Cours 3a – La méthode statistique en médecine

I/ La statistique descriptive

A) Définitions générales

- 2 types d'analyse pour l'étude statistique :

  → **Descriptive** : étude d'une situation à l'aide de paramètres (moyenne, médiane, etc …)

  → **Déductive** : définir si la variabilité d'une donnée est due au hasard ou à une autre explication

- ***Donnée***: Résultat de l'observation d'un individu par un instrument de mesure (poids, taille) ou par les sens de l'observateur

- ***Variable*** : Donnée qui n'est pas strictement équivalente pour chaque individu sur lesquels elle s'observe (elle prend une valeur ou un aspect différent pour chaque individu)

- ***Paramètre*** : Grandeur qui apporte une information résumée sur la variable étudiée (moyenne, médiane …) → seulement pour les variables QUANTITATIVES !!

- ***Variabilité*** : Due → Au hasard (= variabilité intrinsèque)
  → A la physiologie ou à une autre explication

  - ***Intra individuelle*** ⟹ la donnée est différente, varie d'un instant à l'autre pour un même individu (ex : la taille)

  - ***Inter individuelle*** ⟹ la donnée n'est pas équivalente d'un individu à l'autre pour un instant T (ex : la couleur des yeux)

△ **TOUTE OBSERVATION EST SOUMISE A UNE VARIABILITE INTRINSEQUE = HASARD**

→ **Pour plusieurs observations, le résultat est très souvent variable**

On relève la glycémie sur un échantillon représentatif. La valeur pour chaque personne oscille entre 0.75 g/L et 1.05 g/L. Il y a donc bien une variabilité intrinsèque à la personne.

L'OBSERVATION D'UNE DIFFERENCE NE PERMET PAS EN SOI D'EN PRECISER LA CAUSE

→ **Constater une différence statistiquement significative ne donne pas la clé de son interprétation**

Si cette glycémie atteint 1.8 g/L pour un individu donné, est-ce que cela signifie qu'il est diabétique ? C'est une possibilité mais il faut se poser la question de la période durant laquelle a été faite la mesure par exemple (post prandiale ?). L'individu sort peut être d'un goûter chargé de tartines au Nutella ! Dans ce cas cette différence n'est pas significative et cette variation est normale après un repas.

- ***Série statistique*** : Collection d'objets de même nature présentant des caractéristiques différentes (=variables)

- ***Population*** : Série exhaustive de tous les individus que l'on veut étudier (ex: tous les étudiants français) → Une population et un échantillon sont des séries statistiques !

- ***Echantillon*** : Ensemble fini et d'effectif limité extrait de la population (ex : tous les P1 français, échantillon de tous les étudiants français)

1) L'échantillonnage

Pourquoi échantillonner ? ⇨ - Population inaccessible en entier

- Etude sur l'échantillon, on « parie » sur l'extrapolation des résultats à la population

→ Echantillon = connu / Population = inconnue

→ **L'échantillon doit être représentatif de la population**

Pour cela, une solution : **LA RANDOMISATION = TIRAGE AU SORT (TAS)**

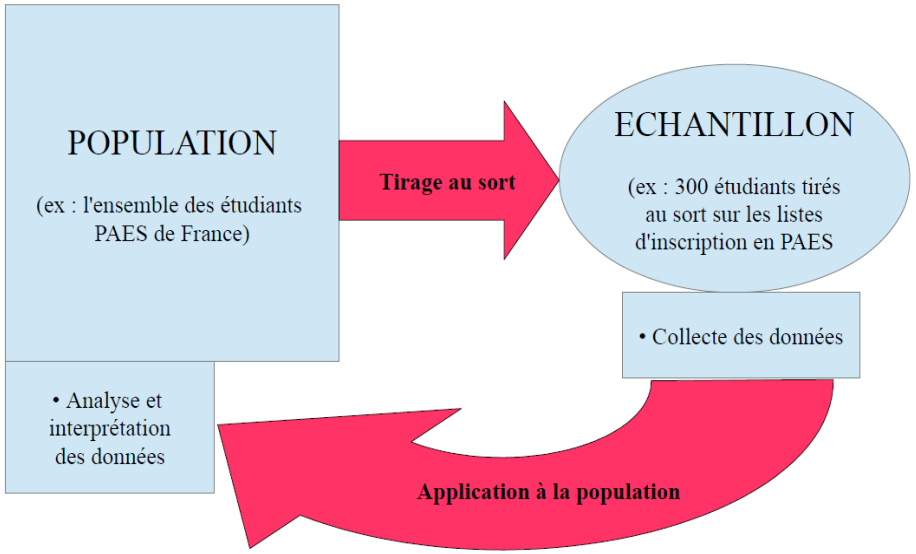
→ Ainsi on élimine les biais de constitution de l'échantillon !

