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.

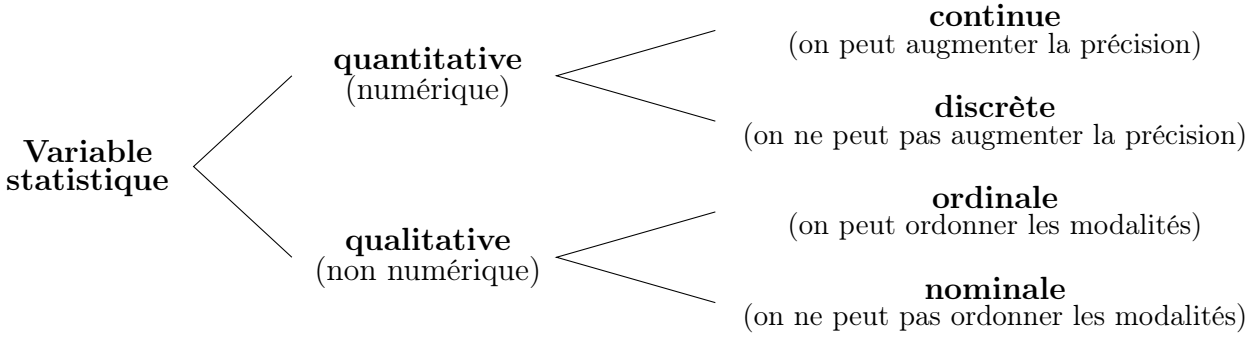
Si l'échantillon est égal à la population entière, on parle d'un **recensement**. Si l'échantillon n'est formé que d'une partie de la population, on parle d'un **sondage**.

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

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

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 :

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

   Echelle :

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 :

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

   Echelle :

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	musique	français	français
français	anglais	musique	anglais	maths

Tableau de distribution :

Répartition des ..... selon .....

matière	nombre d'élèves	fréquence
Total		

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 ..... selon .....

		fréquence
Total		

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

.....
.....
.....
.....

Tableau de distribution :

Répartition des ..... selon .....

Total		

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

(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

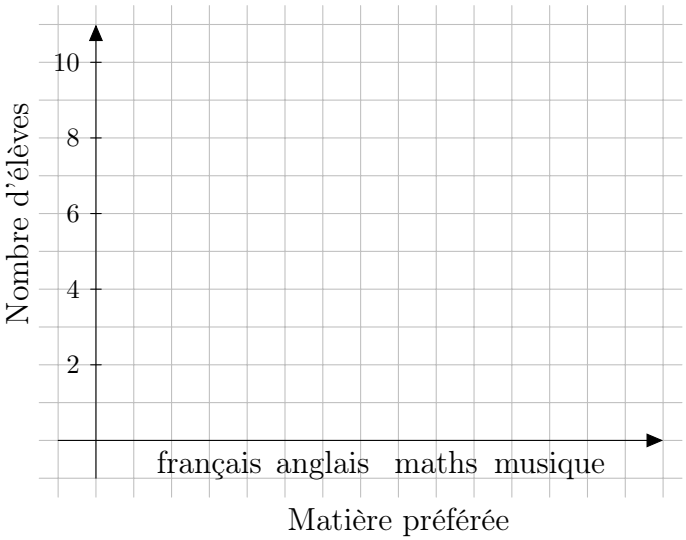
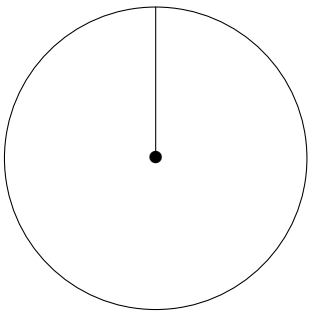


