Module de Mathématiques Statistiques Chapitre 2 : Statistiques Descriptives à une dimension Séance 02

Responsables du cours: **Dr. Metiri Farouk et Dr. Sadoun Ahmed,**
Université de Badji Mokhtar -Annaba-

Département de TCSNV
2020/2021

Mail address: fmetiri@yahoo.fr

L'objectif de la Statistique Descriptive est de décrire de façon synthétique et parlante des données observées pour mieux les analyser. Le terme « statistique » est issu du latin « statisticum », c'est-à-dire qui a trait à l'état. Ce terme a été utilisé, semble-t-il pour la première fois, à l'époque de **Colbert**, par **Claude Bouchu**, intendant de Bourgogne, dans une « Déclaration des biens, charges, dettes et statistiques des communautés de la généralité de Bourgogne de 1666 à 1669».

Par contre, l'apparition du besoin « statistique » de posséder des données chiffrées et précises, précède sa dénomination de plusieurs millénaires. son origine, il est le fait de chefs d'états (ou de ce qui en tient lieu à l'époque) désireux de connaître des éléments de leur puissance: population, potentiel militaire, richesse, . . .

La statistique permet de répondre à de nombreuses questions biologiques.

Exemples :

- Quelle sont les valeurs normales de grandeurs biologiques (taille, poids, glycémie, ...) ?
- Les niveaux d'expression de deux gènes sont-ils différents ?
- Un nouveau traitement est-il plus efficace que le traitement de référence ?
- Peut-on définir de nouvelles typologies de tumeurs ?
- Un test de dépistage est-il fiable ?
- Les modifications de poids d'un individu sont-elles liées aux modifications de cholestérolémie ?
- Dynamique des populations.

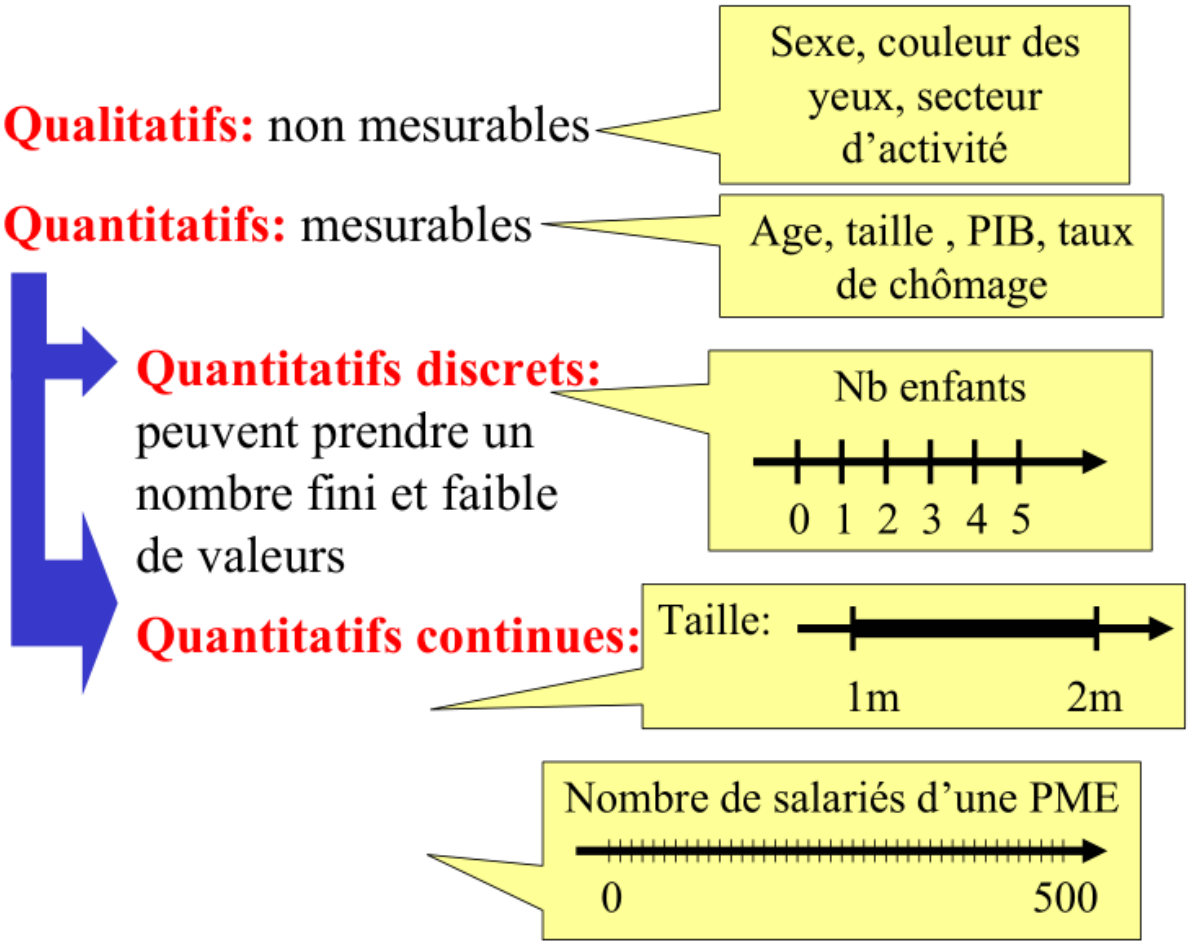
- Economie, assurance, finance : études quantitatives de marchés, prévisions économétriques, analyse de la consommation des ménages, taxation des primes d'assurances et de franchises, gestion de portefeuille, évaluation d'actifs financiers, ...
- Sciences de la terre : prévisions météréologiques, exploration pétrolière, ...
- Sciences humaines : enquêtes d'opinion, sondages, étude de population, ...
- Sciences de l'ingénieur : contrôle de qualité, sûreté de fonctionnement, évaluation des performances, ...
- Sciences de l'information : traitement des images et des signaux, reconnaissance de forme et parole, machine learning. ...

