REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE MOHAMED BOUDIAF - M'SILA

faculté des sciences departement de science de

la nature et de la vie

Cours
Biostatistique

■ Réalisé par:
- Dr.ASLOUM Abdelmadjid Yagoub

Introduction

Tout travail scientifique, qu'il soit expérimental ou clinique, repose sur un protocole bien défini comportant quatre éléments fondamentaux. Ces éléments constituent la base de la conception de l'étude, assurant sa reproductibilité par d'autres chercheurs et renforçant la crédibilité et la comparabilité des résultats obtenus. Ce protocole répond à quatre questions essentielles :

1. **Sur quel sujet a-t-on travaillé ?**
   Il s'agit de définir clairement le matériel d'étude, qu'il s'agisse d'animaux de laboratoire, de sujets humains ou d'autres échantillons analysés.
2. **Quelle méthode a été employée ?**
   Ce point décrit la manière dont les données ont été collectées, que ce soit de manière prospective (selon un plan établi à l'avance) ou rétrospective.
3. **Quel est l'objectif de l'évaluation ?**
   Il est crucial de préciser ce que les chercheurs cherchent à évaluer, comme un examen biologique avec ses valeurs normales, l'efficacité d'un traitement, ou l'impact d'un facteur de risque.
4. **Quels critères de jugement ont été utilisés ?**
   Les critères peuvent inclure la présence ou l'absence de maladie, l'efficacité et la toxicité d'un traitement, le taux de récidive, ou la survie. Les méthodes statistiques employées pour analyser ces critères sont également précisées ici.

Ces éléments essentiels permettent de structurer l'étude de manière rigoureuse et de produire des données fiables, prêtes à être analysées et comparées. Dans ce cadre, les études scientifiques peuvent être classées en deux types principaux : les études non-expérimentales et les études expérimentales.

Les **études non-expérimentales** n'impliquent pas d'interventions directes sur le sujet étudié. Elles se basent sur l'observation des sujets dans leur état naturel, comme dans les études de type enquête, où l'on observe des individus fumeurs et non-fumeurs sur une longue période pour analyser la relation entre le cancer du poumon et le tabac. Ce type d'étude permet d'identifier des corrélations ou des associations sans manipulation des variables.

En revanche, les **études expérimentales** impliquent une intervention ou une manipulation directe. Par exemple, pour tester l'efficacité d'un médicament, les chercheurs divisent les sujets en deux groupes : un groupe reçoit le médicament tandis que l'autre reçoit un

placebo. Cette méthode permet d'évaluer l'effet du médicament en comparant les résultats entre les deux groupes.

En somme, que l'étude soit expérimentale ou non, elle converge vers la production de données analysables, permettant ainsi de répondre aux questions de recherche spécifiques. Cette classification montre comment le choix de la méthodologie est déterminant pour la qualité et la pertinence des résultats obtenus, soulignant l'importance de respecter un protocole bien structuré pour garantir la validité scientifique et la comparabilité des travaux.

Terminologie en Biostatistiques

1. **Modélisation statistique :**

Cette étape est essentielle pour toute étude, qu'elle soit expérimentale ou non, et doit être suivie d'une analyse statistique. La modélisation statistique est le processus d'application de méthodes statistiques pour interpréter et prédire des phénomènes en se basant sur des données collectées.

2. **Statistique descriptive** :

Méthodes utilisées pour résumer et organiser les données, Ces méthodes incluent les tableaux de fréquence, les graphiques, et les mesures de tendance centrale (comme la moyenne)etc.

3. **Statistique inférentielle** :

Méthodes qui permettent de tirer des conclusions sur une population à partir d'un échantillon. Elle repose sur la théorie des probabilités pour estimer les caractéristiques de la population et tester des hypothèses (par exemple, les sondages).

4. **Statistique prédictive :**

La statistique prédictive est une branche de la statistique qui utilise les données historiques et actuelles pour construire des modèles qui permettent de prédire des événements ou des résultats futurs.