Diagramme circulaire



☐ français

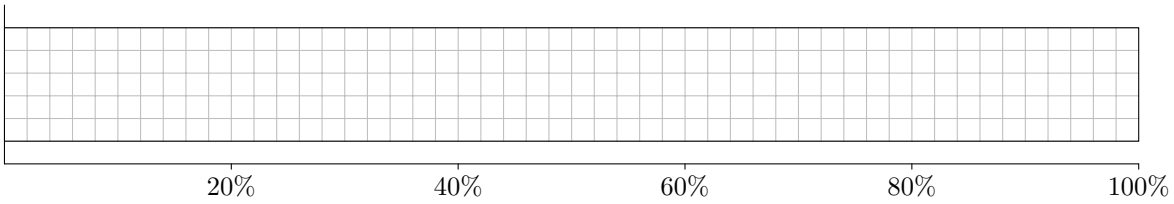
☐ anglais

☐ maths

☐ musique

Diagramme linéaire

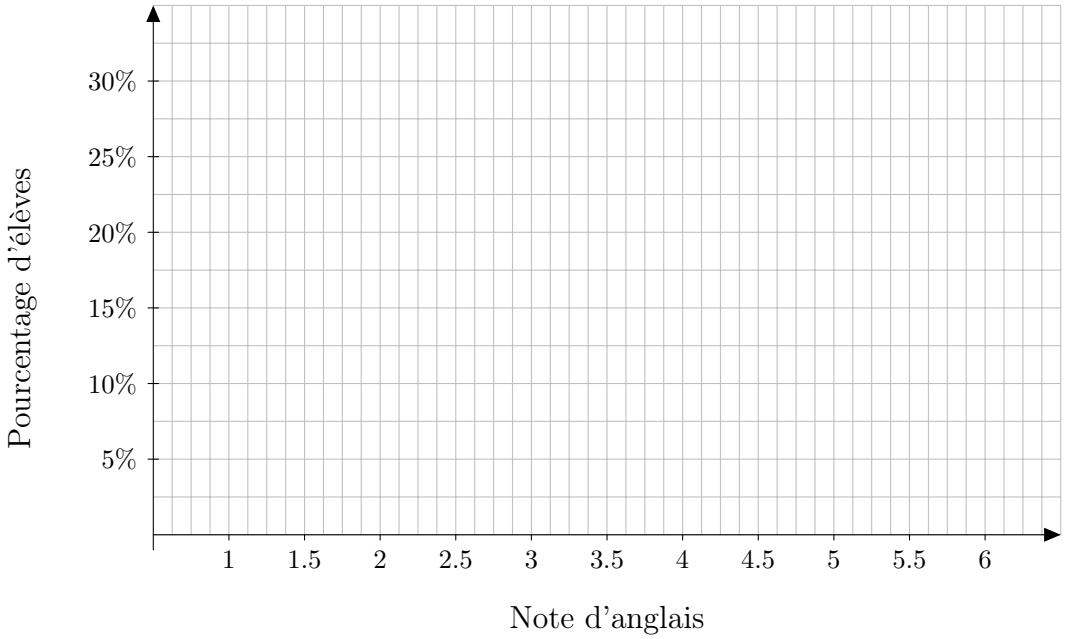
☐ français ☐ anglais ☐ maths ☐ musique



2.2 Représentation graphique d'une variable quantitative discrète

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

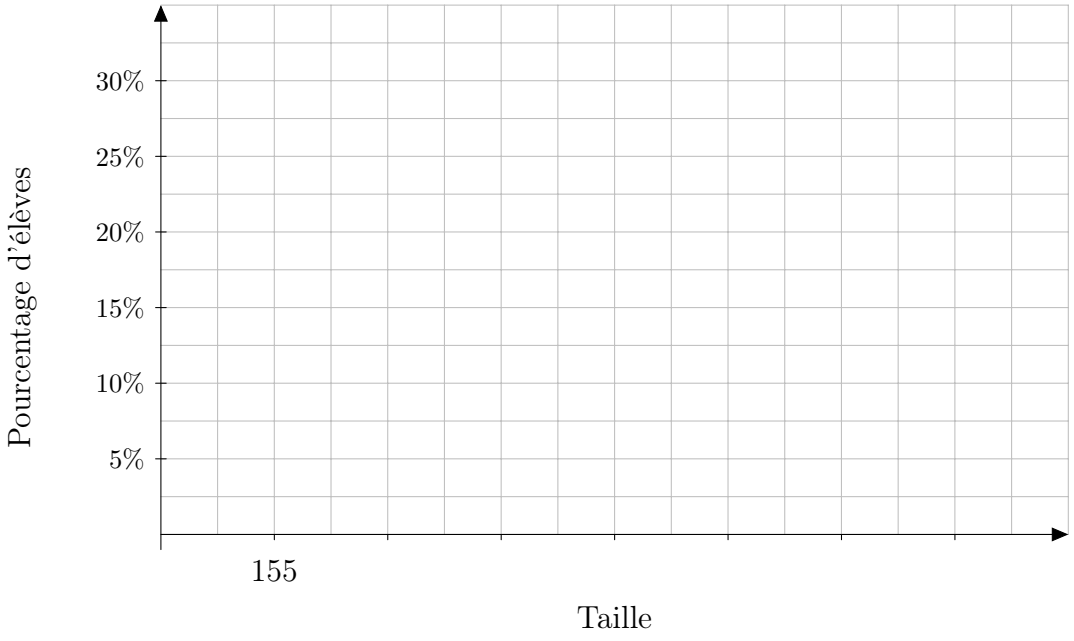
Diagramme en bâtons (ou rectangles)



