Méthode statistique en médecine

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

Définitions :

- **Statistique :** art de collecter, analyser et interpréter des données.

Il en existe 2 types en biostatistiques :

- **Descriptives :** description d'une situation à l'aide de **paramètres**.

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

- **Déductives :** Conclusions à partir d'observations et de mesures : hasard ou explication ?

Ex : on constate que les personnes qui aiment la biostat vont plus souvent 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...

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 d'un 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. On réalise donc 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** +++

Variables et représentation :

Quantitative	Qualitative
Mesurée ou dénombrée Peut être : • Discrète (sans virgule → *âge*) • Continue (avec virgule → *glycémie*)	Non mesurable Peut être : • Binaire (*homme/femme*) • Nominale (*couleur des cheveux*) • Ordinale (*échelle de douleur, stade d'une maladie*) -› pièges

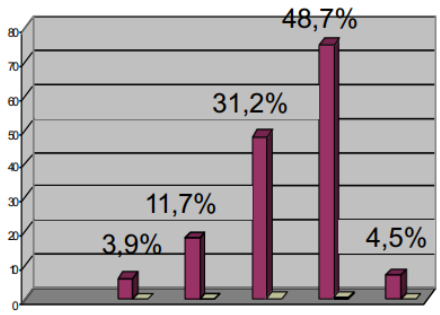
Une variable qualitative ordinale peut être approximée en une variable pseudo quantitative : la variable est qualitative bien qu'on puisse penser qu'elle est quantitative +++

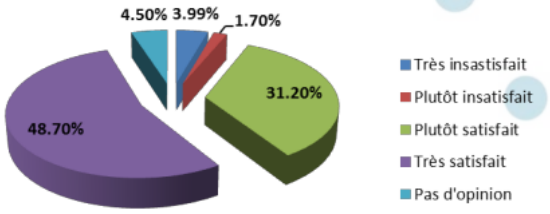
Explication : Si on a une échelle de douleur, de satisfaction... on va les classer de 1 à 10 (c'est pseudo quantitatif), mais cela reste des scores subjectifs qui dépendent d'un ressenti, qui ne sont pas réellement mesurés (donc variable qualitative).

→ Une variable pseudo quantitative reste qualitative !

Représentation : Variables qualitatives → tableau de pourcentages, histogramme, secteurs…

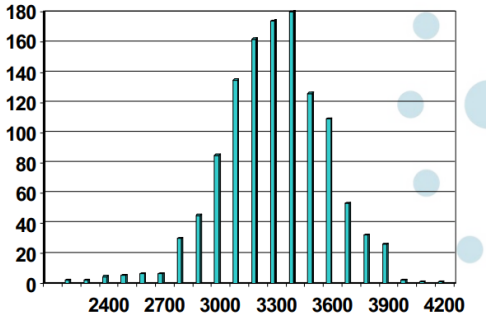
Degré de satisfaction	Nb mères	%
Très insatisfait	6	3,9%
Plutôt insatisfait	18	11,7%
Plutôt satisfait	48	31,2%
Très satisfait	75	48,7%
Pas d'opinion	7	4,5%





Variables quantitatives → tableau, diagramme en bâton ou histogramme

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



Paramètres :

Pour cette partie j'ai préféré vous mettre directement la partie du prof car je trouve que les exemples, schémas et explications sont très bien. Je ne pense pas que l'expression algébrique de la moyenne soit à connaitre par cœur, mais c'est important de savoir la calculer. Si vous avez des questions → forum

Moyenne : cas d'une variable quantitative discrète $m = \frac{\sum_{1}^{n} x_i}{n}$

