11 février 2025

**CFPPH**

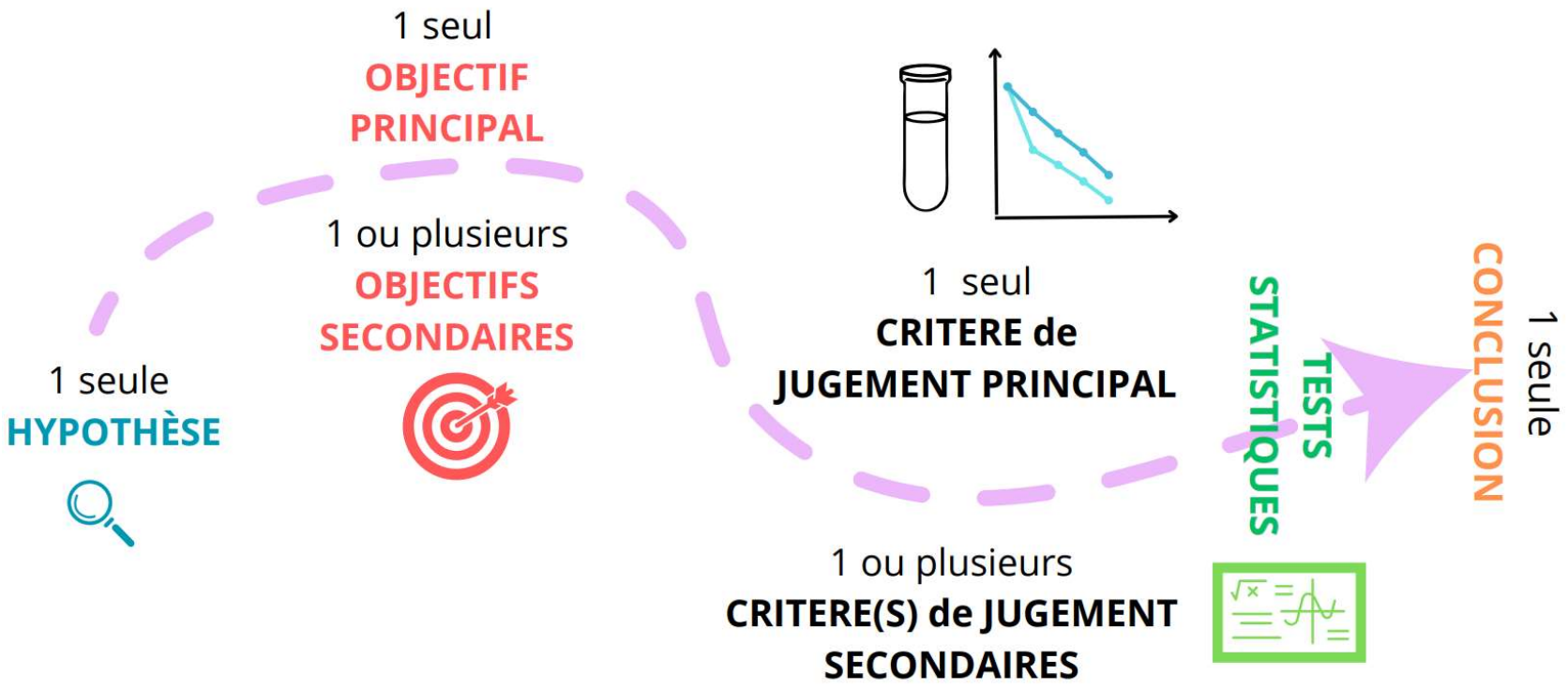
**UE Initiation à la démarche de recherche**

# Introduction aux méthodes biostatistiques de base

Géraldine Leguelinel-Blache

Pharmacien MCU-PH

# Les biostatistiques dans la démarche de recherche



# Objectifs des analyses biostatistiques

- **Lors de la rédaction du protocole**, on les utilise pour :
  - Déterminer le **nombre (minimal) de sujets nécessaires** (NSN) pour vérifier l'hypothèse de recherche
  - Il existe une corrélation entre le NSN et la **puissance** d'un test statistique

- **Après la collecte des données**, il faut les analyser pour :
  - Répondre aux **objectifs** de l'étude (principal et secondaires)
  - Vérifier l'**hypothèse de recherche**

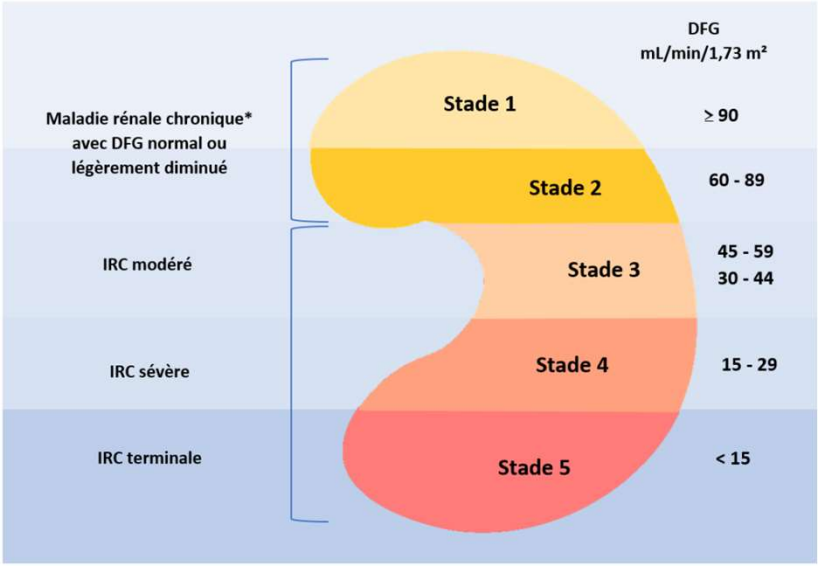
# 1-Identifier les variables

- **Variable aléatoire** = grandeur dont la valeur varie de manière incertaine en fonction du résultat (variation due au hasard)

