Statistiques et probabilités
Cours n°6

Guillaume Postic

Université Paris-Saclay, Univ. Evry
Département informatique

Master 1 MIAGE - 2022/2023

Statistiques descriptives

Objectif : décrire, c'est-à-dire de résumer ou représenter, par des statistiques, les données disponibles quand elles sont nombreuses.

- Indicateurs de tendance centrale : moyenne, médiane
- Indicateurs de dispersion : variance, écart-type, IQR

Inférence statistique (1)

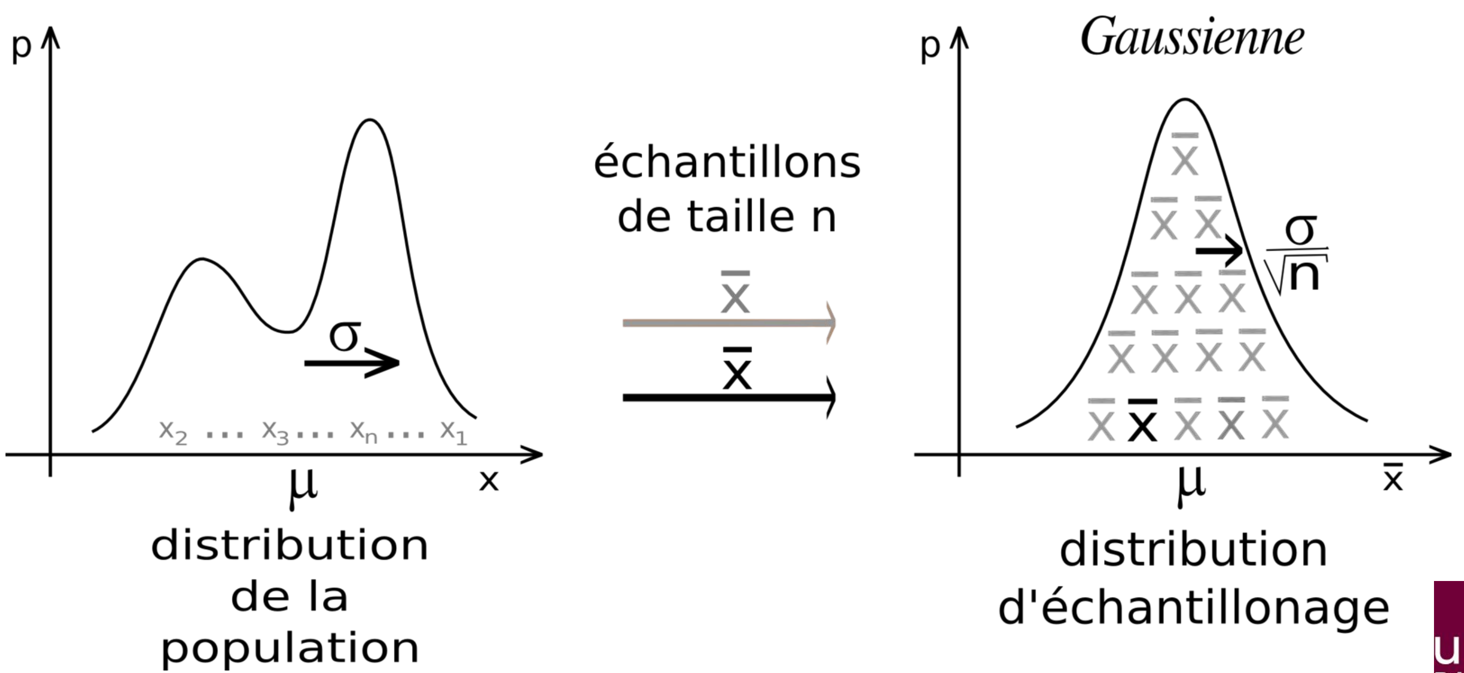
Ensemble des techniques permettant d'**induire les caractéristiques d'un groupe général (la population) à partir de celles d'un groupe particulier (l'échantillon)**, en fournissant une mesure de la certitude de la prédiction : la probabilité d'erreur.

