**MODULE 3 : Statistiques et Probabilités**

**Objectifs :**

* Comprendre les bases des statistiques descriptives.
* Appliquer les probabilités aux données.
* Analyser les corrélations et les distributions.

Les **statistiques descriptives** sont un ensemble de méthodes utilisées pour résumer, organiser et analyser des données afin d’en extraire des informations clés. Elles sont divisées en plusieurs catégories :

**1. Mesures de tendance centrale**

Elles décrivent le point central des données :

**Moyenne** : La somme de toutes les valeurs divisée par le nombre total d'observations.

L’**interprétation à partir de la moyenne** permet de tirer des conclusions sur un ensemble de données en analysant sa valeur centrale. Voici quelques points clés à considérer :

**Interprétation générale**

* Une **moyenne élevée** indique que les valeurs des données sont globalement hautes.
* Une **moyenne faible** signifie que les valeurs sont plutôt basses.
* Si la moyenne est **proche de la médiane**, cela suggère que la distribution est **symétrique**.
* Si la moyenne est **supérieure ou inférieure à la médiane**, cela indique une **asymétrie** dans la distribution des données.

**Comparaison et analyse**

* Comparer la moyenne avec d’autres groupes de données permet d’évaluer les **différences et tendances générales**.
* Une **moyenne influencée par des valeurs extrêmes** (outliers) peut donner une fausse idée de la tendance centrale. Dans ce cas, la médiane est souvent une meilleure mesure.
* En complément, analyser l’**écart-type** permet de savoir si les valeurs sont **proches ou dispersées autour de la moyenne**.

**Exemple d’interprétation**

*Exemple 1 : Notes d'examen*

* Moyenne des notes = **14/20**
* Médiane = **13,5/20**
* Écart-type = **1,5**

Cela signifie que la plupart des élèves ont obtenu des notes autour de **14** avec une dispersion relativement faible.

*Exemple 2 : Revenus mensuels dans une entreprise*

* Moyenne = **3 500€**
* Médiane = **2 500€**
* Écart-type élevé  
   La moyenne est **plus élevée que la médiane**, ce qui suggère la présence de **salaires très élevés** qui augmentent artificiellement la moyenne (asymétrie à droite).

**Médiane** : La valeur qui sépare la moitié supérieure et inférieure des données.

L’interprétation par la médiane est une approche statistique qui consiste à analyser une série de données en se basant sur la médiane plutôt que sur la moyenne.

**Définition de la médiane :**

La médiane est la valeur qui divise une série de données ordonnées en deux parties égales :

* 50 % des valeurs sont inférieures ou égales à la médiane
* 50 % des valeurs sont supérieures ou égales à la médiane

**Pourquoi utiliser la médiane ?**

1. **Moins sensible aux valeurs extrêmes** : Contrairement à la moyenne, qui peut être influencée par des valeurs aberrantes, la médiane donne une mesure plus représentative de la tendance centrale lorsque les données sont asymétriques.
2. **Représentation fidèle des données non symétriques** : Dans une distribution asymétrique (ex. revenus, prix immobiliers), la médiane est souvent plus pertinente que la moyenne.
3. **Utilisation en analyse décisionnelle** : En économie, en sciences sociales et en médecine, la médiane est fréquemment utilisée pour mieux représenter des données hétérogènes.

**Exemple d'interprétation :**

Supposons une série de salaires en euros :  
 **[1 500, 1 800, 2 000, 2 200, 50 000]**

* **Moyenne** : (1 500 + 1 800 + 2 000 + 2 200 + 50 000) ÷ 5 = **11 900 €**
* **Médiane** : La valeur centrale est **2 000 €**

L'interprétation par la médiane indique qu'un individu typique gagne **2 000 €**, alors que la moyenne est fortement biaisée par le salaire extrême de 50 000 €.

**Conclusion**

L’interprétation par la médiane est essentielle lorsqu’on travaille avec des données contenant des valeurs extrêmes ou asymétriques. Elle fournit une mesure plus robuste et fiable de la tendance centrale.

