Statistique Descriptive : Représentations graphiques

Komlan Doigté KATAKOU

March 29, 2025

Représentations graphiques

Importance des représentations graphiques

La visualisation graphique est un outil essentiel en statistique descriptive.

- Elle permet d'explorer la répartition des données.
- Elle aide à identifier des tendances, des anomalies, des relations entre variables et à faire des hypothèses.
- Grâce aux graphiques, les caractéristiques des variables deviennent plus intuitives, facilitant ainsi l'interprétation des résultats.

Dans ces slides, nous faisons un petit tour d'horizon sur les principaux graphiques de base utilisés en statistique.

Histogramme

Définition

L'**histogramme** est un graphique utilisé pour visualiser sous forme de **bandes verticales** la distribution des valeurs d'une variable continue. La hauteur de chaque bande est proportionnelle à la fréquence des données se trouvant dans un dans la **classe** (l'intervalle) correspondant à la bande.

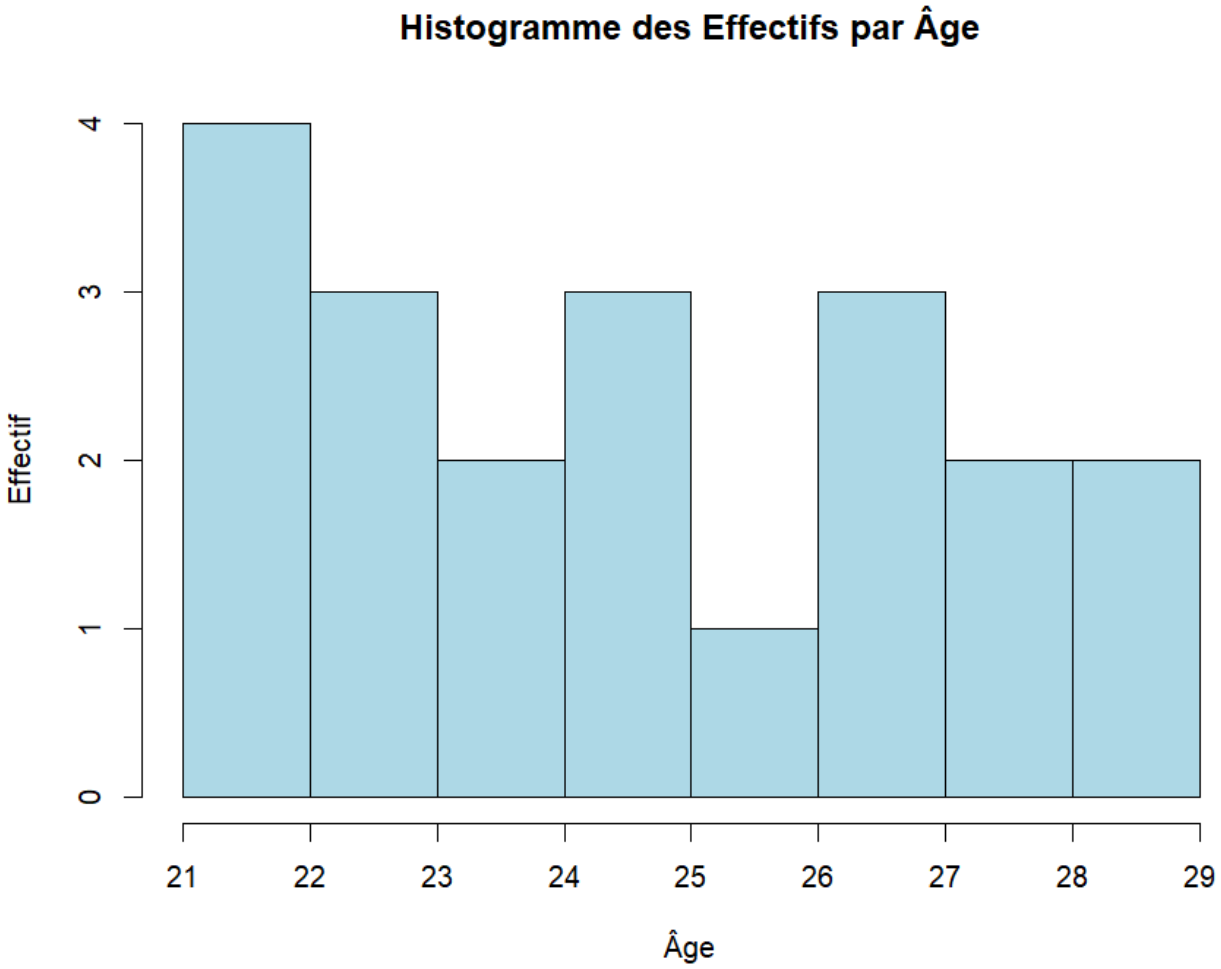
Considérons le jeu de données suivant qui renseigne le nombre de les âges des individus d'une population donnée.

