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).

° **Paramètre** : grandeur apportant une **information résumée** sur la variable étudiée (médiane, quartile, moyenne...)

° **Série statistique** : collection d'objets de **même nature**, avec des caractéristiques différentes d'un objet à l'autre.

° Une variable **quantitative** est mesurable grâce à des **instruments de mesure**, contrairement à une variable **qualitative** qui est **non mesurable** (nominale...)

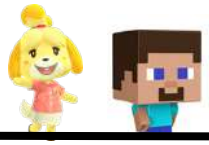
° **Population** : série **exhaustive** de tous les individus étudiés, sur lesquels on veut inférer des décisions (ex: population de France, les étudiants en médecine de Nice...)

° **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).

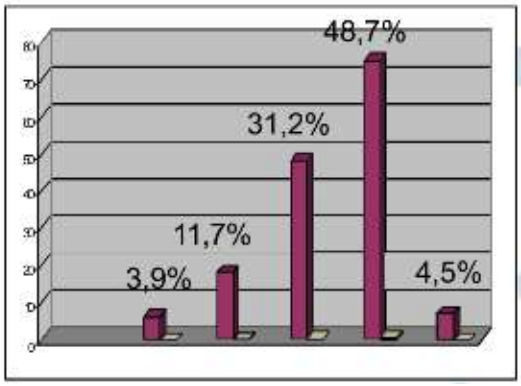
Représentation de variables qualitatives

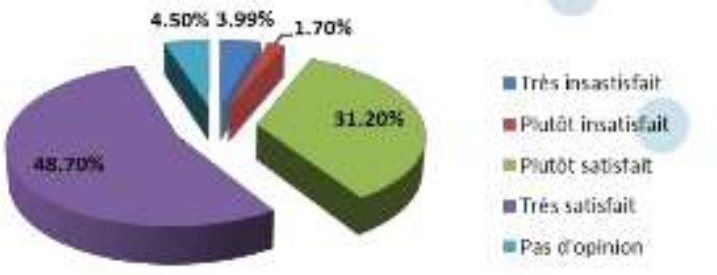
Exemple : Quel est le degré de satisfaction des mères accouchant dans une certaine maternité ?

a) On procède à un **tirage au sort** (TAS) pour former un **échantillon** avec les mères ayant accouché dans cette maternité sur une période donnée (effectif n=154).

b) La variable étudiée est le **degré de satisfaction** (très insatisfait, insatisfait, plutôt satisfait, très satisfait, pas d'opinion) → **variables qualitatives ordinales** : on peu les représenter grâce à des tableaux de pourcentages, des histogrammes, en secteur...

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%





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

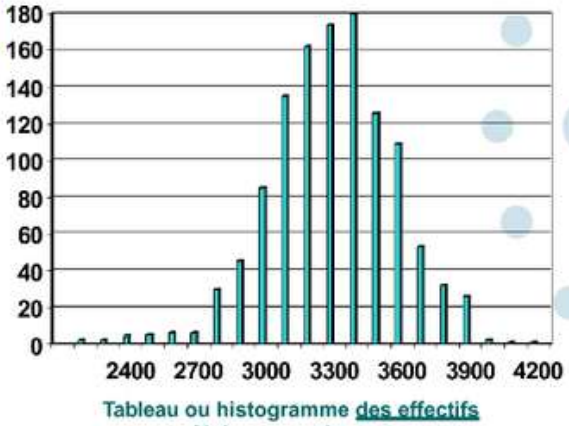


Tableau ou histogramme des effectifs
Mais pas seulement .....

