Chapitre 4 - Analyse descriptive et visualisation des données

Neila Mezghani

(Hiver 2025)

COLLÈGE DE
BOIS-DE-BOULOGNE

Plan du cours

1. Les données
   - Terminologie
   - Représentation des données
2. Types de variables
   - Variable qualitative
   - Variable quantitative
3. Traitement des variables qualitatives
   - Description des variables qualitatives
   - Représentation graphique des données qualitatives
4. Traitement des variables quantitatives
   - Description des variables quantitatives
   - Représentation graphique des données quantitatives

Plan du cours

1. Les données
   - Terminologie
   - Représentation des données
2. Types de variables
   - Variable qualitative
   - Variable quantitative
3. Traitement des variables qualitatives
   - Description des variables qualitatives
   - Représentation graphique des données qualitatives
4. Traitement des variables quantitatives
   - Description des variables quantitatives
   - Représentation graphique des données quantitatives

Plan du cours

1. Les données
   - Terminologie
   - Représentation des données
2. Types de variables
   - Variable qualitative
   - Variable quantitative
3. Traitement des variables qualitatives
   - Description des variables qualitatives
   - Représentation graphique des données qualitatives
4. Traitement des variables quantitatives
   - Description des variables quantitatives
   - Représentation graphique des données quantitatives

Terminologie

Une donnée est...

- un **enregistrement** caractérisé par un ensemble de champs (terminologie des bases de données).
- un **individu** défini par un ensemble de caractéristiques ou de paramètres ou de variables (terminologie issue de la statistique).
- une **instance** caractérisée par un ensemble d'attributs (terminologie orientée objet en informatique).
- un **point** ou un **vecteur** caractérisé par ses coordonnées dans un espace vectoriel (terminologie de l'algèbre).

Plan du cours

1. Les données
   - Terminologie
   - Représentation des données
2. Types de variables
   - Variable qualitative
   - Variable quantitative
3. Traitement des variables qualitatives
   - Description des variables qualitatives
   - Représentation graphique des données qualitatives
4. Traitement des variables quantitatives
   - Description des variables quantitatives
   - Représentation graphique des données quantitatives

Représentation des données

Les données sont généralement représentées sous la forme d'un tableau rectangulaire (ou matrice) à N lignes représentant les individus et K colonnes correspondant aux variables. On note X la matrice de dimension (N, K) contenant les données.

$$X = \begin{pmatrix} x_1^1 & . & . & x_1^K \\ x_2^1 & . & . & x_2^K \\ . & . & . & . \\ x_N^1 & . & . & x_N^K \end{pmatrix}$$

où x_i^j est la valeur de l'individu i pour la variable j.
On notera $\mathbf{x}_i = (x_i^1, ..., x_i^K)'$ le vecteur des variables de l'individu i et $\mathbf{x}^j = (x_1^j, ..., x_n^j)'$ le vecteur des individus de la variable j.

Exemple : Types de données (1/2)

Soit la base de données `heart.txt` qui décrit un ensemble de patients avec leurs caractéristiques cliniques

```
#Lecture de la base de donnees heart.txt
df = pd.read_table("heart.txt",sep = '\t',header = 0)
```

```
#Lecture de la base de donnees carsPreprocessing.xlsx
df = pd.read_excel('carsPreprocessing.xlsx', sheet_name='cars')
```

Exemple : Types de données (2/2)

Pour décrire la base de données, on utilise `df.shape`

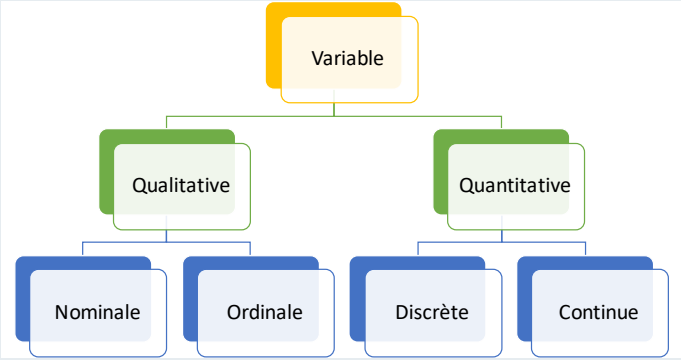
```
# Dimension de la base de donnees
dimension = df.shape
NbrLignes = df.shape[0]
NbrColonnes = df.shape[1]
```

