Statistiques Descriptives et Indicateurs en Épidémiologie : Estimations & Intervalles

I- Rappels

La Statistique est une branche des mathématiques basée sur les observations d'évènements réels à partir desquelles on cherche à établir des hypothèses plausibles en vue de prévisions.

Un problème statistique se décompose en 4 étapes : *recueil des données*, *classement et réduction des données* (statistique descriptive), *analyse des données* (statistique inférentielle), *déduction de prévisions.*

La Statistique Descriptive s'effectue sur une population dont les éléments sont appelés « individus ». Elle consiste à observer et étudier un même aspect de chaque individu, nommé caractère.

L'objectif de la statistique est de connaître les caractéristiques d'une population à partir de données mesurées sur un échantillon obtenu par tirage au sort. Les connaissances récupérées aident à prendre des décisions diagnostiques *(avec l'aide de nos connaissances physiopathologiques)* et thérapeutiques *(après le diagnostic à l'aide de nos connaissances des médicaments)*

II- Différents types de Variables

Donnée : La donnée est le résultat de l'observation d'un individu.

Observer : Réduire un objet infiniment complexe à un nombre limité de caractéristiques.

Variable : Caractéristique mesurable sur plusieurs individus.

A) Variables Qualitatives :

Ce sont des caractères non mesurables. Les individus appartiennent à une seule modalité. La variable qualitative va être soit binaire *(un simple oui/non)*, soit nominale *(quand on ne peut pas les ordonner, par exemple la couleur)*

Quand deux *(ou plus)* variables qualitatives sont mesurées sur le même sujet, les valeurs obtenues sont placées dans un tableau à double entrée = tableau bivarié = **tableau de contingence**

	Problèmes dentaires	*Pas de problèmes dentaires*	*Total*
Hommes	383	2417	2800
Femmes	408	2612	3020
Total	791	5029	5820

B) Variables Ordinales ou Semi-Quantitatives :

Ce sont des variables qualitatives qui peuvent être ordonnées (classées en ordre croissant)

➔ *Par exemple le score APGAR, la mention d'un diplôme, un indice de satisfaction, ect..*

C) Variables Quantitatives :

« Quantitatif »signifie qu'elles sont le résultat de la mesure d'un comptage. On peut associer au caractère un nombre, on peut donc le mesurer. On en distingue 2 types :

« Discrète » : c'est un caractère quantitatif qui ne prend qu'un nombre n_i de valeurs *(le nombre d'enfants)*

« Continue » : c'est un caractère quantitatif qui (théoriquement) peut prendre toutes les valeurs d'un intervalle de l'ensemble des nombres réels *(la taille réelle).* Les variables peuvent être regroupées par classe *(comme la pression artérielle)*

Catégorie	PRESSION Systolique	PRESSION Diastolique
Hypotension	Inférieure à 100	Inférieure à 60
Normal	Entre 100 et 139	Entre 60 et 89
Faible hypertension	Entre 140 et 159	Entre 90 et 99
Hypertension moyenne	Entre 160 et 179	Entre 100 et 109
Hypertension sévère	Supérieure à 180	Supérieure à 110

Une variable continue peut être regroupée en classes : elle devient alors « quantitative discrète » ou « qualitative ordinale » .

III- Présentation des Données

A) Pour des données qualitatives :

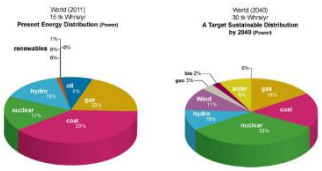
On peut **lister dans un tableau** chaque sujet avec chacun son caractère correspondant, mais on se rend bien compte que dès que l'effectif n devient important, ce mode de présentation des données devient illisible

Patiente #	*Satisf action*
1	Très satisfaite
2	Très satisfaite
3	Plutôt satisfaite
4	Très satisfaite
5	Plutôt satisfaite
6	Plutôt satisfaite
7	Très satisfaite
8	Très satisfaite
9	Très satisfaite
10	Plutôt satisfaite
11	Très satisfaite
12	Très satisfaite
13	Plutôt satisfaite
14	Très satisfaite
15	Très satisfaite
16	Plutôt satisfaite
17	Plutôt satisfaite
18	Très satisfaite
19	Très satisfaite
20	Plutôt satisfaite
21	Très satisfaite
22	Très satisfaite
23	Très satisfaite
24	Plutôt satisfaite
25	Très satisfaite