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)

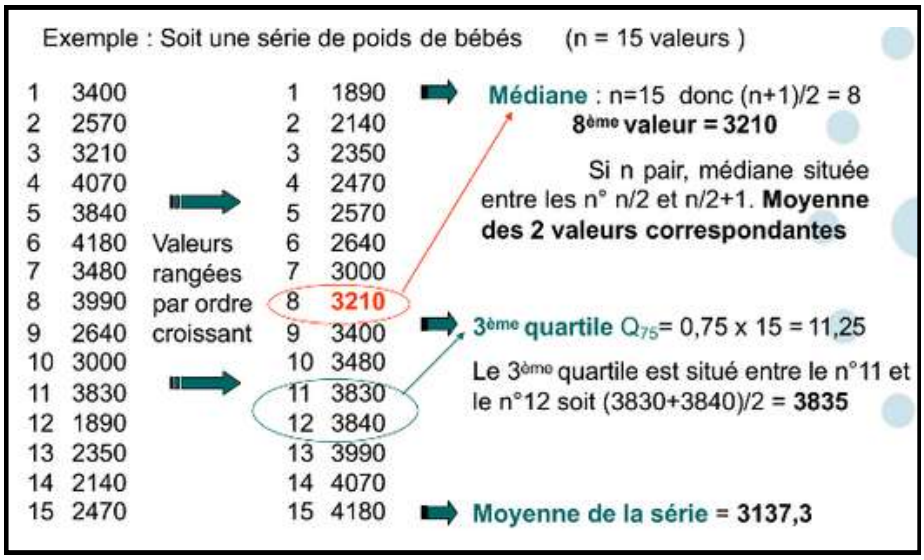
$$m = \frac{\sum_{1}^{n} n_i x_i}{n}$$

° **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 :)



Comparaison entre la moyenne et la médiane

	Avantages	Inconvénients
MOYENNE	° simple à calculer ° facile à manipuler donc adaptée aux calculs statistiques ° très significative si la répartition est assez symétrique avec une faible dispersion	° sensible aux valeurs anormales (minimum et maximum)
MEDIANE	° calcul facile ° peu sensible aux valeurs anormales ° utilisable pour des valeurs ordinales, des classes...	° se prête moins aux calculs statistiques

Explicatut : que veut dire concrètement être sensible aux valeurs anormales ?

° Alors petit exemple pour que tu comprennes : on imagine que les notes à l'examen de biostatistiques soient les suivantes : {10, 11, 12, 14}
Calculons la moyenne : 10+11+12+14/4=11,75
Maintenant imaginons que l'on rajoute une note et que l'élève ait eu 0, on a alors : 10+11+12+14+0/5=9,4
On remarque que la moyenne a chuté, c'est pour cela que la moyenne est sensible aux valeurs anormales.

° Pour la médiane, se sera différent : comme c'est un paramètre de position, que l'on rajoute ou pas une valeur, la médiane sera sensiblement la même. On reprend notre exemple de toute à l'heure avec les notes suivantes : 10-11-12-14. La médiane est entre la 2e et la 3e valeur, c'est à dire 11,5. Si un élève à 0, on aura : 0-10-11-12-14, la médiane est de 11, malgré le 0 de l'élève → la médiane est alors peu sensible aux valeurs anormales.
J'espère que cela vous aidera !

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**.

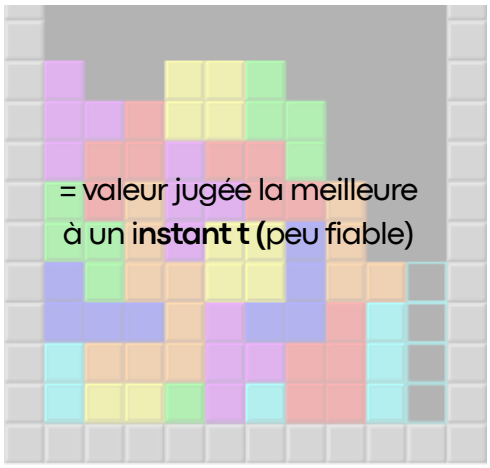
Exemple : en moyenne, le taux de sucre sans le sang chez les sujets normaux est de 1g/L. Un patient avec une glycémie de 1,2 g/L est une valeur normale ou anormale ? Pour le savoir, il faut savoir la variabilité normale de la glycémie, liée aux variabilités individuelles.

