

## 3

## Statistiques descriptives

## I. Vocabulaire

Une **étude statistique** a pour but d'obtenir une information, appelée **caractère**, sur une population à partir de **données** recueillies sur un **échantillon** de cette population.

**Définition 1**

Le **caractère** étudié peut être :

**quantitatif** : les valeurs du caractère s'expriment avec des nombres (ex : températures, pointures, salaires...);

**qualitatif** : les valeurs ne s'expriment pas par des nombres (ex : couleurs, type d'essence...);

**discret** : les valeurs du caractère sont isolés (ex : notes...);

**continu** : les valeurs sont regroupées par classes (ou intervalles de nombre) (par ex : durée, distance parcourue,

**Remarque**

Dans la suite du cours, on considère que le caractère est quantitatif.

**Définition 2**

Lorsque les valeurs d'une série statistique sont regroupés par classe de type  $[a; b[$ , on appelle **centre de classe** le nombre défini par  $\frac{a+b}{2}$ .

## II. Indicateurs de position

## A. Moyenne

**Définition 3**

On considère une série qui possède des valeurs différentes  $x_1, x_2, \dots, x_p$  chacune affectée de leur effectif. L'effectif total est égal à  $N$ .

On appelle **moyenne** d'une série le nombre  $\bar{m}$  tel que :

$$\bar{m} = \frac{n_1 x_1 + n_2 x_2 + n_3 x_3 + \dots + n_p x_p}{n_1 + n_2 + \dots + n_p} = \frac{\sum_{i=1}^p n_i x_i}{N}.$$

**Exemple •** Un élève souhaite calculer la moyennes de ses notes sur 20 : 8 ; 10 ; 14 ; 13 ; 10 ; 14 ; 10.

$$\bar{m} = \frac{8 + 10 + 14 + 13 + 10 + 14 + 10}{7} = \frac{79}{7} \approx 11,29.$$

Parfois, les valeurs sont regroupées dans un tableau :

Notes	8	10	13	14
Effectifs	1	3	1	2

$$\text{On peut alors utiliser la formule : } \bar{m} = \frac{1 \times 8 + 3 \times 10 + 1 \times 13 + 2 \times 14}{1 + 3 + 1 + 2} = \frac{79}{7} \approx 11,29.$$

### Remarques

- 1°) En règle générale, la moyenne n'est pas une valeur de la série. On peut la considérer comme le point d'équilibre des valeurs. Par conséquent, la moyenne est sensible aux valeurs extrêmes.  
Pour reprendre l'exemple, si la prochaine note du contrôle est très élevée alors la moyenne va augmenter.
- 2°) Lorsque les valeurs sont regroupées par classe, au lieu d'utiliser les  $x_i$ , on utilise les centres de classe.

**Exemple •** Le tableau ci-dessous regroupe le temps de parcours des habitants d'un village entre leur domicile et leur lieu de travail. Le maire cherche à calculer le temps moyen.

Durée en min	[0 ; 10[	[10 ; 20[	[20 ; 30[	[30 ; 50[	[50 ; 70[	Total
Effectif	18	35	25	112	80	270
Centre de classe	5	15	25	40	60	

$$\text{On a donc : } \bar{m} = \frac{18 \times 5 + 35 \times 15 + 25 \times 25 + 112 \times 40 + 80 \times 60}{270} = \frac{10\,520}{270} \approx 39 \text{ min.}$$

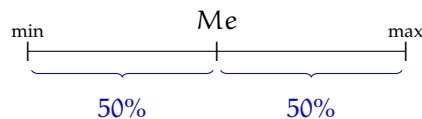
## B. Médiane

### Définition 4

On considère une série statistique dont l'effectif total est égal à  $N$ . Les valeurs sont rangées dans l'ordre croissant.

Une **médiane**  $Me$  est un nombre réel qui permet de partager la série statistique en deux séries de même valeur.

Autrement dit, la moitié (50%) des valeurs de la série est inférieure ou égale à  $Me$  et l'autre moitié est supérieure ou égale à  $Me$ .



