STATISTIQUES DESCRIPTIVES

BIENVENUE DANS LE NIVEAU 1 DU MONDE DES STATS ! VOTRE OBJECTIF : ARRIVER À LA FIN DE CE COURS. VOUS CROISEREZ TOUT PLEIN DE PERSONNAGES SUR VOTRE TRAJET, BON COURAGE !

START

Et coucou les pouss1 ! On se retrouve aujourd'hui pour un petit cours avec des notions que vous connaissez déjà alors pas d'impasse ! Comme vous avez pu le remarquer, vous serez plongé dans le thème des jeux vidéos. J'espère que la mise en page vous plaira ! Sur ce, la partie peut commencer ! (mes remarques seront de cette couleur).

Introduction 

La **biostatistique** nous permet d'appliquer des **théories statistiques** au domaine du vivant : le domaine de la **santé publique** permet de décrire l'état de santé de la population. Pour cela, il faut évaluer les traitements, les techniques, les coûts et mettre en place des **observations épidémiologiques** pour en tirer des conclusions. On procédera d'abord à une analyse **descriptive** puis à une analyse **déductive** (vous verrez ça avec ma cotut incr).

QUELQUES DEFINITIONS :

° **Statistique** : art de collecter et d'interpréter des 'données'

- statistique **descriptive** : description d'une situation à l'aide de **paramètres** (ex: on peut calculer la **moyenne** des étudiants à l'épreuve de biostatistiques).
- statistique **déductive** : conclusions à partir d'**observations** et de **mesures** → ce que l'on observe est-il dû au **hasard ou pas** ? (ex : 2 traitements anti cancéreux donnent à 5 ans une survie de 42% pour l'un et de 48 % pour l'autre. Hasard ou efficacité plus grande pour l'un des deux ?)

° **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,..). Cette donnée n'est intéressante que si on peut l'observer/la **comparer** sur plusieurs individus. Elle ne sera **pas strictement équivalente** d'un individu à l'autre. On parle donc de **variable** (ex : la taille, le poids, le groupe sanguin...). Cette variabilité peut être due au **hasard** ou être **physiologique** (intra ou inter sujets).

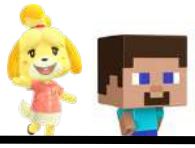
° **Population** : série **exhaustive** de tous les individus étudiés, sur lesquels on veut inférer des décisions (ex: population de France, les étudiants en médecine de Nice...)

° **Echantillon** : ensemble **fini** et d'effectif **limité**, extrait de la population

Pourquoi échantillonner ?

→ car la population est **inaccessible** dans son entièreté pour des raisons d'organisation et de **moyens limités**.
→ on procède à une étude sur **l'échantillon** et on regarde si on peut appliquer ces résultats à la **population**.
→ l'échantillon doit être **représentatif** de la population : on procède alors à un **tirage au sort** (randomisation)
→ on se retrouve alors avec un **échantillon connu** et une **population inconnue** ++

Quelques variables



Qualitative binaire	sexe...
Qualitative nominale	couleur des yeux...
Qualitative ordinale	consommation de tabac (faible 0-9, moyenne 10-19, forte > 19), douleur articulaire (absente, modérée, intense), rang dans un classement...
Quantitative discrète (nombre entier)	âge, nombre d'otites dans la dernière année...
Quantitative continue (nombre à virgule, plus précis)	déficit auditif moyen (ex :11,5 dB), note de biostatistiques arrondi au dixième...

Une variable **qualitative ordinale** peut être considérée comme variable '**pseudo quantitative**' pour certains tests ++ (ex : variation de la douleur sur une échelle de 1 à 3, taux de satisfaction de 1 à 5....) ⚠️ Les **nombres** affectés aux modalités qualitatives ne peuvent **pas** faire l'objet d'**opérations arithmétiques** (logique si sur une échelle de 1 à 10, ma douleur se place à 5, ce chiffre donne un ordre d'idées de l'intensité de la douleur, mais ne veut rien dire en tant que tel).

