Méthode statistique en Médecine

« Il y a trois sortes de mensonges : les petits mensonges, les gros mensonges et les statistiques !" - Benjamin Disraeli → Mal comprise, ou mal utilisée, la statistique peut avoir des conclusions surprenantes, voire absurdes puisqu'on fait dire ce que l'on veut aux chiffres.

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

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 adorent la biostat vont en P2 : 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, ...*

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)

Petit rappel :

- ♥ Variables quantitatives = mesurables avec un instrument de mesure)
- ♥ Variables qualitatives = non mesurables (binaires, nominales...)

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

→ 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.

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

Les Variables :

Type de variable	Caractéristiques	
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

Ex : Le taux de satisfaction d'un client de 1 à 5, 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.

⚠ Une variable pseudo quantitative reste **qualitative** ⚠

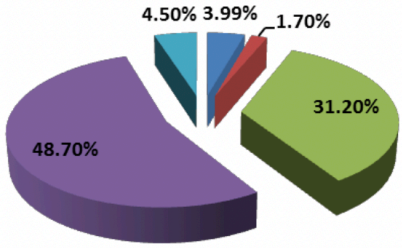
Représentation des variables

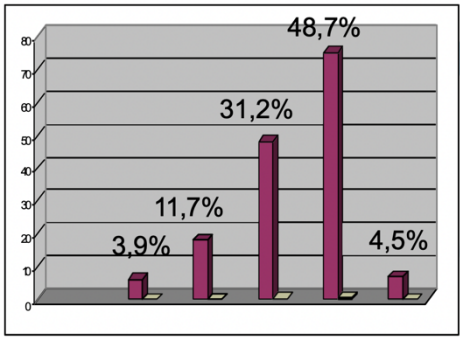
QUALITATIVES :

On les représente sous forme de tableau de % , d'histogramme, de secteurs ...

→ ATTENTION : 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%

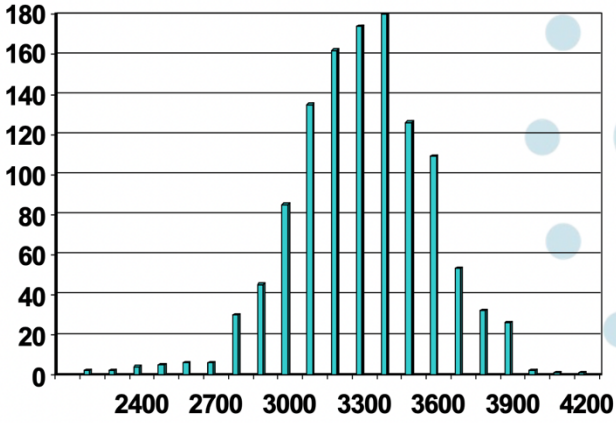




QUANTITATIVES :

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



⇒ Mais aussi sous des formes **Résumées grâce à des paramètres**

Les Paramètres

Paramètres	Caractéristiques	Formule
Moyenne	Variable quantitative discrète :	$m = \frac{\sum x_i}{n}$
	Variable quantitative continue :	$m = \frac{\sum n_i x_i}{n}$
Médiane	Valeur centrale qui sépare la série d'effectif n en 2 sous séries de même effectif	**N pair** : moyenne des $\frac{n}{2}$ et $\frac{n+1}{2}$ valeurs
		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	/

Exemple : les notes de 5 LAS à l'épreuve de biostat : 15, 12, 20, 10, 18

1) *Moyenne : (10+15+12+20+18)/5 = 15*

2) *Médiane → D'abord on classe par **ordre croissant** : 10/12/15/18/20*
*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+10)/2 = 11
25% des PASS seulement ont une note inferieure 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

Variabilité

Comme on l'a dit avant, **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**

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.

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.

