# Analyse Statistique Univariée et Bivariée

Mastère Data & Intelligence artificielle Chef de projet data et IA

**Badmavasan KIROUCHENASSAMY**

Sorbonne Université

15, Octobre, 2025

# Objet de la statistique descriptive

## Objectif

La statistique descriptive sert à décrire une population [ **un (gros) ensemble d'unités statistiques élémentaires** ] à l'aide d'indicateurs numériques ou de techniques graphiques.

# Population, Individu

## Définition

**L'individu** est **l'unité** statistique à laquelle on s'intéresse.

## Définition

**La population** est **l'ensemble fini des individus** statistiques que l'on s'apprête à décrire.

- Les individus statistiques peuvent être des humains, des êtres vivants (animaux, végétaux), mais aussi des objets (voitures, livres, ...), des entités juridiques (entreprises, départements, pays,...)
- **Exemple** : population composée des livres catalogués à la bibliothèque centrale de l'Université de Lille.
- **Notation** : On désigne par N le nombre d'individus (distincts) de la population. Ce nombre est appelé taille de la population.

# Variable

## Définition

**Une variable** statistique est un moyen de **décrire chacun des individus de la population**. Caractère est synonyme de variable.

- Une même population peut être décrite à l'aide de plusieurs variables.
- **Exemple** : chacun des livres catalogués à la bibliothèque peut être décrit par son année de publication, la couleur de sa couverture, son nombre de pages, la discipline dont il traite, ses dimensions, son prix à l'achat, ...
- **Notation** : On désigne les variables statistiques par des lettres latines majuscules. On parlera des variables X, Y, ...

# Modalité ou valeur

## Définition

Une **modalité** d'une variable est une des façons possibles d'effectuer la description d'un individu au moyen de cette variable. Valeur est synonyme de modalité.

- On évitera la terminologie « valeur » et on lui préfèrera « modalité » car une valeur d'une variable statistique n'est pas nécessairement une valeur numérique.
- **Exemple** : si on décrit les livres de la bibliothèque à l'aide de la variable « couleur de la couverture » , une modalité possible est « bleu » . Si on décrit ces livres à l'aide du nombre de pages, une modalité possible est 436.

# L'ensemble des modalités d'une variable

## Définition

L'ensemble des modalités d'une variable est l'ensemble des différentes façons possibles de décrire les individus de la population avec la variable.

- **Notation** : Si **X** dénote une variable statistique, l'ensemble de ses modalités est noté $\boldsymbol{M_X}$. Un élément typique de cet ensemble est noté $x$.
- On dit qu'un individu de la population présente la modalité $x \in \boldsymbol{M_X}$ de la variable **X** si cet individu est décrit par $x$ au moyen de la variable **X**.
- **Exemple** : Si on décrit les livres de la bibliothèque par leur dimension (variable **X**), une modalité $\boldsymbol{x}$ de **X** sera un triplet de nombres positifs $\boldsymbol{x} = (x_b, x_l, x_e)$ dont les composantes désignent la hauteur, la largeur et l'épaisseur du livre.

# La constitution de l'ensemble des modalités

## Principe

Il faut apporter un soin particulier à la constitution de cet ensemble. Pour chaque variable servant à décrire la population, l'ensemble de ses modalités doit être constitué de sorte qu'à chaque individu de la population on puisse assigner une modalité et une seule de la variable.

- Cela signifie deux choses :
  1. Il ne doit pas y avoir d'individu dans la population ne présentant aucune modalité de l'une des variables.
  2. Il ne doit pas y avoir d'individu dans la population présentant plusieurs modalités d'une même variable.
- **Exemple** : Pour la variable « discipline traitée » , il faut choisir la discipline la plus pertinente pour les livres pouvant appartenir à plusieurs disciplines.

# Classification des variables : discrètes

## Variables discrètes

Une variable **X** est dite **discrète** si $M_X$ est un ensemble dénombrable.

- **Exemples** : Les livres décrits selon leur nombre de pages ; les livres décrits selon la discipline à laquelle ils sont rattachés.

# Classification des variables : continues

## Variables continues

Une variable **X** est une variable **continue** si $M_X$ est un ensemble non-dénombrable.

- **Exemple** : Les variables qui décrivent des grandeurs liées au temps, à l'espace, à la masse, etc.
- **Propriété** : Si on choisit deux modalités, toutes les valeurs comprises entre ces deux modalités sont aussi des modalités de la variable.

# Classification des variables : qualitatives

## Variables qualitatives

Une variable est dite **qualitative** si ses modalités ne sont pas quantifiables à l'aide d'une échelle quelconque.

- **Exemple** : Les modalités de la variable « discipline traitée » .

# Classification des variables : quantitatives

## Variables quantitatives

Une variable est dite **quantitative** si ses modalités peuvent être considérées comme des quantités exprimées dans une échelle de valeurs.

- **Exemple** : Les modalités de la variable « nombre de pages » .

# Classification des variables : Variables quantitatives, Variables qualitatives

## Comparaison des approches

Le même phénomène (le prix) peut donner deux variables différentes selon la façon dont on définit les modalités :

- **Variable quantitative** : par une valeur numérique (p. ex. 23,50 €).
- **Variable qualitative** : par des catégories (bon marché / cher / très cher).

# Classification des variables : Variables quantitatives, Variables qualitatives

## Remarque

En pratique, lorsqu'on est en présence d'une variable répondant à la définition d'une variable statistique continue, la précision avec laquelle les modalités de cette variable sont relevées rend la variable discrète.

- **Exemple** : L'épaisseur "vraie" est continue, mais l'instrument de mesure ne l'est pas : il mesure par pas (p. ex. au mm). Toute valeur est donc arrondie au pas de mesure $\rightarrow$ quantification.

# Classification des variables : Variable nominale

## Définition

On dit que **X** est une variable **nominale** s'il n'est pas possible de définir de façon naturelle un ordre sur l'ensemble $\boldsymbol{M_X}$ de ses modalités.

- Pour de telles variables, on pourra toujours sélectionner deux éléments $\boldsymbol{x}$ et $\boldsymbol{x'}$ de $\mathbf{M_X}$ pour lesquels aucune comparaison ayant un sens n'est possible.
- **Exemples** :
  - Pour la variable « discipline » servant à décrire les livres de la bibliothèque, il n'y a aucune façon naturelle de comparer « sociologie » et « littérature anglaise » .
  - Typiquement, toute variable dont les modalités sont formées sur la base d'un système de codage sont des variables nominales : sexe (codage 0,1), catégories socioprofessionnelles (nomenclature de l'INSEE), etc.

# Traitements statistiques : Variable nominale

## Traitements possibles

Les seuls traitements dont les résultats ont un sens sont ceux qui consistent en des opérations de dénombrement (comptage), ou qui reposent sur de telles opérations.

