Analyse Statistique Univariée et Bivariée

Mastère Data & Intelligence artificielle Chef de projet data et IA

Badmavasan KIROUCHENASSAMY

Sorbonne Université

16, Octobre, 2025

Série statistique double (ou bivariée)

Définition

Soit $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$ une population d'effectif n, sur laquelle on étudie deux caractères quantitatifs :

$$X, Y : \Omega \to \mathbb{R},$$

avec X supposé non constant.
Pour tout individu $i \in \{1, \dots, n\}$:

$$x_i = X(\omega_i), \quad y_i = Y(\omega_i).$$

On appelle **série statistique double (ou bivariée)** de la population Ω pour le couple de caractères (X, Y) la donnée du n-uplet :

$$\big((x_i, y_i)\big)_{1 \le i \le n}.$$

Série statistique double (ou bivariée)

Exemple 1.1 Taille et poids de 10 enfants de 6 ans

On note :

$$X \text{ : taille (en cm)}, \quad Y \text{ : poids (en kg)}.$$

Enfant	1	2	3	4	5	6	7	8	9	10
X_i (taille)	121	123	108	118	111	109	114	103	110	115
Y_i (poids)	25	22	19	24	19	18	20	15	20	21

Observation : chaque couple (x_i, y_i) représente un individu (enfant) avec ses deux caractéristiques : sa **taille** et son **poids**. On peut ensuite étudier leur **relation** (corrélation, régression, nuage de points, etc.).

Corrélation : notion générale

Définition

La **corrélation** mesure la **force** et la **direction** de la relation entre deux variables quantitatives.

- Une corrélation ne signifie pas **causalité** : deux variables peuvent être liées sans lien de cause à effet.
- Deux types principaux :
  1. Corrélation de **Pearson** linéaire, paramétrique.
  2. Corrélation de **Spearman** basée sur les rangs, non paramétrique.

Coefficient de corrélation : valeurs et interprétations

Intervalle des valeurs

Un coefficient de corrélation (Pearson r ou Spearman ρ_s) prend ses valeurs dans

$$[-1,\ 1].$$

- **Corrélation négative (proche de -1)** : quand la **variable 1 augmente**, la **variable 2 diminue.** Relation décroissante forte.
- **Corrélation positive (proche de $+1$)** : quand la **variable 1 augmente**, la **variable 2 augmente.** Relation croissante forte.
- **Proche de** 0 : **faible** voire **absence** de corrélation (au sens mesuré). *Attention :* corrélation nulle n'implique pas absence de relation *non linéaire*.

Bonnes pratiques : toujours visualiser (nuage de points), vérifier les **outliers** et la **linéarité** avant d'interpréter un coefficient de corrélation.

Corrélation de Pearson

Définition

Le **coefficient de corrélation linéaire de Pearson**, noté r, mesure l'intensité et la direction d'une relation **linéaire** entre deux variables quantitatives.

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

- $r \in [-1, 1]$
  - $r = 1$ corrélation positive parfaite (linéaire croissante)
  - $r = -1$ corrélation négative parfaite (linéaire décroissante)
  - $r = 0$ absence de corrélation linéaire
- Hypothèse : les deux variables sont **quantitatives continues** et **normalement distribuées.**
- Très sensible aux **valeurs aberrantes.**

Corrélation de Pearson

Storytelling : Un $r = 0.82$ entre les heures travaillées et le revenu hebdomadaire indique une forte relation positive : « **Plus on travaille d'heures, plus le revenu tend à croître** », mais la relation peut être influencée par des exceptions (heures supplémentaires non payées, salaires plafonnés).

Corrélation de Spearman

Définition

La **corrélation de Spearman**, notée ρ_s, est une mesure de dépendance **monotone** entre deux variables. Elle repose sur les **rangs** plutôt que sur les valeurs brutes.

$$\rho_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}, \quad \text{où } d_i = \text{rang}(x_i) - \text{rang}(y_i)$$

- $\rho_s \in [-1, 1]$ comme Pearson, mais :
  - robuste aux valeurs aberrantes ;
  - utilisable pour des **relations non linéaires mais monotones** ;
  - applicable à des **variables ordinales**.
- Hypothèses : pas de distribution normale requise.

Corrélation de Spearman

Exemple (storytelling) : Pour les variables ‘education-num‘ et ‘income‘, une corrélation de Spearman $\rho_s = 0.61$ signifie qu’**à mesure que le niveau d’éducation augmente**, le revenu tend globalement à croître, **même si la relation n’est pas strictement linéaire**.

