STATISTIQUES

I) Médiane et quartiles d'une série statistique quantitative

a) Cas d'une série statistique discrète

Dans ce cas, on dispose d'une famille de réels x_1 ; x_2 ; ... ; x_N que l'on a rangé **dans l'ordre croissant** :

$$x_1 \leqslant x_2 \leqslant ... \leqslant x_N \text{ (Certains de ces réels peuvent être confondus)}$$

Vocabulaire : x_1 s'appelle le terme de rang 1 (ou d'indice 1), x_i le terme de rang (ou d'indice) i ($1 \leqslant i \leqslant N$)

N représente l'effectif total.

On note $(x_i)_{1 \leqslant i \leqslant N}$ cette famille de réels qu'on appelle encore "série statistique".

Exemples :

L'élève *A* a obtenu les 8 notes suivantes :

$$x_1 = 5 \quad x_2 = 5 \quad x_3 = 6 \quad x_4 = 9 \quad x_5 = 10 \quad x_6 = 12 \quad x_7 = 13 \quad x_8 = 13$$

L'élève *B* a obtenu les 9 notes suivantes :

$$x_1 = 2 \quad x_2 = 3 \quad x_3 = 5 \quad x_4 = 6 \quad x_5 = 8 \quad x_6 = 9 \quad x_7 = 9 \quad x_8 = 10 \quad x_9 = 10$$

L'élève *C* a obtenu les 10 notes suivantes :

$$x_1 = 6 \quad x_2 = 6 \quad x_3 = 10 \quad x_4 = 12 \quad x_5 = 12 \quad x_6 = 13 \quad x_7 = 14 \quad x_8 = 15 \quad x_9 = 16 \quad x_{10} = 16$$

L'élève *D* a obtenu les 11 notes suivantes :

$$x_1 = 0 \quad x_2 = 0 \quad x_3 = 1 \quad x_4 = 4 \quad x_5 = 5 \quad x_6 = 8 \quad x_7 = 10 \quad x_8 = 12 \quad x_9 = 13 \quad x_{10} = 16 \quad x_{11} = 17$$

Définition 1 *Médiane*

On appelle médiane tout réel m_e tel que :

au moins 50% des termes de la série ont une valeur inférieure ou égale à m_e

et

au moins 50% des termes de la série ont une valeur supérieure ou égale à m_e

On prouvera, ci-dessous (théorème 1), qu'un tel réel existe toujours !

Remarque : la médiane partage l'ensemble des termes en deux sous ensembles de même effectif. (Enfin presque !)

Exemples :

Pour l'élève *A* ($N = 8$) : $m_e = x_4 = 9$ ($x_5 = 10$ conviendrait également ou, plus généralement, tout réel de [9 ; 10])

Pour l'élève *B* ($N = 9$) : $m_e = x_5 = 8$ (et là, il n'y a pas d'autre choix possible)

Pour l'élève *C* ($N = 10$) : $m_e = 12{,}5$ (ou tout réel de l'intervalle $[x_5 ; x_6] = [12 ; 13]$)

Pour l'élève *D* ($N = 11$) : $m_e = 8$ (et là, il n'y a pas d'autre choix possible)

On constate que la détermination de la médiane est différente suivant que l'effectif total N est **pair** ou **impair** :

- Lorsque l'effectif total N est **impair**, il n'y a pas de difficulté, la médiane m_e est le **terme central**, à savoir le terme de rang $\frac{N+1}{2}$. On a donc : $m_e = x_{\frac{N+1}{2}}$.

- Lorsque l'effectif total N est **pair**, l'usage veut que l'on choisisse pour médiane m_e la **moyenne des deux termes centraux**, à savoir : les termes de rang $\frac{N}{2}$ et $\frac{N}{2}+1$. On a donc : $m_e = \frac{x_{\frac{N}{2}} + x_{\frac{N}{2}+1}}{2}$.
  Mais tout réel de l'intervalle $[\, x_{\frac{N}{2}} ; x_{\frac{N}{2}+1} \,]$ conviendrait également. (En effet, dans certaines situations, la moyenne des deux termes centraux, qui n'est pas une valeur de la série, n'a pas de sens : par exemple, quel est le jour médian du mois de juin ? Le mois de juin comporte 30 jours. Les deux termes centraux sont 15 et 16 ($15^{\text{ème}}$ jour et $16^{\text{ème}}$ jour). Dire que "le jour médian est le $15{,}5^{\text{ème}}$" n'a pas de sens. Mieux vaut dire (dans ce type de situation) : "le jour médian est le $15^{\text{ème}}$ jour" ou "le jour médian est le $16^{\text{ème}}$ jour" (au choix !) ...)

Exemple : si $N = 29$ alors $m_e = x_{15}$; si $N = 42$ alors $m_e = \frac{x_{21} + x_{22}}{2}$.

Exercice : quelle est la médiane de la série suivante : $x_1 = 1$ $x_2 = 1$ $x_3 = 1$ $x_4 = 1$ $x_5 = 1$?

Définition 2 *Quartiles*

On appelle premier quartile tout réel Q_1 tel que :

au moins 25% des termes de la série ont une valeur inférieure ou égale à Q_1

et

au moins 75% des termes de la série ont une valeur supérieure ou égale à Q_1

On appelle troisième quartile tout réel Q_3 tel que :

au moins 75% des termes de la série ont une valeur inférieure ou égale à Q_3

et

au moins 25% des termes de la série ont une valeur supérieure ou égale à Q_3

On prouvera, ci-dessous (théorème 1), que de tels réels existent toujours !

