La méthode statistique en médecine

Dans ce cours, s'il y a des formules il ne faut pas les savoir, le Pr BENOLIEL n'aime pas mettre des calculs au CC. Cependant il faut connaître la formule de l'IC (je mets des ++ à côté)

Les biostatistiques (= statistiques appliquées au domaine de la santé publique) ont 3 objectifs :

1. Descripton d'une populaton par rapport à une maladie
2. Evaluation des traitements, des techniques, des coûts
3. Mise en place des observations épidémiologiques et en tirer des conclusions

Les biostatistiques doivent être capables de décider si une observation peut être due au seul **hasard** ou si elle a une autre explication.

I. DEFINITIONS

Statistique : Art de **collecter**, **d'analyser**, et **d'interpréter** des données. Lorsqu'elle est appliquée au domaine de la **biologie**, on parle de biostatistique.
Il en existe 2 types :

- **Descriptive** : Description d'une situation à l'aide de **paramètres**.
- **Déductive** : une observation est-elle due au **hasard** ? Ou existe-t-il une **explication** ?

Ex : descriptive : Collecte de 2 données sur la population française : taille et couleur des yeux
déductive : On constate que les sujets ayant une taille > 1,70m ont tous les yeux bleus. Hasard ?

Données : Résultat de **l'observation** d'un individu, par l'utilisation d'un **instrument de mesure**, ou par les sens de l'observateur (signes cliniques, biologiques, ..)

Le but d'une donnée est de l'observer ou 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** : intra sujet (comparaison de deux sujets) ou inter sujet (comparaison du sujet à lui-même)
Ex : Taille, Poids, groupe sanguin, température corporelle

Paramètre : Grandeur apportant une **information résumée** (ou synthétisée) sur la variable étudiée.
Ex : moyenne d'une série de valeurs

Série statistique : Collection d'objets de **même nature**, avec des **caractéristiques différentes** d'un objet à l'autre (variables).
Ex : Les hommes et les femmes sont des objets de même nature mais avec des caractéristiques différentes (mince, gros, blond, brun ...)

Variable quantitative : mesurable, obtenu grâce à un appareil de mesure
Ex : la taille d'un individu

Variable qualitative : non mesurable
Ex : la couleur des yeux

Population : Série exhaustive de **tous** les individus étudiés, sur lesquels on veut appliquer (inférer) des décisions
Ex : population de la France

Echantillon : 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)** ++++
Notre échantillon est alors connu, alors que la population est inconnue.
Ex : 10 personnes tirées au sort dans la population française.

II. LES TYPES DE VARIABLES

Il existe deux types de variables ++++

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

Remarque : 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 ++

*Ex : le rang/classement au concours → ce sont des chiffres mais ils n'ont **pas de signification** et ne peuvent **pas faire l'objet d'opérations arithmétiques** (calculer la somme ou la moyenne des classements n'a pas de sens)*
Cette variable est donc qualitative ordinale mais comme on la représente par des chiffres, on dit qu'elle est pseudo-quantitative.

III. REPRESENTATION DES VARIABLES

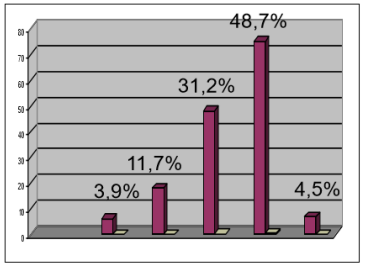
1. Variables qualitatives

On peut les représenter de **2** manières :

- **Diagramme en bâton** ou **histogramme**
- **Tableau**

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



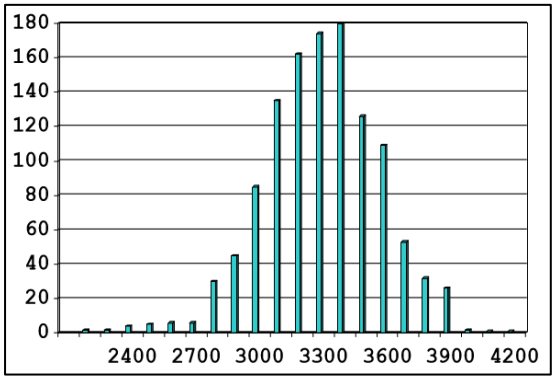
2. Variables quantitatives

On peut les représenter de **3** manières :

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

Ex : poids des nouveaux né dans la maternité

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



IV. PARAMETRES

On peut « résumer » en quelques paramètres les caractéristiques de la série de **données quantitatives**.

