# Data Exploration

## objectif :

« Introduire des méthodes qui nous permet d'extraire de l'information pertinente d'un jeu de données.

A l'issue de cette formation, vous serez capable de :

→ Faire apparaître des comportements particuliers des objets observés (détecter les individus ayant un comportement atypique, trouver des ensembles d'individus ayant un comportement similaire)

→ Trouver des liens entre les variables étudiés

→ Utiliser R.

## Corpus / jeu de données :

un jeu de données est un tableau avec :

- en ligne les individus (observations, unité statistique)
- en colonne les variables (attributs etudiés)

Exp :

| | couleur yeux | couleur cheveux | taille |
|---|---|---|---|
| Karine | Noir | Blond | 164 |
| Olivier | Bleu | Noir | 174 |
| Sarah | Marron | Noir | 158 |
| Paul | Noir | Brun | 180 |

Dans notre exemple:

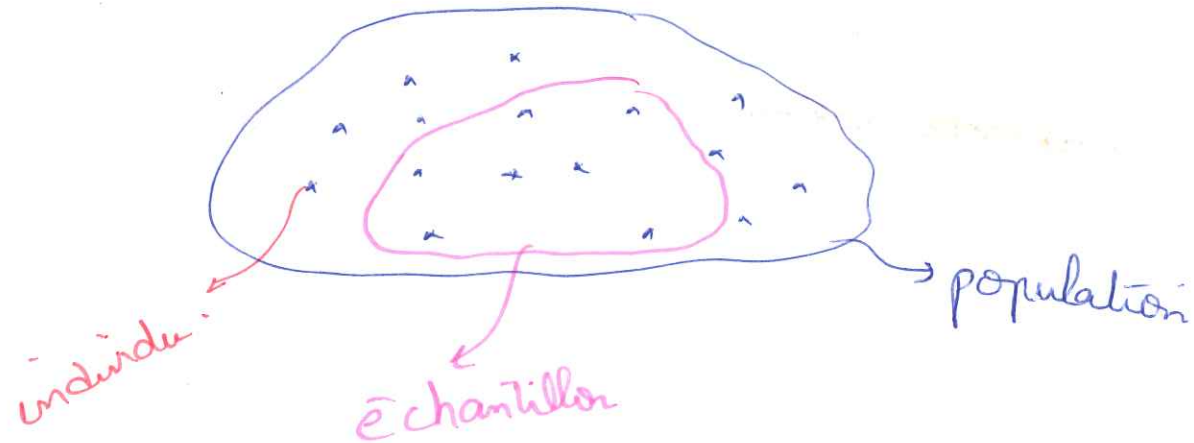
+ les individus sont les étudiants
+ les variables sont: couleur de yeux, couleur de cheveux, taille,

→ A partir de ce tableau, tirer les informations importante pour qu'il soit compréhensible pour tout le monde donc on fait un résumé numérique (Moyenne, Médiane...) ou une représentation graphique (diagramme en barre, histogramme ---) présentation synthétique

## Population et échantillon:

→ population = Ensemble des individus

→ Echantillon = Sous ensemble de la population.




→ Recensement: Etude de tous les individus

→ Sondage: sous ens de la population.

⚠ Si la population infinie alors le Recensement est impossible → on fait un sondage.