2) Les types de variable

Quantitatives (mesure de quantités)		**Qualitatives** (mesure de qualités)	
Discrète	**Continue**	**Nominale (sans ordre)**	**Ordinale**
Nombre d'étudiants dans l'amphi = un nombre entier	*Poids, taille =une mesure, avec des décimales*	*Couleur des yeux* **Binaire (sans ordre)** *Sexe (homme/femme = 2 possibilités exclusives)*	*Degré de douleur, de satisfaction (sur une échelle numérique ou non)*

→ Une variable qualitative binaire est souvent un cas particulier d'une variable qualitative nominale !

→ Une variable qualitative peut se mettre sous forme numérique
Ex : Nombre de cigarettes fumées par jour - moins de 5 (petit)
- entre 5 et 10 (moyen)
- plus de 10 (beaucoup)
} = Qualitative ordinale



B) Paramètres (seulement pour des variables quantitatives)

Indicateurs de position → Moyenne
→ Médiane
→ Quartiles
} position des tendances de la série statistique

Indicateurs de dispersion → Variance
→ Ecart type
} dispersion des données autour d'un indicateur de position (échantillon homogène ou hétérogène)

1) Indicateurs de position

➢ ***La Moyenne :*** $$m = \frac{\sum x_i}{n}$$

➢ ***La Médiane :*** - **Valeur centrale d'une liste ordonnée par ordre croissant**
- Sépare la liste en 2 groupes de même effectif (50% en dessous/ 50% au dessus)

- Effectif n pair → $$m = \frac{x_{n/2} + x_{(n/2)+1}}{2}$$

= moyenne des 2 valeurs

- Effectif n impair → $$m = \frac{x_{n+1}}{2}$$

➢ ***Les quartiles :*** **- Valeurs qui partagent la série ordonnée en 4 groupes de même effectif**

→**Q1 (premier quartile)** sépare les **premiers 25%** de la série
→ **Q2 (deuxième quartile = Médiane)** sépare les **premiers 50%** de la série
→ **Q3 (troisième quartile)** sépare les **premiers 75%** de la série

→ Pour n multiple de 4 : $Q1 = x_{n/4}$ = valeur de la n/4e donnée

$$Q3 = x_{3n/4}$$

→ Pour n non multiple de 4 :

$$Q1 = \frac{x_i + x_j}{2}$$ (avec i et j les 2 valeurs les plus proches de n/4 : i<n/4<j)

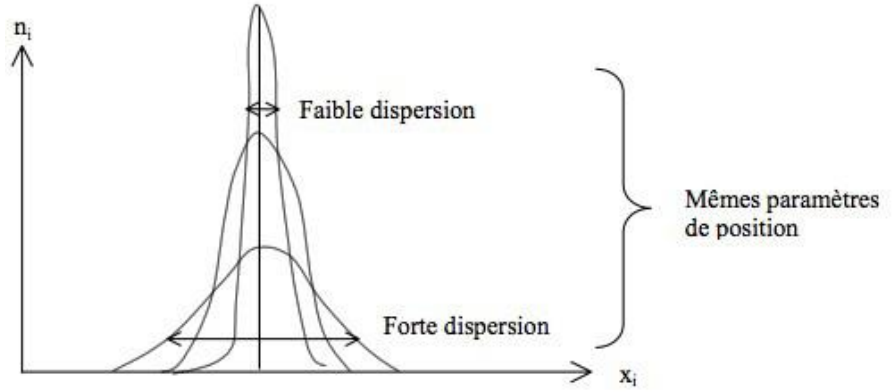
$$Q3 = \frac{x_i + x_j}{2}$$ (avec i et j les 2 valeurs les plus proches de 3n/4 : i<3n/4<j)

2) Indicateurs de dispersion

➢ ***La variance :*** - Indique la dispersion des données autour de la moyenne
- Variance = (Ecart type)²

➢ ***L'écart type :*** - « Moyenne de l'écart à la moyenne »

- Plus les notes d'une classe sont homogènes, plus la dispersion est faible → ecart type faible !!



	Avantages	Inconvénients
Moyenne	→ Facile à calculer → Adaptée aux calculs statistiques → Significative si : - répartition des données symétrique - dispersion faible (= faible écart type)	→ Sensible aux valeurs anormales ++
Médiane	→ Facile à calculer → Peu sensible aux valeurs anormales → Utilisable pour les valeurs ordinales	→ Peu adaptée aux calculs statistiques ...

