Introduction aux bio-statistiques

Emilien Jeannot MPH

UNIVERSITÉ DE GENÈVE
FACULTÉ DE MÉDECINE
Institut de médecine sociale et préventive

Quelques notions de statistiques pour commencer

1. Les statistiques: reflet de la réalité
2. Les différents types de variables
3. Les mesures utilisées en statistiques descriptives
4. Les intervalles de confiance

Les statistiques: reflet de la réalité

- Les statistiques font partie intégrante de l'arsenal de compétences du professionnel de santé notamment de santé publique: médecins, infirmières, psychologues, sociologues, biologistes, économistes utilisent et doivent savoir utiliser les statistiques de base afin de pouvoir estimer des paramètres biologiques ou non, analyser des données, proposer et tester une hypothèse de recherche.

Les statistiques: reflet de la réalité

- Les statistiques sont un outil pour décrire notre réalité car elles permettent notamment de décrire des phénomènes (biologiques, économiques, sociologiques) grâce à des paramètres reflétant la réalité des observations; elles permettent également d'estimer ces paramètres, de les comparer entre eux ou dans diverses population.

Les statistiques: reflet de la réalité

- Elles sont également un élément de planification car elles sont utilisées pour prédire la survenue d'événements d'ordres biologique, économique ou social.
- Leur interprétation, si elle est pertinente, constitue une aide à la décision pour la planification, que se soit de politique ou d'intervention de santé.

Les différents types de variables

- On définit en statistiques une variable comme étant une caractéristique, une qualité, une fonction ou un facteur capable de prendre une infinité de valeurs différentes selon les individus.
- Toutes les variables ont une unité statistique différente.

Les différents types de variables

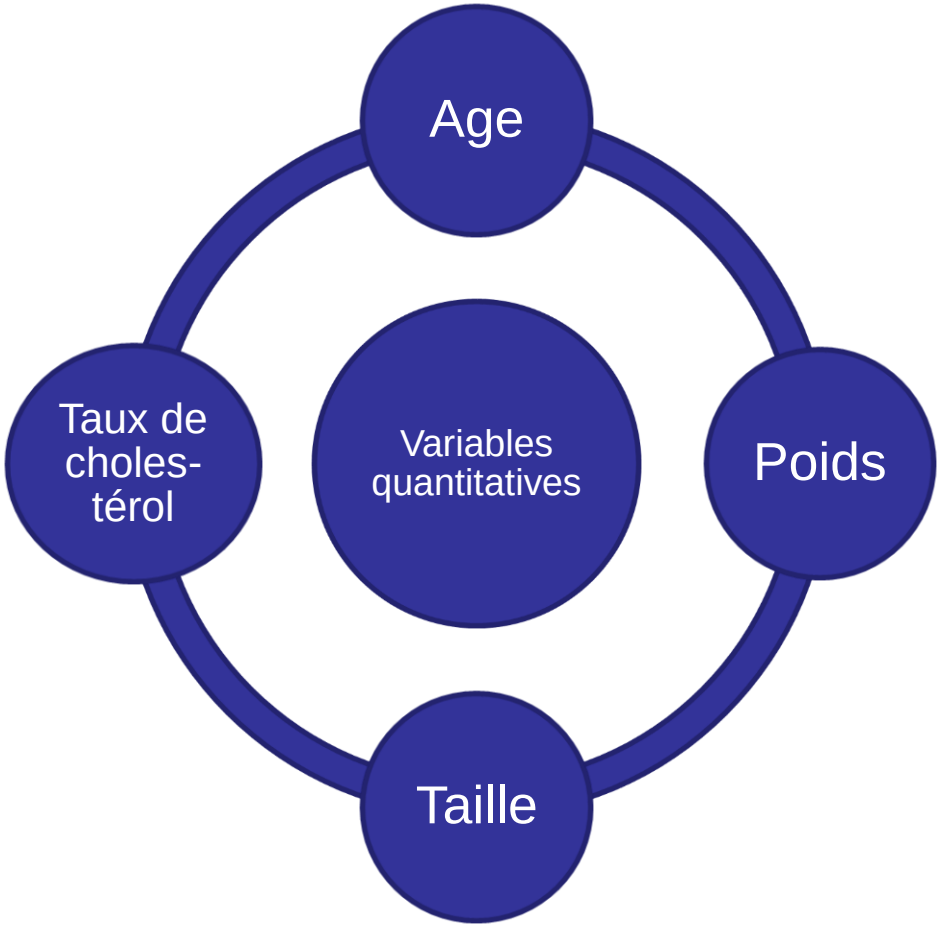
- Le poids, la taille, la couleur des yeux, la pression systolique, le taux de cholestérol, l'âge, le genre sont des variables.
- On distingue 3 grandes sortes de variables, les variables dite **quantitatives**, celles dites **qualitatives** et les variables **binaires**.

Les différents types de variables

– ***Les variables quantitatives***

- Les variables quantitatives ont pour caractéristiques d'être numériques, elles peuvent prendre n'importe quelle valeur numérique dans l'intervalle de leurs observations.
- En théorie, il existe donc une infinité de variables.