**Mode** : La valeur la plus fréquente dans un ensemble de données.

#### **Définition de la mode :**

La mode est la valeur qui apparaît **le plus fréquemment** dans une série de données. Contrairement à la moyenne et à la médiane, qui donnent une mesure de tendance centrale basée sur des calculs mathématiques, la mode identifie simplement l’élément le plus courant dans l’ensemble des données.

#### **Pourquoi utiliser la mode ?**

1. **Pertinente pour les données qualitatives** : La mode est utile lorsque les données ne sont pas numériques (ex. couleur préférée, marque la plus vendue).
2. **Représente la valeur la plus populaire** : Elle indique ce qui est le plus fréquent dans une distribution (ex. pointure de chaussure la plus répandue).
3. **Facile à interpréter** : Contrairement à la moyenne et à la médiane, la mode est immédiatement compréhensible et ne nécessite pas de calculs complexes.

#### **Exemple d’interprétation :**

**1. Données numériques (salaire en euros) :**  
 **[1 500, 1 800, 2 000, 2 000, 2 200, 50 000]**

* La mode est **2 000 €** car ce salaire apparaît deux fois, tandis que les autres valeurs sont uniques.
* Interprétation : **Le salaire le plus courant dans cet échantillon est de 2 000 €**.

**2. Données qualitatives (couleurs de voitures vendues) :**  
 **[Rouge, Bleu, Noir, Noir, Blanc, Noir, Bleu]**

* La mode est **Noir**, car c'est la couleur qui apparaît le plus souvent.
* Interprétation : **La couleur de voiture la plus populaire est le noir.**

#### **Cas particuliers :**

* **Série unimodale** : Une seule valeur est dominante (ex. mode = 2 000 €).
* **Série bimodale** : Deux valeurs apparaissent aussi fréquemment l’une que l’autre (ex. mode = 2 000 € et 3 000 €).
* **Série sans mode** : Si aucune valeur ne se répète, la série n’a pas de mode.

#### **Conclusion**

L’interprétation par la mode est utile pour identifier **les tendances et préférences courantes** dans un ensemble de données. Elle est particulièrement pertinente pour les variables **qualitatives** et permet d’illustrer **ce qui est le plus fréquent dans une population**.

**2. Mesures de dispersion**

Elles indiquent la variabilité des données :

**Étendue** : Différence entre la valeur maximale et la valeur minimale.

#### **1. Définition de la Dispersion**

La dispersion est une mesure statistique qui évalue la **variabilité des données** autour d’une valeur centrale (moyenne, médiane, mode). Plus la dispersion est élevée, plus les valeurs sont **éloignées les unes des autres**.

#### **2. Définition de l’Étendue**

L’étendue est l’une des mesures les plus simples de la dispersion. Elle correspond à la **différence entre la plus grande et la plus petite valeur** d’un ensemble de données.

**Formule :**

Etendue=Valeur maximale−Valeur minimale

#### **3. Interprétation de l’Étendue**

L’étendue indique **l’écart global** entre les valeurs extrêmes d’une distribution.

* **Si l’étendue est faible**, les données sont concentrées dans un intervalle restreint.
* **Si l’étendue est élevée**, les données sont très dispersées, ce qui peut indiquer la présence de valeurs extrêmes (outliers).

#### **4. Exemple d’Étendue**

**Exemple 1 : Salaires en euros**  
Données : **[1 500, 1 800, 2 000, 2 200, 50 000]**

* Valeur max = **50 000 €**
* Valeur min = **1 500 €**
* **Étendue = 50 000 - 1 500 = 48 500 €**  
  L’écart est énorme, indiquant une forte dispersion et la présence probable de valeurs aberrantes.

**Exemple 2 : Notes d’élèves (/20)**  
Données : **[10, 12, 14, 15, 16]**

* Valeur max = **16**
* Valeur min = **10**
* **Étendue = 16 - 10 = 6**  
  Les notes sont relativement regroupées, donc la dispersion est faible.

#### **5. Limites de l’Étendue**

* **Ne tient compte que des valeurs extrêmes** et ignore la répartition des autres données.
* **Sensible aux valeurs aberrantes** : Une seule valeur très grande ou très petite peut fausser l’analyse.
* **Ne permet pas de comparer précisément des distributions** avec des effectifs différents.