5. **Statistique univariée** :

Étude portant sur une seule variable, qu'elle soit quantitative ou qualitative. La statistique univariée fait partie de la statistique descriptive et se concentre sur la distribution et les caractéristiques d'une seule variable.

6. **Population statistique** :

Ensemble complet d'individus ou d'objets sur lequel porte une étude statistique. C'est l'ensemble des éléments observables.

7. **Échantillon** :

Sous-ensemble de la population sélectionnée pour l'étude. L'échantillon est utilisé pour tirer des conclusions sur la population dans son ensemble.

8. **Individu** :

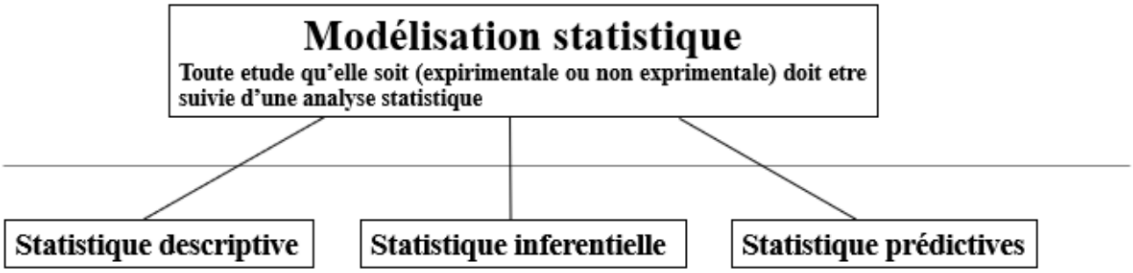
Élément unique de la population statistique étudiée. Chaque individu possède des caractéristiques ou des variables qui seront mesurées.

9. **Variable ou caractère statistique** :

Propriété ou caractéristique mesurable d'un individu de la population étudiée. Les valeurs possibles d'une variable sont appelées **modalités**.

10. La statistique bivariée:

est une branche de la statistique descriptive qui analyse simultanément deux variables afin de comprendre la relation entre elles. L'objectif principal est d'explorer et de quantifier la manière dont une variable peut influencer ou être associée à une autre.



Statistiques descriptives:
Collecte, l'organisation et le traitement des données sur un echantillion.

Démarche à suivre:

1-Collecte des données sondage ou exprience (échantillonage aléatoire)

2-Tableaux individus variable.

3-Methodes statistique tableaux,graphes,parametres.

Vocabulaire statistique

Population statistique: est l'ensemble sur le quel on effectue des observations.

Echantillon: une partie de la population statistique.

Individu: les individus sont les éléments de la population statistique étudiée.

Types de Variables

- **Variable quantitative** :

  Variable exprimée par des nombres et représentant une quantité mesurable.

  - **Variable quantitative continue** :

    Variable qui peut prendre une infinité de valeurs dans un intervalle donné. Exemples : taille, poids.

  - **Variable quantitative discrète** :

    Variable qui prend un nombre fini ou dénombrable de valeurs. Exemples : nombre d'enfants dans une famille, nombre de fleurs sur une plante.

- **Variable qualitative** :
  - Variable qui ne représente pas une quantité numérique mais des catégories ou des qualités.
  - **Variable qualitative nominale** :

    Variable dont les modalités sont des catégories sans ordre particulier. Exemples : couleur des yeux, lieu de naissance.

  - **Variable qualitative ordinale** :

    Variable dont les modalités peuvent être ordonnées de manière logique. Exemples : qualité d'un produit (bon, moyen, mauvais), niveau de satisfaction (faible, moyen, élevé).

Chapitre I

Biostatistique univariée

I. Variables qualitatives

Définition : Une variable qualitative est une variable qui prend des valeurs non numériques et décrit une qualité, une catégorie ou un attribut.

