Méthode statistique en médecine

Introduction

Biostatistiques : statistiques appliquées au domaine de la santé publique
Elles ont 3 objectifs :
- Description d'une maladie par rapport à une population
- Évaluation des traitements, des techniques et des coûts
- Mise en place des observations épidémiologiques et en tirer des conclusions

Les biostatistiques ont pour but de définir si une observation est due au hasard ou si elle a une autre explication.

Définitions :

Statistiques : art de collecter, analyser et interpréter des données. Appliquées au domaine médical, on parle de biostatistiques. Il en existe 2 types :
- Descriptives : description de données à l'aide de paramètres.

Ex : on collecte des données sur la population française : taille, âge, …
- Déductives : l'observation est-elle due au hasard, y a-t-il une autre explication

Ex : on constate que les personnes de 1m60 ont les yeux bleus : est-ce dû au hasard ?

Données : résultat de l'observation d'un individu, grâce à un instrument de mesure, ou par le sens d'un observateur (signes cliniques, biologiques, …)
Une donnée n'est intéressante que si on l'observe ou la compare à d'autres individus. On parle alors de variable car elle prend différentes valeurs selon les individus. *Ex : taille, âge, poids, groupe sanguin, …*

On observe une grande variabilité des données dans le domaine biologique qui peut être due au hasard ou physiologique.
La variabilité peut être :
- inter sujet (=entre 2 sujets) comparaison de 2 sujets
- intra sujet (= pour un même sujet) comparaison du sujet à lui-même

Paramètre : grandeur apportant une information résumée sur la variable étudiée.
Ex : moyenne, médiane, …
Série statistique : collection d'objets de même nature avec des caractéristiques différentes d'un objet à l'autre.
Ex : Les étudiants de LAS de Nice (même nature, caractéristiques différentes)
Population : série exhaustive de **tous** les individus étudiés, sur lesquels on peut appliquer (inférer) des décisions.
Ex : La population française, une école
Échantillon : sous-ensemble fini et d'effectif limité, extrait de la population. Il doit être représentatif de la population, d'où la nécessité de tirage au sort = randomisation
Ex : 100 français tirés au sort

L'échantillon est connu alors que la population est inconnue

Variables :

Il existe 2 grands types de variables :

Variables **qualitatives**	Non mesurables Ex : couleur des yeux, prénom, …	**Binaires** : homme/femme oui/non
		Nominales : couleur des yeux
		Ordinales : échelle de douleur
Variables **quantitatives**	Mesurables (avec appareil de mesure) Ex : taille, poids, …	**Discrètes** : âge
		Continues : poids, glycémie

Une variable qualitative ordinale peut être approximée en une variable pseudo quantitative
ATTENTION : **une variable pseudo quantitative est qualitative**
Ex : le rang/classement à un concours : ce sont des chiffres mais ils n'ont pas de signification et ne peuvent pas faire l'objet d'opération arithmétique. Cette variable est qualitative mais représentée par des chiffres : elle est donc pseudo quantitative.

Paramètres

Moyenne : Variable quantitative discrète : $m = \frac{\sum x_i}{n}$: n l'effectif total et x_i les valeurs prises par la variable

Variable quantitative continue : $m = \frac{\sum n_i x_i}{n}$: n_i l'effectif de chaque valeur x_i

Variance : indique la dispersion des valeurs autour de la moyenne.

Médiane : valeur centrale de l'observation (rangée par ordre croissant) qui sépare la série d'effectif n en 2 sous-séries de même effectif.
Si n est pair, la médiane est la moyenne entre les 2 valeurs $\frac{n}{2}$ et $\frac{n}{2}+1$
Si n est impair la médiane est donnée par $\frac{n+1}{2}$

Quartiles : valeurs de la variable qui séparent la série d'effectif n (rangée en ordre croissant) en 4 sous-séries de même effectif.