Âge	21	22	23	24	25	26	27	28	29
Effectif	2	2	3	2	3	1	3	2	2

Table: Tableau des effectifs des âges

Histogramme

On obtient l'histogramme ci-après.



Interpréter un histogramme

Interprétation

- **Symétrie** : Si les données sont réparties de manière similaire à gauche et à droite autour des barres centrales les plus hautes, cela indique une répartition symétrique des données. (En statistique inférentielle, on pourra modéliser ce type de donées par une loi normale, par exemple)
- **Asymétrie** : Si les données sont concentrées davantage à gauche ou à droite de l'histogramme, cela indique une répartition asymétrique.(Elles pourraient par exemple être modélisées par une loi exponentielle, Gamma, ou log-normale)
- **Bimodalité** : Si l'histogramme présente deux pics distincts, cela suggère probablement l'existence de deux groupes sous-jacents dans la population. (Il s'agit des modèles de mélange dont on parlera plus tard en apprentissage statistique)

Interpréter un histogramme

Interprétation

- **Présence d'outliers** : Si l'on observe des barres peu élevées et assez isolées des autres, cela suggère l'existence de valeurs aberrantes dans les données.
- **Tendance centrale** : La position de la barre la plus élevée donne une idée de la tendance centrale des données (moyenne, médiane).
- **Dispersion** : La largeur de l'histogramme donne une idée de la dispersion des données. Un histogramme étroit indique des données concentrées autour d'une valeur moyenne, tandis qu'un histogramme large indique une plus grande dispersion des données.

Diagramme circulaire (Pie Chart)

Un **diagramme circulaire** est un graphique utilisé pour représenter les différentes modalités d'une **variable catégorielle** sous forme de secteurs. Chaque secteur est proportionnel à la fréquence de la catégorie correspondante, c'est-à-dire la part relative de cette catégorie par rapport à tout l'échantillon.

Interprétation

Si un secteur est beaucoup plus grand que les autres, cela indique une dominance de la catégorie correspondante. Si, à l'inverse, les secteurs sont de taille similaire, cela indique une répartition équilibrée entre les catégories.

Limites

- Il devient difficile à lire avec un grand nombre de catégories.
- Il n'est pas idéal pour comparer des différences petites entre les catégories : une différence de 1% de fréquence entre deux catégories n'est pas forcément visible à l'oeil nu sur le diagramme.

