

Statistiques descriptives

2C / 3C

1 Notions de base

1.1 Contexte

Les statistiques descriptives ont pour but de décrire de manière synthétiques les caractéristiques d'un ensemble d'individus. Les individus peuvent être des personnes, mais aussi des objets ou des faits.

1.2 Vocabulaire

Population : ensemble de tous les individus que l'on souhaite étudier.

Chaque élément d'une population s'appelle une **unité statistique**.

Echantillon : ensemble des individus que l'on interroge, ou dont on connaît les informations.

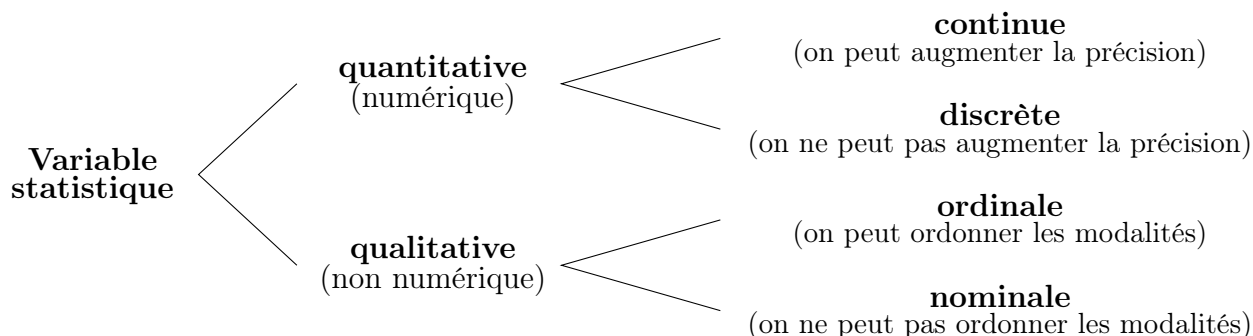
Si l'échantillon est égal à la population entière, on parle d'un **recensement**. Si l'échantillon n'est formé que d'une partie de la population, on parle d'un **sondage**.

Variable statistique : caractéristique que l'on souhaite étudier.

Une variable statistique est dite **quantitative** si les valeurs qu'elle peut prendre sont des nombres. Dans le cas contraire, elle est **qualitative**, et ses valeurs s'appellent des **modalités** ou des **catégories**.

Si les valeurs que peut prendre une variable statistique quantitative peuvent être listées (par exemple 0, 1, 2, 3, ...) et leur précision ne peut pas être augmentée, on dit que la variable est **discrète**. Si la précision des valeurs peut être modifiée (par exemple 2.3 ou 2.31 ou 2.314), on dit qu'elle est **continue**.

Si les valeurs que peut prendre une variable statistique qualitative peuvent être ordonnées (mise dans un certain ordre, par exemple *jamais, parfois, souvent, toujours*), la variable est dite **ordinaire**. Dans le cas contraire, elle est **nominale**.



Exemples

- a) On demande à 200 passagers pris au hasard dans un aéroport de donner leur nationalité.

Population : *passagers de l'aéroport*
Echantillon : *les 200 passagers*
(Unité statistique : *1 passager*)
S'agit-il d'un recensement ou d'un sondage?
Variable statistique : *nationalité*
Type de variable : *qualitatif nominale*
Modalités ou valeurs : *suisse, français, ...*

- b) Durant tout le mois d'avril, on mesure la température maximale de la journée à la Tour-de-Peilz.

Population :
Echantillon :
Unité statistique :
S'agit-il d'un recensement ou d'un sondage?
Variable statistique :
Type de variable :
Modalités ou valeurs :

- c) On demande à 5'000 familles résidant en Suisse le nombre d'enfants dans leur foyer.

Population :
Echantillon :
Unité statistique :
S'agit-il d'un recensement ou d'un sondage?
Variable statistique :
Type de variable :
Modalités ou valeurs :

- d) A la fin d'une période de révision, un enseignant demande à ses élèves de répondre à la question suivante :

Cette période de révision a-t-elle été utile ? ☐ *pas du tout* ☐ *un peu* ☐ *beaucoup*

Population : *les élèves de la classe*
Echantillon : *"*
(Unité statistique : *1 élève*)
S'agit-il d'un recensement ou d'un sondage?
Variable statistique : *l'utilité des révisions*
Type de variable : *qualitatif ordinaire*
Modalités ou valeurs : *"pas du tout", "un peu", "beaucoup".*

1.3 Echelles de mesure

On distingue quatre types d'échelles de mesure.

Echelle nominale

Exemple : 0 (homme) 1 (femme)

- s'utilise pour une variable qualitative nominale
- sert uniquement à différencier les catégories d'une variable ($=$, \neq)
- ne permet pas d'établir une relation d'ordre ($<$, $>$)
- ne permet pas d'effectuer d'opération arithmétique ($+$, $-$, \cdot , $:$)

Echelle ordinale

Exemple : 0 (pas du tout) 1 (un peu) 2 (beaucoup)

- s'utilise pour une variable qualitative ordinale ou quantitative
- sert à différencier les catégories d'une variable ($=$, \neq)
- permet d'ordonner les catégories ($<$, $>$)
- ne permet pas d'effectuer d'opération arithmétique ($+$, $-$, \cdot , $:$)

Echelle d'intervalle

Exemple : les degrés Celsius

- s'utilise pour une variable quantitative
- sert à différencier les valeurs d'une variable ($=$, \neq)
- permet d'ordonner les valeurs ($<$, $>$)
- permet d'effectuer des additions / soustractions, mais pas des multiplications / divisions
- la valeur 0 ne signifie pas l'absence de la caractéristique

Echelle de rapport

Exemple : nombre d'enfants dans une famille : 0, 1, 2, ...

- s'utilise pour une variable quantitative
- sert à différencier les valeurs d'une variable ($=$, \neq)
- permet d'ordonner les valeurs ($<$, $>$)
- permet d'effectuer les quatre opérations arithmétiques ($+$, $-$, \cdot , $:$)
- la valeur 0 signifie une absence de la caractéristique mesurée

Exemple

Voici quatre manières différentes de mesurer la consommation de cigarettes dans un questionnaire :

1. Fumez-vous des cigarettes ?
1. Oui 2. Non

Echelle : *nominale*

2. A quelle fréquence fumez-vous des cigarettes ?
0. Jamais 1. Rarement 2. Occasionnellement 3. Régulièrement

Echelle : *ordinales*

3. En moyenne, combien de cigarettes fumez-vous par jour ?
1. 0 2. De 1 à 5 3. de 6 à 10 4. de 11 à 20 5. Plus de 20

Echelle : *ordinales*

v.s. : nombre de cigarettes quantités discrètes.

$$\begin{array}{ccc} [1; 5] & + & [1; 5] \\ \updownarrow & & \updownarrow \\ 2 & & 2 \end{array} \quad \not\approx \quad \begin{array}{c} [11; 20] \\ \updownarrow \\ 4 \end{array}$$

4. Combien de cigarettes fumez-vous par jour ?
.....

Echelle : *de rapport*

1.4 Tableau de distribution

Une fois les données récoltées, on les regroupe par modalité dans un tableau de distribution.