Estimation statistique

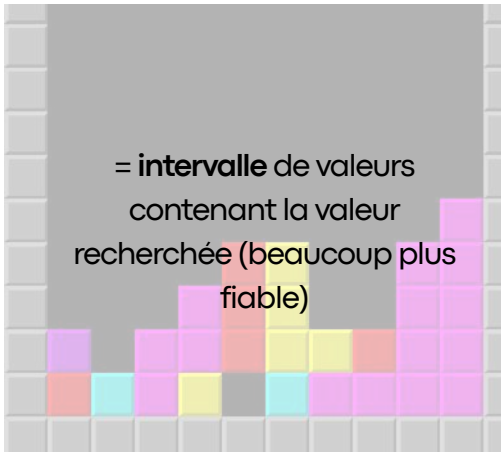
Le **problème** est le suivant : déterminer un **paramètre** au niveau d'une **population** à partir d'observations réalisées sur un **échantillon** de cette population. Après l'étude, on réfléchit à la **légitimité** des résultats et à leur **extrapolation** à la population. Exemple, comment connaître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

Il y a 2 types d'estimation :

ESTIMATION PONCTUELLE



ESTIMATION PAR INTERVALLE



--> 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** de n sujets pour avoir un échantillon représentatif
3- Calcul de l'intervalle de confiance **IC** autour de la valeur inconnue du paramètre
→ 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.
Exemple : 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 avec Claudia)

° **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.

Remarques :
Si on a plusieurs **échantillons**, on aura plusieurs **estimations.**
Si la taille de l'échantillon **augmente**, l'estimation tend vers la **moyenne vraie m**.

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 ⇗ +++

Le nombre de sujets nécessaires «n», pour une précision donnée :

$$n = \frac{\varepsilon^2 s^2}{i^2}$$

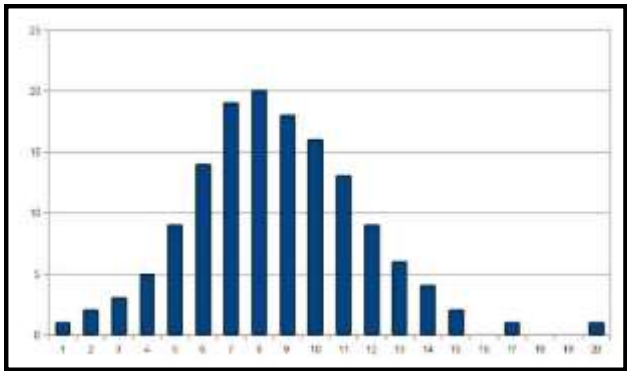
Honnètement la 2e formule ne tombe pas en tant que tel, comprenez surtout cette idée d'indice de précision i.

