

INTRODUCTION GÉNÉRALE À LA BIostatistique

Benchikh Tawfik

Faculté de Médecine, UDL, SBA

1^{ère} année Médecine

26 Septembre 2023



PLAN DU COURS

- 1 INTRODUCTION GÉNÉRALE
- 2 TERMINOLOGIE DE BASE DE LA STATISTIQUE



OBJECTIF DE LA STATISTIQUE

- La statistique est la science dont l'objet est de **collecter**, d'**organiser**, de **traiter**, d'**analyser** et d'**interpréter** des données observées issues de l'observation de phénomènes dont le caractère essentiel est la variabilité (phénomène **aléatoires**, c'est-à-dire dans lesquels le hasard intervient).
- Il s'agit de **résumer** l'information contenue dans les données de façon à en dégager les caractéristiques essentielles sous une forme simple et intelligible.
- L'analyse des données est utilisée pour décrire les phénomènes étudiés, faire des prévisions et prendre des décisions à leur sujet.

STATISTIQUE: DÉFINITION

- On appelle **la statistique** l'ensemble des méthodes (ou encore techniques) permettant d'analyser (on dira plutôt de traiter) des ensembles d'observations où données.

- **La biostatistique** se caractérise principalement par l'application de principes statistiques aux **sciences biologiques/biomédicales**;
Ainsi, la biostatistique fait référence au développement de méthodes statistiques et à l'application de principes statistiques à l'étude des phénomènes biologiques et médicaux.

OBJECTIF DE LA BIO-STATISTIQUE

- La statistique constitue, en médecine, l'outil permettant de répondre à de nombreuses questions qui se posent en permanence au médecin:
 - Quelle est la valeur normale d'une grandeur biologique: taille, poids, glycémie?
 - Quelle est la fiabilité d'une mesure ou d'une observation ?
 - Quel est le risque ou l'avantage d'un traitement ?
 - Quel est le risque de complication d'un état pathologique?
 - Le traitement A est-il plus efficace que le traitement B?



LA VARIABILITÉ ET L'INCERTAIN

- Le point fondamental est que les données sont entâchées d'incertitudes et présentent des **variations** pour plusieurs raisons:
 - le déroulement des phénomènes observés n'est pas prévisible à l'avance avec certitude (par exemple on ne sait pas prévoir avec certitude la présence d'une telle maladie suivant les signes observés).
 - toute mesure est entâchée d'erreur.
 - seuls quelques individus sont observés et on doit extrapoler les conclusions de l'étude à toute une population (contexte des sondages)
 - etc...

LA VARIABILITÉ ET L'INCERTAIN

- Cette variabilité **variabilité totale** est due à la variabilité biologique et à la variabilité métrologique (expérimentale liée au protocole de mesure):

$$\text{variabilité totale} = \text{variabilité biologique} + \text{variabilité métrologique}$$

VARIABILITÉ BIOLOGIQUE:

- La variabilité biologique est une caractéristique du vivant, elle s'exprime par une variabilité inter-individuelle et/ou une variabilité intra-individuelle :

Variabilité biologique = variabilité intra-individuelle + variabilité inter-individuelle .

VARIABILITÉ BIOLOGIQUE:

- **Variabilité inter-individuelle:** certaines caractéristique sont différents d'un individu à l'autre (même dans un même groupe)
 - par exemple: poids, taille, couleur des yeux.
- **Variabilité intra-individuelle:** certaines caractéristiques changent chez un même individu à travers le temps
 - Par exemple, lorsque l'on mesure plusieurs fois la pression artérielle chez un individu, les résultats obtenus sont différents à chaque mesure (en fonction de l'état de stress, du moment de la journée, des efforts effectués avant la mesure, etc.)
 - la mesure de la performance d'un athlète qui n'est pas capable des mêmes performances à chaque essai, mais qui se différencie des autres athlètes (variabilité inter-individuelle).
- En général, la variabilité intra est moindre que la variabilité inter.