Remarques :

- Le deuxième quartile Q_2 ne se défini pas puisqu'il s'agit de la médiane m_e.
- Les trois quartiles partagent l'ensemble des valeurs en quatre sous ensembles de (presque) même effectif.
- On a toujours : $Q_1 \leqslant m_e \leqslant Q_3$.

Exemples :

Pour l'élève *A*, on peut choisir : Q_1 dans $[x_2 ; x_3] = [5 ; 6]$ et Q_3 dans $[x_6 ; x_7] = [12 ; 13]$

Pour l'élève *B*, on a : $Q_1 = x_3 = 5$ et $Q_3 = x_7 = 9$ (pas d'autres choix possibles)

Pour l'élève *C*, on a : $Q_1 = x_3 = 10$ et $Q_3 = x_8 = 15$ (pas d'autres choix possible)

Pour l'élève *D*, on peut choisir : $Q_1 = x_3 = 1$ et $Q_3 = x_9 = 13$ (pas d'autre choix possible)

On constate que la détermination des quartiles est différente suivant que l'effectif total N est un **multiple de 4 ou non** :

- Lorsque l'effectif total N n'est **pas un multiple de 4**, il n'y a pas de difficulté, les quartiles Q_1 et Q_3 sont les termes de rang immédiatement supérieur à $\frac{N}{4}$ et $\frac{3N}{4}$:

$$Q_1 = x_{\left\lfloor \frac{N}{4} \right\rfloor + 1} \qquad Q_3 = x_{\left\lfloor \frac{3N}{4} \right\rfloor + 1}$$

- Lorsque l'effectif total est un **multiple de 4**, alors l'usage veut que l'on choisisse pour quartiles Q_1 et Q_3 les termes de rang $\frac{N}{4}$ et de rang $\frac{3N}{4}$. On a donc $Q_1 = x_{\frac{N}{4}}$ et $Q_3 = x_{\frac{3N}{4}}$. Mais tout réel de l'intervalle $[\, x_{\frac{N}{4}} ; x_{\frac{N}{4}+1} \,]$ conviendrait également pour Q_1 et tout réel de l'intervalle $[\, x_{\frac{3N}{4}} ; x_{\frac{3N}{4}+1} \,]$ conviendrait également pour Q_3.

Exemple : si $N = 29$ alors $Q_1 = x_8$ et $Q_3 = x_{22}$; si $N = 44$ alors $Q_1 = x_{11}$ et $Q_3 = x_{33}$.

Voici un théorème qui donne des formules qui marchent dans tous les cas !

Théorème 1

Soient $N \in \mathbb{N}^*$ et $(x_i)_{1 \leqslant i \leqslant N}$ une famille de réels **ordonnés dans l'ordre croissant**. Les réels :

$$Q_1 = x_{\left\lfloor \frac{N}{4} \right\rfloor + 1} \qquad m_e = x_{\left\lfloor \frac{N}{2} \right\rfloor + 1} \qquad Q_3 = x_{\left\lfloor \frac{3N}{4} \right\rfloor + 1}$$

définissent toujours des valeurs convenables pour le premier quartile, la médiane et le troisième quartile.

Pour démontrer ce théorème, on aura besoin du petit lemme suivant :

Lemme

Soient A et B des éléments de $\mathbb{N}$ avec $A \leqslant B$. L'ensemble $[\![A ; B]\!]$ contient $B - A + 1$ entiers.

Preuve du lemme :

L'ensemble $[\![A ; B]\!]$ contient autant d'entiers que l'ensemble $[\![A - A + 1 ; B - A + 1]\!] = [\![1 ; B - A + 1]\!]$ qui lui même en contient $B - A + 1$.

Démonstration du théorème 1 :

Pour tout réel λ, notons $E(\lambda) = \{i \in [\![1 ; N]\!] \text{ tels que } x_i \leqslant \lambda\}$ et $F(\lambda) = \{i \in [\![1 ; N]\!] \text{ tels que } x_i \geqslant \lambda\}$

$E(\lambda)$ est l'ensemble des indices des termes de la famille $(x_i)_{1 \leqslant i \leqslant N}$ qui sont inférieurs à λ et $F(\lambda)$ est l'ensemble des indices des termes de la famille $(x_i)_{1 \leqslant i \leqslant N}$ qui sont supérieurs à λ.

Posons : $Q_1 = x_{\left\lfloor \frac{N}{4} \right\rfloor + 1}$ $\quad m_e = x_{\left\lfloor \frac{N}{2} \right\rfloor + 1}$ $\quad Q_3 = x_{\left\lfloor \frac{3N}{4} \right\rfloor + 1}$

Montrons que m_e est une valeur convenable pour la médiane : soit $i \in [\![1 ; N]\!]$

$$x_i \leqslant m_e \;\Leftrightarrow\; x_i \leqslant x_{\left[\frac{N}{2}\right]+1} \;\Leftrightarrow\; 1 \leqslant i \leqslant \left[\frac{N}{2}\right]+1 \;\Leftrightarrow\; i \in \left[\!\left[1\,;\,\left[\frac{N}{2}\right]+1\right]\!\right]$$

Or, dans $\left[\!\left[1\,;\,\left[\frac{N}{2}\right]+1\right]\!\right]$ il y a $\left[\frac{N}{2}\right]+1$ entiers.

