I VOCABULAIRE : (VIDÉO 1)

Une étude statistique commence par un recueil de données.
Prenons un exemple classique avec des notes :
Dans le tableau suivant sont regroupées les notes obtenues par les élèves d'une classe de seconde lors du contrôle n° 1 :

4	5	6	6	6	8	8	9	10	11	11	11	12	12	12	12	13
13	14	14	16	16	16	16	16	16	16	17	17	17	17	19	19	19

DÉFINITIONS FONDAMENTALES

- **Série statistique** : Ensemble des valeurs collectées.
  Dans notre exemple, la série statistique est l'ensemble des notes collectées.
- **Population** : Ensemble sur lequel porte l'étude statistique.
  Dans notre exemple, la population est l'ensemble des élèves de seconde .
- **Individus** : Éléments qui composent la population.
  Dans notre exemple, les individus sont chacun des élèves de seconde .
- **Caractère étudié** : Propriété que l'on observe sur les individus.
  Dans notre exemple, le caractère étudié est le résultat obtenu au contrôle n° 1.
- **Modalité** :Les différentes valeurs obtenues sont appelées **valeurs du caractère** ou **modalités**, souvent notées x_1, x_2,… x_p.
  Dans notre exemple, les modalités sont les notes obtenues au contrôle n° 1.
- **Types de caractères** :
  Un caractère peut être **qualitatif** si on étudie une qualité des individus : (situation de famille, sexe, couleur des yeux, type d'habitation…).
  Un caractère peut être **quantitatif** si on mesure une quantité des individus :
  Dans ce cas, il est dit **discret** lorsqu'il ne prend que des valeurs isolées (nombre d'enfants, notes dans une classe...).
  Il est dit **continu** lorsqu'il peut prendre théoriquement toutes les valeurs d'un intervalle (taille, temps d'écoute...) ; dans ce cas, les valeurs sont regroupées en intervalles appelés des **classes**.

  Dans notre exemple, le caractère est quantitatif discret.
- **Effectif** :
  Pour une valeur du caractère (modalité ou classe), on appelle effectif le nombre d'individus de la population ayant cette valeur.
  On note souvent n_1, n_2,… n_p les effectifs respectifs des modalités x_1, x_2,… x_p.
  Dans notre exemple, la valeur x_1 du caractère est 4, la valeur x_2 du caractère est 5, la valeur x_3 du caractère est 6, etc..
  Les effectifs correspondants sont $n_1 = 1$, $n_2 = 2$, $n_3 = 3$, etc..
- **Effectif total** :
  Nombre total d'individus de la population (ou de l'échantillon).
  Il est égal à $n_1 + n_2 + \ldots + n_p$, souvent noté N.
  Dans notre exemple, l'effectif total est le nombre d'élèves de la classe, à savoir 34.

Fréquence :
Pour une valeur du caractère , on appelle fréquence le quotient de l'effectif de cette valeur par l'effectif total. La fréquence peut être exprimée en pourcentage.

$$\text{fréquence} = \frac{\text{effectif de la valeur}}{\text{effectif total}}$$

On note souvent f_1, f_2,..., f_p les fréquences respectives des modalités x_1, x_2,... x_p, donc :

$$f_1 = \frac{n_1}{N} \quad , \quad f_2 = \frac{n_2}{N} \quad , \quad \ldots, \quad f_p = \frac{n_p}{N}.$$

On en déduit que : $0 \leq f_1 \leq 1$, $0 \leq f_2 \leq 1, \ldots,$ $0 \leq f_p \leq 1$
et

$$f_1 + f_2 + \ldots f_p = 1$$

Dans notre exemple, pour une meilleure lisibilité et pour simplifier l'étude, on peut commencer par compter le nombre d'individus ayant obtenu chaque note :

Note	4	5	6	8	9	10	11	12	13	14	16	17	19
Effectif	1	1	3	2	1	1	3	4	2	2	7	4	3
Fréquence à 10^{-2} près	0,03	0,03	0,09	0,06	0,03	0,03	0,09	0,12	0,06	0,06	0,21	0,12	0,09

On lit par exemple que $x_3 = 6$, $n_3 = 3$ et $f_3 \approx 0{,}09$
Remarque
Dans le tableau précédent, la somme des fréquences est supérieure à 1 à cause des arrondis.

