STATISTIQUES DESCRIPTIVES & INDICATEURS EN EPIDEMIOLOGIE

INTRODUCTION & 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*).

I. DIFFÉRENTS TYPES DE VARIABLES

DONNÉE = 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

1. VARIABLES QUALITATIVES

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

Quand 2 (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

2. VARIABLES ORDINALES = SEMI-QUANTITATIVES

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

Ex : score APGAR, mention d'un diplôme, indice de satisfaction, etc…

3. VARIABLES QUANTITATIVES

Ce sont des variables qui 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 ni de valeurs (*ex : 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 (*ex : la taille réelle*). Les variables peuvent être regroupées par classes (*ex : 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 ».

II. PRÉSENTATION DES DONNÉES

1. POUR LES 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.

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

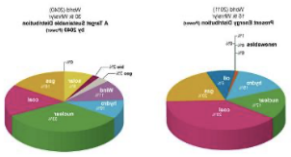
Exemple : Ici, un degré de satisfaction des mères ayant accouché, avec comme modalités « très satisfaite », « plutôt satisfaite », …

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 :

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

• **LE DIAGRAMME EN SECTEUR** = où les surfaces sont proportionnelles. Il est à privilégier pour les variables qualitatives non ordinales.



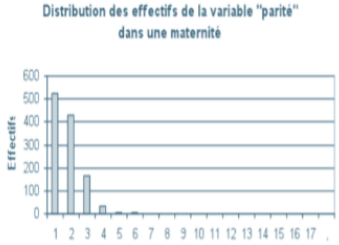
• **L'HISTOGRAMME EN BÂTONS** = on y trouve en abscisse les différentes modalités et en ordonné les effectifs. Une fois de plus, la surface des barres est proportionnelle aux effectifs. Il est à privilégier pour les variables qualitatives ordinales.

2. POUR LES DONNÉES QUANTITATIVES DISCRÈTES

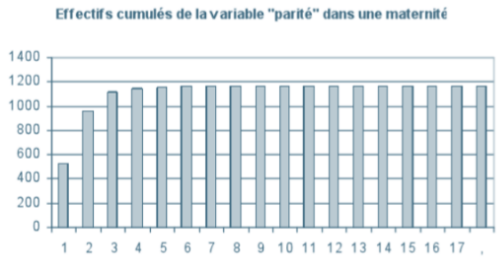
Comme ces données sont quantitatives, ce qui va être intéressant ici c'est de les ranger par **ordre croissant**.

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 ordonné les effectifs.

La surface est proportionnelle à l'effectif si l'origine des effectifs est 0.

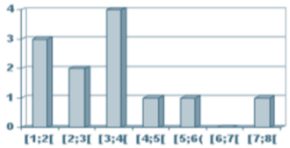


On peut aussi illustrer ce type de variable par les **effectifs cumulés croissants** : le j[ème] effectif cumulé croissant c'est le nombre d'individus dont la variable est inférieure ou égale à xj (*dans la colonne 3, on a les individus qui présentent 0, 1, 2 et 3 fois le caractère étudié*).



3. POUR LES DONNÉES QUANTITATIVES CONTINUES

On va ici regrouper nos données par intervalle, pour pouvoir créer un **histogramme normalisé**. La hauteur et la surface des rectangle 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**, mais pour une variable quantitative continue, on parle d'**histogramme**.

III. PARAMÈTRES STATISTIQUES DES VARIABLES

1. 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 données 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 non pas quartile +++) est la valeur au-dessous de laquelle se trouve q% de ces n observations. Il est noté Q_p. C'est une **généralisation de la médiane**.

- 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.
- Q_{50} = c'est la **médiane**, ou **second quartile**.
- Q_{75} = c'est le **troisième quartile** → on fait n x (3/4) et on prend la valeur supérieure si ça ne tombe pas juste.

Exemple : un échantillon de marathoniens, avec pour variable le temps de course en minutes.

# 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

Quelle est la moyenne ? On prend le temps pour chaque marathonien, on les additionne et on divise le tout par le nombre de marathoniens. m = 204,8 min

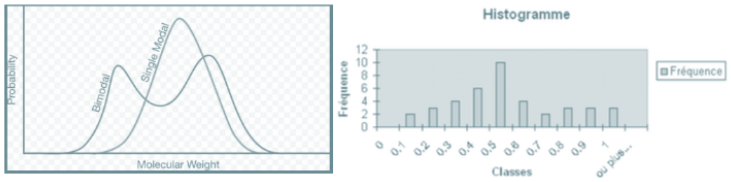
*Quelle est la médiane ? On doit ranger les effectifs par **ordre croissant**. n est pair, donc on va prendre la moyenne des 2 valeurs centrales : (200+203)/2= 201,5*

Quel est le 1er quartile ? On fait 22/4 = 5,5 donc on prend la 6ème valeur = 185

marathonien	
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 (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**.



