# STATISTIQUES

## Rappels de vocabulaire.

**Définitions :**

La population est l'ensemble sur lequel porte l'observation : on étudie un caractère bien précisé sur les individus de cette population. On collecte des données. Un échantillon est une partie de la population. La liste des valeurs prises par le caractère constitue la série statistique.

**Exemple :** Une entreprise fabrique des T-shirts. Elle veut étudier les commandes de ses clients et plus particulièrement la couleur et la taille des T-shirts commandés.

***Population :*** ensemble des T-shirt commandés
***Individu :*** un T-shirt
***Caractère :*** couleur ( valeur : blanc, jaune, rouge, bleu, vert, noir)
taille (valeur : 36, 38, 40, 42, 44, 46, 48)

**Définitions :**

Un caractère est dit quantitatif lorsqu'on peut les mesurer en associant un nombre à chaque individu, sinon il est qualitatif.
Un caractère quantitatif est discret lorsqu'il ne prend que des valeurs isolées. Il est quantitatif continu lorsqu'il peut prendre toutes les valeurs d'un intervalle.

**Exemple :** T-shirt : la couleur : qualitatif
la taille : quantitatif discret
La taille des élèves de la classe est un caractère quantitatif continu.

**Remarques :** Souvent les modalités d'un caractère qualitatif sont codés (exemple : 1 pour janvier, 2 pour février…). Cela n'en fait pas un caractère quantitatif car on ne peut pas faire des opérations (1 + 2 n'a pas de sens).
Il arrive qu'un caractère quantitatif continu soit rendu discret par exemple en remplaçant la taille par son arrondi au centimètre.

**Définitions :**

L'effectif d'une valeur d'un caractère est le nombre d'individu de la population ayant cette valeur.
La fréquence de la valeur est $\frac{\text{effectif de la valeur}}{\text{effectif total}}$.

**Remarque :** La somme des fréquences est toujours égal à 1.

Une fréquence est toujours comprise entre 0 et 1.
On peut également l'exprimer en pourcentage.

Le mode pour un caractère discret est la valeur qui correspond au plus grand effectif.

La classe modale pour un caractère quantitatif continu est la classe qui correspond au plus grand effectif.

L'étendue de la série est la différence entre les valeurs extrêmes. e = $x_{max} - x_{min}$ .

# I Autour de la médiane.

## 1) Rappels

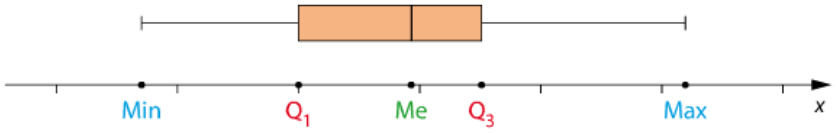
La médiane est une « valeur centrale » d'une série. On lui associe un indicateur de dispersion des valeurs de la série : l'écart interquartile.

**Définitions**

- La **médiane** Me d'une série statistique de $n$ valeurs **ordonnées** est :
– si $n$ est impair, la valeur « du milieu », de rang $\frac{n+1}{2}$.
– si $n$ est pair, la demi-somme des deux valeurs « du milieu », de rangs $\frac{n}{2}$ et $\frac{n}{2}+1$.
- **L'écart interquartile** est la longueur $Q_3 - Q_1$ de **l'intervalle interquartile** $[Q_1 ; Q_3]$ où :
– le **premier quartile $Q_1$** est la plus petite valeur de la série telle qu'au moins **25 %** des valeurs de la série lui soient inférieures ou égales ;
– le **troisième quartile $Q_3$** est la plus petite valeur de la série telle qu'au moins **75 %** des valeurs de la série lui soient inférieures ou égales.

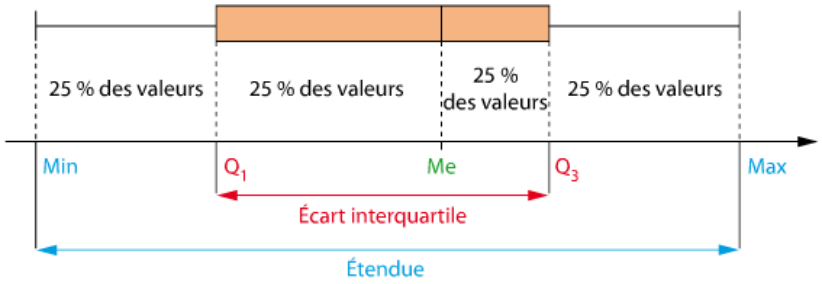
## 2) diagramme en boîte.

Les nombres $Q_1$, Me et $Q_3$, et les valeurs extrêmes de la série, notées Min et Max, offrent un résumé d'une série statistique et une représentation graphique par un diagramme en boîte (*boxplot*).




Construit au-dessus d'un axe gradué, il est constitué :
– d'une boîte délimitée par les 1er et 3e quartiles et partagée par la médiane.
Sa hauteur est souvent arbitraire (elle peut aussi dépendre de l'effectif).
– de deux traits qui relient les quartiles aux valeurs extrêmes de la série.