Effectif cumulé :
Pour une valeur x d'une série statistique quantitative, l'effectif cumulé croissant (respectivement décroissant) de x est la somme des effectifs des valeurs inférieures (respectivement supérieures) ou égales à x. Dans notre exemple :

Note	4	5	6	8	9	10	11	12	13	14	16	17	19
Effectif	1	1	3	2	1	1	3	4	2	2	7	4	3
ECC	1	2	5	7	8	9	12	16	18	20	27	31	34

Les Effectifs Cumulés Croissants (ECC) permettent de déterminer le nombre d'individus ayant une valeur inférieure ou égale à une modalité :
On peut par exemple déduire que 9 élèves ont une note inférieure ou égale à 10.
La dernière valeur des ECC est l'effectif total, puisque toutes les modalités sont inférieures ou égales à la valeur maximum de la série.

Fréquence cumulée :
Pour une valeur x d'une série statistique quantitative, la fréquence cumulée croissante (respectivement décroissante) de x est la somme des fréquences des valeurs inférieures (respectivement supérieures) ou égales à x.

Pour calculer les FCC, pn procède de la même manière que pour les ECC.

Note	4	5	6	8	9	10	11	12	13	14	16	17	19
Effectif	1	1	3	2	1	1	3	4	2	2	7	4	3
FCC 10^{-2} près	0,03	0,06	0,15	0,21	0,24	0,27	0,36	0,48	0,54	0,60	0,81	0,93	1

EXEMPLE D'UNE SÉRIE CONTINUE

On a interrogé en 2008 un échantillon de 4812 Français concernant la durée hebdomadaire d'écoute de la télévision (en heures).
Le caractère étudié, à savoir la durée d'écoute, est quantitatif continu : il peut prendre théoriquement toutes les valeurs de l'intervalle [0 ; 50].
Les données sont regroupées en classes [0 ; 10], [10 ; 15[, [15 ; 20[, [20 ; 30[et [30 ; 50].

Durée	[0 ; 10[	[10 ; 15[	[15 ; 20[	[20 ; 30[	[30 ; 50]
Effectif	972	924	826	1069	1021

II REPRÉSENTATIONS GRAPHIQUES

1 SÉRIES À CARACTÈRE QUANTITATIF DISCRET

DIAGRAMME EN BÂTONS

Dans un **diagramme en bâtons**, on représente une série statistique discrète par des segments dont la hauteur est proportionnelle à l'effectif de la valeur qu'ils représentent.

Exemple On continue à travailler avec les données de l'exemple sur les notes. Voici le diagramme en bâtons de cette série :