cas d'une variable quantitative continue $m = \frac{\sum_{1}^{n} n_i x_i}{n}$

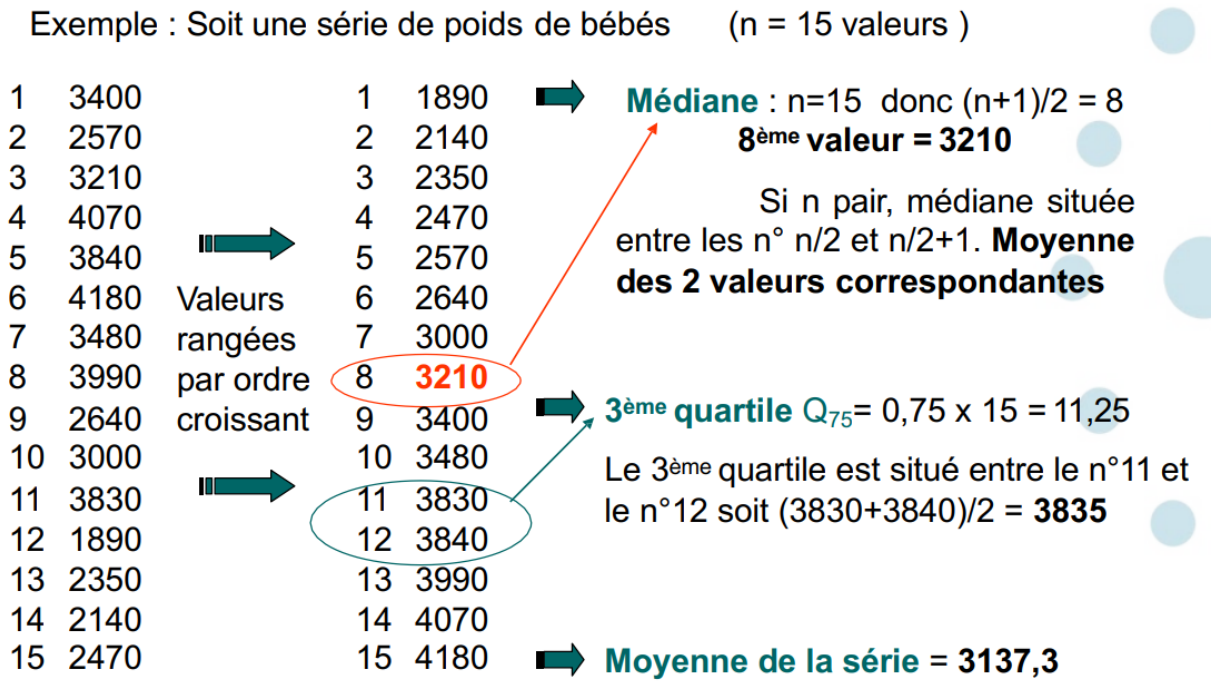
Variance : paramètre indiquant la dispersion des données autour de la moyenne.

Médiane : valeur centrale si rangées par ordre croissant,

50% des valeurs < médiane et 50% des valeurs > médiane

(Exemple : {3,4,**6**,8,10}

Quartiles : Les quartiles partagent la série ordonnée en 4 groupes de même effectif (Exemple Q_{25} = 1er quartile : 25% de la série est < à cette valeur.)



D'autres très bons exemples proposés par Charlotte la tutrice de l'an dernier :

Enoncé : Les notes des LAS en biostat au premier EB : {10, 7, 15, 20, 2}
Calculer la : moyenne, médiane, quartiles :

MOYENNE : (10 + 7 + 15 + 20 + 2) / 5 = 10,8

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 × 5 = 1,25 avec 5 le nombre de valeurs
2) donc Q1 se trouve entre la 1e et la 2e valeur
3) Q1 = (2+7) / 2 = 4,5
4) 25% des LAS seulement ont une note inférieure à 4,5

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

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

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.

On retrouve 2 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**)

→ L'estimation par intervalle est moins précise mais plus juste

Méthodologie pour estimer des données quantitatives :

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 estime la **moyenne**.

- 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
Ex : A l'épreuve de biostat 3 étudiants ont eu 0, 10 et 20, la moyenne est 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. (mis en application dans le cours stats déductives)

- Intervalle de confiance (IC) : C'est l'estimation de la moyenne vraie μ à partir de la moyenne m calculée sur l'échantillon. L'IC est aussi appelé **intervalle au risque α.**
On donne un intervalle auquel μ appartient :

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

Si vous ne comprenez pas certaines lettres, regardez le schéma en bas de la page 5.