Donc $\quad \text{Card}(E(m_e)) = \left[\frac{N}{2}\right]+1$

Or, $\left[\frac{N}{2}\right] \leqslant \frac{N}{2} \leqslant \left[\frac{N}{2}\right]+1$, donc : $\quad \text{Card}(E(m_e)) \geqslant \frac{N}{2}$

De même :

$$x_i \geqslant m_e \;\Leftrightarrow\; x_i \geqslant x_{\left[\frac{N}{2}\right]+1} \;\Leftrightarrow\; N \geqslant i \geqslant \left[\frac{N}{2}\right]+1 \;\Leftrightarrow\; i \in \left[\!\left[\left[\frac{N}{2}\right]+1\,;\,N\right]\!\right]$$

Or, dans $\left[\!\left[\left[\frac{N}{2}\right]+1\,;\,N\right]\!\right]$ il y a $N-\left[\frac{N}{2}\right]$ entiers.

Donc $\quad \text{Card}(F(m_e)) = N-\left[\frac{N}{2}\right]$

Or, $\left[\frac{N}{2}\right] \leqslant \frac{N}{2} \leqslant \left[\frac{N}{2}\right]+1$ donc $-\left[\frac{N}{2}\right] \geqslant -\frac{N}{2}$ et en ajoutant N : $N-\left[\frac{N}{2}\right] \geqslant \frac{N}{2}$ donc $\text{Card}(F(e_m)) \geqslant \frac{N}{2}$.

On a donc bien :

au moins 50% des termes de la série ont une valeur inférieure ou égale à m_e

et

au moins 50% des termes de la série ont une valeur supérieure ou égale à m_e

Donc m_e est bien une valeur médiane de la série.

Montrons que Q_1 est une valeur convenable pour le premier quartile : soit $i \in \left[\!\left[1\,;\,N\right]\!\right]$

$$x_i \leqslant Q_1 \;\Leftrightarrow\; x_i \leqslant x_{\left[\frac{N}{4}\right]+1} \;\Leftrightarrow\; i \leqslant \left[\frac{N}{4}\right]+1 \;\Leftrightarrow\; i \in \left[\!\left[1\,;\,\left[\frac{N}{4}\right]+1\right]\!\right]$$

Or, dans $\left[\!\left[1\,;\,\left[\frac{N}{4}\right]+1\right]\!\right]$ il y a $\left[\frac{N}{4}\right]+1$ entiers.

Donc $\quad \text{Card}(E(Q_1)) = \left[\frac{N}{4}\right]+1$

Or, $\left[\frac{N}{4}\right] \leqslant \frac{N}{4} \leqslant \left[\frac{N}{4}\right]+1$, donc : $\quad \text{Card}(E(Q_1)) \geqslant \frac{N}{4}$

De même :

$$x_i \geqslant Q_1 \;\Leftrightarrow\; x_i \geqslant x_{\left[\frac{N}{4}\right]+1} \;\Leftrightarrow\; i \geqslant \left[\frac{N}{4}\right]+1 \;\Leftrightarrow\; i \in \left[\!\left[\left[\frac{N}{4}\right]+1\,;\,N\right]\!\right]$$

Or, dans $\left[\!\left[\left[\frac{N}{4}\right]+1\,;\,N\right]\!\right]$ il y a $N-\left[\frac{N}{4}\right]$ entiers.

Donc $\text{Card}(F(Q_1)) = N - \left[\frac{N}{4}\right]$

Or, $\left[\frac{N}{4}\right] \leqslant \frac{N}{4} \leqslant \left[\frac{N}{4}\right] + 1$ donc $-\left[\frac{N}{4}\right] \geqslant -\frac{N}{4}$ et en ajoutant N : $N - \left[\frac{N}{4}\right] \geqslant \frac{3N}{4}$ donc $\text{Card}(F(Q_1)) \geqslant \frac{3N}{4}$.

On a donc bien :

au moins 25% des termes de la série ont une valeur inférieure ou égale à Q_1

et

au moins 75% des termes de la série ont une valeur supérieure ou égale à Q_1

Donc Q_1 est bien une valeur du premier quartile de la série.

La démonstration est analogue pour Q_3.

b) Cas d'une série statistique (discrète ou continue) avec regroupement en classes

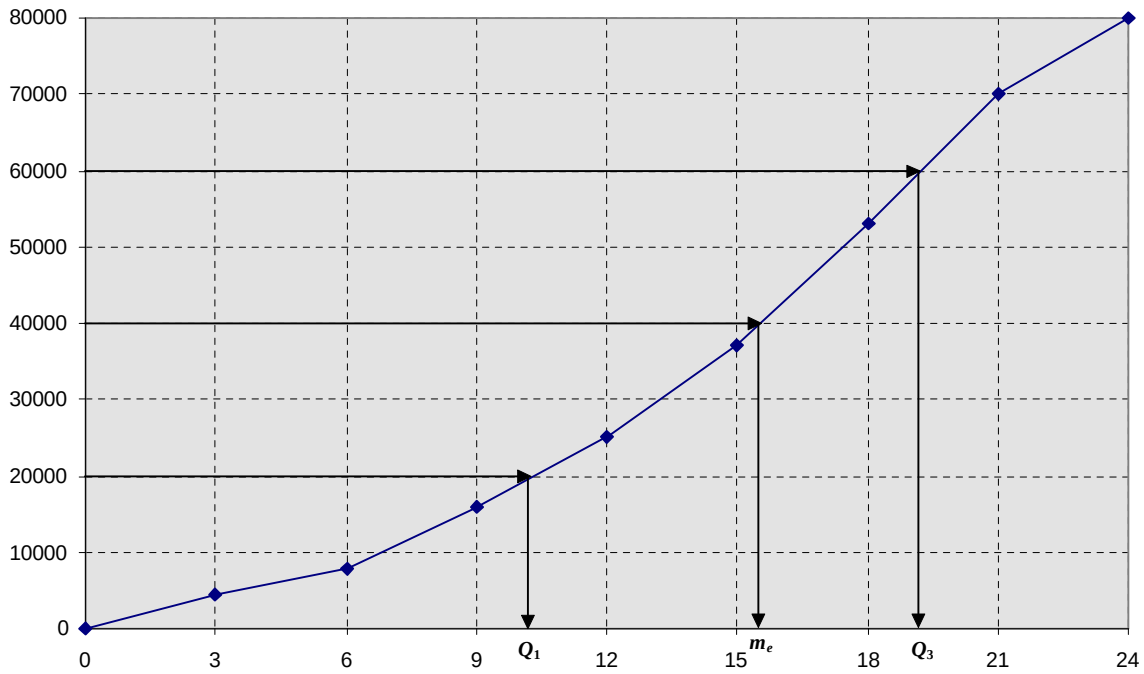
Dans ce cas, médiane et quartiles peuvent se déterminer à l'aide du **polygone des effectifs (ou fréquences) cumulé(e)s croissant(e)s**.

