Statistiques Descriptives En Épidémiologie

I La méthode statistique en médecine :

Les biostatistiques (= statistiques appliquées au domaine de la santé) ont 3 objectifs :

1. description d'une population par rapport à une maladie
2. **Evaluation** des traitements, des techniques, des coûts
3. Mise en place des informations épidémiologiques et en tirer des conclusions

Les biostatistiques doivent être capables de décider si une observation peut être due au seul hasard ou si elle a une autre explication.

Quelques définitions de base :

Terme	Définition	Exemple
Statistique	art de collecter, d'analyser, et d'interpréter des données. Lorsqu'elle est appliquée au domaine de la biologie, on parle de biostatistique. Il en existe 2 types : – descriptive – déductive : une observation est-elle due au hasard ? Ou existe-t-il une explication ?	- Stat descriptive : Collecte de 2 données sur la population française : taille et couleur des yeux - Stat déductive : On constate que les sujets ayant une taille > 1,70m ont tous les yeux bleus. Hasard ? Corrélation ?
Population	Série exhaustive de tous les individus étudiés sur lesquels on veut appliquer des décisions	Ensemble de la population française
Echantillon	Ensemble fini et d'effectif limité extrait de la population, le plus souvent randomisé créé par tirage au sort = TAS) donc représentatif	10 personnes tirées au sort dans la population française.
Variable quantitative	Une donnée quantitative est le résultat de l'observation d'un individu par l'utilisation d'un appareil de mesure, variable d'un individu à l'autre. Lorsque cette donnée est mesurée sur plusieurs individus, on parle de variable quantitative.	Les tailles des 10 individus : 1,62m / 1,63m / 1,66m / 1,66m / 1,68m /1,68m / 1,70m / 1,75m / 1,80m / 1,90m
Variable qualitative	Une donnée qualitative est le résultat de l'observation d'un individu, par les sens de l'observateur. Variable d'un individu à l'autre. Lorsque cette donnée est mesurée sur plusieurs individus, on parle de variable qualitative	La couleur des yeux : bleus / verts / marrons / gris
Paramètre	grandeur apportant une information résumée sur la variable étudiée	La moyenne : m = 1,708m La médiane : 1,68m
Série statistique	Collection d'objets de même nature, présentant des caractéristiques différentes d'un objet à l'autre	Les hommes et les femmes sont des objets de même nature mais avec des caractéristiques différentes (mince, gros, blond, brun ...)
Variabilité	Ensemble des différences inter-individuelles et intra-individuelles. Elles peuvent être : – dues au hasard – physiologiques	- inter-individuelles : **Comparaison** de la taille des sujets **entre eux.** - intra-individuelles : évolution de la taille avec l'âge, **comparaison du sujet à lui-même** à diverses

		périodes

L'étude d'un problème statistique peut se décomposer en 4 étapes : recueil des données, classement, réduction des données (statistique descriptive) et déduction en vue de prévision (statistique déductive)

II Statistique descriptive :

Une étude statistique descriptive s'effectue sur une population dont les éléments sont des individus. Elle consiste à observer et à étudier un même aspect sur chaque individu, nommé caractère ou variable.

A Notion de variabilité :

Toute donnée biologique possède une **variabilité**. Sa connaissance est indispensable pour pouvoir dire si la valeur d'une variable **quantitative** est normale ou pas.

- Une variabilité maîtrisée permet d'établir une estimation.
- Une variabilité non maîtrisée conduit à des biais

Exemple :
La **valeur moyenne de la glycémie (variable quantitative)** chez un sujet normal est de **1g/L +/- 0,25 g/L**
Ceci signifie qu'une glycémie appartenant à l'intervalle : **[0,75 g/L ; 1,25 g/L]** est normale.
.
Vos tuteurs de biostat préférés décident de se contrôler la glycémie un matin à jeun. Ils trouvent :

- Vincent : 1,2 g/L. Il a donc une glycémie **normale**
- Julia : 0,7 g/L. Elle a donc une glycémie **infra – normale**

B La représentation des données :

Il existe divers types de données / variables :

- **Les variables qualitatives** :
  - ✗ binaires : homme / femme ; malade / non malade …
  - ✗ nominales : couleur des yeux, des cheveux …
  - ✗ ordinales : degré de satisfaction des étudiants vis à vis de la tut rentrée :
    peu satisfait / satisfait / très satisfait / il n'y a pas de mots tellement c'était bon !