Représentation de variables quantitatives

Exemple : on s'intéresse aux poids des nouveaux nés dans une maternité :

a) Echantillon : **TAS** des mères ayant accouché dans cette maternité pendant une période donnée (effectif n=1165)

b) Variable étudiée : **poids** du nouveau né --> **variables quantitatives**

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

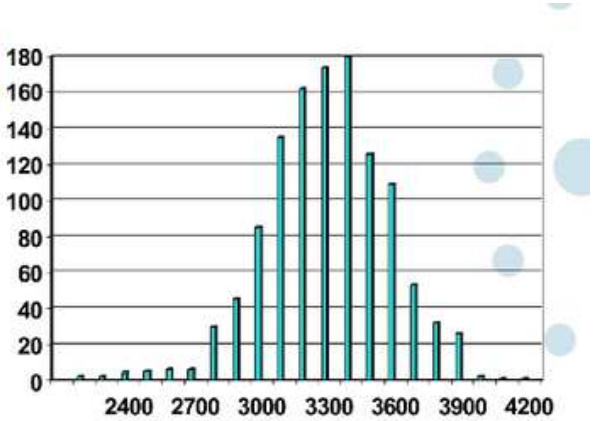


Tableau ou histogramme des effectifs
Mais pas seulement .....

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

° **Moyenne** :

- dans le cas d'une variable quantitative **discrète** (valeur entière)

$$m = \frac{\sum_{1}^{n} x_i}{n}$$

- dans le cas d'une variable quantitative **continue** (valeur à virgule)

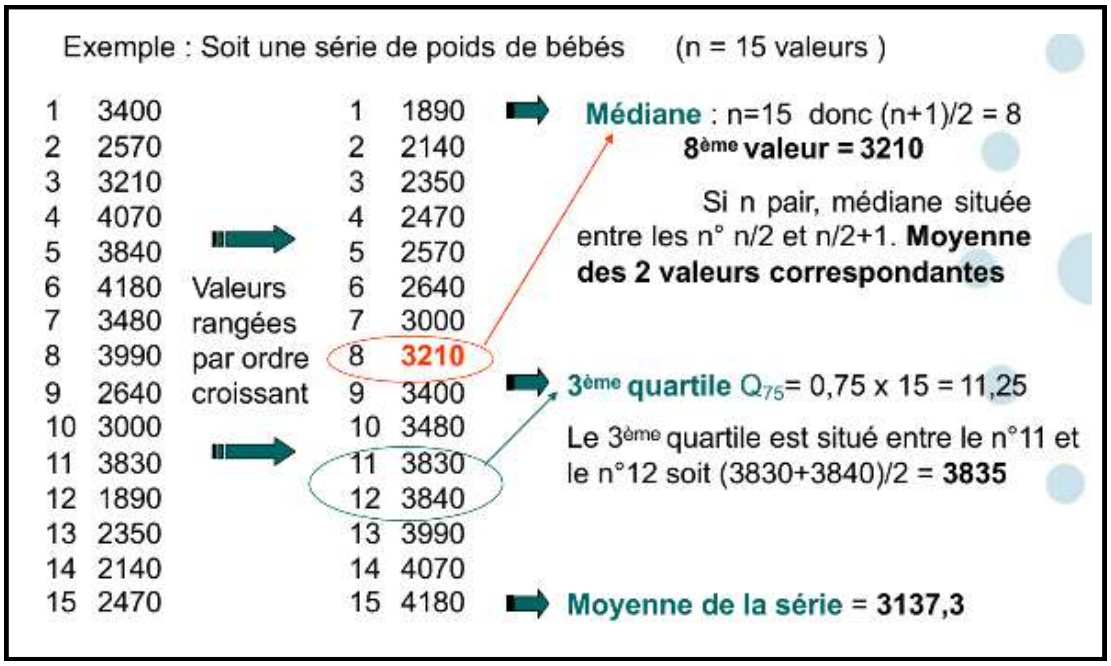
$$m = \frac{\sum_{1}^{n} n_i x_i}{n}$$

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

° **Médiane** : valeur **centrale** si les données sont rangées en ordre **croissant** (attention aux pièges les gars) → 50% des valeurs < médiane ainsi que 50% des valeurs >médiane (ex : {3,5,12,18,21}.

° **Quartiles** : ils partagent la série en 4 groupes de **4 groupes de même effectif** (série en ordre croissant) ex : 1e quartile = 25% de la série est < à cette valeur.

Je vous mets ici la diapo du prof, c'est très bien expliqué et vous avez pas mal de paramètres qui sont calculés : si vous avez des questions, se sera sur le forum :)



Estimation statistique

Le **problème** est le suivant : déterminer un **paramètre** au niveau d'une **population** à partir d'observations réalisées sur un **échantillon** de cette population. Après l'étude, on réfléchit à la **légitimité** des résultats et à leur **extrapolation** à la population. Exemple, comment connaître la durée de séjour moyenne des patients hospitalisés en France, pour une pathologie donnée ?