Moyenne :

- Variable quantitative discrète : $\mathbf{m = \Sigma x_i \ / \ n}$
- Variable quantitative continue : $\mathbf{m = \sum n_i x_i \ / \ n}$

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.**

Ex : On relève les notes de biostat de 6 PACES : 13, 14, 19, 13, 10, 18

1) *Moyenne : (12 + 13x2 + 14 + 18 + 19) / 6 = 14,5*

2) *Médiane : On remet les valeurs dans l'ordre croissant {12, 13, 13, 14, 18, 19}*
(n/ 2) + 1 = 6/2 + 1 = 4 → La 4è valeur correspond à 14
n/2 = 3 → La 3è valeur correspond à 13
La médiane correspond donc à la moyenne de 13 et 14 = ***13,5***

3) *1er quartile : 0,25 x 6 = 1,5 → Q1 se trouve entre la 1ère et la 2è valeur*
Soit Q1 = (12 + 13) / 2 = ***12,5***
25% des PACES ont une note inférieure à 12,5

Propriétés sur la moyenne et la médiane

	Avantages	Inconvénients
Moyenne	- **Facile à calculer** - Se **manipule** facilement dans les tests statistiques (très utilisée) = adaptée aux calculs statistiques - Très **significative** si la répartition des données est assez **symétrique** et la **dispersion faible**	- **Sensible** aux valeurs anormales (mini ou maxi)
Médiane	- **Calcul facile**, **peu sensible** aux valeurs anormales - Utilisable pour les **valeurs ordinales**, les classes, …	- Se prête moins aux calculs statistiques

Statistiques descriptives

I. NOTION DE VARIABILITE

Toute donnée biologique possède une variabilité. Sa connaissance est indispensable pour pouvoir dire si la valeur d'une variable quantitative est **normale** ou **anormale**.

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

Ex : La valeur moyenne de la glycémie (variable quantitative) chez un sujet normal est de ***1g/L +/- 0,25 g/L***
Ceci signifie qu'une glycémie appartenant à l'intervalle : ***[0,75 g/L ; 1,25 g/L]*** *est* ***normale*** *= variabilité (on peut prendre plusieurs valeurs et toujours être normal)*

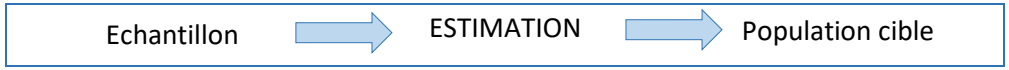
Ainsi un patient avec un glycémie de 1,2g/L (=valeur de la variable) a une glycémie normale alors que celui avec 0,7g/L a une glycémie infra-normale.

II. LES ESTIMATIONS EN STATISTIQUES

1. Définition

En biostatistiques, les études sont réalisées sur un **échantillon** **représentatif** de la population (=**échantillonnage**). A l'issue de ces études, se pose le problème de la légitimité des résultats et de leur possible **extrapolation** à l'ensemble de la **population**. Pour cela, on réalise une **estimation du résultat vrai**, à partir des résultats obtenus sur l'échantillon.

➔ On **détermine donc des paramètres au niveau d'une population à partir d'observations réalisées sur un échantillon de cette population** = estimation ++



Il existe 2 types d'estimations :

- **L'estimation ponctuelle** : valeur **unique** jugée la meilleure à l'instant t (peu fiable)
- **L'estimation par intervalle** : **intervalle** de valeurs contenant la valeur recherchée. Cet intervalle est nommé **intervalle de confiance (= IC)**. (Beaucoup + fiable)

NB : Deux estimations ponctuelles (respectivement. par intervalle) d'une même variable réalisées sur 2 échantillons A et B donneront des valeurs ponctuelles voisines (resp des IC se recouvrant), mais pas nécessairement la même valeur (resp le même IC)

Exemple
Soit un groupe de 220 patients, représentatif d'une population rhumatismale (R). On observe 167 cas de rhumatismes inflammatoires. Quel pourcentage de rhumatismes inflammatoires dans la population R?

*1) Estimation ponctuelle : p=167/220 = 0,76 soit **76%***
2) Estimation par intervalle : a = 5%, donc calcul de $IC_{0,95}$
p = 0,76 donc q = 0,24

$IC_{0,95} = 0,76 \pm 1,96\sqrt{\frac{0,76 * 0,24}{220}}$ = ***[0,70 ; 0,82]***

(voir plus bas pour calculer un intervalle)

➔ L'estimation par intervalle est **moins précise**. Mais si l'on refait ce calcul sur un autre échantillon, cette nouvelle estimation **recouvrira la première**. Ce ne sera pas forcément vrai avec l'estimation ponctuelle.