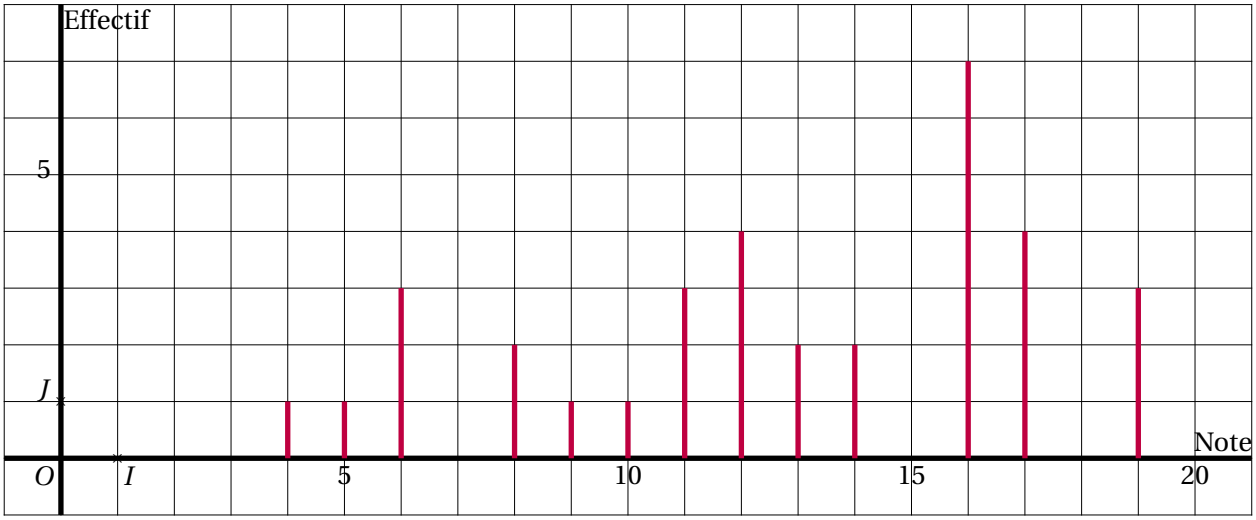


DIAGRAMME CIRCULAIRE

Exemple :
Dans une compétition d'athlétisme, quatre pays s'affrontent : la France, l'Allemagne, la Suède et la Norvège. On note le pourcentage de médailles obtenues par chacun des pays :

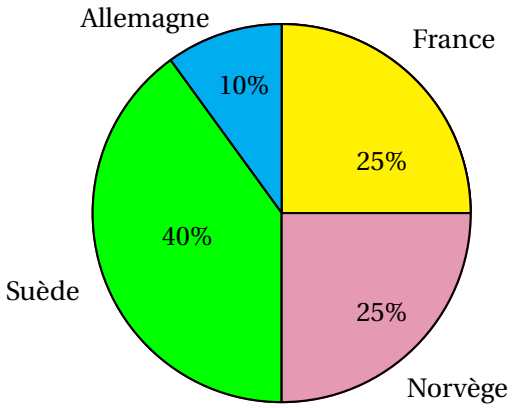
Pays	France	Allemagne	Suède	Norvège
Pourcentage de médailles	25 %	10 %	40 %	25 %

Représenter le diagramme circulaire associé à cette série statistique :

Pays	Total	France	Allemagne	Suède	Norvège
Pourcentage de médailles	100 %	25 %	10 %	40 %	25 %
Angle en degrés	360	90	36	144	90

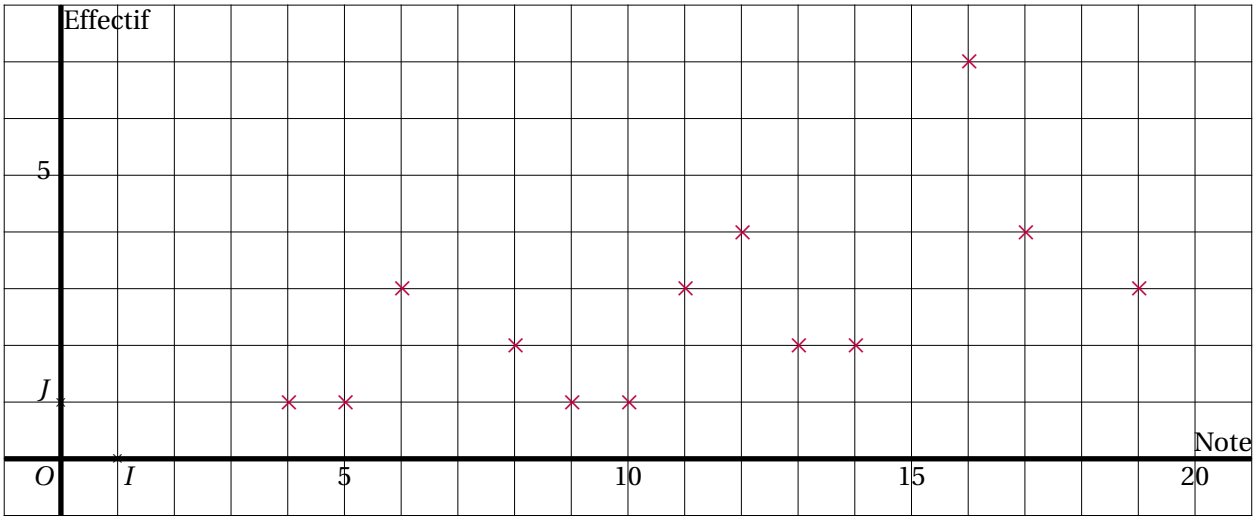
Pour cela, nous avons besoin des angles; nous les calculons par proportionnalité, sachant que 100 % correspondent à 360°.