Les différents types de variables

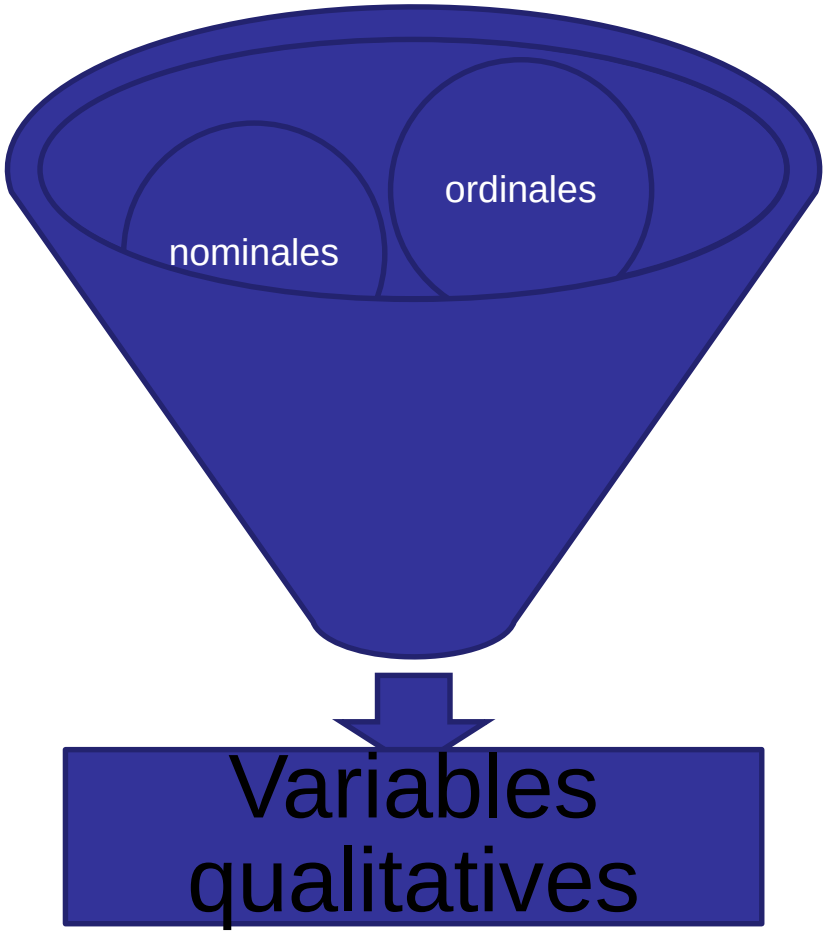


Les différents types de variables

– ***Les variables qualitatives***

- Contrairement aux variables quantitatives, les variables qualitatives n'ont pas de valeur numérique, leurs valeurs sont réparties en différentes classes ou qualités; on peut néanmoins dénombrer les effectifs de chacune des classes préalablement définies.

Les différents types de variables



UNIVERSITÉ DE GENÈVE
FACULTÉ DE MÉDECINE
Institut de médecine sociale et préventive

Les différents types de variables

– ***Les variables qualitatives ordinales***

- Ces variables s'expriment en classes qui peuvent être triées ou hiérarchisées selon une échelle de valeurs. Par exemple, la classe socio-professionnelle d'une personne (ouvrier, artisan, cadre) est une variable de type qualitative ordinale.

Les différents types de variables

- Pour simplifier leur utilisation ont peut transformer/recoder ces variables en variables numériques, par exemple :
- Ouvrier=1
- Artisan=2
- Cadre=3