Plan du cours

1. Les données
   - Terminologie
   - Représentation des données
2. Types de variables
   - Variable qualitative
   - Variable quantitative
3. Traitement des variables qualitatives
   - Description des variables qualitatives
   - Représentation graphique des données qualitatives
4. Traitement des variables quantitatives
   - Description des variables quantitatives
   - Représentation graphique des données quantitatives

Type de variables

La détermination du type de chaque variable est une étape nécessaire en apprentissage machine : Cela détermine notamment les analyses statistiques qu'il est permis d'effectuer et les méthodes d'apprentissage machine à utiliser.



Plan du cours

1. Les données
   - Terminologie
   - Représentation des données
2. Types de variables
   - Variable qualitative
   - Variable quantitative
3. Traitement des variables qualitatives
   - Description des variables qualitatives
   - Représentation graphique des données qualitatives
4. Traitement des variables quantitatives
   - Description des variables quantitatives
   - Représentation graphique des données quantitatives

Variable qualitative (1/2)

- Une variable est dite **qualitative** si ses valeurs ne sont pas mesurables.
- Le sexe, la profession, l'état matrimonial sont quelques exemples de variables qualitatives. Les valeurs d'une variable qualitative sont appelées **modalités**.

Variable qualitative (2/2)

- Une variable qualitative est dite **nominale** si ses modalités ne sont pas ordonnées naturellement. Par exemple, dans une population de personnes actives, la profession est une variable nominale.
- Une variable qualitative est dite **ordinale** si ses modalités suivent une relation d'ordre. Par exemple, une pathologie peut prendre la valeur `légère`, `modérée` ou `sévère`. Ces valeurs peuvent être ordonnées : `légère` $<$ `modérée` $<$ `sévère`.

Plan du cours

1. Les données
   - Terminologie
   - Représentation des données
2. Types de variables
   - Variable qualitative
   - Variable quantitative
3. Traitement des variables qualitatives
   - Description des variables qualitatives
   - Représentation graphique des données qualitatives
4. Traitement des variables quantitatives
   - Description des variables quantitatives
   - Représentation graphique des données quantitatives

Variable quantitative

- Une variable quantitative est dite **discrète** si elle ne peut prendre que des valeurs qui peuvent être énumérées.
- La variable quantitative est dite **continue** si ses valeurs potentielles ne peuvent pas être énumérées.
- Les variables **binaires** sont des variables quantitatives discrètes qui possèdent des propriétés particulières.

Variable quantitative binaire

- **Symétriques** : une variable binaire est dite symétrique si ses deux modalités ont la même importance, c'est-à-dire si celles-ci peuvent être indifféremment codées par 0 ou 1. Par exemple, la variable sexe est une variable symétrique parce qu'elle peut être codée par 0 ou 1 pour masculin de même que pour féminin sans aucune différence.
- **Asymétriques** : une variable binaire est dite asymétrique si les deux modalités n'ont pas la même importance. Par exemple, le résultat d'un examen médical ne peut pas être codé par 0 si l'examen est positif et 1 si l'examen est négatif vu l'importance du résultat attendu de l'examen.