Pourcentage des médailles obtenues par pays :



NUAGE DE POINTS

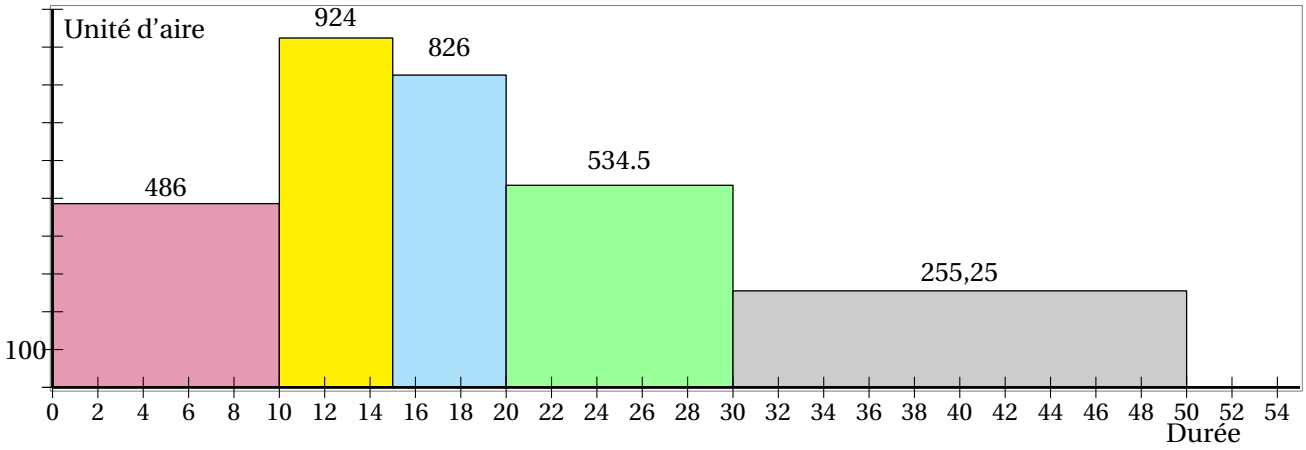
Dans un **nuage de points**, on représente une série statistique discrète par des points dont les abscisses sont les valeurs du caractère, et les ordonnées sont les effectifs correspondants, parfois reliés par des segments.
Exemple On travaille toujours avec les données de l'exemple sur les notes. Voici le nuage de points de cette série :



2 SÉRIES À CARACTÈRE QUANTITATIF CONTINU

HISTOGRAMME

Dans un **histogramme**, on représente une série statistique continue par des rectangles dont la **largeur** correspond à l'**amplitude** de chaque classe et dont l'**aire** est **proportionnelle** à l'**effectif** de la classe.
Exemple
On travaille avec les données de l'exemple sur la durée d'écoute de la télévision. Voici l'histogramme de cette série :



Remarque

Lorsque les classes ont toutes la même amplitude, la hauteur de chaque rectangle est proportionnelle à l'effectif de la classe qu'il représente. On dit alors que l'histogramme est à **pas constant**.

POLYGONE D'EFFECTIFS OU DE FRÉQUENCES CUMULÉS

- Le **polygone des effectifs cumulés croissants** (respectivement **décroissants**) d'une série statistique continue est la ligne brisée qui joint les points du plan dont les abscisses sont les bornes de chaque classe et dont les ordonnées sont les effectifs cumulés croissants (respectivement décroissants) de ces valeurs.
- Le **polygone des fréquences cumulées croissantes** (respectivement **décroissantes**) d'une série statistique continue est la ligne brisée qui joint les points du plan dont les abscisses sont les bornes de chaque classe et dont les ordonnées sont les fréquences cumulées croissantes (respectivement décroissantes) de ces valeurs.