Les différents types de variables

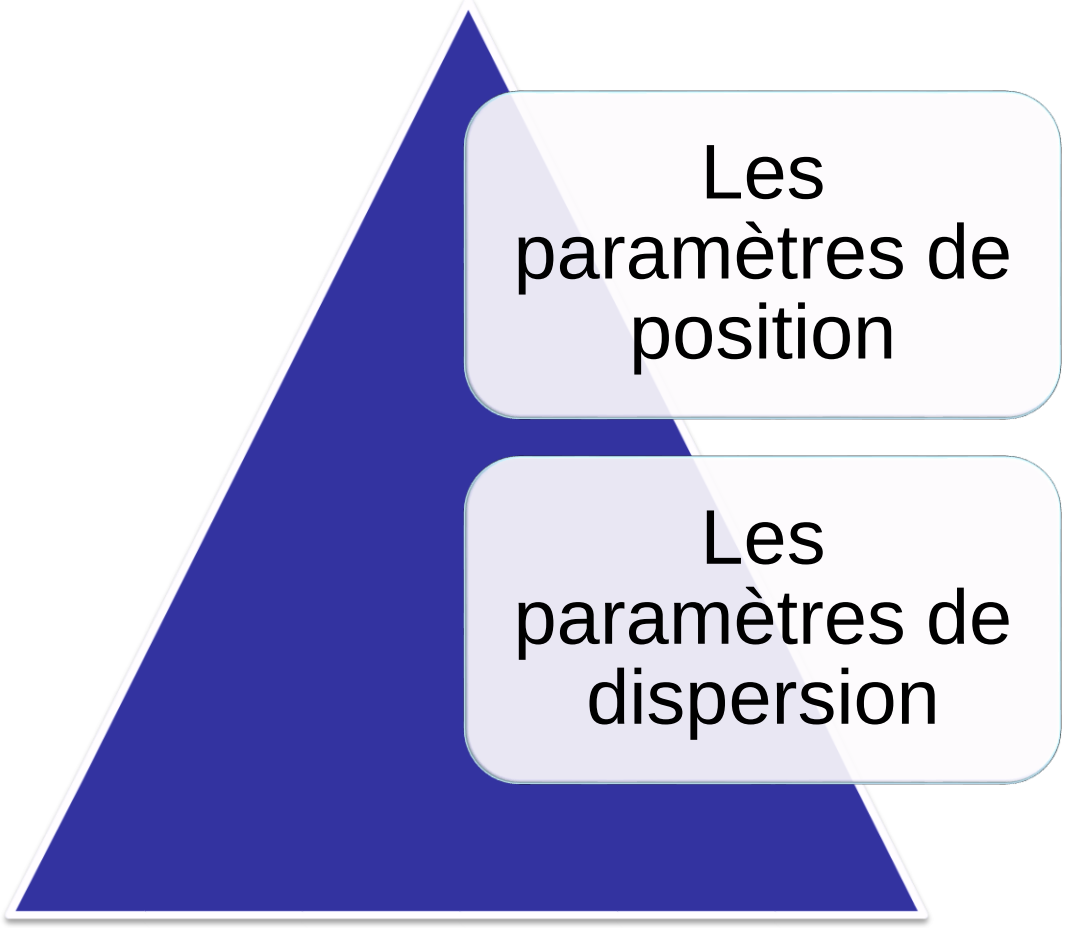
- ***Les variables binaires***

- Les variables binaires sont en fait un type particulier de variables qualitatives nominales où l'on ne retrouve que 2 classes.
- On peut distinguer les variables dichotomiques (présence ou non d'une caractéristique) et les variables booléennes qui ne peuvent prendre pour valeur que vrai ou faux.

Les mesures utilisées en statistiques descriptives

- Nous allons voir comment décrire ces variables, les comparer entre elles ou à d'autres variables venant d'autres séries de données.
- L'élément d'information qui permettra de décrire notre distribution de données s'appelle un **paramètre**.

Les mesures utilisées en statistiques descriptives



Les paramètres de position

– ***La moyenne***

- La moyenne arithmétique est un paramètre de position qui est calculé en faisant la somme des valeurs (x_1) contenues dans une distribution divisée par le nombre (n) de sujets.

- Formule :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Les paramètres de position

- ***La moyenne***

- *Exemple* : si on dispose de l'âge de 10 enfants : 6, 7, 8, 9, 10, 11, 12, 12, 13, 13. La moyenne = 101/10 = **10.1**

Les paramètres de position

La moyenne : attention à son utilisation

Le calcul de la moyenne peut dans certain cas être difficilement interprétable car elle dépend des valeurs extrêmes. Pour éviter une trop grande influence des valeurs extrêmes (basse ou haute), on lui préférera la **médiane**.

UNIVERSITÉ DE GENÈVE
FACULTÉ DE MÉDECINE
Institut de médecine sociale et préventive

Les paramètres de position

– *La moyenne*

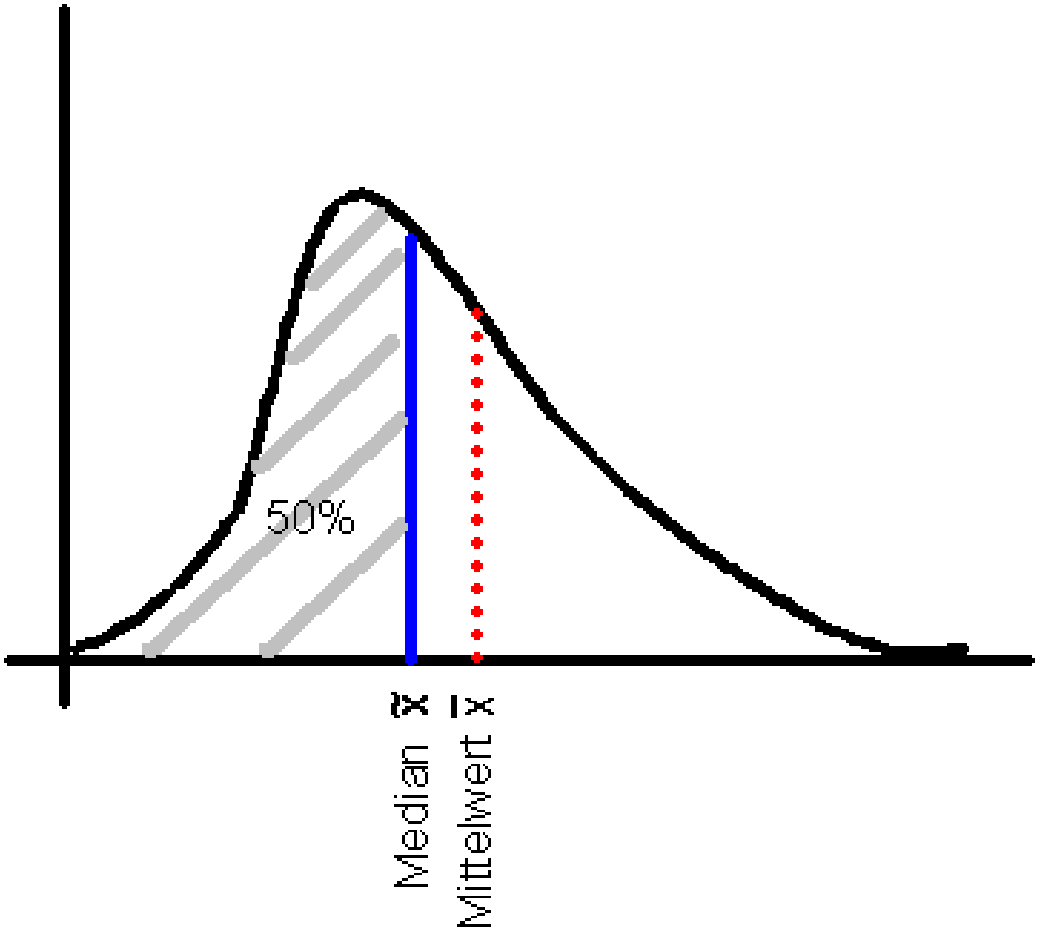
- Il existe d'autres moyennes que la moyenne arithmétique, on peut citer les moyennes géométrique, harmonique, quadratique, énergétique etc… ces différentes moyennes sont peu utilisées en épidémiologie.