- On peut donc calculer des effectifs, des fréquences des modalités de la variable et utiliser le mode de la variable.

# Traitements statistiques : Variable nominale

## Traitements impossibles

L'absence de relation d'ordre implique qu'il est impossible d'utiliser tout outil statistique construit à partir de l'existence d'une relation d'ordre (cumuls, quantiles, ...).

- Les opérations usuelles sur les nombres réels $(+, -, \times)$ ne peuvent pas être définies (par exemple dans le cas des livres décrits selon leur discipline), ou bien produisent des résultats sans signification (par exemple dans le cas des personnes décrites par le numéro minéralogique de leur département de naissance).

# Classification des variables : Variable ordinale

## Définition

On dit que **X** est une variable **ordinale** si :

1. Il est possible de définir un ordre sur l'ensemble $\mathbf{M_X}$ de ses modalités.
2. Les écarts et les relations de proportionnalité entre modalités de **X** n'ont aucune signification.

- **Exemples** :
  - Notes attribuées à des objets en fonction d'un classement.
  - Toute variable mesurant des préférences.

# Traitements statistiques : Variable ordinale

## Traitements possibles

Il est possible d'utiliser les mêmes outils que pour des variables nominales.

- De plus, les objets statistiques établis à partir d'une relation d'ordre sur l'ensemble des modalités de la variable ont un sens.

## Traitements impossibles

On ne peut pas effectuer des traitements qui exploitent autre chose que la relation d'ordre sur l'ensemble des modalités (écarts, ...).

# Classification des variables : Variable numérique

## Définition

On dit que **X** est une variable **numérique** si :

1. Il est possible de définir un ordre sur l'ensemble $\mathbf{M_X}$ de ses modalités.
2. Les écarts et les relations de proportionnalité entre modalités de **X** ont une interprétation.

- **Exemples** :
  - La variable « prix d'achat en € » utilisée pour décrire les livres appartient à l'échelle proportionnelle.
  - De façon générale, toute variable dont les modalités sont exprimées dans une unité de mesure numérique, possédant un zéro traduisant l'absence de phénomène (quantité nulle), est numérique.
  - Les variables dont les modalités sont des nombres réels servant à exprimer des quantités sont donc des variables numériques.

# Classification des variables : Variable numérique vs ordinale

## Variable numérique (quantitative)

- Prix du livre en euros : 12,90 €, 23,50 €, 41,00 €, ...
- Ce que ça veut dire : chaque valeur est un nombre mesuré. On peut calculer des moyennes, des écarts, des ratios (un livre coûte $2\times$ plus qu'un autre), etc.

## Variable ordinale (qualitative ordonnée)

- Niveau de cherté : bon marché $<$ moyen $<$ cher $<$ très cher.
- Ce que ça veut dire : ce sont des catégories avec un ordre, mais sans échelle métrique. On peut dire qu'un livre est plus cher qu'un autre, mais on ne sait pas de combien.

# Effectif

## Définition

L'**effectif** d'une modalité $x$ d'une variable $X$ est le nombre d'individus de la population qui présentent cette modalité. On le note généralement $n_x$.

- **Exemple** : Considérons une bibliothèque de 100 livres décrits par leur discipline :
  - 25 livres en économie, 30 en histoire, 15 en philosophie, 20 en littérature, 10 en sciences.
- **Effectifs** :
  - $n_{\text{économie}} = 25$, $n_{\text{histoire}} = 30$, $n_{\text{philosophie}} = 15$, $n_{\text{littérature}} = 20$, $n_{\text{sciences}} = 10$.

**Remarque** : La somme des effectifs de toutes les modalités est égale à la taille de la population : $\sum_{x \in \mathbf{M}_X} n_x = N$. Dans l'exemple : $25 + 30 + 15 + 20 + 10 = 100 = N$.

# Fréquence

## Définition

La **fréquence** d'une modalité $x$ est le rapport entre l'effectif de cette modalité et la taille de la population. On la note $f_x$ et elle est définie par :

$$f_x = \frac{n_x}{N}$$

- **Exemple** : Pour les 100 livres de la bibliothèque :
  - Fréquence de la modalité "économie" : $f_{\text{économie}} = \frac{25}{100} = 0,25$ (ou 25%).
  - Fréquence de la modalité "histoire" : $f_{\text{histoire}} = \frac{30}{100} = 0,30$ (ou 30%).

# Fréquence

## Interprétation

- L'**effectif** $n_k$ mesure l'importance **absolue** de la sous-population présentant la modalité $x_k$.
- La **fréquence** $f_k$ mesure cette importance **relativement** à la taille $N$ de la population, et est donc indépendante de $N$.

On appelle **distribution statistique** de la variable $X$ dans la population $P$ la donnée des $K$ couples $(x_k, f_k)$, pour $k = 1, \ldots, K$.

- Ces couples indiquent avec quelle importance chacune des modalités de $X$ est représentée au sein de la population.
- La distribution statistique constitue le point de départ de toute analyse statistique.

# Fréquence

## Remarque

La somme des fréquences de toutes les modalités est égale à 1 (ou 100%) :

$$\sum_{x \in \mathbf{M}_X} f_x = 1$$

# Effectif cumulé croissant : Définition

## Définition formelle

On appelle **effectif cumulé croissant** de (ou associé à) la modalité $x_k$ d'une variable $X$ ordonnée le nombre noté $N_k$ et défini par :

$$N_k = \sum_{j=1}^{k} n_j$$

où :

- $n_j$ est l'effectif de la modalité $x_j$,
- $x_1 \leq x_2 \leq \ldots \leq x_k \leq \ldots \leq x_K$ (modalités ordonnées).

# Effectif cumulé croissant

## Interprétation

$N_k$ représente le nombre total d'individus de la population présentant une modalité **inférieure ou égale** à $x_k$.

## Exemple introductif

Pour une variable $X$ = "Note sur 20" avec modalités ordonnées $x_1 = 10$, $x_2 = 12$, $x_3 = 14$ :

- Si $n_1 = 2$, $n_2 = 5$, $n_3 = 3$,
- alors $N_1 = 2$, $N_2 = 2 + 5 = 7$, $N_3 = 7 + 3 = 10$.

# Effectif cumulé croissant : Propriétés et remarques

**Propriétés mathématiques**

- **Croissance** : $N_1 \leq N_2 \leq \ldots \leq N_K$
- **Égalité finale** : $N_K = N$ (taille totale de la population)
- **Relation avec les fréquences** : $F_k = \frac{N_k}{N}$ (fréquence cumulée)

**Remarques importantes** :

- Applicable uniquement aux variables **ordinales ou numériques** (modalités ordonnables)
- Permet de calculer des **quantiles** (médiane, quartiles, etc.)
- Base pour construire des **courbes cumulatives** et **diagrammes de Pareto**
- Indépendant des valeurs absolues des effectifs (contrairement aux $n_k$)

**Attention**

Pour les variables **qualitatives nominales** (non ordonnables), la notion d'effectif cumulé n'a **aucun sens**.

