2^{nde} : Statistiques

Table des matières

I Vocabulaire

Une étude statistique commence par un recueil de données. On utilise le vocabulaire suivant pour décrire cette étude :

- **Série statistique** : Ensemble des valeurs collectées.
- **Population** : Ensemble sur lequel porte l'étude statistique.
- **Individus** : Éléments qui composent la population.
- **échantillon** : Partie de la population.
- **Caractère étudié** : Propriété que l'on observe sur les individus. Les différentes valeurs obtenues sont appelées **valeurs du caractère** ou **modalités**, souvent notées x_1, $x_2, \dots x_p$. On distingue deux types de caractères.
  - ◇ Un caractère peut être **qualitatif** (situation de famille, sexe, couleur des yeux, type d'habitation...).
  - ◇ Un caractère peut être **quantitatif**. 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**.
- **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, $n_2, \dots n_p$ les effectifs respectifs des modalités x_1, $x_2, \dots x_p$.
- **Effectif total** : Nombre total d'individus de la population (ou de l'échantillon). Il est égal à $n_1 + n_2 + \dots + n_p$, souvent noté N.
- **Fréquence** : Pour une valeur du caractère (modalité ou classe), on appelle fréquence le quotient de l'effectif de cette valeur par l'effectif total. On note souvent f_1, $f_2, \dots,$ f_p les fréquences respectives des modalités x_1, $x_2, \dots x_p$, donc :
$$f_1 = \frac{n_1}{N},\ f_2 = \frac{n_2}{N}, \dots,\ f_p = \frac{n_p}{N}.$$
  On en déduit que $0 \leq f_1 \leq 1$, $0 \leq f_2 \leq 1, \dots, 0 \leq f_p \leq 1$, et $f_1 + f_2 + \dots f_p = 1$.
- **Valeurs extrêmes** : Valeurs minimales et maximales d'un caractère quantitatif.
- **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.
- **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.

Exemple avec des notes :

Dans le tableau suivant sont regroupées les notes obtenues par les élèves d'une seconde lors du contrôle n° 1 (éventuellement arrondies pour simplifier l'étude) :

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

Dans cet exemple :

- La série statistique est l'ensemble des notes collectées.
- La population est l'ensemble des élèves de seconde .
- Les individus sont chacun des élèves de seconde Turner.
- Le caractère étudié est le résultat obtenu au contrôle n° 1.
- Les modalités sont les valeurs chiffrées des notes obtenues au contrôle n° 1.
- L'effectif total est le nombre d'élèves de la classe, à savoir 34.
- Les valeurs extrêmes sont 4 et 19.

Exemple avec des notes : 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

Remarque
Dans le tableau précédent, la somme des fréquences est supérieure à 1 à cause des arrondis.

Exemple : série continue On a interrogé en 2008 un échantillon de 4812 Français concernant la durée hebdomadaire d'écoute de la télévision (en heures) :

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

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].

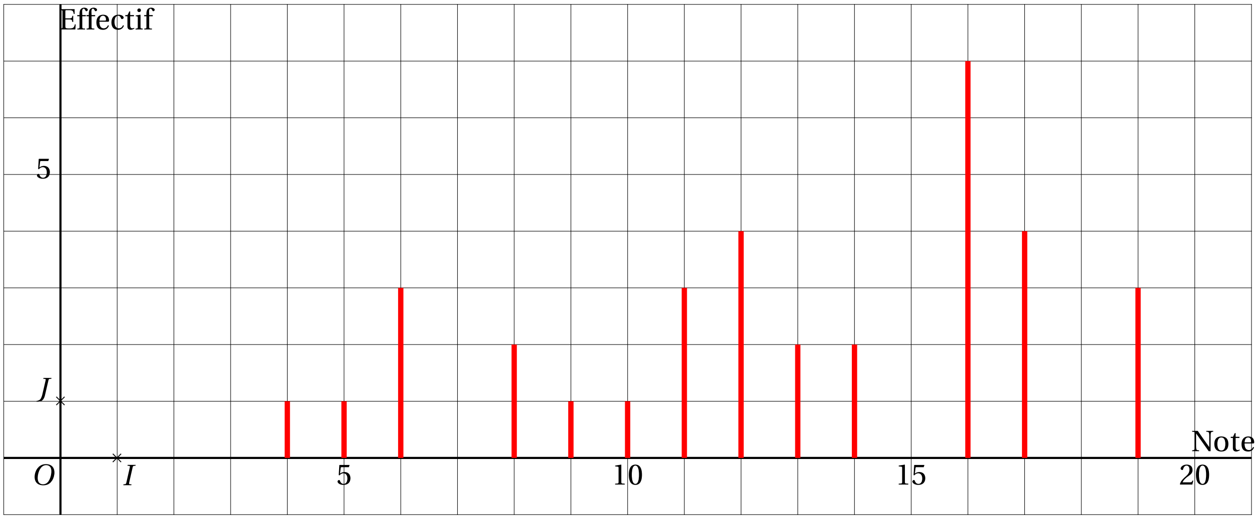
II Représentations graphiques

II.1 Séries à caractère quantitatif discret

II 1.1 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 :



