La méthode statistique en médecine

Hello, le cours est basé sur ce qui a été fait l'année dernière, il est donc à jour pour les PACES mais pas encore pour les PASS. Je vous conseille quand même de le bosser, ça m'étonnerait qu'il change. Dites moi si la fiche vous plait. Bonne lecture !

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 des traitements, des techniques, des coûts
➳ Mise en place des observations épidémiologiques, conclusions
Les biostatistiques doivent être capables de décider si une observation peut être due au **hasard** ou si elle a **une autre explication**.

I) Définitions

Statistique : art de collecter, analyser et 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 : l'observation est-elle due au hasard ? Existe-t-il une autre 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 supérieure à 1m70 ont les yeux bleus. Est-ce dû au 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 de 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 : inter sujet (comparaison de 2 sujets) ou intra 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*

Variable quantitative : **mesurable+++**, obtenue grâce à un **appareil de mesure** *Ex : 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 du 💗💗 tirage au sort = randomisation 💗💗** L'échantillon est connu, alors que la population est inconnue.
Ex : 10 personnes tirées au sort dans la population française

II) Les types de variables

Il existe 2 types de variables 💗💗:

Variables qualitatives	**Variables quantitatives**
Binaires : homme/femme Nominales : couleur des yeux Ordinales : douleur articulaire	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 ++++

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. Cette variable est donc qualitative mais comme on la représente par des chiffres on dit qu'elle est pseudo quantitative.

III) Représentation des variables

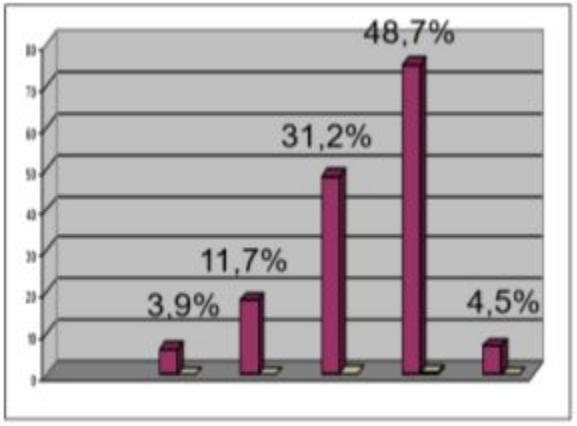
A) Variables qualitatives :

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)

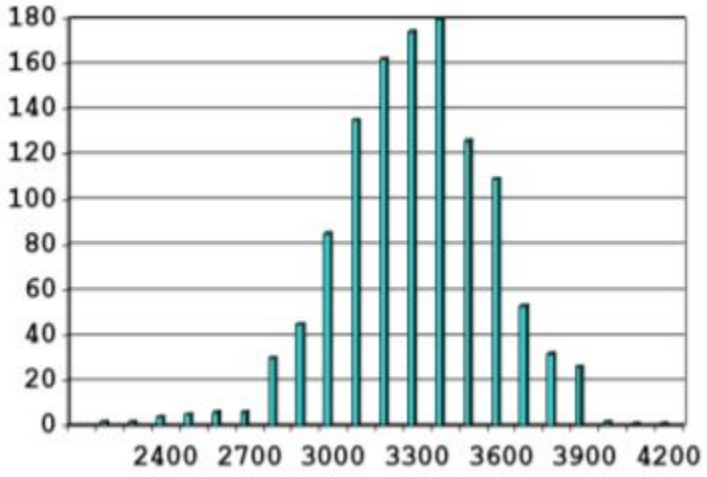
B) Variables quantitatives :

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



IV) Paramètres

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

- Moyenne :

-Variable quantitative discrète : $m = \Sigma x_i / n$

-Variable quantitative continue : $m = \Sigma n_i x_i / n$

- 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 : les notes de 7 PACES à l'épreuve d'UE4 : 15/12/13/9/18/15/10
1) moyenne : m = (15*2 + 12 + 13 + 9 + 18 + 15 + 10) / 7 = 13,1
2) médiane : on classe par ordre croissant : 9 ; 10 ; 12 ; 13 ; 15 ; 15 ; 18 7 notes → impair donc on prend la valeur (n+1) / 2 = (7+1)/2 = 8/2 = 4 C'est donc la 4ème valeur donc médiane = 13
3) 1er quartile : 0,25 * 7 = 1,75 → Q1 se trouve entre la 1ère et la 2ème valeur Donc Q1 = (9+10)/2 = 9,5 25% des PACES ont une note inférieure à 9,5.