- **Les variables quantitatives** :
  - ✗ discrètes : Il s'agit de variables quantitatives qui, théoriquement, ne prennent qu'un nombre ni de valeurs.
    Exemple : âge des étudiants en PAES (à 1 an près). En effet, on dit qu'un étudiant a 19 ans, 20 ans etc. mais on ne tient pas compte de ce qui se passe entre les années.

  - ✗ continues : Il s'agit de variables quantitatives qui, théoriquement, peuvent prendre toutes les valeurs d'un intervalle de l'ensemble de nombres réels. Les variables pourront être regroupées par classes (= discrétisées) et transformées en variables qualitatives nominales ou ordinales.
    Exemple : poids, glycémie, ...

Les variables qualitatives peuvent être représentées de plusieurs manières :

- diagramme en bâton ou histogramme. Les barres sont proportionnelles aux valeurs représentées
- tableau
- secteur (= camember) : les surfaces sont proportionnelles aux valeurs représentées

Exemple : On relève la couleur des yeux de 90 bébés à la sortie de la maternité. On constate que :

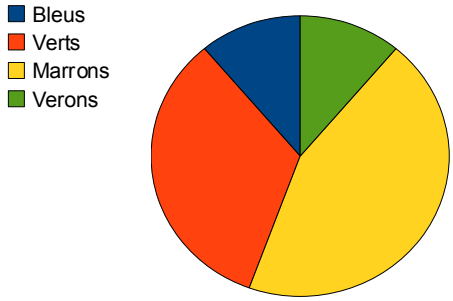
- 10 ont les yeux bleus
- 30 ont les yeux verts
- 40 ont les yeux marrons
- 10 ont les yeux verrons (un oeil vert et l'autre marron)

Modalité = Couleur des yeux	Nombre de bébés	Fréquence = proportion
bleus	10	11,11%
verts	30	33,33%
marrons	40	44,44%
verrons	10	11,11%

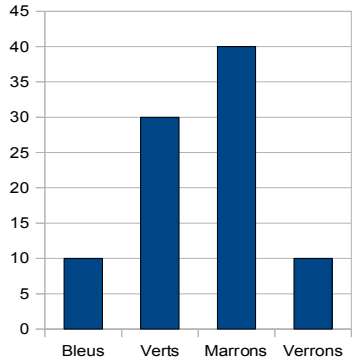
NB : La fréquence de chaque modalité se calcule de la manière suivante :

fi = ni/n

avec ni = l'effectif de la modalité concernée i



Secteur



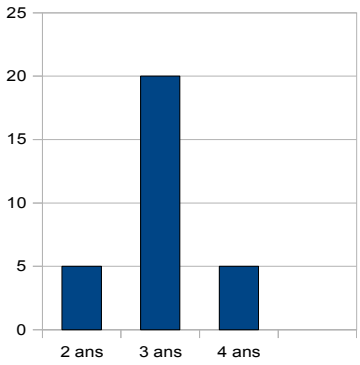
Histogramme

Les variables quantitatives peuvent être représentées/synthétisées de plusieurs manières :

- diagramme en bâton ou histogramme
- tableau. Si il n'y a que 2 variables quantitatives, on utilise un tableau à double entrée dit table de contingence.
- synthétisées grâce à des paramètres

Exemple : On relève l'âge de 30 enfants d'une classe de 1ère année de maternelle :

- 5 ont 2 ans, c'est à dire [2 ans ; 2 ans et 363 jours]
- 20 ont 3 ans, c'est à dire [3 ans ; 3 ans et 363 jours]
- 5 ont 4 ans, c'est à dire [4 ans ; 4 ans et 363 jours]



Âge	Nombre d'enfants	proportion
2 ans	5	16,67%
3 ans	20	66,67%
4 ans	5	16,67%

Quelques propriétés concernant l'histogramme :

- Il est normalisé, c'est à dire que la surface de chaque rectangle est égale à la proportion des données qu'il représente. Ainis, la surface totale de l'histogramme est égale à 1.
  Exemple : Ici, on a : 0,1667 + 0,6667 + 0,1667 = 1

- L'histogramme est une approximation de la densité de probabilité de la variable

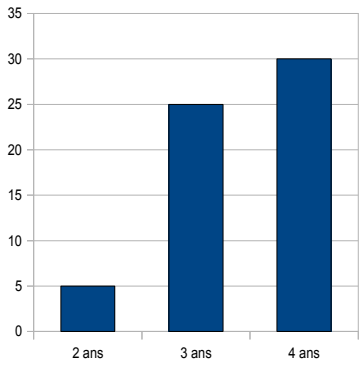
- Soient pi la propotion des données comprise dans un intervalle [ai; bi] et hi la hauteur du rectangle de base [ai; bi], on a : pi =surface du rectangle = hi x (bi-ai).