Les paramètres de position

– *La médiane*

- La médiane est la valeur qui permet de partager en 2 groupes d'effectifs égaux les valeurs d'une distribution. Autrement dit, c'est la valeur centrale qui est définie par le fait que 50% de la distribution est inférieure à cette valeur et 50% de la distribution est supérieure à cette valeur.

Les paramètres de position



Les paramètres de position

La médiane : son utilité

Contrairement à la moyenne arithmétique, la valeur **médiane** permet d'atténuer l'influence perturbatrice des valeurs extrêmes enregistrées.

Les paramètres de position

- Exemple:
- Supposons 19 personnes ayant un faible revenu et un milliardaire se trouvent dans la même pièce. Tous prennent l'argent de leur poche et le déposent sur une table. Chaque personne ayant un faible revenu dépose 5 francs, alors que le milliardaire met 1 milliard de francs.
- Le montant total est 1 000 000 095 de francs.
- Si l'on calcule la moyenne de l'apport de chaque personne =

  1 000 000 095 / 20 = **50 000 004,75** francs.
- Cependant, la valeur **médiane** est de 5 francs, puisque le groupe peut être divisé en deux parties égales de 10 personnes.
- Le calcul de la médiane corrige la valeur extrême causée par le milliardaire. Elle est souvent utilisée pour évaluer et comparer les salaires.

Les paramètres de position

Quartiles

Les quartiles sont les 3 valeurs qui partagent une distribution en 4

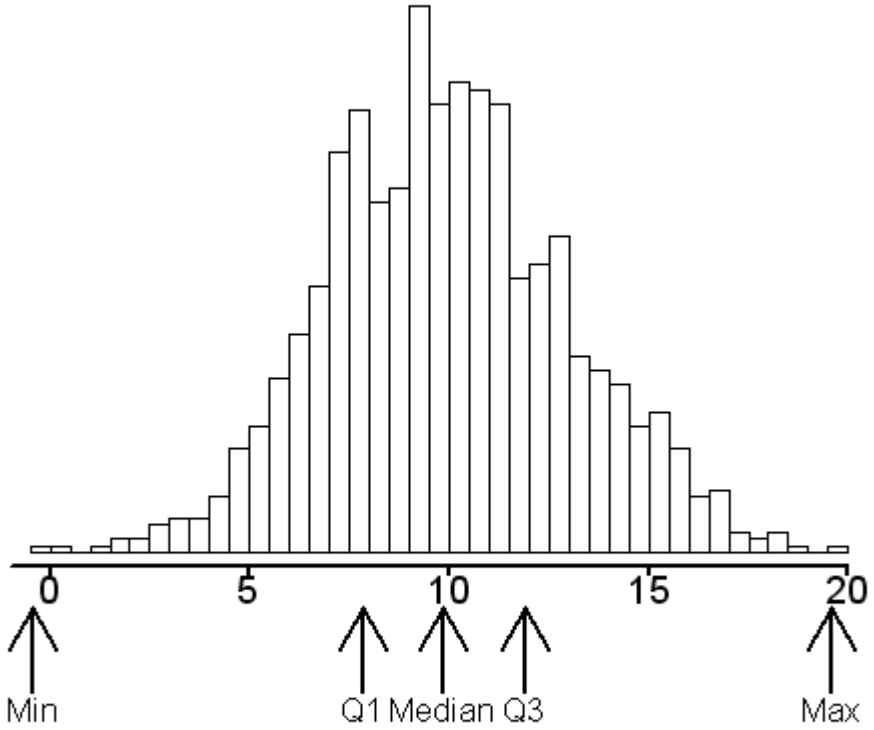
Le premier quartile est la valeur où 25 % de la distribution est inférieure à cette valeur et 75% supérieure à cette valeur

Le second quartile est la médiane

Le troisième quartile est la valeur où 75 % de la distribution se trouve en dessous de cette valeur et 25% est au dessus

UNIVERSITÉ DE GENÈVE
FACULTÉ DE MÉDECINE
Institut de médecine sociale et préventive

Les paramètres de position



Les paramètres de position

Les percentiles sont les 9 valeurs qui partagent une distribution en 10 groupes, chacun des groupes contennant 10% des effectifs

Le percentile 10 est la valeur ou 10 % de la distribution se trouve en dessous de cette valeur et 90% au dessus

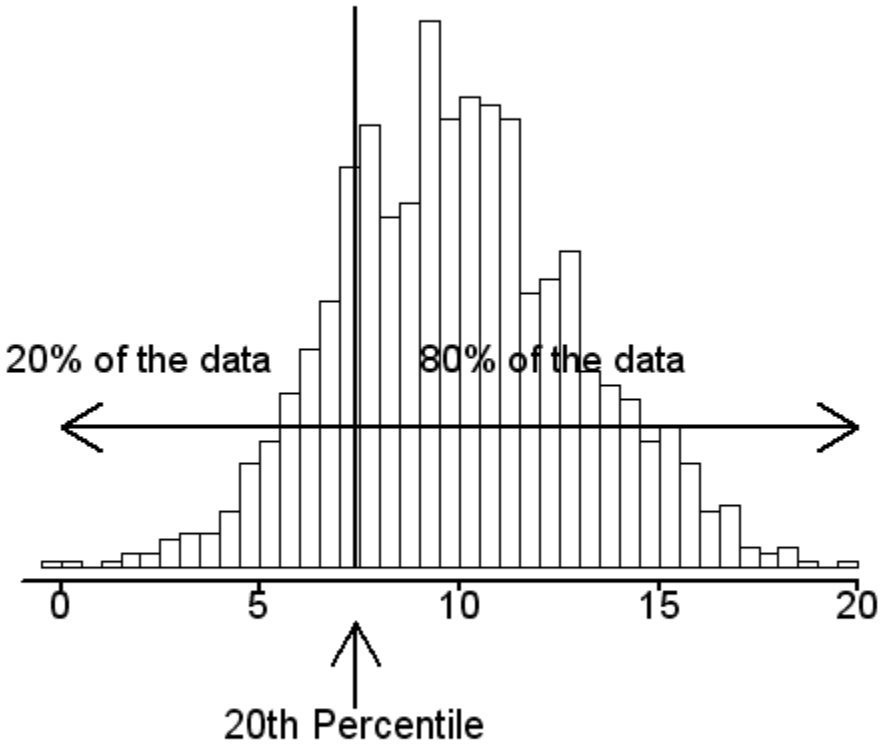
Le percentile 20 est la valeur ou 20 % de la distribution se trouve en dessous de cette valeur et 80% au dessus

Etc….