Propriétés sur la moyenne et la médiane : 💗💗💗

	Avantages	**Inconvénients**
Moyenne	♥ Facile à calculer ♥ Se manipule facilement dans les tests statistiques = adaptée aux calculs statistiques ♥ Très significative si la répartition des données est assez symétrique et la dispersion faible	♠ Sensible aux valeurs anormales (max et min)
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 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**.

On retrouve deux types d'estimations :

➳ 1. **L'estimation ponctuelle** : valeur unique jugée la meilleure à l'instant t (peu fiable+++).

➳ 2. **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** 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

a) 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 (plus les valeurs sont proches de la moyenne).

b) 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

Les 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 (le bouff). 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.

c) Intervalle de Confiance :

L'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.

$$\mu \in [m \mp \varepsilon s/\sqrt{n}\,]$$

L'IC est aussi appelé intervalle au risque α. **Le risque α c'est le risque d'erreur dans l'estimation de μ+++** (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 représente 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

d) Précision de l'estimation :

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

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
 *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*

Ici on visualise l'intervalle de confiance comme une cible

Explications : 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.

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

D'après la formule de l'IC vue juste avant, l'IC est compris : entre $[m - \varepsilon s/\sqrt{n}]$ et $[m + \varepsilon s/\sqrt{n}]$ donc 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 : $n = \varepsilon^2 s^2/i^2$

Récap :

❤ 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 (distribution des données autour de la moyenne).
❤ Les variations du risque **α** déterminent la précision de l'estimation.
❤ **i** représente la largeur de l'IC : $i = \varepsilon s/\sqrt{n}$
❤ IC = [m ∓ 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. Si c'est toujours pas capté envoyez moi un mp.

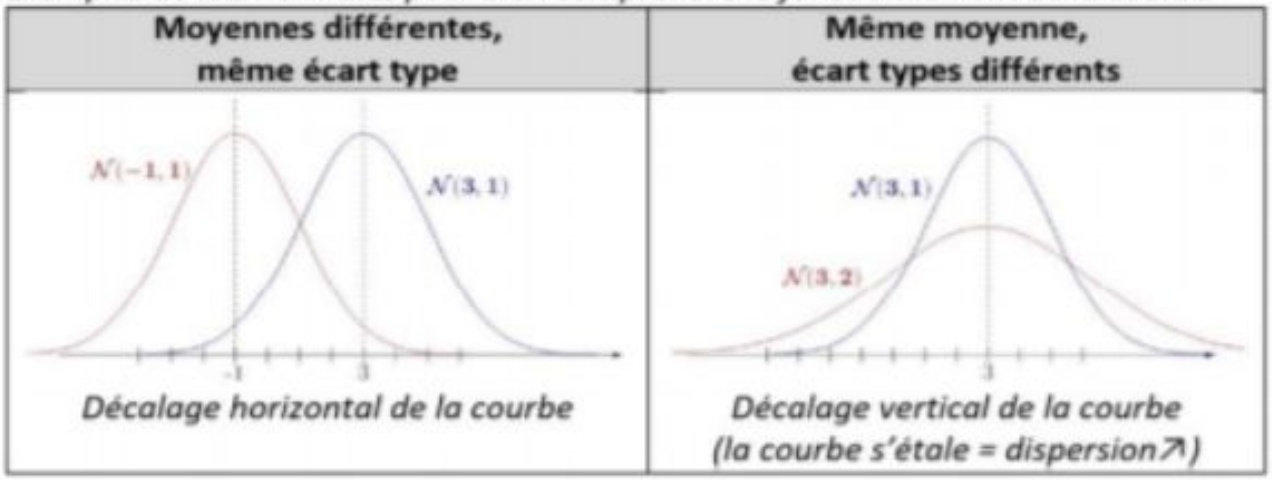
e) Loi de Gauss ou loi Normale :

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 :