La « fluctuation d'échantillonnage » désigne la variabilité des résultats provenant de la prise d'échantillon.

Inférence statistique (2)

Exemple : **estimation** de la moyenne et TLC

Plus *n* est grand, plus la dispersion autour de la moyenne (*i.e.* variance ou écart type) est petite, *i.e.* **moins il y a de fluctuation due à l'échantillonnage**.



Test statistique : définition

Un test, ou **test d'hypothèse**, est une procédure de décision entre deux hypothèses. Il s'agit d'une démarche consistant à **rejeter ou à ne pas rejeter** une hypothèse statistique, appelée hypothèse nulle, en fonction d'un échantillon de données.

H_0 : hypothèse nulle

H_1 : hypothèse alternative

Test statistique : objectifs

- **Inférence** : comparer des indicateurs mesurés sur un échantillon à ceux d'une distribution théorique (population)

Mais aussi…

- Comparer plusieurs groupes par des indicateurs mesurés sur des échantillons
- Prédiction (p. ex., régression)

Exemple : test Z (1)

- But : comparer deux distributions par leurs moyennes
  - Échantillon vs population (de paramètres μ_0 et σ)
- H_0 : les moyennes ne sont pas significativement différentes
- H_1 : les moyennes sont différentes

Les fluctuations d'échantillonnage font que la moyenne de l'échantillon ne sera jamais parfaitement égale à celle de la population.

La différence observée est-elle alors

- **uniquement dûe aux fluctuations d'échantillonnage ?**
- **ou bien significative ?**

Exemple : mesurer l'effet d'un médicament

Exemple : test Z (2)

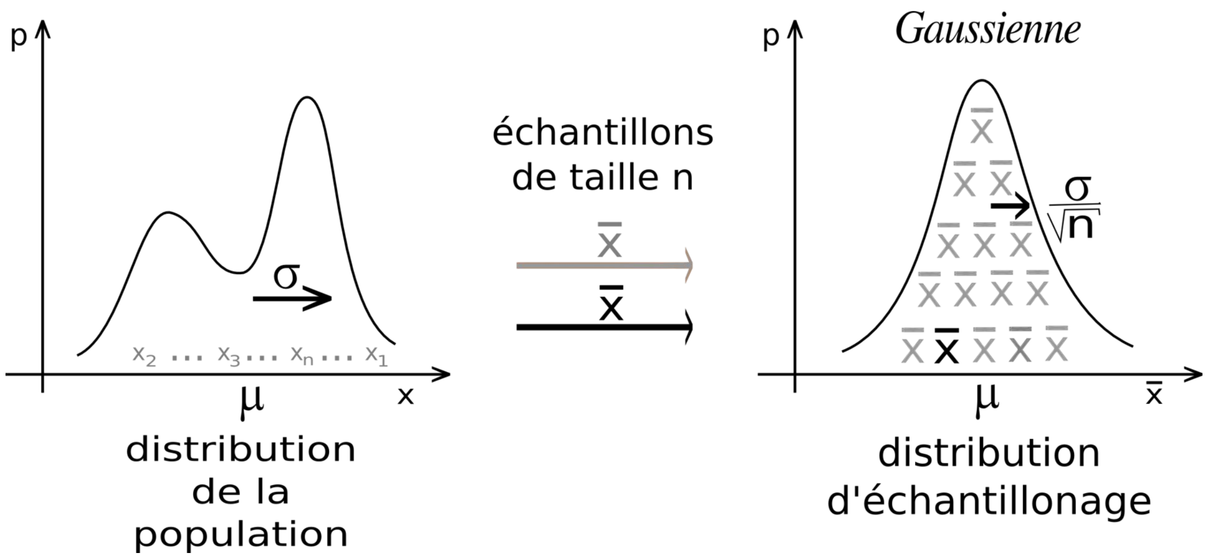
À partir des données, on calcule la **différence standardisée entre les deux moyennes** :

$$Z = \frac{(\bar{X} - \mu_0)}{\sigma}$$

Cette valeur est ce que l'on appelle la « statistique de test », ou encore « **variable de décision** ».