Le percentile 90 est la valeur ou 90 % de la distribution se trouve en dessous de cette valeur et 10% au dessus

A noter que le percentile 25 correspond au premier quartile, le percentile 50 représente quant-à lui la médiane, le percentile 75 correspond au troisième quartile

Les paramètres de position



Les paramètres de position

– *Le mode*

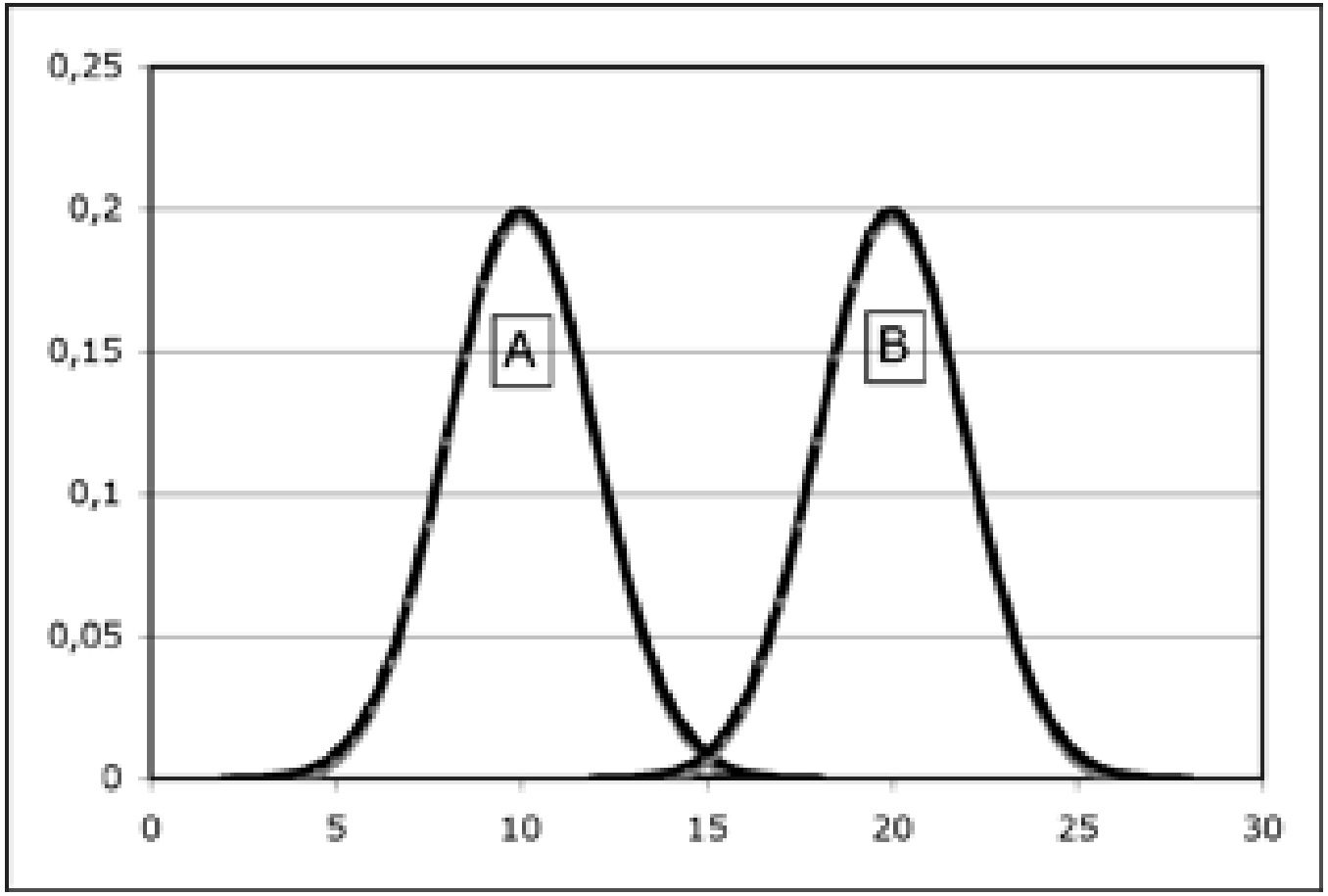
- Le mode est la valeur d'une distribution la plus répandue, si nous reprenons notre exemple avec l'âge des enfants avec une distribution de 9 âge : 6, 7, 8, 9, 10, 11, 12, 12, 13, le mode est la valeur 12 car elle revient 2 fois sur 9.

Les paramètres de position

– *Le mode*

- Si dans une distribution il n'existe qu'un seul mode avec un pic bien défini, on qualifiera la distribution d'**uni-modale**. S'il y a un 2ème pic de valeurs dans une autre partie de la distribution on appellera cette distribution **bimodale.**

Les paramètres de position



Les paramètres de dispersion

- Les paramètres de dispersion sont calculés également pour les variables statistiques quantitatives.
- Ils complètent les informations données par les paramètres de position sur la distribution d'une série de données en fonction d'une variable.
- Ils ne donnent pas une information complète sur une variable statistique X : en effet, deux variables qui ont la même moyenne peuvent se présenter avec des dispersions très différentes.