Exemples

- a) On a demandé à tous les élèves d'une classe quelle était leur matière préférée parmi les matières suivantes : français, anglais, maths et musique.

musique français français anglais français
anglais musique maths musique musique
musique musique français musique
français anglais musique français maths

Tableau de distribution :

Répartition des élèves selon leur matière préférée

	matière	nombre d'élèves	fréquence
	français	6	$6/20 = 0,3 = 30\%$
	musique	8	40%
	anglais	4	20%
	math	2	10%
	Total	20	100%

- b) On a demandé à ces mêmes élèves leur dernière note d'anglais.

5 4.5 3.5 5 6 3.5 4 2.5 4 4.5
4 4.5 4.5 4.5 3 4 4.5 5 3.5 4

Tableau de distribution :

Répartition des élèves selon leur note d'anglais (à leur film)

	Notes	Effectif	fréquence
	1	0	0%
	1.5	0	0%
	2	0	0%
	2.5	1	5%
	3	1	5%
	3.5	3	15%
	4	5	25%
	4.5	6	30%
	5	3	15%
	5.5	0	0%
	6	1	5%
	Total	20	100%

c) Enfin, on a demandé à ces élèves leur taille en centimètre.

172 157 162 156 167 179 173 173 178 160
168 171 165 166 184 170 165 164 160 175

C'est une variable statistique quantitative continue.
 ⇒ C'est très rare d'avoir plusieurs fois la même valeur.
 ⇒ On regroupe les données par classes.

Tableau de distribution :

Répartition des élèves selon leur taille.

pas nécessaire (→)

155 est dans la classe 160 ne l'est pas

	Taille	Effectif	Fréquence
	[150 ; 155 [0	0
	[155 ; 160 [2	$\frac{2}{20} = 10\%$
	[160 ; 165 [4	20%
	[165 ; 170 [5	25%
	[170 ; 175 [5	25%
	[175 ; 180 [3	15%
	[180 ; 185 [1	5%
	Total	N = 20	100%

Aide pour le choix du nombre de classes : Table de Sturges

Nombre de données	Nombre approximatif de classes
Entre 10 et 22	5
Entre 23 et 44	6
Entre 45 et 90	7
Entre 91 et 180	8
Entre 181 et 360	9
Entre 361 et 720	10

on a pris 6 classes pour avoir des classes "pratiques"

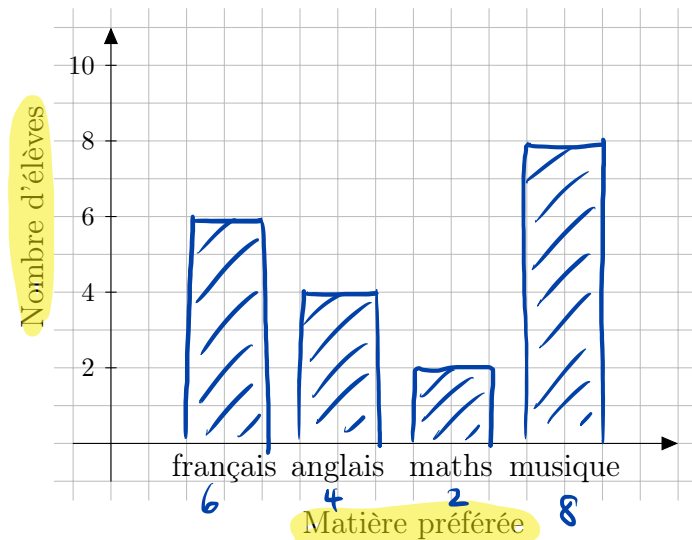
(Formule : $1 + \log_2(n)$)

2 Représentations graphiques

2.1 Représentation graphique d'une variable qualitative

Répartition des 20 élèves d'une classe selon leur matières préférées (exemple p.5).

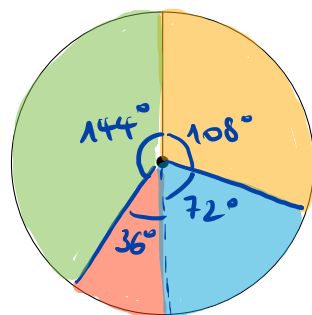
Diagramme à rectangles = histogramme



On laisse un espace entre les rectangles.

$$\begin{array}{lcl} 30 & \leftrightarrow & 100 \\ ? & \leftrightarrow & 360^\circ \end{array}$$

