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 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 :

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 personnes, 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

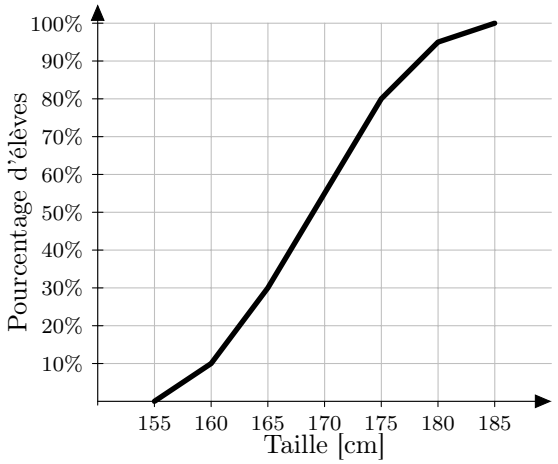
$\tilde{x} =$

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[	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%	

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 : .....

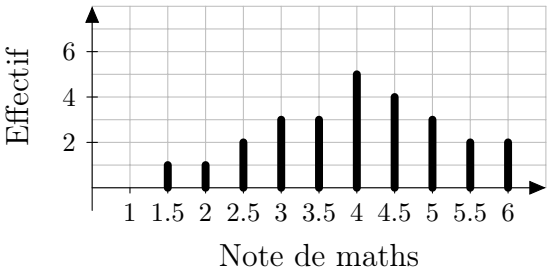
.....

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

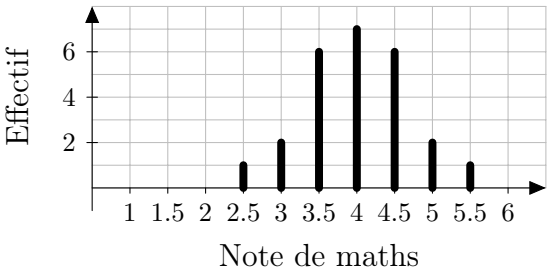


moyenne :

médiane :

mode :

Classe B

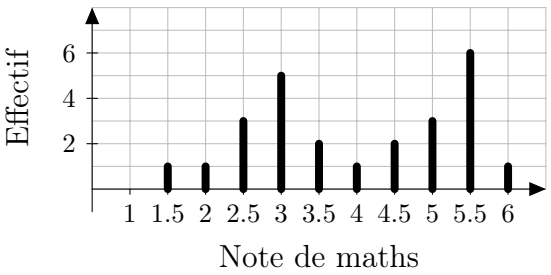


moyenne :

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

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

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

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

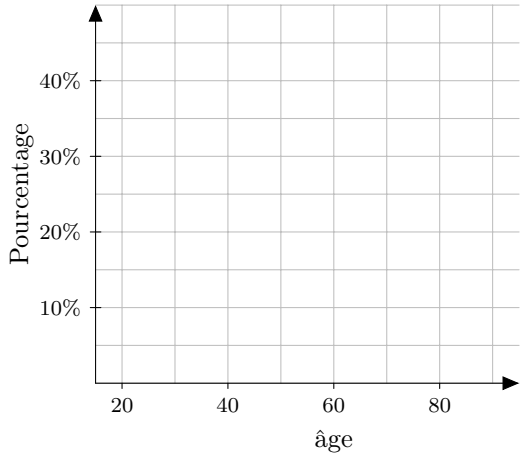
Pour que les quantiles aient du sens, il faut que l'échantillon soit suffisamment grand. On ne calculera jamais le premier décile d'une distribution composée d'une dizaine de valeurs !

Exemple

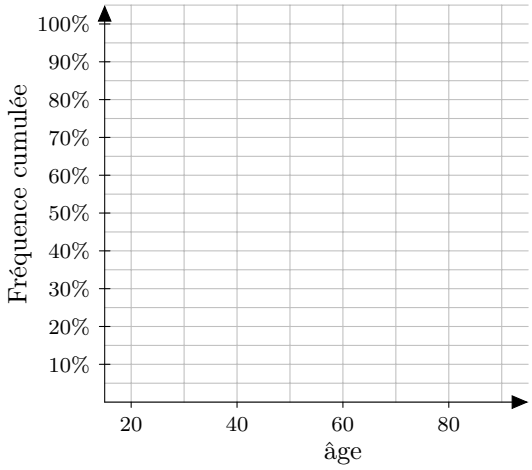
Selon une étude, l'âge des propriétaires de PME peut se résumer par le tableau suivant :

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 :

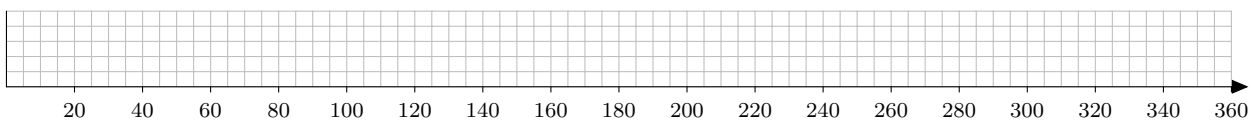
Périodes d'absence	[0 ; 10[	[10 ; 30[	[30 ;60[	[60 ;90[	[90 ; 120[	[120 ; 360[	Total
Fréquence	14%	60 %	15 %	7%	2%	2%	100%

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 ?