LA VARIABILITÉ MÉTROLOGIQUE

- **Variabilité métrologique = variabilité expérimentale + variabilité appareil de mesure,**
- Exemple: la mesure de la pression artérielle peut grandement varier sur un individu donné suivant les conditions de cette mesure; il est ainsi recommandé de la mesurer après un repos d'au moins 15 minutes, allongé, en mettant le patient dans des conditions de calme maximal. Cette recommandation vise à minimiser la variabilité due aux conditions expérimentales. La précision de l'appareil de mesure est une donnée intrinsèque de l'appareil, et est fournie par le constructeur.

LA VARIABILITÉ MÉTROLOGIQUE

- La statistique permet d'énoncer des lois qui intègrent l'incertitude créée par la variabilité biologique.
- En effet, lors de la description d'une série de valeurs ou de la comparaison de plusieurs séries de valeurs, la variabilité impose le recours à une méthode particulière permettant de prendre en compte la notion d'incertitude. Cette méthode, la statistique, est basée sur le concept de probabilités.
- C'est pourquoi la statistique et les probabilités sont indispensables à l'étude des problèmes dans le domaine de la santé.

LA DÉCISION DANS L'INCERTAIN

- Pour prendre une **décision diagnostique ou thérapeutique** le médecin doit avoir des éléments lui permettant de prendre en compte cette variabilité naturelle, pour distinguer ce qui est **normal** de ce qui est **pathologique** (décision à propos d'un patient) et pour évaluer la qualité d'un nouvel examen, ou d'une nouvelle thérapeutique (décision thérapeutique).
- La compréhension des méthodes statistiques, de leur puissance et de leurs limites, est essentielle pour un médecin de nos jours.
- Tout résultat de recherche médicale résulte d'une expérimentation (clinique ou biologique) qui s'appuie sur une méthodologie statistique rigoureuse, et dont les résultats sont analysés en termes statistiques.

DÉMARCHE GÉNÉRALE POUR LA STATISTIQUE DESCRIPTIVE

Toute étude statistique peut être décomposée en deux phases au moins:

- Le recueil ou collecte des données,
- L'analyse et l'interprétation des données:
 - **Statistique descriptive**: résumer et présenter les données observées de la manière la plus pertinente possible. Les deux principaux outils de la statistique descriptive sont les
 - Indicateurs statistiques,
 - Représentations graphiques;
 - **Statistique inférentielle**: extrapoler les résultats liés à un échantillon à une population sous-jacente.

L'INFÉRENCE ET RISQUE STATISTIQUE

- Le principe général de l'inférence statistique consiste à tirer des conclusions **fiables** concernant une **population** auquel on ne peut pas accéder directement (généralement parce qu'il est trop grand. à partir d'un échantillon auquel on a accès et que l'on considère comme un échantillon aléatoire provenant de cette population. Cette généralisation se fait au **risque du statisticien**.
- Il s'agit de **statistique probabiliste**, fondée sur la prise en considération des lois de probabilité auxquelles obéissent les phénomènes naturels, objets d'observation.



L'INFÉRENCE ET RISQUE STATISTIQUE: REMARQUE

- **Déduction:** prédire, à partir d'une population connue ou supposée connue, les caractéristiques des échantillons qui en seront prélevés.
- **Induction (inférence):** prédire les caractéristiques d'une population inconnue à partir des statistiques déterminées dans un échantillon représentatif de cette population.

L'IMPORTANCE DE LA DÉMARCHE STATISTIQUE:

- Pour améliorer la pratique médicale dans ses aspects décisionnels, à savoir choisir le meilleur examen (clinique ou para-clinique) pour aboutir le plus rapidement et le plus sûrement au diagnostic.
- Pour optimiser la thérapeutique, choisir le traitement le mieux adapté à un patient donné (choix du médicament, posologie, etc).

- ❶ **Population statistique Ω** : ensemble (au sens mathématique de terme) concerné par une étude statistique: ensemble d'objets ou de personnes homogènes étudiées.
- ❷ **Individu ou unité statistique**: tous élément de la population ($\omega \in \Omega$).
- ❸ **Échantillon**: groupe restreint ou sous-ensemble, issu de la population sur lequel sont effectivement réaliser les observations.
- ❹ **Taille de l'échantillon n** : cardinal du sous-ensemble correspondant.