Exemple :

La répartition des accidents corporels de la route selon les heures de la journée est décrite par le tableau suivant, pour l'année 1999.

Tranche horaire	[0 ; 3[	[3 ; 6[	[6 ; 9[	[9 ; 12[	[12 ; 15[	[15 ; 18[	[18 ; 21[	[21 ; 24[	Total
Nombre d'accidents	4550	3230	8220	9050	12040	16040	16820	10050	80000
Effectifs cumulés croissants	4550	7780	16000	25050	37090	53130	69950	80000	

On trace ensuite le polygone des effectifs cumulés croissants :



Si N est l'effectif total et f la fonction affine par morceaux correspondant au polygone des effectifs cumulés croissants. Alors, on considère que les quartiles et la médiane sont définis par :

$$Q_1 = f^{-1}\left(\frac{N}{4}\right) \qquad m_e = f^{-1}\left(\frac{N}{2}\right) \qquad Q_3 = f^{-1}\left(\frac{3N}{4}\right)$$

Dans notre cas $N = 80000$.

Calculons $Q_1 = f^{-1}(20000)$:

Posons $A(9 ; 16000)$, $B(12 ; 25050)$ et $M_1(Q_1 ; 20000)$.

Comme les vecteurs $\overrightarrow{AB}\begin{vmatrix}3\\9050\end{vmatrix}$ et $\overrightarrow{AM_1}\begin{vmatrix}Q_1-9\\4000\end{vmatrix}$ sont colinéaires, on a : $3 \times 4000 - 9050(Q_1 - 9) = 0$

D'où $Q_1 = \dfrac{1869}{131} \simeq 10{,}3$ (à 10^{-1} près. Inutile de donner un résultat plus précis, cela n'aurait pas de sens car le regroupement en classe gomme déjà beaucoup de la précision)

Interprétation : un quart des accidents corporels quotidiens ont lieu entre 0h00 et 10h20 du matin.

On calcule de même $m_e = f^{-1}\left(\dfrac{N}{2}\right)$ et $Q_3 = f^{-1}\left(\dfrac{3N}{4}\right)$ à l'aide des points $C(15 ; 37090)$, $M(m_e ; 40000)$, $D(18 ; 53130)$, $M_3(Q_3 ; 60000)$ et $E(21 ; 69950)$:

Comme les vecteurs $\overrightarrow{CD}\begin{vmatrix}3\\16040\end{vmatrix}$ et $\overrightarrow{CM}\begin{vmatrix}m_e-15\\2910\end{vmatrix}$ sont colinéaires, on a : $3 \times 2910 - 16040(m_e - 15) = 0$

D'où $m_e = \dfrac{24933}{1604} \simeq 15{,}6$ (à 10^{-1} près)

Interprétation : la moitié des accidents corporels quotidiens ont lieu entre 0h00 et 15h40.

Comme les vecteurs $\overrightarrow{DE}\begin{vmatrix}3\\16820\end{vmatrix}$ et $\overrightarrow{DM_3}\begin{vmatrix}Q_3-18\\6870\end{vmatrix}$ sont colinéaires, on a : $3 \times 6870 - 16820(Q_3 - 18) = 0$

D'où $Q_3 = \dfrac{32337}{1682} \simeq 19{,}2$ (à 10^{-1} près)

Remarques :

- Une simple lecture graphique donne souvent une précision satisfaisante.
- Si on construit le polygone des fréquences cumulées croissantes alors Q_1, m_e et Q_3 sont les antécédents respectifs de 0,25 ; 0,5 et 0,75.
- Dans le cas d'un regroupement en classe, les statisticiens parlent rarement de valeur médiane mais plutôt de **classe médiane**.

c) Propriété de la médiane et des quartiles

Propriété 1

Soient $N \geqslant 5$ et $(x_i)_{1 \leqslant i \leqslant N}$ une famille de réels **ordonnés dans l'ordre croissant**.

Soient Q_1, Q_3 et m_e les quartiles et la médiane de la série $(x_i)_{1 \leqslant i \leqslant N}$.

Soit m et M le minimum et le maximum de la série $(x_i)_{1 \leqslant i \leqslant N}$.

Si l'on remplace m par un réel de $]-\infty ; Q_1[$ ou M par un réel de $]Q_3 ; +\infty[$ alors les quartiles restent inchangés.

Si l'on remplace m par un réel de $]-\infty ; m_e[$ ou M par un réel de $]m_e ; +\infty[$ alors la médiane reste inchangée.

