

# Statistiques et probabilités

## Cours n°6

Guillaume Postic

Université Paris-Saclay, Univ. Evry  
Département informatique

Master 1 MIAE - 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

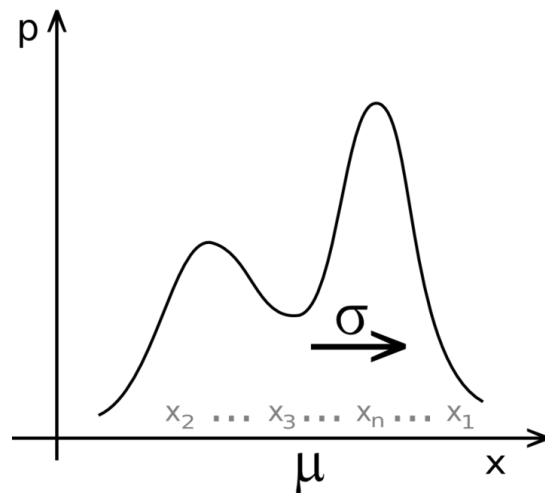
# Inférence statistique (1)

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.

# Inférence statistique (2)

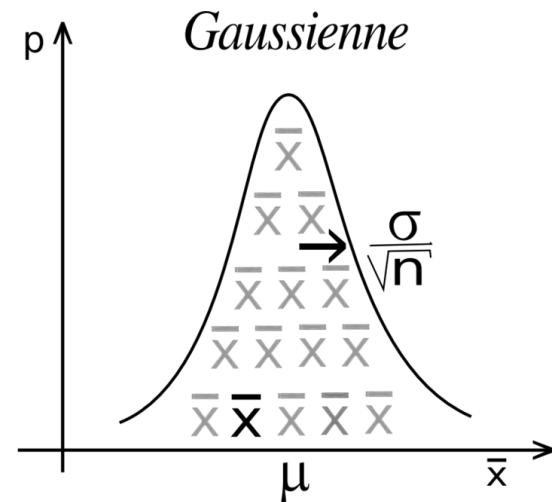
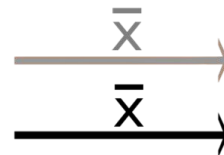
Exemple : **estimation** de la moyenne et TLC

Plus  $n$  est grand, plus la dispersion autour de la moyenne (*i.e.* variance ou écart type) est petite.



distribution  
de la  
population

échantillons  
de taille  $n$



distribution  
d'échantillonnage

# Test statistique (1)

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, complémentaire à  $H_0$

# Test statistique (2)

## Objectifs

- Inférence : comparer des indicateurs mesurés sur un échantillon à ceux d'une distribution théorique (population)
- Comparer plusieurs groupes par des indicateurs mesurés sur des échantillons
- Prédiction (p. ex., régression)

# Exemple : test Z

- But : comparer deux distributions par leurs moyennes
  - Échantillon vs population (de paramètres  $\mu_0$  et  $\sigma$ )
- $H_0$  : les moyennes ne sont pas significativement différentes
- $H_1$  : les moyennes sont différentes

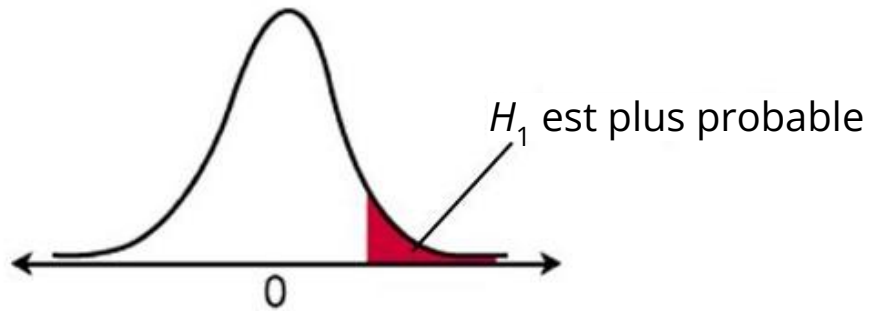
À partir des données, on calcule la « statistique de test », ou « variable de décision » :