Exemple : test Z (3)

D'après le théorème de la limite centrale, les moyennes des échantillons suivent une loi normale, de moyenne μ égale à celle de la population.

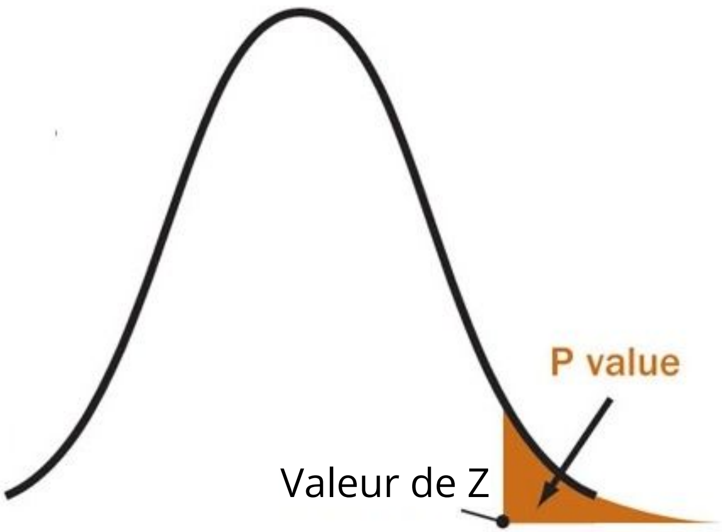


Il est ainsi possible de calculer la vraisemblance d'un échantillon (probabilité d'observer un échantillon extrait de cette population), à partir de sa moyenne.

Exemple : test Z (4)

Plus un échantillon s'éloigne de la moyenne théorique

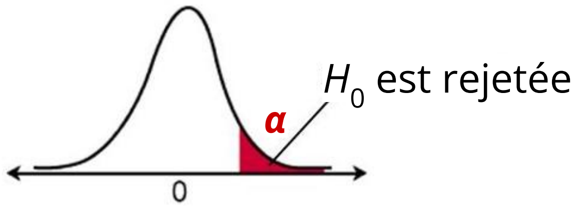
- plus Z est grand
- moins l'échantillon est vraisemblable (sous H_0)
- plus on peut rejeter H_0
- plus la différence est significative



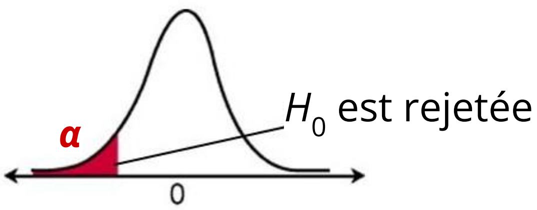
La probabilité d'observer un tel échantillon sous H0 est appelée **valeur p ou *p value***. Elle indique le niveau de significativité de la différence observée.

Exemple : test Z (5)

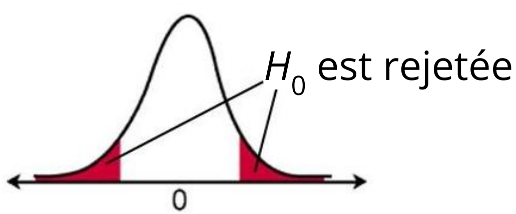
Pour pouvoir prendre la décision de rejeter ou non H_0, c. -à-d. pour déterminer si la différence est significative ou non
$\longrightarrow$ on choisit un **seuil de significativité α**
$\longrightarrow$ cette valeur définit ainsi une **région de rejet de H_0**