# Représentations graphiques des distributions statistiques

## Objectif des graphiques

Les graphiques usuels représentent la répartition des modalités d'une variable $X$ au sein d'une population $P$ en associant à chaque modalité un élément graphique dont la taille (longueur, hauteur, surface, etc.) est proportionnelle à son importance dans la population.

- **Principe des tailles absolues** :
  - À chaque modalité $x_k$, on associe un élément graphique dont la taille est **égale à l'effectif $n_k$ ou à la fréquence $f_k$** de la modalité
  - La taille totale du graphique dépend de la taille $N$ de la population
  - **Exemple** : Diagrammes en bâtons
  - **Caractéristique** : La hauteur de chaque bâton est proportionnelle à $n_k$ ou $f_k$

# Représentations graphiques des distributions statistiques

- **Principe de normalisation** :
  - La taille **totale** du graphique est normalisée à 1 unité (100%)
  - À chaque modalité $x_k$, on associe un élément dont la taille est **égale à la fréquence** $f_k$
  - **Exemples** :
    - Diagrammes en secteurs (camembert) : chaque secteur a un angle de $360^\circ \times f_k$
    - Histogrammes : la surface totale des rectangles = 1
- **Remarques importantes** :
  - Le principe (a) est adapté quand on veut visualiser les **effectifs réels**
  - Le principe (b) permet des **comparaisons entre populations** de tailles différentes
  - Pour les histogrammes, c'est la **surface** des rectangles qui représente $f_k$, pas seulement la hauteur

# Diagramme en secteurs : Principes de construction

## Principe de normalisation de la surface

Le graphique consiste en un disque dont la surface est normalisée à 1 unité de surface. Ce disque est divisé en plusieurs secteurs (parts), chacun étant associé à une et seule modalité de la variable $X$.

- La proportion de la surface totale du disque occupée par le secteur associé à la modalité $x_k$ est égale à $f_k$.

## Principe de normalisation du périmètre

Au lieu de normaliser la surface du disque à 1, on peut normaliser son périmètre à 1 unité de longueur.

- Dans ce cas, la longueur de l'arc de cercle décrit par un secteur associé à une modalité $x_k$ constitue une proportion du périmètre égale à $f_k$.

# Diagrammes en bâtons : Principes de construction

## Principe de construction

Les diagrammes en bâtons consistent en des représentations des couples $(x_k, f_k)$ (ou $(x_k, n_k)$), $k = 1, \ldots, K$, où pour chaque modalité repérée sur l'axe horizontal, on représente la fréquence (ou l'effectif) par un bâton, dont la hauteur repérée sur l'axe vertical est égale à la valeur de la fréquence (ou de l'effectif).

# Diagrammes en bâtons : Principes de construction

## Variable nominale

L'ordre dans lequel on place ses modalités le long de l'axe horizontal n'a aucune importance. On peut alors choisir par exemple l'ordre croissant ou décroissant des fréquences.

## Variable ordinale ou numérique

Il existe un ordre sur les modalités et on suppose alors que $x_1 < x_2 < \ldots < x_K$. Cet ordre doit être utilisé pour représenter les modalités sur l'axe horizontal. Ainsi, de gauche à droite sur l'axe horizontal, on placera d'abord $x_1$, puis $x_2$, etc., puis $x_K$.

# Histogramme : Définition et principes

## Utilisation de l'histogramme

L'histogramme des fréquences ne s'utilise que pour une variable numérique. Ce type de représentation n'a d'intérêt que si les modalités de la variable ont été regroupées par classes. De plus, la situation dans laquelle l'histogramme présente un intérêt particulier par rapport au diagramme en bâtons est celle où les classes sont d'amplitudes inégales.

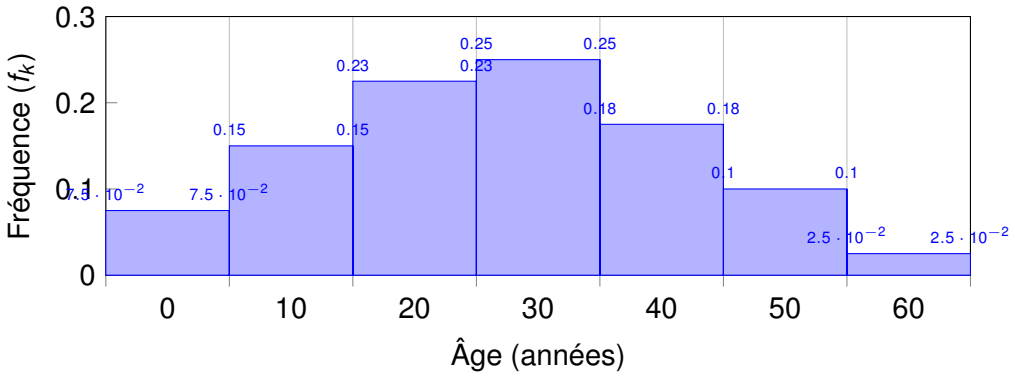
## Principe de construction

Le principe de l'histogramme consiste à normaliser la surface totale du graphique à 1 unité. Cette surface est découpée en rectangles. Chaque rectangle est associé à une classe de modalités de sorte que le rectangle associé à la classe $C_k$ représente une proportion de la surface totale égale à la fréquence $f_k$ de $C_k$.

# Histogramme : Exemple et interprétation

| Classe d'âge (années) | Effectif ($n_k$) | Fréquence ($f_k$) |
|---|---|---|
| [0-10[ | 15 | 0.075 |
| [10-20[ | 30 | 0.15 |
| [20-30[ | 45 | 0.225 |
| [30-40[ | 50 | 0.25 |
| [40-50[ | 35 | 0.175 |
| [50-60[ | 20 | 0.10 |
| [60-70] | 5 | 0.025 |

# Histogramme : Exemple et interprétation



# Histogramme : Exemple et interprétation

## Interprétation de l'histogramme

- **Forme générale** : Distribution unimodale avec un pic autour de 30-40 ans
- **Classe modale** : $[30 - 40[$ (25% de la population)
- **Concentration** : 65% de la population a entre 20 et 50 ans
- **Classes extrêmes** :
  - $[0 - 10[$ et $[60 - 70]$ représentent seulement 10% de la population
  - Sous-représentation des très jeunes et seniors
- **Amplitudes inégales** : Bien que les classes aient ici la même amplitude (10 ans), l'histogramme montre clairement que :
  - La surface du rectangle [30-40[ est la plus grande (25%)
  - Les surfaces sont proportionnelles aux fréquences

# Généralités sur la description numérique univariée

## 1. Contexte de la description numérique

Dans un contexte univarié, la description numérique des données consiste à résumer à l'aide de modalités d'une variable $X$ certaines des caractéristiques de la population décrite par $X$ (ou encore de la distribution statistique de $X$, telle que définie à la section 4.2).