# Identifier et caractériser les variables

- **Variable qualitative** avec plusieurs modalités :

- Sexe : femme, homme *(variable binaire)*
- Personne de plus de 65 ans : oui, non
- Diabète : oui, non, ne sait pas
- Insuffisance rénale : non, faible, modérée, sévère, terminale *(variable ordinale)*
- Groupe sanguin : O+, AB-, B+, … *(variable nominale)*



# Identifier et caractériser les variables

- **Variable quantitative** avec plusieurs modalités :

- Nombre de caries dépistées chez le dentiste : 1, 2, 3, 4, … (nombres entiers, *variable discrète*)

- Nombre de médicaments prescrits : 1, 2, 3, 4, …

- Taux de cholestérol total dans le sang : entre 0 et une valeur maximale extrêmement élevée (nombre de valeurs possibles théoriquement infinies, *variable continue*)

- Age : entre 0 et 122 ans (ex : 4 ans et 5 mois = 4,42 ans)

# Identifier et caractériser les variables

- Une étude de mortalité est réalisée dans les régions A et B.
- Les taux de mortalité par tranche d'âge sont étudiés.
- Il est dénombré 2 000 cas d'infarctus du myocarde (IDM) dans la région A et 2800 cas dans la région B.
- Le taux de cholestérol total des patients ayant fait un IDM était en moyenne de 2,9 g/L dans la région A et de 3,4 g/L dans la région B.

# Identifier les variables

- Une étude de mortalité est réalisée dans les régions A et B.
- Les taux de **mortalité** par **tranche d'âge** sont étudiés.
- Il est dénombré 2 000 cas d'**infarctus du myocarde** (IDM) dans la région A et 2800 cas dans la région B.
- Le **taux de cholestérol total** des patients ayant fait un IDM était en moyenne de 2,9 g/L dans la région A et de 3,4 g/L dans la région B.

# Caractériser les variables

- **Mortalité** : variable qualitative binaire (oui/non)
- **Tranches d'âge** : variable qualitative
- **Infarctus du myocarde** : variable qualitative binaire (oui/non)
- **Taux de cholestérol total** : variable quantitative continue

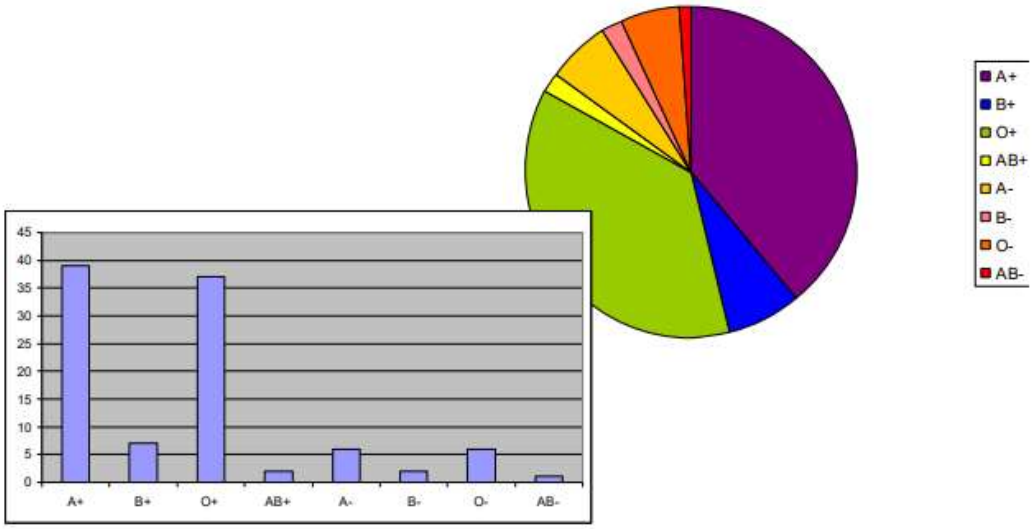
# 2-Décrire les variables

- **Variable qualitative** :

- Description de la répartition de l'effectif dans chaque modalité

- Soit en **nombre** soit en **pourcentage (%)** = sous-effectif / effectif total

| Groupe sanguin | Fréquence (%) |
|---|---|
| A+ | 39 |
| B+ | 7 |
| O+ | 37 |
| AB+ | 2 |
| A- | 6 |
| B- | 2 |
| O- | 6 |
| AB- | 1 |



# 2-Décrire les variables

- **Calcul d'un pourcentage** :

On effectue le dosage en mg/L d'un principe actif dans des poches de chimiothérapie. On réalise 10 mesures :

| **Valeurs** | **96** | **98** | **100** | **102** | **104** |
|---|---|---|---|---|---|
| **Effectif** | 1 | 2 | 4 | 2 | 1 |

# 2-Décrire les variables

- **Calcul d'un pourcentage** :