Exemple :

Considérons la série suivante :

$x_1 = 1$ $x_2 = 5$ $x_3 = 8$ $x_4 = 15$ $x_5 = 29$ $x_6 = 35$

On a : $Q_1 = x_2 = 5$; $m_e = \frac{1}{2}(x_3 + x_4) = 11{,}5$; $Q_3 = x_5 = 29$.

Si l'on remplace $m = x_1 = 1$ par un réel de $]-\infty ; 5[$, cela ne changera pas les valeurs de Q_1 ; m_e et Q_3. (Même si la série est à réordonner)

Par contre, si l'on remplace m par un réel supérieur à Q_1, par exemple par 9.

En réordonnant la série, on obtient :

$y_1 = 5$ $y_2 = 8$ $y_3 = 9$ $y_4 = 15$ $y_5 = 29$ $y_6 = 35$

On constate que Q_1 devient égal à $y_2 = 8$ et m_e devient égal à $\frac{1}{2}(y_3 + y_4) = 12$.

Remarque :

On dit parfois que la médiane et les quartiles sont insensibles aux termes extrêmes.

Démonstration de la propriété :

En remplaçant x_1 par un réel de $]-\infty ; Q_1[$, on ne change pas le nombre de termes de la série qui ont une valeur inférieure ou égale à Q_1 (il y en aura donc toujours au moins 25%) ni le nombre de termes de la série qui ont une valeur supérieure ou égale à Q_1 (il y en aura donc toujours au moins 75%). Donc Q_1 reste une valeur convenable du premier quartile de la série.

Même raisonnement pour le reste...

d) Diagrammes en boîtes (ou boîtes à moustaches)

Définition 3

Soient $N \in \mathbb{N}^*$ et $(x_i)_{1 \leqslant i \leqslant N}$ une famille de réels **ordonnés dans l'ordre croissant**.

(Ainsi $x_1 = \min_i x_i$ et $x_N = \max_i x_i$)

Soient m_e, Q_1 et Q_3 la médiane et les quartiles de $(x_i)_{1 \leqslant i \leqslant N}$.

- On appelle étendue la différence $x_N - x_1$. (Différence entre les termes extrêmes de la série)
- On appelle interquartile la différence $Q_3 - Q_1$.
- On appelle intervalle interquartile l'intervalle $[Q_1 ; Q_3]$.
- Lorsque $m_e \neq 0$, on définit l'interquartile relatif par le quotient : $\frac{Q_3 - Q_1}{m_e}$. (Grandeur sans unité)

Remarque : l'interquartile est un indicateur de dispersion (au même titre que l'étendue ou l'écart-type). Son avantage est qu'il ne tient compte que de 50% de la population, ce qui a pour effet d'ignorer les valeurs extrêmes souvent marginales. Il est donc assez utilisé car considéré comme "standard".

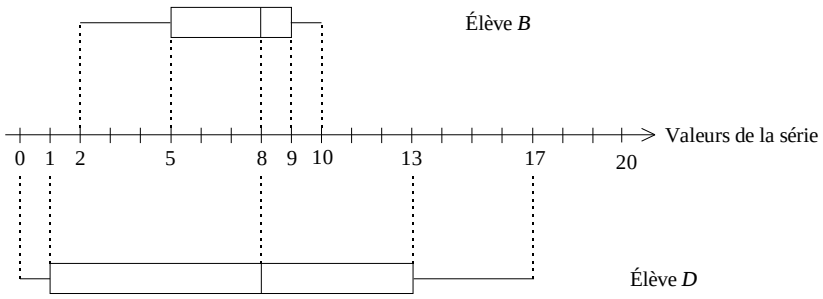
Exemple :

Pour l'élève *B*, l'étendue est $e = 8$, l'intervalle interquartile est $[5 ; 9]$.

Pour l'élève *D*, l'étendue est $e = 17$, l'intervalle interquartile est $[1 ; 13]$.

Le diagramme en boîte permet de visualiser les éléments suivants :

minimum premier quartile médiane troisième quartile maximum



La boîte (de largeur arbitraire) représente 50% (au moins) de l'effectif total.

De cette boîte s'étirent deux moustaches (représentées par des traits) jusqu'au minimum et au maximum.

Ces diagrammes permettent une interprétation visuelle et rapide de la dispersion des séries statistiques. Ils permettent également d'apprécier des différences entre des séries. (Lorsqu'elles ont des ordres de grandeurs comparables ; sinon, on utilise l'interquartile relatif, voir II)b) exemple 2).

Dans notre exemple, nos deux élèves *B* et *D* ont la même note médiane (8) mais les résultats de *D* sont bien plus dispersés que ceux de *B*.

e) Effet d'un changement affine

Théorème 2

Soit $N \in \mathbb{N}^*$

Soit $(x_i)_{1 \leqslant i \leqslant N}$ une famille de réels **ordonnés dans l'ordre croissant** de médiane m_e et de quartiles Q_1 et Q_3.

Soient $a \in \mathbb{R}^*$ et $b \in \mathbb{R}$. Soit $(y_i)_{1 \leqslant i \leqslant N}$ la famille de réels définis par : $y_i = ax_i + b$ pour tout $i \in [\![1 ; N]\!]$.

Si $a > 0$ alors la famille $(y_i)_{1 \leqslant i \leqslant N}$ est **ordonnée dans l'ordre croissant**. Les réels suivants :

$$m_e' = am_e + b \qquad Q_1' = aQ_1 + b \qquad Q_3' = aQ_3 + b$$

sont des valeurs convenables de la médiane et des quartiles de la famille $(y_i)_{1 \leqslant i \leqslant N}$.