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

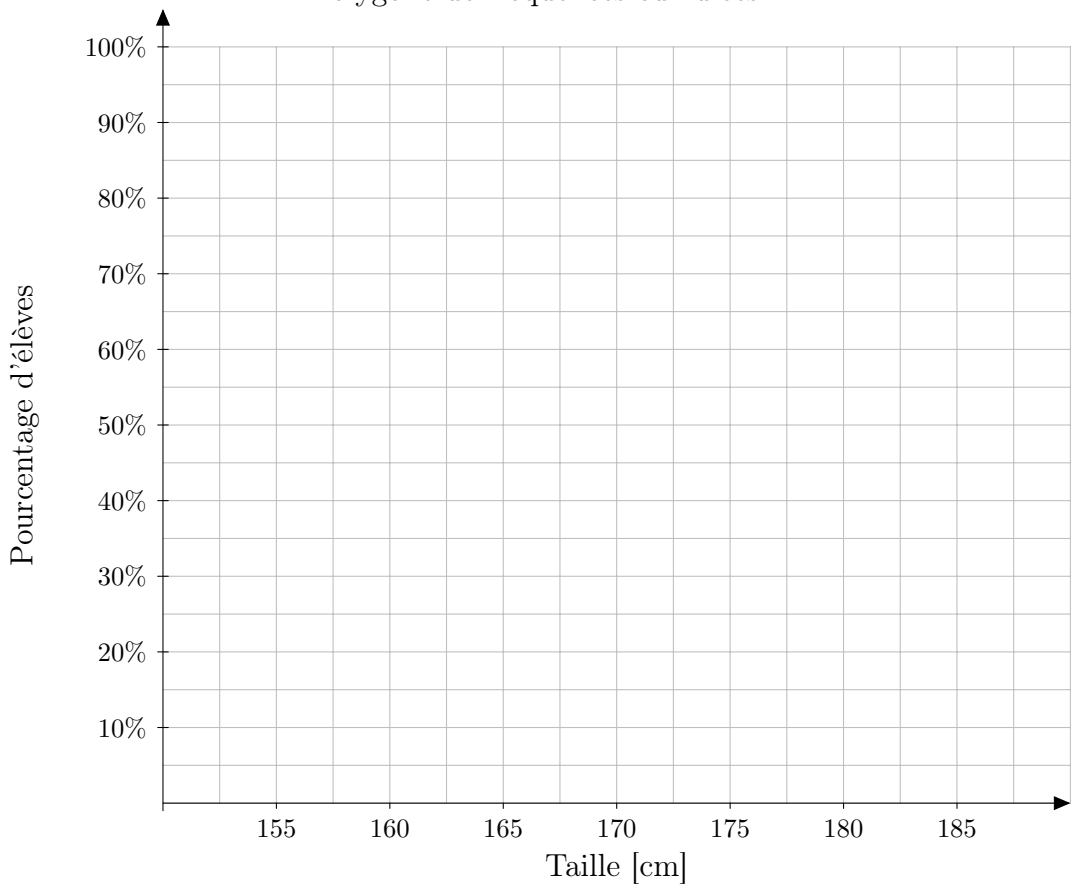


Polygone de fréquences cumulées (ou courbe de fréquences cumulées)

Répartition des élèves selon leur taille

taille	effectif	fréquence	fréquence cumulée
[155 ; 160[	2	10%	
[160 ; 165[	4	20%	
[165 ; 170[	5	25%	
[170 ; 175[	5	25%	
[175 ; 180[	3	15%	
[180 ; 185[	1	5%	
Total	20	100%	

Polygone de fréquences cumulées



30% des élèves mesurent .....

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

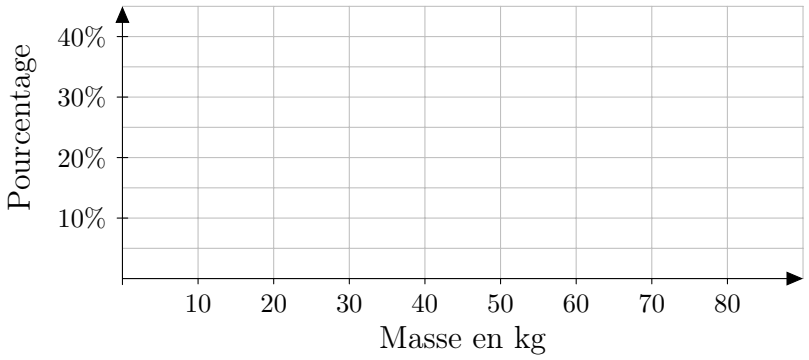
5% des élèves mesurent au moins .....