population : un ensemble d'éléments homogènes auxquels on s'intéresse. Par exemple, les étudiants d'une classe, les ménages Algériennes. . .

2. Les éléments de la population sont appelés les **individus** ou **unités statistiques.**

3. Des observations concernant un thème particulier ont été effectuées sur ces individus. La série de ces observations forme ce que l'on appelle une **caractère statistique.** Par exemple, les Notes des Etudiants à l'Examen de Statistique, les Mentions qu'ils ont obtenues à leur Bac, leur Sexe, les Couleurs de leurs Yeux, le Chiffre d'Affaire d'une entreprise, le Nombre d'Enfants par Ménage, le taux de glycémie . . .

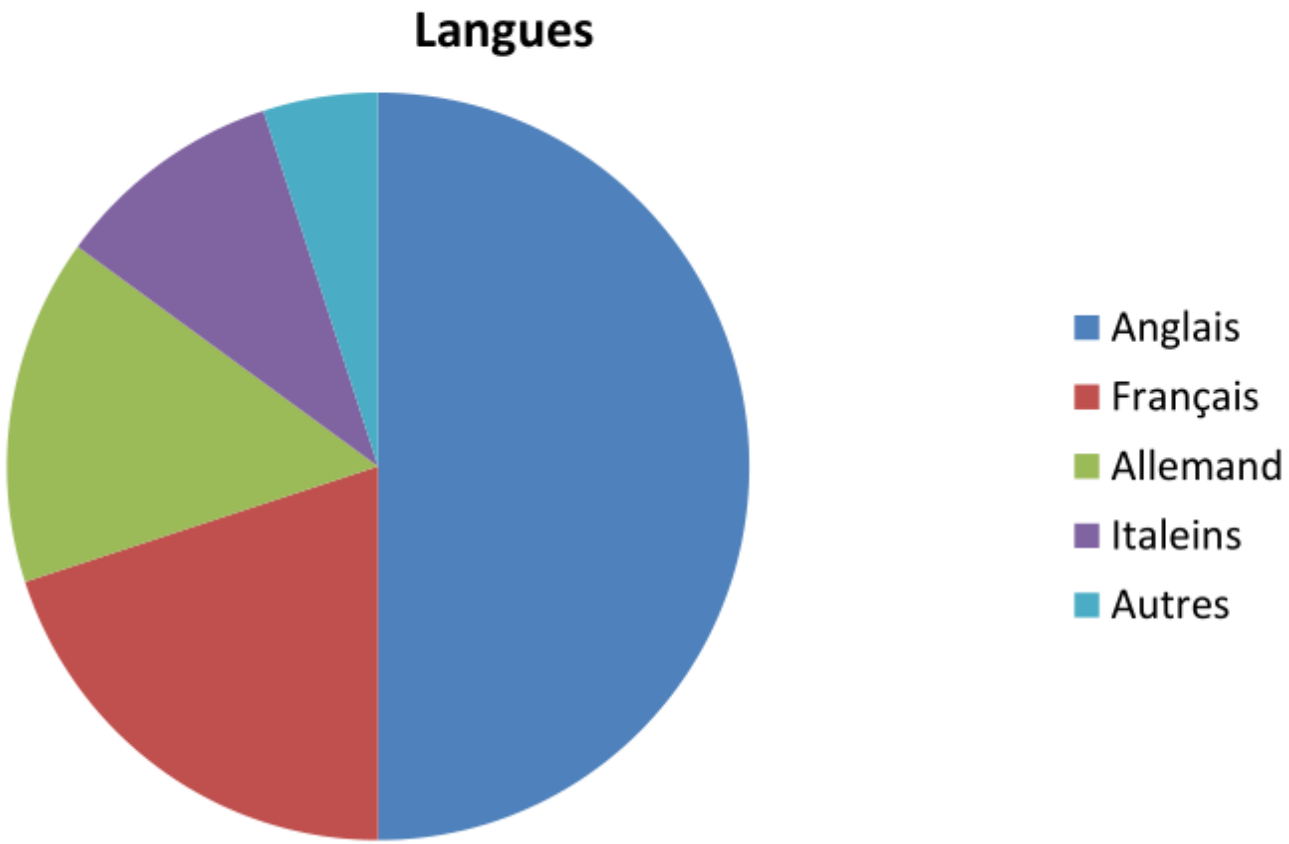
On distingue deux types de caractères:



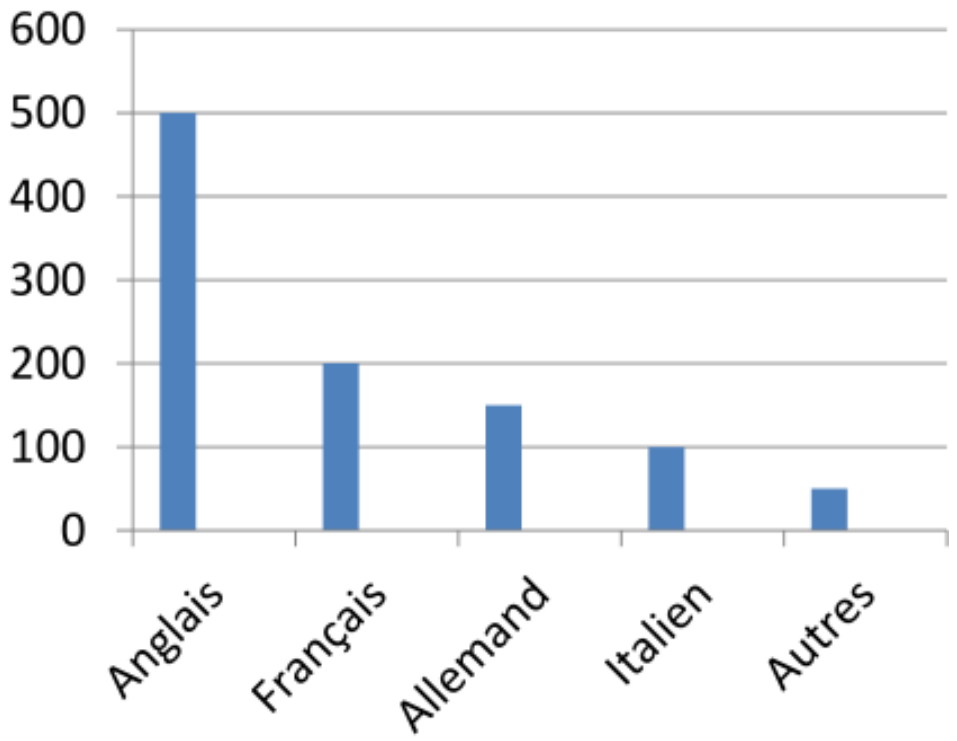
On distingue deux types de variables (**caractères**) qualitatives :
a- **Les variables qualitatives ordinales** : dans ce type, les modalités peuvent être classées dans un certain ordre naturel, c'est par exemple le cas de la variable *Mention au Bac* (passable, assez bien, bien, excellent...)
b- **Les variables qualitatives nominales** :
Une variable qualitative est dite nominale, lorsque les modalités ne peuvent pas être ordonnées de façon naturelle (c'est par exemple le cas de la variable *Couleur des Yeux* ou encore de la variable *Sexe*).

Représentation graphique d'un caractère qualitatif :
On a deux représentations possibles:
1)**Graphiques circulaires (camembert) :**
on dessine sur un disque des sections correspondants au modalités du caractère, et dont les angles au centre du disque sont proportionnels au pourcentages.
Exemple : Dans un lycée international, on s'intéresse à la langue étrangère étudiée pour 1000 étudiants. On a obtenu les résultats suivants:

Langue	Modalité
Anglais	500
Français	200
Allemand	150
Italien	100
Autre	50


Langues
Anglais
Français
Allemand
Italeins
Autres

2) Les bandes subdivisées :



Série statistique :
La forme la plus simple de présentation des données statistiques relatives à un seul caractère ou variable, consiste à une simple énumération des valeurs prises par le caractère.
Où la valeur N représente le nombre total des observations, aussi dit $effectif\ total$.
Exemple 01: Dans un immeuble habité par dix familles, on a compté le nombre d'enfants par foyer, les résultats suivants ont été obtenues:

$$1, 2, 3, 2, 4, 5, 3, 6, 2, 1$$

- L'**effectif** d'une valeur est le nombre de fois que cette valeur apparait dans la population, notée n_i
• On l'appelle aussi **fréquence absolue** ou tout simplement **fréquence**.

Effectifs Cumulés Croissants $(\mathbf{n_i^c}\uparrow)$ et Décroissants $(\mathbf{n_i^c}\downarrow)$:

• L'effectif cumulé croissant jusqu'à k est la somme des effectifs associés aux valeurs du caractère qui sont inférieurs ou égales à k.

• L'effectif cumulé décroissant associé à k est la somme des effectifs associés aux valeurs du caractère qui supérieures ou égales à k dans la série.

Pour faciliter la lecture des données, on regroupe toutes les données de la série statistique dans un tableau indiquant la répartition des individus selon le caractère étudié.

Classe statistique : Pour un caractère quantitatif continu, on considère les intervalles de type $[a, b[$ que l'on appelle Classes statistiques. La longueur $(b - a)$ est dite **amplitude** de la classe. La fréquence d'une classe est le *nombre d'observations* qui y sont continues.