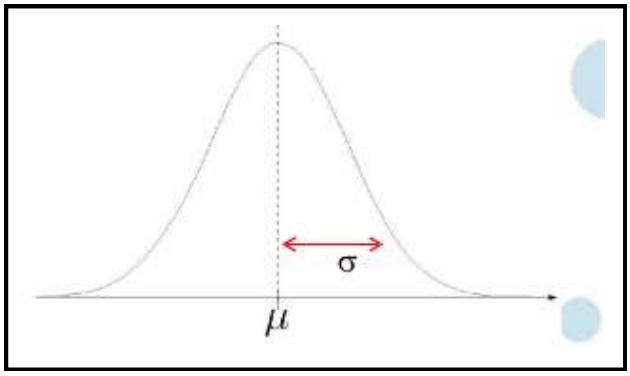
Loi de Gauss ou loi normale

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.**

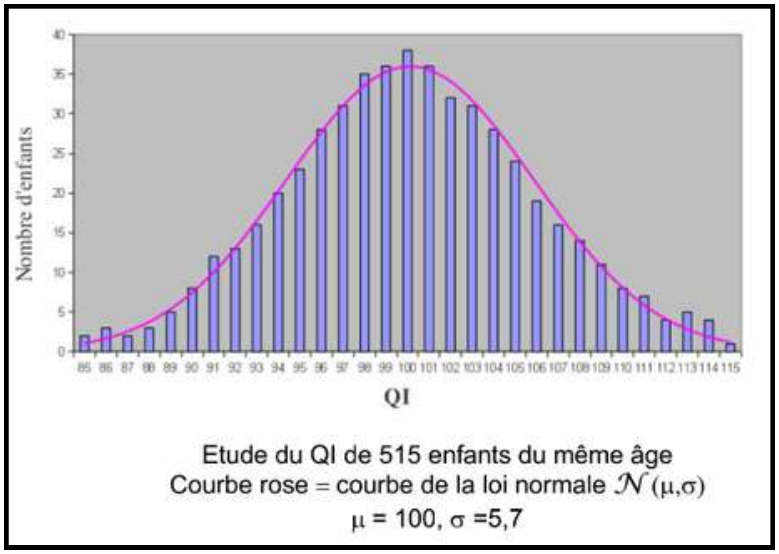
- En **abscisse** [m ± εs] donc l'IC
- En **ordonnée** : l'effectif pour chaque valeur
- **L'aire** sous la courbe, le % de la population concerné

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)





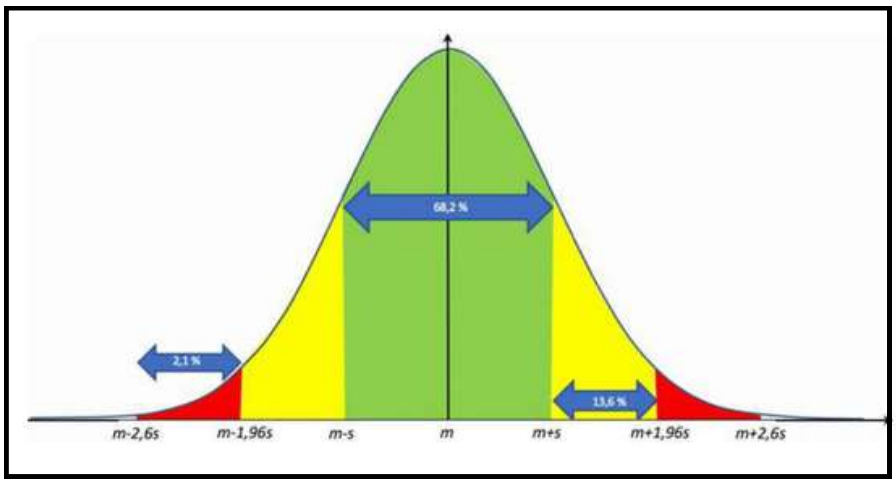
Exemple :



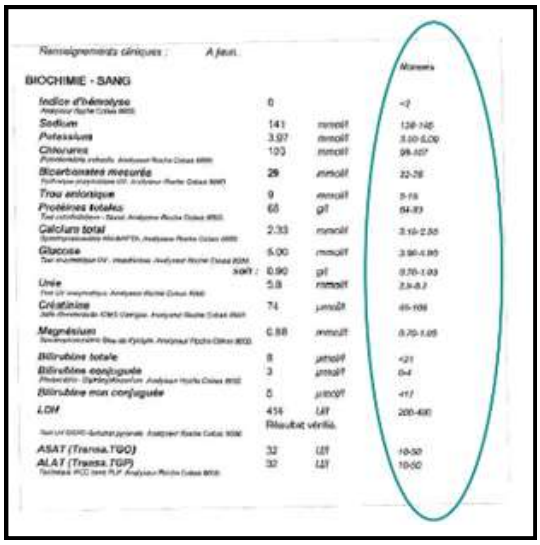
Etude du QI de 515 enfants du même âge
Courbe rose = courbe de la loi normale $\mathcal{N}(\mu,\sigma)$
μ = 100, σ =5,7

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 ++



Pour savoir si la valeur obtenue est '**normale**' ou pas, il faut connaître l'intervalle de confiance du dosage, c'est-à-dire les **normes**.



Estimation de données qualitatives : fluctuation d'un pourcentage

Exemple : dans une population, quel % d'individus présentent un caractère donné ?