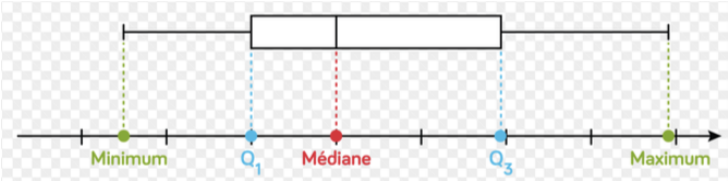
EXTREMA = ce sont les plus petites et plus grandes valeurs de l'échantillon. Elles ont peu de valeurs en tant que paramètres de position.

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

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 **symétrique**.

Dans une distribution asymétrique, médiane et quantiles sont importants.

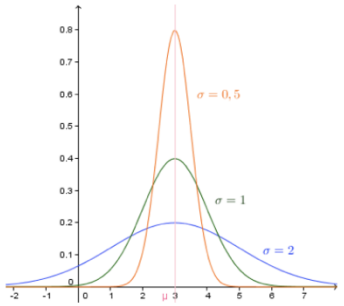


2. PARAMETRES DE 2ÈME ORDRE = DE DISPERSION

Ils apprécient la tendance des données à s'étaler 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**.

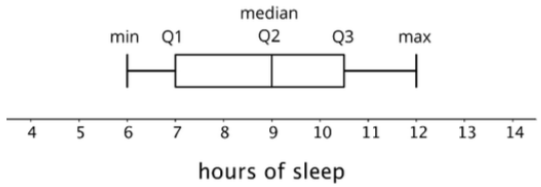
Deux échantillons peuvent avoir la même moyenne mais des écarts-types différents.



ÉTENDUE = c'est la soustraction de la valeur maximale par la valeur minimale :

$$xmax - xmin$$

DISTANCES 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 sensibles aux valeurs anormales.



IV. 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
- **EVALUATIVE** = 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-démographiques (*pyramide des âges, fécondité, éducation*)
- sanitaire (*mortalité, morbidité, espérance de vie*)
- d'utilisation des services de santé
- d'activité et d'évaluation

1. 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 **2 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

COTE = 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 COTE = **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 du 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.

2. INDICATEURS DE FRÉQUENCE

PREVALENCE = 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 de l'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).

3. INDICATEURS DE MORTALITE / MORBIDITE

TAUX DE MORBIDITE = c'est le rapport du **nombre de malades / la population totale.**

TAUX DE PREVALENCE = 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 statiques 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 » ce serait par contre à compter FAUX.

TAUX BRUT DE MORTALITE = c'est le **nombre de décès** (peu importe la cause) / **personnes** / **année**

TAUX SPECIFIQUE DE MORTALITE = c'est le **nombre de décès dû à la maladie étudiée** / **population totale** / **année**

Ces taux ont un intérêt pour les maladies chroniques.

Voilà pour ce cours assez simple, avec beaucoup de définitions et des notions qui seront reprises dans d'autres cours.
Comme d'hab', si vous avez un problème, go sur le forum. Bon courage les gars !

Grosse dédi à mes co-tuts parce qu'on est vraiment la meilleure team de tout le tut.
Dédi aux Chefs Tut' parce qu'ils sont tous trop chous.
Dédi à mes vieux que j'aime d'un amour incommensurable.
Dédi à Yanis, à son père et à son appart, à Victor, Cloé et Alexis (la chimie dans le cœur), à Amélie le Coma, à EP qui tient tellement pas l'alcool, à Carl (t'es mon champion, je suis si fière de toi), à Sarah parce que c'est une bonne rameuse, à Justine, à Charlotte, à Lucie, à Oumi la resta, à Blanblan, à Yanousa (ça va bien se passer), à Gio.gio.r, à Elena, à Diegz, à Tristan (ce dieu du shotgun), à Virgile (¡veux une D1 comme ça) et Eventut.
Dédi à ma co-marraine Mathilde et à tous mes fillots.