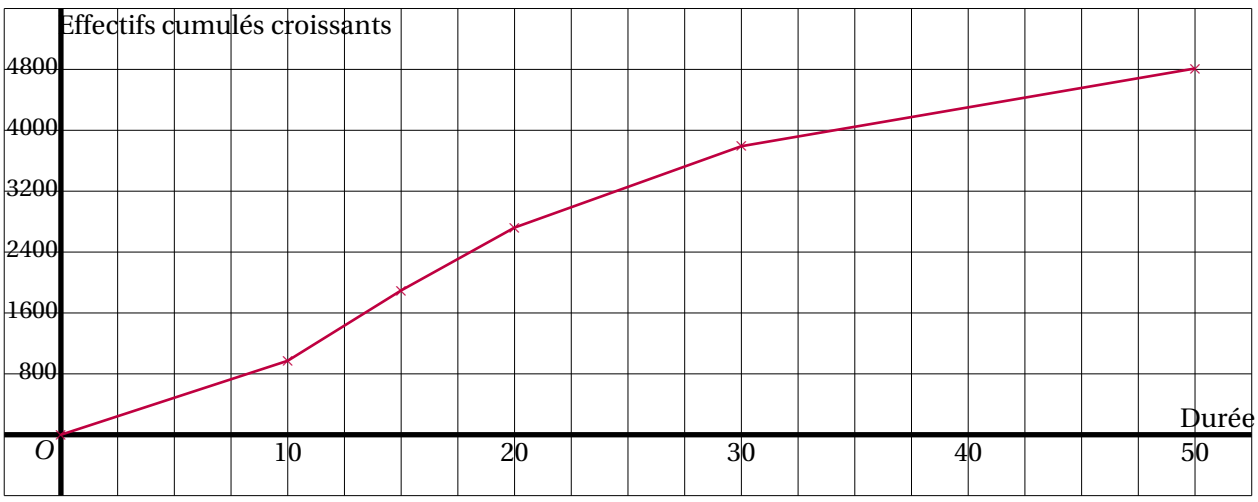
Ces représentations donnent l'allure de la répartition des valeurs de la série.

Exemple

La situation est toujours celle de l'exemple de la page 3 sur le temps d'écoute de la télévision. Le tableau des effectifs cumulés croissants est le suivant :

Durée	0	10	15	20	30	50
ECC	0	972	1896	2722	3791	4812

D'où le polygone des effectifs cumulés croissants :



Cela permet de répondre aux questions du type :
« *Combien de personnes regardent moins 20 heures la télévision ?* »

Traitons à présent le cas des fréquences cumulées décroissantes :

Durée ≥	0	10	15	20	30	50
Effectif	4812	3840	2916	2090	1021	0
Fréquence	1	0,8	0,61	0,43	0,21	0

D'où le polygone des fréquences cumulées décroissantes :

Cela permet de répondre aux questions du type :
« *Quel est le pourcentage de personnes regardant plus de 30 heures la télévision ?* »

III PARAMÈTRES DE POSITION :

1 LA MOYENNE :

On considère une série statistique donnée par le tableau suivant :

Valeur	x_1	x_2	x_3	…	x_{p-1}	x_p
Effectif	n_1	n_2	n_3	…	n_{p-1}	n_p
Fréquence	f_1	f_2	f_3	…	f_{p-1}	f_p

DÉFINITION

La **moyenne** de cette série statistique est le réel noté $\overline{x}$ défini par

$$\overline{x} = \frac{n_1x_1 + n_2x_2 + \ldots + n_px_p}{n_1 + n_2 + \ldots n_p} = \frac{n_1x_1 + n_2x_2 + \ldots + n_px_p}{N}$$

en notant $N = n_1 + n_2 + \ldots + n_p$ l'effectif total de la série.

PROPRIÉTÉ :

On peut également calculer la moyenne à l'aide des fréquences :

$$\overline{x} = x_1 f_1 + x_2 f_2 + \ldots + x_p f_p.$$

EXEMPLE

Dans un service de maintenance, on a répertorié le nombre d'interventions par jour sur un mois. On a obtenu la distribution suivante :

Nombre d'interventions x_i	3	5	6	7	8	9
Nombre de jours n_i	2	4	9	6	3	1

Le nombre moyen d'interventions par jour est :

$$\overline{x} = \frac{2\times 3 + 4\times 5 + 9\times 6 + 6\times 7 + 3\times 8 + 1\times 9}{25} = 6{,}2$$

ou en utilisant les fréquences :

Nombre d'interventions x_i	3	5	6	7	8	9
Nombre de jours n_i	2	4	9	6	3	1
Fréquence f_i	0,08	0,16	0,36	0,24	0,12	0,04

$$\overline{x} = 0{,}08\times 3 + 0{,}16\times 5 + 0{,}36\times 6 + 0{,}24\times 7 + 0{,}12\times 8 + 0{,}04\times 9 = 6{,}2$$