II 1.2 Diagramme circulaire

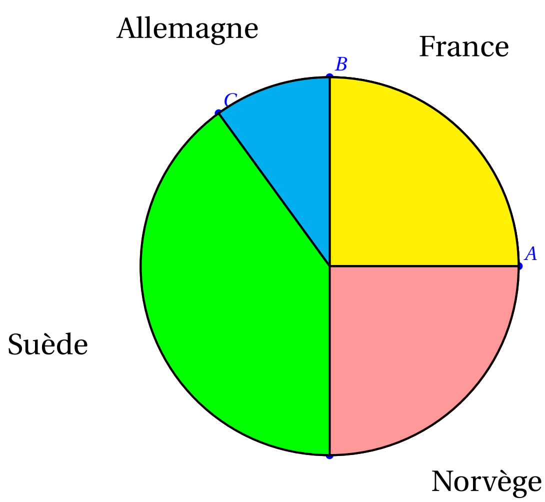
Exemple :
Dans une compétition d'athlétisme, quatre pays s'affontent : la France, l'Allemagne, la Suède et la Norvège. On note le pourcentage de médailles obtenues par chacun des pays :

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 °	360	90	36	144	90

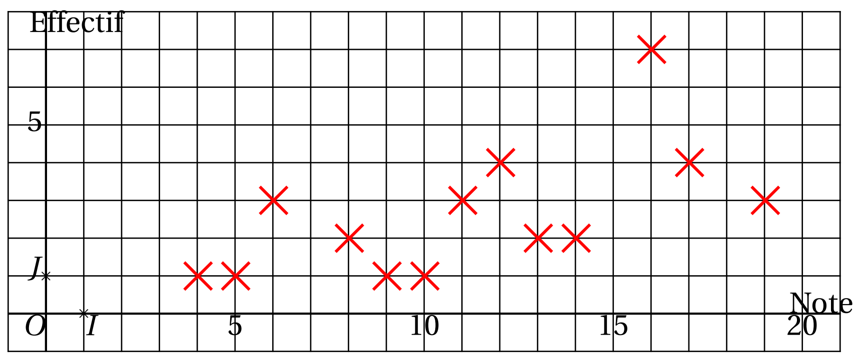
Pour cela, nous avons besoin des angles ; nous les calculons par proportionnalité, sachant que 100 % correspondent à 360°.



II 1.3 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 :



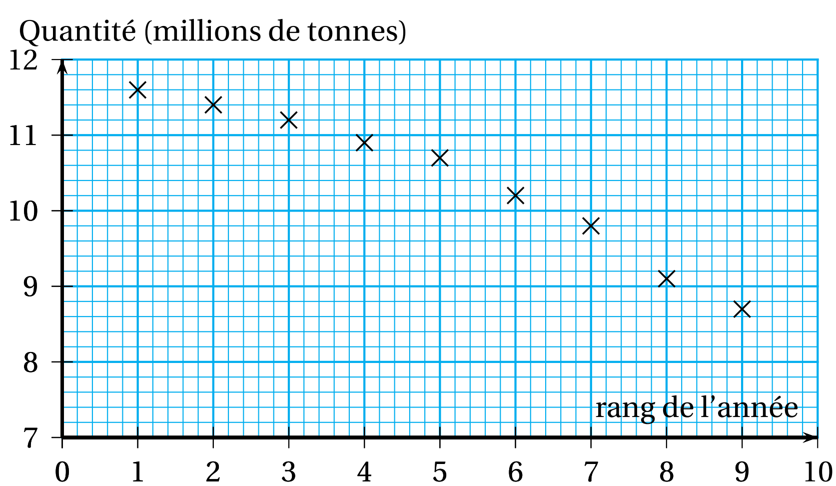
Autre exemple, tiré du sujet de baccalauréat ES Polynésie juin 2012 :

Le tableau ci-dessous donne les quantités de super sans plomb livrées et vendues en France de 2001 à 2009 (les quantités sont exprimées en millions de tonnes).

