METHODE STATISTIQUE EN MEDECINE

méthode statistique en médecine

INTRODUCTION

Biostatistiques = statistiques appliquées au domaine de la santé publique

Elles ont 3 objectifs :

- **Description** d'une population par rapport à une maladie
- **Evaluation** traitements, techniques, coûts
- Mise en place des **observations** épidémiologiques, **conclusions**

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

DEFINITIONS

Statistique = art de collecter, analyser et interpréter des données.

Lorsque l'on applique les statistiques au domaine de la biologie/médecine, on parle de **biostatistiques**.

Il en existe 2 types :

- **Descriptives** : description d'une situation à l'aide de **paramètres**
  par exemple on collecte des données sur la population française : taille et âge
- **Déductives** : l'observation est-elle due au **hasard ?** Existe-t-il une autre explication ?
  par exemple on constate que les personnes de moins d'1m65 sont brunes. Est-ce dû au hasard ?

Données = résultat de l'observation d'un individu, grâce à un instrument de mesure, ou par les sens de l'observateur *(signes cliniques, biologiques…)*

Le but d'une donnée est de l'observer ou de la comparer sur plusieurs individus. On parle donc de **variable**.
La variable prend une valeur pour un individu, une autre valeur pour un autre individu etc…

On observe une **grande variabilité des données** dans le domaine biologique qui peut être due au hasard ou qui peut être physiologique :

- **inter sujet** (=entre deux sujets) comparaison de 2 sujets
- ou **intra sujet** (=pour un même sujet) comparaison du sujet à lui-même

par exemple des données peuvent être la taille, le poids, l'âge, le groupe sanguin….

Paramètre	grandeur apportant une information résumée *(ou synthétisée)* sur la variable étudiée	*par exemple moyenne d'une série de valeur, écart-type…*
Série statistique	collection d'objets de même nature, avec des caractéristiques différentes d'un objet à l'autre *(variables)*	*par exemple les hommes et les femmes (même nature, caractéristiques différentes)*
Variable quantitative	mesurable, obtenue grâce à un appareil de mesure	*par exemple taille d'un individu, poids d'un individu*
Variable qualitative	non mesurable	*par exemple la couleur des yeux, couleur des cheveux*
Population	série exhaustive de tous les individus étudiés, sur lesquels on veut appliquer (inférer) des décisions	*par exemple population de la France, une école*

Echantillon	sous ensemble fini et d'effectif limité, extrait de la population. Il doit être **représentatif** de la population d'où la nécessité du **tirage au sort = randomisation**.	*par exemple 10 personnes tirées au sort dans la population française, une classe tirée au sort dans l'école*

L'échantillon est connu, alors que la population est inconnue

Types de variables

Il existe 2 types de variables :

Variables qualitatives	*Binaires* : homme/femme
	Nominales : couleur des yeux
	Ordinales : douleur
Variables quantitatives	*Discrètes* : âge
	Continues : poids, glycémie

Une variable qualitative ordinale peut être approximée en une **variable pseudo quantitative.** *la variable est qualitative mais ressemble à une quantitative*

+++ ATTENTION : une variable pseudo quantitative est qualitative ++++

par exemple le rang/classement au concours, les scores en médecine : ce sont des chiffres mais ils n'ont pas de signification et ne peuvent pas faire l'objet d'opérations arithmétiques. Cette variable est donc qualitative mais comme on la représente par des chiffres on dit qu'elle est pseudo quantitative.