2 MÉDIANE

DÉFINITION

La **médiane** M d'une série statistique est un réel qui partage cette série en deux parties telles que :

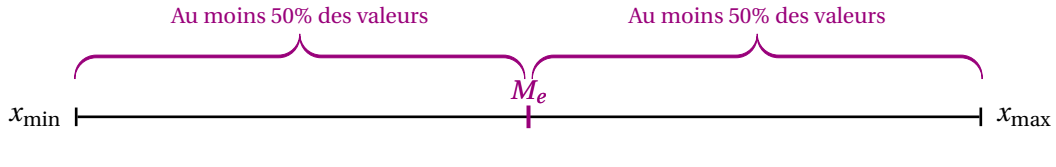
- Au moins 50 % des valeurs sont inférieures ou égales à la médiane ;
- Au moins 50 % des valeurs sont supérieures ou égales à la médiane.

PROPRIÉTÉ

En pratique, on adopte la démarche suivante pour déterminer la médiane M d'une série statistiques d'effectif total N :

- On range d'abord les N valeurs du caractère par ordre croissant.
- Si N est pair, M est la moyenne des deux valeurs « centrales » de la série.
- Si N est impair, M est la valeur centrale de la série.

ILLUSTRATION :



EXEMPLE

Dans la série précédente sur le nombre d'interventions par jour du service de maintenance, l'effectif total $N = 25$ donc l'effectif est impair.
La médiane est la valeur centrale de la série, celle du caractère de rang 13 soit $Me = 6$.
Ce qui signifie qu'au moins la moitié du temps, le nombre d'interventions par jour est inférieur ou égal à 6.

Dans l'exemple des notes du début du cours (page 2), l'effectif total est 34, c'est-à -dire pair.
La médiane est donc la moyenne des deux valeurs centrales de la série, à savoir les 17e et 18e valeurs.
Donc $M = \dfrac{13+13}{2} = 13$, ce qui signifie qu'au moins la moitié des notes est inférieure ou égale à 12 (en réalité 18 notes), et qu'au moins la moitié des notes est supérieure ou égale à 12 (en réalité 18 notes également).

ATTENTION :

Il ne faut pas confondre la **valeur** de la médiane et son **rang**.
Dans l'exemple sur les notes, le **rang** de la médiane est entre la 17e et 18e valeurs, mais sa **valeur** est 12.
On doit donc bien distinguer ces deux éléments pour la médiane : On cherche d'abord le **rang** pour déterminer ensuite la **valeur**.

MÉDIANE OU MOYENNE ?

La moyenne est très sensible à des valeurs extrêmes.
La série représente la répartition des salaires dans une entreprise :

salaire en € x_i	1200	1500	1800	2000	2200	25000
Nombre de salariés n_i	2	4	9	6	3	1

Le salaire moyen dans l'entreprise est :

$$\overline{x} = \frac{2\times 1200 + 4\times 1500 + 9\times 1800 + 6\times 2000 + 3\times 2200 + 1\times 25000}{26} = 2728$$

La moyenne à 2728€ est très sensible à la valeur extrême de 25 000€. Il faut donc être prudent quand on interprète une moyenne, qui ne donne aucune information sur la répartition des valeurs.
La moyenne est un paramètre de position et ne doit être interprété que comme tel.
Dans cet exemple, la médiane est de 1800 € (On cherche la valeur de rang 13). Elle est moins sensible aux valeurs extrêmes.
C'est pour cela que pour caractériser la position des salaires d'un pays, le salaire médian est plus pertinent que le salaire moyen.
Mais la médiane est aussi un paramètre de position et ne permet pas d'étudier la dispersion des valeurs autour de sa valeur.

Source : .