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 :

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

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



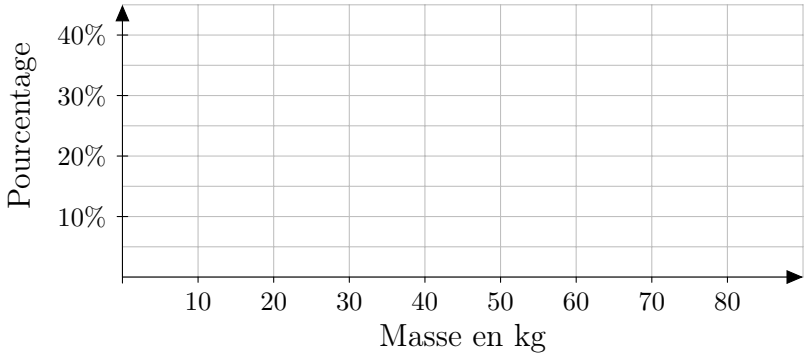
Cet histogramme donne-t-il une information correcte ou faussée ? .....

Principe de proportionnalité des aires :

.....

.....

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



3 Mesures de tendance centrale

But : résumer une série statistique par une seule valeur.

3.1 Moyenne

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

Exemple 1

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 :

Calcul de la moyenne à partir du tableau de distribution (formule plus rapide) :

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%

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

Exemple 2

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

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

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

Calcul de la moyenne à partir du tableau de distribution :

Répartition des 20 élèves d'une classe selon leur taille

taille	valeur centrale	effectif	fréquence	fréqu. cum.
[155 ; 160[		2	10%	10%
[160 ; 165[		4	20%	30%
[165 ; 170[		5	25%	55%
[170 ; 175[		5	25%	80%
[175 ; 180[		3	15%	95%
[180 ; 185[		1	5%	100%
Total		20	100%	

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

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 $\tilde{x}$.

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

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 impair, 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 :

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 : $\bar{x} =$

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 : $\tilde{x} =$

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

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

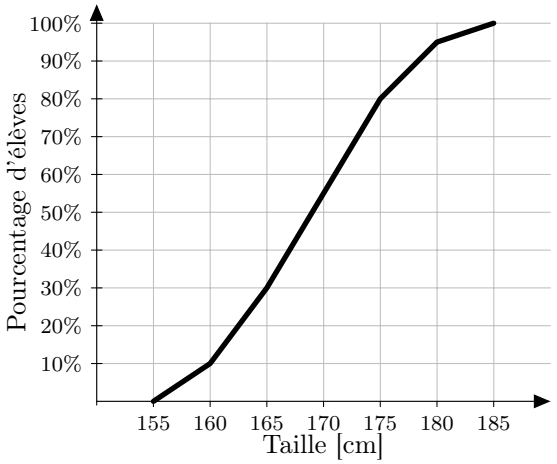
$\tilde{x} =$

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%	

Polygone de fréquences cumulées



Classe médiane :

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

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 .....

Interprétation : .....

.....

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

Interprétation : .....

.....

c) Dans l'exemple des tailles, la classe modale est .....

Interprétation : .....

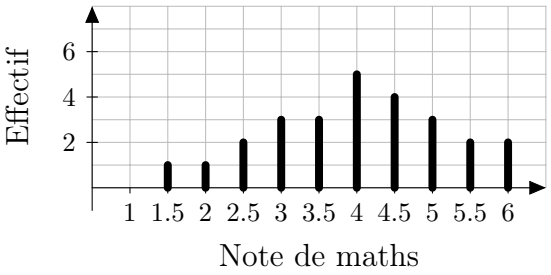
.....