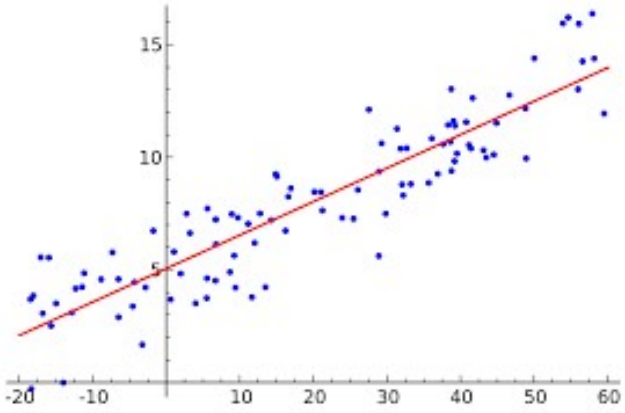
On effectue le dosage en mg/L d'un principe actif dans des poches de chimiothérapie. On réalise 10 mesures :

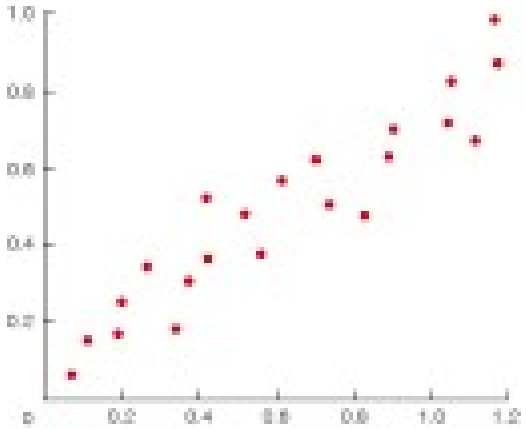
| **Valeurs** | **96** | **98** | **100** | **102** | **104** |
|---|---|---|---|---|---|
| **Effectif** | 1 | 2 | 4 | 2 | 1 |
| **%** | 1/10<br>= 0,1<br>ou 10% | 2/10<br>= 0,2<br>ou 20% | 4/10<br>= 0,4<br>ou 40% | 20% | 10% |

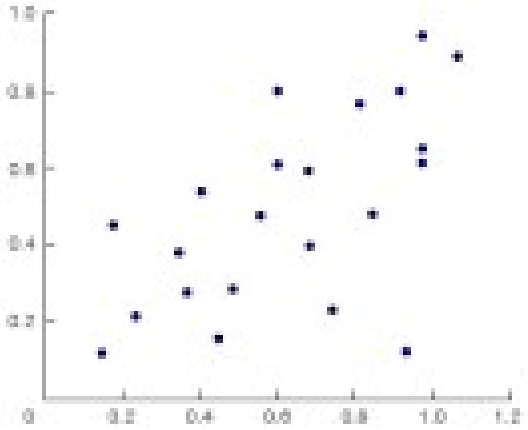
# 2-Décrire les variables

- **Variable quantitative** :

- Description de la répartition des valeurs prises par la variable (nuage de points)



# 2-Décrire les variables

- **Variable quantitative** :

- Description de la répartition des valeurs prises par la variable (nuage de points)

- Selon la **position** :

- **Moyenne :** somme des valeurs multipliées par les sous-effectifs et divisée par l'effectif total

- **Médiane :** valeur qui partage l'effectif total en 2 sous-effectifs égaux

- **Mode :** valeur la plus fréquente

# 2-Décrire les variables

- **Calcul de la moyenne** :

On effectue le dosage en mg/L d'un principe actif dans des poches de chimiothérapie. On réalise 10 mesures :

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **102** | **98** | **102** | **104** | **100** | **100** | **96** | **98** | **100** | **100** |

# 2-Décrire les variables

- **Calcul de la moyenne** :

On calcule la moyenne en additionnant les 10 mesures et en les divisant par l'effectif total (N=10) :

(102+98+102+104+100+100+96+98+100+100) / 10 = 100 mg/L

On trouve une moyenne de 100 mg/L.

***Attention à ne pas oublier l'unité (ici les mg/L)***

# 2-Décrire les variables

- **Calcul de la moyenne** :

On effectue le dosage en mg/L d'un principe actif dans des poches de chimiothérapie. On réalise 10 mesures :

| **Valeurs** | **96** | **98** | **100** | **102** | **104** |
|---|---|---|---|---|---|
| **Effectif** | 1 | 2 | 4 | 2 | 1 |

# 2-Décrire les variables

- **Calcul de la moyenne** :

On calcule la moyenne en multipliant chacune des 5 valeurs par leur sous-effectif et on divise la somme de ces 5 produits par l'effectif total (N=10) :

[(96x1)+(98x2)+(100x4)+(102x2)+(104x1)] / 10 = 100 mg/L

On trouve une moyenne de 100 mg/L.

# 2-Décrire les variables

- **Calcul de la médiane** :

On effectue le dosage en mg/L d'un principe actif dans des poches de chimiothérapie. On réalise 10 mesures :

| **Valeurs** | **96** | **98** | **100** | **102** | **104** |
|---|---|---|---|---|---|
| **Effectif** | 1 | 2 | 4 | 2 | 1 |

# 2-Décrire les variables

- **Calcul de la médiane** :
- On calcule la valeur de la moitié de l'effectif = 10/2 = 5
- On regarde en effectif cumulé où se situe cette valeur.

| Valeurs | 96 | 98 | 100 | 102 | 104 |
|---|---|---|---|---|---|
| Effectif | 1 | 2 | 4 | 2 | 1 |
| Effectif cumulé | 1 | 3 | 7 | 9 | 10 |

- La valeur de la variable correspondante (100 mg/L) correspond à la médiane.
- C'est la valeur qui partage l'effectif total en 2 sous-effectifs égaux.