Si $a < 0$ alors la famille $(y_i)_{1 \leqslant i \leqslant N}$ est **ordonnée dans l'ordre décroissant** Les réels suivants :

$$m_e' = am_e + b \qquad Q_1' = aQ_3 + b \qquad Q_3' = aQ_1 + b$$

sont des valeurs convenables de la médiane et des quartiles de la famille $(y_i)_{1 \leqslant i \leqslant N}$.

Démonstration :

Lorsque $a > 0$, la fonction affine $f : t \mapsto at + b$ est croissante. On a alors :

$$x_i \leqslant Q_3 \Leftrightarrow f(x_i) \leqslant f(Q_1) \Leftrightarrow ax_i + b \leqslant aQ_1 + b \Leftrightarrow y_i \leqslant Q_1'$$

Donc

$$\{i \in [\![1 ; N]\!] \text{ tels que } x_i \leqslant Q_1\} = \{i \in [\![1 ; N]\!] \text{ tels que } y_i \leqslant Q_1'\}$$

Et puisque ces ensembles d'indices sont identiques, ils ont a fortiori le même nombre d'éléments :

$$\text{Card}\{i \in [\![1 ; N]\!] \text{ tels que } x_i \leqslant Q_1\} = \text{Card}\{i \in [\![1 ; N]\!] \text{ tels que } y_i \leqslant Q_1'\}$$

Or, Card$\{i \in [\![1 ; N]\!]$ tels que $x_i \leqslant Q_1\} \geqslant \dfrac{N}{4}$ puisque Q_1 est le premier quartile de $(x_i)_{1 \leqslant i \leqslant N}$.

Donc Card$\{i \in [\![1 ; N]\!]$ tels que $y_i \leqslant Q_1'\} \geqslant \dfrac{N}{4}$.

On démontre de même que Card$\{i \in [\![1 ; N]\!]$ tels que $y_i \geqslant Q_1'\} \geqslant \dfrac{3N}{4}$.

On en déduit, d'après la définition 2 que $Q_1' = aQ_1 + b$ est le premier quartile de $(y_i)_{1 \leqslant i \leqslant N}$.

Lorsque $a < 0$, la fonction affine $f : t \mapsto at + b$ est décroissante. On a alors :

$$x_i \leqslant Q_3 \Leftrightarrow f(x_i) \geqslant f(Q_3) \Leftrightarrow ax_i + b \geqslant aQ_3 + b \Leftrightarrow y_i \geqslant Q_1'$$

Donc

$$\{i \in [\![1 ; N]\!] \text{ tels que } x_i \leqslant Q_3\} = \{i \in [\![1 ; N]\!] \text{ tels que } y_i \geqslant Q_1'\}$$

Et puisque ces ensembles d'indices sont identiques, ils ont a fortiori le même nombre d'éléments :

$$\text{Card}\{i \in [\![1 ; N]\!] \text{ tels que } x_i \leqslant Q_3\} = \text{Card}\{i \in [\![1 ; N]\!] \text{ tels que } y_i \geqslant Q_1'\}$$

Or, Card$\{i \in [\![1 ; N]\!]$ tels que $x_i \leqslant Q_3\} \geqslant \dfrac{3N}{4}$ puisque Q_3 est le troisième quartile de $(x_i)_{1 \leqslant i \leqslant N}$.

Donc Card$\{i \in [\![1 ; N]\!]$ tels que $y_i \geqslant Q_1'\} \geqslant \dfrac{N}{4}$.

On démontre de même que Card$\{i \in [\![1 ; N]\!]$ tels que $y_i \leqslant Q_1'\} \geqslant \dfrac{3N}{4}$.

On en déduit, d'après la définition 2 que $Q_1' = aQ_3 + b$ est le premier quartile de $(y_i)_{1 \leqslant i \leqslant N}$.

La démonstration est analogue pour m_e' et Q_3'.

Exemple :

Dans une entreprise les salaires sont résumés par :

	Minimum	Premier quartile	Médiane	Moyenne	Troisième quartile	Maximum
Salaires en €	$m = 1020$	$Q_1 = 1200$	$m_e = 1400$	$\bar{x} = 1450$	$Q_3 = 1800$	$M = 3800$

Le conseil d'administration décide d'une augmentation des salaires de 2% auquel s'ajoute encore une indemnité de 10 €.

Cela se traduit par la transformation affine f définie par : $f(x) = 1{,}02x + 10$. (Ici $a > 0$)

Cela donne : $f(m) = 1050{,}4$; $f(M) = 3886$ pour le minimum et le maximum.

D'après le théorème 2, cela donne : $f(Q_1) = 1234$; $f(m_e) = 1438$ et $f(Q_3) = 1846$.

Enfin, la nouvelle moyenne est donnée par $f(\bar{x})$. En effet :

Notons $(x_i)_{1 \leqslant i \leqslant N}$ la série des salaires initiaux et posons $y_i = f(x_i)$, pour $i \in [\![1 ; N]\!]$. La série $(y_i)_{1 \leqslant i \leqslant N}$ correspond aux nouveaux salaires. La moyenne $\bar{y}$ des nouveaux salaires est :

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i = \frac{1}{N}\sum_{i=1}^{N}\left(ax_i + b\right) = \frac{1}{N}\left(a\sum_{i=1}^{N} x_i + \sum_{i=1}^{N} b\right) = \frac{1}{N}(aN\bar{x} + Nb) = a\,\bar{x} + b = f(\bar{x})$$

Dans notre cas, cela donne : $\bar{y} = 1489$.