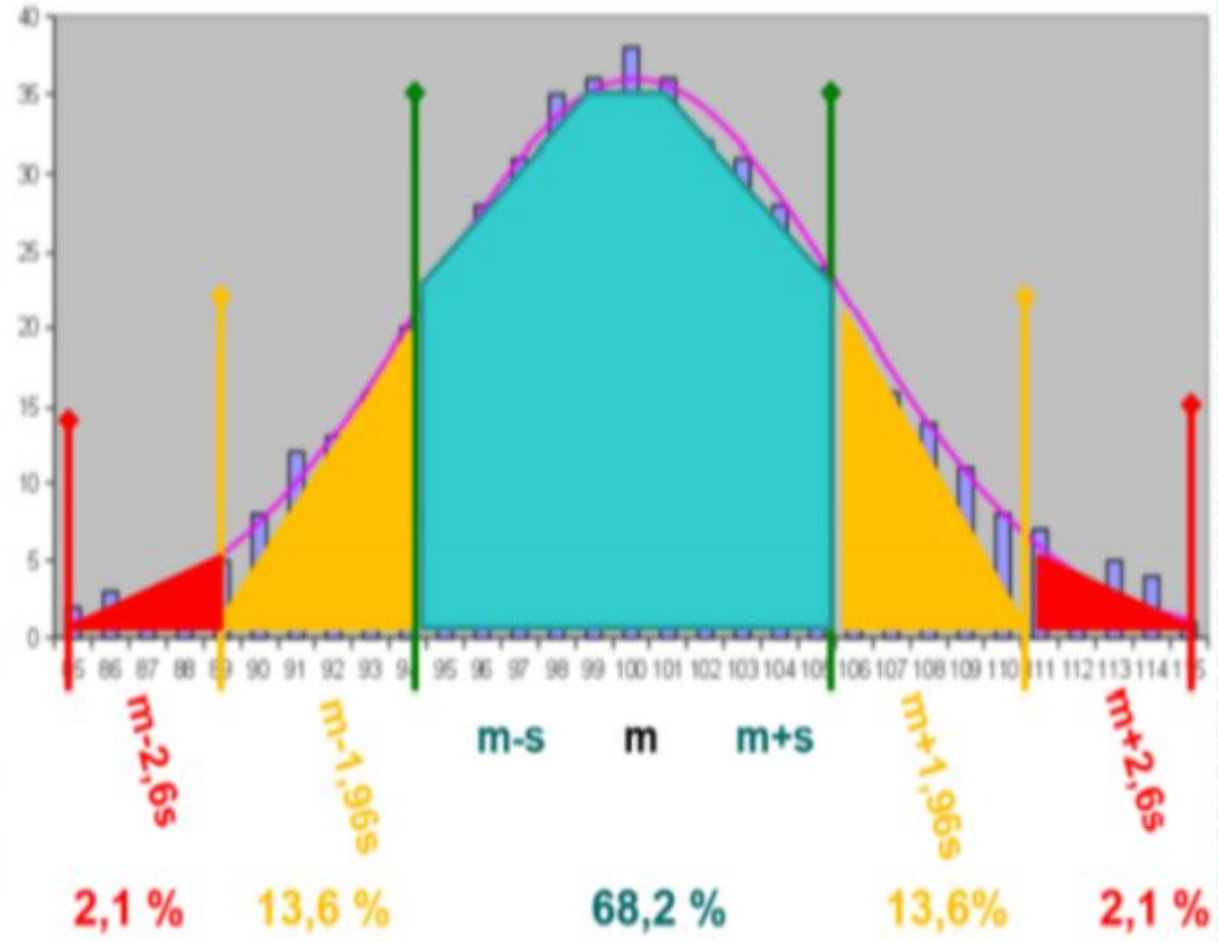
- En abscisse : [m +/- εs], donc **l'IC**
- En ordonnée : **n**
- L'aire sous la courbe : **le % de la population concernée**

La loi de Gauss 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 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



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%)

C) Estimation des données qualitatives :

Méthodologie :
1) Constitution d'un échantillon représentatif par TAS
2) Calcul du pourcentage $pobs$ 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 données qualitatives, on va estimer un pourcentage !**💗

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 même procédé que pour les données quantitatives, seuls changeront les paramètres utilisés et donc les formules qui en découlent.

a) Écart-type :

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

Il est donné par $s = \sqrt{(pobs.qobs/n)}$ avec $qobs = 1 - pobs$

b) Intervalle de Confiance :

L'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.

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

c) Précision de l'estimation :

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

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

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 : $n = \varepsilon^2 pq\, i^2$

d) 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 :

➸ *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 ↗*

Et voilà pour ce cours très dense et assez compliqué, j'espère que vous êtes toujours vivants (évidemment puisque vous êtes des boss). Ce cours est une "intro" au prochain cours Statistiques déductives qu'on ne vous a pas fait à la tut rentrée pour pas vous surcharger. Bossez-le bien il y a plein de notions importantes pour la Biostat et SURTOUT, comprenez le avant de l'apprendre. Je reste toujours disponible si vous avez des questions !

“Un héros n’est pas celui qui ne tombe jamais. C’est celui qui se lève, encore et encore, sans jamais perdre la vue de ses rêves.”BON COURAGE POUR CETTE ANNÉE !!!! ❤️❤️