# 2-Décrire les variables

- **Calcul du mode** :

On effectue le dosage en mg/L d'un principe actif dans des poches de chimiothérapie. On réalise 10 mesures :

| **Valeurs** | **96** | **98** | **100** | **102** | **104** |
|---|---|---|---|---|---|
| **Effectif** | 1 | 2 | 4 | 2 | 1 |

# 2-Décrire les variables

- **Calcul du mode** :

Le mode est la valeur la plus fréquente dans l'effectif total.
C'est celle qui a le sous-effectif le plus grand.

| Valeurs | 96 | 98 | 100 | 102 | 104 |
|---|---|---|---|---|---|
| Effectif | 1 | 2 | 4 | 2 | 1 |

# 2-Décrire les variables

- **Calcul du mode** :

Le mode est 100 mg/L.

| Valeurs | 96 | 98 | 100 | 102 | 104 |
|---|---|---|---|---|---|
| Effectif | 1 | 2 | 4 | 2 | 1 |

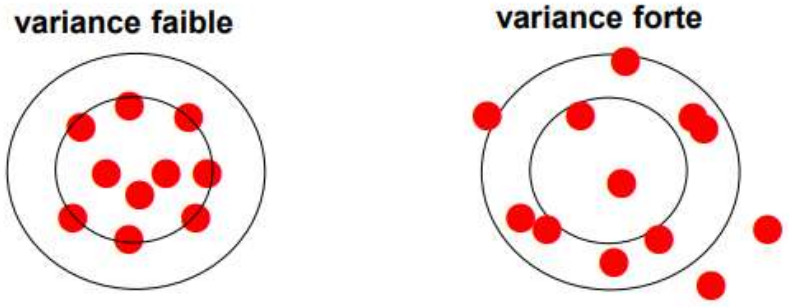
- *Attention il est possible qu'une variable ait plusieurs modes : 100 et 102 mg/L*

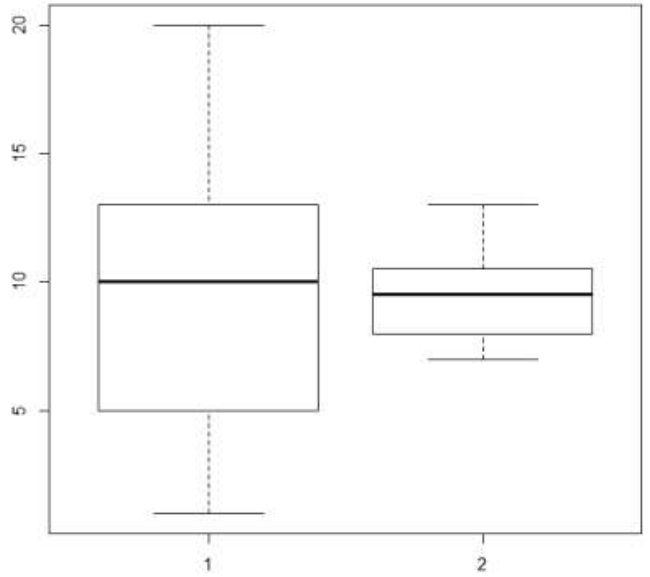
| Valeurs | 96 | 98 | 100 | 102 | 104 |
|---|---|---|---|---|---|
| Effectif | 1 | 2 | 4 | 4 | 1 |

# 2-Décrire les variables

- **Variable quantitative** :

- Description de la répartition des valeurs prises par la variable (boites à moustaches)



# 2-Décrire les variables

- **Variable quantitative** :

- Description de la répartition des valeurs prises par la variable (boites à moustaches)

- Selon la **dispersion** :

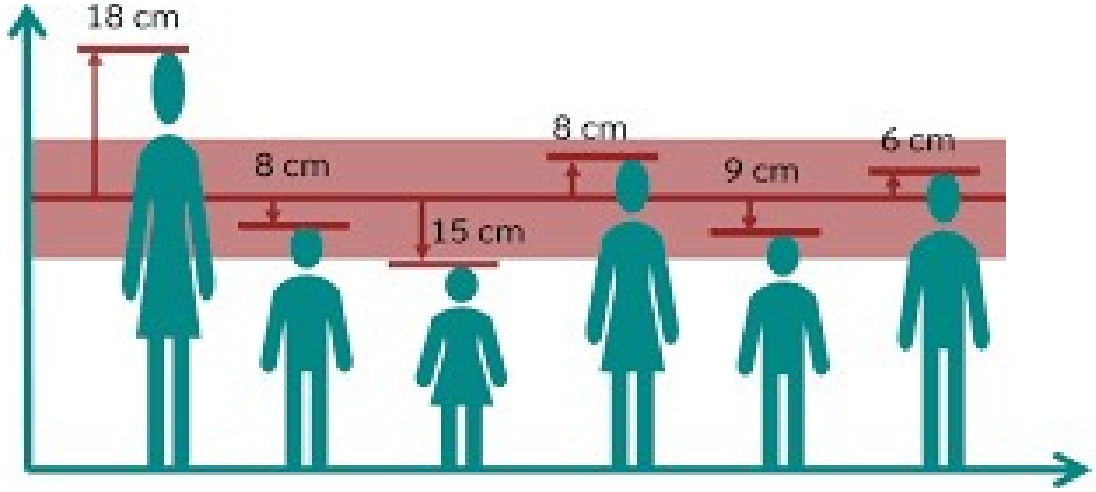
- **Ecart-type** **:** somme des écarts au carré entre les valeurs de la variable et la moyenne, divisée par l'effectif total (caractérise la dispersion d'une **moyenne**)

- **Intervalle interquartile** **:** intervalle compris entre le 1$^{er}$ quartile Q1 (1$^{er}$ quart si on partage en 4 parties égales l'effectif total) et le 3$^{ème}$ quartile (3$^{ème}$ quart) (caractérise la dispersion d'une **médiane**)