Fréquences relatives f_i:

• La fréquence relative d'une valeur ou d'une classe est le quotient de l'effectif n_i par l'effectif total N,

On a

$$f_i = \frac{n_i}{N} = \frac{n_i}{\sum_i n_i} \text{ avec } \sum_i f_i = 1.$$

Nbre pers. à charge x_{oj}	Eff. n_j	Eff. cumulé N_j	Fréq. f_j	Fréq. cumulée F_j
0	10	10	0.10	0.10
1	17	27	0.17	0.27
2	33	60	0.33	0.60
3	20	80	0.20	0.80
4	15	95	0.15	0.95
5	5	100	0.05	1
Total	$n =$ 100		1	

Étude d'une variable quantitative discrète :

Exemple 02 :

Une enquête réalisée dans un village porte sur le nombre d'enfants à charge par famille.
On note X le nombre d'enfants, les résultats sont résumés dans ce tableau :

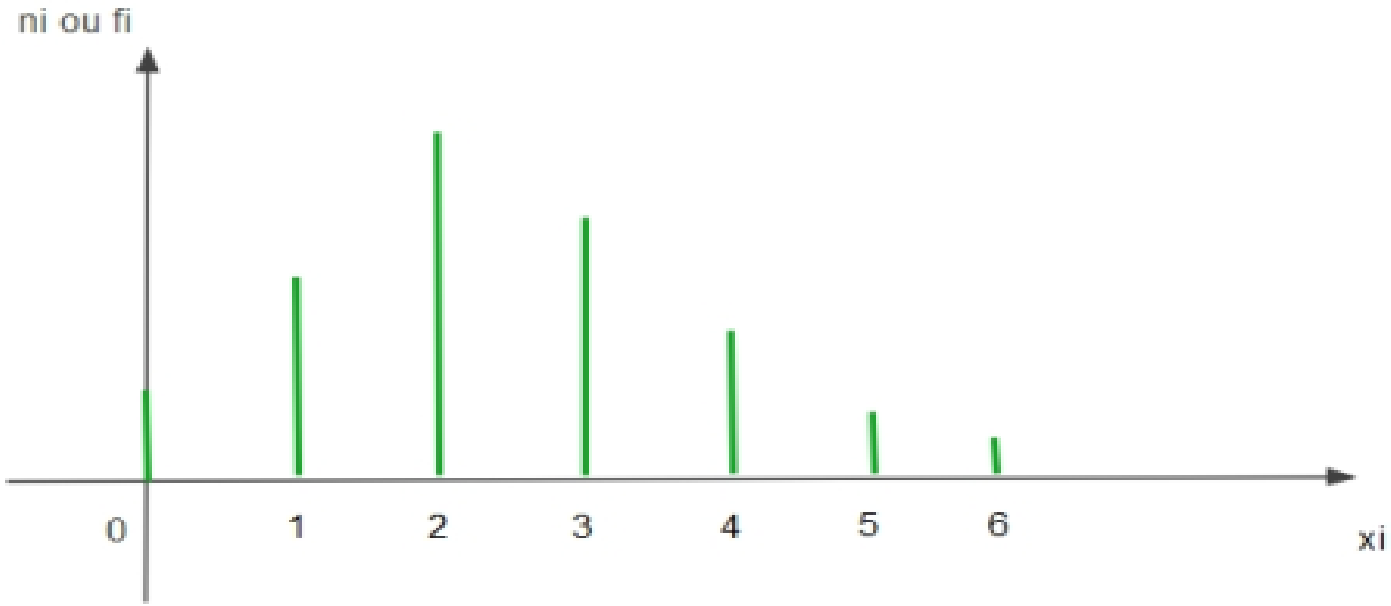
x_i	0	1	2	3	4	5	6
n_i	18	32	66	41	32	9	2

- On a fait une étude sur une polulation composée de 200 familles.
- Dans l'exemple précédent, 66 est le nombre de familles qui ont 2 enfants.
- Le caractère étudié est le *nombre d'enfants à charge par famille*, il est un caractère quantitatif discret.