Parametres

Differents parametres

On peut résumer en quelques paramètres les caractéristiques de la série de données quantitatives :

Moyenne	Variable **quantitative discrète** : $m = \frac{\Sigma xi}{n}$	
	Variable **quantitative continue** : $m = \frac{\Sigma nixi}{n}$	
Variance	indique la **dispersion** des données autour de la moyenne	
Médiane	valeur de l'observation centrale qui **sépare la série d'un effectif n en 2 sous séries de même effectif**	Si n est pair : la médiane est donnée par la **moyenne des deux valeurs correspondantes à n/2 et (n/2)+1**
		Si n est impair : la médiane est donnée par **(n+1)/2**
Quartiles	valeurs de la variable qui **partagent la série d'effectif n en 4 sous séries de même effectif**	

Exemple : les notes de 5 PASS à l'épreuve de biostats : 14/15/12/20/18
*1) **Moyenne** : (14+15+12+20+18)/5 = 15,8*
*2) **Médiane***
D'abord on classe par ordre croissant : 12/14/15/18/20
Ensuite on compte le nombre de notes : 5 → nombre impair
On prend la note qui est la (5+1)/2 = 3
La 3e note c'est 15 donc la médiane = 15
*3) **1e quartile***
On fait 1/4 x 5 = 1 ,25
Donc Q1 se trouve entre la 1e et la 2e note
Donc Q1 = (12+14)/2 = 13
25% des PASS seulement ont une note inférieure à 13

	Avantages	**Inconvénients**
Moyenne	- simple à calculer - facile à manipuler dans les test stats donc adaptées aux calculs statistiques - très significative si la répartition des données est assez symétrique et avec une faible dispersion	- sensible aux valeurs anormales (max et min)
Médiane	- calcul facile - peu sensible aux valeurs anormales - utilisable pour les valeurs ordinales, les classes…	- se prête moins aux calculs statistiques

ESTIMATIONS EN STATISTIQUES

Les études en biostatistique sont réalisées sur un échantillon représentatif de la population après « échantillonnage ».

Après l'étude on doit réfléchir à la légitimité des résultats et à leur extrapolation potentielle à l'ensemble de la population. Pour ça on réalise une **estimation du résultat vrai** à partir des données obtenues sur l'échantillon.

On retrouve deux types d'estimations :

- **L'estimation ponctuelle** : valeur unique jugée la meilleure à l'instant t, peu fiable
- **L'estimation par intervalle** : il y a un intervalle de valeurs comprenant la valeur recherchée, c'est l'Intervalle de Confiance ou IC, beaucoup plus fiable

***Méthodologie** :*
1) Définition précise de la population étudiée = Population cible
2) Tirage au sort (TAS) d'un échantillon représentatif
3) Calcul de l'intervalle de confiance
*Pour les données quantitatives, on va estimer la **moyenne** !*
L'estimation assure la correspondance entre ce qu'il se passe au niveau de l'échantillon et ce qu'il se passe au niveau de la population

ECART-TYPE

***Ecart-type* = 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** *(plus les valeurs sont proches de la moyenne)*.

DEGRE DE LIBERTE (DDL)

Les ddl = nombre de valeurs nécessaires à connaître pour pouvoir résoudre l'équation et connaître toutes les valeurs de la série

par exemple Julie a eu 4 notes mais a perdu une feuille… elle se souvient qu'elle a eu un 15, un 17 et un 13 et elle se souvient que sa moyenne est de 15. Ce qui est top c'est qu'elle a 3 notes sur les 4 (soit n-1 valeurs) donc on peut trouver la dernière. Mais comment ça ? En utilisant la moyenne par exemple.
La moyenne c'est la somme de toutes les valeurs divisée par l'effectif (le nombre de valeurs).

$$moyenne = 15 = \frac{15 + 17 + 13 + x}{4}$$

$$60 = 45 + x$$

$$15 = x$$

Donc sa dernière note était un 15.

INTERVALLE DE CONFIANCE

L'IC = 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**.

Risque α = risque d'erreur dans l'estimation de μ+++

Autrement dit le risque que notre IC ne comprenne pas la valeur vraie de μ.

On prend en général **α = 5%.**

ε l'écart-réduit = valeur qui dépend du risque α, ils varient en sens inverse, si α augmente, ε diminue.

Un écart-réduit mesure de combien d'écarts-types une observation particulière est éloignée de la population.