# 2-Décrire les variables

- **Calcul de l'écart-type** :

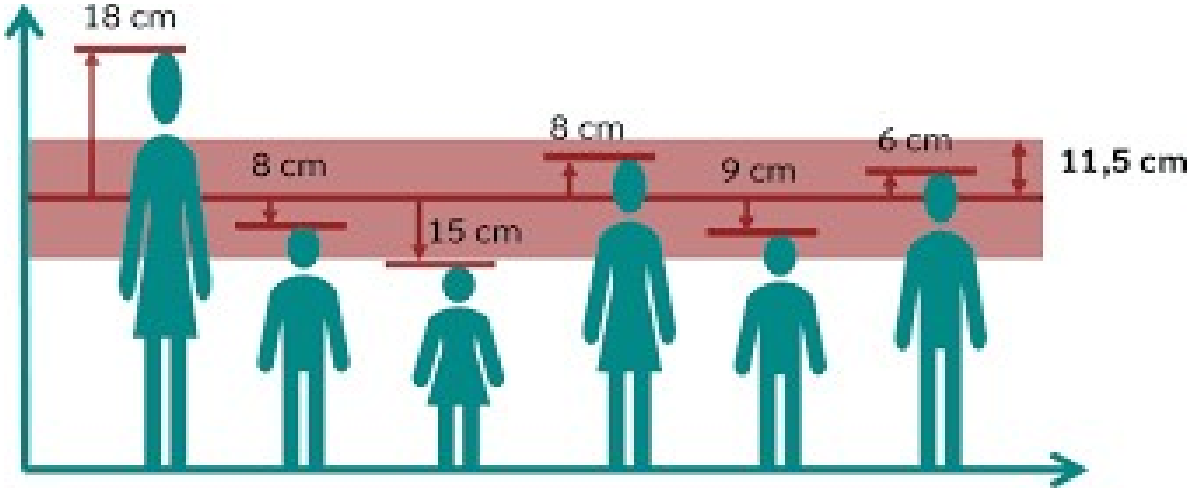
- somme des carrés des écarts entre les valeurs de la variable (ici la taille des individus) et la moyenne, divisée par l'effectif total (n=6)



# 2-Décrire les variables

- **Calcul de l'écart-type** :

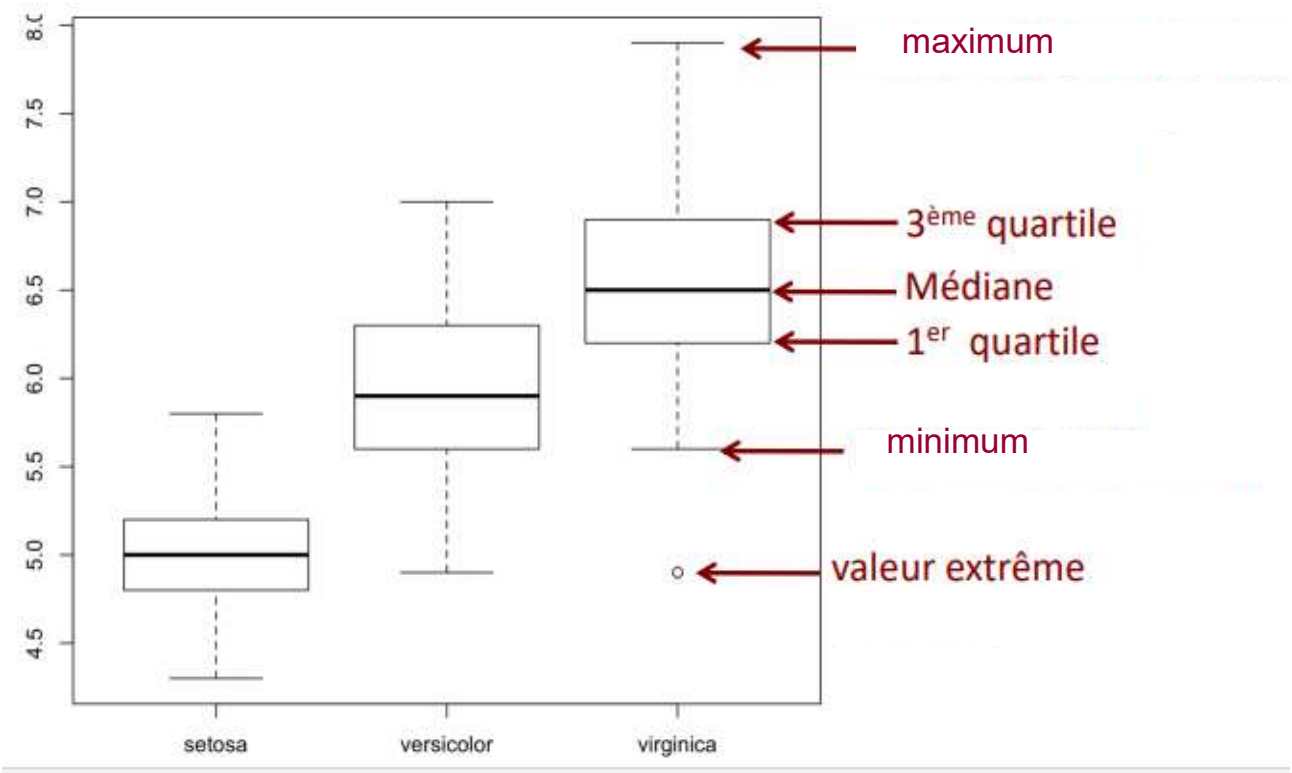
- $(18^2+8^2+15^2+8^2+9^2+6^2) / 6 = 11,5$ cm




**Écart-type**

# 2-Décrire les variables

- **Calcul de l'intervalle interquartile [Q1;Q3]** :



# 2-Décrire les variables

- **Calcul de l'intervalle interquartile [Q1;Q3]** :
- On calcule la valeur du quart de l'effectif = 10/4 = 2,5
- On regarde en effectif cumulé où se situe cette valeur.

| Valeurs | 96 | 98 | 100 | 102 | 104 |
|---|---|---|---|---|---|
| Effectif | 1 | 2 | 4 | 2 | 1 |
| Effectif cumulé | 1 | 3 | 7 | 9 | 10 |

- La valeur de la variable correspondante (98 mg/L) correspond à Q1.