Diagramme circulaire



français

anglais

maths

musique

$$\frac{6}{20} = 30\%$$

$$\frac{4}{20} = 20\%$$

$$\frac{2}{20} = 10\%$$

$$\frac{8}{20} = 40\%$$

angles

$$108^\circ$$

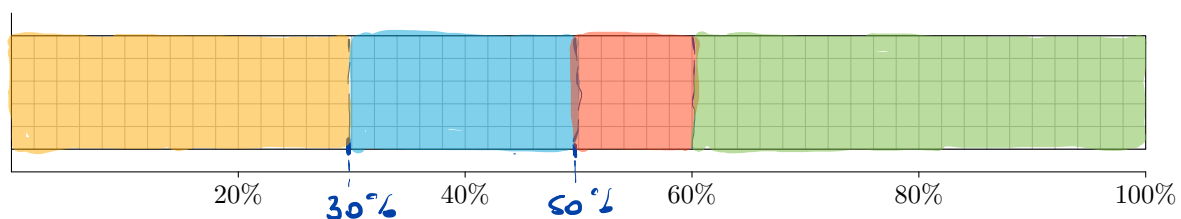
$$72^\circ$$

$$36^\circ$$

$$144^\circ$$

Diagramme linéaire

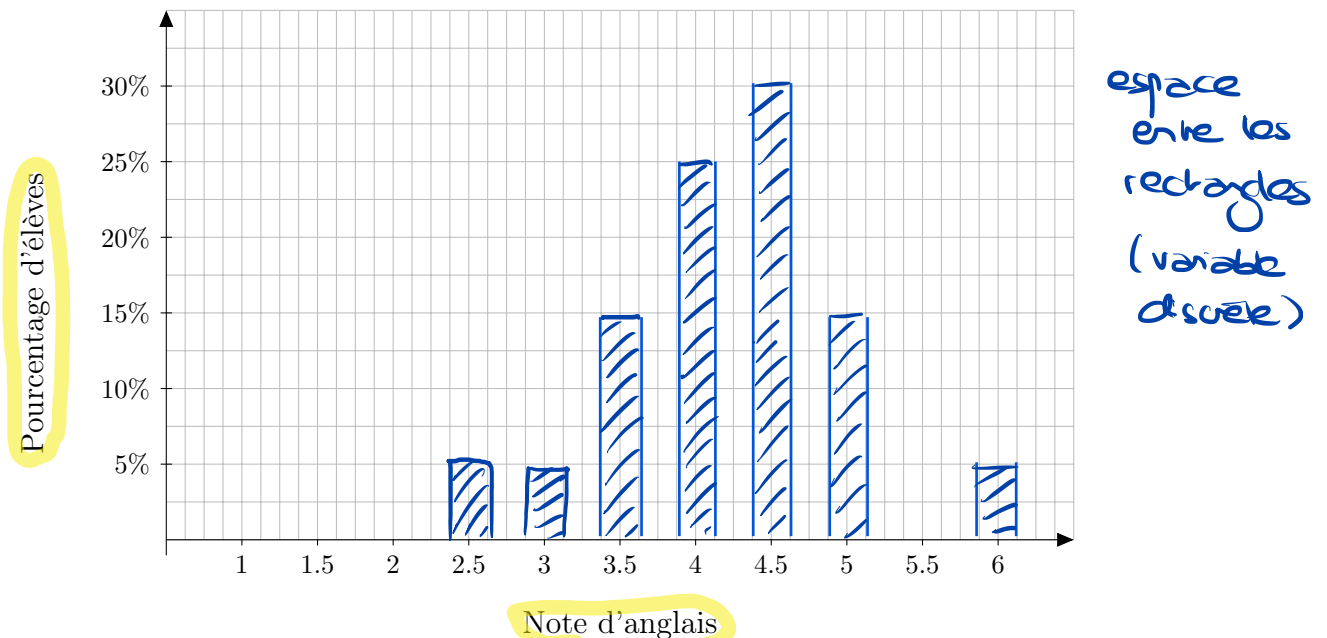
français anglais maths musique



2.2 Représentation graphique d'une variable quantitative discrète

Répartition des 20 élèves d'une classe selon leur note d'anglais (exemple p.5).

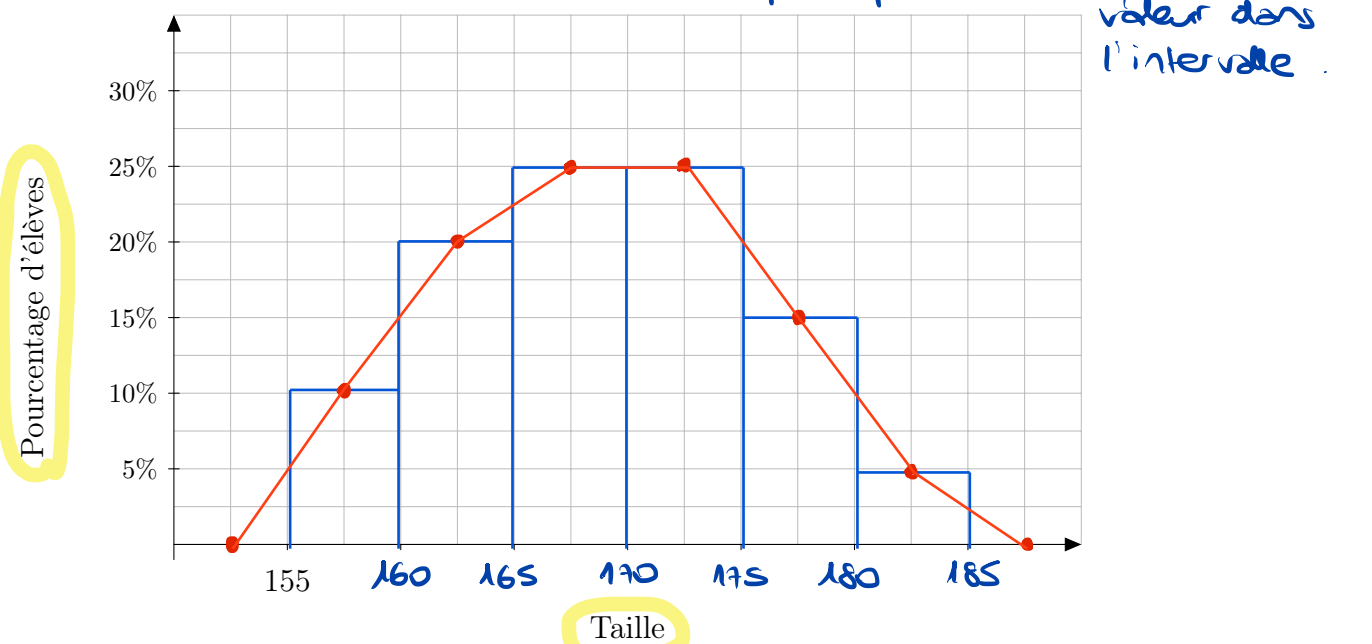
Diagramme en bâtons (ou rectangles)



2.3 Représentation graphique d'une variable quantitative continue

Répartition des 20 élèves d'une classe selon leur taille (exemple p.6).