- Exemples :
  - **Genre** (Homme, Femme)
  - **Groupe sanguin** (A, B, AB, O)
  - **Présence ou absence** d'une maladie (Oui, Non)

I.1. Types de variables qualitatives

I.1.1. Qualitative nominale

Les catégories n'ont pas d'ordre ou de hiérarchie.

Exemples : Couleurs des yeux (Bleu, Marron, Vert)

I.1.2. Qualitative ordinale

Les catégories peuvent être classées selon un ordre logique.

Exemples : Degré de douleur (Léger, Modéré, Sévère)

I.2. Mesures descriptives pour une variable qualitative

I.2.1. Effectif et fréquence

- **Effectif** n_i: Nombre d'observations dans chaque catégorie.
  - Exemple : Dans une étude sur le groupe sanguin de 100 personnes :
    - Groupe A : 40 personnes
    - Groupe B : 30 personnes
    - Groupe O : 20 personnes
    - Groupe AB : 10 personnes
- **fréquence relative** (*fi*) : Proportion d'observations dans chaque catégorie

$$f_i = \frac{n_i}{N}$$

- **Proportion en pourcentage (*P%*) :** Pour calculer le pourcentage, il faut multiplier la fréquence des observations dans chaque catégorie par 100.
    - Exemple :
        - Groupe A :

$$f_A = \frac{40}{100} = 0.4$$

$$P\%_A = \frac{40}{100} \cdot 100$$

- **Distribution cumulée (surtout pour les variables qualitatives ordinales)**

C'est la somme des fréquences (absolues ou relatives) pour les catégories jusqu'à une certaine valeur.

- Exemple : Pour des niveaux d'éducation :
    - Primaire : 20
    - Secondaire : 30
    - Universitaire : 50
    - Distribution cumulée :
        - Primaire : 20
        - Secondaire : 20 + 30 = 50
        - Universitaire : 50 + 50 = 100

- **Le tableau statistique:**

Modalités	Effectif	Effectif cumulé	fréquence	Proportion %
A	40	40	0.4	40
B	30	70	0.3	30
AB	10	80	0.1	10
O	20	100	0.2	20

- **Mode**

  La **valeur la plus fréquente** (la catégorie ayant la plus grande fréquence).

  - Exemple : Si 40 personnes sont de groupe sanguin A, A est le **mode**.

II. Variables quantitatives

Une **variable quantitative** est une variable qui prend des valeurs **numériques** mesurables et qui permettent d'effectuer des calculs mathématiques ou statistiques. Elle se distingue des variables qualitatives par le fait qu'elle exprime une quantité ou une mesure.

II.1. Caractéristiques d'une variable quantitative

- Elle permet de mesurer ou de compter.
  - Les opérations mathématiques (addition, soustraction, moyenne, écart-type, etc.) sont possibles.
  - La différence entre deux valeurs a un sens (par exemple, 5 kg de différence entre deux poids).

II.2. Types de variables quantitatives

II.2.1. Quantitative continue

Prend des valeurs dans un intervalle **infini** (décimales possibles).

Mesurée avec une précision arbitraire.

Exemple :

- Taille (170,2 cm)
- Poids (65,7 kg)
- Température (36,8 °C)

II.2.2. Quantitative discrète

- Prend des valeurs **finies ou dénombrables** (souvent des entiers).
- Exemple :
  - Nombre d'enfants dans une famille (0, 1, 2, 3...)
  - Nombre de graines germées dans un échantillon.