Corrélation Causalité

Idées clés

- **Corrélation** : deux variables varient **ensemble** (relation statistique), sans direction.
- **Causalité** : une variable **influence** directement l'autre ($X \to Y$).
- Une corrélation peut être due à : **hasard**, **biais de mesure**, **variable confondante** Z, ou **réciprocité** ($X \leftrightarrow Y$).

Exemples

- *Ventes de glaces* $\uparrow$ et *noyades* $\uparrow$: corrélées par la **température** Z (été), pas causales.
- *Publicité* $\leftrightarrow$ *ventes* : relation possiblement bidirectionnelle (causalité inversée).

Corrélation Causalité

Établir la causalité (indices)

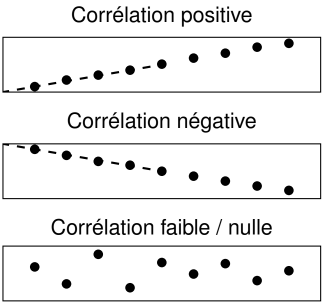
- **Antériorité temporelle** : X précède Y.
- **Contrôle des confounders** : appariement, régression, *propensity score*.
- **Expérimentation** : A/B test, randomisation.
- **Causal Graphs** : DAGs (Z bloque/ouvre des chemins).

Règle d'or : une corrélation **suggère** une hypothèse causale, elle ne la **prouve** pas. Il faut un **design** (expérimental/quasi-expérimental) et une **analyse** adaptés.

Nuages de points : visualiser la relation entre deux variables

Définition

Un **nuage de points** représente chaque individu par un point de coordonnées (x_i, y_i), afin de visualiser la **forme**, la **direction** et la **force** de la relation entre deux variables quantitatives.



Lecture et bonnes pratiques

- **Tendance** : croissante (positive), décroissante (négative) ou aucune tendance.
- **Forme** : linéaire vs non linéaire (courbure, effet seuil).
- **Dispersion** : points serrés (relation forte) vs dispersés (faible).
- **Points atypiques** (outliers) : peuvent **influencer** fortement les mesures.
- Ajouter au besoin : **droite de tendance** / **lissage** pour guider l'oeil.

Nuages de points : visualiser la relation entre deux variables

À retenir : le nuage de points est **le** réflexe visuel avant de mesurer la corrélation (Pearson/Spearman) : il révèle la **linéarité**, les **non-linéarités**, et les **outliers** qui guideront votre analyse.

Régression linéaire et lien avec la corrélation

Idée générale

La **régression linéaire simple** cherche à modéliser la relation entre deux variables quantitatives X et Y à l'aide d'une **droite de tendance** :

$$\hat{Y} = aX + b$$

où :

- a = pente de la droite (variation moyenne de Y lorsque X augmente d'une unité) ;
- b = ordonnée à l'origine (valeur estimée de Y lorsque $X = 0$).

Régression linéaire et lien avec la corrélation

Lien avec la corrélation de Pearson

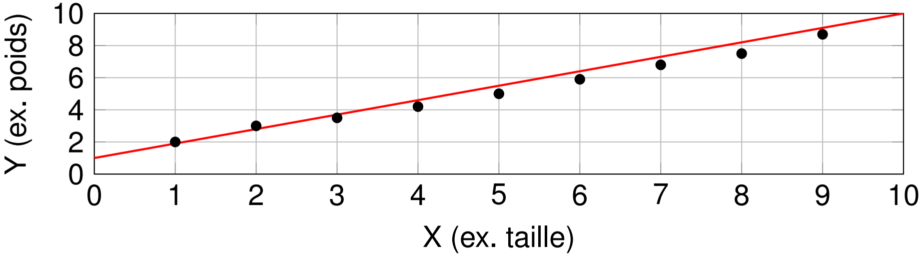
La pente *a* de la droite de régression s'écrit :

$$a = r_{XY} \times \frac{s_Y}{s_X},$$

où r_{XY} est le coefficient de corrélation linéaire. Ainsi :

- Si $r_{XY} > 0$, la pente est positive relation croissante ;
- Si $r_{XY} < 0$, la pente est négative relation décroissante ;
- Si $r_{XY} \approx 0$, la pente est quasi nulle pas de relation linéaire marquée.

Régression linéaire et lien avec la corrélation



Interprétation (storytelling) : Si la corrélation entre taille et poids est forte et positive ($r = 0.9$), la régression estime la tendance générale : *« plus un enfant est grand, plus il pèse lourd »*, tout en tenant compte des écarts individuels autour de la droite.