Exemple : Types de données (2/3)

Le type des données est donné par :

```
print(df.dtypes)
```

On obtient :

```
age              int64
sexe            object
type_douleur    object
pression         int64
cholester        int64
sucre           object
electro         object
```

Exemple : Types de données (2/3)

On peut également utiliser :

```
print(df.info())
```

Dans ce cas on obtient :

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 270 entries, 0 to 269
Data columns (total 13 columns):
#   Column         Non-Null Count  Dtype
---  ------         --------------  -----
0   age            270 non-null    int64
1   sexe           270 non-null    object
2   type_douleur   270 non-null    object
3   pression       270 non-null    int64
```

Exemple : Types de données (3/3)

L'affichage suivant est plus d'un niveau de programmation.

```
age              int64
sexe            object
type_douleur    object
pression         int64
cholester        int64
sucre           object
electro         object
```

Ce langage doit être traduit en disant par exemple : la variable `type_douleur` est une variable qualitative.

Plan du cours

1. Les données
   - Terminologie
   - Représentation des données
2. Types de variables
   - Variable qualitative
   - Variable quantitative
3. Traitement des variables qualitatives
   - Description des variables qualitatives
   - Représentation graphique des données qualitatives
4. Traitement des variables quantitatives
   - Description des variables quantitatives
   - Représentation graphique des données quantitatives

Plan du cours

1. Les données
   - Terminologie
   - Représentation des données
2. Types de variables
   - Variable qualitative
   - Variable quantitative
3. Traitement des variables qualitatives
   - Description des variables qualitatives
   - Représentation graphique des données qualitatives
4. Traitement des variables quantitatives
   - Description des variables quantitatives
   - Représentation graphique des données quantitatives

Description des variables qualitatives (1)

Soit un individu décrit par une variable qualitative x pouvant prendre c modalités $(a_1, a_2, ..., a_i, ...a_c)$.

- **L'effectif**
  L'effectif, aussi appelé fréquence absolue, est le nombre d'individus n_i dont la variable x présente la modalité a_i.
- **La fréquence**
  La fréquence de la modalité est le rapport entre l'effectif et le nombre total d'individus.
  $$f_j = \frac{n_i}{N}$$
  N est le nombre total d'individus.
- `DataFrame.value_counts`

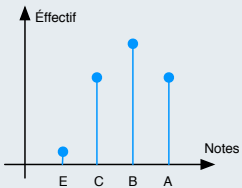
Plan du cours

1. Les données
   - Terminologie
   - Représentation des données
2. Types de variables
   - Variable qualitative
   - Variable quantitative
3. Traitement des variables qualitatives
   - Description des variables qualitatives
   - Représentation graphique des données qualitatives
4. Traitement des variables quantitatives
   - Description des variables quantitatives
   - Représentation graphique des données quantitatives

Représentation graphique des données qualitatives (1/2)

Diagramme en bâtons