Les paramètres de dispersion

- ***Les extrêmes***

- Les extrêmes correspondent aux valeurs minimum et maximum d'une distribution.
- Exemple si on dispose de l'âge de 10 enfants : 6, 7, 8, 9, 10, 11, 12, 12, 13, 13. Les extrêmes seront **6** et **13.**

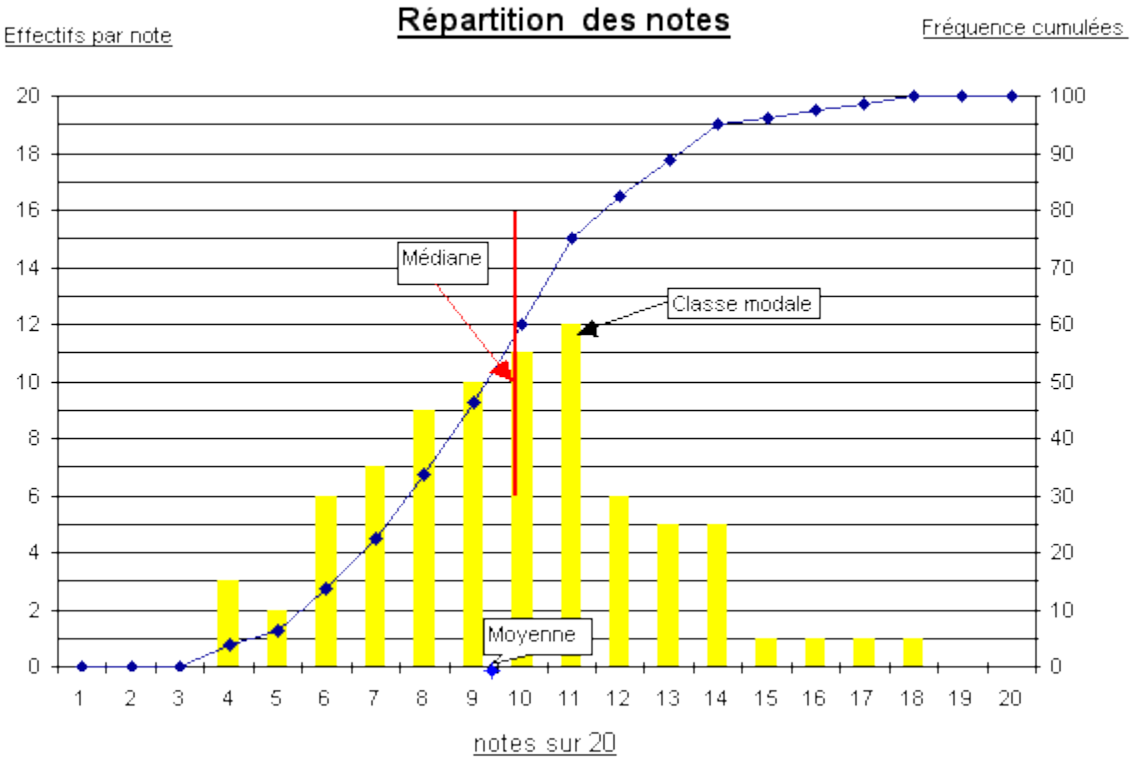
Les paramètres de dispersion

- L'étendue.

- L'étendue correspond à la différence entre les valeurs extrêmes maximum et minimum.

UNIVERSITÉ DE GENÈVE
FACULTÉ DE MÉDECINE
Institut de médecine sociale et préventive

Les paramètres de position



Le mode est égal à 11/20 : c'est la note la plus fréquente : 12 élèves l'ont obtenue.

L' **étendue** est de 18 - 4 = 14 points.

Les paramètres de dispersion

- La variance

- La variance est le paramètre de dispersion souvent utilisé, son calcul est basé sur la moyenne arithmétique des carrés des écarts à la moyenne pour chacune des valeurs d'une distribution.
- Exemple : Pour une variable X, nous disposons de 3 valeurs, 1, 2 ,3. Pour cette exemple le calcul de la moyenne donne 2 -> (1+2+3)/3. Le calcul de la variance est le suivant
- $VAR(X) = [(1 - 2)^2 + (2 - 2)^2 + (3 - 2)^2] \div 3 =$ **0,667**
- VAR(X)=variance = [écart au carré moyen] ÷ nombre d'observations

Les paramètres de dispersion

- ***L'écart type.***

- L'écart type noté *standard déviation (SD)* dans les ouvrages anglos-saxons est l'autre paramètre de dispersion le plus souvent utilisé, car celui-ci intervient dans le calcul des intervalles de confiance autour d'une moyenne dans la loi normale.
- L'écart type correspond à la racine carré de la variance.