Si on compte les effectifs de chaque caractère possible, on peut calculer la fréquence de ces caractères

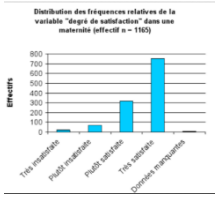
➔ *Ici un degré de satisfaction des mères ayant accouché, avec comme modalités « très satisfaite, plutôt satisfaite, ect…)*

On va d'ailleurs souvent transformer ça en **tableau des effectifs.**



Pour simplifier la lecture des données on va utiliser d'autres formes de graphiques : Le **diagramme en secteur**, où les surfaces sont proportionnelles. *Il est à privilégier pour les variables qualitatives non ordinales.*

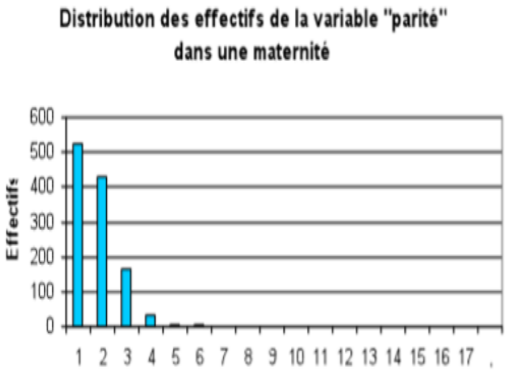
On a aussi **l'histogramme en bâtons** : on y trouve en abscisse les différentes modalités, en ordonnée les effectifs. Une fois de plus ma surface des barres est proportionnelle aux effectifs. *Il est lui à privilégier pour les variables qualitatives ordinales.*



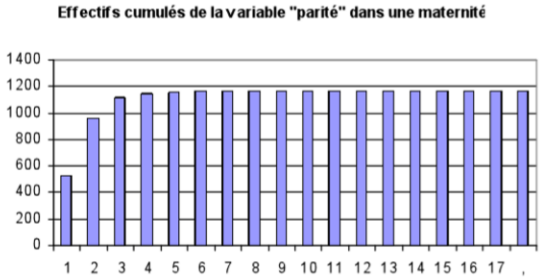
B) Pour des données quantitatives discrètes :

Comme ces données sont quantitatives, ce qui va être intéressant ici est de les ranger **par ordre croissant**.

Un tableau listé reste illisible mais on y retrouve aussi le tableau des effectifs, et le **diagramme en bâtons** avec encore en abscisse les différentes valeurs et en ordonnée les effectifs. La surface est proportionnelle à l'effectif si l'origine des effectifs est 0.



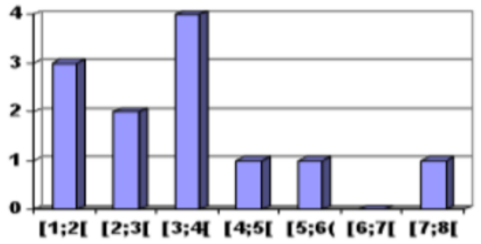
On peut aussi illustrer ce type de variable par les **effectifs cumulés croissants** : le j^{ème} effectif cumulé croissant, c'est le nombre d'individus dont la variable est inférieure ou égale à xj *(dans la colonne 3 on a les individus qui présentent 0, 1, 2 et 3 fois le caractère étudié)*.



C) Pour des données quantitatives continues :

On va ici regrouper nos données par intervalle, pour pouvoir créer un **histogramme normalisé**. La hauteur et la surface des rectangles sont proportionnelles aux effectifs.

On va considérer que la surface totale de l'histogramme est égale à 1. On peut donc dire que l'histogramme devient une approximation de la densité de probabilité de la variable.



Mathématiquement, on peut écrire :

hauteur x (borne sup – borne inf de l'intervalle) = probabilité

Attention, pour une variable quantitative discrète on parle de diagramme en bâtons, pour une variable qualitative continue on parle d'histogramme !