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

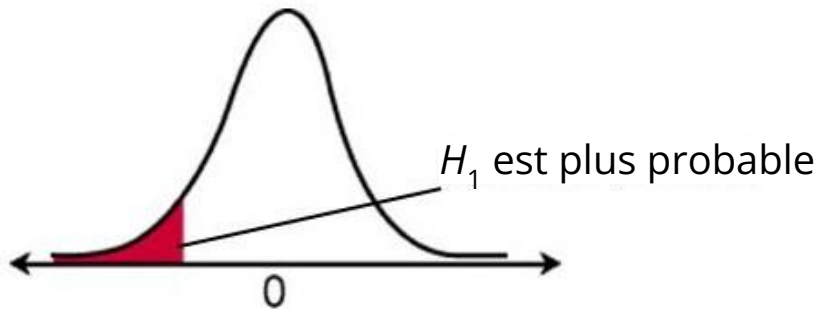
Cette valeur du test s'interprète comme une **différence standardisée entre les deux moyennes**.

On reporte cette valeur sur la distribution représentant l'hypothèse nulle ( $H_0$ ) : la distribution Z.

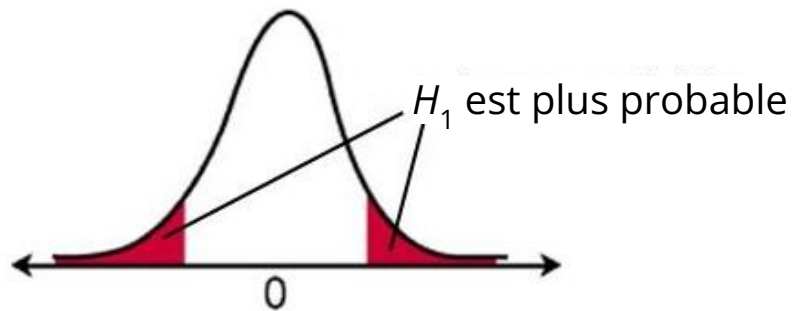
# Régions de rejet (et non-rejet)



Test unilatéral droit



Test unilatéral gauche



Test bilatéral



# Valeur p (*p-value*)

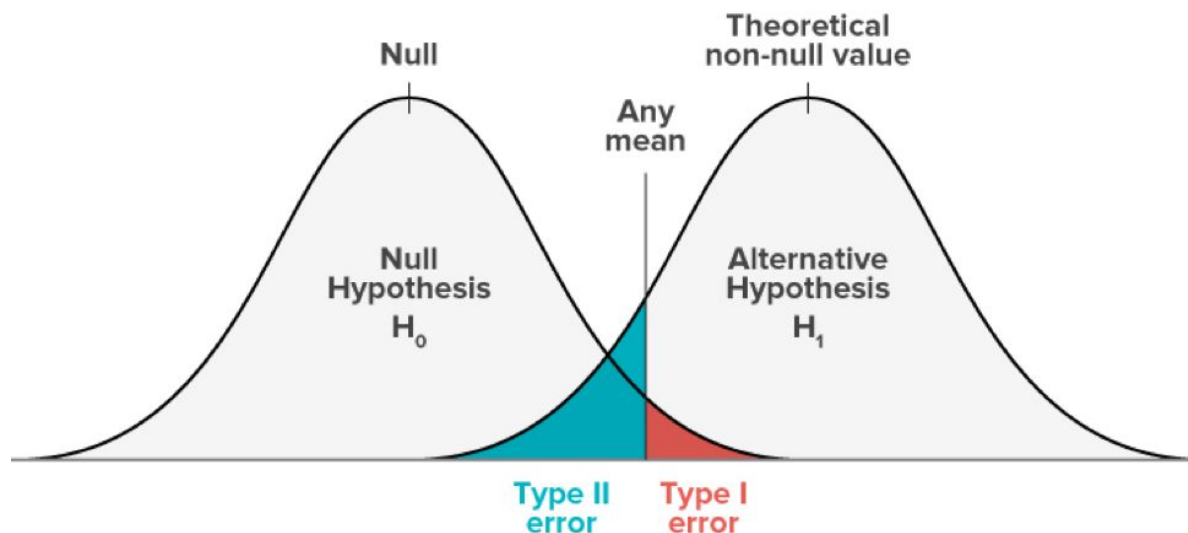
Plus on s'éloigne de 0, c'est à dire plus la différence entre les moyennes est grande, plus la **probabilité d'observer une différence au moins aussi grande** (quand on considère les moyennes comme étant égales, car on se place sous  $H_0$ ) est faible.