Histogramme et polygone de fréquences



Polygone de fréquences cumulées (ou courbe de fréquences cumulées)

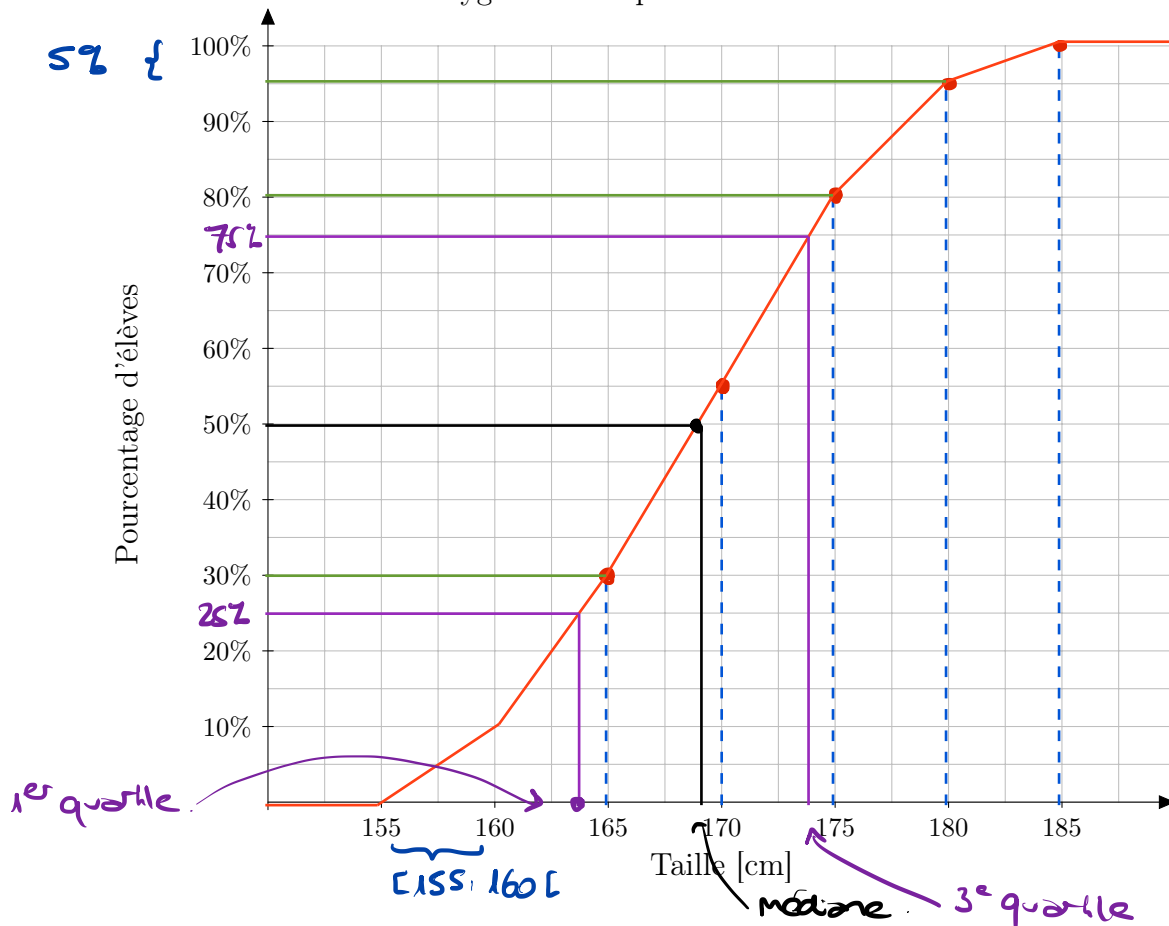
cette classe couvre de 0 à 10%

Répartition des élèves selon leur taille

taille	effectif	fréquence f_i	fréquence cumulée F_i
[155 ; 160[2	10%	10%
[160 ; 165[4	20%	10% + 20% = 30%
[165 ; 170[5	25%	30% + 25% = 55%
[170 ; 175[5	25%	80%
[175 ; 180[3	15%	95%
[180 ; 185[1	5%	100%
Total	20	100%	

pourcentage d'élèves dont la taille est < 170.

Polygone de fréquences cumulées



30% des élèves mesurent *moins de 1.65m = 165 cm*

... *80* ... % des élèves mesurent moins de 1.75 mètre.

5% des élèves mesurent au moins *180 cm*

Histogramme dans le cas de classes inégales

On veut représenter par un histogramme la répartition des roches dans une rocaille de fleurs selon leur masse. Cette répartition est donnée par le tableau suivant :

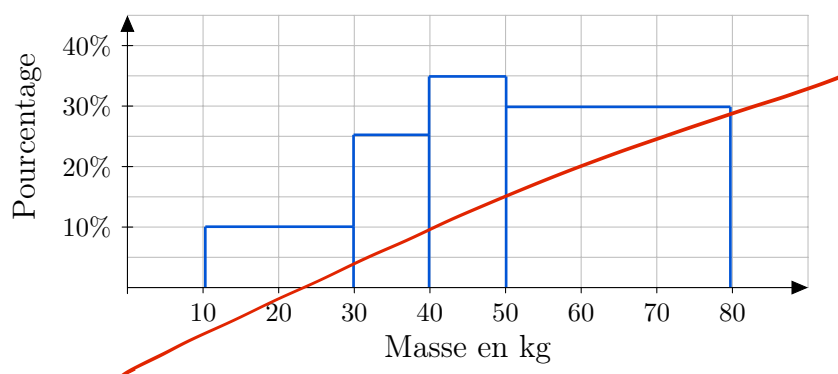
$l = \text{longueur}$

$l=20 \rightarrow$
 $l=10 \rightarrow$
 $l=10 \rightarrow$
 $l=30 \rightarrow$

masse en kg	pourcentage de roches
[10 ; 30[10%
[30 ; 40[25%
[40 ; 50[35%
[50 ; 80[30%
Total	100%

$[50;60[$ 15%
 $[60;70[$ 10%
 $[70;80[$ 5%
 10%
 10%
 10%
 ?
 ?

Histogramme obtenu en utilisant directement les fréquences données dans le tableau :

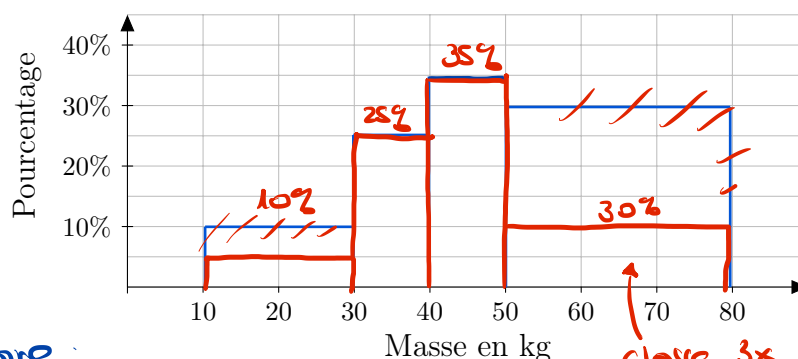


Cet histogramme donne-t-il une information correcte ou faussée ? faussée

Principe de proportionnalité des aires :

L'aire du rectangle doit être proportionnelle
 au pourcentage (et non sa hauteur).