- C'est une représentation graphique de la distribution d'une variable statistique qualitative à l'aide de segments verticaux ou horizontaux.
- Il s'apparente à un diagramme à bandes. Toutefois, le diagramme à bandes est plutôt utilisé pour une variable quantitative alors que le diagramme à bâtons est utilisé dans le cas d'une variable qualitative.
- `DataFrame.plot.bar`



Représentation graphique des données qualitatives (2/2)

Diagramme circulaire

- Un **diagramme circulaire**, aussi appelé **camembert** ou **tarte**, est une représentation graphique de données qualitatives sous la forme d'un disque partagé en secteurs
- À chaque modalité étudiée correspond un secteur. Les mesures des secteurs sont proportionnelles aux effectifs représentés.
- `DataFrame.plot.pie`

Plan du cours

1. Les données
   - Terminologie
   - Représentation des données
2. Types de variables
   - Variable qualitative
   - Variable quantitative
3. Traitement des variables qualitatives
   - Description des variables qualitatives
   - Représentation graphique des données qualitatives
4. Traitement des variables quantitatives
   - Description des variables quantitatives
   - Représentation graphique des données quantitatives

Plan du cours

1. Les données
   - Terminologie
   - Représentation des données
2. Types de variables
   - Variable qualitative
   - Variable quantitative
3. Traitement des variables qualitatives
   - Description des variables qualitatives
   - Représentation graphique des données qualitatives
4. Traitement des variables quantitatives
   - Description des variables quantitatives
   - Représentation graphique des données quantitatives

Description des données quantitatives

Nous distinguons trois grandes familles :

- **Caractéristiques de tendances centrales** : indiquent l'ordre de grandeur des données et de leur valeur centrale, c'est-à-dire la position autour de laquelle se rassemblent ces valeurs.
- **Caractéristiques de dispersion** : servent à évaluer la variabilité des données et à résumer l'éloignement de l'ensemble des individus par rapport à leur tendance centrale.
- **Caractéristiques de formes** : De nombreux phénomènes physiques se distinguent par des variables quantitatives qui suivent une loi normale. Ces principales caractéristiques de formes sont le coefficient d'asymétrie et l'aplatissement de Fisher.

Description des données quantitatives

Caractéristiques de tendances centrales

- La moyenne arithmétique
- La moyenne arithmétique pondérée
- La médiane
- Les quartiles

Caractéristiques de dispersions

- L'étendue
- La variance et écart-type

Caractéristiques de formes

- Asymétrie
- Aplatissement

Caractéristiques de tendances centrales (1/4)

La moyenne arithmétique

- La moyenne arithmétique μ est la somme des valeurs de la variable j pour tous les individus i, $i = 1, ..., N$.

$$\mu_j = \frac{1}{N} \sum_{i=1}^{N} x_i^j \tag{1}$$

  La moyenne arithmétique est sensible aux valeurs aberrantes.

- `DataFrame.mean`

Caractéristiques de tendances centrales (2/4)

La moyenne arithmétique pondérée