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

PRECISION DE L'ESTIMATION

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↓), on a un **intervalle de confiance plus grand** (ε↑), on a **plus de chances que la moyenne soit dedans**.

Indice de précision « i » = 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 vue juste avant, l'IC est compris :

entre $[m - \frac{\varepsilon s}{\sqrt{n}}]$ et $[m + \frac{\varepsilon s}{\sqrt{n}}]$ donc entre $[m - i]$ et $[m + i]$.

D'après la formule de l'indice de précision :

si n↑ alors i↓ donc l'IC↓ donc la précision ↑

On peut conclure que **plus la taille de l'échantillon augmente**, **plus la précision augmente**. « n » le nombre de sujets nécessaires : $n = \frac{\varepsilon^2 s^2}{i^2}$

♥ RECAP ♥

L'IC = **estimation de la moyenne vraie μ** à partir de la **moyenne m** calculée sur l'échantillon. Il est aussi appelé **"intervalle au risque α"**.

Risque α = **risque d'erreur** dans l'estimation de **μ**.

ε = écart-réduit **distribution des données** autour de la **moyenne**.

Les variations du risque α déterminent la précision de l'estimation.

i représente la **largeur de l'IC** : $i = \frac{\varepsilon s}{\sqrt{n}}$

IC = [m ∓ i]

Si n↑ alors i↓ donc l'IC↓ donc la précision↑

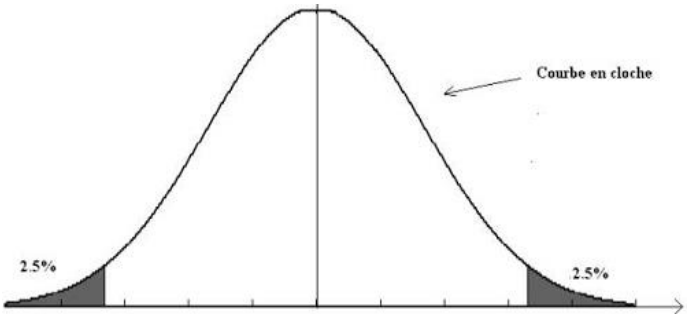
Si α↑ alors ε↓ donc i↓ donc l'IC↓ donc la précision↑

Loi de Gauss ou loi Normale

En sciences humaines, on observe souvent des distributions des variables plutôt symétriques autour de la moyenne avec une forme de cloche : c'est la **courbe de Gauss**.

La représentation graphique de données par la loi de Gauss donne une courbe en cloche avec :

- En abscisse : $[m \pm \varepsilon s]$, donc **l'IC**
- En ordonnée : **n**
- L'aire sous la courbe : le **%** de la population concernée



La loi 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 va supposer que notre variable X (quantitative continue) suit une distribution « modèle » : la **loi Normale**. Ainsi, Pour chaque (μ,σ) il existe une loi normale de moyenne μ et d'écart type σ : on la note **N (μ, σ)**

Estimation des données qualitatives

Méthodologie *:*

1) Constitution d'un échantillon représentatif par TAS

2) Calcul du pourcentage $pobs$ de l'échantillon présentant un caractère A et de l'écart-type « s »

3) Estimation de la valeur vraie « p » du pourcentage de la population présentant A et de l'écart-type « σ »

Pour les données qualitatives, on va estimer un **pourcentage** !

L'estimation assure la correspondance entre ce qui se passe au niveau de **l'échantillon** et au niveau de la **population**. Seuls changeront les paramètres utilisés et donc les formules qui en découlent.

Intervalle de Confiance

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

$$p \in [pobs \pm \varepsilon s]$$

FIN

Il était une fois,

Un P1 rassuré d'avoir fini ce cours de biostatistiques

Mais les biostatistiques reviendront encore plus fortes bientôt…

Sauf que nos P1 acharnés et de par l'aide de leur tuteurs finiront pas vaincre cette menace !

Et de leur apprentissage ils tireront leur plus grande force.

!!! *en gros* ON LACHE RIEN LES GARS !!!