x_i	n_i	$f_i = \frac{n_i}{n}$	$f_i^c \uparrow$	$n_i^c \uparrow$	$n_i^c \downarrow$
0	*18*	$\frac{18}{200} = 0.09$	0.09	*18*	200
1	32	$\frac{32}{200} = 0.16$	0.09+0.16=0.25	18+32=50	182
2	66	$\frac{66}{200} = 0.33$	0.25+0.33=0.58	50+66=116	150
3	41	$\frac{41}{200} = 0.205$	0.58+0.205=0.785	116+41=157	84
4	32	$\frac{32}{200} = 0.16$	0.785+0.16=0.945	157+32=189	43
5	9	$\frac{9}{200} = 0.045$	0.945+0.045=0.99	189+9=198	11
6	2	$\frac{2}{200} = 0.01$	0.99+0.01=1	198+2=200	2
	$\sum_i n_i = 200$	$\sum_i f_i = 1$			

Dans l'exemple précédent, il y a $0,33 = 33\%$ de familles dont le nombre d'enfants égale à 2.

Dans le paragraphe suivant, nous allons voir comment on traduit le tableau ci-dessus en graphique permettant aussi de résumer d'une manière visuelle les données.
Le *diagramme en bâtons* permet de représenter une variable quantitative discrète.
On veut représenter cette répartition sous la forme d'un diagramme en bâtons. Chaque marque correspond un bâton. Les hauteurs des bâtons sont proportionnelles aux effectifs représentés.



Paramètres de position (caractéristique de tendance centrale)

Les indicateurs statistiques de tendance centrale (dits aussi de position) considérés fréquemment sont *la moyenne, la médiane et le mode.*

1– Le mode

Le mode d'un caractère statistique est la valeur x_i qui a le plus grand effectif partiel (ou la plus grande fréquence partielle) et il est dénoté par M_0.

Dans l'exemple **02**, le mode est égal à 2 qui correspond au plus grand effectif.

NB : Le mode ou la classe modale n'est pas obligatoirement unique.

2– La moyenne

On appelle moyenne de X, la quantité $\bar{x} = \frac{1}{N}\sum_i n_i x_i$.

Si $\bar{x} = 2.46$, alors nous avons au moyenne 2.46 enfants pour chaque famille dans ce quartier.

3– La médiane

La médiane est un paramètre de position tel que la moitié des observations lui sont inférieures ou égales et l'autre moitié supérieures ou égales (la médiane partage la série statistique en deux groupes de même effectif).

- Quand le nombre d'observations N est *pair*, la médiane est la moyenne arithmétique des valeurs $x_{\frac{N}{2}}$ et $x_{\frac{N}{2}+1}$.

$$Me = Q_2 = \frac{x_{\frac{N}{2}} + x_{\frac{N}{2}+1}}{2}$$

Et si N est *impair*, alors

$$Me = x_{\frac{N+1}{2}}$$

4 – Les quartiles :

a- **Le premier quartile(noté Q_1)** : est la valeur d'une série qui est supérieure ou égale à au moins 25% des données de la série ordonnée de valeurs statistiques.

Appelons N le nombre total des valeurs d'une série, et calculons $\frac{N}{4}$.

Lorsque $\frac{N}{4}$ est entier, la valeur représentant le premier quartile est la $0,25$e valeur.

Lorsque $\frac{N}{4}$ est un décimal non entier, on l'arrondit à l'**entier supérieur** p et alors la valeur représentant le premier quartile est la $\mathbf{p}^{ième}$ valeur.