- ***Écart-type (S)*** $$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$

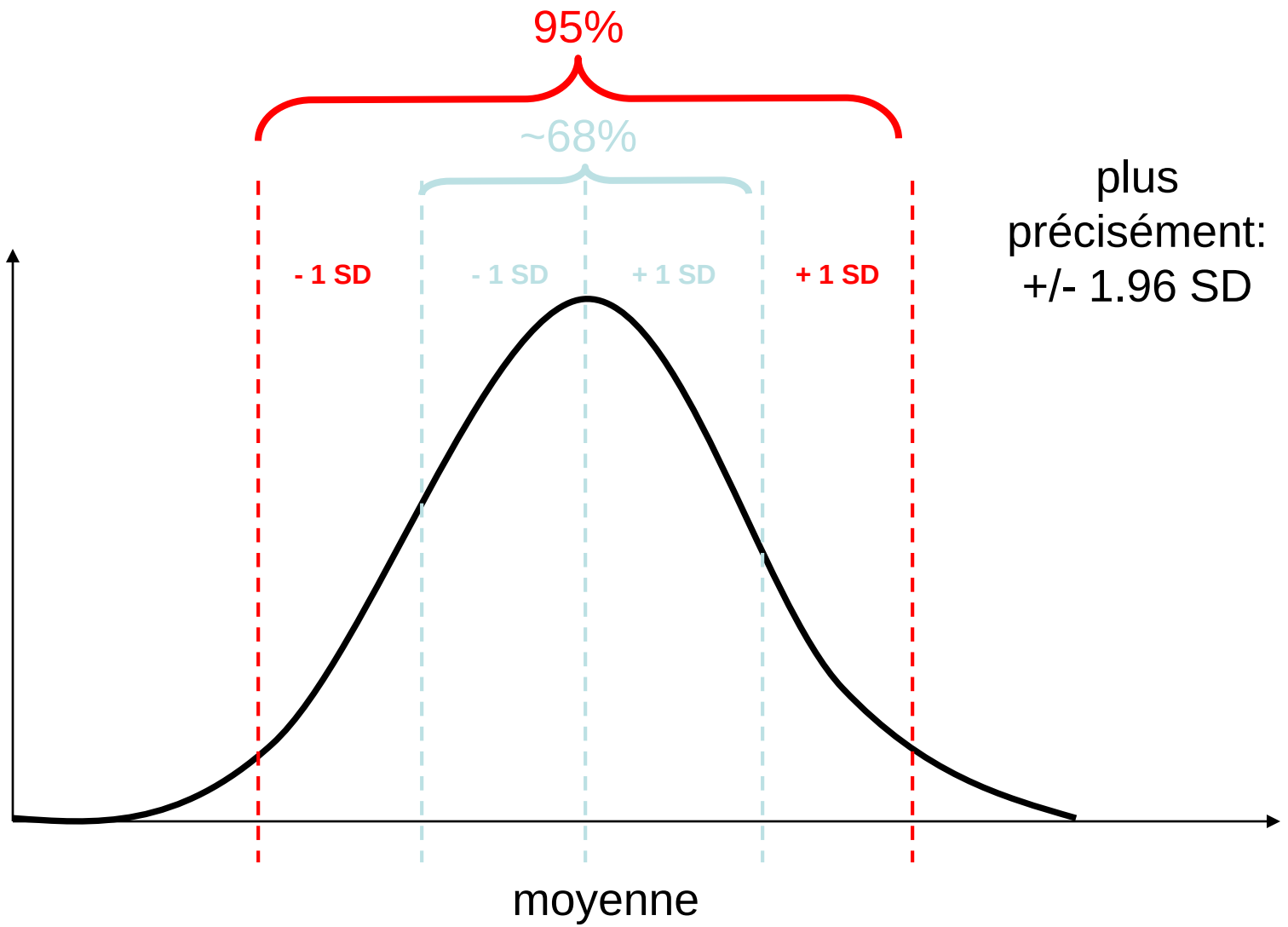
La loi normale

- C’est la loi la plus fréquemment utilisée car elle permet de décrire la majorité des phénomènes biologiques. En effet, on observe en général que la distribution d’une variable X se trouve autour d’une valeur moyenne et que les autres valeurs s’accroissent et décroissent de façon homogène et symétrique autour de cette valeur moyenne.

La loi normale

- Cette distribution prend la forme d'une courbe en cloche aussi appelée courbe de Gauss hommage a **Johann Carl Friedrich Gauss,** mathématicien, astronome et physicien allemand du 19^{e} siècle.

Ecarts-types



plus précisément: +/- 1.96 SD

Les intervalles de confiance

- En épidémiologie, il n'est pas possible d'interroger l'exhaustivité d'une population pour connaître un paramètre physique ou biologique.

- Par exemple, pour connaître le pourcentage d'enfants vaccinés contre la rougeole en Suisse, il n'est pas possible pour des raisons évidentes de coût et de logistique de demander le statut vaccinal de tous les enfants : la couverture vaccinale sera donc évaluée sur un échantillon.

Les intervalles de confiance

- Le fait d'étudier un échantillon et non pas l'ensemble d'une population entraîne certaines contraintes.

- En effet, le calcul des paramètres de position tels que la moyenne et le pourcentage d'une certaine catégorie de la population se retrouve lié a l'échantillonnage. On doit donc estimer ces paramètres avec une certaine marge d'erreur.

Les intervalles de confiance

- Cette marge d'erreur est appelé intervalle de confiance.

- Elle correspond à la zone où l'on sait pour une probabilité donnée que se trouvera la moyenne ou le pourcentage d'une valeur étudiée.