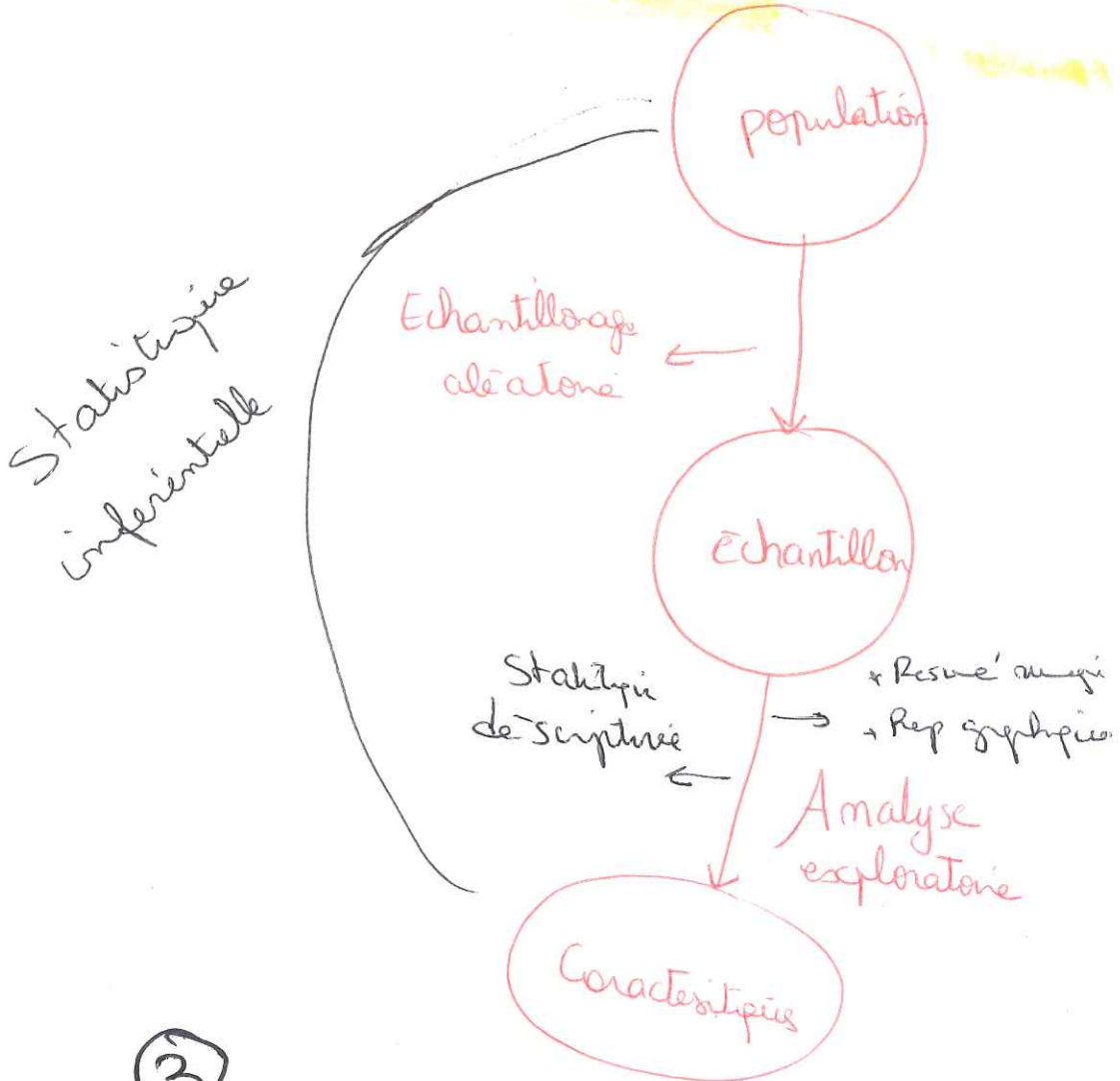
# 2 types d'études statistique

## Analyse exploratoire (Statistique descriptive)

- (Tableau de contingence)
- Résumé numérique
- Représentation graphique
- Recherche de sous groupes homogènes

## Statistique inférentielle

- Trouver des estimateurs sans biais et efficace pour passer de l'échantillon à la population.



→ Chaque individu est décrit par un ensemble de caractéristiques appelé "Variable" ou attributs.

## Nature Variable

### Quantitatives
prend des valeurs numériques

- **discrète** → Exp: nb enfants dans une famille
- **Continue** → Exp: taille, poids

### Qualitatives
prend des modalités

- **ordinal** → Mentions Bac, Collège, résistance d'un produit
- **Nominales** → couleur de yeux

on a aussi dichotomique: Exp: a fumer ou non fumeur, atteint maladie ou Non

# Types d'analyse :

## Statistiques uni variées

↓

On s'intéresse à une seule variable

## Statistiques Bi variées

↓

On s'intéresse à deux variables simultanément et on fait étude selon la nature de ces deux variables.

## Statistiques multivariées

↓

On s'intéresse à plusieurs ou $p$ variables

Dans la suite on s'intéresse à :

# Statistiques univariée

- Représentat° syntétique
- Représentat° graphique
- Résumé numérique

# Representation Synthetique

Exp:

| étudiant | Mention Bac | couleur Yeux | couleur cheveux |
|---|---|---|---|
| A | Bien | Vert | Marron |
| B | Bien | Noir | Noir |
| C | Bien | Bleu | Noir |
| D | TB | Vert | Noir |
| E | AB | Noir | Noir |
| F | Bien | Noir | Marron |
| G | AB | Marron | Noir |
| H | T.B | Noir | Marron |
| I | Bien | Bleu | Blond |
| J | T.B | Noir | Noir |

↝ jeu de donnée

A partir de "Jeu de donnés", on peut créer pour chaque variable un "tableau de contingence"

Exp:
- Variable "Mention en Bac":
  - les modalités sont: AB - Bien - TB
  - On compte pour chaque modalité, le nombre des individus (ici etudiants) ont ce modalité.

