Méthode statistique en médecine

1) Introduction

- Biostatistiques : statistiques appliquées au domaine de la santé publique

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

2) Définitions

- Statistique : art de collecter, analyser et interpréter des données.
Il en existe 2 *types* en biostatistiques :

- Descriptives : description de données à l'aide de paramètres.

Ex : on collecte des données sur les étudiants en LAS : QI, âge, taille, note en biostat ...

- Déductives : l'observation est-elle due au hasard, y a-t-il une autre explication

Ex : on constate que les personnes qui aiment la biostat ont une meilleure espérance de vie : est-ce dû au hasard ?

- Données : c'est le 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 est différente selon les individus.

Ex : taille, âge, poids, groupe sanguin...

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

On revient aux variables après tkt

- **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 LAS tirés au sort

La population est *inaccessible* dans son entièreté pour des raisons d'organisation et de moyens limités. Du coup on réalise l'étude sur l'échantillon puis on fait un « *pari* » sur l'application des résultats à la population.

Logique, on va pas prendre tous les étudiants en médecine de France pour tester leur niveau de stress, alors qu'on peut prendre un échantillon à Nice, du coup faut faire gaffe à la fiabilité du résultat !

L'ÉCHANTILLON EST CONNU, ALORS QUE LA POPULATION EST INCONNUE.

Ca c'est très important please

3a) Variables

Variable **qualitative**	Variable **quantitative**
Non mesurable, *Couleur des yeux, prénom...*	Mesurable, *Taille, poids...*
Binaires *Femme/homme*	Discrètes *Age (sans virgule)*
Nominales *Couleur des cheveux*	Continues *Poids, glycémie (avec virgule)*
Ordinales *Echelle de douleur de 1 à 10*	

Tips : Une variable discrète est trop timide pour utiliser des virgules !

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

En gros, si on a une échelle par exemple de douleur, de satisfaction... on va les classer de 1 à 10 (donc ça c'est pseudo quantitatif), mais ça reste des scores subjectifs qui dépendent de la personne donc qu'on peut pas vraiment mesurer (donc qualitatif)

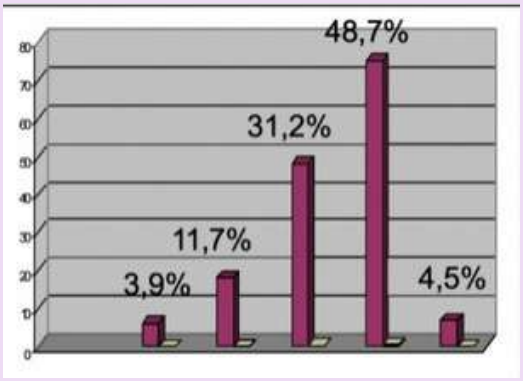
UNE VARIABLE PSEUDO QUANTITATIVE RESTE QUALITATIVE !!

Ca aussi grrrr

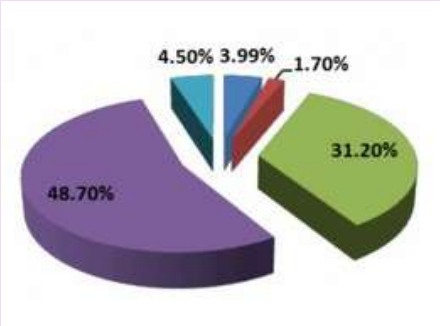
3b) Représentation des variables

Qualitatives :

On les représente sous forme de tableau de % , d'histogramme, de secteurs ... *un pourcentage est une variable qualitative*



Degré de satisfaction	Nb 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%

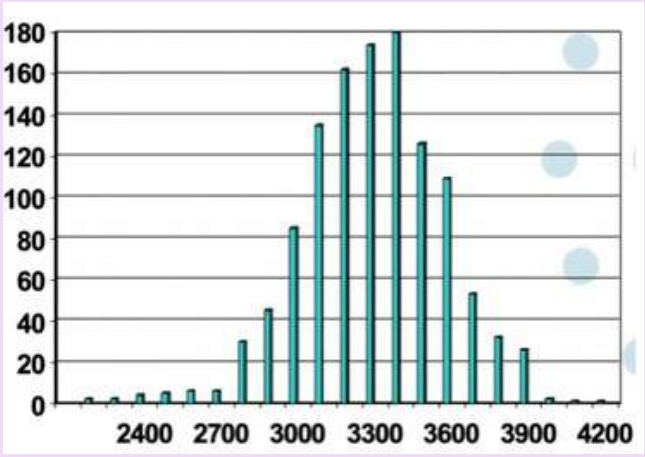


(N'apprenez pas ces images, c'est pour illustrer <3)