# 2-Décrire les variables

- **Calcul de l'intervalle interquartile [Q1;Q3]** :
- On calcule la valeur des 3 quarts de l'effectif = 10x3/4 = 7,5
- On regarde en effectif cumulé où se situe cette valeur.

| Valeurs | 96 | 98 | 100 | 102 | 104 |
|---|---|---|---|---|---|
| Effectif | 1 | 2 | 4 | 2 | 1 |
| Effectif cumulé | 1 | 3 | 7 | 9 | 10 |

- La valeur de la variable correspondante (102 mg/L) correspond à Q3.
- L'intervalle interquartile est [98;102] mg/L

# 3-Comparer les variables

- Il est difficile d'étudier une variable au sein d'une population car l'effectif est souvent trop grand pour qu'on puisse faire une étude sur l'ensemble de la population.

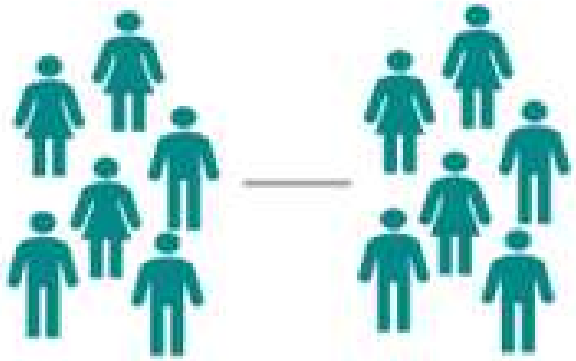
- On va donc extraire de la population un **échantillon** pour l'étudier et estimer ce qu'il se passe dans la population.
- Un échantillon est **représentatif** de la population lorsqu'il est tiré au sort.

# 3-Comparer les variables

- Pour étudier les variables, plusieurs designs d'étude sont possibles.
- On peut comparer :
  - Des individus différents issus de 2 **échantillons indépendants**
  - Les mêmes individus issus de 2 **échantillons appariés**
- Les tests statistiques utilisés vont dépendre de la nature des échantillons.

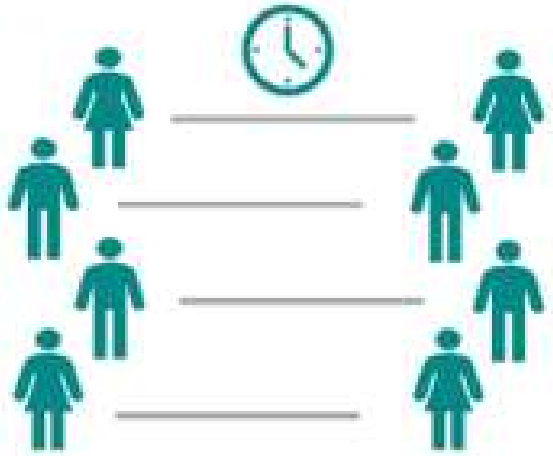
# 3-Comparer les variables

**Echantillons indépendants**



Y a-t-il une différence entre deux groupes ?

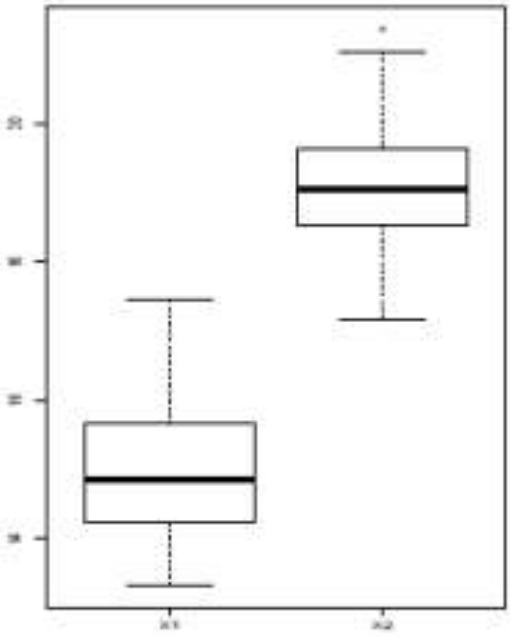
**Echantillons appariés**



Y a-t-il une différence au sein d'un groupe entre deux moments dans le temps ?

