Statistiques

On utilisera la série statistique suivante pour les exemples :
Dans une classe de première L comptant 26 élèves, on a collecté le temps consacré à la lecture chaque semaine. Les données ont été rassemblées dans un tableau :

Durée (h)	2	3	4	5	6	7	8	9
Effectifs	3	4	6	2	4	4	1	2

I) Paramètres de position : moyenne, médiane, quartiles

définitions :
Soit une série statistique X d'effectif total N, à caractère quantitatif, dont les valeurs du caractère ont été rangées dans l'ordre croissant $x_1 \leqslant x_2 \ldots\ldots\ldots \leqslant x_p$.
L' effectif correspondant à chaque valeur est noté $n_1, n_2, \ldots\ldots n_p$.

$N = n_1 + n_2 + \ldots. + n_p$!

► La **moyenne** de la série statistique est le nombre réel, noté $\overline{x}$, tel que :

$$\overline{x} = \frac{n_1x_1 + n_2x_2 + \ldots\ldots\ldots + n_px_p}{N}$$

ou

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{i=p} n_ix_i$$

$\sum_{i=1}^{i=p} x_i$ signifie $x_1 + x_2 + \ldots.. + x_p$
c'est la somme des x_i de i=1 à i = p !

Ex : Calculons la moyenne de la série statistique utilisée comme exemple.

En notant x_i les valeurs et n_i les effectifs correspondants, on a

$$\overline{x} = \frac{n_1x_1 + n_2x_2 + \ldots\ldots\ldots + n_px_p}{N} = \frac{3 \times 2 + 4 \times 3 + 6 \times 4 + 2 \times 5 + 4 \times 6 + 4 \times 7 + 1 \times 8 + 2 \times 9}{26} = 5 \text{ h}$$

► La **médiane** de la série statistique est une valeur, notée Me, tel que :
- 50% au moins des individus ont une valeur du caractère inférieure ou égale à Me
- 50% au moins des individus ont une valeur du caractère supérieure ou égale à Me

rappels : La médiane permet de couper la population étudiée en deux groupes comprenant le même nombre d'individus.
Soit une série statistique de N valeurs ordonnées par ordre croissant :
- **si N est impair** (N=2n+1), la médiane est **la donnée de rang n + 1**

Ex : Un élève a eu les notes suivantes en Mathématiques pendant un trimestre : 7;7;8;9;11;13;13;14;15
L'effectif total impair est égal à 9 (2 x **4** + 1). La médiane est la note correspondante au rang 5 (**4**+1) soit 11
- **si N est pair** (N=2n), la médiane est **la demi-somme des données de rang n et de rang n + 1**

Ex : Un élève a eu les notes suivantes en Français pendant un trimestre : 8; 9;10;12;13;14;15;17
L'effectif total pair est égal à 8 (2 x **4)**.
La médiane est la demi-somme des notes de rang 4 et de rang 5 soit $\frac{12 + 13}{2} = 12{,}5$

Ex : Calculons la médiane de la série statistique utilisée comme exemple.
L'effectif total est 26 qui est pair. 26 = 13 x 2.
La médiane **Me** de la série est donc $\frac{4 + 5}{2}$ = **4,5 h**

► **Le premier quartile Q_1** est la plus petite valeur Q_1 de la série telle qu'au moins un quart des valeurs de la liste sont inférieures ou égales à Q_1
Ex : Déterminons le premier quartile Q_1 de la série statistique utilisée comme exemple.
L'effectif total est 26 et $\frac{1}{4}$ x 26 = 6,5. On cherche **la plus petite valeur de la série** pour laquelle **au moins 7 données lui sont inférieures ou égales**. Donc Q_1 = **3**.

► **Le troisième quartile Q_3** est la plus petite valeur Q_3 de la série telle qu'au moins les trois quarts des valeurs de la liste sont inférieures ou égales à Q_3
Ex : Déterminons le troisième quartile Q_3 de la série statistique utilisée comme exemple.
L'effectif total est 26 et $\frac{3}{4}$ x 26 = 19,5. On cherche **la plus petite valeur de la série** pour laquelle **au moins 20 données lui sont inférieures ou égales**. Donc Q_3 = **7**.

II) Paramètres de dispersion : étendue, écart interquartile, variance et écart-type

définitions :
Soit une série statistique X d'effectif total N, à caractère quantitatif, dont les valeurs du caractère ont été rangées dans l'ordre croissant $x_1 \leqslant x_2 \ldots\ldots \leqslant x_p$.
L' effectif correspondant à chaque valeur est noté n_1, n_2, ... n_p.
La moyenne de la série est notée $\bar{x}$

$N = n_1 + n_2 + \ldots. + n_p$!

► **L' étendue** de la série statistique est la différence entre ses valeurs extrêmes.
Ex : Calculons l'étendue de la série statistique utilisée comme exemple.
L'étendue de la série est 9 – 2 = **7**

► **L' écart interquartile** de la série statistique est la différence $Q_3 - Q_1$ entre le troisième quartile et le premier quartile.
Ex : Calculons l'écart interquartile de la série statistique utilisée comme exemple.
L'écart interquartile est $Q_3 - Q_1$ = 7 – 3 = **4**

► **La variance** de la série statistique est le nombre réel **V** défini par :

$$V = \frac{n_1(x_1 - \bar{x})^2 + n_2(x_2 - \bar{x})^2 + \ldots\ldots\ldots + n_p(x_p - \bar{x})^2}{N}$$

La variance est la moyenne des carrés des «écarts à la moyenne» !!

