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 $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ 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 $i \in \{1, \ldots, n\}$:

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

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

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

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).

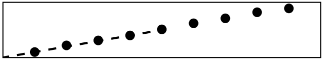
Règle d'or : une corrélation **suggère** une hypothèse causale, elle ne la **prouve** pas. Il faut un **design** (expérimental/quasi-expérimental) et une **analyse** adaptés.

Nuages de points : visualiser la relation entre deux variables

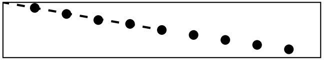
Définition

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.

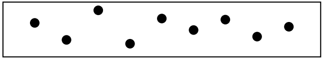
Corrélation positive



Corrélation négative



Corrélation faible / nulle



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

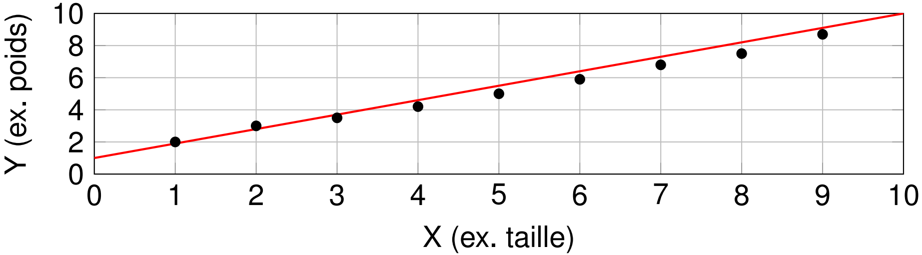
La pente *a* de la droite de régression s'écrit :

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

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.

Table de contingence (ou tableau croisé)

Définition

Une **table de contingence** permet de résumer la répartition conjointe de deux variables qualitatives. Elle présente les **effectifs croisés** observés pour chaque combinaison de modalités.

- En ligne : les modalités de la variable X.
- En colonne : les modalités de la variable Y.
- Chaque cellule : nombre d'individus ayant la combinaison (X_i, Y_j).
- Les marges (totaux par ligne et par colonne) donnent les distributions marginales.

Table de contingence (ou tableau croisé)

Exemple Sexe et niveau de diplôme On interroge 20 personnes sur leur sexe et leur niveau de diplôme :

$$\text{Variables : } X = \text{Sexe (H/F)}, \quad Y = \text{Diplôme (Secondaire, Supérieur)}.$$

	Secondaire	**Supérieur**	**Total ligne**
Hommes (H)	5	7	12
Femmes (F)	6	2	8
Total colonne	11	9	20

Lecture : 5 hommes ont un diplôme secondaire, 7 un diplôme supérieur, etc. On peut ensuite calculer :

- les **fréquences conjointes** $f_{ij} = n_{ij}/n$,
- les **fréquences marginales** (par ligne et par colonne),
- et tester l'**indépendance** entre X et Y (test du χ^2).

Remarque : ce type de tableau est la base des analyses de dépendance entre variables qualitatives.

Définition de l'indépendance entre deux variables

Définition

Deux variables qualitatives X et Y sont dites **statistiquement indépendantes** si la répartition des modalités de Y est la même, quelle que soit la modalité de X.

Formellement :

$$f_{ij} = f_{i.} \times f_{.j}$$

où :

- f_{ij} = fréquence conjointe de la cellule (i, j),
- $f_{i.}$ = fréquence marginale de la modalité X_i,
- $f_{.j}$ = fréquence marginale de la modalité Y_j.

Corrélation vs Indépendance

Différences conceptuelles

- **Corrélation** : mesure la **force et la direction d'une relation linéaire** entre deux variables quantitatives (ou monotone pour Spearman).
- **Indépendance** : notion plus générale ; signifie qu'il n'existe **aucune dépendance statistique** entre deux variables, quelle qu'en soit la forme.

Tableau comparatif

	Corrélation	**Indépendance**
Type de variable	Quantitatives (Pearson, Spearman)	Qualitatives, quantitatives ou mixtes
Mesure	Force d'une **relation linéaire ou monotone**	Absence de **relation statistique quelconque**

Corrélation vs Indépendance

À retenir :

L'indépendance implique absence de corrélation, mais une corrélation nulle n'exclut pas une relation **non linéaire** ou **structure cachée**.

Démonstration de l'indépendance entre variables

Condition mathématique d'indépendance

Soient deux variables qualitatives X et Y et leur table de contingence (n_{ij}) de taille $p \times q$, avec $n = \sum_{i,j} n_{ij}$.

$$\text{Si} \quad \forall(i,j), \quad f_{ij} = f_{i.} \times f_{.j}, \quad \text{alors } X \text{ et } Y \text{ sont indépendantes.}$$

où $f_{ij} = n_{ij}/n$.

Outil de vérification : le test du χ^2

$$\chi^2 = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{(n_{ij} - n_{ij}^{(th)})^2}{n_{ij}^{(th)}}$$

Si χ^2 est supérieur à la valeur critique (ddl = $(p-1)(q-1)$), on **rejette l'hypothèse d'indépendance** entre X et Y.

Mesurer lassociation entre deux variables : le V de Cramer

Définition

Le **coefficient V de Cramer** est une mesure de la **force dassociation** entre deux variables qualitatives, calculée à partir du **test du Chi-deux** dindépendance.