**Méthode de calcul :** Par définition, la médiane dépend de l'effectif de la série :

- Si  $N$  est impair, alors on calcule  $\frac{N+1}{2}$  et le résultat correspond à la position de la médiane choisie dans la série.
- Si  $N$  est pair, alors la médiane choisie est égale à la moyenne de la valeur située à la position  $\frac{N}{2}$  et la valeur suivante.

**Exemple •** On considère le relevé des températures en janvier et en février dans une ville :

Janvier									
Valeurs	-3°	-2°	-1°	0°	1°	2°	3°	4°	Total
Effectifs	3	5	8	5	4	3	2	1	31
Février									
Valeurs	-3°	-2°	-1°	0°	1°	2°	3°	4°	Total
Effectifs	1	2	3	3	5	9	3	2	28

**En janvier :**  $N = 31$  donc  $\frac{N+1}{2} = 16$  : une médiane possible est la 16<sup>e</sup> valeur donc  $Me = -1^\circ$ . En janvier, il a fait moins de  $-1^\circ$  la moitié du mois.

**En février :**  $N = 28$  donc  $\frac{N}{2} = 14$  : la 14<sup>e</sup> est  $1^\circ$  et la suivante est égale à  $2^\circ$ . Donc la moyenne des deux est  $\frac{1+2}{2} = 1,5$ . Une médiane possible est égale à  $1,5^\circ$  donc durant la moitié du mois, la température a été supérieure à  $1,5^\circ$ .

### Remarque

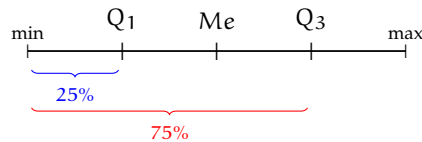
Contrairement à la moyenne, la médiane n'est pas sensible aux valeurs extrêmes. En effet, dans notre exemple, si la dernière température était égale à 20° au lieu de 4°, la moyenne augmenterait alors que la médiane resterait identique car l'effectif total n'a pas changé.

## C. Quartiles

### Définition 5

On considère une série statistique  $S$  dont l'effectif total est égal à  $N$ . Les valeurs sont rangées dans l'ordre croissant.

- Le **premier quartile**  $Q_1$  de  $S$  est le plus petit élément  $a$  de  $S$  tel qu'au moins 25% des données soient inférieures ou égales à  $a$ .
- Le **troisième quartile**  $Q_3$  de  $S$  est le plus petit élément  $b$  de  $S$  tel qu'au moins 75% des données soient inférieures ou égales à  $b$ .



**Méthode de calcul :** Par définition, les quartiles dépendent de l'effectif de la série :

**Premier quartile :** On arrondit le nombre  $\frac{N}{4}$  à l'unité par excès et cela donne la position de  $Q_1$  dans la série  $S$ .

**Troisième quartile :** On arrondit le nombre  $3 \times \frac{N}{4}$  à l'unité par excès et cela donne la position de  $Q_3$  dans la série  $S$ .

**Exemple •** On reprend les températures du mois de janvier de l'exemple précédent.

**Premier quartile :**  $N = 31$  donc  $\frac{N}{4} = 7,75 \approx 8$  donc la 8<sup>e</sup> valeur est  $Q_1 = -2^\circ$ .

**Troisième quartile :**  $N = 31$  donc  $3 \times \frac{N}{4} = 23,25 \approx 24$  donc la 24<sup>e</sup> valeur donne  $Q_3 = 1^\circ$ .

Ainsi, 25% des valeurs sont inférieures ou égales à  $-2^\circ$  et 75% des valeurs sont inférieures ou égales à  $1^\circ$ . On peut dire aussi que 25% des valeurs sont supérieures ou égales à  $1^\circ$ .

## III. Indicateurs de dispersion

Dans les classes antérieures, l'**étendue** était un indicateur de dispersion utilisé dont voici rappelée la définition :

### Définition 6

Dans une série statistique, l'**étendue** est la différence entre la valeur maximale et la valeur minimale.