Diagramme circulaire (Pie Chart)

Catégorie	Effectif
A	10
B	15
C	20
D	12
E	8

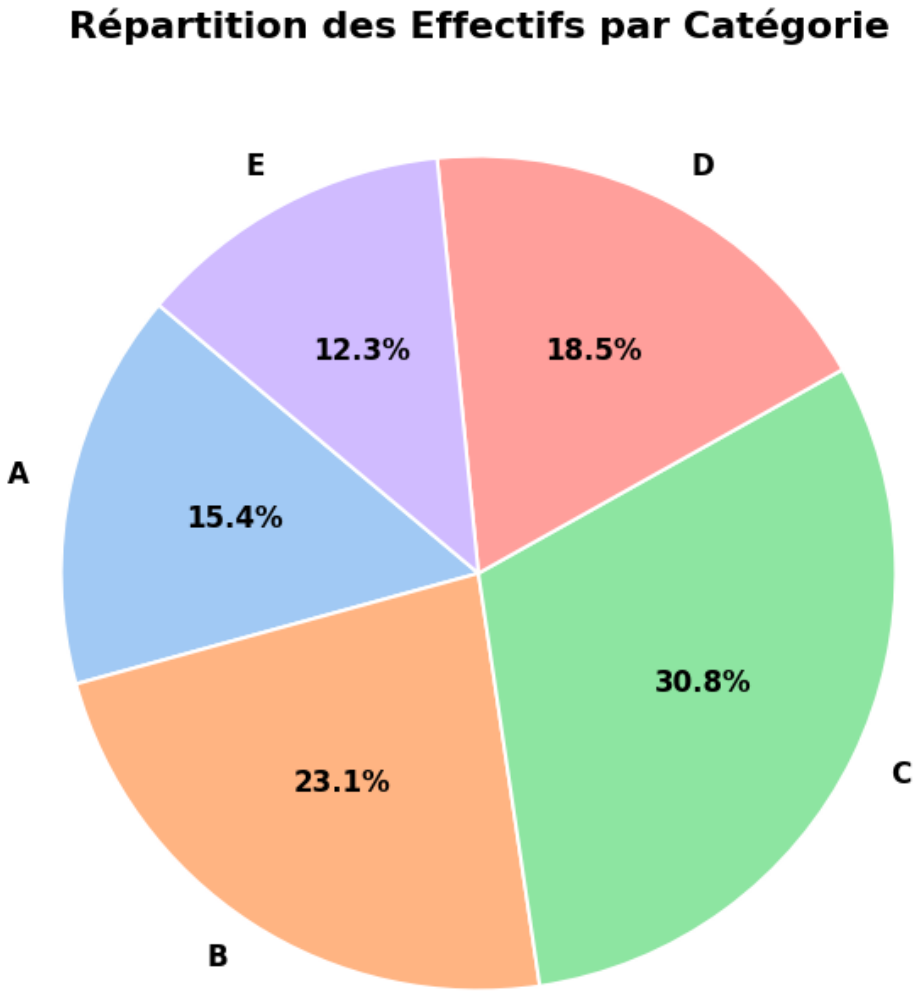


Diagramme en barres (ou en bâtons)

Un **diagramme en barres** est un graphique qui utilise des barres pour représenter des **données catégorielles ou discrètes**. La hauteur de chaque barre est proportionnelle à la fréquence de sa catégorie.

- **Interprétation** : Son interprétation est similaire à celle du diagramme circulaire.
- **Limites** :
  - Peu adapté pour un grand nombre de catégories.
  - Non adapté aux données continues (préférer un histogramme).
- **Exemple**

Catégorie	A	B	C	D	E
Effectif	10	15	20	12	8

Table: Tableau des effectifs des catégories (format horizontal)