Attention : Pour le calcul des quartiles, la série doit d'abord être ordonnée selon l'ordre croissant (du plus petit au plus grand)

Exemple: Soit la série suivante:

$$10; 111; 30; 110; 41; 70; 55; 50; 42; 101; 40; 25$$

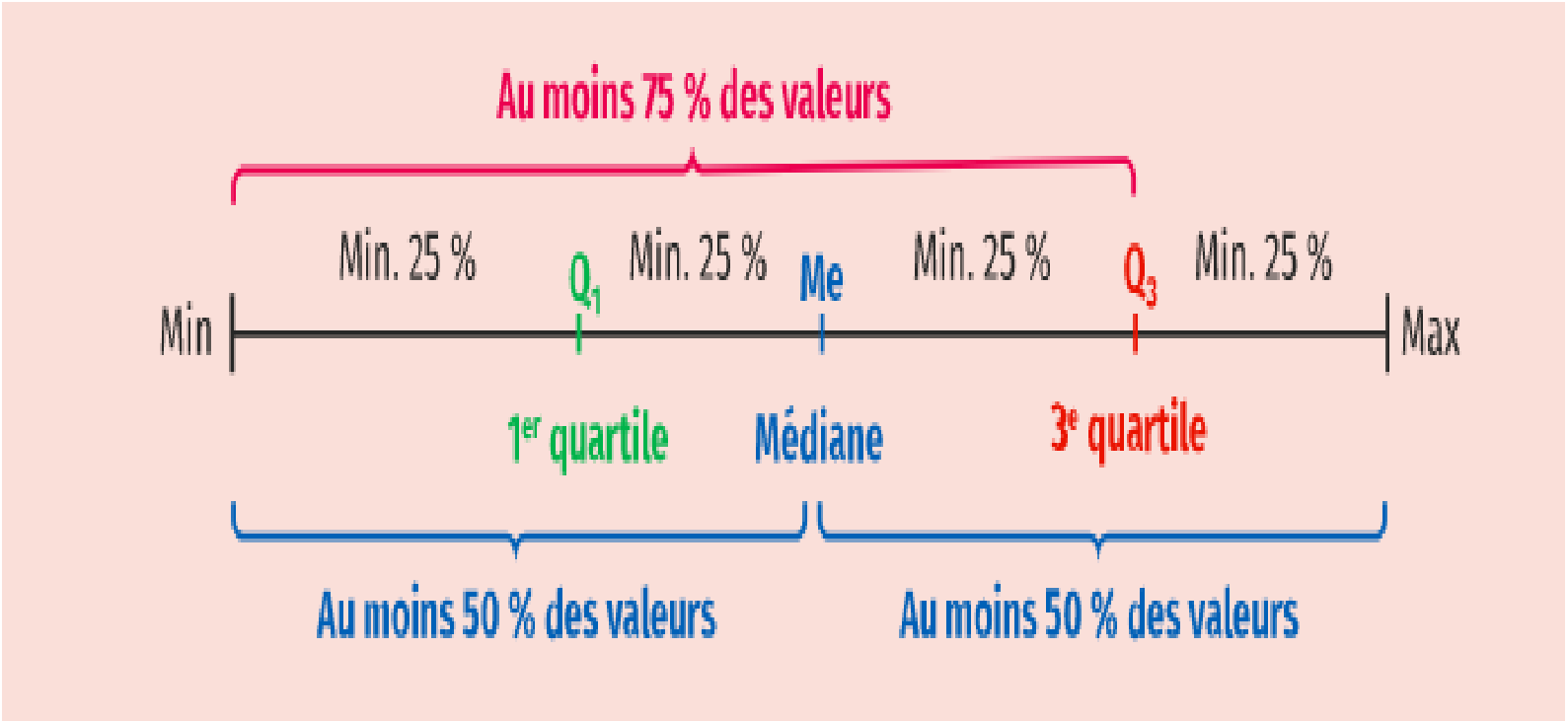
On ordonnee d'abord les valeurs:

$$10; 25; 30; 40; 41; 42; 50; 55; 70; 101; 110; 111$$

le premier quartile est 30. En effet, il y a 12 nombres dans cette série, et $\frac{12}{4} = 3$. Le premier quartile est donc la 3e valeur, soit 30.