Quantitatives :

On les représente sous forme de tableau, de diagramme en bâton ou d'histogramme

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



4) Paramètres

MOYENNE

Variable quantitative **discrète**

$$m = \frac{\sum x_i}{n}$$

Variable quantitative **continue**

$$m = \frac{\sum n_i x_i}{n}$$

MÉDIANE

N pair : moyenne des $\frac{n}{2}$ et $\frac{n+1}{2}$ valeurs

Valeur centrale qui **sépare** la série d'effectif n en 2 sous séries de même effectif

N impair : $\frac{n}{2}$ eme valeur

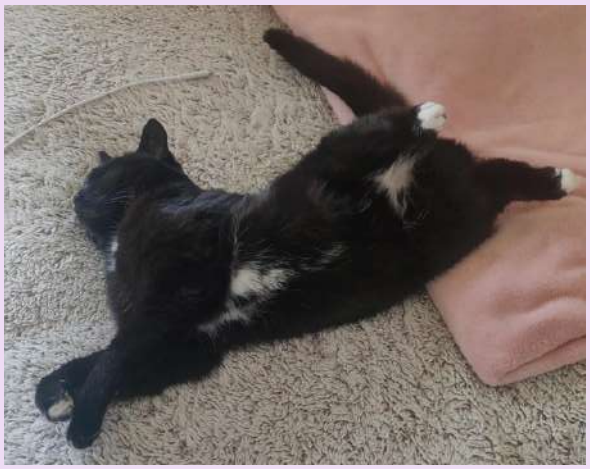
VARIANCE

Indique la dispersion des valeurs autour de la moyenne.

QUARTILES

Valeurs de la variable qui partagent la série d'effectif n en 4 sous séries de même effectif

Pas de formules yayy



Soyez comme réré <3

Exemples :

Enoncé : Les notes des LAS en biostat au premier EB (tintintinnn) : 10, 7, 15, 20, 2

Calculer la : moyenne, médiane, quartiles :

MOYENNE : (10 + 7 + 15 + 20 + 2) / 5 = 10,8 ça c'est easy

MEDIANE :
1) remettre dans l'ordre la suite : 2, 7, 10, 15, 20
2) parité de la suite : ici impair car 5 valeurs
3) application : on prend la valeur du milieu : 10

QUARTILES :
1) premier quartile on fait 1/4 x 5 = **1 ,25** avec 5 le nombre de valeurs
2) donc Q1 se trouve entre la **1e et la 2e note**
3) Q1 = **(2+7)**/2 = 4,5
4) 25% des LAS seulement ont une note inferieure à 4,5

Perso j'ai mis du temps à comprendre les quartiles donc si vous arrivez pas go fofo ;)

	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

1) Variabilité

Toutes les données biologiques possèdent une variabilité.
Il faut la connaitre pour pouvoir classer nos données comme « normales » ou « anormales » :

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

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

2) Estimation statistique

- Les études en biostatistique 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.

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

Ce charabia résumé

On retrouve <u>deux types d'estimations :</u>

- L'estimation ponctuelle : valeur unique jugée la meilleure à l'instant t (PEU FIABLE)
- L'estimation par intervalle : un intervalle de valeurs comprenant la valeur recherchée, c'est **l'Intervalle de Confiance ou IC** (BEAUCOUP + 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.**

> 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

L'ESTIMATION PAR INTERVALLE EST MOINS PRÉCISE MAIS PLUS JUSTE

3) Estimation des 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 qu'il se passe au niveau de l'échantillon et ce qu'il se passe au niveau de la population.

ECART-TYPE	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 : A l'épreuve de biostar 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.
En gros plus les valeurs sont éloignées plus l'écart-type est grand, et inversement.

DEGRÉ DE LIBERTÉ DDL	Le nombre de valeurs nécessaires à **connaître** pour pouvoir résoudre l'équation et connaître toutes les valeurs de la série.