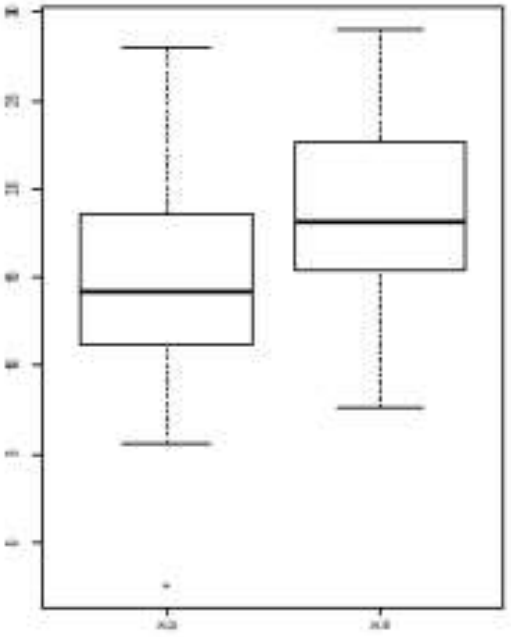
# 3-Comparer les variables

**Différentes**



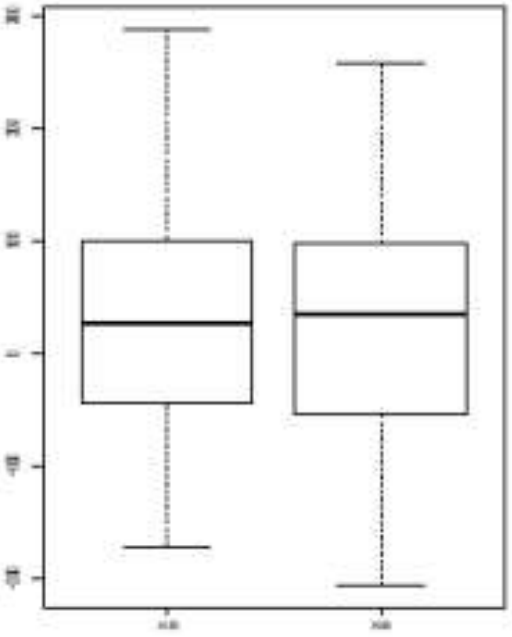
✔

**On ne sait pas juger sans test**



?

**Comparables**



✗

# 3-Comparer les variables

- Les tests statistiques vont nous permettre de savoir si les valeurs mesurées sont comparables ou différentes.

- On va formuler 2 hypothèses :

  - **H0** : les moyennes sont **comparables** (variabilité due au hasard)
  - **H1** : (variabilité expliquée par un facteur d'intérêt)
    - les moyennes sont **différentes** = ***test bilatéral***
    - une moyenne est **inférieure ou supérieure** à la moyenne de référence = ***test unilatéral d'infériorité ou de supériorité***

# 3-Comparer les variables

- En faisant ce test statistique, on risque se tromper.
- Il existe un **risque d'erreur α** (alpha) de rejeter à tort l'hypothèse H0.
- C'est le risque de **faux positif**.
- Il est souvent fixé à **5%** ou 0,05 mais il peut prendre n'importe quelle valeur.
- C'est l'expérimentateur qui va le choisir.

# 3-Comparer les variables

- On va identifier le test à utiliser en fonction de la **nature des variables** (qualitative ou quantitative) que l'on compare et de la **nature des échantillons** comparés (appariés ou indépendants).

- Les tests statistiques sont réalisés à l'aide de **logiciels informatiques**.
- Nous utiliserons le logiciel **BiostaTGV®** disponible gratuitement en ligne :

https://biostatgv.sentiweb.fr/?module=tests

# 3-Comparer les variables

Type de test à mettre en évidence

Type de test : -Tous-

| Facteur d'étude | | Variable de réponse : Qualitative nominale (2 groupes) | Qualitative nominale (plus de 2 groupes) | Qualitative ordinale | Quantitative |
|---|---|---|---|---|---|
| Qualitatif (deux groupes) | Indépendants | *Z de comparaison de proportions.** | Chi² (χ2.) | *Test de Cochran-Armitage** | Test de Mann-Whitney. |
| | | Chi² (χ2.) | | | t de Student. |
| | | Test exact de Fisher. | | | *Test de Welch.** |
| | Appariés | Test de McNemar. | *Q de Cochran.** | *Tests des signes.** | t de Student pour données appariées. |
| | | Test exact de Fisher. | | Tests des rangs signés de Wilcoxon. | Tests des rangs signés de Wilcoxon. |
| Qualitatif (plus de deux groupes) | Indépendants | Chi² (χ2.) | Chi² (χ2.) | Test de Kruskal-Wallis. (ordinal) | Analyse de la variance. |
| | | | | | Test de Kruskal-Wallis. (échelle quanti) |
| | Appariés | *Q de Cochran.** | *Q de Cochran.** | Test de Friedman. | Test de Friedman. |
| Quantitatif | | *Régression logistique** | *Régression logistique multinomiale** | Corrélation de Spearman. | Corrélation de Pearson. |
| | | | | Tau de Kendall. | *Régression linéaire.** |

* : La réalisation de ces tests n'est actuellement pas disponible sur biostaTGV

# 3-Comparer les variables

- L'interprétation des résultats du test statistique se fera en calculant la p-value.

- Si la **p-value < 0,05**, alors H0 est fausse avec un risque α de 5% **: il existe une différence, une supériorité ou une infériorité** (en fonction de H1) significative entre les échantillons.
- Si la **p-value ≥ 0,05**, alors H0 est vraie avec un risque α de 5% : il n'existe pas de différence, de supériorité ou d'infériorité (en fonction de H1) significative entre les échantillons ; les **échantillons sont statistiquement comparables**.

# Exercice

- Une étude de mortalité est réalisée dans les régions A et B.
- Les taux de mortalité par tranche d'âge sont étudiés.
- Il est dénombré 2 000 cas d'infarctus du myocarde (IDM) dans la région A et 2800 cas dans la région B.
- Le taux de cholestérol total des patients ayant fait un IDM était en moyenne de 2,9 g/L dans la région A et de 3,4 g/L dans la région B.