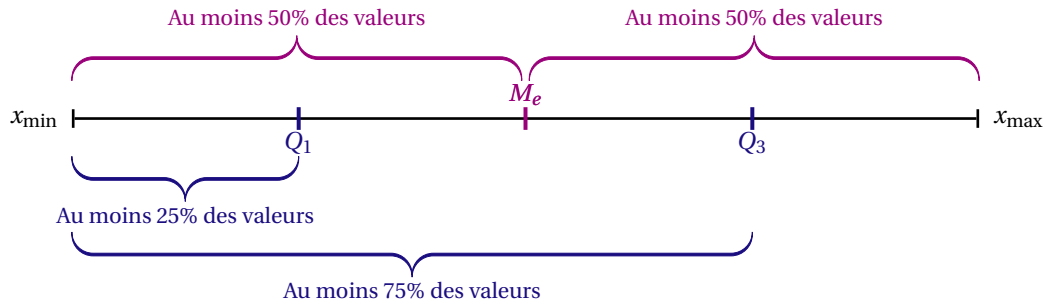
3 QUARTILES

DÉFINITION

On considère une série statistique.

- Le premier **quartile** Q_1 est la plus petite valeur de la série telle qu'au moins 25 % des données soient inférieures ou égales à Q_1.
- Le troisième **quartile** Q_3 est la plus petite valeur de la série telle qu'au moins 75 % des données soient inférieures ou égales à Q_3.

ILLUSTRATION :



MÉTHODE

Soit une série statistique d'effectif total N :
En pratique, on calcule le quart de l'effectif, soit $\frac{N}{4}$, puis $\frac{3N}{4}$.

- Si quotient est un nombre entier, il donne respectivement le **rang** de Q_1 ou de Q_3, attention, pas sa **valeur**, son **rang**!!
- Si le quotient n'est pas un nombre entier, le **rang** du quartile est arrondi par excès à son entier supérieur.

EXEMPLE

On considère toujours les données de l'exemple des notes du début de cours :

- $\frac{34}{4} = 8{,}5$ donc le **rang** de Q_1 est 9. D'où $Q_1 = 10$, ce qui signifie qu'au moins un quart des notes sont inférieures ou égales à 10 .
- $\frac{3\times 34}{4} = 25{,}5$ donc le **rang** de Q_3 est 26, d'où $Q_3 = 16$, ce qui signifie qu'au moins trois quarts des notes sont inférieures ou égales à 16 .

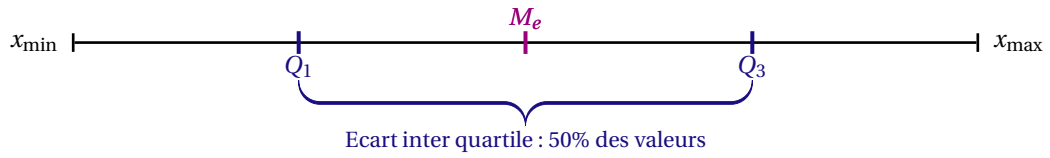
IV PARAMÈTRES DE DISPERSION

1 ECART INTER QUARTILE :

DÉFINITION :

On appelle écart interquartile la différence entre le troisième et le premier quartiles : $Q_3 - Q_1$

ILLUSTRATION :



ASSOCIATION :

En pratique, on associe l'**écart inter quartile**, paramètre de dispersion, à la **médiane**, paramètre de position.

BOÎTES À MOUSTACHES

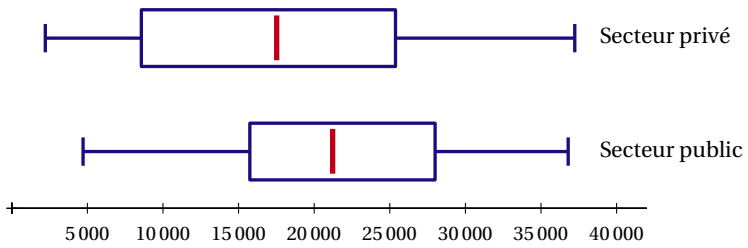
Il est commode d'illustrer la médiane et les quartiles d'une série par un diagramme, appelé **diagramme en boîte**, ou « boîte à moustaches »
Pour une catégorie donnée, on construit :

- Un **axe** permettant de repérer les valeurs de la variable étudiée,
- Un **rectangle** dont la longueur est égale à l'**écart interquartile** $Q_3 - Q_1$,
- Un **trait** représentant la **médiane**.
- Deux **traits** représentant les valeurs extrêmes de la série.