## 2. Utilisation des indicateurs

Ces descriptions sont obtenues à l'aide d'indicateurs. Chacun de ces indicateurs est construit de façon à pouvoir capturer et résumer une et une seule des caractéristiques de la distribution de $X$. Par conséquent, si on s'intéresse à plusieurs des propriétés de cette distribution, on est amené à utiliser simultanément plusieurs indicateurs.

# Généralités sur la description numérique univariée

## 3. Spécificité des indicateurs

En revanche, pour décrire l'une des caractéristiques d'une distribution, on peut utiliser plusieurs indicateurs. Chacun aura des propriétés propres qui constitueront des avantages ou des inconvénients par rapport aux autres. Quelles que soient ces propriétés, tout indicateur doit être construit de façon à effectivement capturer la caractéristique souhaitée. Un indicateur ne peut donc s'utiliser de façon interchangeable pour capturer plusieurs caractéristiques possibles.

# Indicateurs de position

## Tendance Centrale

Les indicateurs de position ou indicateurs de tendance centrale permettent de résumer par une seule modalité la valeur de toutes les modalités observées de $X$. Ils sont construits pour pouvoir décrire l'endroit de l'ensemble des modalités de $X$ autour duquel se positionnent les données $X(1), \ldots, X(N)$. De façon générale, cet endroit est appelé tendance centrale.
Les indicateurs de position les plus couramment utilisés sont le mode, la médiane et la moyenne.

# Le mode

## Définition

Le mode de la variable $X$ est la modalité la plus fréquemment observée dans la population $P$. On note $\mathrm{Mo}(X)$ le mode de $X$.

## Interprétation

Cet indicateur mesure la position des données en utilisant un principe de représentation (relativement) majoritaire : la position des données est décrite par la modalité la plus souvent observée.

# La médiane

La médiane d'une variable $X$ est un indicateur déterminé à partir d'un classement des individus selon un critère d'ordre des modalités de $X$. Ceci n'est évidemment possible que s'il existe un ordre sur $\mathcal{M}_X$. Par conséquent la médiane est un indicateur qui ne s'utilise que pour des variables ordinales ou numériques.

## Définitions

Soit $X$ une variable statistique et soit $i$ un individu de la population. Le rang de $i$, noté $R(i)$, est sa position dans un classement des individus de la population par ordre croissant des modalités de $X$.

# Définitions de la médiane

## Définition

L'individu médian est un individu désigné par $i_{\text{Me}}$ tel que $R(i_{\text{Me}}) = \lceil \frac{N}{2} \rceil$, où pour tout nombre réel $y$, $\lceil y \rceil$ désigne le plus petit nombre entier supérieur ou égal à $y$.[1]
Ainsi on a :

- $R(i_{\text{Me}}) = \frac{N+1}{2}$ si $N$ est impair ;
- $R(i_{\text{Me}}) = \frac{N}{2}$ si $N$ est pair.

On déduit immédiatement de la définition que :

a) $\frac{N-1}{2}$ individus précèdent $i_{\text{Me}}$ et $\frac{N-1}{2}$ suivent $i_{\text{Me}}$, si $N$ est impair ;

b) $\frac{N}{2} - 1$ individus précèdent $i_{\text{Me}}$ et $\frac{N}{2}$ individus suivent $i_{\text{Me}}$ si $N$ est pair.

# Définitions de la médiane

## Définition

La médiane de $X$, que l'on note $\text{Me}(X)$, est la modalité de $X$ présentée par l'individu médian : $\text{Me}(X) = X(i_{\text{Me}})$.

## Cas particulier pour N pair

Lorsque $N$ est pair, la médiane de $X$ est assez souvent définie comme le milieu de l'intervalle qui sépare les modalités des individus de rangs $\frac{N}{2}$ et $\frac{N}{2} + 1$. Ce milieu est obtenu en faisant la moyenne de ces deux modalités.

# La moyenne arithmétique

La moyenne arithmétique est un indicateur de position qui est déterminée sur la base d'une répartition égalitaire.

## Définition

La moyenne arithmétique (moyenne par la suite) de la variable $X$ est le nombre noté $\overline{X}$ et défini par

$$\overline{X} = \frac{1}{N} \sum_{i=1}^{N} X(i).$$

La moyenne, pour pouvoir être interprétée comme un indicateur de position, requiert de pouvoir additionner plusieurs modalités de la variable $X$. La moyenne ne s'utilise donc que pour des variables numériques.

# Comparaison : Mode, Médiane et Moyenne

## Interprétations

- **Mode** : Valeur la plus fréquente
- **Médiane** : Valeur centrale (50% des données de chaque côté)
- **Moyenne** : Répartition égalitaire (somme des valeurs divisée par N)

| | **Mode** | **Médiane** | **Moyenne** |
|---|---|---|---|
| Type de variable | Toutes | Ordinale/Numérique | Numérique |
| Sensibilité aux valeurs extrêmes | Non | Non | Oui |
| Unicité | Non | Oui | Oui |
| Existence | Toujours | Toujours | Toujours |

# Comparaison : Mode, Médiane et Moyenne

## Exemple numérique

Pour les données : 1, 2, 2, 3, 4, 7, 9

- **Mode** = 2 (valeur la plus fréquente)
- **Médiane** = 3 (4ème valeur sur 7)
- **Moyenne** = $\frac{1+2+2+3+4+7+9}{7} = 4$

# Limites des indicateurs de position et nécessité des indicateurs de dispersion

- Les indicateurs de position étudiés dans la section précédente servent à désigner l'endroit de la distribution de $X$ autour duquel se positionnent les modalités observées de $X$.
- Utilisé séparément des autres, aucun ne fournit de renseignement sur la façon dont les modalités se positionnent autour de cet endroit.
- Des caractéristiques intéressantes de la distribution de $X$ sont :
  1. la plus ou moins grande dispersion des modalités observées dans leur ensemble ;
  2. la plus ou moins grande dispersion des modalités dans leur regroupement autour d'une tendance centrale.

# L'étendue

## Définition

On appelle étendue de (la distribution de) la variable $X$ le nombre noté ETD($X$) et défini par

$$\text{ETD}(X) = X_M - X_m$$

avec

$$X_M = \max\{X(i), i \in P\},$$

$$X_m = \min\{X(i), i \in P\}.$$

## Interprétation

L'étendue mesure la dispersion de la distribution de $X$ par la distance maximale qu'on observe entre deux modalités de $X$. Plus l'étendue est grande, plus la dispersion est grande.