Exp: on a: 5 etudiants ont mention Bien en Bac ou [5] c'est l'effectif de la modalité Bien

- On fait pareil pour les 2 autres modalités "AB" et "TB" et on stocke tout dans un tableau →

| | Mention | effectif | freq |
|---|---|---|---|
| $i=1$ | AB | 2 ~ $n_1=2$ | $\frac{2}{10}=0.2$ ou 20% |
| $i=2$ | B | 5 ~ $n_2=5$ | $\frac{5}{10}=0.5$ ou 50% |
| $i=3$ | TB | 3 ~ $n_3=3$ | $\frac{3}{10}=0.3$ ou 30% |
| | total | 10 ~ car on a 10 étudiants | |

Fréquence = $f_i = \frac{n_i}{n}$ ~ eff chq modalité / ~ total

ou ~ freq en pourcentage = $f_i \times 100$

Interpretation : Il ya 50% des étudiants qui ont mention "Bien" en Bac.

⚠

## Représentation graphique

### ① Variables qualitatives :

① Variables nominales : diag en barre ou diagramme circulaire

② Variables ordinales : Diag en bâtons

# ② Variable ordinale : Diagramme en bâton

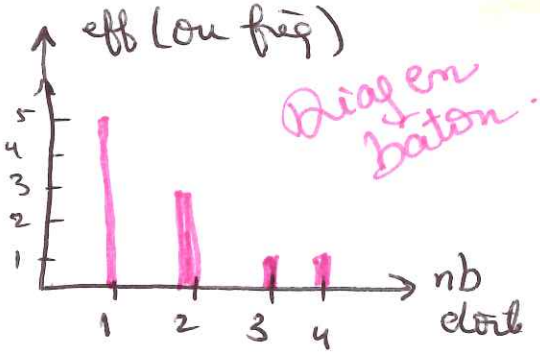
- Variable quantitative :

→ Discrète → diag en bâton

→ Continue → histogramme.

Exp : nb d'ent attribué à 10 hôtels parisiens

| nb d'ent | Effectif |
|---|---|
| 1 | 5 |
| 2 | 3 |
| 3 | 1 |
| 4 | 1 |
| total | 10 |




Diag en bâton.

Exp : Ici les valeurs sont regroupés par des intervalles.

Exp : Âge enfants crèche municipale de 10 enfants (en mois)

| centre classe | âge | Effectif |
|---|---|---|
| 3 | [0, 6[ | 2 |
| 9 | [6, 12[ | 1 |
| 15 | [12, 18[ | 3 |
| 21 | [18, 24[ | 4 |
| | total | 10 |


Histogramme

⚠ Effectif (ou fréquence) augmente ⇓ Hauteur du rectangle aussi

# Résumés Numériques.

⚠ Que pour les variables quantitatives. ⚠

On distingue deux types de Résumé numérique

**Indicateur de position**

(Moyenne, mode, médiane, quartiles...),
Ils positionnent la série de valeurs observés autour d'une tendance centrale,
Exp: Moy. Âge des étudiants

**Indicateur de dispersion**

(Variance, écat type, étendue interquartile ...)
Indiquent la fluctuat° des valeurs de la série autour d'une tendance centrale en générale.

## ★ Indicateur de position:

① Mode: La valeur observée qui a le plus grand effectif.

Exp: nb étoile : Mode est 1 (qui a eff le plus haut = 5)

- Si donnée sont quanti continue, on parle alors de classe modale.

Exp: crèche, classe modale est: [18, 24[

⚠

⚠ • "La mode" est le seul indicateur où on peut calculer pour les variables qualitatives aussi:

Exp: Mention, le Mode est "Bien"

② Moyenne:

① Discrète: $\bar{x} = \frac{1}{n} \sum_i n_i x_i$

Exp: (Etoile):

$$\bar{x} = \frac{1}{10}\left[(1\times5)+(2\times3)+(3\times1)+(1\times4)\right]$$

$$= 1.8$$

② Continue:

$$\bar{x} = \frac{1}{n} \sum_i n_i . x_i$$

Ici le $x_i$ est le centre du classe.

⚠ classe $[a,b]$ ⇒ centre $(x_i) = \frac{a+b}{2}$.

Exp: (Crèche): $\bar{x} = \frac{1}{10}\left[(3\times2)+(9\times1)+(15\times3)+(4\times21)\right]$

$= 14,4$ mois

## ③ Quantiles:

- **Médiane**: c'est la valeur qui sépare une population en deux groupes d'effectifs égaux. Elle n'a pas de sens que si les données sont rangées par ordre croissant.

Ex: Médiane notes élèves en Maths est 13: c'_à_d:

50% des étudiants ont une note inférieur à 13.

et 50% (la moitié) ont une note supérieur à 13.