Diagramme en barres

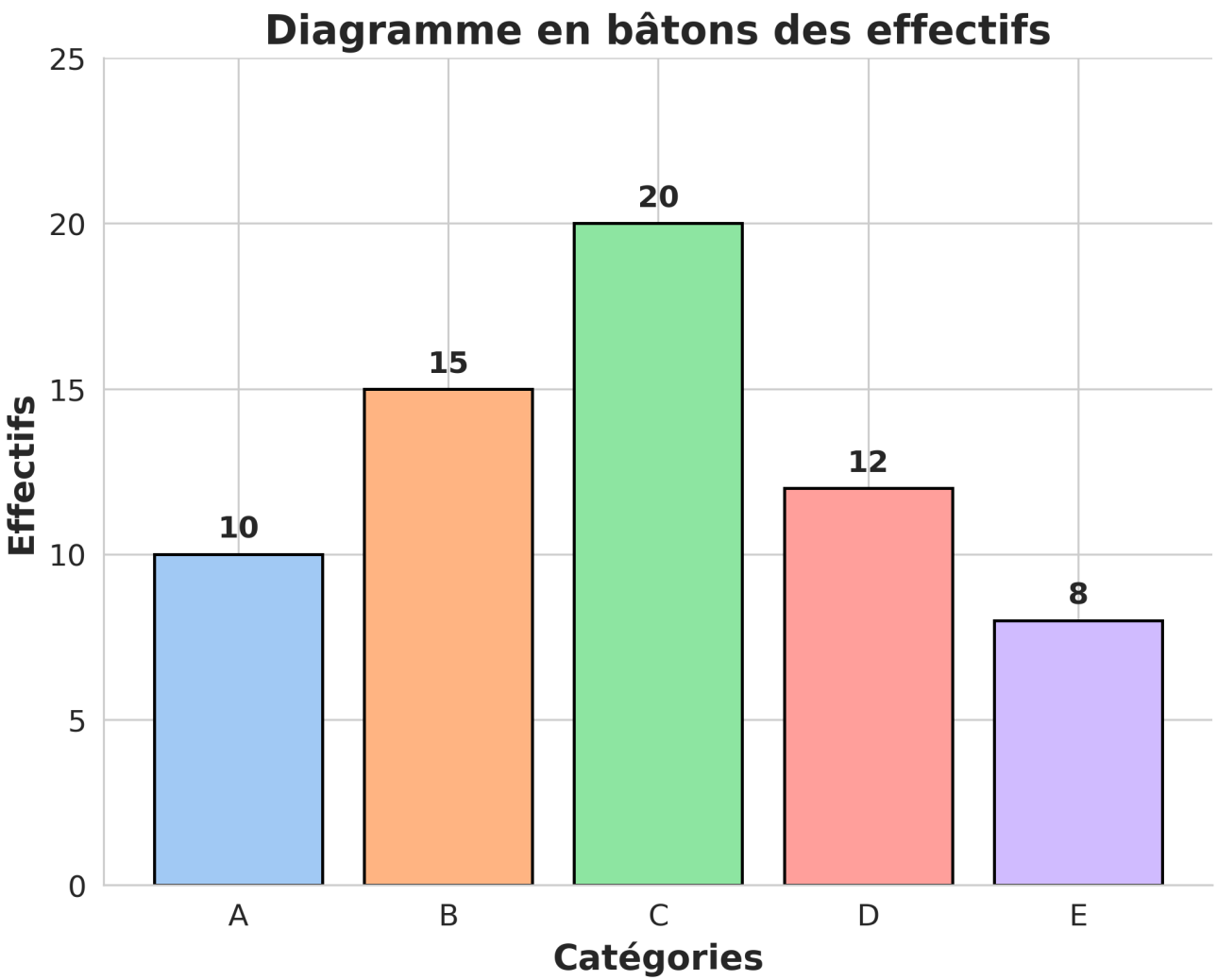


Diagramme à moustaches (Boxplot)

Un **diagramme à moustaches** est un graphique permettant de visualiser sous forme d'une boîte rectangulaire la répartition d'une **variable quantitative** en montrant **la médiane, les quartiles** et les éventuelles **valeurs aberrantes**.

- La boîte représente l'intervalle interquartile (valeurs entre Q1 et Q3).
- La ligne à l'intérieur de la boîte donne la valeur de la médiane (Q2). Cette ligne est horizontale s'il s'agit d'un boxplot vertical.
- Les moustaches (limites supérieure et inférieure des graphiques) montrent l'étendue des données jusqu'à 1,5 fois l'écart interquartile, c'est-à-dire l'intervalle
  $[Q_1 - 1.5 \times (Q_3 - Q_1),\ Q_3 + 1.5 \times (Q_3 - Q_1)]$
- Les éventuels points isolés (hors de la boîte) représentent les valeurs aberrantes .