2. L'estimation des données quantitatives : estimation de la moyenne

Méthodologie
1. Détermination précise de la population étudiée = **population cible**
2. **Tirage au sort (TAS)** de n sujets ++++ (c'est le TAS qui rend l'échantillon représentatif++)
3. Calcul de l'intervalle de confiance

L'estimation assure donc la correspondance entre ce qui se passe au niveau de l'échantillon et au niveau de la population.

ESTIMATION

Paramètre	Echantillon	Population
Moyenne	**m** = estimateur de la moyenne vraie μ au niveau de l'échantillon	**μ** = moyenne vraie
Ecart type	**s** = estimateur de l'écart type vrai σ au niveau de l'échantillon	**σ** = écart type vrai
Effectif	**n**	**N**

a) L'écart type

Il mesure la **dispersion d'un ensemble de données autour de la moyenne**. Il s'agit donc de la variabilité des mesures entre elles et par rapport à la moyenne. Plus il est **faible**, plus le caractère étudié est **homogène** (+ les données sont proches de la moyenne) *et vice versa*.

b) Les degrés de liberté

On a : m=moyenne ; xi=valeurs dont on veut faire la moyenne ; n=effectif ; (xi-m)=écarts

- ⇨ Il y a n écarts
- ⇨ Il y a (n-1) écarts indépendants à la moyenne = nombre de degrés de liberté

La somme des n écarts **vaut 0** : $\sum(\boldsymbol{xi} - \boldsymbol{m}) = \sum \boldsymbol{xi} - \boldsymbol{n}.\boldsymbol{m} = \mathbf{0}$

$= (x1 - m) + (x2 - m) + \cdots + (xi - m)$

Dans l'équation qu'on a posée si on connait la moyenne **m**, l'effectif **n** et toutes les valeurs que prend **x** sauf une, on se retrouve avec une équation avec une seule inconnue. On pourra donc résoudre l'équation et retrouver cette dernière valeur qu'il nous manque.
Il y a donc n-1 (x_i) indépendants et donc **n-1 degrés de liberté**.

Rq : On dit qu'il y a « n – 1 » degré de liberté car il suffit de connaitre n-1 valeurs pour connaitre toutes les valeurs de la série.

Exemple : Paul a eu trois notes mais sur une de ses évaluations il y a de la moutarde sur la note il sait qu'il a eu 12 et 13 et il connait sa moyenne 14.
m=14 ; x={12,13,} ; n=3
Il peut donc avec n-1 valeurs, cad 2 valeurs, trouver la troisième, Il y a 2 techniques :

1) *Avec la moyenne*

$moyenne = 14 = \frac{12+13+x}{3}$

soit $x = \ 14 * 3 - (12 + 13) = 17.$

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.

2) *Avec les écarts*

Notes	12	13	x
xi-m	-2	-1	x - 14

$\sum(\boldsymbol{xi} - \boldsymbol{m}) = \mathbf{0}$

-2 – 1 + (x – 14) = 0
x – 14 = 3
x = 17

c) Intervalle de confiance ++++

L'IC est **l'estimation de la moyenne vraie μ** à partir de la moyenne m calculée sur l'échantillon *(on suppose que la vraie moyenne se trouve dans cet intervalle déterminé à partir de m)* :

$$\mu \in [m +/- \frac{\varepsilon s}{\sqrt{n}}]$$

- L'IC est aussi appelé **intervalle au risque α** avec **α** le **risque d'erreur** dans l'estimation de μ : *risque que l'IC ne contienne pas la vraie valeur de la moyenne μ.* On prend en général **α=5%** *(Il y a 95% de chance que la moyenne μ appartienne à l'IC)*

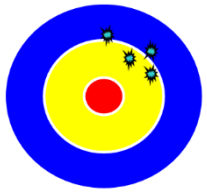
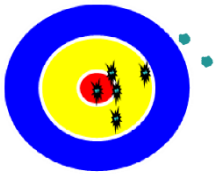
- **ε** représente **l'écart réduit**. La valeur d'ε dépend de la valeur d'α (ε et α varient en **sens inverse**).