# L'écart absolu moyen

## Définition

Pour une variable statistique $X$, l'écart absolu moyen (EAM) est la quantité définie par

$$\mathrm{EAM}(X) = \frac{1}{N} \sum_{i=1}^{N} |X(i) - \overline{X}|.$$

## Interprétation

Pour chaque $i \in P$, $|X(i) - \overline{X}|$ est la distance usuelle entre $X(i)$ et $\overline{X}$. L'EAM de $X$ s'obtient directement comme la moyenne de ces distances.
L'EAM de $X$ est donc la moyenne des distances entre les modalités observées de $X$ et leur moyenne. Plus $\mathrm{EAM}(X)$ est élevé, plus cette dispersion est grande.

# La variance

## Définition 6.15

Pour une variable statistique $X$, la variance est la quantité définie par

$$\mathrm{Var}(X) = \frac{1}{N}\sum_{i=1}^{N}(X(i) - \overline{X})^2.$$

## Propriétés

- La variance mesure la dispersion des valeurs autour de la moyenne
- Toujours positive ou nulle : $\mathrm{Var}(X) \geq 0$
- Unité : carré de l'unité de mesure de $X$
- Sensible aux valeurs extrêmes

# La variance

## Interprétation

La variance représente la moyenne des carrés des écarts à la moyenne. Plus $\text{Var}(X)$ est élevée, plus les valeurs de $X$ sont dispersées autour de $\overline{X}$.

## Formule alternative

$$\text{Var}(X) = \frac{1}{N}\sum_{i=1}^{N} X(i)^2 - \overline{X}^2$$

# L'écart-type

## Définition 6.16

Pour une variable statistique $X$, l'écart-type est la quantité définie par

$$\sigma(X) = \sqrt{\mathrm{Var}(X)} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X(i) - \overline{X})^2}.$$

## Propriétés

- Mesure la dispersion dans les mêmes unités que $X$
- Toujours positif ou nul : $\sigma(X) \geq 0$
- Moins sensible aux valeurs extrêmes que la variance
- Permet de comparer la dispersion de variables de même unité

# L'écart-type

## Interprétation

L'écart-type représente la dispersion "moyenne" des valeurs autour de la moyenne. Plus $\sigma(X)$ est élevé, plus les valeurs de $X$ sont dispersées autour de $\overline{X}$.

- Si $\sigma(X) = 0$ : toutes les valeurs sont identiques
- Si $\sigma(X)$ est petit : les valeurs sont proches de la moyenne
- Si $\sigma(X)$ est grand : les valeurs sont très dispersées

## Relation avec la variance

L'écart-type est simplement la racine carrée de la variance :

$$\sigma(X) = \sqrt{\text{Var}(X)}$$

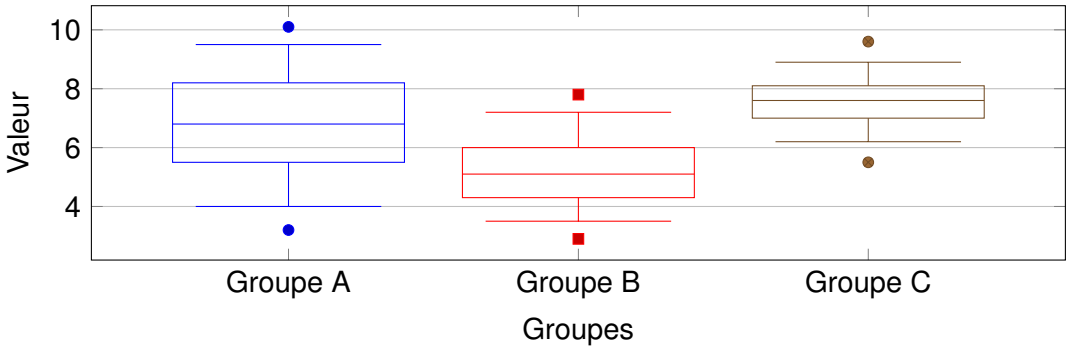
# Diagramme en boîte (boîte à moustaches)

## Définition

Représentation graphique d'une **variable quantitative** résumant sa distribution à l'aide des **quartiles** et des **valeurs extrêmes non aberrantes**.

- **Boîte** : s'étend de $Q_1$ à $Q_3$ (l'intervalle interquartile $IQR = Q_3 - Q_1$).
- **Trait médian** : la médiane $\tilde{x}$ au sein de la boîte.
- **Moustaches** : vont jusqu'aux plus petites/grandes valeurs situées dans $[\, Q_1 - 1.5\, IQR,\ Q_3 + 1.5\, IQR\,]$ (valeurs non aberrantes).
- **Valeurs aberrantes** : points au-delà des moustaches ($< Q_1 - 1.5\, IQR$ ou $> Q_3 + 1.5\, IQR$).
- **Lecture** :
  - $\tilde{x}$ proche du centre de la boîte $\Rightarrow$ distribution plutôt *symétrique*.
  - Boîte/moustache plus longues d'un côté $\Rightarrow$ *asymétrie* (queue plus étendue).
  - Permet la **comparaison** rapide de plusieurs groupes côte à côte.

# Exemple : diagrammes en boîte (trois groupes)




Les boîtes vont de $Q_1$ à $Q_3$ (IQR), le trait central est la médiane ; les moustaches s'étendent jusqu'aux valeurs non aberrantes. Les points en-dehors sont les valeurs aberrantes.

# Définition de la distribution

## Définition

La **distribution** d'une variable statistique $X$ décrit la façon dont les différentes **modalités (ou valeurs)** de $X$ se répartissent dans la population étudiée.

- Elle indique les effectifs ou fréquences associés à chaque modalité.
- Elle peut être représentée sous forme de tableau, d'histogramme ou de diagramme en boîte.
- Elle constitue la base de tout résumé numérique (moyenne, médiane, variance, etc.).

# Indicateurs de forme : symétrie et asymétrie

## Objectif

Les **indicateurs de forme** complètent les mesures de tendance centrale et de dispersion. Ils permettent de caractériser la **forme générale de la distribution** d'une variable :

- **Symétrie** : répartition équilibrée autour du centre.
- **Asymétrie** (ou *skewness*) : Asymétrie entre les deux côtés de la distribution.
- D'autres indicateurs décrivent aussi l'**aplatissement** (kurtosis), mais nous nous concentrerons ici sur la symétrie.

# Symétrie et asymétrie d'une distribution

## Symétrie

Une distribution est dite **symétrique** lorsque ses valeurs sont réparties de manière identique de part et d'autre d'un point central (souvent la moyenne ou la médiane).

## Asymétrie

Une distribution est **asymétrique** lorsqu'un des côtés est plus étalé que l'autre :

- **Asymétrie à droite (positive)** : longue queue vers les grandes valeurs.
- **Asymétrie à gauche (négative)** : longue queue vers les petites valeurs.