Année	2001	2002	2003	2004	2005	2006	2007	2008	2009
Rang de l'année x_i	1	2	3	4	5	6	7	8	9
Quantité (millions de tonnes) y_i	11,6	11,4	11,2	10,9	10,7	10,2	9,8	9,1	8,7

Source : INSEE

On représente alors les données sous forme de nuage de points en mettant en abscisses le rang de l'année et en ordonnées les quantités en (millions de tonnes)



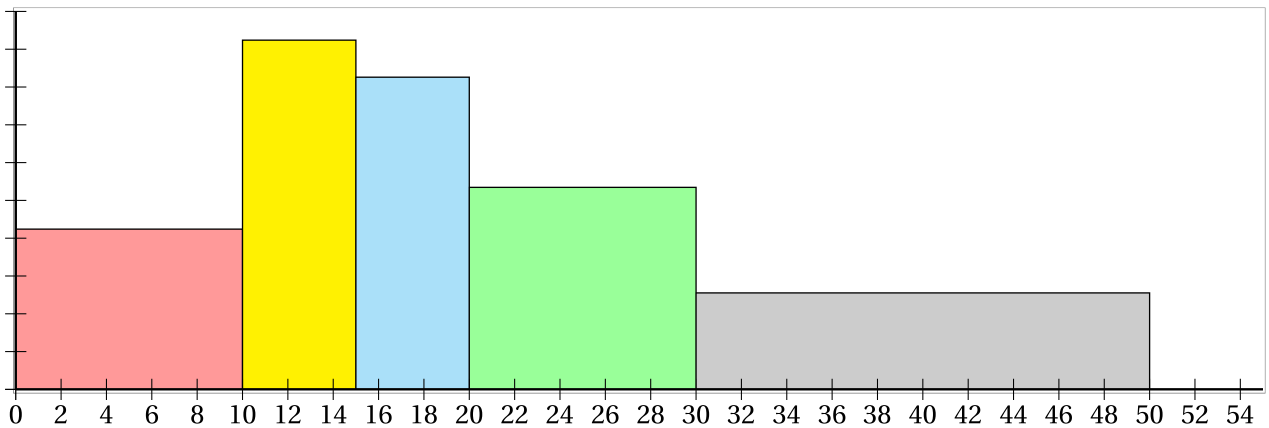
II.2 Séries à caractère quantitatif continu

II 2.1 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. (voir ici)
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**.

II 2.2 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.

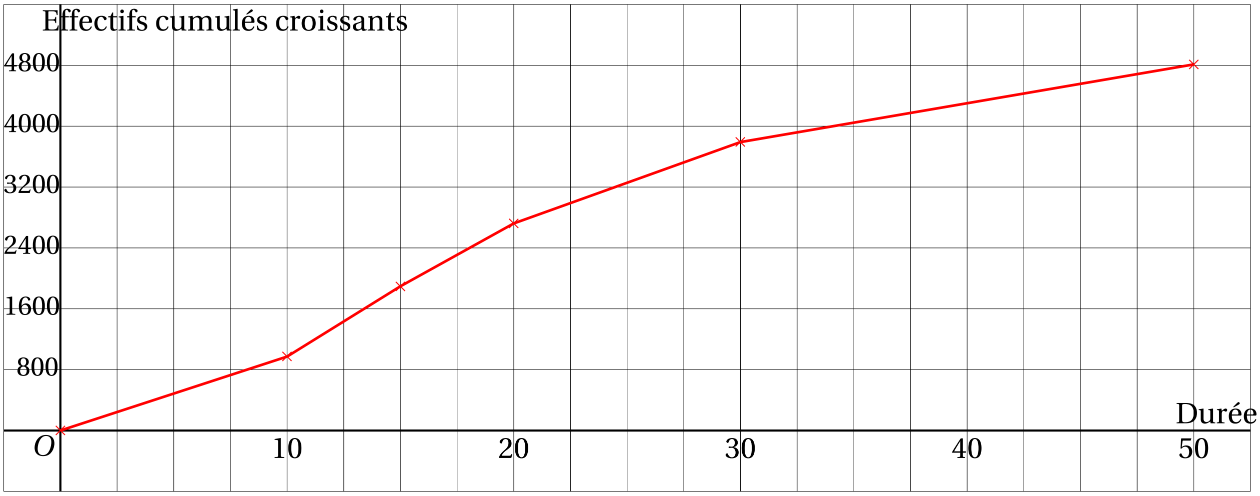
Ces représentations donnent l'allure de la répartition des valeurs de la série.

Exemple

La situation est toujours celle de l'exemple I. Le tableau des effectifs cumulés croissants est le suivant :