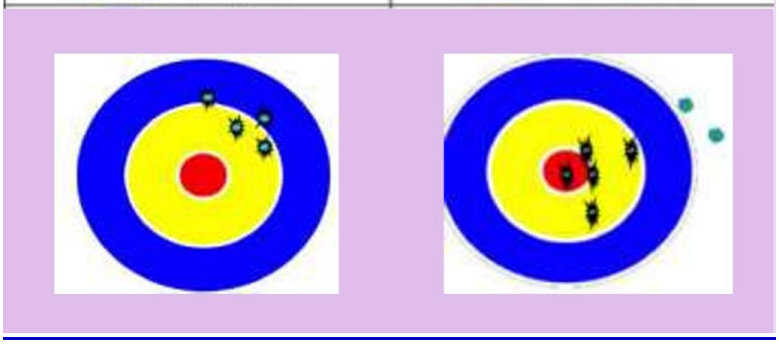
- Risque α : C'est le risque d'erreur dans l'estimation de μ (le risque que notre IC ne contienne pas μ). On prend en général un risque α = 5% (on a 95% de chance que la moyenne vraie soit dans notre IC)

- Ecart 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.

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

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

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

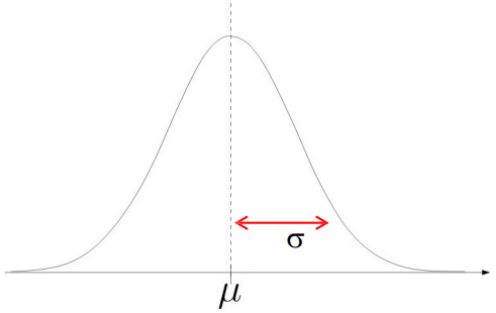
Loi de Gauss ou loi normale :

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

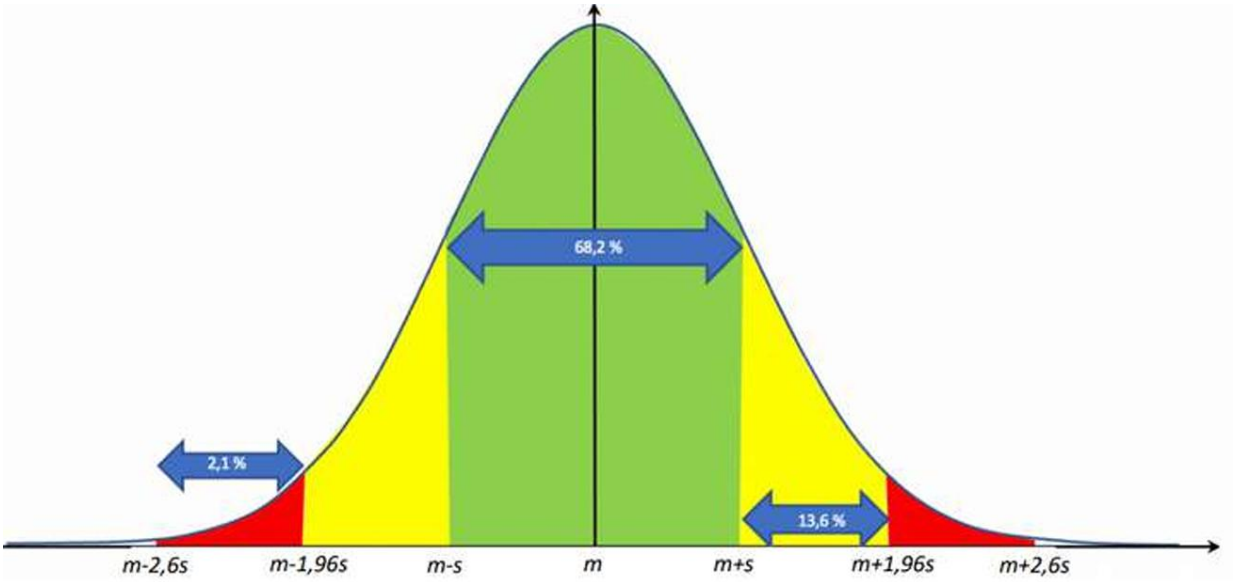
La courbe de Gauss permet de **visualiser l'IC** autour de la moyenne, **l'écart-type**, la dispersion autour de cette valeur moyenne et la moyenne.

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



A partir de la **loi normale** (= loi 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



Estimation de données 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 : si n ⇗, i ⇘ donc l'IC ⇘ donc la précision ⇗
La précision dépend de la taille de l'échantillon, et de l'écart-type « s ».

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 ⇗