- Lorsque les variables n'ont pas la même importance, on attribue un poids à chacune d'entre elles. Dans ce cas, on calcule la moyenne arithmétique pondérée :

$$\lambda_j = \frac{\sum_{i=1}^{N} w_i x_i^j}{\sum_{i=1}^{N} w_i} \tag{2}$$

Caractéristiques de tendances centrales (3/4)

La médiane

- Soit un ensemble de N données rangées par ordre croissant. La médiane est la valeur de la variable qui partage l'ensemble des données en deux parties de même effectif.
  - Si N est impair ($N = 2n + 1$), alors la médiane est la donnée de rang n.
  - Si N est pair ($N = 2n$), alors la médiane est la donnée de rang n ou de rang $n + 1$ ou bien la moyenne des deux.
- Dans le cas d'une distribution normale, la médiane et la moyenne sont égales.
- `DataFrame.median`

Caractéristiques de tendances centrales (4/4)

Les quartiles

- Soit un ensemble de N données rangées par ordre croissant.
- Les quartiles sont les valeurs Q1, Q2, Q3 de la variable qui partagent l'effectif en quatre sous-ensembles de même effectif.
- Le premier quartile (Q1) est la plus petite donnée de cet ensemble telle qu'au moins un quart des données sont inférieures ou égales à Q1. Le troisième quartile (Q3) est la plus petite donnée de cet ensemble telle qu'au moins les trois quarts des données de l'ensemble de données ordonnées sont inférieurs ou égaux à Q3.
- Le deuxième quartile Q2 correspond à la médiane.
- `DataFrame.quantile`

Description des données quantitatives

Caractéristiques de tendances centrales

- La moyenne arithmétique
- La moyenne arithmétique pondérée
- La médiane
- Les quartiles

Caractéristiques de dispersions

- L'étendue
- La variance et écart-type

Caractéristiques de formes

- Asymétrie
- Aplatissement

Caractéristiques de dispersion (1/2)

L'étendue

- L'étendue est l'écart entre la plus grande et la plus petite des valeurs.
- Cette caractéristique, qui dépend des valeurs aberrantes, est par conséquent peu fiable.

Caractéristiques de dispersion (2/2)

La variance et écart-type

- La variance (`DataFrame.var`)est la moyenne des carrés des écarts à la moyenne.

$$var(x) = \sigma_x^2 = \frac{1}{n}\sum_{i=1}^{N}(x_i - \mu)^2$$

avec μ la moyenne arithmétique de l'ensemble de données.

- L'écart-type σ_x (`DataFrame.std`) est la racine carrée positive de la variance :

$$\sigma_x = \sqrt{\frac{1}{n}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Description des données quantitatives

Caractéristiques de tendances centrales

- La moyenne arithmétique
- La moyenne arithmétique pondérée
- La médiane
- Les quartiles

Caractéristiques de dispersions

- L'étendue
- La variance et écart-type

Caractéristiques de formes

- Asymétrie
- Aplatissement

Le coefficient d'asymétrie (1/3)

- Le coefficient d'asymétrie (*skewness* en anglais) correspond au moment d'ordre trois de la variable centrée réduite. Pour une distribution uniforme x, le coefficient d'asymétrie est donné par la formule :

$$\gamma_1 = E\left[\left(\frac{x-\mu}{\sigma}\right)^3\right]$$

où E désigne l'espérance de x, μ la moyenne et σ l'écart type. Ou encore

$$\gamma_1 = \frac{\mu_3}{{\mu_2}^{3/2}}$$

avec μ_i les moments centrés d'ordre i.

Le coefficient d'asymétrie (2/3)