#### **6. Autres mesures de dispersion**

Pour mieux analyser la dispersion, on utilise souvent d’autres indicateurs :

* **Écart-type** : mesure la dispersion autour de la moyenne.
* **Variance** : moyenne des carrés des écarts à la moyenne.
* **Intervalle interquartile (IQR)** : moins sensible aux valeurs extrêmes, basé sur les quartiles.

#### **Conclusion**

L’étendue est une **mesure simple et rapide** pour évaluer la dispersion des données, mais elle doit être complétée par d’autres indicateurs pour une analyse plus précise.

**Variance** : Moyenne des carrés des écarts à la moyenne.

#### **1. Définition de la Variance**

La variance est une **mesure de dispersion** qui indique **à quel point les valeurs d’un ensemble de données s’écartent de la moyenne**. Plus la variance est élevée, plus les données sont dispersées.

#### **2. Interprétation de la Variance**

* **Si la variance est faible**, les valeurs sont proches de la moyenne → la distribution est peu dispersée.
* **Si la variance est élevée**, les valeurs sont éloignées de la moyenne → les données sont très dispersées.

#### **3. Exemple d’Interprétation**

**Exemple 1 : Notes d’élèves (/20)**

* Classe A : **[10, 11, 12, 13, 14]**
* Classe B : **[5, 8, 12, 16, 19]**
* Les deux classes ont la même moyenne (**12**), mais la **variance est plus grande pour la classe B** car les notes sont plus dispersées.

**Exemple 2 : Revenus mensuels (en €)**

* Ville A : **[2 200, 2 500, 2 600, 2 700, 2 800]**
* Ville B : **[1 500, 2 000, 2 500, 3 500, 5 000]**
* **La moyenne est similaire**, mais la variance est plus élevée pour la Ville B, montrant une plus grande inégalité des revenus.

#### **4. Limites de la Variance**

* **Unité au carré** : La variance est exprimée dans une unité différente de celle des données initiales (ex. €² au lieu de €). Pour une interprétation plus intuitive, on utilise **l’écart-type**, qui est la racine carrée de la variance.
* **Sensibilité aux valeurs extrêmes** : Une seule valeur très grande ou très petite peut fortement influencer la variance.

#### **5. Conclusion**

La variance est un outil fondamental en statistiques pour mesurer la **dispersion des données autour de la moyenne**. Elle est souvent utilisée en économie, en finance et en sciences pour comprendre **la variabilité et l’homogénéité d’un ensemble de données**. Cependant, pour une meilleure interprétation, on préfère souvent utiliser l’écart-type.

**Écart-type** : Racine carrée de la variance, exprimée dans la même unité que les données.

#### **1. Définition de l'Écart-Type**

L’écart-type est une **mesure de dispersion** qui indique à quel point les valeurs d’un ensemble de données **s’écartent de la moyenne**. C’est la **racine carrée de la variance**, ce qui le rend plus facile à interpréter, car il est exprimé dans la même unité que les données.

#### **2. Interprétation de l'Écart-Type**

* **Si l’écart-type est faible**, les valeurs sont proches de la moyenne → **faible dispersion**.
* **Si l’écart-type est élevé**, les valeurs sont éloignées de la moyenne → **forte dispersion**.

**L’écart-type permet de quantifier la variabilité d’un ensemble de données** et d’évaluer si les valeurs sont homogènes ou dispersées.

#### **3. Exemples d’Interprétation**

**Exemple 1 : Notes d’élèves (/20)**

* Classe A : **[10, 11, 12, 13, 14]** → **Écart-type faible** (les notes sont proches de la moyenne).
* Classe B : **[5, 8, 12, 16, 19]** → **Écart-type élevé** (les notes sont très dispersées).

**Les élèves de la classe A ont des performances plus homogènes, alors que ceux de la classe B ont des résultats plus variés.**

**Exemple 2 : Revenus mensuels (en €)**

* Ville A : **[2 200, 2 500, 2 600, 2 700, 2 800]** → **Écart-type faible** (revenus homogènes).
* Ville B : **[1 500, 2 000, 2 500, 3 500, 5 000]** → **Écart-type élevé** (revenus très inégaux).