Cette probabilité est la *p-value*. **Plus elle est faible, plus on peut rejeter l'hypothèse  $H_0$** , donc plus la différence entre les moyennes est significative. Sa valeur peut être calculée en intégrant la fonction de distribution.

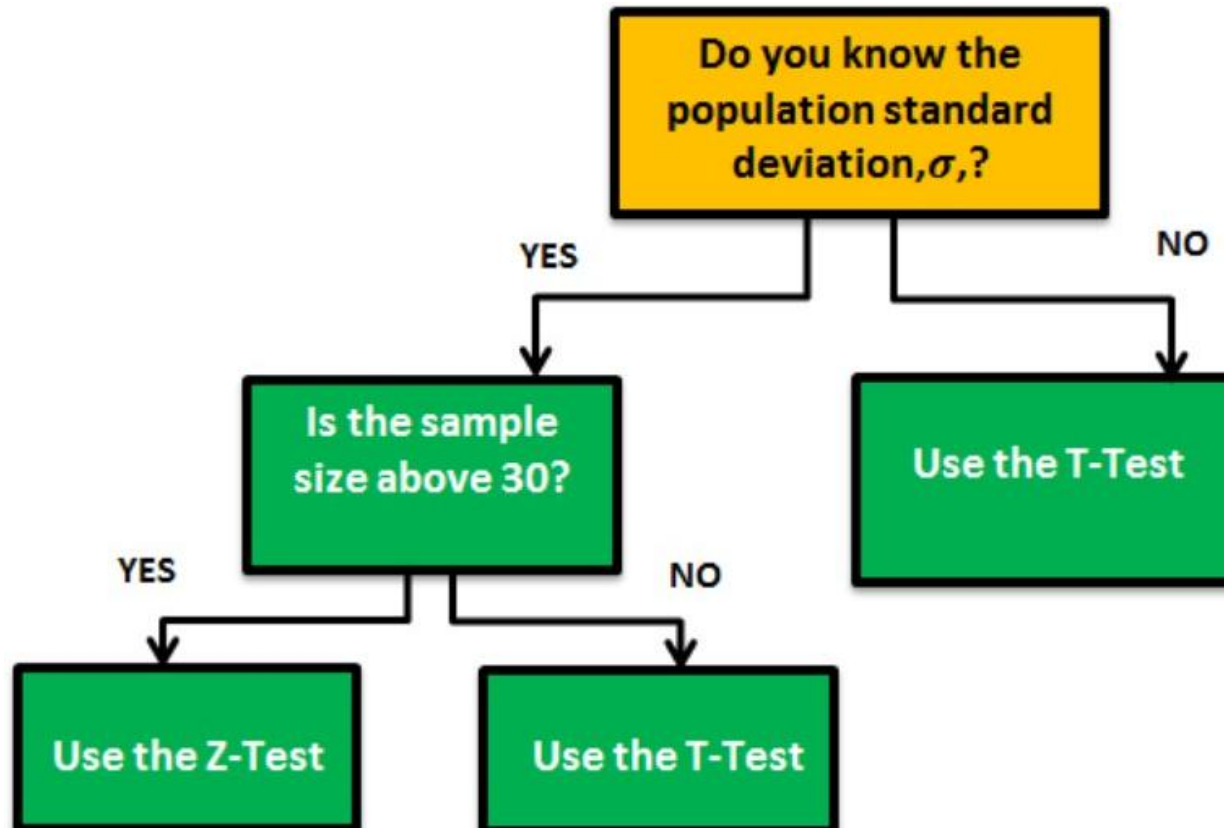
# Erreurs de types I et II

On compare la *p-value* à une valeur seuil : la **probabilité  $\alpha$  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  $\beta$  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  $1-\beta$** .