# Exercice

## 1) Quelle est la nature des échantillons ?

- Une étude de mortalité est réalisée dans les régions A et B.
- Les taux de mortalité par tranche d'âge sont étudiés.
- Il est dénombré 2 000 cas d'infarctus du myocarde (IDM) dans la région A et 2800 cas dans la région B.
- Le taux de cholestérol total des patients ayant fait un IDM était en moyenne de 2,9 g/L dans la région A et de 3,4 g/L dans la région B.

# Exercice

## 1) Quelle est la nature des échantillons ?

- Une étude de mortalité est réalisée **dans les régions A et B**.
- Les taux de mortalité par tranche d'âge sont étudiés.
- Il est dénombré 2 000 cas d'infarctus du myocarde (IDM) dans la région A et 2800 cas dans la région B.
- Le taux de cholestérol total des patients ayant fait un IDM était en moyenne de 2,9 g/L dans la région A et de 3,4 g/L dans la région B.

# Exercice

## 1) Quelle est la nature des échantillons ?

- *Une étude de mortalité est réalisée dans les régions A et B.*
- Les **échantillons sont indépendants** car on compare 2 régions différentes.

# Exercice

**1) Quel est l'indicateur qui permettra de caractériser la dispersion du taux moyen de cholestérol total ?**

- Une étude de mortalité est réalisée dans les régions A et B.
- Les taux de mortalité par tranche d'âge sont étudiés.
- Il est dénombré 2 000 cas d'infarctus du myocarde (IDM) dans la région A et 2800 cas dans la région B.
- Le taux de cholestérol total des patients ayant fait un IDM était en moyenne de 2,9 g/L dans la région A et de 3,4 g/L dans la région B.

# Exercice

**2) Quel est l'indicateur qui permettra de caractériser la dispersion du taux moyen de cholestérol total ?**

- Le taux moyen de cholestérol total est une moyenne. Sa dispersion sera caractérisée par un **écart-type**.
- Si le taux médian de cholestérol total avait été calculé, sa dispersion aurait été caractérisée par un intervalle interquartile [Q1;Q3].

# Exercice

**3) Quelles sont les 2 hypothèses qui seront formulées pour comparer les taux de cholestérol total dans les 2 régions ?**

- Une étude de mortalité est réalisée dans les régions A et B.
- Les taux de mortalité par tranche d'âge sont étudiés.
- Il est dénombré 2 000 cas d'infarctus du myocarde (IDM) dans la région A et 2800 cas dans la région B.
- Le taux de cholestérol total des patients ayant fait un IDM était en moyenne de 2,9 g/L dans la région A et de 3,4 g/L dans la région B.

# Exercice

**3) Quelles sont les 2 hypothèses qui seront formulées pour comparer les taux de cholestérol total dans les 2 régions ?**

- **H0** : les taux moyens de cholestérol total sont **comparables**.
- **H1** : les taux moyens de cholestérol total sont **différents**.

Il n'y a aucune raison de penser qu'une région puisse favoriser un taux de cholestérol supérieur ou inférieur à une autre région. Si un doute existait (gradient Nord-Sud), il serait précisé dans le texte.

# Exercice

**4) Quel va être le test statistique utilisé pour comparer les taux de cholestérol total dans les 2 régions A et B ?**

- Une étude de mortalité est réalisée dans les régions A et B.
- Les taux de mortalité par tranche d'âge sont étudiés.
- Il est dénombré 2 000 cas d'infarctus du myocarde (IDM) dans la région A et 2800 cas dans la région B.
- Le taux de cholestérol total des patients ayant fait un IDM était en moyenne de 2,9 g/L dans la région A et de 3,4 g/L dans la région B.

# Exercice

**4) Quel va être le test statistique utilisé pour comparer les taux de cholestérol total dans les 2 régions A et B ?**

- On compare **2 variables quantitatives** dans **2 échantillons indépendants**.

# BiostaTGV

Taux de cholestérol

Régions A et B

| Type de test à mettre en évidence : Type de test = -Tous- | | | Variable de réponse : Qualitative nominale (2 groupes) | Qualitative nominale (plus de 2 groupes) | Qualitative ordinale | Quantitative |
|---|---|---|---|---|---|---|
| Facteur d'étude | Qualitatif (deux groupes) | Indépendants | *Z de comparaison de proportions.** | Chi² (χ2.) | *Test de Cochran-Armitage** | Test de Mann-Whitney. |
| | | | Chi² (χ2.) | | | t de Student. |
| | | | Test exact de Fisher. | | | *Test de Welch.** |
| | | Appariés | Test de McNemar. | *Q de Cochran.** | *Tests des signes.** | t de Student pour données appariées. |
| | | | Test exact de Fisher. | | Tests des rangs signés de Wilcoxon. | Tests des rangs signés de Wilcoxon. |
| | Qualitatif (plus de deux groupes) | Indépendants | Chi² (χ2.) | Chi² (χ2.) | Test de Kruskal-Wallis. (ordinal) | Analyse de la variance. |
| | | | | | | Test de Kruskal-Wallis. (échelle quanti) |
| | | Appariés | *Q de Cochran.** | *Q de Cochran.** | Test de Friedman. | Test de Friedman. |
| | Quantitatif | | *Régression logistique** | *Régression logistique multinomiale** | Corrélation de Spearman. | Corrélation de Pearson. |
| | | | | | Tau de Kendall. | *Régression linéaire.** |

* : La réalisation de ces tests n'est actuellement pas disponible sur biostaTGV

# Exercice

**4) Quel va être le test statistique utilisé pour comparer les taux de cholestérol total dans les 2 régions A et B ?**

- On compare **2 variables quantitatives** dans **2 échantillons indépendants**.
- D'après le tableau des tests, on choisira :
  - **Test de Mann-Whitney**
  - **Test T de Student**
  - **Test de Welch**