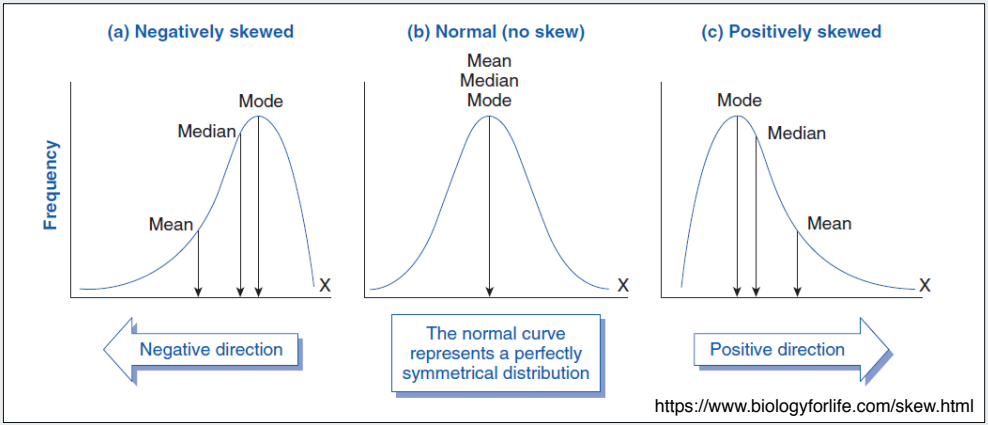
Le moment centré d'un échantillon de donnée est fournit par :

$$\mu_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^k$$

Trois cas se présentent :

- $\gamma_1 < 0$ si la distribution de la variable est étalée à la gauche (la queue est à gauche) et la moyenne est inférieure à la médiane
- $\gamma_1 = 0$ si la distribution de la variable est symétrique
- $\gamma_1 > 0$ si la distribution de la variable est étalée vers la droite (la queue est à droite) et la moyenne est supérieure à la médiane
- `DataFrame.skew`

Le coefficient d'asymétrie (3/3)



https://www.biologyforlife.com/skew.html

Le coefficient d'aplatissement (1/3)

- Le coefficient d'aplatissement (*kurtosis* en anglais) correspond au quotient du moment d'ordre quatre de la variable centrée réduite par la puissance quatrième de l'écart type. Il est donné par l'équation suivante :

$$\gamma_2 = E\left[\left(\frac{X-\mu}{\sigma_x}\right)^4\right]$$

- `DataFrame.kurtosis`

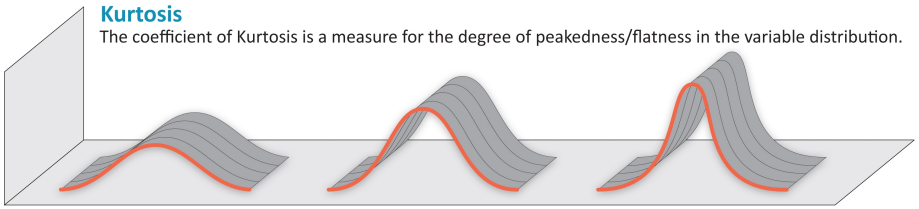
Le coefficient d'aplatissement (2/3)

- **Le coefficient d'aplatissement** :

$$\gamma_2 = E\left[\left(\frac{X-\mu}{\sigma_x}\right)^4\right]$$

  - Si $\gamma_2 = 0$ alors la distribution est dite **musokurtique**.
  - Si $\gamma_2 > 0$, alors la distribution est plus concentrée que la normale ; elle est dite **leptokurtique**. Dans le domaine de la finance, ce coefficient sert à déterminer des valeurs anormales plus fréquentes.
  - Si $\gamma_2 < 0$, alors la distribution est plus aplatie que la normale. Elle est dite **platikurtique**. Dans le domaine de la finance, ce coefficient sert à déterminer des valeurs anormales plus fréquentes.

Le coefficient d'aplatissement (3/3)



https://www.oreilly.com/library/view/machine-learning-for/9781786469878/f2cbeb0b-4094-4cb9-be54-2d046ff79d7f.xhtml

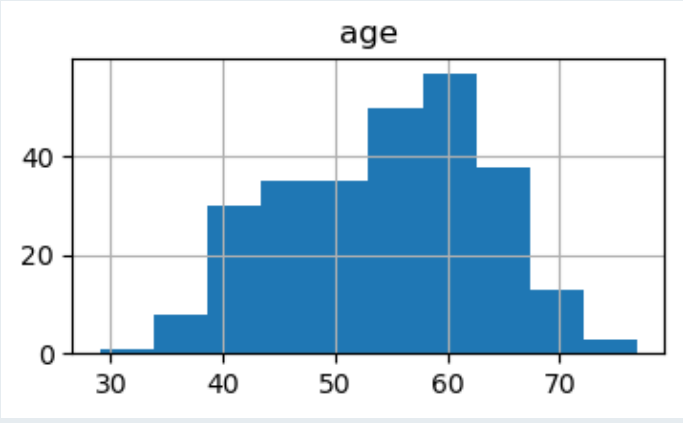
Plan du cours

1. Les données
   - Terminologie
   - Représentation des données
2. Types de variables
   - Variable qualitative
   - Variable quantitative
3. Traitement des variables qualitatives
   - Description des variables qualitatives
   - Représentation graphique des données qualitatives
4. Traitement des variables quantitatives
   - Description des variables quantitatives
   - Représentation graphique des données quantitatives

Histogramme (1/2)

- C'est un diagramme qui permet de représenter par des bandes juxtaposées une distribution d'une variable statistique quantitative.
- Pour tracer un histogramme de variables continues, nous procédons à leurs regroupement en classes.
- `DataFrame.hist`

Histogramme (2/2)



Boite à moustache (1/2)

- Le **diagramme en boîte** ou **boîte à moustache** d'un ensemble de données est une représentation graphique de ses caractéristiques statistiques à savoir la médiane (Q2), les quartiles (Q1, Q2 et Q3) et l'étendue.
- `DataFrame.boxplot`

Boite à moustache (2/2)

- Les limites du rectangle correspondent au premier et troisième quartile Dans le domaine de la finance,
- Les limites des segments correspondent à la valeur minimale et maximale de l'ensemble des données.