Test unilatéral droit



Test unilatéral gauche

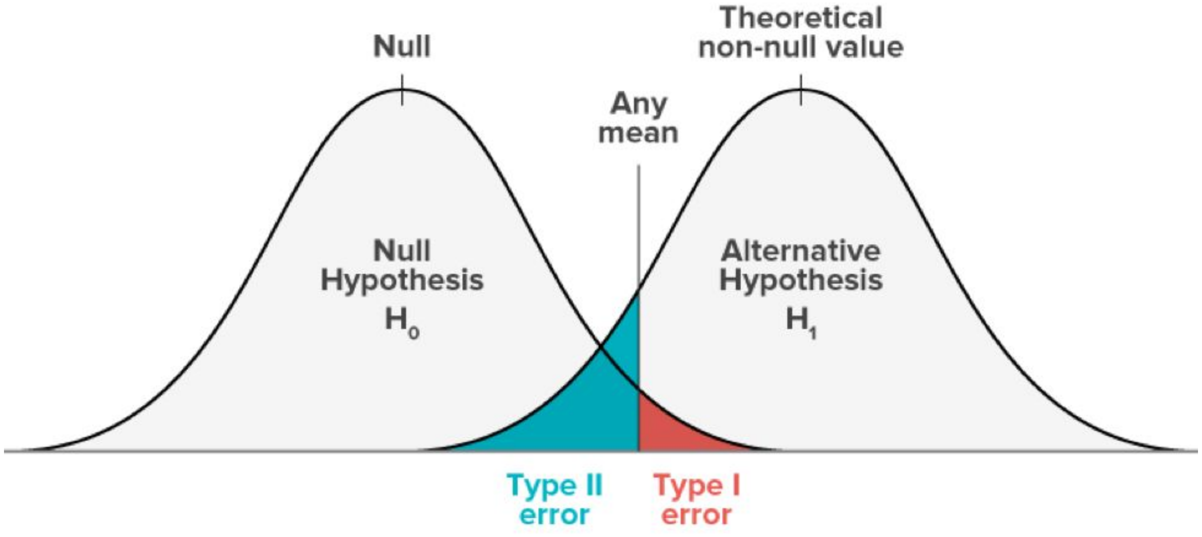


Test bilatéral

Erreurs de types I et II

On compare la *p-value* à une valeur seuil : la **probabilité α de se tromper en rejetant H_0 (erreur de type I ou risque de première espèce)**. Si $p < \alpha$, la différence entre les moyennes est significative (on rejette H_0).

