STATISTIQUES	
Cours	***Première S***

Dans ce chapitre, on considère des séries à caractères quantitatifs.

Notations :
$x_1, x_2, \ldots, x_p$ sont les valeurs ou les centres des classes si ces valeurs sont regroupées en classe.
$n_1, n_2, \ldots, n_p$ sont les effectifs respectifs des valeurs $x_1, x_2, \ldots, x_p$
N est l'effectif total : $N = n_1 + n_2 + \ldots + n_p$

1. Paramètres de position d'une série statistique

Les paramètres de position (ou indicateurs de tendance centrale) d'une série statistique sont des valeurs numériques qui « résument » la série en caractérisant l'ordre de grandeur des observations. Ils s'expriment dans la même unité que ces dernières.

1) Moyenne

Définition 1 : La **moyenne** d'une série statistique est le nombre, noté $\bar{x}$, défini par

$$\bar{x} = \frac{n_1x_1 + n_2x_2 + \ldots + n_px_p}{N} = \frac{\sum_{i=1}^{p} n_ix_i}{N}$$

Exemple : Dans tout ce chapitre, tous les exemples s'appuient sur la série statistique suivante : on considère une liste de températures (en °C), relevées sous abri, à différents moments d'une journée (de 0 h à 23 h) le 31 octobre 2014 à Nouakchott
(source : http://www.infoclimat.fr/observations-meteo/archives/31/octobre/2014/nouakchott/61442.html).

Les données sont les suivantes :
23 - 22 - 22 - 22,7 - 22 - 22 - 22,6 - 22 - 23 - 25,4 - 25 - 26 - 28,6 - 27 - 27 - 29 - 28 - 27 - 26,7 - 24 - 24 - 23,9 - 23,9 - 23.

$$\bar{x} = \frac{5\times 22 + 22{,}6 + 22{,}7 + 3\times 23 + 2\times 23{,}9 + 2\times 24 + 25 + 25{,}4 + 26 + 26{,}7 + 3\times 27 + 28 + 28{,}6 + 29}{24}$$

$$\bar{x} = \frac{589{,}8}{24} \approx 24{,}6$$

La température moyenne était alors d'environ 24,6 °C le 31 octobre 2014 à Nouakchott.

2) Médiane

Définition 2 : La **médiane** d'une série statistique est un nombre M tel qu'il y ait autant de valeurs du caractère étudié inférieures à M que de valeurs supérieures à M.
La médiane partage ainsi la population en deux parties d'effectifs égaux.

Remarque : Pour déterminer la médiane d'une série statistique, on commence par ranger les valeurs dans l'ordre croissant.

Si N est impair, alors la médiane est la valeur de la série située au rang $\frac{N+1}{2}$.

Si N est pair, alors la médiane est la demi somme des valeurs de la série situées au rang $\frac{N}{2}$ et au rang $\frac{N}{2}+1$.

C. Lainé

Exemple : On range les températures dans l'ordre croissant : 22 - 22 - 22 - 22 - 22 - 22,6 - 22,7 - 23 - 23 - 23 - 23,9 - 23,9 - 24 - 24 - 25 - 25,4 - 26 - 26,7 - 27 - 27 - 27 - 28 - 28,6 - 29.

Comme $N = 24$ est pair, alors la médiane est la demi somme des valeurs de la série situées au rang 12 et au rang 13.
La 12ème valeur est 23,9 et la 13ème valeur est 24.
Donc la médiane de cette série est 23,95.
La température médiane de cette série est 23,95 °C.

3) Quartiles

On considère une série ordonnée par ordre croissant.

Définition 3 : On appelle premier quartile Q_1 d'une série statistique la plus petite valeur de la série telle qu'au moins 25 % des valeurs de celle-ci lui sont inférieures ou égales.

On appelle troisième quartile Q_3 d'une série statistique la plus petite valeur de la série telle qu'au moins 75 % des valeurs de celle-ci lui sont inférieures ou égales.

Méthode de recherche des quartiles :

• Si $\frac{N}{4}$ est un entier, le premier quartile Q_1 est la valeur qui dans cette liste occupe le rang $\frac{N}{4}$ et le troisième quartile Q_3 est la valeur qui dans cette liste occupe le rang $\frac{3N}{4}$.

• Si $\frac{N}{4}$ n'est pas un entier, le premier quartile Q_1 est la valeur qui dans cette liste occupe le rang immédiatement supérieur à $\frac{N}{4}$ et le troisième quartile Q_3 est le valeur qui dans cette liste occupe le rang immédiatement supérieur à $\frac{3N}{4}$.