- **Quartiles**:

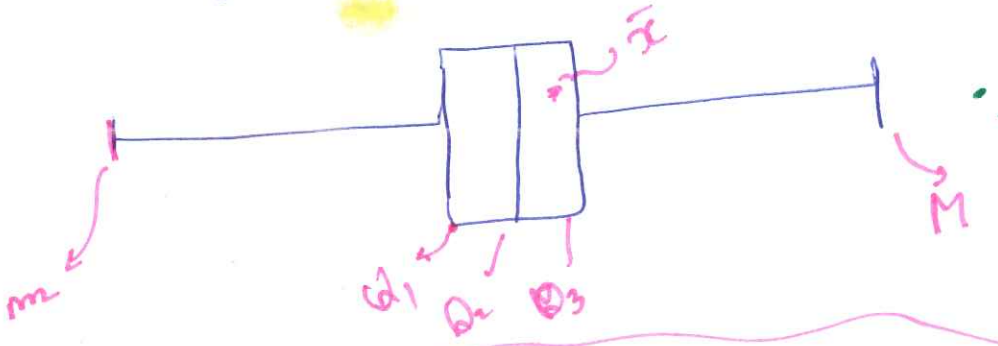
① 1er quartile $Q_1$: 25% des valeurs sont inf à $Q_1$ (ou les quarts de valeurs sont inf à $Q_1$ ; donc le $\frac{3}{4}$ restant sont sup à $Q_1$)

② 2e quartile $Q_2$: 50% des valeurs sont inf à $Q_2$ $\Rightarrow$ Médiane

③ 3e quartile $Q_3$: 75% des valeurs sont inf à $Q_3$: [ou $\frac{3}{4}$ des valeurs inf à $Q_3$ alors le $\frac{1}{4}$ restant sont sup à $Q_3$.]

# Boîte de Tuckey (Boîte à Moustaches. Boxplot).

Représente graphiquement les indicateurs de position :




m et M sont les moustaches :

$$\begin{cases} m = Q_1 - 1{,}5 \cdot (Q_3 - Q_1) \\ M = Q_3 + 1{,}5 \cdot (Q_3 - Q_1) \end{cases}$$

Les valeurs de la série en dehors des moustaches est considérée comme "atypique" ou valeur aberrante.

⚠ Boxplot nous permet d'avoir une aperçue graphique rapide de la distribution des valeurs de la série.

Exp:

| Note | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Elevé L | 9 | 10 | 8 | 7 | 10 | 9 | 11 |
| Elevé P | 14 | 2 | 16 | 5 | 6 | 5 | 16 |

Série ordonnée :

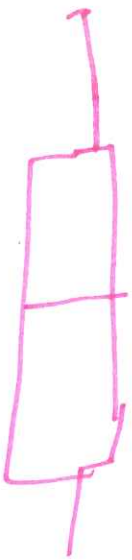
| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Elevé L : | 7 | 8 | 9 | (9) ← Me | 10 | 10 | 11 |
| Elevé P : | 2 | 5 | 5 | (6) ← Me | 14 | 16 | 16 |

**Elevé L :**

- Moyenne $\bar{x} = \frac{1}{7}(7+8+\dots+11) = 9.1$
- Médiane $Me = 9$
- $Q_1 = 8$
- $Q_3 = 10$
- $m = Q_1 - 1,5(Q_3 - Q_1)$
  $= 8 - 1,5(10-8) = 5$
- $M = Q_3 + 1,5(Q_3 - Q_1)$
  $= 10 + 1,5(10-8) = 13$


⚠ pas de valeurs aberrantes car tous les notes de l'elevé L sont entre (5) et (13).

**Elevé P :**



$\bar{x} = 9,1$
$Me = 6$
$Q_1 = 5$
$Q_3 = 14$
$m = 0$
$M = 27,5$

# Indicateurs de dispersion:

① **Variance**: Mesure l'écart au carré entre les valeurs de la série et leur moyenne.

$$V = S^2 = \frac{1}{n} \sum n_i (x_i - \bar{x})^2$$

⚠ tout le temps positif.

Exp: (étoile)

$$S^2 = \frac{1}{10} \left[ 5(1-1,8)^2 + 3(2-1,8)^2 + 1(3-1,8)^2 + 1(4-1,8)^2 \right]$$

- **Ecart-type**:

$$S = \sqrt{S^2}$$

- **Ecart-Médian**: Mesure l'écart entre les valeurs de la série et leur médiane:

$$e_m = \frac{1}{n} \sum_i n_i |x_i - \text{med}|$$

- **Etendue**: $\max\{x_i\} - \min\{x_i\}$

Exp (étoile): $4 - 1 = 3$

- **Ecart-inter-quartiles**: $Q_3 - Q_1$

# Variables centrées - réduites.

On définit la série centrée réduite de la façon suivante :

$$\tilde{x}_i = \left( \frac{x_i - \bar{x}}{s_x} \right)$$

($\bar{x}$ : moyenne ; $s_x$ : écart-type)

→ La série est dite :

. Centrée car elle a Moyenne nulle.

. Réduite car de variance 1.