EXEMPLE

Le tableau suivant donne la distribution du revenu salarial dans deux entreprises :

	Minimum	Q1	Médiane	Q3	Maximum
Secteur privé	2 218	8 570	17 520	25 377	37 234
Secteur public	4 716	15 744	21 221	27 996	36 797



REMARQUE

Le fait que le partage théorique en 25 %, 50 % et 75 % de la série statistique à l'aide des indicateurs Q_1, M et Q_3 ne soit pas tout à fait exact provient du fait que la série comporte des valeurs identiques. Ce phénomène a tendance à s'amoindrir lors d'une étude sur une population plus importante avec un caractère dont les modalités sont plus disparates.

2 VARIANCE ET ECART-TYPE

PRINCIPE

L'idée est de définir un paramètre de dispersion, qui associé à la moyenne, permettrait de définir un couple de paramètre pratique pour étudier une série statistiques.
Prenons les cas extrêmes de deux classes, une où tous les élèves ont 10/20, l'autre ou la moitié à 0/20 et l'autre moitié 20/20.
Les deux classes ont la même moyenne de 10/20 mais on comprend bien que la deuxième a une dispersion des notes bien plus importante que la première.
Pour mesurer cette dispersion, l'idée est de calculer **l'écart de chaque valeur avec la moyenne**.
On calcule donc $x_i - \overline{x}$ pour chaque modalité.
On obtient $0 - 10 = -10$ pour les élèves qui ont 0, et $20 - 10 = 10$ pour les élèves qui ont 20.
Si on ajoute les écarts à la moyenne, les valeurs positives vont compenser les valeurs négatives. On ne quantifiera donc pas la dispersion.
Pour éviter cela, une solution est de les élever au carré avant de les ajouter, pour n'avoir que des nombres positifs qui se cumulent.
Puis de diviser par l'effectif total pour se ramener à une valeur correspondant à un individu. C'est ce qu'on appelle la **variance** de la série.
Et enfin, pour rendre le résultat plus cohérent avec la série, pour compenser le fait qu'on ait élevé au carré, on calcule la racine carrée du résultat. C'est ce qu'on appelle **l'écart-type** de la série.

DÉFINITIONS :

La variance V d'une série statistique de moyenne $\bar{x}$ dont les valeurs caractère sont x_1, x_2, x_3, …, x_k et les effectifs correspondants sont n_1, n_2, n_3, …, n_k est égale à

$$V = \frac{n_1 \times (x_1 - \bar{x})^2 + n_2 \times (x_2 - \bar{x})^2 + \cdots + n_k \times (x_k - \bar{x})^2}{n_1 + n_2 + \cdots + n_k}$$

L'écart-type σ d'une série statistique de variance V est égal à : $\sigma = \sqrt{V}$

EXEMPLE

Dans l'exemple du service de maintenance, on avait calculé et le nombre moyen d'interventions par jour :$\overline{x} = 6{,}2$
Rajoutons une ligne au tableau de valeurs :

Nombre d'interventions x_i	3	5	6	7	8	9
Nombre de jours n_i	2	4	9	6	3	1
$x_i - \overline{x}$	-3,2	-1,2	-0,2	0,8	1,8	2,8
$(x_i - \overline{x})^2$	10,24	1,4	0,04	0,64	3,24	7,84
$n_i(x_i - \overline{x})^2$	20,48	5,6	0,36	3,84	9,72	7,84

$$\begin{aligned} V &= \frac{n_1 \times (x_1 - \bar{x})^2 + n_2 \times (x_2 - \bar{x})^2 + \cdots + n_k \times (x_k - \bar{x})^2}{n_1 + n_2 + \cdots + n_k} \\ &= \frac{20{,}48 + 5{,}6 + 0{,}36 + 3{,}84 + 9{,}72 + 7{,}84}{25} \\ &= \frac{42{,}84}{25} \\ &= 1{,}7136 \end{aligned}$$

$$\begin{aligned} \sigma &= \sqrt{V} \\ &= \sqrt{1{,}7136} \\ &\approx 1{,}31 \end{aligned}$$

REMARQUE :

La variance n'est utilisée à notre niveau que comme un outil qui permet de calculer l'écart-type.
Son calcul étant fastidieux, pour les séries qui comportent trop de modalités, on se contente d'une valeur donnée à la calculatrice. Voir tutoriel sur mathsguyon.fr