α = 5% ε = 1,96
α = 1% ε = 2,6

Remarque : + la taille de l'échantillon augmente, + l'estimation tend vers la moyenne vraie

d) Précision de l'estimation

- Les **variations de α** conditionnent la **précision de l'estimation** et la **largeur de l'IC**. On peut voir l'IC comme une cible :

IC large	IC resserré
Si α↘ alors ε↗ donc l'IC↗ : ➔ On a + de chances que la **moyenne μ soit dans l'IC** ➔ On risque de rater la bonne estimation = **mauvaise précision de l'estimation**	Si α↗ alors ε↘ donc l'IC↘ : ➔ La **bonne valeur de μ pourra être ratée** ➔ La **précision augmente**
 *La précision est mauvaise parce que les flèches ne sont pas au centre mais il n'y a pas de valeurs qui ne sont pas dans l'IC*	 *La précision est meilleure puisque les flèches sont + proches du centre mais les points verts ne sont pas dans l'IC*

- **Indice de précision** : Cet indice permet de calculer la **précision de l'estimation** de μ. Cette valeur représente la **largeur de l'IC**.

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

Selon la formule de l'IC vu plus haut, l'IC correspond donc à ***[m +/- i]***

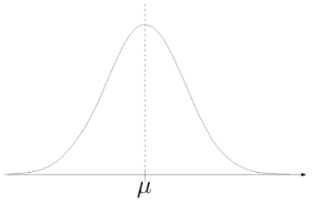
Il faut savoir manipuler cette formule +++

- ❖ n↗ alors i↘
- ❖ i↘ alors IC↘ = précision↗

Plus la taille n de l'échantillon augmente, et plus la précision augmente : c'est à dire que l'estimation tend de plus en plus vers la valeur vraie.

- **Nombre de sujets nécessaires** pour une précision donnée : $n = \varepsilon^2 \frac{s^2}{i^2}$

e) Loi de Gauss ou loi Normale



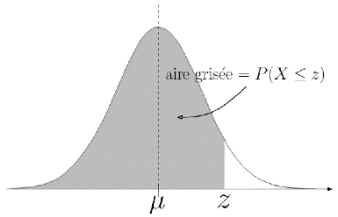
En sciences humaines, on observe souvent des distributions des variables plutôt **symétriques autour de la moyenne** avec une forme de **cloche** : c'est la **courbe de Gauss**.

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

- en abscisse : m +/- εs, donc l'IC
- en ordonnée : n
- l'aire sous la courbe : le % de la population concernée

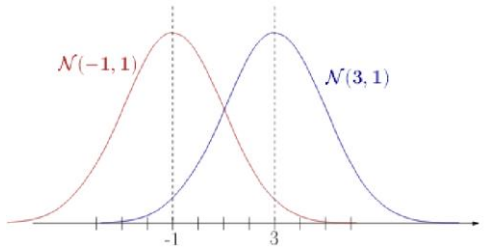
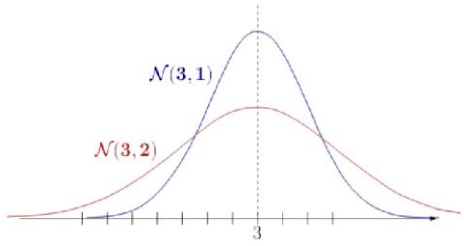
La loi de Gauss est une loi qui permet de visualiser :

- la notion **d'IC autour de la moyenne**
- la notion **d'écart type**
- la notion de **dispersion** autour de cette valeur moyenne
- la **moyenne**