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

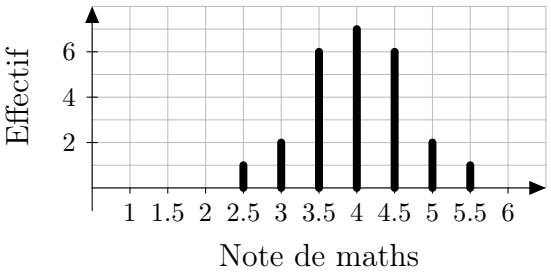


moyenne :

médiane :

mode :

Classe B

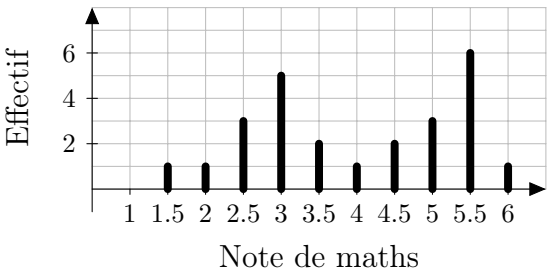


moyenne :

médiane :

mode :

Classe C



moyenne :

médiane :

mode :

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			

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

4.2 Variance et écart-type

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

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

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

$$\begin{aligned} s^2 &= \frac{n_1 \cdot (c_1 - \bar{x})^2 + n_2 \cdot (c_2 - \bar{x})^2 + \cdots + n_k \cdot (c_k - \bar{x})^2}{n} \\ &= f_1 \cdot (c_1 - \bar{x})^2 + f_2 \cdot (c_2 - \bar{x})^2 + \cdots + f_k \cdot (c_k - \bar{x})^2 \end{aligned}$$

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.

Classe	A	B	C
Variance			
Ecart-type			

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

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 ($D_1, D_2, ..., D_9$) sont les quantiles à 10%, 20%, ... , 90%. Ils partagent les données en 10 parties égales.
- Les centiles ($C_1, C_2, ..., C_{99}$) 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

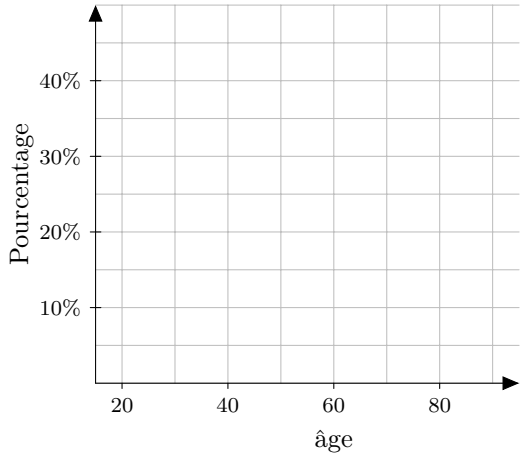
Pour que les quantiles aient du sens, il faut que l'échantillon soit suffisamment grand. On ne calculera jamais le premier décile d'une distribution composée d'une dizaine de valeurs !

Exemple

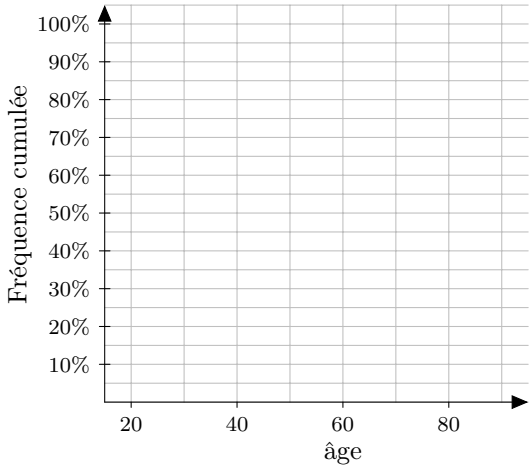
Selon une étude, l'âge des propriétaires de PME peut se résumer par le tableau suivant :

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

Histogramme



Courbe de fréquences cumulées



Calcul de la médiane :

Calcul du deuxième décile :

Calcul du quantile à 79% :

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 :

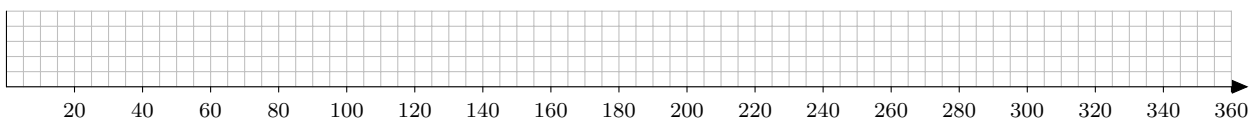
Périodes d'absence	[0 ; 10[	[10 ; 30[	[30 ;60[	[60 ;90[	[90 ; 120[	[120 ; 360[	Total
Fréquence	14%	60 %	15 %	7%	2%	2%	100%

Calcul de la médiane :

Calcul du premier quartile :

Calcul du troisième quartile :

Boxplot



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 .....

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

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

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

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

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

Cote Z de Loïc :

Cote Z de Muriel :

Cote Z d'Antonin :

Cote Z d'Eloïse :

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 ?

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 ?