Échantillon → 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 : un intervalle de valeurs comprenant la valeur recherchée, c'est **l'Intervalle de Confiance** ou **IC** *(beaucoup plus fiable+++)*

2 estimations **ponctuelles** réalisées sur 2 échantillons donneront des résultats **proches mais différents**
Alors que 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** ⚠

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.

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

- 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 degré de liberté

Ex : Un élève a eu 4 notes : 12, 15, 16 et une copie perdue dont il veut connaitre la note. Il connait 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*

INTERVALLE DE CONFIANCE

- C'est l'estimation de la moyenne vraie μ à partir de la moyenne m calculée sur l'échantillon.
- Avec « n » l'effectif et « s » l'écart-type on a :

FORMULE :

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

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

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 a = **5%** (on a 95% de chance que la moyenne vraie soit dans notre IC)

L'écart-réduit ε : valeur qui dépend du risque a → ils varient en **sens inverse**. Un écart-réduit mesure de combien d'écart-type une observation particulière est éloignée de la population.

Valeurs à connaitre par cœur :
Pour α = 5% → ε = 1,96
Pour 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 ↘) 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.

FORMULE :

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

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

⚠ *Quand l'indice de précision **diminue** la précision **augmente*** ⚠

Le nombre de sujets nécessaires «n» : pour une précision donnée → $n = \frac{\varepsilon^2 s^2}{i^2}$

Petit rappel :

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

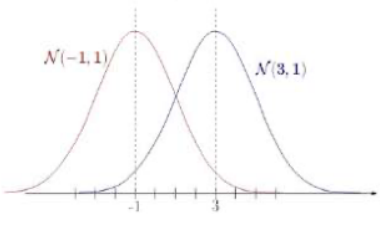
LOI DE GAUSS OU LOI NORMALE

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

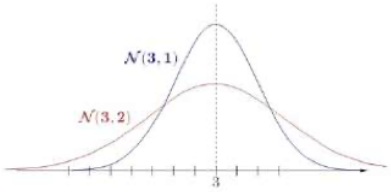
La représentation graphique de données suivant la courbe de Gauss est une courbe en cloche avec :

- 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

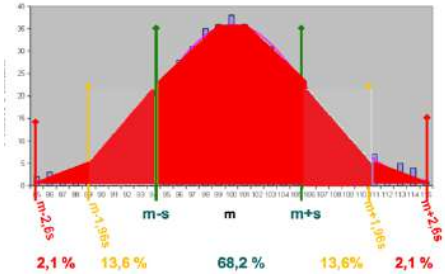
La courbe de Gauss permet de **visualiser l'IC** autour de la moyenne, **l'écart-type**, la dispersion autour de cette valeur moyenne et **la moyenne**.
Pour pouvoir faire des calculs on suppose que notre variable X (quantitative continue) suit une distribution modèle : la **loi Normale**.
Ainsi, pour chaque couple (μ, s), il existe une loi normale de moyenne μ et d'écart-type s notée **N(μ,s)**