Echantillon
effectif = n
moyenne = m
écart type = s

→ ESTIMATION →

Population cible
effectif = N
moyenne = μ
écart type = σ

Il y a 2 types d'estimation :

ESTIMATION PONCTUELLE

= valeur jugée la meilleure à un **instant t** (peu fiable)

ESTIMATION PAR INTERVALLE

= **intervalle** de valeurs contenant la valeur recherchée (beaucoup plus fiable)

--> L'ESTIMATION PAR INTERVALLE EST MOINS PRÉCISE MAIS PLUS JUSTE ++

° **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 :

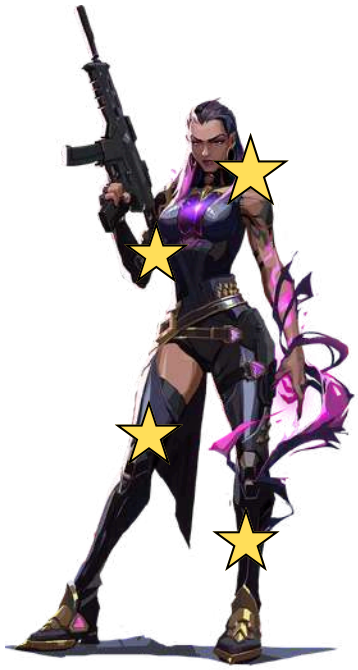
$$\mu \in \left[m \pm \frac{\varepsilon s}{\sqrt{n}} \right]$$ --> INTERVALLE AU RISQUE α

° **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 un risque **α = 5%** (on a 95% de chance que la moyenne vraie soit dans notre IC). Plus **α** est **petit**, plus l'intervalle de **confiance** est **grand** : on réussit plus souvent. On s'expose aussi au risque de **rater** la "bonne" estimation.

° **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.

Précision de l'estimation

L'intervalle de confiance peut être vu comme une **cible** (ici les centres des cibles seront les têtes de Reyna et de Killjoy pour les connaisseurs ;).

IC large	IC resserré
→ plus de chances **d'atteindre** la cible, mais on a une **mauvaise précision** de l'estimation (au lieu de tirer dans la tête, on tire sur son corps donc moins précis)	→ on a un **risque de rater**, certaines balles seront à l'extérieur, mais on a une **meilleure précision** de l'estimation (donc plus de balles toucheront la tête)
	

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**.

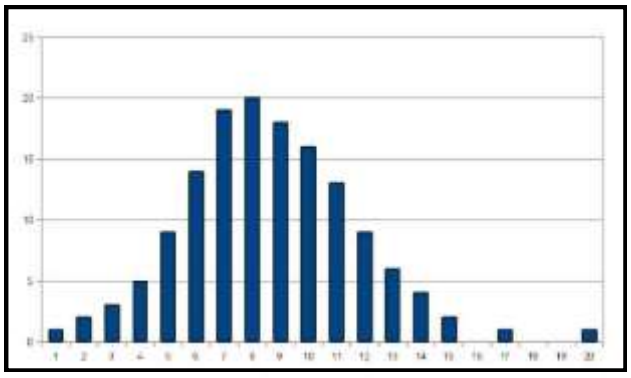
D'après la formule de l'indice de précision :
Quand n ⇗ , i ⇘ donc l'IC ⇘ donc la précision ⇗ +++