D) Paramètres Statistiques des Variables

A) Paramètres de 1er ordre = de position

Moyenne : Elle est facile à calculer, mais elle est sensible aux erreurs et aux valeurs aberrantes.

Médiane : C'est la valeur au-dessous de laquelle se trouvent 50% des valeurs. Elle est peu sensible aux erreurs mais nécessite de classer les donner par ordre.

Il faut ranger les n valeurs dans un ordre croissant. Si n est impair, la médiane est au rang (n+1)/2. Si n est pair, la médiane est la moyenne des 2 valeurs centrales.

Quantiles : le q^{ème} quantile (et pas quartile !) est la valeur au-dessous de laquelle on trouve q% de ces n observations. Il est noté Q_p. **C'est une généralisation de la médiane.**

- $\boldsymbol{Q_{25}}$: c'est le **premier quartile** *(pour le trouver on prend le nombre de l'effectif n et on le divise par 4. Si ça ne tombe pas sur une valeur on prend la supérieure)*
- $\boldsymbol{Q_{50}}$: c'est la **médiane**, ou **second quartile**
- $\boldsymbol{Q_{75}}$: c'est le **troisième quartile** *(on fait ici n x (3/4) et on prend la valeur sup si ça ne tombe pas juste)*

# marathonien	Temps en min
1	216
2	220
3	176
4	183
5	195
6	195
7	235
8	203
9	197
10	213
11	230
12	229
13	185
14	179
15	215
16	175
17	227
18	196
19	200
20	273
21	211
22	153

Par exemple cet échantillon de marathoniens, avec pour variable le temps de course en minutes.

Quelle est ma moyenne ? Je prends le temps pour chaque marathonien, je les additionne et je divise le tout par le nombre de marathoniens : m = 204,8 min

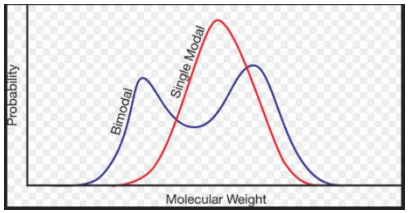
Quelle est ma médiane ? Ici je dois ranger mes effectifs par ordre croissant ! n est pair donc on va prendre la moyenne des 2 valeurs centrales : (200 + 203) / 2 = 201,5

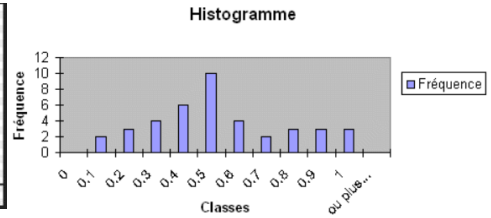
Quel est le premier quartile ? on fait 22/4 = 5,5 donc on prend la 6ème valeur : 185

1	153
2	175
3	176
4	179
5	183
6	185
7	195
8	195
9	196
10	197
11	200
12	203
13	211
14	213
15	215
16	216
17	220
18	227
19	229
20	230
21	235
22	273

Mode : C'est la valeur centrale d'une classe dont **l'effectif est un maximum local** (relativement aux autres effectifs qui sont plus petits). Moyenne et mode sont des indicateurs de tendance centrale.

Une distribution peut être monomodale ou bimodale (ou plurimodale). *L'histogramme à droite est monomodal*. Sur un histogramme la distribution modale est d'ailleurs très facile à repérer, c'est le plus haut rectangle. Le plus intéressant dans un mode est **le nombre de modes**.



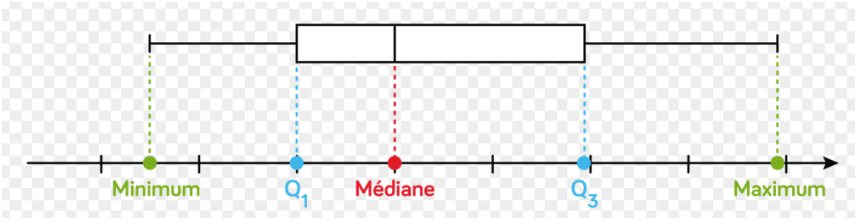


Extrema : Les extrema sont les valeurs la plus petite et la plus grande de l'échantillon. Ils ont peu de valeur en temps que paramètres de position

Pour l'échantillon des marathoniens, les extrema sont 153 et 273 min.