Diagramme à moustaches (Boxplot)

Importance

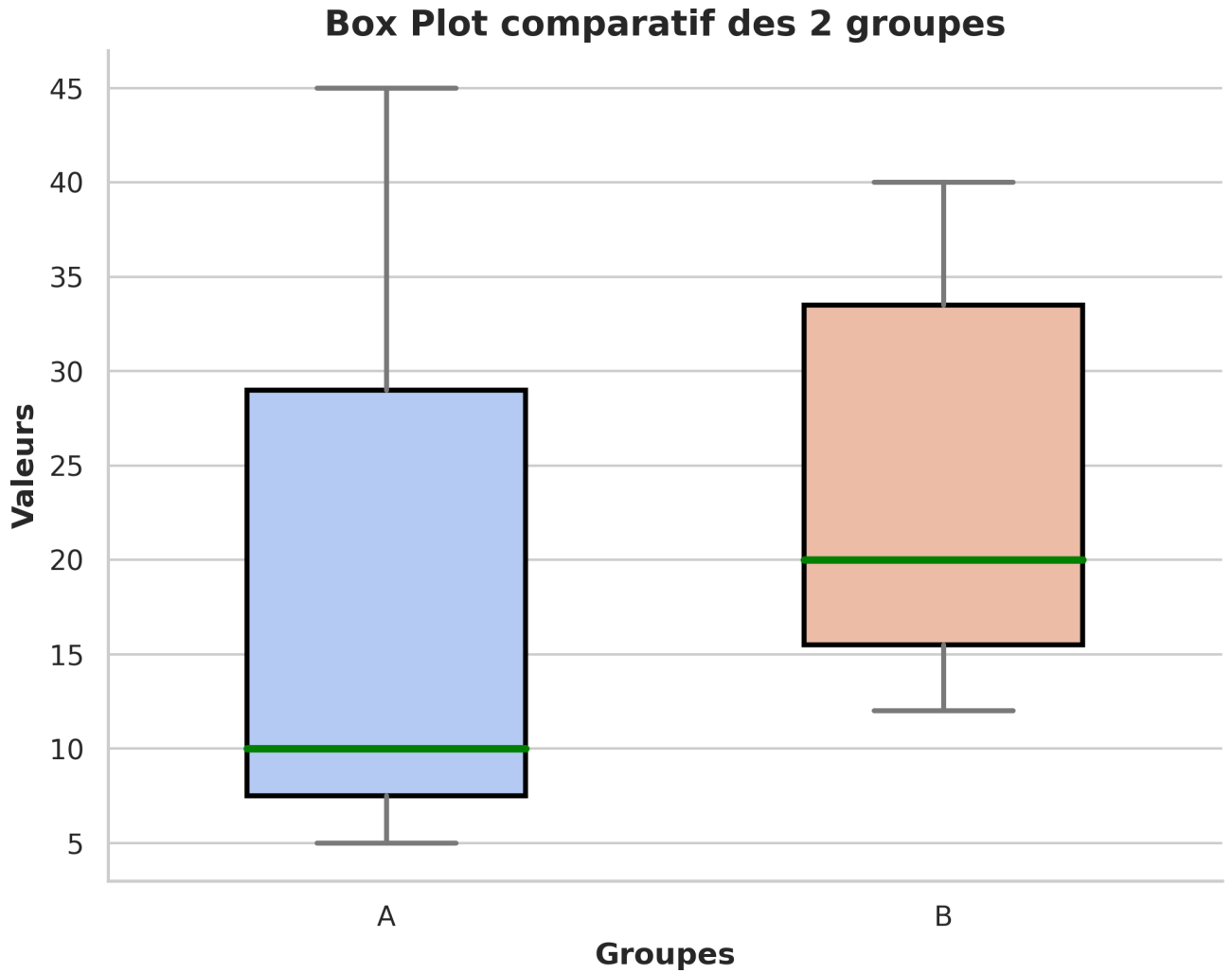
- Les boxplots sont souvent utilisés pour une analyse comparative de la distribution de deux variables.
- Il renseigne sur la symétrie ou l'asymétrie de la distribution, la dispersion des données et permet de détecter la présence de valeurs aberrantes dans les données.