# Interprétation et propriétés de la symétrie

- Si la distribution est **symétrique**, les trois mesures de tendance centrale coïncident :

$$\text{Moyenne} = \text{Médiane} = \text{Mode}.$$

- Si elle est **asymétrique**, ces trois valeurs diffèrent.
- La direction de l'asymétrie indique la position relative de ces mesures :
  - Queue à gauche $\Rightarrow$ asymétrie négative.
  - Queue à droite $\Rightarrow$ asymétrie positive.

# Positions relatives du mode, de la médiane et de la moyenne

- Lorsque la distribution est **asymétrique à gauche (négative)** :

$$Mo(X) < Me(X) < \overline{X}$$

- Lorsque la distribution est **asymétrique à droite (positive)** :

$$Mo(X) > Me(X) > \overline{X}$$

- Cette observation permet de construire des **indicateurs d'asymétrie** numériques.

# Mesures d'asymétrie (Pearson) : principes

## Définition générale

Les **mesures d'asymétrie** décrivent numériquement le degré et la direction de la dissymétrie d'une distribution. On appelle coefficients d'asymétrie de Pearson les deux coefficients définis par :

- Ces indicateurs comparent les positions relatives de la moyenne, de la médiane et du mode.
- Plusieurs formules existent, par exemple :

$$P_1(X) = \frac{\overline{X} - Mo(X)}{s} \quad \text{ou} \quad P_2(X) = \frac{3(\overline{X} - Me(X))}{s}$$

  où $s$ est l'écart-type.
- Plus $P_i(X)$ est proche de 0, plus la distribution est symétrique.

# Mesures d'asymétrie : interprétation

- Si la distribution est **symétrique**, alors :

$$P_1(X) = P_2(X) = 0$$

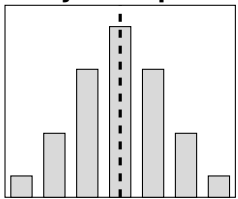
- Si $P_2(X) < 0$, alors $\overline{X} < Me(X)$ : **asymétrie à droite (positive)**.
- Si $P_2(X) > 0$, alors $\overline{X} > Me(X)$ : **asymétrie à gauche (négative)**.
- On peut montrer que :

$$-1 \leq P_2(X) \leq 1$$

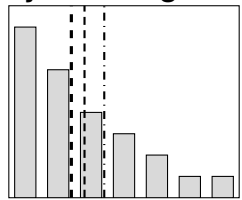
- Ces mesures permettent de quantifier précisément la forme de la distribution.

# Symétrie et asymétrie : illustrations

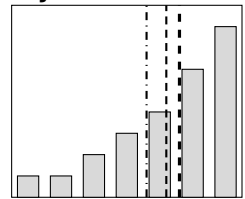
**Symétrique**



**Asymétrie à gauche**



**Asymétrie à droite**



**Rappel :** Asymétrie à gauche $\Rightarrow Mo < Me < \overline{X}$ ; Asymétrie à droite $\Rightarrow Mo > Me > \overline{X}$ ; Symétrique $\Rightarrow Mo = Me = \overline{X}$.

# Coefficient d'asymétrie $\gamma_1$

## Définition

À partir des résultats du point 12 (sec. 6.4.1.1), le **coefficient d'asymétrie** est défini par :

$$\gamma_1(X) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{X_i - \overline{X}}{\sigma(X)}\right)^3$$

où $\overline{X}$ est la moyenne et $\sigma(X)$ l'écart-type.

- $\gamma_1(X) = 0$ pour une distribution parfaitement **symétrique** (ex. gaussienne centrée).
- $\gamma_1(X) > 0$ : **asymétrie à droite** (queue vers les grandes valeurs).
- $\gamma_1(X) < 0$ : **asymétrie à gauche** (queue vers les petites valeurs).
- Sensible aux **valeurs extrêmes** (mesure basée sur les moments).

# Coefficient d'asymétrie de Yule–Bowley

## Définition

On définit le **coefficient d'asymétrie de Yule–Bowley** :

$$B(X) \;=\; \frac{\big[Q_3(X) - Q_2(X)\big] - \big[Q_2(X) - Q_1(X)\big]}{Q_3(X) - Q_1(X)} \;=\; \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1},$$

où $Q_1, Q_2, Q_3$ sont respectivement les **quartiles** (médiane $Q_2$).

- $B(X) = 0$ pour une distribution **symétrique**.
- $B(X) > 0 \Rightarrow$ **asymétrie à droite** ; $B(X) < 0 \Rightarrow$ **asymétrie à gauche**.
- $|B(X)| \leq 1$ en général ; mesure **robuste** (basée sur les quartiles).
- Utile quand on souhaite limiter l'influence des **outliers**.

# Tableau comparatif des coefficients d'asymétrie

| Coefficient | Formule | Données utilisées | Sensibilité aux valeurs extrêmes | Interprétation / Cas d'usage |
|---|---|---|---|---|
| $P_2(X) = \frac{3(\overline{X} - Me(X))}{s}$ | Compare la moyenne et la médiane pour estimer la dissymétrie. | $\overline{X}, Me, s$ | Moyenne | Simple à calculer. Utilisé pour des distributions modérément asymétriques. |
| $\gamma_1(X) = \frac{1}{N}\sum\left(\frac{X_i - \overline{X}}{s}\right)^3$ | Mesure fondée sur le **moment d'ordre 3** standardisé. | Données complètes | Très forte | Utilisé en analyse théorique ou quand la distribution est bien connue (ex. normales tronquées). |
| $B(X) = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1}$ | Basé sur les **quartiles**. Mesure robuste de l'asymétrie. | $Q_1, Q_2, Q_3$ | Faible | Préféré pour les échantillons petits ou avec valeurs aberrantes. |

# Tableau comparatif des coefficients d'asymétrie

**Exemple d'interprétation :** Pour un échantillon de revenus (en €) :

$$\overline{X} = 2800, \quad Me = 2000, \quad s = 900 \Rightarrow P_2(X) = \frac{3(2800 - 2000)}{900} \approx 2.67 > 0$$

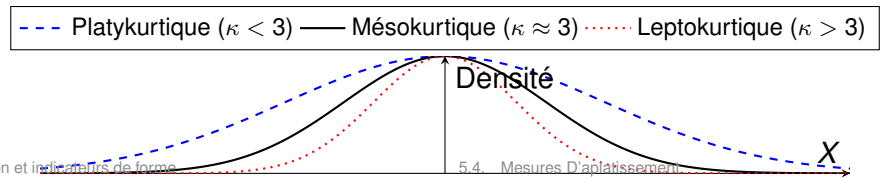
$\Rightarrow$ **Distribution asymétrique à droite** : la plupart des individus ont des revenus modestes, mais quelques-uns gagnent beaucoup plus.