D'où le nouveau tableau :

	Minimum	Premier quartile	Médiane	Moyenne	Troisième quartile	Maximum
Nouveaux Salaires en €	$m = 1050{,}4$	$Q_1 = 1234$	$m_e = 1438$	$\bar{x} = 1489$	$Q_3 = 1846$	$M = 3886$

II) Moyenne, variance et écart-type

Dans ce paragraphe, nous utiliserons une nouvelle notation. Soit $(z_i)_{1\leqslant i\leqslant N}$ une série statistique. Certains de ces réels peuvent être confondus. Notons p le nombre de valeurs de la série ($1 \leqslant p \leqslant N$) et, pour tout $i \in [\![1\,;\,p]\!]$, notons x_i ces valeurs et n_i l'effectif de x_i. On notera $(x_i, n_i)_{1\leqslant i\leqslant p}$ la série statistique ainsi obtenue où les x_i sont distincts deux à deux.

a) Définitions

Définition 4

La moyenne d'une série statistique $(x_i, n_i)_{1\leqslant i\leqslant p}$ est le nombre $\bar{x}$ défini par :

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{p} n_i x_i \quad \text{où} \quad N = \sum_{i=1}^{p} n_i \ \text{(Effectif total)}$$

La variance d'une série statistique $(x_i, n_i)_{1\leqslant i\leqslant p}$ est le nombre noté V et défini par :

$$V = \frac{1}{N}\sum_{i=1}^{p} n_i\left(x_i - \bar{x}\right)^2$$

la variance est la moyenne des carrés des écarts à la moyenne

L'écart-type d'une série statistique $(x_i, n_i)_{1\leqslant i\leqslant p}$ est le nombre noté s (ou σ) et défini par :

$$s = \sqrt{V}$$

Remarques :

- La variance est une somme de carrés. C'est donc une quantité positive. L'écart-type est donc bien défini. Et il s'exprime dans la même unité que la caractère étudié.
- Si on note $f_i = \frac{n_i}{N}$ la fréquence de x_i, les formules deviennent : $\bar{x} = \sum_{i=1}^{p} f_i x_i$ et $V = \sum_{i=1}^{p} f_i\left(x_i - \bar{x}\right)^2$.
- Dans le cas d'un regroupement en classe, les calculs sont effectués en choisissant x_i au centre de chaque classe (c'est l'hypothèse de répartition uniforme de chaque classe)

Pour calculer la variance, on dispose d'une formule un peu plus pratique :

Théorème 3

La variance d'une série statistique $(x_i, n_i)_{1 \leqslant i \leqslant p}$ peut se calculer avec la relation suivante :

$$V = \frac{1}{N} \sum_{i=1}^{p} n_i x_i^2 - \bar{x}^2$$

la variance est l'écart entre la moyenne des carrés et le carré de la moyenne

Démonstration :

$$\sum_{i=1}^{p} f_i (x_i - \bar{x})^2 = \sum_{i=1}^{p} f_i \left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right) = \sum_{i=1}^{p} f_i x_i^2 - 2\bar{x} \sum_{i=1}^{p} f_i x_i + \bar{x}^2 = \sum_{i=1}^{p} f_i x_i^2 - 2\bar{x}^2 + \bar{x}^2 = \sum_{i=1}^{p} f_i x_i^2 - \bar{x}^2$$

b) Interprétation de l'écart-type

La variance est la moyenne des carrés des écarts à la moyenne. Elle mesure donc la dispersion des valeurs autour de la moyenne. Elle n'est pas très parlante car elle s'exprime dans le carré de l'unité du caractère. L'écart-type a l'avantage de s'exprimer dans la même unité que le caractère.

L'écart-type permet de comparer la dispersion de deux séries. Contrairement à l'interquartile, il tient compte de l'ensemble de la population.

Exemple 1 : cas de séries dont les ordres de grandeurs sont comparables (et de moyennes voisines)

L'élève *A* a obtenu les dix notes suivantes : 10 15 16 13 8 11 12 12 13 15

L'élève *B* a obtenu les dix notes suivantes : 11 9 9 10 15 7 12 12 14 13

Calculer les moyennes de *A* et *B*. Quel est l'élève qui a les résultats les plus homogènes ?

Moyenne de *A* : $m_A = 12{,}5$; moyenne de *B* : $m_B = 11{,}2$.

Variance de *A* (théorème 3) : $V_A = \frac{1}{10}(10^2 + 2 \times 15^2 + 16^2 + 2 \times 13^2 + 8^2 + 11^2 + 2 \times 12^2) - 12{,}5^2 = 5{,}45$

D'où l'écart-type de *A* : $s_A = 2{,}33$ (à 10^{-2}près)

De même : $V_B = \frac{1}{10}(7^2 + 2 \times 9^2 + 10^2 + 11^2 + 2 \times 12^2 + 13^2 + 14^2 + 15^2) - 11{,}2^2 = 5{,}56$

D'où l'écart-type de *B* : $s_B = 2{,}36$ (à 10^{-2} près)

Les élèves *A* et *B* ont des résultats d'homogénéité comparable.

(Remarque : l'interquartile de *A* est 15 – 11 = 4 ; celui de *B* est 13 – 9 = 4)

Exemple 2 : cas de séries dont les ordres de grandeurs sont différents.

Dans ce cas, l'écart-type du caractère prenant les plus grandes valeurs sera certainement supérieur au second. Mais cela ne signifie pas, pour autant, que ses valeurs soient plus dispersées. On introduit alors un nouvel indicateur, appelé coefficient de variation C_v :

$$C_v = \frac{s_x}{\bar{x}} \text{ (défini pour des séries dont la moyenne } \bar{x} \text{ est non nulle)}$$

Le coefficient de variation a pour effet de relativiser l'écart-type par rapport à la moyenne.