Pour une table de contingence de taille $p \times q$:

$$V = \sqrt{\frac{\chi^2}{n \times \min(p-1,\ q-1)}}$$

où :

- χ^2 = statistique du test du χ^2,
- n = effectif total de léchantillon,
- p, q = nombre de modalités de chaque variable.

Mesurer lassociation entre deux variables : le V de Cramer

Interprétation (valeurs possibles)

$$0 \leq V \leq 1$$

- $V = 0$ indépendance parfaite (aucune association),
- $V = 1$ liaison parfaite entre les variables,
- Plus V est grand, plus la dépendance est forte.

Exemple : Pour une table Sexe x Niveau de diplôme, si $\chi^2 = 6.48$, $n = 200$ et $\min(p-1, q-1) = 1$,

$$V = \sqrt{\frac{6.48}{200}} = 0.18.$$

Association faible mais non nulle : la répartition des diplômes diffère légèrement selon le sexe.

Mesurer lassociation entre deux variables : le V de Cramer

- le chi-deux indique sil existe une dépendance, tandis que V mesure lintensité de cette dépendance.
- On peut construire une échelle dinterprétation :
  - $V < 0.1$ liaison très faible
  - $0.1 V < 0.3$ liaison modérée
  - $V 0.3$ liaison forte

Étudier limpact dune variable qualitative sur une variable quantitative

Objectif

Analyser comment la variable quantitative Y varie selon les modalités dune variable qualitative X. Exemples :

- $Y =$ revenu mensuel, $X =$ niveau déducation
- $Y =$ âge, $X =$ sexe

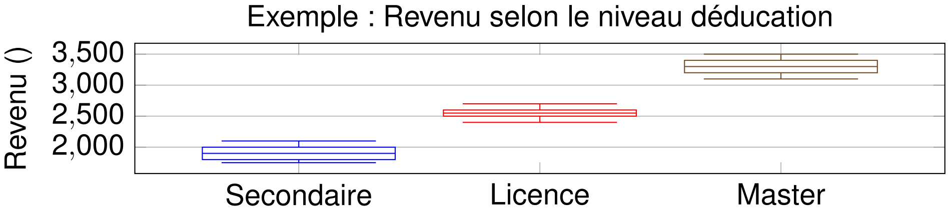
- **Visualisation** : diagrammes en boîte (*boxplots*)
- **Tests statistiques** :
  - Test de **Student** pour 2 groupes (ex. hommes/femmes)
  - Test **ANOVA** pour plus de 2 groupes (ex. 3 niveaux détudes)

Principe : On cherche à savoir si la moyenne (ou la distribution) de Y diffère significativement selon les catégories de X.

Visualisation : diagramme en boîte (boxplot)

Définition

Le **boxplot** permet de visualiser la **dispersion** et la **tendance centrale** dune variable quantitative pour chaque modalité dune variable qualitative.



Interprétation :

- La médiane (Q_2) sépare la moitié inférieure et supérieure des revenus.
- Lécart interquartile (IQR) mesure la dispersion.
- On observe ici une **tendance croissante** du revenu avec le niveau déducation.
- Les points extrêmes (outliers) signalent des cas atypiques.

Test de Student : comparaison de deux moyennes

Objectif

Vérifier si la moyenne dune variable quantitative Y est la même dans deux groupes définis par une variable qualitative X.

$$H_0 : \mu_1 = \mu_2 \quad \text{vs} \quad H_1 : \mu_1 \neq \mu_2$$

- μ_1 et μ_2 : moyennes de Y dans les deux groupes.
- Hypothèse : Y suit une distribution normale dans chaque groupe et les variances sont comparables.

Test de Student : comparaison de deux moyennes

Statistique du test

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \text{où} \quad s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Interprétation :

- Si p-value < 0.05 on **rejette** H_0 : la différence de moyennes est significative.
- Sinon, aucune différence notable nest détectée.

Exemple : comparer le revenu moyen des hommes et des femmes.

ANOVA : comparaison de plusieurs moyennes

Objectif

LANOVA (Analysis of Variance) généralise le test de Student à $k > 2$ groupes. Elle permet de vérifier si les moyennes de Y diffèrent selon les modalités dune variable qualitative X.

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \quad \text{vs} \quad H_1 : \text{au moins une moyenne est différente.}$$

ANOVA : comparaison de plusieurs moyennes

Principe

On décompose la variance totale :

$$s_T^2 = s_B^2 + s_W^2$$

où :

- s_B^2 = variance **entre** les groupes (effet de X),
- s_W^2 = variance **à lintérieur** des groupes.

ANOVA : comparaison de plusieurs moyennes

Statistique du test

$$F = \frac{s_B^2}{s_W^2}$$

On rejette H_0 si la **p-valeur** associée à F est inférieure à 0.05.

Interprétation :

- Si $p < 0.05$ la variable qualitative a un **effet significatif** sur Y.
- Sinon pas de différence significative entre les moyennes des groupes.

ANOVA : comparaison de plusieurs moyennes

- On commence toujours par visualiser les groupes (boxplots)
- On teste ensuite les différences : Student pour 2 groupes, ANOVA pour plusieurs.
- Si le test est significatif, on conclut que la variable qualitative influence la variable quantitative.
- Ces analyses permettent de quantifier limpact dun facteur catégoriel (sexe, diplôme, statut, etc.) sur une mesure continue (revenu, âge, score, etc.)