**La ville A a une répartition des revenus plus équilibrée, tandis que la ville B présente des inégalités plus marquées.**

#### **4. Utilisation de l'Écart-Type**

L’écart-type est couramment utilisé en :

* **Économie et finance** : pour mesurer la volatilité des prix ou des actions en bourse.
* **Sciences et ingénierie** : pour évaluer la précision des mesures expérimentales.
* **Psychologie et sociologie** : pour analyser la dispersion des scores dans des études ou des enquêtes.

#### **5. Relation avec la Loi Normale**

Dans une distribution normale (**courbe de Gauss**) :

* **68 %** des valeurs sont comprises dans l’intervalle **[moyenne ± 1 écart-type]**.
* **95 %** des valeurs sont dans **[moyenne ± 2 écarts-types]**.
* **99,7 %** des valeurs sont dans **[moyenne ± 3 écarts-types]**.

**L’écart-type permet donc d’estimer dans quelle mesure les données s’écartent de la moyenne dans une distribution normale.**

#### **6. Conclusion**

L’écart-type est une mesure clé pour comprendre **la variabilité des données**. Contrairement à l’étendue, qui ne tient compte que des valeurs extrêmes, **l’écart-type prend en compte toute la distribution**. Il est souvent utilisé avec la moyenne pour mieux interpréter la dispersion des données.

**Coefficient de variation** : Rapport entre l’écart-type et la moyenne, exprimé en pourcentage.

#### **1. Définition du Coefficient de Variation**

Le **coefficient de variation (CV)** est une mesure **relative** de dispersion qui permet de comparer la variabilité de plusieurs ensembles de données, même s’ils ont des unités ou des moyennes différentes.

#### **2. Interprétation du Coefficient de Variation**

* **CV faible** → **Les données sont peu dispersées** autour de la moyenne (faible variabilité).
* **CV modéré** → **Dispersion moyenne** des données.
* **CV élevé** → **Forte variabilité** des données, la moyenne est moins représentative.

**Un CV plus faible indique une meilleure stabilité des données.**

#### **3. Exemples d’Interprétation**

**Exemple 1 : Comparaison des salaires dans deux entreprises**

| **Entreprise** | **Moyenne des salaires (€)** | **Écart-type (€)** | **Coefficient de Variation** |
| --- | --- | --- | --- |
| A | 2 500 | 250 | 10% |
| B | 2 500 | 800 | 32% |

**Interprétation :**

* **L’entreprise A a une répartition plus homogène des salaires** (CV = 10 %).
* **L’entreprise B présente une plus grande inégalité** dans les salaires (CV = 32 %).

**Exemple 2 : Comparaison de la performance de deux machines**

| **Machine** | **Temps moyen de production (min)** | **Écart-type (min)** | **CV (%)** |
| --- | --- | --- | --- |
| M1 | 50 | 5 | 10 % |
| M2 | 30 | 7 | 23 % |

**Interprétation :**

* **M1 est plus fiable** car son CV est plus faible.
* **M2 est plus instable**, car la variation du temps de production est plus grande par rapport à la moyenne.

#### **4. Utilisation du Coefficient de Variation**

Le CV est souvent utilisé dans :

* **Économie et Finance** : comparer la volatilité des actions en bourse.
* **Industrie** : évaluer la stabilité des machines et des processus.
* **Médecine** : analyser la variabilité des dosages médicaux.
* **Statistiques et Recherche** : comparer des ensembles de données ayant des unités différentes.

#### **5. Conclusion**

Le **coefficient de variation** est une **mesure standardisée** qui permet de comparer **la dispersion de données de natures différentes**. Plus le CV est **faible**, plus les données sont **stables et homogènes**. Un CV élevé indique **une forte variabilité et une faible représentativité de la moyenne**.