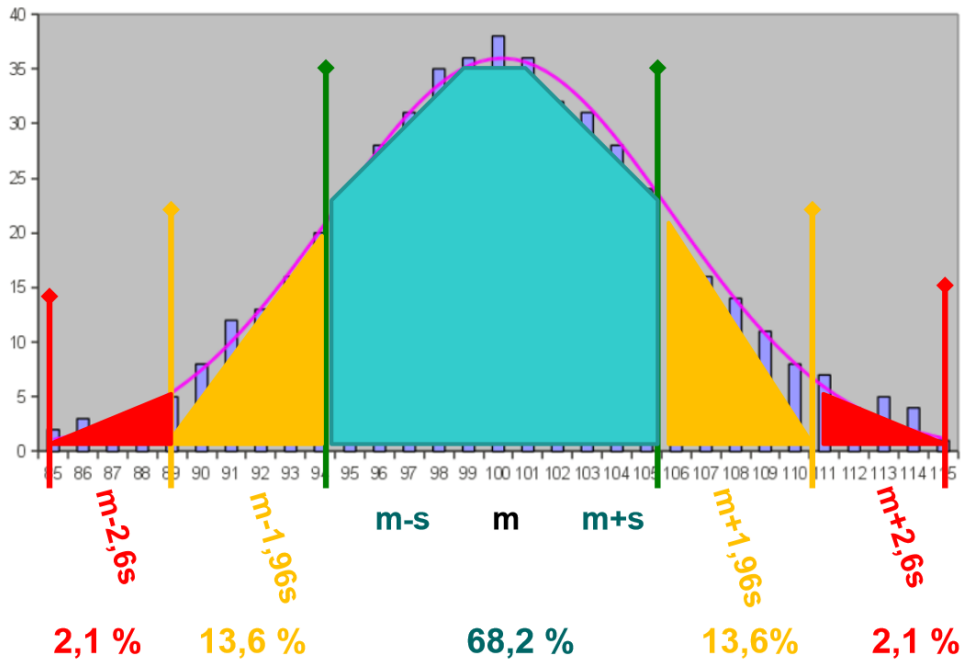


Pour pouvoir faire des calculs on va supposer que notre variable X (quantitative continue) suit une distribution « modèle » : la **loi Normale**. Ainsi, Pour chaque (μ,σ) il existe une loi normale de moyenne μ et d'écart type σ : on la note **N (μ, σ)**.

Exemples de lois normales pour bien comprendre le fonctionnement de la courbe

Moyennes différentes, même écart type	Même moyenne, écart types différents
	
Décalage horizontal de la courbe	*Décalage vertical de la courbe (la courbe s'étale = dispersion↗)*

A partir de la loi Normale ou loi de Gauss on peut trouver des IC ++++



[m - 1 s ; m + 1 s] contient 68,2% de la population
[m - 1,96 s ; m + 1,96 s] contient 95,4% de la population (95%)
[m - 2,6 s ; m + 2,6 s] contient 99,6% de la population (99%)

Exemple :
Sur un échantillon de 114 personnes tirées au sort, on a calculé le taux moyen de cholestérol et l'écart type au niveau de cet échantillon :
m = 195,4 cg
s = 45,6 cg

L'IC à 95% est : [195,4 ± (1,96 x 45,6 / $\sqrt{114}$]

3. L'estimation des données qualitatives : estimation d'un pourcentage

Dans l'estimation de données **qualitatives**, on s'intéresse au pourcentage d'individus d'une population présentant un caractère donné A. *(on ne s'intéresse plus à la moyenne)*

Méthodologie
1. Constitution d'un échantillon représentatif par **TAS**
2. Calcul du pourcentage p_{obs} de l'échantillon présentant le 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**

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

	Echantillon	Population
Proportion (%)	$p_{obs} = p_{observé}$ = estimateur du pourcentage inconnu p	p = pourcentage vrai
Ecart type	s = estimateur de l'écart type vrai σ au niveau de l'échantillon	σ = écart type vrai
Effectif	n	N

a) L'écart type

Il a les **mêmes caractéristiques** pour une variable qualitative que pour une variable quantitative.

Il est donné par $s = \sqrt{\frac{p(obs) * q(obs)}{n}}$ avec q(obs) = 1 – p(obs)

b) Intervalle de confiance

$$p \in [p_{obs} +/- \varepsilon s]$$

Les valeur d'α et de ε sont les mêmes que pour les variables quantitatives, ainsi que leurs variations.

Pour faire une **bonne estimation** il faut +++
- Un **échantillon représentatif** donc constitué par **TAS**
- Pas de biais lors de la sélection
- Un **IC** qui **accompagne TOUJOURS l'estimation** (il montre la variabilité des données)
- Une **taille importante de l'échantillon** : si n↗ alors la précision↗

c) Précision de l'estimation

- **Indice de précision** : il représente la largeur de l'IC

$$i = \varepsilon \sqrt{\frac{p(1-p)}{n}} = \varepsilon s$$

Il faut savoir manipuler cette formule aussi **+++**
- ❖ n↗ alors i↘
- ❖ i↘ alors IC↘ = précision↗

La précision dépend de la **taille de l'échantillon n**. Plus l'effectif de l'échantillon est grand et plus la précision sera bonne. Cependant, elle dépend au de l'**écart type s**.

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

- **Nombre de sujets nécessaires** : $n = \varepsilon^2 \frac{pq}{i^2}$

d) Le sondage

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