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

(Ca c'est du cours pur, si vous comprenez l'exemple c'est carrée)

Ex : Un élève a eu 4 notes : 12, 15, 16 et une copie perdue (grr) dont il veut connaitre la note.
Il connait sa moyenne de 15.
Donc on fait une petite équation, et hop !
> (12 + 15 + 16 + ?)/4 = 15
> 43 + ? = 60
Sa dernière note est donc 17

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 :

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

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

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'écarts-types une observation particulière est éloignée de la population.

CA C'EST PAR <3

J'insiste ça va vous servir pour les autres cours !!

Pour α = 5% ; ε = 1,96
Pour α = 1% ; ε = 2,60

4) PRECISION DE L'ESTIMATION

IC Large	IC Resserré
Si α↘ alors ε↗ donc l'IC↗	Si α↗ alors ε↘ donc l'IC↘
➔ On a plus de chances que μ soit comprise dans l'IC ➔ Par contre on perd en précision	➔ On a moins de chance que μ soit dans l'IC ➔ Mais on diminue l'IC, on gagne en précision

> 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 , on a un intervalle de confiance plus grand , on a plus de chances que la moyenne soit dedans, (et inversement).

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 donc compris entre **[m + i] et [m - i]**

> Plus la taille de **l'échantillon augmente**, plus la **précision augment**

> Quand **l'indice de précision diminue** la **précision augmente.**

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

n ⇗ , i ⇘ donc l'IC ⇘ donc la précision ⇗

Le **nombre de sujets** nécessaires **«n»,** pour une précision donnée :

$$n = \frac{\varepsilon^2 s^2}{i^2}$$

RECAP DU TURFU :

★ **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**
★ **IC= [m±i]**

DONC :

(encrez moi ça dans vos petites têtes)

★ Si **n** ⇗**, i** ⇘ donc **l'IC** ⇘ donc la **précision** ⇗
★ Si **α** ⇗ **alors ε**⇘ donc **i** ⇘ donc **l'IC se resserre** donc la **précision** ⇗

5) LOI DE GAUSS OU LOI NORMALE

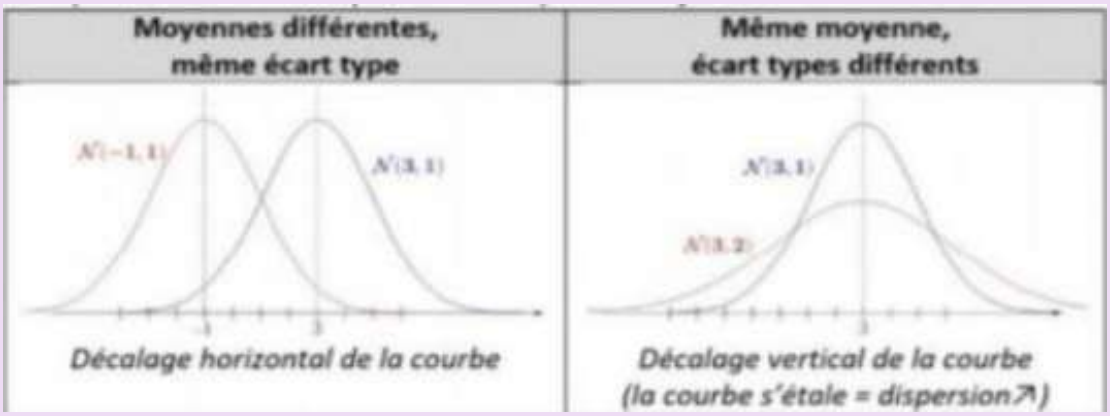
En sciences humaines, on observe souvent des distributions des variables assez symétriques autour de la moyenne : c'est **la courbe de Gauss**