**3. Mesures de position**

Elles permettent de situer une donnée par rapport aux autres :

**Quartiles (Q1, Q2, Q3)** : Valeurs qui divisent les données en quatre parties égales.

#### **1. Définition des Quartiles**

Les **quartiles** sont des valeurs qui divisent un ensemble de données triées en **quatre parties égales**. Chaque quartile représente **25 % des données**.

Les trois quartiles principaux sont :

* **Q1 (Premier quartile - 25%)** : Valeur en dessous de laquelle **25 % des données** se trouvent.
* **Q2 (Médiane - 50%)** : Valeur en dessous de laquelle **50 % des données** se trouvent (médiane).
* **Q3 (Troisième quartile - 75%)** : Valeur en dessous de laquelle **75 % des données** se trouvent.

**Q2 est aussi la médiane**, qui divise les données en deux parties égales.

#### **2. Calcul des Quartiles**

1. **Trier les données** en ordre croissant.
2. **Déterminer Q1, Q2 et Q3** :
   * **Q1** = Médiane de la première moitié des données.
   * **Q2** = Médiane de l’ensemble des données.
   * **Q3** = Médiane de la seconde moitié des données.

#### **3. Interprétation des Quartiles**

**Exemple : Notes d’élèves (/20)**  
Données : **[5, 8, 10, 12, 14, 15, 16, 18, 19]**

* **Q1 (25%)** = **10** → **25 % des élèves ont une note inférieure ou égale à 10**.
* **Q2 (Médiane - 50%)** = **14** → **50 % des élèves ont une note inférieure ou égale à 14**.
* **Q3 (75%)** = **16** → **75 % des élèves ont une note inférieure ou égale à 16**.

**Interprétation** :

* La majorité des élèves ont des notes comprises entre **Q1 (10) et Q3 (16)**.
* Seuls **25 % des élèves ont plus de 16**, ce qui montre une répartition équilibrée.

#### **4. Utilisation des Quartiles**

* **Analyse des distributions** : Permet de voir comment les données sont réparties.
* **Étude des inégalités** : En finance, on peut analyser la dispersion des salaires.
* **Boîte à moustaches (Boxplot)** : Visualisation graphique qui montre Q1, Q2, Q3 et les valeurs extrêmes.
* **Détection des valeurs aberrantes** : Une valeur est considérée comme extrême si elle est trop éloignée de la moyenne.

#### **5. Conclusion**

Les quartiles permettent de **résumer rapidement la répartition des données** et d’identifier les tendances générales et les valeurs extrêmes. Ils sont largement utilisés en **statistiques, économie, finance et sciences sociales** pour analyser des distributions de données.

**Percentiles** : Valeurs qui divisent les données en 100 parties égales.

#### **1. Définition des Percentiles**

Les **percentiles** sont des valeurs qui divisent un ensemble de données **triées en 100 parties égales**. Chaque percentile représente **1 % des données**.

Un **percentile Pk** signifie que **k % des observations** sont inférieures ou égales à cette valeur.

**Exemples courants de percentiles :**

* **P25 (25ᵉ percentile) = Q1 (Premier quartile)** → 25 % des données sont inférieures.
* **P50 (50ᵉ percentile) = Médiane** → 50 % des données sont inférieures.
* **P75 (75ᵉ percentile) = Q3 (Troisième quartile)** → 75 % des données sont inférieures.

#### **2. Interprétation des Percentiles**

**Exemple 1 : Scores à un examen**  
Données triées : **[45, 50, 55, 60, 65, 70, 75, 80, 85, 90]**

* **P10 = 50** → 10 % des étudiants ont une note inférieure ou égale à 50.
* **P50 (Médiane) = 70** → 50 % des étudiants ont une note inférieure ou égale à 70.
* **P90 = 85** → 90 % des étudiants ont une note inférieure ou égale à 85.

**Interprétation :**

* Si un élève obtient **80**, il est au-dessus du **75ᵉ percentile**, ce qui signifie qu’il a une meilleure note que **75 % des autres étudiants**.
* Un score proche de **P10** indique une performance faible, alors qu’un score proche de **P90** indique une performance élevée.