Remarques :
• Une série admet trois quartiles ; le deuxième, dont on ne fait pas usage en première, est associé à la valeur 50 %.

• De nombreuses calculatrices considèrent les quartiles comme les médianes des deux séries obtenues après avoir partagé la série initiale par sa médiane … ce qui explique les différences constatées. Dans la pratique, ces différences ont peu d'importance vu la taille des séries.

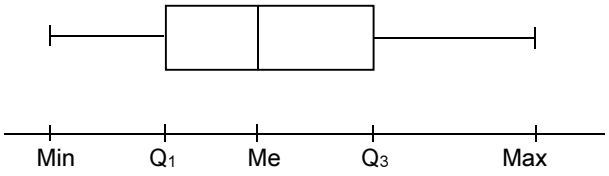
• De la même façon, on peut définir les déciles d'une série statistique.
Le premier décile d_1 d'une série statistique la plus petite valeur de la série telle qu'au moins 10 % des valeurs de celle-ci lui sont inférieures ou égales.
Le neuvième décile d_9 d'une série statistique la plus petite valeur de la série telle qu'au moins 90 % des valeurs de celle-ci lui sont inférieures ou égales.

Exemple : La série contient 24 valeurs. Alors $\frac{N}{4} = 6$.
Le premier quartile Q_1 est la 6ème valeur ; d'où $Q_1 = 22,6$ °C .
Le troisième quartile Q_3 est la 18ème valeur ; d'où $Q_3 = 26,7$ °C .

C. Lainé

2. Diagramme en boîte (ou boîte à moustaches)

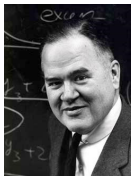
Définition 4 : **Un diagramme en boite** est un rectangle délimité par le premier quartile et le troisième quartile.
Pour l'obtenir, on trace un axe horizontal (ou vertical) sur lequel on place les valeurs de Q_1, Q_3 et Me. L'un des côtés du rectangle a pour longueur l'écart interquartile, l'autre est quelconque. On complète ce diagramme en traçant deux traits horizontaux : l'un joignant Q_1 au minimum de la série et l'autre joignant Q_3 au maximum de la série.



Remarques :
• Lorsque l'effectif est important, ou que les valeurs extrêmes ne sont pas connues, on raccourcit les « moustaches » aux déciles 1 et 9, les valeurs restantes étant indiquées par des points isolés.

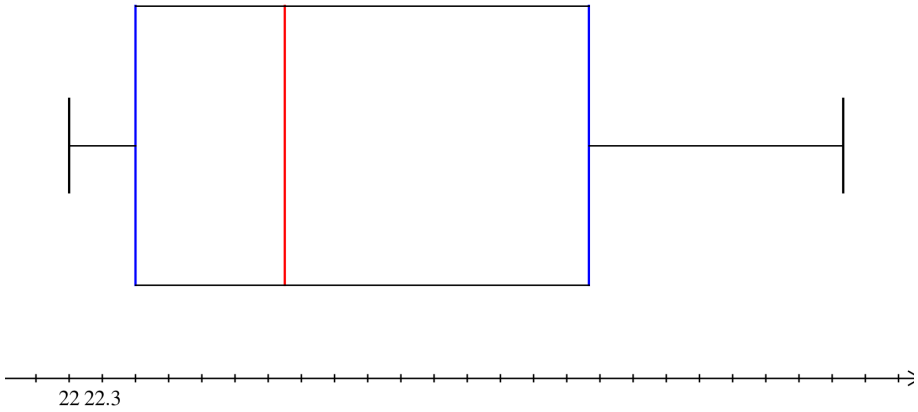
• La largeur du rectangle est quelconque.

• On peut aussi construire ce diagramme verticalement.

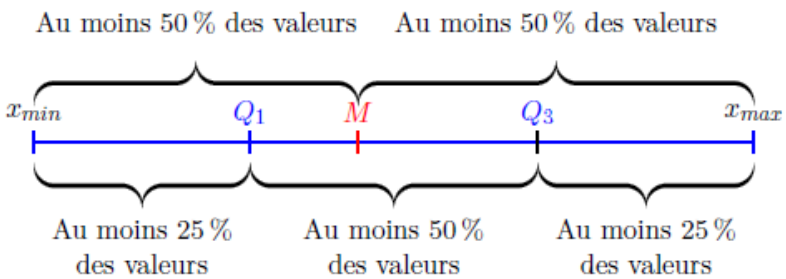


Ce type diagramme porte également le nom de *boîte à moustaches* ou *diagramme de Tukey*.
John Wilder Tukey (1915 – 2000) était un statisticien américain.