Attention, le coefficient de variation n'a pas d'unité !

Étudions un cas concret : cinq sportifs ont couru un 1500m et un 5000m. Leurs temps sont donnés dans le tableau suivant :

	Coureur 1	Coureur 2	Coureur 3	Coureur 4	Coureur 5
1500 m	3'58"17	4'05"48	4'12"97	4'08"29	4'00"12
5000 m	14'58"12	14'47"08	15'37"85	13'57"70	14'48"34

Laquelle des deux courses a les temps les plus homogènes ?

Pour le 1500 m : (on convertit tous les temps en secondes pour un calcul plus aisé)

- moyenne : $m = \frac{1}{5}(238{,}17 + 245{,}48 + 252{,}97 + 248{,}29 + 240{,}12) = 245{,}006$ secondes (soit environ 4'05"01)
- variance : $V = \frac{1}{5}(238{,}17^2 + 245{,}48^2 + 252{,}97^2 + 248{,}29^2 + 240{,}12^2) - 245{,}006^2 \simeq 29{,}0$ d'où un écart-type $s \simeq 5{,}39$ secondes
- coefficient de variation : $C_v = \frac{s}{m} \simeq 0{,}022$.

Pour le 5000 m :

- moyenne : $m' = \frac{1}{5}(898{,}12 + 887{,}08 + 937{,}85 + 837{,}70 + 888{,}34) = 889{,}818$ secondes (soit environ 14'49"82)
- variance : $V' = \frac{1}{5}(898{,}12^2 + 887{,}08^2 + 937{,}85^2 + 837{,}70^2 + 888{,}34^2) - 889{,}818^2 \simeq 1020{,}4$ d'où un écart-type $s' \simeq 31{,}94$ secondes
- coefficient de variation : $C_v' = \frac{s'}{m'} \simeq 0{,}036$.

Conclusion : le 1500 m a été plus homogène car $C_v < C_v'$.

On peut également, dans ce type de situation, utiliser l'interquartile relatif.

Pour le 1500 m, on a $\frac{Q_3 - Q_1}{m_e} = \frac{248{,}29 - 240{,}12}{245{,}48} \simeq 0{,}033$.

Pour le 5000 m, on a : $\frac{Q_3' - Q_1'}{m_e'} = \frac{898{,}12 - 887{,}08}{888{,}34} \simeq 0{,}012...$

Conclusion : le 5000 m a été plus homogène que le 1500 m.

Moralité : surtout lorsque les effectifs sont petits, le coefficient de variation et l'interquartile relatif n'aboutissent pas toujours aux mêmes conclusions. (Rappel : l'interquartile ne tient compte que de 50% de la population)

c) Effet d'un changement affine

Théorème 4

Soit $(x_i, n_i)_{1 \leqslant i \leqslant p}$ une série statistique de variance V_x et d'écart-type s_x.

Soient $a \in \mathbb{R}^*$ et $b \in \mathbb{R}$.

Soit $(y_i, n_i)_{1 \leqslant i \leqslant p}$ la série statistique définie par $y_i = ax_i + b$, pour tout $i \in [\![1 ; p]\!]$.

Notons V_y sa variance et s_y son écart-type.

Alors : $V_y = a^2 V_x$ et $s_y = |a| s_x$

Démonstration :

On rappelle que $\bar{y} = a\bar{x} + b$. (En effet : $\bar{y} = \sum_{i=1}^{p} f_i y_i = \sum_{i=1}^{p} f_i (ax_i + b) = a\sum_{i=1}^{p} f_i x_i + b\sum_{i=1}^{p} f_i = a\bar{x} + b$)

$$V_y = \sum_{i=1}^{p} f_i \left(y_i - \bar{y}\right)^2 = \sum_{i=1}^{p} f_i \left(ax_i + b - a\bar{x} - b\right)^2 = a^2 \sum_{i=1}^{p} f_i \left(x_i - \bar{x}\right)^2 = a^2 V_x$$

Et comme $\sqrt{a^2} = |a|$: $\qquad s_y = |a| s_x$

Exemple :

Soit $(x_i, n_i)_{1 \leqslant i \leqslant p}$ une série statistique de moyenne $\bar{x}$ et d'écart-type s_x.

On définit une nouvelle série statistique $(y_i, n_i)_{1 \leqslant i \leqslant p}$ par : $y_i = \dfrac{x_i - \bar{x}}{s_x}$ pour tout $i \in [\![1 ; p]\!]$.

Calculer la moyenne $\bar{y}$ et l'écart-type s_y de $(y_i, n_i)_{1 \leqslant i \leqslant p}$.

On a donc un changement affine ($y = ax + b$) avec $a = \dfrac{1}{s_x}$ et $b = -\dfrac{\bar{x}}{s_x}$.

On sait déjà que $\bar{y} = a\bar{x} + b = \dfrac{1}{s_x}\bar{x} - \dfrac{\bar{x}}{s_x} = 0$.

D'après le théorème 4, $s_y = |a| s_x = \dfrac{1}{s_x} \times s_x = 1$.

La série statistique $(y_i, n_i)_{1 \leqslant i \leqslant p}$ a donc une moyenne nulle et un écart-type égal à 1.

(On dit que l'on a "centré et réduit" la série statistique $(y_i, n_i)_{1 \leqslant i \leqslant p}$ ou encore que l'on a standardisé les données).