Ex : les notes de 5 LAS à l'épreuve de biostats : 15, 12, 20, 10, 18
La moyenne : $\frac{15+12+20+10+18}{5}$ = 15
La médiane : On classe par ordre croissant : 10 < 12 < 15 < 18 < 20
Il y a 5 notes : nombre impair : on prend la note qui est $\frac{5+1}{2}$ = 3 : on prend la 3eme note donc 15.
La médiane est 15
Les quartiles : le 1er : $\frac{1}{4}$ x 5 = 1,25 donc Q1 se trouve entre la 1ere et la 2eme valeur. $\frac{10+12}{2}$ = 11
Le 1er quartile est 11 : 25 % des LAS ont une note inférieure à 11

	Avantages	Inconvénients
Moyenne	+ Simple à calculer + Facile à manipuler dans des tests stats donc adaptée aux calculs statistiques + Très significative si la répartition des données est assez symétrique avec une faible dispersion	- Sensible aux valeurs anormales (max et min)
Médiane	+ Calcul facile + Peu sensible aux valeurs anormales + Utilisable pour des valeurs ordinales, des classes	- Se prête moins aux calculs statistiques

Statistiques descriptives

Estimation en statistiques

Les études en statistiques sont réalisées sur un échantillon représentatif de la population après échantillonnage.

Après l'étude on réfléchit à la légitimité des résultats et à leur extrapolation à la population. On réalise donc une estimation du résultat vrai à partir des données de l'échantillon.

Il y a 2 types d'estimations :
- **L'estimation ponctuelle** : valeur unique jugée la meilleure à un instant t, peu fiable
- **L'estimation par intervalle** : un intervalle de valeur comprend la valeur recherchée : c'est l'intervalle de confiance (IC), beaucoup plus fiable.

2 estimations ponctuelles réalisées sur 2 échantillons donneront des résultats proches mais différents
2 estimations par intervalles réalisées sur 2 échantillons donneront 2 IC se recouvrant mais pas nécessairement le même IC

L'estimation par intervalle est moins précise mais plus juste

Toutes les données biologiques possèdent une **variabilité**. Il est nécessaire de connaître cette variabilité pour classer les données comme « normales » ou « anormales »

Si la variabilité des résultats n'est pas maîtrisée, cela conduit à des biais.
Si cette variabilité est maîtrisée, cela permet une estimation.

I. Données quantitatives :

Méthodologie :
1. *Détermination précise de la population étudiée (=population cible)*
2. *Tirage au sort (TAS) d'un échantillon représentatif (n sujets)*
3. *Calcul de l'intervalle de confiance*

Pour les données quantitatives on va estimer la moyenne

L'estimation assure la correspondance entre ce qui se passe au niveau de l'échantillon et ce qui se passe au niveau de la population

Écart type : il mesure la dispersion d'un ensemble de données autour de la moyenne. C'est la variabilité des mesures entre elles et par rapport à la moyenne.
Plus l'écart type est faible plus le caractère étudié est homogène (les valeurs sont proches de la moyenne).
Ex : À un examen 3 étudiants ont eu 0, 10 et 20, la moyenne est de 10, la médiane et de 10. Ici c'est l'écart-type qui permettra le mieux de résumer la dispersion de la série.
Si les étudiants avaient eu 9, 10 et 11 la moyenne et la médiane seraient les mêmes, l'écart-type serait plus petit.

Degré de liberté ou ddl : c'est le nombre de valeur à connaitre pour résoudre une équation et connaître toutes les valeurs de la série.

Si « m » la moyenne, « x_i » les valeurs dont on veut faire la moyenne, « n » l'effectif et « $x_i - m$ » les écarts
Il y a n écarts
Il y a (n - 1) écarts indépendants à la moyenne ou degré de liberté
Ex : Un élève a eu 4 notes : 12, 15, 16 et une copie perdue dont il veut connaître la note. Il connaît sa moyenne de 15. Le degré de liberté est (n -1) donc 3 et il a 3 copies il peut donc retrouver sa dernière note.
Par exemple en faisant : $\frac{12+15+16+?}{4}$ *= 15. 43 + ? = 60 Donc sa dernière note était 17*

L'intervalle de confiance : c'est l'estimation de la moyenne vraie μ à partir de la moyenne m calculée sur l'échantillon.
On donne un intervalle auquel μ appartient :

$$\mu \in [m \pm \frac{\varepsilon s}{\sqrt{n}}]$$

L'IC est aussi appelé intervalle au risque α

Avec « n » l'effectif et « s » l'écart-type

Le risque α : c'est le risque d'erreur dans l'estimation de μ (le risque que notre IC ne contienne pas μ)
On prend en général **α = 5%** (on a 95% de chance que la moyenne vraie soit dans notre IC)

L'écart-réduit ε : c'est une valeur qui dépend du risque α : ils varient en sens inverse : si α augmente ε diminue.
Un écart-réduit mesure de combien d'écart-type une observation particulière est éloignée de la population.