# Conditions d'application



# Test de Student (1)

- But : comparer deux distributions par leurs moyennes
  - Échantillon vs population (de paramètres  $\mu_0$ )
- $H_0$  : les moyennes ne sont pas significativement différentes
- $H_1$  : les moyennes sont différentes

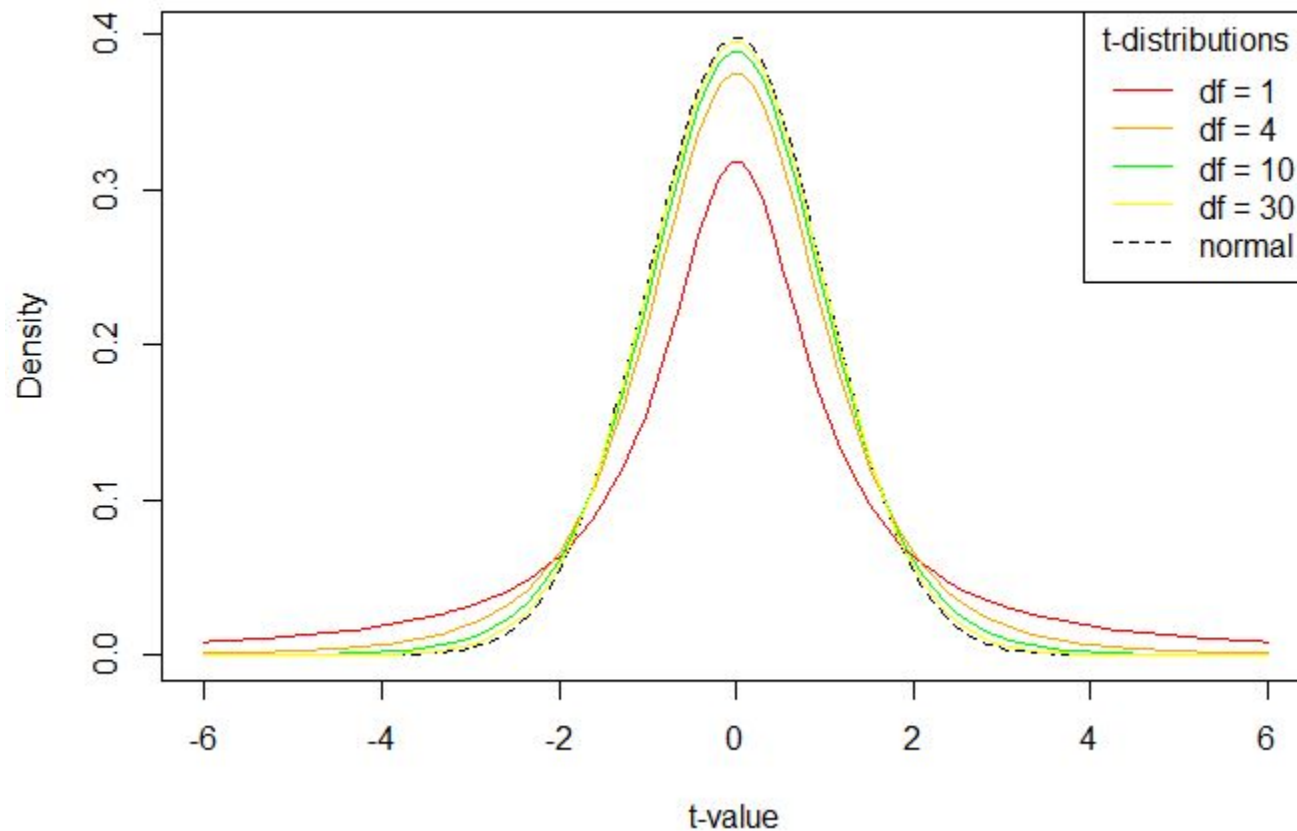
À partir des données, on calcule un **estimateur de  $\sigma$** , puis la statistique de test :