Histogramme correct (respectant le principe de proportionnalité des aires) :



A choix :

→ Revar Combinatoire :
ex 13.1 à 13.43

→ Avancer dans le polycopié de stat 10 pp. 11 à 15, ex 7.15 à 7.17

→ Faire des ex. suppl. de stat. 7.1 à 7.14

classe 3x plus longue
 ⇒ hauteur divisée
 par 3

3 Mesures de tendance centrale

"où est le milieu ?"
 ← moyenne
 ← médiane
 ← mode

But : résumer une série statistique par une seule valeur.

3.1 Moyenne

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n}$$

Exemple 1

Σ = sigma

On reprend les dernières notes d'anglais des 20 élèves d'une classe (exemple p.5).

5 4.5 3.5 5 6 3.5 4 2.5 4 4.5
 4 4.5 4.5 4.5 3 4 4.5 5 3.5 4

Calcul de la moyenne :

$$\bar{x} = \frac{5 + 4.5 + 3.5 + \dots + 3.5 + 4}{20} = 4.2$$

Calculatrice :
 5 $\boxed{\Sigma+}$ 4.5 $\boxed{\Sigma+}$ 3.5 $\boxed{\Sigma+}$
 ... $\boxed{\Sigma+}$ 4 (affiche $n=20$)
 $\boxed{2nd} \boxed{\times 2} \Rightarrow 4.2$

Calcul de la moyenne à partir du tableau de distribution (formule plus rapide) : $\boxed{\bar{x}}$

Répartition des 20 élèves d'une classe selon leur note d'anglais

note	effectif	fréquence
1 / 1.5 / 2	0	0%
2.5	1	5%
3	1	5%
3.5	3	15%
4	5	25%
4.5	6	30%
5	3	15%
5.5	0	0%
6	1	5%
Total	20	100%

$$\bar{x} = \frac{n_1 \cdot c_1 + n_2 \cdot c_2 + \dots + n_k \cdot c_k}{n} = f_1 \cdot c_1 + f_2 \cdot c_2 + \dots + f_k \cdot c_k$$

$$x = \frac{1 \cdot 2.5 + 1 \cdot 3 + 3 \cdot 3.5 + \dots + 1 \cdot 6}{20} = \underbrace{\frac{1}{20}}_{f_1} \cdot 2.5 + \underbrace{\frac{1}{20}}_{f_2} \cdot 3 + \underbrace{\frac{3}{20}}_{f_3} \cdot 3.5 + \dots$$

Calculatrice : 2.5 $\boxed{2nd} \boxed{1/x}$ 1 $\boxed{\Sigma \times}$...
 \boxed{FRQ}

156 ; 157 ; 160 ; 160 ; ...

La valeur 10 et la valeur 11
 Médiane ← 167,5 cm

Exemple 2

On reprend les tailles des 20 élèves d'une classe (exemple p.6).

172 157 162 156 167^{10^e} 179 173 173 178 160
 168^{11^e} 171 165^{7^e} 166^{9^e} 184 170 165^{8^e} 164 160 175

Calcul direct de la moyenne (à partir des données brutes) :

$$\bar{x} = \frac{172 + 157 + 162 + \dots + 175}{20} = 168,25$$

Calcul de la moyenne à partir du tableau de distribution :

Répartition des 20 élèves d'une classe selon leur taille
 Cette classe contient la médiane

taille	valeur centrale	effectif	fréquence	fréq. cum.
[155 ; 160[157,5	2	10%	10%
[160 ; 165[162,5	4	20%	30%
[165 ; 170[167,5	5	25%	55%
[170 ; 175[172,5	5	25%	80%
[175 ; 180[177,5	3	15%	95%
[180 ; 185[182,5	1	5%	100%
Total		20	100%	

$$\bar{x} = \frac{n_1 \cdot c_1 + n_2 \cdot c_2 \dots + n_k \cdot c_k}{n} = f_1 \cdot c_1 + f_2 \cdot c_2 \dots + f_k \cdot c_k$$

$$\bar{x} = \frac{2 \cdot 157,5 + 4 \cdot 162,5 + \dots + 1 \cdot 182,5}{20} = 169 \text{ cm} \quad \text{ou}$$

$$\bar{x} = 0,1 \cdot 157,5 + 0,2 \cdot 162,5 + 0,25 \cdot 167,5 + \dots + 0,05 \cdot 182,5 = 169 \text{ cm}$$

Remarque

La valeur de la moyenne n'est pas la même selon la méthode utilisée. Dans le deuxième cas, on ne dispose plus des données brutes, et on doit donc estimer la valeur de \bar{x} avec l'information disponible.

3.2 Médiane

La médiane partage une série de données **triées** en deux parties égales.

Si \tilde{x} est la médiane d'une série statistique, il y a donc 50% des données qui sont plus petites ou égales à \tilde{x} , et 50% qui sont plus grandes ou égales à \tilde{x} .

$$\tilde{x} = \begin{cases} x_{\frac{n+1}{2}} & \text{si } n \text{ est impair} \\ \frac{1}{2} \left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1} \right) & \text{si } n \text{ est pair} \end{cases}$$

Remarque

Il y a deux formules différentes, car si n est impair et les données sont triées, il y a **une** donnée au milieu de la série. Si n est pair, par contre, il y a **deux** données qui sont au milieu, et on utilise donc la moyenne de ces deux valeurs.

Exemple

Calcul de la médiane dans l'exemple des notes d'anglais :

On trie les notes :

1	2	3	4	5	6	7	8	9	10	11
2,5	3	3,5	3,5	3,5	4	4	4	4	4	4,5
4,5	4,5	4,5	4,5	4,5	5	5	5	5	6	
12	13	14	15	16	17	18	19	20		

milieu

tilde $\rightarrow \tilde{x} = \frac{4 + 4,5}{2} = 4,25$ La moitié des élèves ont fait 4 ou moins l'autre moitié ont fait 4,5 ou plus.

Remarque importante

Cette mesure de tendance centrale est plus **robuste** que la moyenne, elle est moins affectée par les valeurs extrêmes.

L'exemple suivant illustre cette propriété :

Dans une entreprise de 35 employés, supposons que le patron gagne 40'000 francs par mois, alors que les 34 employés gagnent 3'000 francs par mois.