Limites

Il est moins informatif pour des données asymétriques ou d'étendues très petites.

Diagramme à moustaches (Boxplot)

Voici un exemple où l'on trace des boxplots pour comparer la répartition (valeur moyenne et variabilité) des âges dans deux groupes A et B.

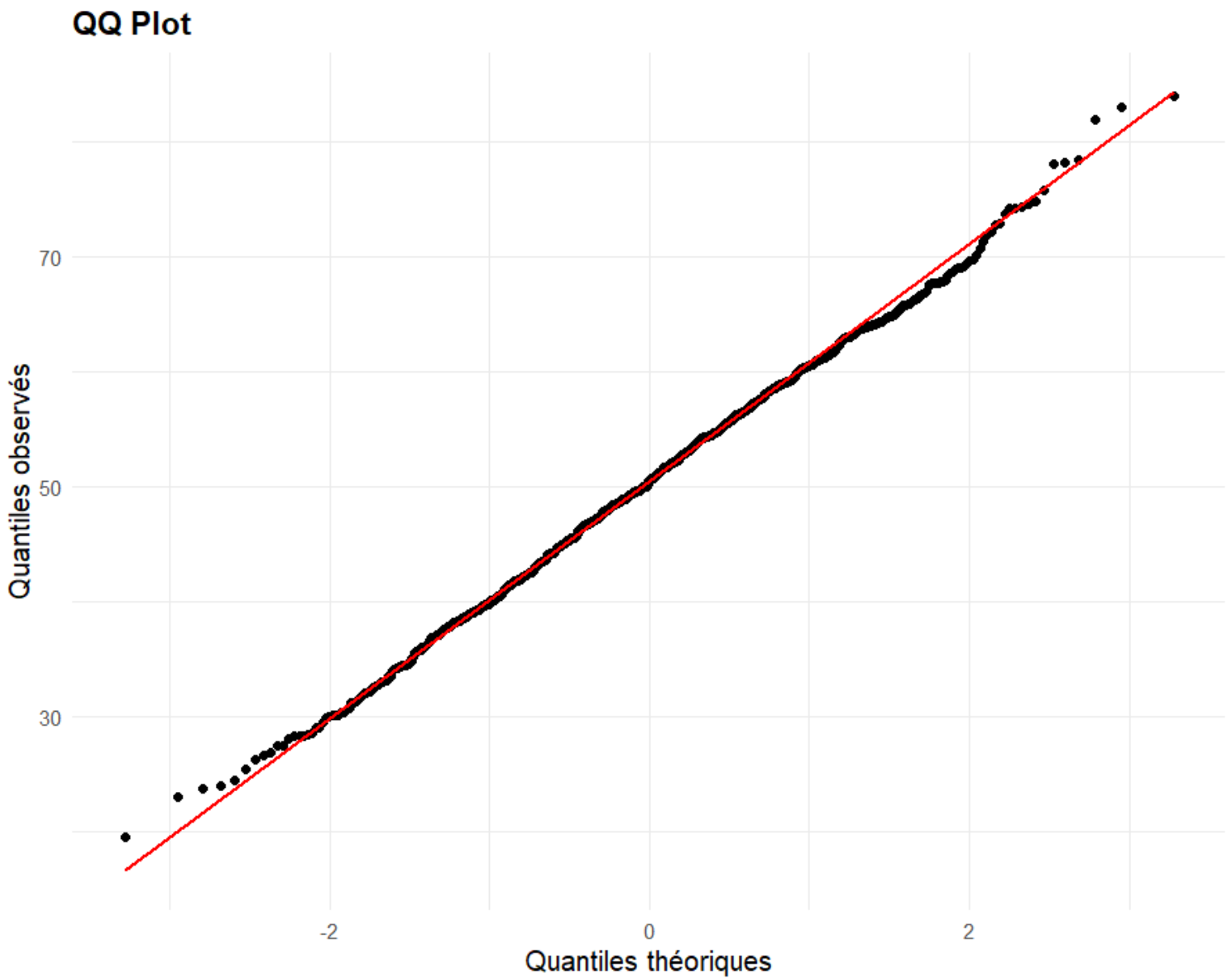


QQ - Plot

Le **QQ Plot** est un graphique qui permet de comparer la distribution d'un jeu de données avec une distribution théorique, souvent la loi normale. Il sert principalement à vérifier visuellement si les données suivent la distribution théorique en question.

- Les quantiles du jeu de données sont tracés contre les quantiles de la distribution théorique.
- **Interprétation** :
  - Un alignement des quantiles empiriques et des quantiles théoriques sur la droite diagonale (**qqline**) suggère une similarité entre les deux distributions.
  - Des écarts aux extrémités suggèrent que l'on est en présence d'une distribution à queue lourde ou légère.
- **Limites** :
  Le QQ Plot ne fournit qu'une visualisation . Des tests statistiques sont nécessaires pour confirmer la normalité. (On en reparle quand on abordera la statistique inférentielle)

QQ -Plot


QQ Plot
Quantiles observés
70
50
30
-2
0
2
Quantiles théoriques

Formes de Distribution : Asymétrie et Aplatissement

Soit $y = (y_1, \ldots, y_n)$ un échantillon de n mesures d'une variable continue. La variance empirique et la variance empirique corrigée sont :

$$s_n^2(y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \quad \text{et} \quad s_{n-1}^2(y) = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2,$$

On peut calculer des indicateurs complémentaires, qui permettent de mieux cerner la distribution des données.