L' écart-type de la série statistique est le nombre réel **σ** défini par :

$$\sigma = \sqrt{V}$$

Ex : Calculons l' écart-type de la série statistique utilisée comme exemple.

$$V = \frac{3 \times (2-5)^2 + 4 \times (3-5)^2 + 6 \times (4-5)^2 + 2 \times (5-5)^2 + 4 \times (6-5)^2 + 4 \times (7-5)^2 + 1 \times (8-5)^2 + 2 \times (9-5)^2}{26}$$

$\simeq$ 4,23

$\sigma = \sqrt{V} \simeq \sqrt{4,23} \simeq$ 2,06 h

L'unité de l' écart-type est celui du caractère étudié. C'est son principal intérêt ! (la variance, elle, n'a pas d'unité)

propriété :
La variance V de la série statistique peut également se définir par :

$$V = \frac{n_1x_1^2 + n_2x_2^2 + ... + n_p\,x_p^2}{N} - \overline{x}^2$$

V est la moyenne des carrés des valeurs x_i moins le carré de la moyenne $\overline{x}$!
Avec cette expression, il y a moins d'opérations à faire !

► **démonstration**
Par définition de la variance, on a

$$V = \frac{n_1(x_1 - \overline{x})^2 + n_2(x_2 - \overline{x})^2 + ...+ n_p(x_p - \overline{x})^2}{N}$$

$$NV = n_1(x_1 - \overline{x})^2 + n_2(x_2 - \overline{x})^2 + ...+ n_p(x_p - \overline{x})^2$$

$$NV = n_1(x_1^2 - 2x_1\overline{x} + \overline{x}^2) + n_2(x_2^2 - 2x_2\overline{x} + \overline{x}^2) + ...+ n_p(x_p^2 - 2x_p\overline{x} + \overline{x}^2)$$

$$NV = (n_1x_1^2 + n_2x_2^2 +....+ n_px_p^2) + (-2\,n_1x_1\overline{x} - 2\,n_2x_2\overline{x} -...- 2\,n_px_p\overline{x}) +(n_1\overline{x}^2 + n_2\overline{x}^2 + ...+ n_p\overline{x}^2)$$

$$NV = (n_1x_1^2 + n_2x_2^2 +..... n_px_p^2) - 2\overline{x}(n_1x_1+ n_2x_2 +.....+ n_px_p) + \overline{x}^2(n_1 + n_2 + ...+ n_p)$$

Or, $n_1x_1+ n_2x_2 +.....+ n_px_p = N\overline{x}$ (par définition de la moyenne) et $n_1 + n_2 + ...+ n_p = N$

Donc, $NV = (n_1x_1^2 + n_2x_2^2 +..... n_px_p^2) - 2\overline{x}N\overline{x} + \overline{x}^2N = (n_1x_1^2 + n_2x_2^2 +..... n_px_p^2) - \overline{x}^2N$

Par suite, en divisant par N, on obtient $V = \frac{n_1x_1^2 + n_2x_2^2 +..... n_px_p^2}{N} - \overline{x}^2$

Ex : Recalculons avec cette expression la variance de la série statistique utilisée comme exemple.

$$V = \frac{3 \times 2^2 + 4 \times 3^2 + 6 \times 4^2 + 2 \times 5^2 + 4 \times 6^2 + 4 \times 7^2 + 1 \times 8^2 + 2 \times 9^2}{26} - 5^2 \simeq 4{,}23$$

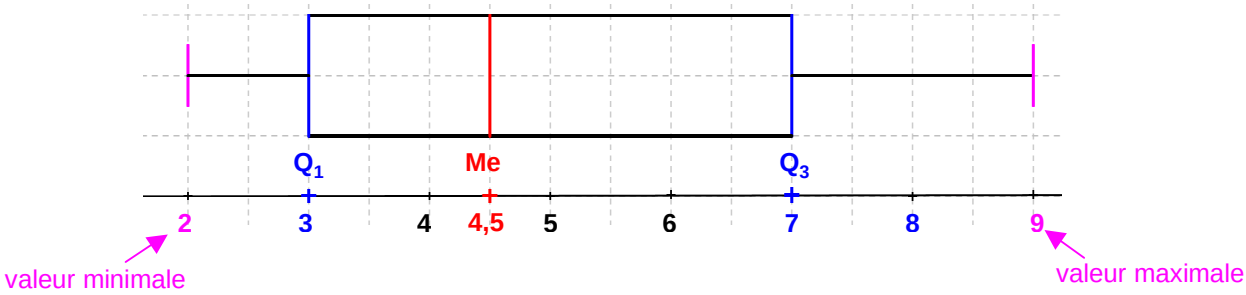
III) Diagramme en boîte - Résumé d'une série statistique

a) diagramme en boîte

La répartition des données peut être représentée par un diagramme en boîte.
Il résume le caractère étudié par les valeurs extrêmes, la médiane, les quartiles.

on l'appelle aussi **diagramme de Tukey** du nom de son créateur ou **diagramme en boîte à moustaches** à cause de sa forme !

Représentons le diagramme en boîte de la série statistique utilisée comme exemple.



b) résumé d'une série statistique

Résumer une série statistique revient à indiquer la répartition des données.
On utilise le plus souvent deux indicateurs :
- un paramètre de tendance centrale (médiane ou moyenne)
- un paramètre de dispersion (écart interquartile ou écart-type)

Deux choix sont fréquemment utilisés :
► Le couple constitué de la moyenne et de l' écart-type ($\overline{x}$; σ). Il est sensible aux valeurs extrêmes.
► Le couple constitué de la médiane et de l'écart interquartile (Me ; $Q_3 - Q_1$) est lui peu sensible aux valeurs extrêmes mais sa détermination est moins facile que le précédent.