# Aplatissement (kurtosis) : interprétation et importance

## Définition intuitive

L'**aplatissement** (ou *kurtosis*) mesure la **forme verticale** d'une distribution : il indique si celle-ci est plus ou moins **concentrée autour de la moyenne**.

- **Kurtosis élevé** ($\kappa(X) > 3$) : distribution **leptokurtique** forte pointe, queues épaisses. (ex. données financières avec valeurs extrêmes)
- **Kurtosis moyen** ($\kappa(X) \approx 3$) : distribution **mésokurtique** semblable à la loi normale.
- **Kurtosis faible** ($\kappa(X) < 3$) : distribution **platykurtique** plus aplatie, moins concentrée autour du centre.



# Indice d'aplatissement (kurtosis)

## Définition 6.22

On appelle **indice d'aplatissement** (kurtosis) de $X$ le nombre noté $\kappa(X)$ défini par

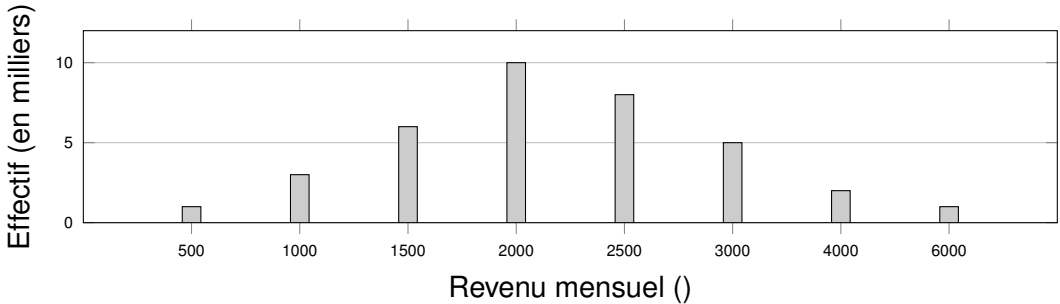
$$\kappa(X) \;=\; \frac{1}{N}\sum_{i=1}^{N}\left(\frac{X_i - \overline{X}}{\sigma(X)}\right)^4 .$$

## Propriété 6.14

L'indice $\kappa(X)$ est d'autant plus grand que les modalités de $X$ sont **concentrées autour** des valeurs $\overline{X} - \sigma(X)$ et $\overline{X} + \sigma(X)$. De plus, $\kappa(X)$ atteint sa **valeur minimale** 1 *ssi* pour tout individu $i = 1, \ldots, N$,

$$X_i \in \{\, \overline{X} - \sigma(X),\ \overline{X} + \sigma(X) \,\}.$$

# Exemple : distribution des revenus mensuels (en )




**Lecture :**

- Moyenne : $\overline{X} = 2600$
- Médiane : $Me = 2200$
- Mode : $Mo = 2000$
- $\Rightarrow Mo < Me < \overline{X}$

**Asymétrie :**

- **Asymétrie à droite** ($\gamma_1 > 0$) : la majorité gagne entre 1500 et 3000 , mais quelques individus ont des salaires très élevés.

**Aplatissement :**

- Distribution **leptokurtique** ($\kappa > 3$) : forte concentration autour de la médiane et queues épaisses.

# Exemple : distribution des revenus mensuels (en )

**Data storytelling :** Cette distribution raconte une histoire de *déséquilibre économique*. La majorité vit avec des revenus modestes, tandis qu'une minorité tire la moyenne vers le haut. L'aplatissement marqué montre une société où les revenus sont concentrés autour d'un niveau central, mais quelques exceptions extrêmes façonnent la perception globale de la richesse.

# Valeurs aberrantes : c'est quoi et pourquoi s'en soucier ?

## Définition

Une **valeur aberrante (outlier)** est une observation **inhabituelle** qui s'écarte nettement du **schéma général** de la distribution (beaucoup plus grande ou plus petite que les autres).

- **Origines** : erreurs de saisie/mesure, rareté réelle, changement de régime, sous-population.
- **Pourquoi c'est important ?**
  - Biaisent la **moyenne** et l'**écart-type**, perturbent les modèles (régression, clustering).
  - Influencent les **tests statistiques** et la **prise de décision**.
  - Nécessitent un **diagnostic** (corriger, exclure, ou modéliser séparément selon le contexte).

# IQR et Z-score : définitions et quand les utiliser

## IQR (InterQuartile Range)

$$Q_1 = \text{1er quartile}, \quad Q_3 = \text{3e quartile}, \quad IQR = Q_3 - Q_1$$

**Règle des 1.5 IQR** : une observation est aberrante si

$$x < Q_1 - 1.5\,IQR \quad \text{ou} \quad x > Q_3 + 1.5\,IQR.$$

*Atout* : **robuste** (basé sur les quantiles), peu sensible aux extrêmes.

# IQR et Z-score : définitions et quand les utiliser

## Z-score (score standardisé)

$$z_i = \frac{x_i - \overline{X}}{s} \quad (\overline{X} = \text{moyenne},\ s = \text{écart-type})$$

Seuils usuels : $|z_i| > 3$ (strict) ou $|z_i| > 2.5$ (plus sensible). *Atout* : simple et lié aux méthodes gaussiennes ; *limite* : **sensible** aux extrêmes.

**Conseil** : privilégier **IQR** en présence d'asymétrie/queues épaisses ; **Z-score** si la distribution est proche de la normale et que l'on veut une mesure compatible avec des modèles paramétriques.

# Exemple : IQR vs Z-score

**Données (9 obs.)** : $\{5, 6, 6, 7, 7, 8, 9, 12, 25\}$

**Méthode IQR**

$$Q_1 = 6, \quad Q_3 = 9, \quad IQR = 3$$

Bornes (règle 1.5 IQR) :

$$\underbrace{6 - 1.5 \times 3}_{=1.5} \text{ et } \underbrace{9 + 1.5 \times 3}_{=13.5}$$

$\Rightarrow$ **Outliers IQR** : toutes valeurs $> 13.5$. $\Rightarrow \boxed{25}$ est **aberrante** (les autres ne le sont pas).

**Méthode Z-score**

$$\overline{X} \approx 9.44, \quad s \approx 5.83$$

Z-scores (arrondis) :

$\{-0.76, -0.59, -0.59, -0.42, -0.42, -0.25, -$

$|z| > 3 \Rightarrow$ **aucun** outlier. $|z| > 2.5 \Rightarrow \boxed{25}$ est **aberrante**.

**Conclusion (storytelling)** : les deux méthodes racontent une histoire cohérente un *cas isolé très élevé* (25) s'écarte du cur de la distribution. L'IQR (robuste) le détecte immédiatement ; le Z-score le détecte si l'on choisit un seuil plus sensible (2.5). Le choix du seuil dépend du **contexte métier** (tolérance au risque, coût d'un faux positif/négatif) et de la **forme** de la distribution (asymétrie, aplatissement).