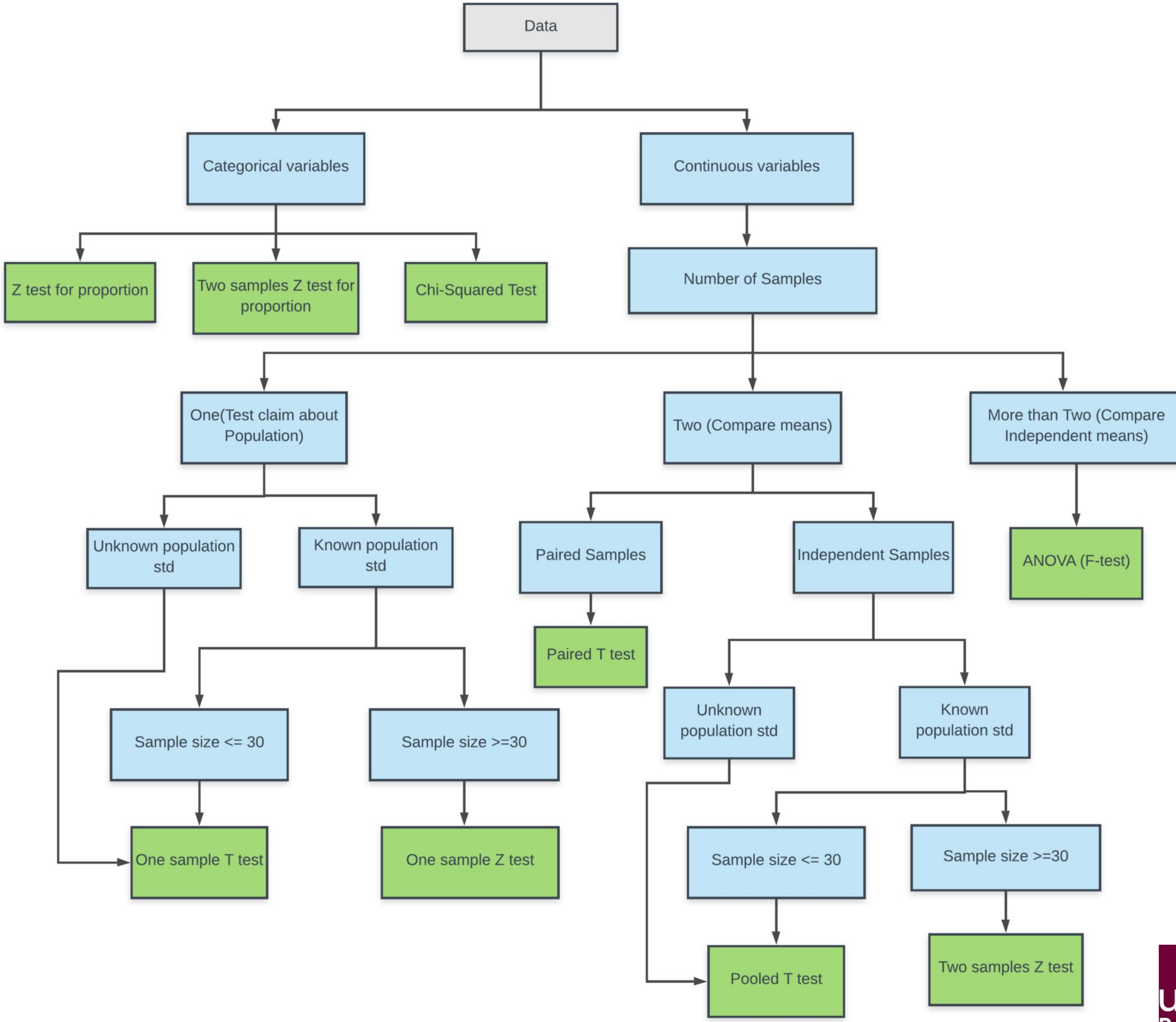
Il est possible de calculer une **probabilité β de se tromper en rejetant H_1**. Cela nécessite donc une seconde distribution représentant l'hypothèse alternative H_1. La moyenne de cette seconde distribution sera arbitrairement définie. Enfin, la **puissance du test sera calculée par $1-\beta$**.



Choix du test

Il se fait selon

- La question posée
  - Échantillon vs population
  - Comparaison d’échantillon
- Le type de variables
  - Quantitatives
  - Qualitatives
- Conditions d’applications
  - Normalité de la distribution
  - Taille de l’échantillon
  - Egalité des variances
  - …


Data
Categorical variables
Continuous variables
Z test for proportion
Two samples Z test for proportion
Chi-Squared Test
Number of Samples
One(Test claim about Population)
Two (Compare means)
More than Two (Compare Independent means)
Unknown population std
Known population std
Paired Samples
Independent Samples
ANOVA (F-test)
Paired T test
Unknown population std
Known population std
Sample size <= 30
Sample size >=30
One sample T test
One sample Z test
Sample size <= 30
Sample size >=30
Pooled T test
Two samples Z test

Test de Student (1)

- But : comparer deux distributions par leurs moyennes
  - Échantillon vs population (de paramètres μ_0)
- H_0 : les moyennes ne sont pas significativement différentes
- H_1 : les moyennes sont différentes