**Exemple 2 : Revenus mensuels dans une population**

* **P20 = 1 500 €** → 20 % des personnes gagnent moins de 1 500 €.
* **P50 (médiane) = 2 500 €** → 50 % des personnes gagnent moins de 2 500 €.
* **P90 = 5 000 €** → 90 % des personnes gagnent moins de 5 000 €.

**Interprétation :**

* Si une personne gagne **4 000 €**, elle se situe **au-dessus du 75ᵉ percentile**, donc elle gagne plus que 75 % de la population.
* Si une personne gagne **1 200 €**, elle est **en dessous du 20ᵉ percentile**, ce qui signifie qu’elle fait partie des revenus les plus bas.

#### **3. Utilisation des Percentiles**

**Les percentiles sont utilisés dans plusieurs domaines :**

* **Statistiques et analyse des données** : Pour étudier la répartition des données.
* **Médecine** : Par exemple, pour évaluer la **courbe de croissance des enfants** (un bébé dans le 90ᵉ percentile de taille est plus grand que 90 % des bébés de son âge).
* **Éducation** : Scores aux examens standardisés (ex : un élève dans le 80ᵉ percentile a mieux réussi que 80 % des autres candidats).
* **Économie et finance** : Études sur la répartition des salaires et des richesses.

#### **4. Conclusion**

Les percentiles sont des outils puissants pour **comparer des individus par rapport à un ensemble de données**. Ils permettent d’identifier les tendances, d’évaluer les écarts et d’analyser les distributions de manière plus détaillée qu’une simple moyenne ou médiane.

**4. Représentations graphiques**

**Histogramme** : Représentation de la distribution des fréquences.

L'interprétation des histogrammes en data science est essentielle pour analyser la distribution des données et extraire des insights utiles. Voici les principaux éléments à prendre en compte :

**1. Forme de la distribution**

**Distribution normale (gaussienne)** : Courbe en cloche symétrique autour de la moyenne.

* La distribution normale est un **modèle clé** en data science, car elle permet d’appliquer de nombreuses méthodes statistiques et algorithmiques.
* Elle simplifie les calculs et aide à **tirer des conclusions** sur les données.

Si les données ne sont pas normales, des **ajustements ou transformations** sont nécessaires.

**Distribution asymétrique** :

* + **Asymétrie à gauche (négative)** : La queue est plus longue à gauche (ex. : salaires avec un plafond maximum).
* **Queue plus longue à gauche** : Les valeurs extrêmes faibles sont plus fréquentes.
* **Moyenne < Médiane < Mode** : La moyenne est plus petite que la médiane, qui elle-même est plus petite que le mode.
* **Exemple visuel** : L'histogramme montre une descente progressive vers la gauche, avec un pic vers les valeurs plus élevées.

**Solutions** :

* **Utiliser la médiane** au lieu de la moyenne.
* **Appliquer une transformation des données** (comme la transformation de puissance de Box-Cox).
* **Utiliser des tests statistiques non paramétriques** qui ne supposent pas de normalité.
  + **Asymétrie à droite (positive)** : La queue est plus longue à droite (ex. : revenus des grandes entreprises).

L'**asymétrie à droite** (ou **asymétrie positive**) désigne une distribution où la queue droite (valeurs élevées) est plus longue ou plus étalée que la queue gauche. Cela signifie que la majorité des valeurs se situent à gauche de la moyenne, avec quelques valeurs extrêmes plus grandes qui tirent la moyenne vers la droite.

**Moyenne > Médiane > Mode**

* **Distribution uniforme** : Les barres ont des hauteurs similaires, indiquant une répartition égale.
* **Distribution bimodale ou multimodale** : Présence de plusieurs pics, ce qui peut signifier plusieurs groupes distincts dans les données.

**2. Dispersion et étalement**

* Une plage étendue indique une grande variance des valeurs.
* Une plage resserrée indique des valeurs concentrées autour d’une moyenne.

**3. Présence d’outliers**

* Des barres isolées à l'extrémité de l'histogramme peuvent indiquer des valeurs aberrantes.

**4. Taille des classes (bins)**

* Un choix trop large peut masquer des détails importants.
* Un choix trop petit peut introduire du bruit inutile.