Les effectifs cumulés croissants (ECC) : Soit xj, la jème valeur observée de la variable d'intérêt X. Le

jème effectif cumulé croissant, noté ECCj, représente le nombre d'individus dont a valeur de la variable est inférieure à xj. Donc **$ECC_j = \sum n_i$ pour i ≤ j**.

Contruisons ici le tableau et l'histogramme correspondant des ECC :

Âge	Nombre d'enfants	ECC
2 ans	5	5
3 ans	20	25
4 ans	5	30



Notion de paramètre : Un paramètre est une grandeur apportant une information résumée sur la variable étudiée. Il existe des paramètres de positions et de dispersion.

Concernant les **paramètres de position**, on a :

1) La moyenne :
   - Pour une variable quantitative **discrète** : **$m = \sum x_i/n$**
   - Pour une variable quantitative **continue** : **$m = \sum n_i\, x_i/n$**

2) La médiane = **observation centrale des valeurs** 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 valeur de la variable correspondant au rang :
     n/2 ou [n/2 + (n+1) /2] / 2 → pour n pair, La valeur de la médiane correspond à [$X_{n/2} + X_{[(n/2)+1]}$] /2.
     La première méthode étant un rappel de celle donnée dans le cours du Professeur Benoliel

   - Si n est impair, la médiane est donnée par la valeur de la variable correspondant au rang :
     (n+1) /2

3) Les quartiles : Ce sont les valeurs **de la variable** qui partagent la série d'effectif n en **4 sous-séries** de même effectif.
   - le premier quartile = Q1 = Q25 est donné par la valeur correspondant à l'effectif de rang : n/4
   - le deuxième quartile = Q2 = Q50 est la médiane
   - le troisième quartile = Q3 = Q75 est donné par la valeur correspondant à l'effectif de rang : 3n/4

4) Les quantiles : Ce sont les valeurs **de la variable** qui partagent la série d'effectif n en **q sous-séries** de même effectif. Ainsi, le qème quantile est la valeur au dessous de laquelle se trouvent q% de l'effectif n.

5) Le mode : Il s'agit de la valeur de la variable ou de la valeur centrale d'une classe de la variable dont **l'effectif** est un **maximum local.** A noter qu'il existe des distributions bi modales, c'est à dire où il y a *2 valeurs de la variable pour lesquelles l'effectif est un maximum local.* Le nombre de modes est un paramètre intéressant.

6) Les extrema : Il s'agit des valeurs maximales et minimales de la **variable**. Ils ont peu de valeur.

7) Les symétries / asymétries :
   - Dans les distributions asymétriques, médianes et moyennes peuvent être éloignées
   - Dans les distributions symétriques, médianes et moyennes sont proches

Concernant les **paramètres de dispersion** : Ils complètent les informations données par les paramètres de psition, en permettant d'apprécier la tendance plus ou moins forte des données à "s'étaler" de part et d'autre des valeurs centrales.

1) La variance : ou (écart type)^2. Elle indique la **dispersion** des données autour de la moyenne. Elle s'exprime dans une unité qui est le carré de l'unité de la variable. Elle est donnée par la formule suivante :

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

2) L'étendue : (x_{max} - x_{min})

3) La distance inter quartile : IQ_{75} - Q_{25}| Il s'agit de l'étendue des données après élimination de 25 %| des valeurs les plus faibles et plus élevées, ce qui permet d'éliminer d'éventuelles valeurs aberrantes.

Exemple : On relève le poids de 5 enfants d'une classe d'un service de pédiatrie. On trouve :
- 30 kg
- 31 kg
- 31 kg
- 32 kg
- 33 kg

1) La moyenne : Le poids est une variable quantitative continue.
m = [(30x1) + (31x2) + (32x1) + (33x1)] / 5 = 31,4kg

2) La médiane : Soient les valeurs suivantes : 30 / 31 / **31** / 32 / 33.
Rappels sur le Professeur Benoliel : (n+1) / 2 = 6/2 = 3
La médiane est donc la **3ème valeur des poids rangés par ordre croissant** : 31 kg.