Diagramme en boîte = boîte à moustaches = box-plot : C'est une représentation directe de la distribution, qui permet de savoir si elle est symétrique ou non.

Si la médiane et la moyenne sont éloignées, la distribution est dite **asymétrique**. Si médiane et moyenne sont rapprochées, la distribution est dite **symétrique**. *(dans une distribution asymétrique, médiane & quantiles sont importantes).*

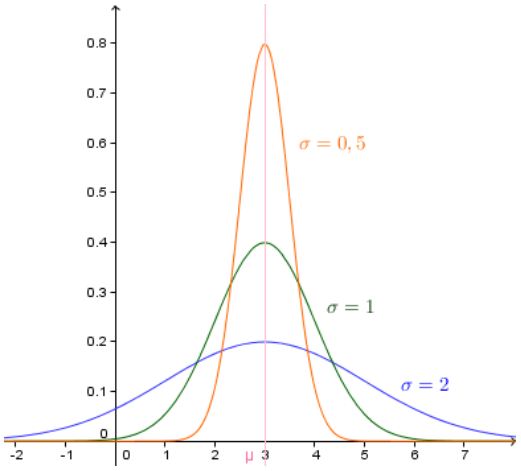


B) Paramètres de 2ème ordre = de dispersion

Ils apprécient la tendance des données à s'étaler à autour de la valeur centrale.

Variance : C'est la somme des carrés des écarts à la moyenne. La racine carrée de la variance est **l'écart-type**.

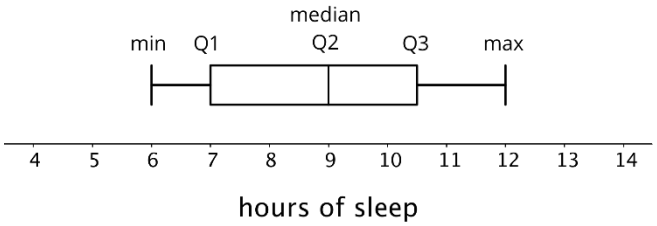
Deux échantillons peuvent avoir la même moyenne mais des écarts-types différents



Étendue : C'est la soustraction de la valeur maximale par la valeur minimale : $x_{max} - x_{min}$

Distance inter-quartiles : $|Q_{75} - Q_{25}|$

C'est comme l'étendue, après qu'on se soit débarrassé de 25% des valeurs les plus faibles et 25% des valeurs les plus élevées. Elle est plus robuste que l'étendue car moins sensible aux valeurs anormales.



E) Indicateurs en Épidémiologie

Épidémiologie : Étude de la fréquence et de la répartition dans le temps et l'espace des problèmes de santé dans les populations humaines et des facteurs qui les déterminent.

On a plusieurs types d'épidémiologie :

- **Descriptive :** Étudie la fréquence et la répartition des problèmes de santé en fonction des caractéristiques des personnes *(âge, sexe…)*, de la géographie et de l'évolution dans le temps.
- **Explicative = Analytique :** Recherche les causes des problèmes de santé, étudie l'exposition à des facteurs pouvant favoriser leur apparition.
- **Évaluative :** Apprécie les résultats d'une action de santé dans la collectivité.

On va dans cette discipline utiliser des **Indicateurs** : Ce sont des variables pouvant être mesurées et permettant de mesurer l'état de santé d'une population.

On a différents types d'indicateurs :

Socio-démographique *(pyramide des âges, fécondité, éducation)*
Sanitaire *(mortalité, morbidité, espérance de vie)*
D'utilisation des services de santé
D'activité et d'évaluation

A) Mesures en épidémiologie

Proportion : Le numérateur est une part du dénominateur, et surtout **les deux sont de même nature**. C'est toujours **compris entre 0 et 1** et ça peut être exprimé en pourcentages.

Exemple : la proportion de garçons daltoniens en France est de 1/10 ou 10%

Ratio : C'est le **rapport** entre les effectifs des **deux classes d'une même variable.** Ce nombre est sans unité.