II.3. Mesures descriptives pour une variable quantitative

II.3.1. Mesures de tendance centrale

- **Moyenne** : la valeur que devrait avoir chaque individu de façon équitable.
    - **Variable quantitative discrète:**

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i n_i$$

    - **Variable quantitative continue:**

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} c_i n_i$$

- **Médiane** : Valeur qui sépare la moitié inférieure et supérieure des données.
    - **Variable quantitative discrète :**

Si *N* paire.

$$N = 2P \rightarrow Me = \frac{(X_p + X_{p+1})}{2}$$

Si *N* impaire.

$$N = 2P + 1 \rightarrow Me = X_{p+1}$$

    - **Variable quantitative continue:**

$$Me = BI + (BS - BI)\ \frac{\frac{N}{2} + n_{ic-1}}{n_{ic} - n_{ic-1}}$$

- **Les quartiles:** Un quartile est l'une des trois valeurs dénoté Q_1,Q_2 et Q_3 qui divise les données en quatre parties égales de sorte que chaque partie contient le quart des données.
    - **Variable quantitative discrète:**

$$Rang_{Q1} = \frac{N}{4} = N \cdot 0.25 = N25\%$$

$$Rang_{Q3} = \frac{3N}{4} = N \cdot 0.75 = N75\%$$

- **Variable quantitative continue:**

$$Q_1 = BI + (BS - BI)\frac{\left(\frac{N}{4} - n_{ic-1}\right)}{n_{ic} - n_{ic-1}}$$

$$Q_3 = BI + (BS - BI)\frac{\left(\frac{3N}{4} - n_{ic-1}\right)}{n_{ic} - n_{ic-1}}$$

Remarque: Q_2=Me

- **Mode :** Valeur la plus fréquente,la modalité xi dont l'effectif est le plus grand.

Remarque: pour la variable quantitative continue.

$$Mod = BI + a_n \frac{(n_{ico} - n_{ico-1})}{(n_{ico} - n_{ic-1}) + (n_{ico} - n_{ic+1})}$$

II.3.2. Mesures de dispersion

- **Étendue** : Différence entre la plus grande et la plus petite valeur.

$$e = x_{\max} - x_{\min}$$

- **Variance et écart-type** : Indiquent la variabilité des données autour de la moyenne.

$$v(x) = \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Remarque: la variance peut aussi s'écrire:

$$v(x) = \bar{x^2} - \bar{x}^2$$

- Ecart-type:

$$\sigma = \sqrt{v(x)}$$

- **Le coefficient de variation (CV):**

Le coefficient de variation (CV) est une mesure statistique qui exprime l'écart-type en proportion de la moyenne. Il est souvent utilisé pour comparer la variabilité de différentes séries de données, même si leurs unités ou leurs échelles sont différentes.

$$MCV = \frac{\sigma}{\bar{x}}$$

L'interprétation du **coefficient de variation (CV)** peut être classée en différentes catégories pour évaluer la variabilité relative des données. Voici une catégorisation courante :

Chapitre II
Biostatistique Bivariée

introduction

Les variables *x* et *y* peuvent être analysées séparément. On peut calculer

tous les paramètres dont les moyennes et les variances :

Remarque: Ces paramètres sont appelés paramètres marginaux : variances marginales, moyennes marginales, écarts-types marginaux, quantiles marginaux, etc.. . .

II. Analyse des couples (*xi*, *yi*)

II.1. Relation entre *X* et *Y*

L'objectif principal est d'analyser si *X* et *Y* sont :

Indépendants : Pas de lien significatif entre les deux variables.

Corrélés : Une relation linéaire ou non linéaire existe entre les deux variables.

II.2. Organisation des données

Un tableau statistique regroupe les *n* observations sous forme de **couples**.

II.3. Calculs sur la série double

II.3.1. Moyenne et variance pour chaque variable

- **Moyenne pour *X*:**

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{k} x_i$$

- **Variance pour X :**

$$v(x) = \bar{x^2} - \bar{x}^2$$

- **Moyenne pour Y:**

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{k} y_i$$

- **Variance pour *Y* :**

$$v(y) = \bar{y^2} - \bar{y}^2$$

- **Covariance entre *X* et *Y***

$$cov(x;y) = \frac{1}{N}\sum_{i=1}^{k} x \cdot y \ + \bar{x} \cdot \bar{y}$$

Si Cov (xy)>0 : Relation positive entre *X* et *Y*.