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

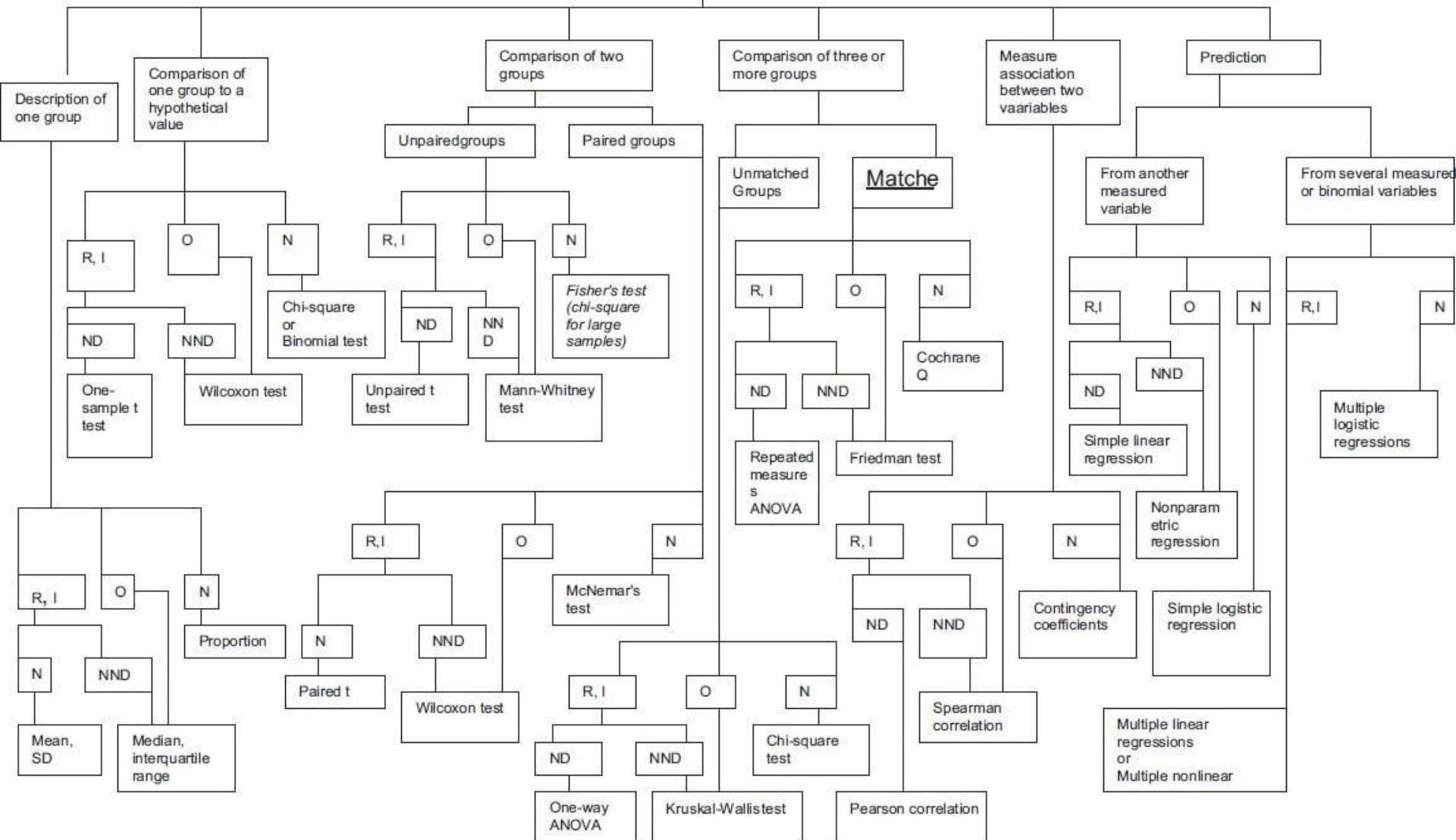
La variable de décision  $t$  suit une **loi de Student à  $n-1$  degrés de liberté** sous  $H_0$  (théorème de Cochran).

# Test de Student (2)

df : *degrees of freedom*



Goal



R, I = Ratio and Interval data    O = Ordinal data    N = Nominal data

N = Normal distribution    NND = Non normal distribution

# 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.