Asymétrie

Le coefficient de dissymétrie (**skewness**) est un paramètre de forme qui mesure l'asymétrie d'un échantillon.

$$\gamma_1 = \frac{1}{s_n^3}\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^3\right)$$

Formes de Distribution : Asymétrie et Aplatissement

Interprétation

Un coefficient de dissymétrie négatif indique une concentration des données à droite de la moyenne et donc une queue de distribution qui s'étale vers la gauche. S'il est positif, les données sont plutôt concentrées à gauche de la moyenne et la queue s'étale vers la droite. Si ce coefficient est nul ou proche de 0, la distribution est symétrique.

Aplatissement

Le coefficient d'aplatissement (**kurtosis**) est un paramètre de forme qui mesure, à écart-type égal, si les données se regroupent de manière pointue ou de plate autour de la moyenne. Il est défini comme suit :

$$\gamma_2 = \frac{1}{s_n^4}\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^4\right)$$

Formes de Distribution : Asymétrie et Aplatissement

Interprétation

Un coefficient d'aplatissement inférieur à 3 indique des données relativement aplaties. Un coefficient d'aplatissement supérieur à 3 indique des données plus pointues que la normale avec des valeurs extrêmes.

Les deux graphiques ci-après illustrent différentes formes de distribution des données.

- Sur la figure de gauche, nous avons une distribution asymétrique (skewed), où les données sont plus concentrées à gauche avec une queue de distribution qui s'étale vers la droite.
- À droite, nous observons une distribution avec un aplatissement élevé, caractérisée par des valeurs extrêmes et une concentration très marquée autour de la moyenne.

