METHODE STATISTIQUE EN MEDECINE

I. Introduction

Les **biostatistiques** sont les **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.

II. Définitions

Statistiques	correspondent à l'art de collecter, d'analyser et d'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	correspondent au 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**.

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.

III. Les 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 ++++

IV. Représentation des variables

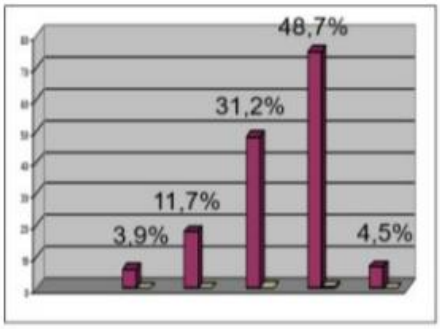
A. Variables qualitatives

On peut les représenter de 2 manières :

- **Tableau**
- **Diagramme en bâtons** ou **histogramme**

Ex : degré de satisfaction des mères accouchant dans une maternité

Degré de satisfaction	Nb de mères	%
Très insatisfait	6	3,9
Plutôt insatisfait	18	11,7
Plutôt satisfait	48	31,2
Très satisfait	75	48,7
Pas d'opinion	7	4,5



ATTENTION : un pourcentage est une variable qualitative

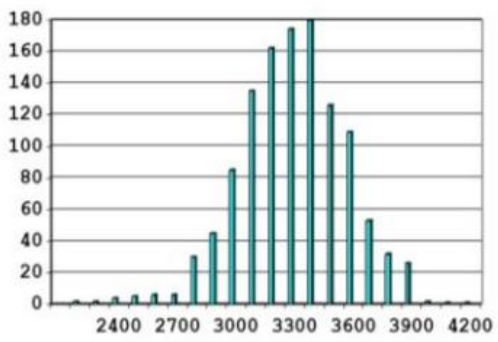
B. Variables quantitatives

On peut les représenter de 3 manières :

- **Tableau**
- **Diagramme en bâton** ou **histogramme**
- Résumée grâce à des **paramètres**

Ex : poids des nouveaux nés dans la maternité

Poids (g)	Nb bébés
2200	2
2300	2
2400	4
2500	5
...	...
3100	121
3200	150
3300	162
3400	170



V. Paramètres

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

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

<u>Exemple</u> : 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 3^{e} note c'est 15 donc la médiane = 15
*3) **1^{e} quartile***
On fait 1/4 x 5 = 1 ,25
Donc Q1 se trouve entre la 1^{e} et la 2^{e} note
Donc Q1 = (12+14)/2 = 13
25% des PASS seulement ont une note inférieure à 13

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

STATISTIQUES DESCRIPTIVES

I. Notion de variabilité

Toutes les données biologiques possèdent une variabilité.

La connaissance de cette variabilité est nécessaire pour pouvoir classer nos données comme « normales » ou « anormales ».

- ☞ Une **variabilité maîtrisée** permet une **estimation**
- ☞ Une **variabilité non maîtrisée** conduit à des **biais**

Par exemple les valeurs normales de la glycémie sont comprises entre 0,75 et 1,25 g/L. Si on est en dessous de 0,75 g/L on a une valeur anormale, on est en hypoglycémie.

II. Estimations en statistiques

A. Définition

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 détermine des paramètres au niveau d'une population à partir d'observations réalisées sur un échantillon de cette population.

Echantillon ➔ estimation ➔ population cible

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+++)*

Deux estimations **ponctuelles** d'une même variable réalisées sur 2 échantillons A & B donneront des valeurs ponctuelles proches mais pas nécessairement la même valeur.
Deux estimations **par intervalle** d'une même variable réalisées sur 2 échantillons A & B donneront des IC se recouvrant *(car proches)* mais pas nécessairement le même IC.

L'estimation par intervalle est moins précise

Cependant, si on refait la même estimation sur un autre échantillon, elle recouvrira la première, ce qui ne serait sûrement pas le cas avec des valeurs ponctuelles.

Donc l'estimation par intervalle est plus juste, d'où son intérêt.

B. Estimation des données quantitatives

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

1) Écart-type :

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

2) Degrés de liberté :

On définit « m » la moyenne, « xi » les valeurs dont on veut faire la moyenne, « n » l'effectif, « $xi - m$ » les écarts.