- Echantillon représentatif par **tirage au sort** (n sujets)
- Calcul d'un **%** qui tend vers la proportion cherchée, mais s'en écarte suivant une variabilité liée au **hasard**
- Avec un autre échantillon, on obtient un autre %

Estimateur du pourcentage inconnu p → p_{obs}

Estimateur de l'écart type inconnu σ → $s = \sqrt{\frac{p_0 q_0}{n}}$ $q_0 = 1 - p_0$

INTERVALLE DE CONFIANCE

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

Exemple : **précision d'un sondage**

900 personnes ont été interrogées sur leur intention de vote à une élection présidentielle qui oppose 2 candidats A et B.

52% (p=0,52) ont déclaré qu'elles **voteraient A**.

Les journaux annoncent que le candidat A arrive en tête.

Vérification statistique de cette affirmation

Pour A $IC_{0,95} = [0,52 \pm 1,96 \sqrt{\frac{0,52 \times 0,48}{900}}] = [0,49 ; 0,55]$

Pour B $IC_{0,95} = [0,48 \pm 1,96 \sqrt{\frac{0,52 \times 0,48}{900}}] = [0,45 ; 0,51]$

52% et 48 % possèdent des IC contenant 50%

Les 2 candidats peuvent être considérés comme à égalité !

Comparaison estimation ponctuelle/ estimation par intervalle

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 ?

Estimation ponctuelle	**p=167/220=0,76=76%**
Estimation par intervalle	Nous choisissons le risque α = 5%, donc calcul de $IC_{0,95}$ p=0,76 donc q=0,24 $IC_{0,95} = \left[0,76 \pm 1,96\sqrt{\frac{0,76 \times 0,24}{220}}\right]$ $IC_{0,95} = [0,70; 0,82]$

L'estimation par intervalle semble **moins précise.** Mais si l'on refait ce calcul sur un autre échantillon, cette nouvelle estimation sera dans **l'IC.** Ce ne sera **pas** forcément **vrai** avec l'estimation **ponctuelle**.

Autre exemple :

Ainsi, plus la taille de l'échantillon augmente, plus la précision augmente ++

Soit P la population des ouvriers travaillant dans une usine
Nous voulons estimer le pourcentage **p** d'hommes dans cette population.
Considérons un échantillon TAS de 10 ouvriers : 7 hommes, soit p_o=70%
Estimation au niveau de P, au risque α=1% ?

s= $\sqrt{\frac{0,7 \times 0,3}{10}}$ = 0,144 IC $_{99\%}$ = [0,7 ± 2,6 x 0,144] = [32,6% ; 100 %]

Considérons un échantillon de 1000 ouvriers : même % d'hommes p_0 = 70%

s= $\sqrt{\frac{0,7 \times 0,3}{1000}}$ = 0,014 IC $_{99\%}$ = [0,7 ± 2,6 x 0,014] = [66,4% ; 73,6%]

TUT RECAP :

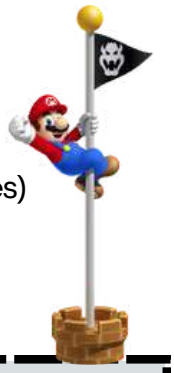
effectif n augmente → IC se resserre → précision augmente

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**.

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 ⇗**



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 est fini, vous avez terminé le niveau 1 ! J'espère que le thème vous a plu, dites moi si vous trouvez ça trop chargé... Le cours peut paraître compliqué au début, mais vous verrez les questions demandées sont plutôt simples. N'oubliez pas, la biostat c'est des QRU, donc si vous doutez faites par élimination ! Bon courage pour votre année, accrochez vous et écoutez vous ! Niveau 2 : les statistiques déductives avec ma cotut :)

Dédi à Jad et Lukas qui m'ont accompagné en p1
Dédi à Rahma qui voulait continuer sa licence wtf
Dédi à mon ardoise et mes airpods qui m'ont beaucoup trop carry
Dédi aux cafés noisette >>
Grosse dédi à la TTR du s1 c'était trop bien