# Pourquoi tester la normalité des données ?

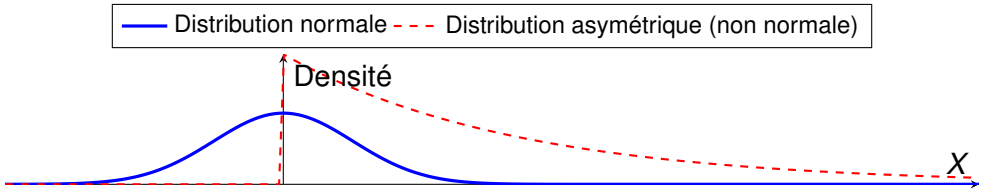
## Définition

Tester la **normalité** consiste à vérifier si la distribution d'une variable suit (ou s'écarte) d'une **loi normale** (courbe en cloche centrée et symétrique).

- **Pourquoi c'est important ?**
  - De nombreux tests statistiques (paramétriques) supposent que les données suivent une loi normale.
  - Cela conditionne la **validité des conclusions** : intervalles de confiance, tests de comparaison, régressions linéaires.
  - Identifier des **écarts à la normalité** permet d'adapter la méthode : transformation, test non paramétrique ou exclusion d'outliers.
- **Tests sensibles à la normalité :**
  - Test de Student (*t*-test), ANOVA, corrélation de Pearson, régression linéaire, PCA.
  - Tests non paramétriques alternatifs : MannWhitney, KruskalWallis, Spearman.

En résumé, le test de normalité nous dit si nos outils statistiques« classiques » sont fiables ou s'il faut recourir à des méthodes plus robustes.

# Visualiser la normalité : exemple graphique et interprétation




**Lecture graphique :**

- La courbe bleue (normale) est symétrique, centrée, sans queue étendue.
- La rouge (non normale) montre une **asymétrie à droite** : longue queue de grandes valeurs.

**Storytelling :**

- Si ces données représentent les *revenus* d'un échantillon : la majorité des individus se situe autour du revenu moyen, mais quelques cas extrêmes (grands revenus) tirent la moyenne vers le haut. Ce **déséquilibre** invalide certains tests paramétriques : on préférera des méthodes **non paramétriques** ou une transformation logarithmique.

# Tests de normalités

**Exemples de tests de normalité** : ShapiroWilk, AndersonDarling. Ces tests permettent de confirmer visuellement ou numériquement si la loi normale est adaptée.

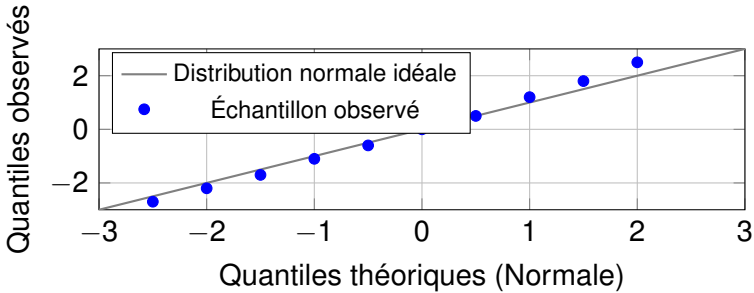
# QQ-Plot (QuantileQuantile Plot)

## Définition

Le **QQ-Plot** (QuantileQuantile Plot) compare les **quantiles observés** d'un échantillon avec les **quantiles théoriques** d'une distribution normale.

- Si les points s'alignent sur une diagonale droite  la distribution est **proche de la normale**.
- Des écarts systématiques (courbure, queue allongée) indiquent une **asymétrie** ou des **valeurs extrêmes**.
- Méthode visuelle rapide avant d'appliquer un test formel.

# QQ-Plot (QuantileQuantile Plot)




**Interprétation storytelling :** Si les points s'écartent dans les queues, cela raconte que nos données ont plus d'extrêmes que prévu  la distribution n'est pas« aussi normale » qu'elle en a l'air.

# Test de ShapiroWilk

## Principe

Le **test de ShapiroWilk** évalue la normalité en comparant les valeurs observées aux valeurs attendues sous une loi normale. Il calcule une statistique $W$ : plus $W$ est proche de 1, plus les données sont normales.

- **Hypothèses :**

  $H_0$ : Les données suivent une loi normale  $H_1$ : Les données ne sont pas normales

- Si la **p-valeur** < 0.05  on rejette $H_0$  la normalité est improbable.
- Recommandé pour **petits ou moyens échantillons** (n < 2000).
- Très sensible aux **valeurs aberrantes**.

**Exemple :** Pour 30 observations, $W = 0.94, p = 0.03 \Rightarrow$ on rejette la normalité à 5
Storytelling : Nos données ne dessinent pas la cloche parfaite  elles cachent une dissymétrie.

# Test d'AndersonDarling

## Principe

Le **test d'AndersonDarling** mesure la distance entre la distribution cumulée empirique et celle d'une loi normale. Il accorde une **pondération plus forte aux queues** de la distribution.

- **Hypothèses :**

  $H_0$ : Les données suivent d'une loi normale  $H_1$ : Les données ne sont pas normal

- Statistique $A^2$ : plus elle est grande, plus les écarts à la normalité sont forts.
- Fournit des **valeurs critiques** selon le seuil de signification (5
- Particulièrement utile pour **grands échantillons** (n > 500) ou données avec queues épaisses.

**Interprétation :** Si $A^2 > A^2_{critique}$ rejet de $H_0$. L'AndersonDarling "écoute" les extrêmes : il détecte des déviations que ShapiroWilk pourrait ignorer.

# Comparer ShapiroWilk et AndersonDarling

| Critère | ShapiroWilk | AndersonDarling | Quand préférer ? |
|---|---|---|---|
| Taille d'échantillon | Petits à moyens (n < 2000) | Grands échantillons (n > 500) | Données volumineuses ou distributions complexes |
| Sensibilité | Forte aux **valeurs centrales** | Forte aux **queues** (extrêmes) | Quand les extrêmes comptent (finance, fiabilité) |
| Sortie du test | Statistique *W* et *p-valeur* | Statistique $A^2$ et *valeurs critiques* | Selon ton logiciel (R, Python) et ton besoin d'interprétation |
| Hypothèse nulle | Normalité | Normalité | Identique, seul le poids des écarts diffère |

**Storytelling :** - Le **ShapiroWilk** vérifie si le cur de tes données forme une cloche régulière. - L'**AndersonDarling** va plus loin  il observe les extrêmes, cherchant les histoires cachées dans les queues. *Choisis selon ton contexte : petite enquête  Shapiro ; données massives ou financières  Anderson.*