- 1 **Enquête statistique:** opération consiste à observer (ou mesurer, ou questionner, ... etc) l'ensemble des individus d'un échantillon.
- 2 **Recensement:** enquête dans laquelle l'échantillon observé est la population tout entière (enquête exhaustive).
- 3 **Sondage:** enquête dans laquelle l'échantillon observé est un sous-ensemble strict (une partie) de la population (enquête non exhaustive)

VARIABLE STATISTIQUE: DÉFINITION

- c'est une caractéristique définie sur la population (liée à l'enquête statistique) et observé sur l'échantillon, par exemple: âge, salaire, taille, sexe, ... etc.
- Mathématiquement, il s'agit d'une application définie sur l'échantillon:

$$X : \Omega \rightarrow \begin{cases} \mathbb{R} & \text{si quantitative} \\ \mathcal{E} & \text{si qualitative.} \end{cases}$$

TYPES DES VARIABLE STATISTIQUE

- Si la variable est à valeurs dans $\mathbb{R}$ (celle prenant des valeurs numériques), ou partie de $\mathbb{R}$, comme $\mathbb{N}$ ou $\mathbb{Z}$, ou un ensemble de partie de $\mathbb{R}$, elle est dite **quantitative**, par exemple: âge, salaire, taille, poids, nombre d'enfant dans une famille, ...etc.
- Sinon elle est dite **qualitative** (celles prenant des valeurs non numériques), par exemple: sexe, couleur des yeux, catégorie socio-professionnelle, ... etc.

NOTATIONS

- Une variable statistique (ou aléatoire) est notée par une lettre majuscule X , Y , et les valeurs qu'elle prend par des lettres minuscules x_1 , x_2 , ..., y_1 , y_2 , ...
- un seul caractère étudié: série numérique à une dimension,
- deux caractères étudiés: série numérique à deux dimensions (ou bidimensionnelle),
- plus de deux caractères: on doit utiliser les techniques de l'analyse multidimensionnelle.

TYPES DES VARIABLES QUANTITATIVES

- Une variable quantitative est appelée:
 - **discrète** si elle prend un nombre fini (dénombrable) de valeurs souvent entières, exemple: nombre d'enfants d'un ménage,
 - **continue** si elle prend toutes les valeurs d'un intervalle fini ou infini (les données sont en général regroupées en classes), Exemple: taille. En effet, entre une personne mesurant 160 cm et 161cm, on peut imaginer une infinité de valeurs (ce qui n'existe pas entre 1 et 2 enfants par exemple). Ce sont la précision des instruments de mesure et les conventions qui font que la taille est traitée comme une variable discrète.

TYPES DES VARIABLES QUALITATIVES

- Une variable qualitative **ordinaire** prend des valeurs qui sont ordonnées, hiérarchisées. On peut classer les modalités les unes par rapport aux autres mais on ne peut pas dire à partir de cet ordre de "combien" est la différence entre deux modalités.
Exemple: Les réponses à un sondage, du type "pas du tout", "un peu", "assez", "beaucoup".
- Sinon la variable est dite **nominale**.

MODALITÉS: DÉFINITION

- On appelle modalités, les valeurs possibles de cette variable.
L'ensemble des modalités est noté $\mathcal{E} = \{e_1, \dots, e_m\}$.
- Par exemple, si la variable est le nombre d'enfants par famille, l'ensemble des modalités est $\mathcal{E} = \{0, 1, 2, \dots, 9\}$.
- si la variable est la couleur des yeux d'un individu, l'ensemble des modalités est $\mathcal{E} = \{vert, bleu, brun, gris, noir, \dots\}$.