3) Le 1er quartile : Soient les valeurs suivantes : 30 / 31 / 31 / 32 / 33.
Le 1er quartile ou Q1 est la **valeur des poids qui délimite les premiers 25%** de la série.
Q1 = 0,25 x 5 = 1,25 → Q1 se trouve entre la 1ère et la 2ème valeur des poids.
Soit : (30 + 31) / 2 = **30,5 kg**.

4) Le 3ème quartile : Soient les valeurs suivantes : 30 / 31 / 31 / 32 / 33.
Le 3ème quartile ou Q3 est la **valeur des poids qui délimite les premiers 75%** de la série.
Q3 = 0,75 x 5 = 3,75 → Q3 se trouve entre la 3ème et la 4ème valeur des poids.
Soit : (31 + 32) / 2 = **31,5 kg**.

5) Le mode : La valeur de la variable pour laquelle **l'effectif est le plus élevé** est **31 kg**. En effet, 2 bébés pèsent 31 kg, alors que les autres poids ne concernent qu'un seul bébé. **31 kg** est le **mode** de cette série statistique.

6) Les extrema :
   - 30 kg est un minimum local
   - 33 kg est un maximum local

7) Les symétries / asymétries : Ici, la médiane (31 kg) et la moyenne (31,4 kg) sont assez **proches**, donc la distribution est **symétrique**.

8) La variance : Le calcul de la variance donne : **161,8 kg^2**. Donc l'écart type vaut **12,72 kg.**

9) L'étendue : 33 - 30 = **3 kg**
10) La distance inter quartiles : 31,5 - 30,5 = **1 kg**

	avantages	inconvénients
moyenne	– Facile à calculer, à manipuler – significative si la répartition symétrique des données – dispersion faible	– Sensibles aux valeurs anormales et aux erreurs – sensible aux minimum – sensible aux maximum
médiane	– Facile à calculer – peu sensible aux valeurs anormales et aux erreurs – utilisable pour les valeurs ordinales	– Moins adéquat pour les calculs statistiques – nécessite de classer les données par ordre croissant

C Application des statistiques au domaine de l'épidémiologie :

1) Définitions :

L'épidémiologie est l'étude de la fréquence et de la répartition dans le temps des problèmes de santé dans des populations humaines et du rôle des facteurs qui les détermine. Il y a différents domaines :

1. L'épidémiologie descriptive : Elle étudie la **fréquence** et de la **répartition dans le temps** des problèmes de santé dans des populations humaines, en fonction des caractéristiques des personnes, de la répartition géographique et de leur évolution dans le temps.

2. L'épidémiologie explicative ou analytique : Elle recherche les **causes** des problèmes de santé, le rôle de l'**"exposition à des facteurs de risque** etc.

3. L'épidémiologie évaluative : Elle apprécie les **résultats** d'une action de santé sur la collectivité.

2) Les indicateurs de santé :

Un indicateur de santé est une variable mesurable permettant d'apprécier l'état de santé d'une population. Le but est donc de **mesurer l'état de santé** d'une population, de **classer** les problèmes de santé, de **dégager des choix** prioritaires et de présenter aux populations et aux décideurs des **solutions**. Parmi ces indicateurs de santé, on distingue :

- Les indicateurs socio démographiques : âge, natalité, fécondité ...
- Les indicateurs socio économiques : éducatio, revenus, comportements sociaux ...
- Les indicateurs sanitaires : mortalité, morbidité, espérance de vie, mortalité évitable ...
- Les indicateurs d'utilisation des services de santé
- Les mesures d'activité et d'évaluation
- Les indicateurs de fréquence : Ils s'intéressent à la survenue d'un évènement dans le temps.

  - **Le taux d'incidence** par exemple est la vitesse de production de **nouveaux** cas d'une maladie dans la population pendant une année de temps / nombre de personnes suivies pendant cette unité de temps.

$$TI = \frac{I}{PT}$$

$$PT = \sum_{k=1}^{k=N} \Delta t_k$$

  - Le prévalence = incidence x durée de la maladie.