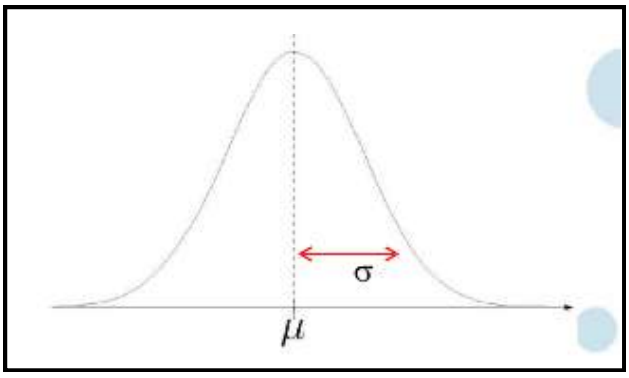
Loi de Gauss ou loi normale

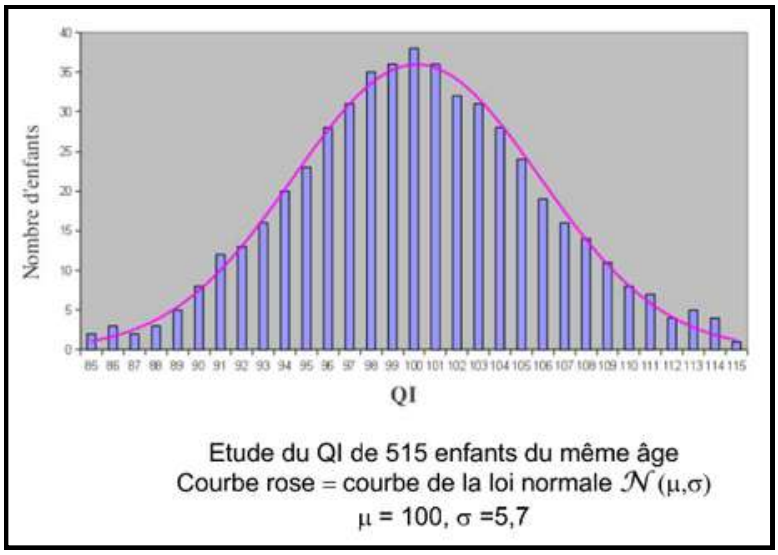
En **sciences humaines** on observe souvent des **distributions** (X) plutôt symétriques autour de la moyenne avec une forme de **cloche** pour pouvoir faire des calculs, on va supposer que X suit une distribution « modèle », pour des variables quantitatives continues : **la Loi Normale.**

- En **abscisse** [m ± εs] donc l'IC
- En **ordonnée** : 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)

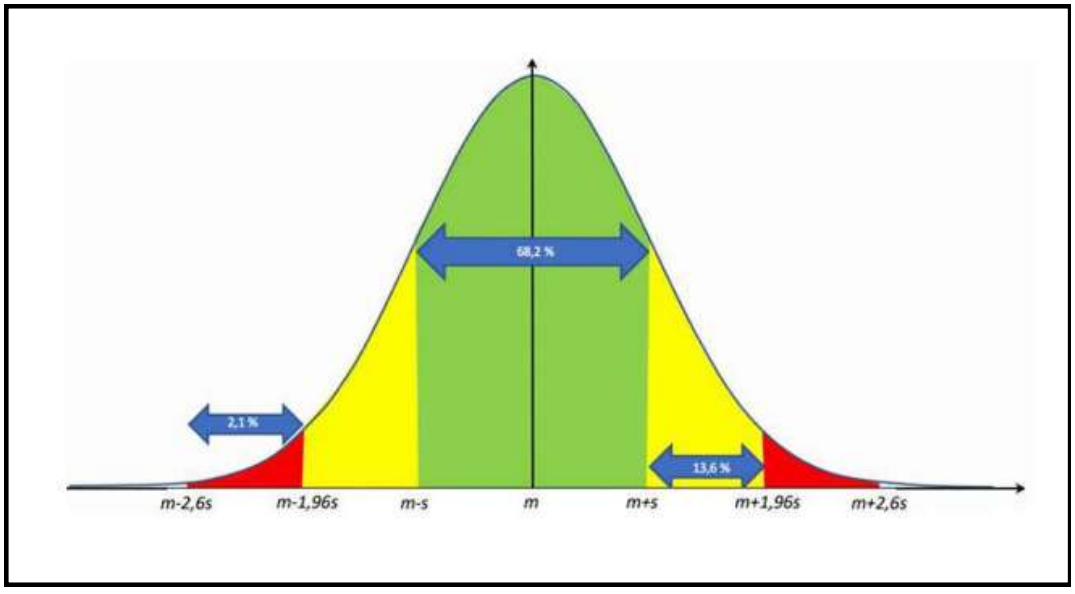






A partir de la loi normale (= loi de Gauss), on précise les intervalles de confiance :

- [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 ++



TUT RECAP :

effectif n augmente → IC se resserre → précision augmente

Contre exemple :

Un chirurgien écrit à 1000 de ses patients afin de connaitre leurs suites chirurgicales et sur 100 réponses : 75 vont très bien, 25 ont des séquelles handicapantes. Or, il y a eu 900 non-réponses. On ne peut pas préjuger de l'état de ces 900 patients. Ils sont peut être décédés des suites opératoires, ou bien très mécontents du chirurgien, ou tout au contraire sont très satisfaits et ne jugent pas utile de répondre. Cet échantillon est **BIAISÉ**.

RECAP :

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

Et voilà les loulous le cours version TTR est terminé ! Je sortirai très rapidement la fiche complète car il manque des infos mais vous avez vu le plus gros ! Prenez le temps de bien comprendre, entraînez-vous bien, posez des questions fin des trucs qu'on vous rabâche depuis le début finalement 🙂 Bonne continuation et ne lâchez jamais rien les bg !!