Exemple en PACES on a 1 garçon pour 3 filles

Indice : C'est un **rapport** entre **deux effectifs de nature différente**. Ce sont souvent des indicateurs de fonctionnement de l'économie de la santé.

Exemple : le nombre de lits par hôpital, le nombre de grains de beauté par individu, le nombre de médecins par région)

Côte : C'est le **rapport** de la **probabilité de survenue d'un évènement sur sa probabilité de non-survenue**. En pratique on va plutôt écrire le nombre de survenues sur le nombre de non-survenues. Elle marche pour des **variables binaires**, « survenue de l'évènement » et « non-survenue ».

Exemple : Lors d'une épidémie d'intoxication alimentaire, sur 75 cas observés, on a 53 cas qui ont consommé de l'aliment X et 22 cas qui n'en ont pas consommé ; ici on apparente la « survenue » à : avoir consommé l'aliment sachant qu'on a une intoxication alimentaire.
La côte d'exposition est de 53/22, soit 2,4 cas exposés à l'aliment X pour 1 cas non exposé

Rapport de côtes = Odds-ratio : C'est le **rapport d'une côte sur une autre.** Il exprime la **quantification du risque.** C'est une approximation du risque relatif.

Exemple : on prend 2 groupes. Le groupe A a été exposé à un aliment Y, et on y observe 30 cas d'intoxication alimentaire pour 15 cas où aucune intoxication ne s'est déclarée. Le groupe B n'a pas été exposé à l'aliment Y, et on y dénombre 40 cas sans intoxication et 2 cas d'intoxication. L'odds ratio est de (30/15) / (2/40) = 2 / 0,05 = 40

Taux : Ce rapport prend en compte la **notion de temps**. C'est la probabilité de survenue d'un évènement en fonction du temps.

Exemple : en 4 jours, 5 chats sur les 25 admis à la SPA ont été adoptés

B) Indicateurs de fréquence

Prévalence : C'est le rapport du **nombre de malades / la population observée**. C'est un **indicateur statique** qui ne prend pas en compte l'évolution de la maladie dans le temps.
Prévalence = incidence x durée de la maladie

Incidence : C'est **le nombre de nouveaux cas dans un intervalle de temps / le nombre de personnes observées dans cet intervalle de temps**. L'incidence est donc un **taux**. C'est donc un **indicateur dynamique**, qui prend en compte la durée d'observation.

Taux d'incidence *: Vitesse de production des nouveaux cas d'une maladie (ou de décès) dans la population (nombre de nouveaux cas pendant une année de temps / nombre de personnes suivies pendant cette unité de temps.*

C) Indicateurs de mortalité / morbidité

Taux de morbidité : C'est le rapport du **nombre de malades / la population totale**.

Taux de prévalence : C'est comme le taux de morbidité, mais en prenant compte **1 maladie en particulier**.

On dit « taux de prévalence » et « taux de morbidité » alors que ces deux indicateurs sont bien statistiques et ne prennent pas le temps en compte, c'est un abus de langage.

- ➔ *Si on mentionne un « taux de prévalence/morbidité » dans l'énoncé, ne pas le prendre comme un piège.*
- ➔ *Un item « la prévalence est un taux » serait par contre à compter FAUX*

Taux brut de mortalité : C'est le **nombre de décès** (peu importe la maladie ou l'accident) **/ personnes / année**

Taux spécifique de mortalité : C'est le **nombre de décès dû à la maladie étudiée / population totale / population totale**

Ces taux ont un intérêt pour les maladies chroniques.

La Biostats vous zem <3

Place à la dédicace !

Au Pr. Lupi, please répond à mes mails <3

À mes co-tuts toujours aussi inébranlables et glowy

À nos vieilles qui il faut l'admettre, sont assez sympatoches

À Antho & Basile allez fooooncez

Paula, u're da best accroche toi

Mes fillots tous malades mais tous talentueux : Lamia, Inès, Elodie, Timotey, Andréa, Teresa <3 et ma co-marraine so sexy and you know it Marion le s

Mes deux mamans, Alex & Enza

À ma dream team 5 des défis en ville

À Mr. Maignant qui est tout chou

À la vie

À …… , fonce ne t'arrête pas je crois en toi
(cale ton prénom là et profite :)*