Calcul du revenu mensuel moyen : $\bar{x} = \frac{1 \cdot 40'000 + 34 \cdot 3'000}{35} = 4057 \text{ fr/mois}$

Cette moyenne ne reflète en rien la réalité des travailleurs de cette entreprise. La valeur extrême du salaire du patron a un impact trop grand sur la moyenne.

Calcul de la médiane : $\tilde{x} = 3000 \text{ fr/mois}$

3000 3000 3000 ... | ... 3000 40000

Il est correct de dire que le salaire moyen dans cette entreprise est de **4057** par mois, mais il vaudrait mieux dire que le salaire médian est de **3000** par mois, donc qu'au moins la moitié des employés gagnent **3000 fr ou moins**.

Calcul de la médiane dans le cas continu

Dans l'exemple des tailles des 20 élèves d'une classe, la médiane peut se calculer à partir des données brutes. On obtient alors

10^e valeur : 167, 11^e : 168
des tailles

$$\tilde{x} = 167,5 \text{ cm}$$

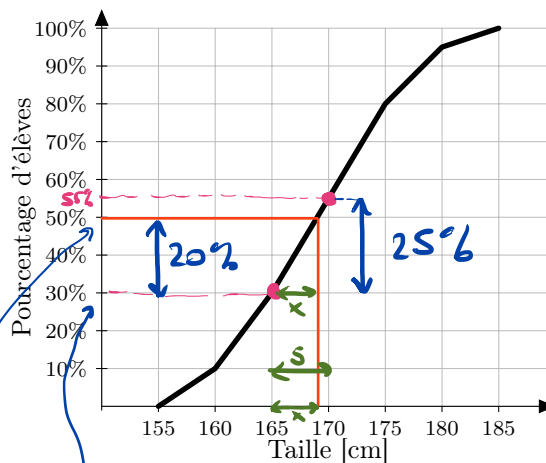
Si on ne dispose que du tableau de distribution, on doit estimer la médiane autrement.

Répartition des 20 élèves d'une classe selon leur taille

taille	val. centr.	effectif	fréqu.	fréqu. cum.
[155 ; 160[1.575	2	10%	10%
[160 ; 165[1.625	4	20%	30%
[165 ; 170[1.675	5	25%	55%
[170 ; 175[1.725	5	25%	80%
[175 ; 180[1.775	3	15%	95%
[180 ; 185[1.825	1	5%	100%
Total		20	100%	

\tilde{x} est dans cet intervalle

Polygone de fréquences cumulées



Classe médiane : la première classe dont la fréquence cumulée dépasse 50% .
 \rightarrow [165 ; 170[

Calcul de la médiane par proportionnalité entre les fréquences et les valeurs :

Fréquence cumulée pour 165 cm : 30%
 " " " 170 cm : 55% } 25% pour [165 ; 170[

25% | 5 cm
 20% | x cm

$$\Rightarrow x = 4 \text{ cm}$$

$$\Rightarrow \tilde{x} = 165 + 4 = 169 \text{ cm}$$

Fréquence cumulée pour \tilde{x} : 50%

\Rightarrow 20% pour [165 ; \tilde{x}]

Pour mercredi :
 Ex. 7.15

Remarque

La valeur de la médiane n'est pas la même selon la méthode utilisée. Dans le deuxième cas, on ne dispose plus des données brutes, et on doit donc estimer la valeur de \tilde{x} avec l'information disponible.

3.3 Mode et classe modale

Le **mode** est la **valeur** (ou la catégorie dans le cas qualitatif) **qui revient le plus souvent** dans une série statistique.

La **classe modale** est la **classe qui regroupe le plus de données** dans le cas d'une variable **continue**.

Remarques

1. Le mode ou la classe modale ne sont significatifs que si leur effectif est largement plus grand que celui des autres modalités ou des autres classes.
2. Le mode est la seule mesure de tendance centrale qui peut être utilisée pour une variable qualitative.

Exemples

a) Dans l'exemple des matières préférées, le mode est ... **la musique**

Interprétation : ... **une majorité d'élèves préfère la musique**

b) Dans l'exemple des notes d'anglais, le mode est ... **4.5**

Interprétation : ... **une majorité d'élèves a eu 4.5**

c) Dans l'exemple des tailles, la classe modale est ... **[165 ; 170[** ou **[170 ; 175[**

Interprétation : ... **non pertinent : 2 classes égales et pas**

... **beaucoup plus hautes que les autres**

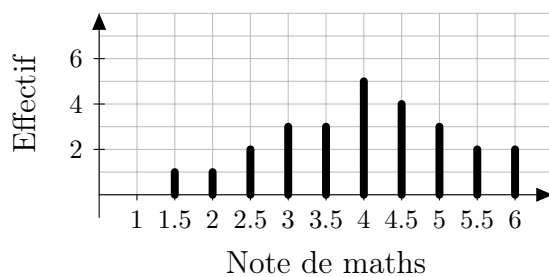
ex. 7.15 - 7.16 - 7.17

4 Mesures de dispersion

Lorsqu'on résume une série statistique par une mesure de tendance centrale (souvent la moyenne), on ne donne aucune information sur la manière dont les données se répartissent autour de cette valeur : sont-elles toutes assez proches de la moyenne, ou trouve-t-on des valeurs très dispersées autour de celle-ci ? Cette question nécessite de donner une valeur supplémentaire, appelée mesure de dispersion.

Pour illustrer ces mesures de dispersion, nous allons nous baser sur les notes de maths de trois classes parallèles, données par des diagrammes en bâtons.

Classe A

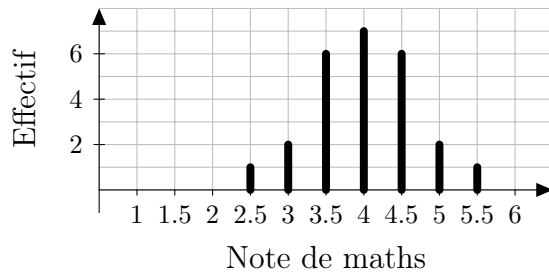


moyenne : $\bar{x} = 4$

médiane : $\tilde{x} = 4$

mode : 4

Classe B

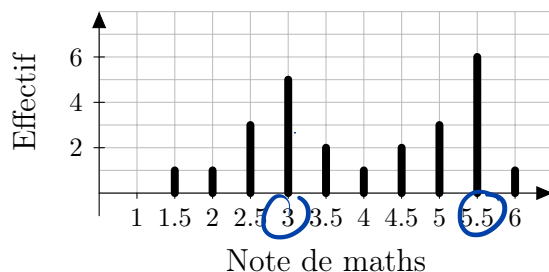


moyenne : $\bar{x} = 4$

médiane : $\tilde{x} = 4$

mode : 4

Classe C



moyenne : $\bar{x} = 4$

médiane : $\tilde{x} = 4.00$

mode : 5.5 et 3
peu significatif

13^e

Remarque

Ces mesures de tendance centrale ne suffisent pas à décrire les différences entre ces trois classes.