SÉRIE STATISTIQUE: DÉFINITION

- On appelle série statistique la suite des valeurs prises par une variable X sur les unités d'observation.
- Le nombre d'unités d'observation est noté n (ou N), c'est la taille de l'échantillon .
- Les valeurs de la variable X sont notés

$$x_1, \dots, x_i, \dots, x_n .$$

SÉRIE STATISTIQUE: EXEMPLE

- On s'intéresse à la variable "état-civil" notée X () et à la série statistique des valeurs prises par X sur 20 personnes. La codification est:

C: célibataire

M: marié(e)

V: veuf(ve)

D: divorcée

SÉRIE STATISTIQUE: EXEMPLE

- La population est "les personnes";
- L'échantillon est de 20 personnes;
- La variable statistique est l'"état-civil";
- On obtient la série statistique "brut" suivante:

M	M	D	C	C	M	C	C	C	M
C	M	V	M	V	D	C	C	C	M

Ici, $n = 20$, $x_1 = M$, $x_2 = M$, $x_3 = D$, $x_4 = C$, $x_5 = C$, $\dots$, $x_{20} = M$.

- Les modalités de la variable X est $\{C, M, V, D\}$.

DONNÉES STATISTIQUES: 1

- Il désigne l'ensemble des individus observés (ceux de l'échantillon), l'ensemble des variables considérées et les observations de ces variables sur ces individus.
- Les données sont en général présentées sous forme de tableaux (individus en lignes et variables en colonnes) et stockées dans un fichier informatique.

DONNÉES STATISTIQUES: 2

	X_1	X_2	$\dots$	X_j	$\dots$	X_p
individu 1 (ω_1)	x_{11}	x_{12}	$\dots$	x_{1j}	$\dots$	x_{1p}
individu 2 (ω_2)	x_{21}	x_{22}	$\dots$	x_{2j}	$\dots$	x_{2p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
individu i (ω_i)	x_{i1}	x_{i2}	$\dots$	x_{ij}	$\dots$	x_{ip}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
individu n (ω_n)	x_{n1}	x_{n2}	$\dots$	x_{nj}	$\dots$	x_{np}

$$x_{ij} = X_j(\omega_i) .$$

DONNÉES STATISTIQUES: EXEMPLE

- Voici un tout petit exemple de données:

	sexe	âge	revenu mensuel net
<i>individu 1</i>	1	55	20680
<i>individu 2</i>	1	41	46870
<i>individu 3</i>	1	28	12350
<i>individu 4</i>	2	64	19410
<i>individu 5</i>	2	32	24560

LE RECUEIL DES DONNÉES

- L'enquête statistique est toujours précédée d'une phase où sont déterminés les différents caractères à étudier et le choix de la population à étudier.
- Il se pose alors le problème de l'échantillonnage : choix de l'ensemble des individus à étudier (la population d'étude), la taille de la population et sa représentativité.
- Après il faut établir des protocoles d'étude: Qui fait quoi, comment et quand ? Pré-traitement des données : transformation



LE RECUEIL DES DONNÉES

- Nous appellerons données les valeurs obtenus et référencées suite à une **enquête**, une **investigation** ou une **étude réalisée** (mesures, observations, questionnaire; ...).
- Le recueil des données peut être réalisé soit par la simple **observation** des phénomènes, soit par l'**expérimentation**, c'est-à-dire en provoquant volontairement l'apparition de certains phénomènes contrôlés.

LE RECUEIL DES DONNÉES: REMARQUE

- Le choix de l'échantillon, le recueil des données nécessaires à l'étude, constituent la partie fondamentale, la plus longue, de l'étude.
- De plus, un échantillon doit être **représentatif** de la population dont il est extrait (constituer une image réduite de l'ensemble de la population dont on veut étudier un caractère bien défini) Dans ce cas, ses caractéristiques sont similaires à celles de la population. S'il n'est pas représentatif, on dit qu'il est biaisé. Afin de disposer d'un échantillon représentatif, il faut le constituer d'une manière "aléatoire": on peut pour cela procéder à un véritable tirage au sort (tirage avec remise).

L'IDENTIFICATION DU PROBLÈME

- La phase préliminaire à toute approche statistique vise à déterminer et identifier le problème par un ensemble de questionnements qui permettront de délimiter les investigations et les différentes approches:
 - Quels sont les objectifs ?
 - Quelle est la population ou l'échantillon à étudier ?
 - Quels sont les caractéristiques et les variables ?
 - Que pourra apporter une étude statistique ?