L’analyse d’un histogramme permet donc d’identifier des tendances et de mieux comprendre la distribution d’un jeu de données avant d’appliquer des modèles prédictifs ou des transformations.

* **Boîte à moustaches (boxplot)** : Permet de visualiser la médiane, les quartiles et les valeurs extrêmes.
* **Nuage de points** : Visualisation des relations entre deux variables.

Le **nuage de points** est un type de graphique qui représente la relation entre deux variables numériques. Chaque point du graphique correspond à une **observation** et est placé en fonction de ses coordonnées (x,y)(x, y)(x,y).

### ****1. Objectif d'un Nuage de Points****

Le scatter plot est utilisé pour :

* **Détecter des corrélations** entre deux variables (positive, négative ou inexistante).
* **Visualiser les tendances** (ex. : relation entre l'âge et le salaire).
* **Identifier les valeurs aberrantes** (points éloignés du groupe principal).
* **Représenter des clusters (groupes)** dans un ensemble de données.

### ****2. Types de Corrélations Observables****

**Corrélation positive** : Quand xxx augmente, yyy augmente aussi. (Ex. : Nombre d’heures d’étude et note obtenue)  
**Corrélation négative** : Quand xxx augmente, yyy diminue. (Ex. : Vitesse d’une voiture et consommation d’essence)  
**Aucune corrélation** : Pas de lien clair entre xxx et yyy. (Ex. : Numéro de téléphone et taille des individus)

### ****3. Exemple de Nuage de Points en Python****

**Visualisation avec Matplotlib et Seaborn**

*import numpy as np*

*import matplotlib.pyplot as plt*

*import seaborn as sns*

*# Génération de données aléatoires*

*np.random.seed(42)*

*x = np.random.rand(100) \* 10 # Variable indépendante*

*y = 2 \* x + np.random.randn(100) \* 3 # Relation linéaire avec bruit*

*# Création du scatter plot*

*plt.figure(figsize=(6, 4))*

*sns.scatterplot(x=x, y=y, color="blue", alpha=0.7)*

*plt.xlabel("Variable X")*

*plt.ylabel("Variable Y")*

*plt.title("Nuage de points")*

*plt.show()*

### ****4. Améliorations Possibles****

* **Ajout d’une courbe de tendance** (régression linéaire avec sns.regplot()).
* **Coloration des points selon une 3ᵉ variable** (ex. : catégorie des individus).
* **Taille des points variable** (ex. : importance d’une observation).

### ****5. Applications du Nuage de Points****

* **Analyse de marché** : Relation entre le prix et la demande d’un produit.
* **Médecine** : Corrélation entre la consommation de sel et la pression artérielle.
* **Finance** : Lien entre les dépenses publicitaires et les ventes.

### ****Comment Interpréter un Nuage de Points****

L'interprétation d'un nuage de points dépend de la forme et de la distribution des points dans le graphique. Voici les principales caractéristiques à observer lors de l'analyse d'un scatter plot (nuage de points) :

### ****1. Corrélation entre les Variables****

Les points du nuage de points peuvent révéler s'il existe une **relation** entre les deux variables. Voici les types de relations possibles :

#### **a) Corrélation Positive**

* **Observation** : Les points du nuage de points forment une pente montante, c'est-à-dire qu'à mesure que la variable xxx augmente, la variable yyy augmente également.
* **Exemple** : Plus une personne fait d'exercice, plus sa condition physique (ou son score sur un test de forme physique) s'améliore.
* **Indicateur visuel** : La distribution des points suit une ligne droite ascendante.

#### **b) Corrélation Négative**

* **Observation** : Les points du nuage de points forment une pente descendante, indiquant que lorsque la variable xxx augmente, la variable yyy diminue.
* **Exemple** : Plus la vitesse d’une voiture augmente, plus sa consommation de carburant par km diminue.
* **Indicateur visuel** : Les points suivent une ligne droite descendante.