4.1 Etendue

L'étendue est la "distance" entre la plus petite et la plus grande valeur.

Classe	A	B	C
Etendue	4.5	3	4.5

Cette mesure permet de différencier les situations des classes ...A...C... et ...B..., mais pas des classes ...A... et ...C...

4.2 Variance et écart-type

La variance est l'écart quadratique moyen à la moyenne. Elle se calcule par la formule suivante :

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n}$$

Dans le cas de données regroupées par modalités ou par classes, la formule devient

valeur du milieu ↓ *effectif* ←

$$s^2 = \frac{n_1 \cdot (c_1 - \bar{x})^2 + n_2 \cdot (c_2 - \bar{x})^2 + \dots + n_k \cdot (c_k - \bar{x})^2}{n}$$

$$= f_1 \cdot (c_1 - \bar{x})^2 + f_2 \cdot (c_2 - \bar{x})^2 + \dots + f_k \cdot (c_k - \bar{x})^2$$

[c1, c2, ..., ck] | [c1, c2, ..., ck] | [n1, n2, ..., nk]

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

Comme la variance est calculée à partir de grandeurs au carré, on définit l'écart-type, noté s , comme la racine de la variance. On obtient ainsi une mesure de la dispersion dans la même unité que les mesures initiales.

Calculatrice :

- entrer les données avec $\Sigma+$
- σ s'obtient avec $2nd \div$

on peut utiliser la touche \boxed{FRQ} si on a plusieurs fois la même donnée.

Classe	A	B	C
Variance	$\sim 1,35$	0,46	1,74
Ecart-type	$\sim 1,16$	$\sim 0,68$	$\sim 1,32$

Grâce à ces nouvelles mesures, on peut maintenant affirmer que les notes de la classe ...B... sont les moins dispersées autour de la moyenne, et que celles de la classe ...C... sont les plus dispersées.

ex. 7.22, 7.23, 7.24, 7.25, 7.26
→ mardi 8.6

5 Mesures de position

5.1 Quantiles

La médiane est une valeur qui partage les données de l'échantillon en deux groupes de taille égale : 50% des données sont inférieures ou égales à la médiane, et 50% des données lui sont supérieures ou égales.

Cette idée se généralise pour n'importe quel pourcentage. Par exemple, quelle est la valeur qui sépare les 25% les plus petits des 75% les plus grands ?

Un **quantile** à $p\%$ est une valeur qui est supérieure ou égale aux $p\%$ des données les plus petites, et inférieure ou égale au reste des données. On le note $q_p\%$.

Cas particuliers

- Les quartiles (Q_1, Q_2, Q_3) sont les quantiles à 25%, 50% et 75%. Ils partagent les données en quatre parties égales. Le deuxième quartile (Q_2) est égal à la médiane.
- Les quintiles (V_1, V_2, V_3, V_4) sont les quantiles à 20%, 40%, 60% et 80%. Ils partagent les données en cinq parties égales.
- les déciles (D_1, D_2, \dots, D_9) sont les quantiles à 10%, 20%, ..., 90%. Ils partagent les données en 10 parties égales.
- Les centiles (C_1, C_2, \dots, C_{99}) sont les quantiles à 1%, 2%, ..., 99%. Ils partagent les données en cent parties égales.

Les quantiles se déterminent en utilisant le même principe que pour la médiane.

Remarque

Pour que les quantiles aient du sens, il faut que l'échantillon soit suffisamment grand. On ne calculera jamais le premier décile d'une distribution composée d'une dizaine de valeurs !

Exemple

Selon une étude, l'âge des propriétaires de PME peut se résumer par le tableau suivant :

âge	fréquence
[20 ans ; 40 ans[12 %
[40 ans ; 50 ans[27 %
[50 ans ; 60 ans[40 %
[60 ans ; 90 ans[21 %
Total	100%

fréquence cumulée

12%

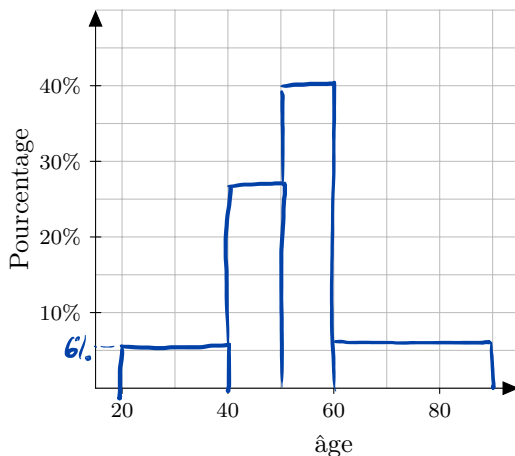
27+12 = 39% ← il manque 11%

79%

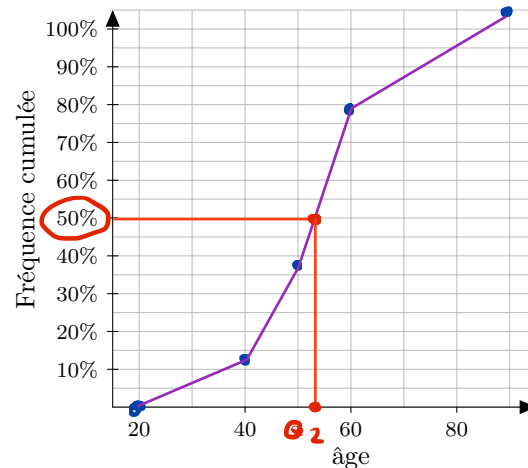
100%

Histogramme

Médiane ?



Courbe de fréquences cumulées



Calcul de la médiane : Elle est dans la classe [50 ; 60[.