UNIVERSITÉ DE GENÈVE
FACULTÉ DE MÉDECINE
Institut de médecine sociale et préventive

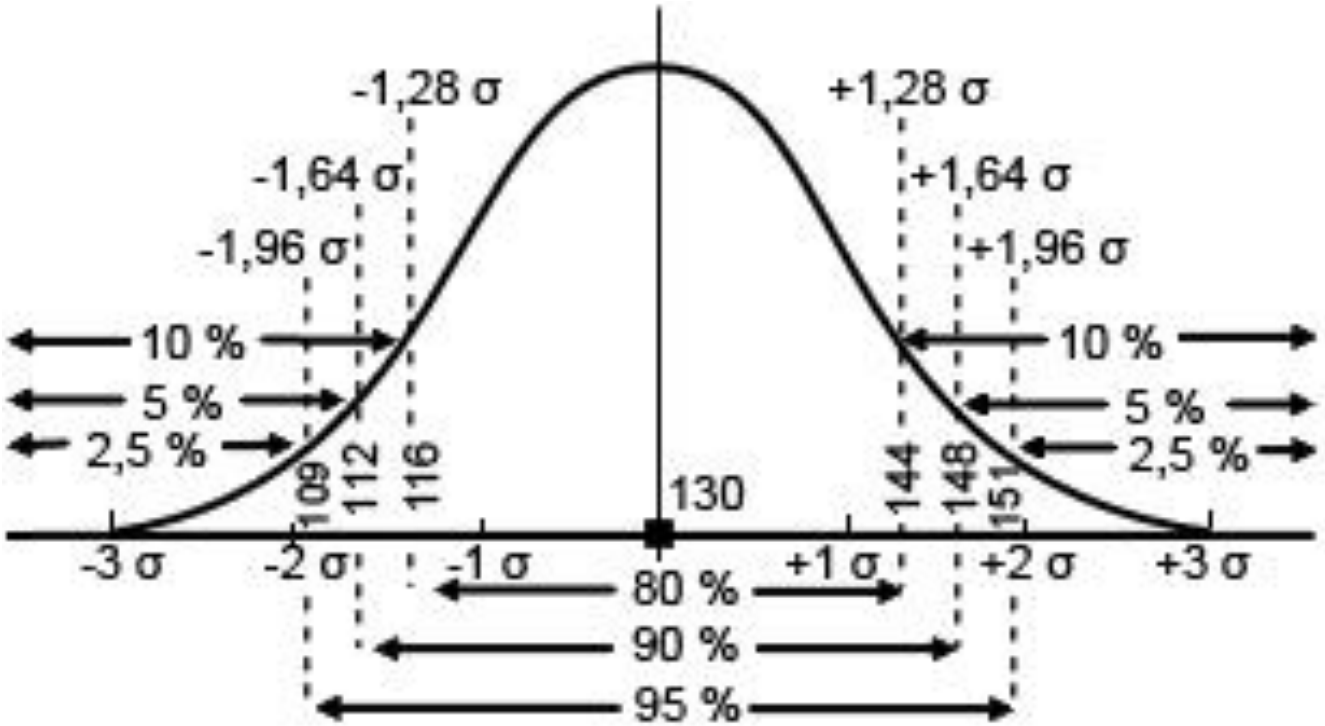
Les intervalles de confiance

- ***Intervalle de confiance d'une moyenne***

- On veut estimer la taille moyenne de 200 collégiens, pour ce faire nous devons choisir un risque d'erreur consentie ou risque α. Ce risque α correspond à l'erreur que nous acceptons pour évaluer notre moyenne. Si, par exemple, on veut être sûr que notre moyenne calculée ait 95 % de chance d'être dans notre intervalle de confiance, nous choisirons un risque α de 5%.

Les intervalles de confiance

- Intervalle de confiance d'une moyenne



UNIVERSITÉ DE GENÈVE
FACULTÉ DE MÉDECINE
Institut de médecine sociale et préventive

Les intervalles de confiance

$$\left]\overline{x} - 1,96\frac{\sigma(X)}{\sqrt{n}}; \overline{x} + 1,96\frac{\sigma(X)}{\sqrt{n}}\right[$$

- Avec
- $\overline{x}$ = *moyenne estimé de l'échantillon*
- *n = taille de l'échantillon*
- Et l'écart type des valeurs de l'échantillon ou *σ*

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

UNIVERSITÉ DE GENÈVE
FACULTÉ DE MÉDECINE
Institut de médecine sociale et préventive

Les intervalles de confiance

- ***Intervalle de confiance d'une moyenne***

- **Moyenne calculée = 1.60**
- Intervalle de confiance à 95% calculé = [1.56 - 1.63]

- **Interprétation:**
- On peut en conclure qu'il y a 95 % de chance pour que la vraie valeur de la moyenne soit comprise entre 1.56 et 1.63.

Les intervalles de confiance

- Intervalle de confiance d'un pourcentage

- C'est le même principe de réflexion avec cette fois ci la formule suivante:
- $p +/- 1.96 * \sqrt{(p*(1-p)/n)}$
- Avec
- P = pourcentage calculé de l'échantillon
- N= taille de l'échantillon

UNIVERSITÉ DE GENÈVE
FACULTÉ DE MÉDECINE
Institut de médecine sociale et préventive

Les intervalles de confiance

- *Intervalle de confiance d'un pourcentage*

- Sur un échantillon de 200 élèves, nous avons calculé le pourcentage d'enfants vaccinés contre la rougeole. Le pourcentage de couverture calculé était de p = 83%. On peut calculer que l'intervalle de confiance à 95 % sera le suivant.
- $IC_{95\%}$= [76.7 – 87.3]

Les intervalles de confiance

- **Interprétation?**

- On peut donc en conclure qu'il y a 95 % de chance pour que le vrai taux de couverture vaccinale contre la rougeole soit compris entre 76.7 % et 87.3%

Les intervalles de confiance

- On constate que pour le calcul de ces intervalles de confiance, **la taille de l'échantillon** joue un rôle important. En effet, plus l'échantillon sera grand et plus la taille de son intervalle de confiance, donc la différence entre ses bornes supérieure et inférieure, sera petite. Inversement, plus l'échantillon sera petit plus l'étendue de l'intervalle de confiance sera grand.
- http://www.sphinxonline.com/suristat/simu1.htm

Merci de votre attention

UNIVERSITÉ DE GENÈVE
FACULTÉ DE MÉDECINE
Institut de médecine sociale et préventive