Si Cov(xy)<0 : Relation négative entre *X* et *Y*.

Si Cov(xy)=0 : Indépendance linéaire.

II.4. Analyse des Tableaux Croisés

L'analyse des tableaux croisés, également appelée analyse des contingences ou analyse des tables de contingence, est une méthode statistique utilisée pour examiner les relations entre deux variables qualitatives ou catégoriques. Elle consiste à organiser les données sous forme d'un tableau où les lignes représentent les modalités d'une variable et les colonnes celles d'une autre variable. Les cellules du tableau indiquent les fréquences (ou effectifs) observées pour chaque combinaison de modalités.

II.4.1. Analyse des Tableaux Croisés Deux Variables Quantitatives

	y_1	y_2	.....	y_j	total
x_1	n_{11}	n_{12}	.....	n_{1j}	$n_{1.}$
x_2	n_{21}	n_{22}	.....	n_{2j}	$n_{2.}$
x_i	n_{i1}	n_{i2}	.....	n_{ij}	$n_{.j}$
	$n_{.1}$	$n_{.2}$	.....	$n_{.j}$	$n_{..}$

Tableau statistique présente simultanément de manière croisée 2 séries statistiques.

Les $n_{i.}$ et $n_{j.}$ sont appelés les effectifs marginaux. Dans ce tableau :

- $n_{i.}$ représente le nombre total d'occurrences de la modalité x_i pour toutes les colonnes.
- $n_{j.}$ représente le nombre total d'occurrences de la modalité y_j pour toutes les lignes.
- n_{ij} représente le nombre d'occurrences où la modalité x_i est associée à la modalité y_j.

On à :

- $\sum_{i=1}^{k}\sum_{j=1}^{p} n_{ij} = n_{\cdot\cdot} = N$
- $n_{i\cdot} = \sum_{j=1}^{p} n_{ij}$
- $n_{\cdot j} = \sum_{i=1}^{k} n_{ij}$
- $\sum_{i=1}^{k} n_{i\cdot} = N \; et \; \sum_{j=1}^{p} n_{j\cdot} = N$

Les totaux marginaux permettent d'analyser les distributions univariées de chaque variable, tandis que les effectifs n_{ij} servent à étudier les relations entre les deux variables.

- **Moyennes des distributions marginales:**

$$\bar{\bar{x}} = \frac{1}{n_{\cdot\cdot}} \sum_{i=1}^{k} n_{i\cdot} x_i$$

$$\bar{\bar{y}} = \frac{1}{n_{\cdot\cdot}} \sum_{j=1}^{p} n_{\cdot j} y_j$$

- **La covariance:**

La covariance de *x* et *y* est le nombre réel défini par

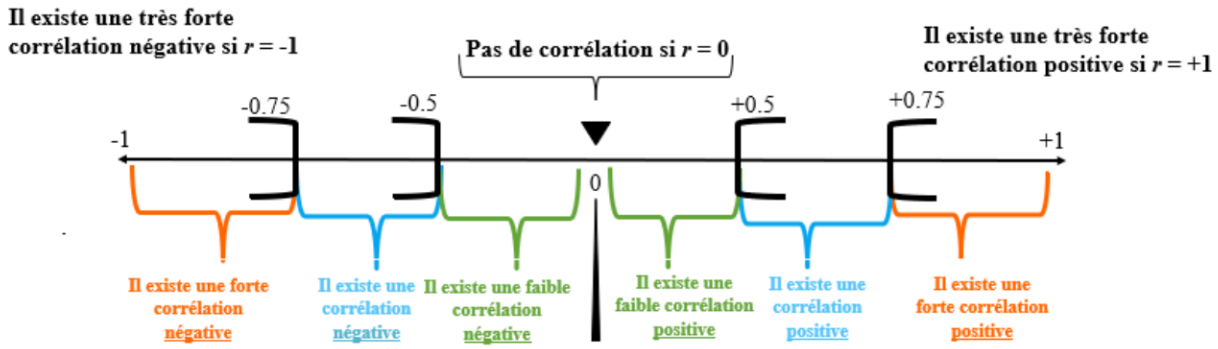
$$cov\,(x;y) = \frac{1}{n_{\cdot\cdot}} \sum_{i=1}^{k}\sum_{j=1}^{p} n_{ij} x_{ij} y_{ij} - \bar{\bar{x}} \cdot \bar{\bar{y}}$$