Il manque 11% à 39% pour arriver à 50%.

$$l = 10 \quad \left| \quad \begin{array}{l} 40\% \\ 11\% \end{array} \Rightarrow x = \frac{10 \cdot 11}{40} = 2,75$$

$$\tilde{x} = Q_2 = 52,75 \text{ ans.}$$

50% des propriétaires de PME ont moins de 53 ans (ou plus de 52 ans)

Calcul du deuxième décile :

on veut 20% : Dans la classe [40 ; 50[

$$l = 10 \quad \left| \quad \begin{array}{l} 27\% \\ 8\% \end{array} \Rightarrow x = \frac{10 \cdot 8}{27} \approx 2,96$$

fréquence cumulée : 12%
il manque 8%

$$\Rightarrow D_2 = 40 + 2,96 \approx 43 \text{ ans.}$$

Calcul du quantile à 79% : 60 ans (pas besoin de calcul).

5.2 Boxplot

Un boxplot (ou boîte à moustache) est une manière de représenter graphiquement la distribution d'une variable statistique en faisant apparaître la médiane, les quartiles et les deux valeurs extrêmes (la plus petite et la plus grande).

Exemple

On suppose que le nombre de périodes d'absences par année des élèves d'un gymnase se répartit de la manière suivante :

Etendue = 360

Périodes d'absence	[0 ; 10[(10 ; 30]	[30 ; 60[[60 ; 90[[90 ; 120[[120 ; 360[Total
Fréquence	14%	60%	15%	7%	2%	2%	100%

Fréq. cumulée

14%

74%

89%

96%

98%

100%

Calcul de la médiane :

[10 ; 30[
↑
freq. cum: 14%
il manque: 36%

$$l = 20 \quad \left| \quad \begin{array}{l} 60\% \\ x \end{array} \right| \quad 36\%$$

$$\Rightarrow x = \frac{20 \cdot 36}{60} = 12$$

$Q_2 = \text{médiane}$
"
 $\Rightarrow \tilde{x} = 10 + 12 = 22$

Calcul du premier quartile :
↳ 25%

freq. cum: 14%
il manque: 11%
↓
 $Q_1 : [10 ; 30[$

$$l = 20 \quad \left| \quad \begin{array}{l} 60\% \\ x \end{array} \right| \quad 11\% \quad \Rightarrow x = \frac{20 \cdot 11}{60} \approx 3,7$$

$$Q_1 = 13,7$$

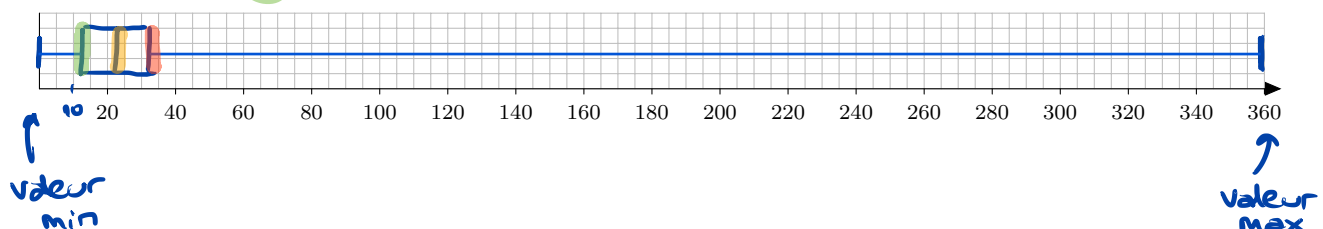
Calcul du troisième quartile :

$$Q_3 \text{ (exercice)} = 32$$

Pour demain: relire tout
cela + calcul de Q_3 .

Boxplot

$$Q_1 = 13,7 \quad Q_2 = 22 \quad Q_3 = 32$$

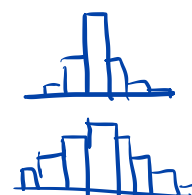


5.3 Cote Z

Mise en situation :

Un gymnase souhaite engager un ancien étudiant pour donner des cours d'appui de mathématiques. Les quatre candidats ont suivi leur troisième année dans quatre gymnases différents, mais on souhaite tout de même déterminer le meilleur étudiant en fonction de ses résultats à l'examen de maturité.

Candidat	Note de l'élève	Note moyenne de son gymnase	Ecart-type de son gymnase
Loïc	4.5	3.7	1.1
Muriel	5	4.1	0.6
Antonin	5.5	5.1	0.4
Eloïse	5	4.0	0.9



Si le gymnase ne se fie qu'à la note de l'élève, il devrait engager ... Antonin

S'il tient aussi compte de la moyenne du gymnase, il engagera plutôt ... Eloïse

Enfin, en tenant compte de l'écart-type du gymnase, il choisira alors ... Muriel

Pour décrire la position d'une donnée par rapport à une distribution, on utilise la cote Z.

$$\text{Cote Z de } x_i = \frac{x_i - \bar{x}}{s}$$

La cote Z mesure la distance d'une valeur à la moyenne, mesurée en nombre d'écart-type.

Cote Z de Loïc : $\frac{4.5 - 3.7}{1.1} \approx 0.7$

Cote Z de Muriel : $\frac{5 - 4.1}{0.6} = 1.5$ meilleure note !

Cote Z d'Antonin : $\frac{5.5 - 5.1}{0.4} \approx 1$

Cote Z d'Eloïse : $\frac{5 - 4}{0.9} \approx 1.1$

Interprétation de la cote Z

Une cote Z positive signifie que la valeur est supérieure à la moyenne, alors qu'une cote Z négative indique qu'elle est en dessous de la moyenne.

Une cote Z de 3 ou plus, ou de -3 ou moins indique une valeur très rare. La cote Z permet donc d'identifier des situations exceptionnelles ou peu plausibles.

Exemple

Un cinéma accueille en moyenne 120 spectateurs les soirs de semaine, avec un écart-type de 14 spectateurs. Il décide de proposer une offre spéciale le mardi soir, avec des places à tarif réduit. Le mardi suivant, 172 spectateurs assistent à la projection.

Peut-on déduire que l'offre spéciale a eu de l'effet ?

$$\text{Cote } z \text{ de } 172 : \frac{172 - 120}{14} \approx 3,7 > 3$$

↳ exceptionnelle ! L'offre a bien fonctionné.

Un lundi soir, une exposition a lieu tout près du cinéma. Ce même soir, le cinéma vend 104 billets. Le gérant se plaint de l'effet négatif de l'exposition, qui lui aurait "volé" des clients. Est-ce justifié ?

$$\text{Cote } z \text{ de } 104 : \frac{104 - 120}{14} \approx -1,1$$

Ce n'est pas une valeur rare, ce n'est donc pas justifié.

exercices : 7.26, 7.28 à 7.35.