Exemple :



Résumé :



C. Lainé

3. Paramètres de dispersion

Les paramètres de dispersion d'une série statistique sont des valeurs numériques qui permettent de « mesurer » la répartition des valeurs d'une série statistique.

1) L'écart interquartile

Définition 54 : **L'écart interquartile** d'une série statistique est le nombre $Q_3 - Q_1$.

Exemple : D'après le **2.**, l'écart interquartile est égale à $Q_3 - Q_1 = 26,7 - 22,6 = 4,1$ °C .

Remarques :
• L'écart interquartile mesure la dispersion des valeurs autour de la médiane ; plus l'écart est petit, plus les valeurs de la série appartenant à l'intervalle interquartile sont concentrées autour de la médiane.

• Contrairement à **l'étendue** (notée e) qui mesure l'écart entre la plus grande et la plus petite valeur, l'écart interquartile élimine les valeurs extrêmes qui peuvent être douteuses, cependant il ne tient compte que de 50% de l'effectif …

On peut correctement résumer une série statistique par le couple : **(médiane ; intervalle interquartile)**

3) Variance et écart-type

a) Variance d'une série statistique

Définition 6 : La variance V d'une série statistique de valeurs $(x_1 , x_2 , ..., x_p)$ ayant pour effectifs respectifs $(n_1 , n_2 , ..., n_p)$ est le nombre défini par : $V = \dfrac{\sum_{i=1}^{p} n_i \times \left(x_i - \overline{x}\right)^2}{N}$.

Remarques :
• V est la moyenne des carrées des écarts des valeurs x_i à la moyenne $\overline{x}$. Elle permet de mesurer la dispersion des valeurs autour de la moyenne.

• Une variance est toujours un réel positif.

Exemple : Dans l'exemple précédent,

$$V = \frac{5\times(22-24,6)^2 + (22,6-24,6)^2 + (22,7-24,6)^2 + 3\times(23-24,6)^2 + 2\times(23,9-24,6)^2}{24}$$
$$\frac{+2\times(24-24,6)^2 + (25-24,6)^2 + (25,4-24,6)^2 + (26-24,6)^2 + (26,7-24,6)^2 + 3\times(27-24,6)^2}{24}$$
$$\frac{+(28-24,6)^2 + (28,6-24,6)^2 + (29-24,6)^2}{24} = \frac{122,16}{24} = 5,09 .$$

b) Autre expression de la variance

Propriété : $V = \dfrac{\sum_{i=1}^{p} n_i \ x_i^2}{N} - \overline{x}^2$

Démonstration : On a $V = \dfrac{1}{N}\left[\left(n_1 x_1^2 - 2n_1 x_1 \overline{x} + n_1 \overline{x}^2\right) + ... + \left(n_p x_p^2 - 2n_p x_p \overline{x} + n_p \overline{x}^2\right)\right]$

C. Lainé

D'où : $V = \frac{1}{N}\sum_{i=1}^{p} n_i x_i^2 - \frac{1}{N}\sum_{i=1}^{p} 2n_i x_i \bar{x} + \frac{1}{N}\sum_{i=1}^{p} n_i \bar{x}^2 = \frac{1}{N}\sum_{i=1}^{p} n_i x_i^2 - \frac{2\bar{x}}{N}\sum_{i=1}^{p} n_i x_i + \frac{\bar{x}^2}{N}\sum_{i=1}^{p} n_i$

Or $\frac{1}{N}\sum_{i=1}^{p} n_i x_i = \bar{x}$ et $\frac{1}{N}\sum_{i=1}^{p} n_i = \frac{N}{N} = 1$.

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

c) Écart type

Définition 7 : L'écart type s est la racine carrée de la variance : $s = \sqrt{V}$.

Exemple : Dans l'exemple précédent, $s = \sqrt{\frac{122,16}{24}} \approx 2,26$.

Remarques :

- L'écart type est un paramètre plus fin que l'étendue, car il tient compte de la répartition des valeurs.
- L'écart type à la même unité que les valeurs de la série étudiée.
- L'écart type mesure la dispersion des valeurs de la série autour de la moyenne. Plus l'écart type est petit, plus les valeurs de la série sont concentrées autour de la moyenne.

On peut correctement résumer une série statistique par le couple : **(moyenne ; écart type)**.