Même écart-type, moyennes différentes



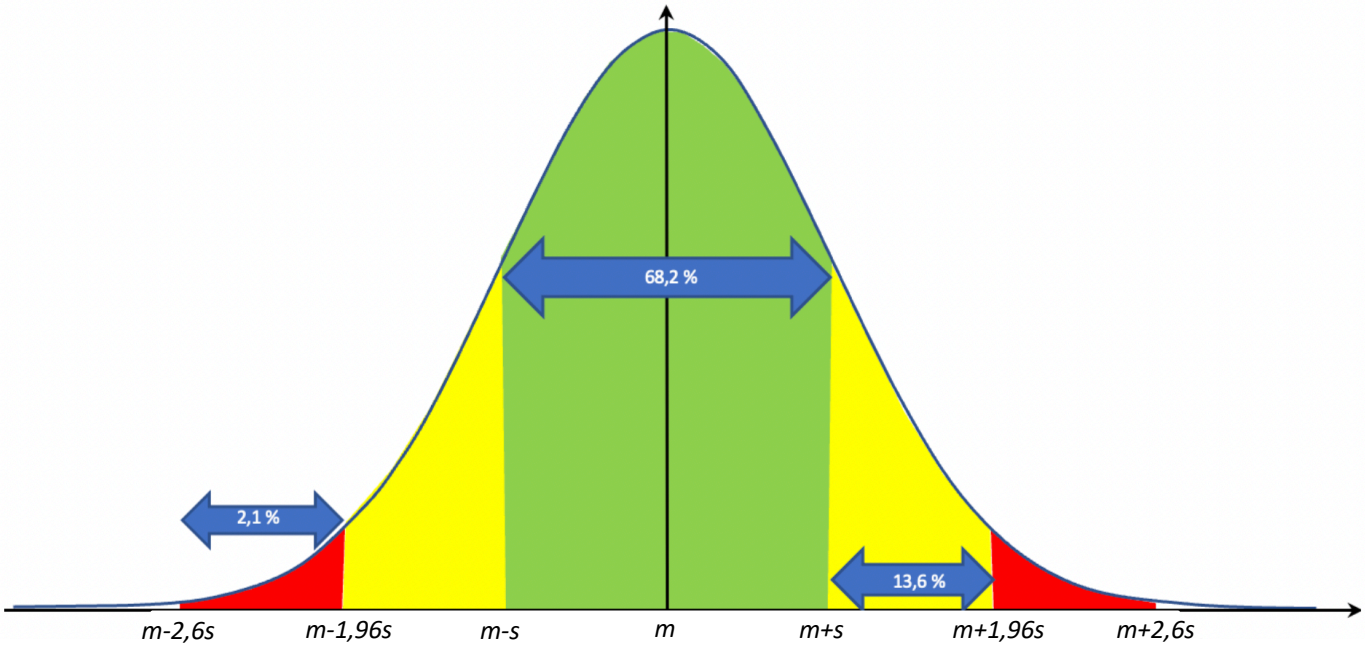
Même moyenne, écarts-types différents (dispersion)



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

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



Apprenez cette petite courbe par cœur <3

Estimation des données qualitatives

ÉCART-TYPE

➢ Il a les **mêmes caractéristiques** que la variable soit **qualitative** ou **quantitative**

FORMULE :

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

avec $qobs = 1 - pobs$

INTERVALLE DE CONFIANCE

➢ C'est l'estimation de la moyenne vraie μ à partir de la moyenne m calculée sur l'échantillon.

FORMULE :

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

PRECISION DE L'ESTIMATION

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

FORMULE :

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

⇨ Si **n est multiplié par 100**, alors **s est divisé par 10** et donc la **précision augmente d'un facteur 10**
⇨ Si n ⇗ , i ⇘ donc l'IC ⇘ donc la précision ⇗
⇨ La précision dépend de la taille de l'échantillon, et de l'écart-type « s ».

Plus la taille de l'échantillon augmente, plus la précision augmente.

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

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 ⇗

Dernière fiche snif snif, place aux dédis :

- ♥ *Dédi à Valentine, t'es dans ta phase de rush, t'arrêtes pas je crois en toi <3*
- ♥ *Dédi à Oskour ce boss*
- ♥ *Dédi à Ray (aka rayane) du discord, t'es un goat*
- ♥ *Dédi à MinAss*
- ♥ *Dédi à Ram- ification et à paparacelse , foncez les boss*
- ♥ *Dédi à tous mes potos las 1/2/3 : Manon, sarah, Bidoli, Ghait, Emma, Elly, Tonystonks aka baptiste*
- ♥ *Dédi à Killian mon incroyable fillot*
- ♥ *Dédi à lou, gabriela et ophélie, vive vous et le SV gang*
- ♥ *Dédi à mes futurs pioux quand même (faites tuts de biostat hihi)*
- ♥ *Dédi à vous tous pour finir, c'est la dernière ligne droite foncez tête baissée et croyez en vous*