- Si $\frac{N}{4}$=4,25, Q_1 est égale à la cinquième valeur (attention, ce n'est pas 5).

b- **Le troisième quartile (noté $\mathbf{Q_3}$) :**
C'est la valeur de la série qui est supérieure ou égale à au moins 75% des données de la série ordonnée de valeurs statistiques.
Lorsque $\frac{3N}{4}$ est entier, la valeur représentant le premier quartile est la $0,75$e valeur.
Lorsque $\frac{3N}{4}$ est un décimal non entier, on l'arrondit à l'entier supérieur p et alors la valeur représentant le troisième quartile est la $p^{ième}$ valeur.
Exemple: Dans la même série: 10; 25; 30; 40; 41; 42; 50; 55; 70; 101; 110; 111,
Le troisième quartile $\mathbf{Q_3}$ est 70. En effet, il y a 12 nombres dans cette série, et $\frac{3.12}{4} = 9$. Le troisième quartile est donc la 9e valeur, soit 70.
Exemple : si $\frac{3N}{4} = 15,25$, Q_3 est égale à la *seizième* valeur (attention, ce n'est pas 16).


Au moins 75 % des valeurs
Min. 25 %
Min. 25 %
Min. 25 %
Min. 25 %
Min
Q1
Me
Q3
Max
1er quartile
Médiane
3e quartile
Au moins 50 % des valeurs
Au moins 50 % des valeurs

Paramètres de dispersion (variabilité):

Les indicateurs statistiques de dispersion usuels sont la distance interquartile, l'étendue, la variance, l'écart type et le coefficient de variation.

1- **La distance interquartile :** représente la différence entre Q_3 et Q_1, elle indique la moitié des valeurs centrales.

Exemple : Dans la même série 10; 25; 30; 40; 41; 42; 50; 55; 70; 101; 110; 111,

l'écart interquartile est 40. En effet, Q_3 valant 70 et Q_1 valant 30, il suffit de calculer $70 - 30$.

2 – La variance $(V(x))$ et l'écart type :

La variance est la moyenne des carrés des écarts à la moyenne

Si x a pour unité la personne, alors σ^2 a pour unité personne^2

$$\sigma^2 = \frac{1}{N}\sum_i n_i \left(x_i - \bar{x}\right)^2 = \sum_i f_i \left(x_i - \bar{x}\right)^2$$

L'écart-type est la racine carré de la variance

Même unité que le caractère

$$\sigma = \sqrt{\sigma^2}$$

Entre $\bar{x} - 2\sigma$ et $\bar{x} + 2\sigma$ il y a au moins 75% de la population

L'écart type **mesure donc la dispersion autour de la moyenne.**
En raison de ses liens étroits avec la moyenne, l'écart-type peut être grandement influencé si la moyenne donne une mauvaise mesure de tendance centrale.
N.B : L'écart-type et la variance **ne sont jamais négatifs.**

3 – L′étendue :
La différence entre la plus grande valeur et la plus petite valeur du caractère, donnée par la quantité

$$e = x_{max} - x_{min},$$

s'appelle l'étendue de la V.S X. Le calcul de l'étendue est très simple. Il donne une première idée de la dispersion des observations mais il reste un indicateur très rudimentaire.

4 – Le coefficient de variation :
Le coefficient de variation (désigné par CV) se définit par la relation suivante :

$$CV = \frac{L'écart\ type}{La\ moyenne}.100\% = \frac{\sigma}{\bar{x}}.100\%$$

Le coefficient de variation est une mesure relative de dispersion (puisque l'écart-type est rapporté à la moyenne).
on dit que la série est homogène c'est à dire . Sinon elle est hétérogène.
Donc :
Si $CVX < 25\%$: on a une faible dispersion = la série est *homogène* = les observations sont regroupées autour de la moyenne.
Dans les deux cas suivants, la série est *hétérogène* :

$$\left.\begin{array}{c} 25\% < CV < 80\% : \text{les observations sont assez dispersées} \\ CVX > 80\% : \text{une très forte dispersion= les observations sont dispersées ou éloignées par rapport à la moyenne.} \end{array}\right\}$$