- **Coefficient de corrélation linéaire (*r*)**

Le coefficient de corrélation de Pearson est donné par :

$$r\,(x;y) = \frac{\mathrm{Cov}(x;y)}{\sigma(x)\sigma(y)}$$

- **La description de cette dépendance:**



Remarque:

Le coefficient de corrélation mesure la dépendance linéaire entre deux variables : $-1 \leq r \leq 1$,

- Quand $x_i = y_i$, pour tout $i = 1, \ldots, n$, la covariance est égale à la variance.
- Si le coefficient de corrélation est nul ou proche de zéro, il n'y a pas de dépendance linéaire. On peut cependant avoir une dépendance non-linéaire avec un coefficient de corrélation nul.
- Si le coefficient de corrélation est positif, les points sont alignés le long d'une droite croissante.
- Si le coefficient de corrélation est négatif, les points sont alignés le long d'une droite décroissante.

II.4.2. Analyse des Tableaux Croisés le test de χ^2:

Objectif du test :

Le test de χ^2 est utilisé pour analyser les relations entre deux variables qualitatives dans un tableau croisé. Il permet de tester si les deux variables sont indépendantes (aucune association) ou dépendantes (existence d'une association significative).

II.4.2.1. Le test des Hypothèses :

- **Hypothèse nulle (H_0) :**

  Les deux variables sont indépendantes.

Il n'existe aucune relation entre les modalités des deux variables.

- **Hypothèse alternative (H_1) :**

  Les deux variables sont dépendantes.

Il existe une relation significative entre les modalités des deux variables.

- **Effectifs observés (n_{ij}) :**

Ce sont les données réelles du tableau croisé, correspondant au nombre d'occurrences observées pour chaque combinaison de modalités.

- **Effectifs attendus (n_{ij} *) :**

$$\chi^2 = \sum \frac{n_{ij}^2}{n_{ij}^*} - n_{..}$$

II.4.2.2. Statistique du test χ^2:

$$\chi^2 = \sum \frac{n_{ij}^2}{n_{ij}^*} - n_{..}$$

- **Degrés de liberté :**

$$df = (L-1) \cdot (C-1)$$

- **Interprétation :**
  - **Comparer la valeur de χ^2 calculée :**

Avec une valeur critique issue de la table du χ^2 , selon un seuil de signification (α, souvent 0.05) et les degrés de liberté (*df*) .Si χ^2 calculée > χ^2 critique : Rejeter H_0, les variables sont dépendantes.Sinon : Ne pas rejeter *H*0, les variables sont indépendantes.

P-valeur :Si la p-valeur associée à χ^2 est inférieure à α (par exemple 0.05), rejeter H_0.

II.4.2.2. Le coefficient de contingence (*C*):

Le coefficient de contingence (*C*) est calculé à partir de la statistique du test χ^2 selon la formule suivante :

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n..}}$$

Où :

χ^2: est la statistique du test de χ^2.

n..: est la taille totale de l'échantillon (le total général du tableau croisé).

Propriétés :

C est une valeur comprise entre 0 et 1 :

C=0 : Les deux variables sont complètement indépendantes.

C proche de 1 : Les deux variables sont fortement associées.

Le coefficient de contingence dépend de la taille du tableau croisé. Il ne peut jamais atteindre 1, sauf si le tableau est carré.

Plus la taille du tableau augmente (nombre de lignes et colonnes), plus la valeur maximale de C diminue.