- ☞ Il y a **n écarts**
- ☞ Il y a **(n – 1) écarts indépendants à la moyenne**, ou **degrés de liberté**

Degrés de liberté ou ddl	c'est le nombre de valeurs nécessaires à connaître pour pouvoir résoudre l'équation et connaître toutes les valeurs de la série. *(Si je connais ma moyenne et toutes mes valeurs sauf une, je peux trouver la valeur manquante).*

Exemple : Paul a eu 3 notes mais une de ses évaluations est tachée (pas cool). Il sait qu'il a eu 12 et 13 et il connaît sa moyenne : 14. m=14 ; x={12,13,y} ; n=3
Il peut donc avec n-1 valeurs, c'est à dire 2 valeurs, trouver la troisième, il y a différentes techniques, comme par exemple avec la moyenne: $moyenne$ = 14 = (12+13+y)/3 donc y = 17
On retrouve bien sa note à partir des autres, cependant s'il manquait deux notes on n'aurait pas pu déterminer la deuxième c'est pourquoi il y a n-1 et pas n-2 ddl.

3) *Intervalle de Confiance :*

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**.

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

L'IC est aussi appelé **intervalle au risque α**.

Risque α	c'est le risque d'erreur dans l'estimation de μ+++ *(autrement dit le risque que notre intervalle de confiance ne comprenne pas la valeur vraie de μ)*.

On prend en général α = 5% ***(donc on a 95% de chances que la moyenne vraie appartienne bien à l'IC).***

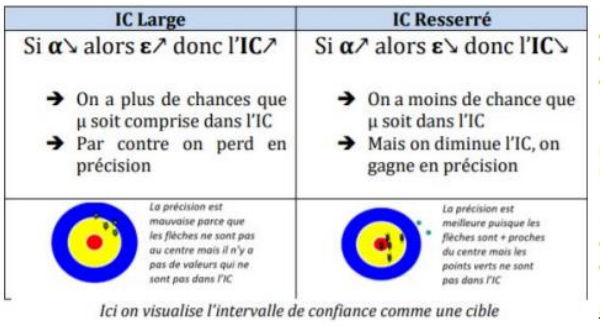
ε = l'écart-réduit	C'est une valeur qui dépend du risque α : ils **varient en sens inverse**, si α augmente, ε diminue

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

Pour **α = 5%** ; **ε = 1,96**

Pour **α = 1%** ; **ε = 2,60**

4) *Précision de l'estimation :*



Ici on visualise l'intervalle de confiance comme une cible

Les variations du risque α vont conditionner la **précision de l'estimation** et la **largeur de l'intervalle de confiance**.

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. *(et inversement)*

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 vue juste avant, l'IC est compris : entre $[m - \frac{\varepsilon s}{\sqrt{n}}]$ et $[m + \frac{\varepsilon s}{\sqrt{n}}]$ donc :

IC compris 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 ↗

Attention : quand l'indice de précision diminue la précision augmente !

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** c'est **l'estimation de la moyenne vraie μ** à partir de la moyenne m calculée sur l'échantillon. Il est aussi appelé **"intervalle au risque α"**.
- Le **risque α** c'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** : $i = \frac{\varepsilon s}{\sqrt{n}}$
- $IC = [m \pm i]$

DONC +++

Si n↗ alors i↘ donc l'IC↘ donc la précision↗
Si α↗ alors ε↘ donc i↘ donc l'IC↘ donc la précision↗

J'espère que vous avez bien compris, c'est hyper important de connaître les différentes variations en fonction des autres facteurs, ça tombe souvent au concours. Sinon envoyez moi un message sur le forum ou sur discord.

5) Loi de Gauss ou loi Normale :

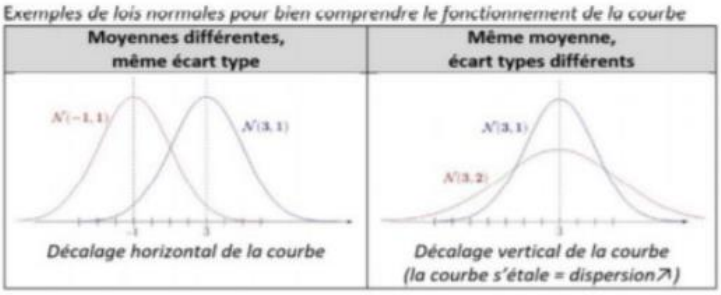
En sciences humaines, on observe souvent des **distributions des variables plutôt symétriques** autour de la moyenne avec une forme de cloche : c'est la **courbe de Gauss**.

