forte relation.
- Solutions :
  - Utiliser des **nuages de points**.
  - Employer la **corrélation de Spearman** pour relations monotones (à voir plus tard).

Pourquoi association n'est pas causalité

- Deux variables associées ne signifient pas que l'une cause l'autre.
- La corrélation décrit le mouvement conjoint, pas l'explication.
- Exemple : les étudiants qui boivent plus de café réussissent mieux. Le café **cause-t-il** de meilleures notes ? Pas forcément.

Variables confondantes : le troisième facteur caché

- Une **variable confondante** influence les deux variables observées.
- Exemple :
  - Ventes de glace et noyades sont corrélées.
  - Mais manger une glace ne provoque pas de noyade.
  - Cause réelle : **la chaleur estivale**.
- Importance de contrôler les variables confondantes.

Corrélations fallacieuses : le hasard est trompeur

- Certaines corrélations existent par **coïncidence**.
- Exemple :
  - Consommation de chocolat par habitant et nombre de prix Nobel par pays.
  - Pure coïncidence, pas causalité.
- Solution :
  - Vérifier la **significativité statistique**.
  - Employer des méthodes d'**inférence causale**.

Comment établir la causalité ?

- Pour affirmer une relation causale, il faut plus qu'une corrélation.
- Toujours se demander : **Quoi d'autre pourrait expliquer cette relation ?**

Conditions de la causalité (I)

- Trois conditions classiques doivent être réunies pour affirmer qu'une variable X cause une variable Y :
  1. **Association empirique** : X et Y doivent être corrélés.
  2. **Ordre temporel** : la cause doit précéder l'effet dans le temps.
  3. **Non-spuriousness** : l'association ne doit pas être due à une variable confondante.
- Sans ces conditions, il est impossible d'établir un lien de causalité solide.

Conditions de la causalité (II)

- Au-delà des conditions minimales, des critères plus exigeants permettent de renforcer l'inférence causale :
  - **Mécanisme théorique** : existence d'un processus explicatif reliant X à Y.
  - **Réplication empirique** : le lien se vérifie dans différents contextes et échantillons.
  - **Expérimentation ou quasi-expérimentation** : assignation aléatoire ou méthodes d'inférence causale pour isoler l'effet de X.
- Ainsi, la causalité est toujours un **jugement raisonné**, reposant à la fois sur la théorie et sur des preuves empiriques.

Étude de cas : le choléra à Londres (1854)

- Au XIXe siècle, Londres connaît plusieurs épidémies meurtrières de **choléra**.
- La théorie dominante est celle des miasmes : l'air vicié serait responsable des maladies.
- Le médecin **John Snow** doute de cette explication et soupçonne une autre cause : **l'eau contaminée**.
- Son enquête sur le quartier de Soho devient un exemple classique de raisonnement causal en sciences sociales et médicales.

Observation : la pompe de Broad Street

- Snow cartographie les cas de choléra autour des pompes publiques.
- Constat frappant : la majorité des décès se concentre autour de la **pompe de Broad Street**.
- Intervention : il fait retirer la poignée de la pompe.
- Résultat : les cas de choléra diminuent immédiatement dans le quartier.

Vers une logique d'inférence causale

- Trois conditions de la causalité sont réunies :
  1. **Association** : zones proches de la pompe = forte incidence du choléra.
  2. **Ordre temporel** : la fermeture de la pompe précède la baisse des cas.
  3. **Non-spuriousness** : d'autres facteurs (météo, air) ne suffisent pas à expliquer la différence.
- Comparaison implicite :
  - **Avant / après** : évolution des cas à Soho avant et après la fermeture.
  - **Avec / sans** : quartiers avec égouts modernes vs sans égouts.
- Cette logique anticipe une approche **difference-in-differences** : mesurer l'effet d'un traitement (égouts / fermeture de pompe) en comparant l'évolution de deux groupes.

Illustration de la dispersion

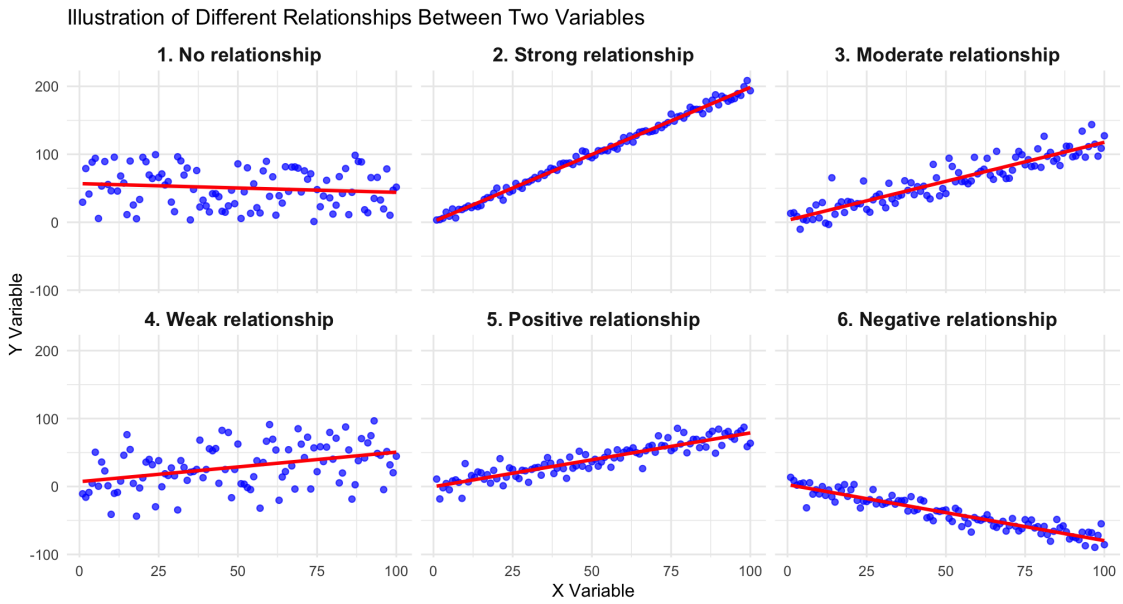


Figure: Illustration de différentes relations entre deux variables

Retour à la diapositive précédente