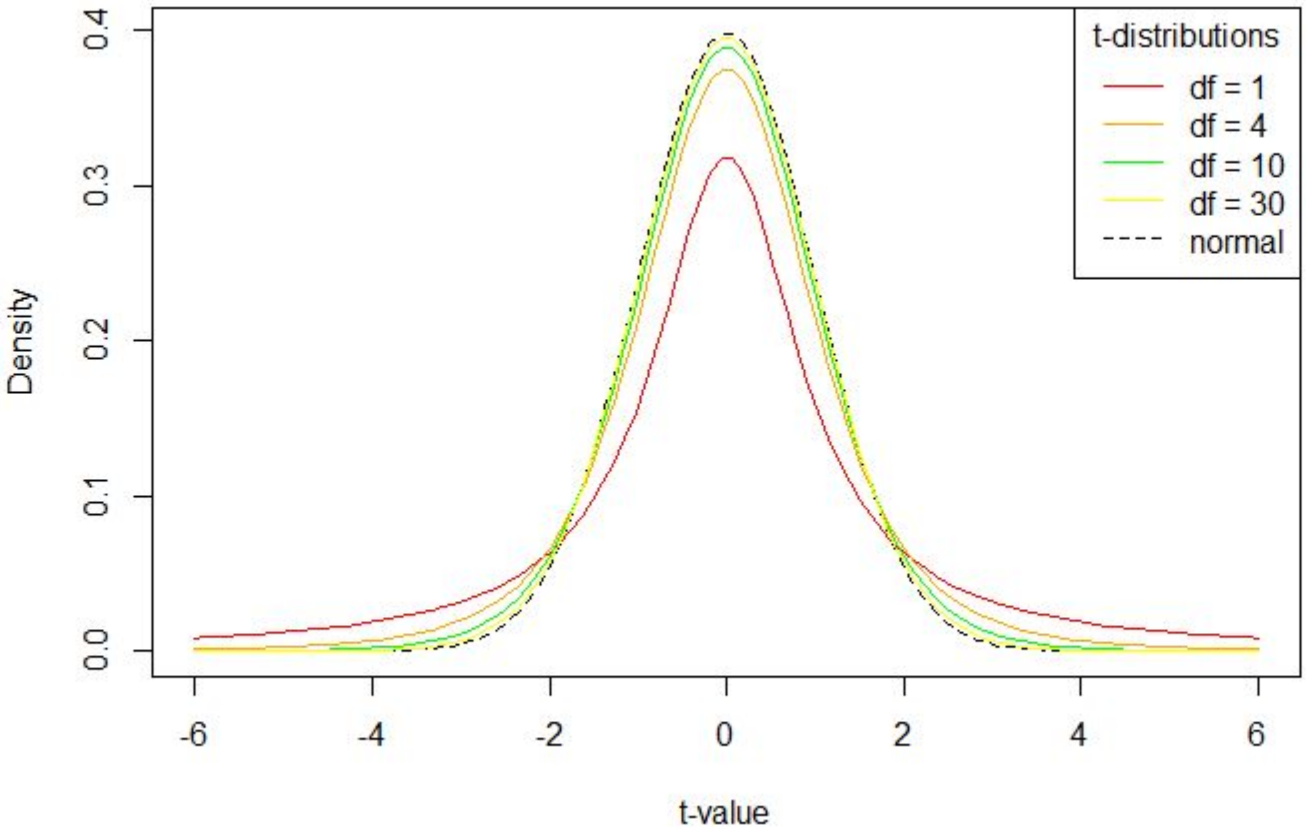
À partir des données, on calcule un **estimateur de *σ***, puis la statistique de test :

$$t = \frac{Z}{s} = \frac{\bar{X} - \mu}{\widehat{\sigma}/\sqrt{n}}$$

La variable de décision *t* suit une **loi de Student à n–1 degrés de liberté** sous H_0 (théorème de Cochran).

Test de Student (2)

df : *degrees of freedom*



Tests non-paramétriques

- Un test paramétrique est un test pour lequel on fait une hypothèse paramétrique sur la loi des données sous H_0 (loi normale, loi de Poisson…). Les hypothèses du test concernent alors les paramètres de cette loi.
- **Un test non-paramétrique est un test ne nécessitant pas d'hypothèse sur la loi des données**.