Pour **α = 5% ε = 1, 96**
Pour **α = 1%. ε = 2, 60**

Précision de l'intervalle :

Les variations du risque α vont conditionner la précision de l'estimation et la largeur de l'IC

Si on prend moins de risque (α ↓) l'intervalle de confiance augmente (car ε↑). On a plus de chance que la moyenne soit dedans mais l'estimation est moins précise.

Si on prend plus de risque (α↑) l'IC diminue (car ε↓)
L'estimation est plus précise mais il y a plus de chance que la moyenne ne soit pas dans l'IC

L'indice de précision i : il permet de calculer la précision de l'estimation de μ. Cette valeur représente la largeur de l'IC.

$$i = \frac{\varepsilon s}{\sqrt{n}}$$

D'après la formule de l'IC vu avant l'**IC** est compris **entre [m + i] et [m - i]**

D'après la formule de l'indice de précision, si n ↑, i ↓ donc l'IC ↓ donc la précision ↑
Plus la taille de l'échantillon augmente, plus la précision augmente.

Le nombre de sujets nécessaires « n » pour une précision donnée : $n = \frac{\varepsilon^2 s^2}{i^2}$ (la même formule que i)

RECAP

- L'IC est l'estimation de la moyenne vraie μ à partir de la moyenne m calculée sur l'échantillon
- Le risque α est le risque d'erreur dans l'estimation de μ
- ε représente l'écart-réduit
- Les variations du risque α déterminent la précision de l'estimation
- i représente la largeur de l'IC
- IC = [m ± i]

DONC +++

⇒ Si n↑ alors i↓ donc l'IC se resserre donc la précision ↑
⇒ Si α↑ alors ε↓ donc i↓ donc l'IC se resserre donc la précision ↑

Loi de Gauss ou loi normale :

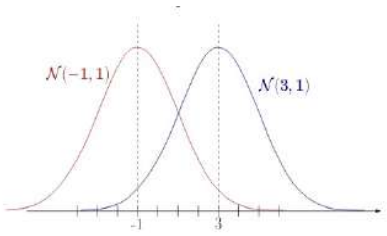
En sciences humaines, on observe souvent des distributions des variables assez symétriques autour de la moyenne : c'est la **courbe de Gauss**

La représentation graphique de données suivant la courbe de Gauss est une courbe en cloche avec :

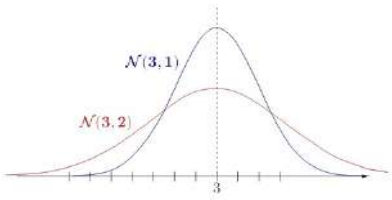
- En abscisse [m ± εs] donc l'IC
- En ordonnée n_i : l'effectif pour chaque valeur
- L'aire sous la courbe, le % de la population concerné

La courbe de Gauss permet de **visualiser l'IC** autour de la moyenne, **l'écart-type**, la dispersion autour de cette valeur moyenne et **la moyenne**.

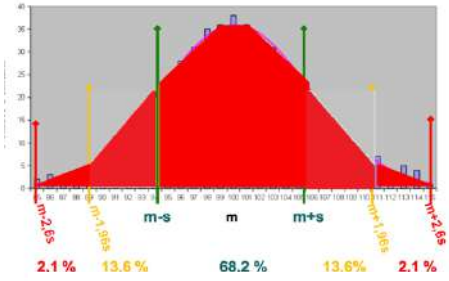
Pour pouvoir faire des calculs on suppose que notre variable X (quantitative continue) suit une distribution modèle : la **loi Normale**.
Ainsi, pour chaque couple (μ, σ), il existe une loi normale de moyenne μ et d'écart-type σ notée **N(μ,σ)**



Même écart-type, moyennes différentes



Même moyenne, écarts-types différents (dispersion ↑)



À partir de la loi normale ou de Gauss on peut retrouver des IC