#### **c) Aucune Corrélation**

* **Observation** : Les points sont dispersés de manière aléatoire et ne montrent aucun modèle évident, indiquant qu'il n'y a pas de relation entre les deux variables.
* **Exemple** : L’âge d'une personne et la couleur de ses cheveux.
* **Indicateur visuel** : Les points ne suivent aucun schéma particulier, ils sont dispersés dans toutes les directions.

### ****2. Force de la Corrélation****

* **Corrélation forte** : Les points sont proches d'une ligne droite (positive ou négative), ce qui signifie que la relation entre les deux variables est forte.
  + **Exemple** : Le temps d’étude et les résultats aux examens. Les points seraient proches d’une droite ascendante.
* **Corrélation faible** : Les points sont dispersés autour d’une ligne de tendance, ce qui suggère une relation plus faible entre les variables.
  + **Exemple** : Le nombre d’heures de sommeil et la performance au travail. Les points sont moins proches d’une droite, mais montrent encore une tendance générale.

### ****3. Valeurs Aberrantes (Outliers)****

* **Observation** : Un ou plusieurs points sont loin de la masse principale des autres points. Ces points peuvent représenter des valeurs extrêmes qui ne suivent pas la tendance générale.
* **Exemple** : Un étudiant qui a étudié 10 heures pour un examen et a obtenu une note anormalement basse pourrait être un outlier dans un nuage de points représentant la relation entre les heures d’étude et la performance.
* **Indicateur visuel** : Les points situés loin de la région principale du nuage.

### ****4. Tendance ou Forme Non Linéaire****

* **Observation** : Parfois, les relations entre les variables ne sont pas linéaires, mais plutôt courbées. Un nuage de points peut suivre une courbe (parabolique, exponentielle, etc.).
* **Exemple** : La relation entre la vitesse d'une voiture et le freinage peut être mieux modélisée par une courbe (les premiers augmentations de vitesse ayant plus d'impact sur la distance de freinage).
* **Indicateur visuel** : Les points suivent une courbe plutôt qu'une droite.

### ****5. Groupes ou Clusters****

* **Observation** : Si le nuage de points montre des regroupements distincts de points, cela peut indiquer la présence de **sous-groupes** ou de **clusters** dans les données.
* **Exemple** : Les résultats d'un test de performance, où un groupe d’étudiants obtient des scores faibles, un autre des scores moyens, et un autre des scores élevés.
* **Indicateur visuel** : Des clusters de points qui ne se mélangent pas, chaque groupe ayant une forme distincte.

### ****6. Symétrie et Disposition****

* **Observation** : Le nuage de points peut aussi révéler si les données sont symétriques ou présentent une **asymétrie**.
* **Exemple** : Si les points sont répartis de manière symétrique autour de la ligne de tendance, cela peut suggérer une distribution normale. Si les points sont concentrés d'un côté, cela pourrait indiquer une asymétrie.

### ****Exemple Concret en Python****

Imaginons un nuage de points qui montre la relation entre les heures d’étude et les notes obtenues à un examen :

*import numpy as np*

*import matplotlib.pyplot as plt*

*# Données fictives*

*hours\_studied = np.array([1, 2, 3, 4, 5, 6, 7, 8, 9, 10])*

*exam\_scores = np.array([45, 50, 55, 60, 65, 70, 75, 80, 85, 90])*

*# Création du nuage de points*

*plt.scatter(hours\_studied, exam\_scores)*

*plt.xlabel("Heures étudiées")*

*plt.ylabel("Scores à l'examen")*

*plt.title("Relation entre les heures étudiées et les scores")*

*plt.show()*

Dans cet exemple, on peut observer une **corrélation positive** entre les heures d'étude et les scores obtenus. Plus on étudie, plus les scores augmentent.

### ****En Résumé :****

* **Corrélation positive** : Les points suivent une pente ascendante.
* **Corrélation négative** : Les points suivent une pente descendante.
* **Aucune corrélation** : Les points sont dispersés sans modèle clair.
* **Outliers** : Des points éloignés du groupe principal des données.
* **Tendance non linéaire** : Les points suivent une courbe.

Ces outils sont essentiels pour analyser des données avant d'appliquer des modèles plus complexes en statistiques inférentielles.