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

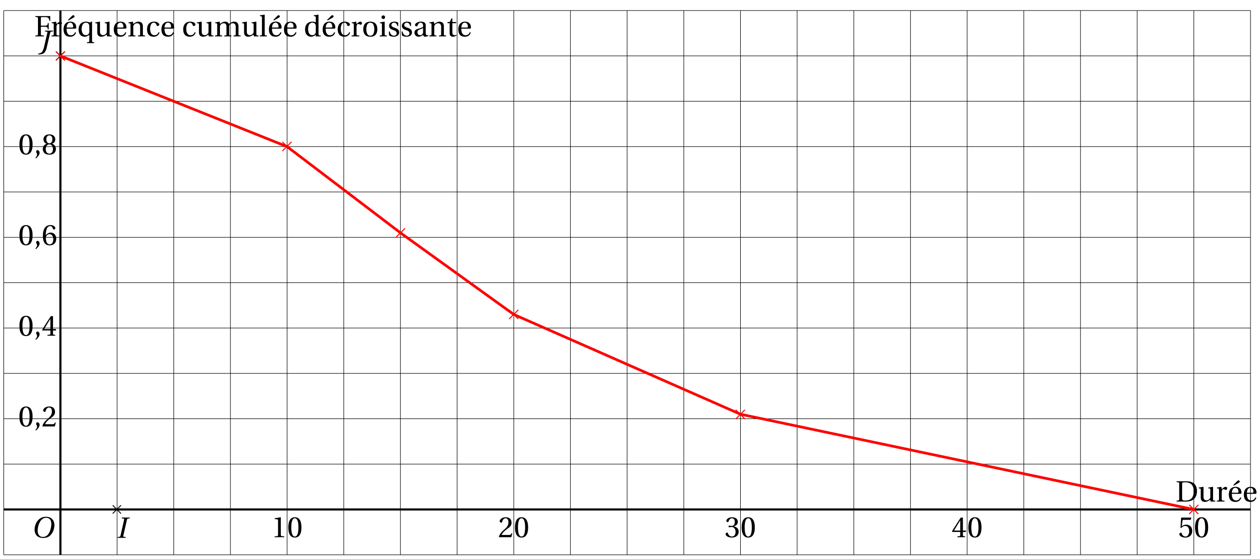
D'où le polygone des effectifs cumulés croissants :



Traitons à présent le cas des fréquences cumulées décroissantes :

Durée $\geq$	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 :



III Paramètres statistiques

III.1 Paramètres de position

III 1.1 Mode

Définition :
Un **mode** d'une série statistique est une valeur de la série dont l'effectif est strictement supérieur à celui des autres valeurs.

Remarque
Dans une série statistique, il peut y avoir plusieurs modes.
Exemple
Dans l'exemple sur les notes, le mode est 16.

III 1.2 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 :

$$\overline{x} = x1f_1 + x_2f_2 + \ldots + x_pf_p.$$

Exemple
Avec les données de l'exemple I, la moyenne de la classe est :

$$\overline{x} = \frac{1\times4+1\times5+3\times6+2\times8+1\times9+1\times10+3\times11+4\times12+2\times13+2\times14+7\times16+4\times17+3\times19}{34} = \frac{434}{34}$$

Une valeur approchée de $\overline{x}$ à 10^{-2} près est 12,76.

III 1.3 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, $N = 2p$, M est la moyenne des deux valeurs « centrales » de la série, c'est-à-dire $M = \dfrac{x_N + x_{N+1}}{2}$.
- Si N est impair, $N = 2p+1$, M est la valeur centrale de la série, c'est-à-dire x_{p+1}.

Exemple

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

Exemple

En France, en 2005, dans le secteur privé et semi-public (SNCF, Poste, Caisse d'épargne...), le salaire net mensuel médian est de 1528 €, (ce qui signifie que la moitié des salariés de ces entreprises touchent moins que cette valeur) alors que le salaire net mensuel moyen est de 1903,5 €. (somme de tous les salaires divisée par le nombre de salariés)

Source : .

III.2 Paramètres de dispersion

III 2.1 Étendue

Définition

L'**étendue** d'une série statistique est la différence entre sa plus grande et sa plus petite valeur.

Exemple

Dans l'exemple sur les notes, l'étendue est égale à $19 - 4 = 15$.

III 2.2 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.

Exemple

On considère toujours les données de l'exemple I.

- $34 \times \dfrac{25}{100} = 8,5$ donc le premier quartile Q_1 de la série est la 9^e^valeur, d'où $Q_1 = 10$, ce qui signifie qu'au moins un quart des notes sont inférieures ou égales à 10 (en réalité 9 notes, soit environ 26 %).
- $34 \times \dfrac{75}{100} = 25,5$ donc le troisième quartile Q_3 de la série est la 26^e^ valeur, d'où $Q_3 = 16$, ce qui signifie qu'au moins trois quarts des notes sont inférieures ou égales à 16 (en réalité 27 notes, soit environ 79 %).

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.

III 2.3 Écart interquartile

Définition

On considère une série statistique de premier quartile Q_1 et de troisième quartile Q_3.

- On appelle **intervalle interquartile** l'intervalle $[Q_1 ; Q_3]$.
- On appelle **écart interquartile** la différence $Q_3 - Q_1$.

Exemple

Dans l'exemple I, l'intervalle interquartile est $[10 ; 16]$ et l'écart interquartile est $Q_3 - Q_1 = 16 - 10 = 6$.