Asymétrie et Aplatissement

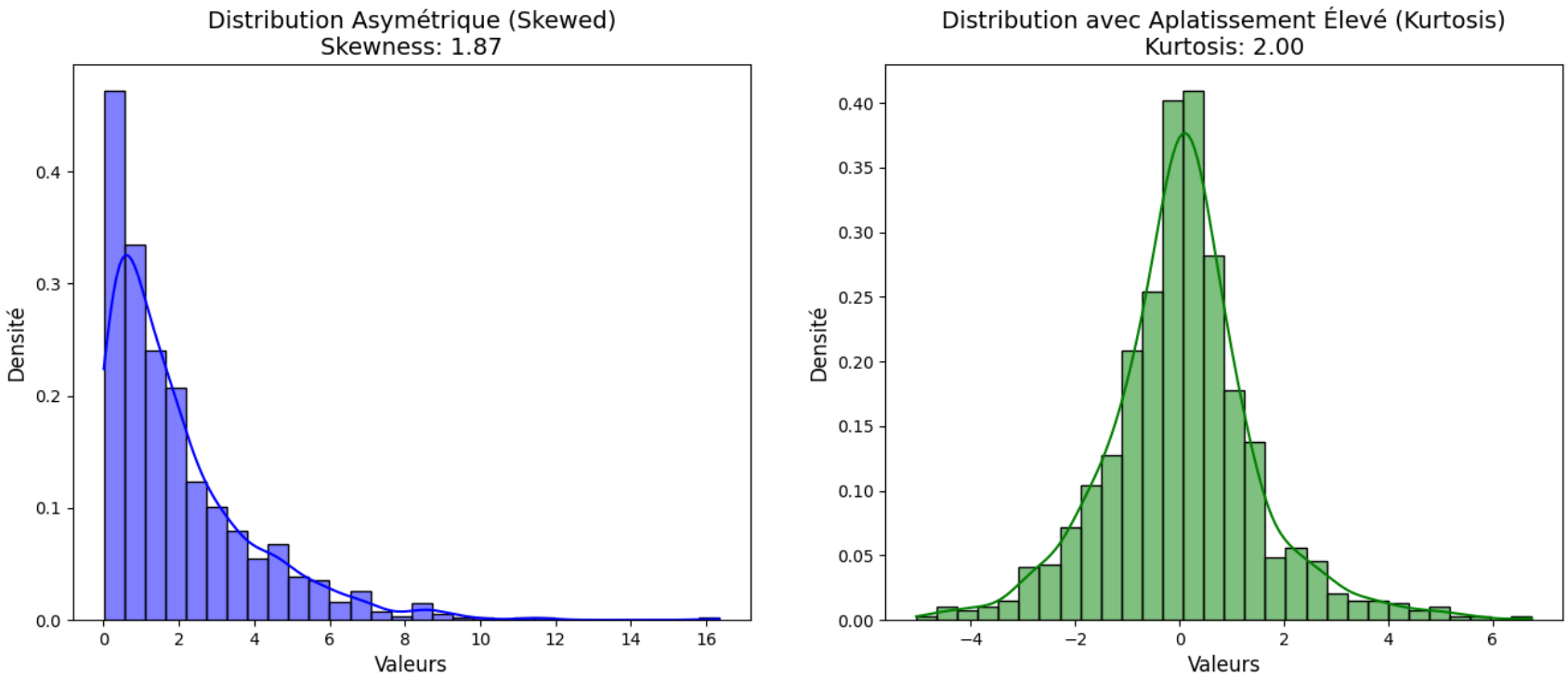


Figure: Illustration de l'asymétrie et de l'aplatissement

À suivre ...

Curieux d'en savoir plus ?
Ce qui vous attend prochainement :

- **Relations entre deux variables** : Nuage de points, Corrélation, ...
- **Une mise en pratique avec un projet de Statistique descriptive**
- **Inférence statistique**

A très bientôt !