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

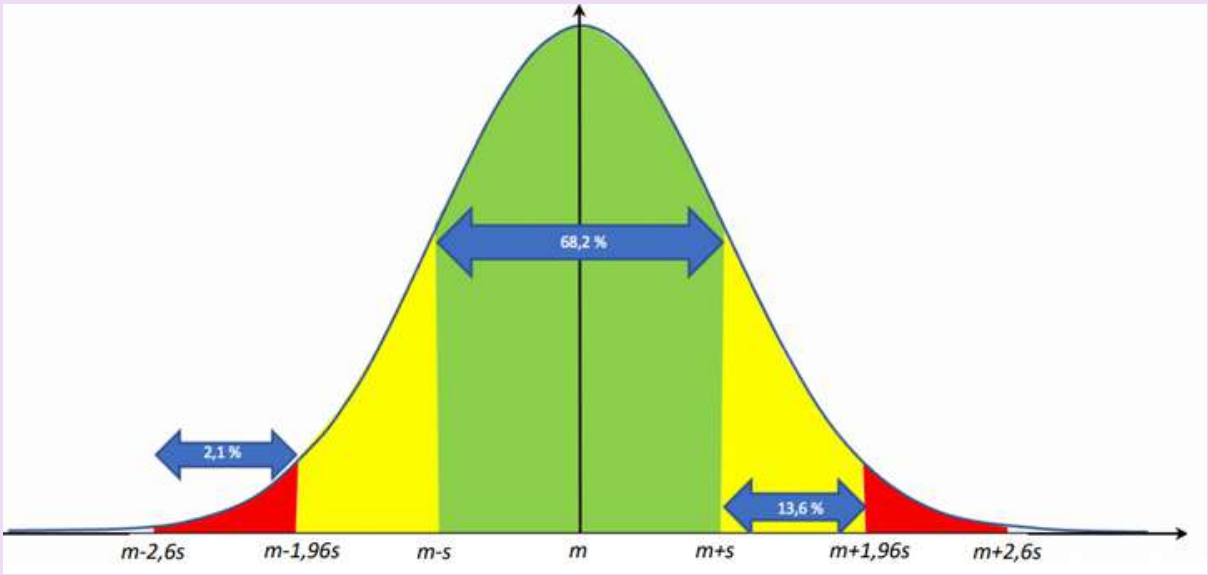
- En abscisse [m ± εs] donc l'IC
- En ordonnée ni : l'effectif pour chaque valeur
- L'aire sous la courbe, le % de la population concerné

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, s), il existe une loi normale de moyenne μ et d'écart-type s notée **N(μ,s)**



A partir de la Loi Normale ou de GAUSS, on précise les intervalles de confiance
[m – 1 s ; m + 1s] contient 68,2% de la population
[m – 1,96 s ; m + 1,96s] contient 95,4% de la population
[m – 2,6 s ; m + 2,6s] contient 99,6% de la population



6) ESTIMATION DES DONNEES QUALITATIVES

ÉCART-TYPE	Il a les mêmes caractéristiques que la variable soit qualitative ou quantitative	$s = \sqrt{pobs.\frac{qobs}{n}}$
INTERVALLE DE CONFIANCE	C'est l'estimation de la moyenne vraie μ à partir de la moyenne m calculée sur l'échantillon	$p \in [pobs \pm \varepsilon s]$
INDICE DE PRECISION "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 :

n ⇗, i ↘ donc l'IC ↘ donc la précision ⇗

Comme avant en gros

> 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 : $n = \varepsilon^2\ pq$ ❖

7) 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 ⇗

TADAAAAAM

C'est finito pour ce cours qui peut paraître compliqué mais les QCMs sont trèeeees abordables (surtout avec les DM que je vais vous sortir hehe), force à vous les potos

Maintenant c'est avec joie que je peux ENFIN faire mes premières dédisssssss : (hors ttr)

Dédi à mon chat, mon soutient émotionnel que j'aime de tout mon coeur,
Dédi à ma famille, à mes parents et ma petite soeur (coeur coeur),
Dédi à mes grands-parents qui seraient méga fières de moi snif
Dédi à mes copines, Héloise, Victoria, Clara, Célia et Mélanight
Dédi à ma meilleure amie Ines, je t'aime fort ma vida loca
Dédi à mes giga copines du tut, Marina votre sage-femme pref, Iris (les 2), Manon mon petit microbe
Dédi à Ramram mon chouchou, Houcine et Yacine mes stars, JP et Mathys gros love
Dédi à mes co tut (ya trop de monde au tut, on souffle)
Dédi à mes vieilles/vieux, Juliete ma canapêche, Madeline mon petit sucre, Camcam, Aymeric, et tous les fossiles (la dynastie est longue..)
Dédi aux autres vieux et surtout à la team biostat/BDR <3
Dédi a la P1 qui a faillit avoir mon âme
Dédi a moi qui vit seule parce que c'est un carnage
Dédi a la P2 parce que c'est trop cool, accrochez-vous ;)
Dédi à mon unique pioux (pour le moment, j'attend vos CV) Aurélia déchire tout je croit en toi

★ET DEDI A VOUS MES CHOUCHOUX ! LA BIOSTAR CROIT EN VOUS★