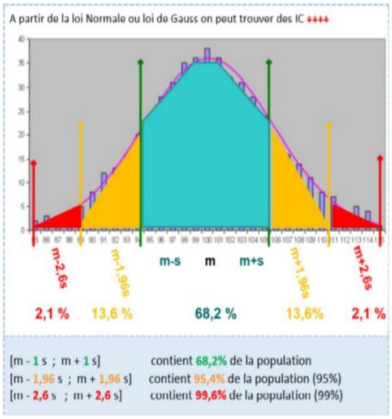
La <u>représentation graphique de données</u> par la loi de Gauss donne une courbe en cloche avec :

- ♥ En <u>abscisse</u> : $[\boldsymbol{m} \pm \boldsymbol{\varepsilon s}]$, donc l'IC
- ♥ En <u>ordonnée</u> : **n**
- ♥ <u>L'aire sous la courbe</u> : le **% de la population concernée**

La <u>loi de Gauss</u> permet de <u>visualiser l'IC autour de la moyenne</u>, <u>l'écart type</u>, la <u>dispersion</u> autour de cette valeur moyenne et la <u>moyenne</u>.

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





C. <u>Estimation des données qualitatives</u>

<u>Méthodologie :</u>
1) Constitution d'un **échantillon représentatif** par **TAS**
2) Calcul du **pourcentage $\boldsymbol{pobs}$ 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 <u>données qualitatives</u>, on va estimer un <u>pourcentage</u>

Comme précédemment, **l'estimation assure la correspondance entre ce qui se passe au niveau de l'échantillon et au niveau de la population**

Tout ce qui va suivre sera le <u>même procédé que pour les données quantitatives</u>, seuls changeront les paramètres utilisés et donc les formules qui en découlent.

1) Écart-type :

Ecart-type	Il a les mêmes caractéristiques que la variable soit qualitative ou quantitative

Il est donné par :

$$s = \sqrt{pobs.\frac{qobs}{n}}$$

avec $qobs = 1 - pobs$

2) Intervalle de Confiance :

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 :

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

3) Précision de l'estimation :

Indice de précision « i »	Il représente toujours la largeur de l'IC

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

Si **n est multiplié par 100**, alors **s est divisé par 10** et donc la **précision augmente d'un facteur 10**

On peut aussi conclure sans problème la même chose :

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**. La précision dépend de la taille de l'échantillon, et de l'écart-type « s ».

« n » le nombre de sujets nécessaires : $n = \varepsilon^2 pq\, i$

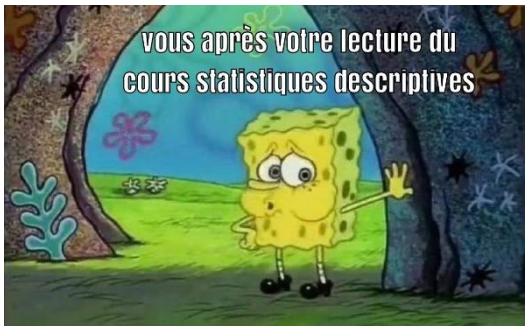
4) Sondages :

Le **sondage** est une **application directe de l'IC** calculée sur des données qualitatives. Tout résultat de sondage doit être accompagné d'un IC.

Pour une bonne estimation il nous faut donc :

- Un **échantillon représentatif** constitué par TAS
- **Pas de biais** pendant la sélection
- Un **IC** qui accompagne toujours l'estimation *(il montre la variabilité des données)*
- Une **taille importante** de l'échantillon : Si n↗ la précision↗

LES DEDIS, LES DEDIS…



Et voilà les boss, fin du cours qui comme vous avez pu le voir reprend les bases de la TTR et les approfondi. Ce cours est à bien comprendre pour appréhender comme il faut le cours d'Olivier *(ce génie)* sur les statistiques déductives.

On le sait c'est pas facile, mais sachez-le tous vos efforts seront tôt ou tard récompensés. Pour le moment, tout vous parait flou et vous pensez couler sous la tonne de cours qui sortent. Trouvez-votre méthode de travail et si besoin on est là pour ça aussi et ne perdez pas votre temps. N'oubliez pas que vos tuteurs de biostat sont là pour votre moral quand vous êtes au plus bas *(et pas que pour déconner)*. Vous allez devenir des machines de guerre dans peu de temps <3

Enfin, on le dira jamais assez mais le tutorat c'est plein de love alors je vais envoyer ma dose d'amour :

♥ A mes co-tuts qui vous font des fiches / fiches récap / QRU incroyables

♥ Aux tuteurs de cette année (big up à Oskour, clochonou, Tagada et Dydou avec qui on a chanté jusqu'au bout de la nuit)

♥ A toutes les matières du S2 (et à pousssezMadamepouC ++)

♥ A Estechaise le bg suprême

♥ Et **à vous les bg** qui venez de finir cette pseudo-ronéo interminable

LA BIOSTAT VOUS AIME