C) L'estimation statistique

→ **Déterminer une grandeur définie sur une population à partir d'observations effectuées sur un échantillon représentatif de cette population.**

Exemple : combien d'heures travaille un P1 en moyenne à Nice ?

- ***Nombre de degrés de liberté :*** - Nombre des écarts indépendants (Xi – m)
  - Le nombre de degré de liberté (ou ddl) se traduit par le nombre minimal de données qu'il est nécessaire de connaître afin de pouvoir déduire toutes les données manquantes.
  - Quand on veut remplir un tableau à N lignes et n colonnes, il faut connaître au minimum (N-1) x (n-1) données afin d'avoir toutes les données de ce tableau
  - Il y a n écarts (Xi – m)
  - Leur somme est égale à 0
  - Il suffit d'en connaître (n-1) pour tous les connaître → n-1 degrés de liberté
  - Exemple : si l'on cherche deux chiffres dont la somme est 12, aucun des deux chiffres ne peut être directement déterminé par la simple équation X + Y = 12 .
    X peut être choisi arbitrairement, mais alors pour *Y* il n'y a plus le choix. Ainsi, si vous choisissez 11 comme valeur pour *X*,
    Y vaut obligatoirement 1. Il y a donc deux variables aléatoires (X et Y) mais un seul degré de liberté.

1) Les 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**
  - On admet un **risque α** souvent **égal à 5%**

- On l'appelle **« Intervalle de confiance » (IC)** ou **« Intervalle au risque α »**
- Beaucoup plus fiable !

2) Méthodologie

- **Déterminer** précisément **la population à étudier** = **Population cible**
- **Tirage au sort (TAS)** d'un **échantillon représentatif**
- **Etude** de **l'échantillon**
- **Extrapolation** à la **population**

***Estimation** par intervalle*

→ Soient A et B deux échantillons représentatifs d'une population :

- **Deux estimations ponctuelles** d'une même variable réalisées sur les échantillons A et B donneront des **valeurs ponctuelles voisines**, mais pas nécessairement les mêmes valeurs.
- **Deux estimations par intervalles** d'une même variable réalisées sur les échantillons A et B donneront des **Intervalles de confiance (IC) qui se recouvrent**, mais pas nécessairement les mêmes.

Exemple : 2 échantillons représentatifs A et B

	Echantillon A	Echantillon B
Estimation ponctuelle	0,95 g/L	1,03 g/L
Estimation par intervalle (95%)	[0,90 g/L ; 1,04 g/L]	[0,95 g/L ; 1,10 g/L]

- **Les estimations ponctuelles sont proches**
- **Les intervalles de confiance se recouvrent**

3) L'intervalle de confiance

$$\mu \in [\, m \pm \frac{\varepsilon s}{\sqrt{n}} \,] \Rightarrow \text{Intervalle au risque } \alpha$$

α = Probabilité de se tromper dans l'estimation de la moyenne μ

ε = Ecart réduit (différent pour chaque α choisi)

A connaitre ++ : **α = 5% → ε = 1.96**
α = 1% → ε = 2.6

❖ **Si α diminue, ε augmente !!**

❖ **Plus α est petit, plus l'intervalle est grand !!**

(car ε augmente, son addition fait augmenter la valeur, sa soustraction la fait diminuer → l'IC est plus large !)

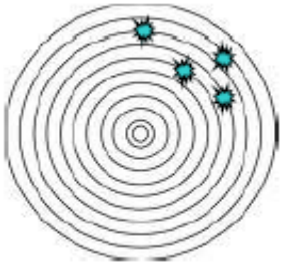
→on réussit plus souvent mais on prend un plus grand risque de se tromper !

❖ **Si la taille de l'échantillon augmente, la précision augmente !!**

❖ **Plus l'IC est large, moins il est précis !!**

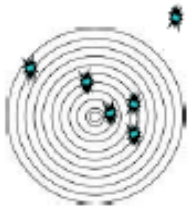
→ ***Large =*** **plus de chance de l'atteindre, mauvaise précision de l'estimation**

= risque α faible = ε élevé = IC plus large



→ ***Resserré =*** **meilleure précision de l'estimation, moins de chance de l'atteindre**

= plus grand risque α = plus petit ε = IC moins large



❖ **Indice de précision i**

→ **Calcule la précision de l'estimation de la moyenne m**

→ **C'est la largeur de l'IC**

→ **La précision augmente quand i diminue !!**

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

ε = écart réduit
s = sigma = écart type
n = effectif de l'échantillon

❖ Nombre de sujets nécessaires

→ Calcule le nombre de sujets nécessaires à l'échantillon pour un écart type et une précision donné

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

ε = écart réduit
s = sigma = écart type
i = indice de précision

Enfin fini ! Bon courage à tous !