### Remarque

Une valeur élevée de l'étendue signifie qu'au moins une des valeurs extrêmes de la série est éloignée de la médiane et le risque de dispersion des valeurs est donc plus important.

Dans l'exemple précédent, l'étendue de la série janvier était identique à celle de la série février ce qui montre les besoins d'avoir d'autres indicateurs.

Les quartiles nous permettent d'obtenir d'autres indicateurs de dispersion liés à la médiane.

## A. Intervalle et écart interquartile

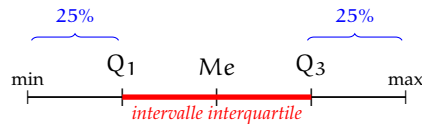


### Définition 7

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



### Remarque



La figure ci-dessus nous permet de comprendre pourquoi l'intervalle interquartile contient 50% de valeurs de la série, c'est-à-dire la moitié. Ainsi, un écart interquartile faible impose que les valeurs soient regroupées proches de la médiane et donc peu dispersées pour la moitié d'entre elles.

Voici maintenant un indicateur de dispersion lié à la moyenne.

## B. Écart-type

**Exemple •** Voici les notes obtenues par Paul dans quelques matières lors des deux premiers trimestres de l'année.

**Trimestre 1 :** 7 ; 8 ; 11 ; 12 ; 13 ; 13 ; 13.

**Trimestre 2 :** 4 ; 7 ; 9 ; 12 ; 13 ; 13 ; 19.

Bien qu'au deuxième trimestre Paul ait obtenu une très bonne note, on remarque qu'il a aussi obtenu une mauvaise note. Ses parents lui explique qu'il a été moins régulier ce trimestre et par conséquent, moins sérieux.

Paul se défend en calculant la moyenne et la médiane. Dans chaque cas, il trouve une moyenne égale à 11 et une médiane égale à 12 donc à chaque trimestre, la moitié de ses notes est supérieure à 12. Ses parents lui font tout de même constater que l'étendue du premier trimestre est égale à 6 ( $13 - 7$ ) alors qu'elle vaut 15 ( $19 - 4$ ) au deuxième trimestre.

Il réplique en indiquant que les écarts interquartiles des deux trimestres sont peu différents.

Pour faire entendre raison à leur fils, les parents doivent trouver un autre indicateur. Ils s'intéressent à l'écart de chaque note par rapport à la moyenne et constatent que cet écart est plus important au deuxième trimestre, ce qui montre des notes plus dispersées, moins régulières. Ces écarts avec la moyenne leur permettent de calculer tout d'abord ce que l'on appelle la **variance** puis enfin l'**écart-type**.



### Définition 8

L'**écart-type** est un nombre réel positif qui caractérise la dispersion des valeurs d'une série statistiques par rapport à la moyenne.  
Plus l'écart-type est petit et plus les valeurs sont proches de la moyenne (dispersion faible). Au contraire, si l'écart-type est grand alors les valeurs sont éloignées de la moyenne (dispersion élevée).



### Remarque

En 1<sup>ère</sup> STMG, on utilisera la calculatrice ou le tableur pour calculer directement l'écart-type.

**Exemple •** Dans un tableur, dans la colonne A, on inscrit les notes obtenues par Paul lors du premier trimestre. À la suite de ses notes, on inscrit la formule `=ECARTYPEP(A1:A7)` pour obtenir la valeur arrondie au dixième 2,3. Cet écart-type, seul, nous permet de dire que les valeurs sont peu dispersées autour de la moyenne. Mais l'écart-type a surtout un intérêt comparatif entre deux séries dont les valeurs sont exprimées dans une même unité de mesure.

Dans la colonne B, on inscrit maintenant les notes du deuxième trimestre puis, en dessous, on inscrit `=ECARTYPEP(B1:B7)`. Le tableur renvoie la valeur arrondie 4,5.

Cette fois, on peut affirmer que les notes du deuxième trimestre sont plus dispersées que celles du premier trimestre par rapport à la moyenne.



### Remarque

Sur les nouvelles versions d'EXCEL, il est préférable d'utiliser la fonction `ECARTYPE.PEARSON`.