Il permet de visualiser de nombreux renseignements :




La superposition de diagrammes en boîte se révèle pertinente pour comparer plusieurs séries étudiant le même caractère sur des populations différentes.

**METHODES SUR LES MEDIANES et QUARTILES**

Le but est de découper la liste en deux sous-listes de même effectif et ensuite en quatre sous-listes de même effectif.

**Le cas où la liste possède un nombre pair d'éléments**

On considère la série suivante :

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 45 | 34 | 6 | 15 | 25 | 67 | 38 | 87 | 27 | 14 |

**Méthode :**

- Classer la série dans l'ordre croissant :

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 6 | 14 | 15 | 25 | 27 | 34 | 38 | 45 | 67 | 87 |

**Recherche de la médiane.**

Il y a 10 valeurs.

La médiane sépare les 10 valeurs en deux sous-listes de 5 valeurs. La médiane est donc une valeur entre la 5 ème et la $6^{ème}$ valeur. La médiane est donc ente 27 et 34. On prend (27+34)/2 = 30,5.

**La médiane vaut 30,5**

**Recherche du premier quartile Q1.**

On calcule sa position : 25% de 10 = 0.25*10 = 2,5. On prend la troisième valeur de la liste, soit 15.

**Q1 = 15**

**Recherche du troisième quartile Q3.**

On calcule sa position : 75% de 10 = 0.75*10 = 7,5. On prend la huitième valeur de la liste, soit 45.

**Q3=45**

**<u>Le cas où la liste possède un nombre impair d'éléments</u>**

On considère la série suivante :

| | | | | | | |
|---|---|---|---|---|---|---|
| 45 | 34 | 6 | 15 | 25 | 67 | 38 |

**Méthode :**

- Classer la série dans l'ordre croissant :

| | | | | | | |
|---|---|---|---|---|---|---|
| 6 | 15 | 25 | 34 | 38 | 45 | 67 |

**Recherche de la médiane.**

Il y a 7 valeurs.

La médiane sépare les 7 valeurs en deux sous-listes de 3 valeurs. La médiane est donc la quatrième valeur.

**La médiane vaut 34**

**Recherche du premier quartile Q1.**

On calcule sa position : 25% de 7 = 0.25*7 = 1,75. On prend la deuxième valeur de la liste, soit 15.

**Q1 = 15**

**Recherche du troisième quartile Q3.**

On calcule sa position : 75% de 7 = 0.75*7 = 5,25. On prend la sixième valeur de la liste, soit 45.

**Q3=45**

***ATTENTION : les calculatrices donnent ces valeurs, mais il peut exister quelques variantes dans les définitions.***

# II Autour de la moyenne.

## 1) La moyenne.

**Définition**

Soit $(x_k ; n_k)$ où $1 \leqslant k \leqslant r$ une série statistique dont les valeurs distinctes $x_1, x_2, \ldots, x_r$ ont pour effectifs $n_1, n_2, \ldots, n_r$ et pour fréquences $f_1, f_2, \ldots, f_r$.
Son effectif total est $N = n_1 + n_2 + \ldots + n_r$.

| Valeurs | $x_1$ | $x_2$ | … |
|---|---|---|---|
| Effectifs | $n_1$ | $n_2$ | … |
| Fréquences | $f_1$ | $f_2$ | … |

La **moyenne** de la série statistique $(x_k ; n_k)$ où $1 \leqslant k \leqslant r$ est le nombre $\bar{x}$ (ou $m$) :

$$\bar{x} = \frac{1}{N}(n_1 \times x_1 + n_2 \times x_2 + \ldots + n_r \times x_r) = f_1 \times x_1 + f_2 \times x_2 + \ldots + f_r \times x_r$$

**Propriété 1**

Une série statistique partagée en $p$ sous-séries disjointes de moyennes et d'effectifs respectifs $(\overline{x_1}, n_1), (\overline{x_2}, n_2), \ldots, (\overline{x_p}, n_p)$ a pour moyenne :

$$\bar{x} = \frac{n_1 \times \overline{x_1} + n_2 \overline{x_2} + \ldots + n_p \times \overline{x_p}}{n_1 + n_2 + \ldots n_p}$$

**Exemple**

Une entreprise emploie 150 ouvriers de salaire moyen 1 800 €, 12 cadres de salaire moyen 2 900 € et 3 cadres supérieurs de salaire moyen 4 600 €.

Le salaire moyen dans l'entreprise est $\frac{150 \times 1\,800 + 12 \times 2\,900 + 3 \times 4\,600}{150 + 12 + 3} \approx 1\,931$ €.

## 2) Variance et écart type

**Définition**

- Soit la série statistique $(x_k ; n_k)$ où $1 \leqslant k \leqslant r$, d'effectif total N et de moyenne $\bar{x}$.

Le réel $V = \frac{1}{N}[n_1(x_1 - \bar{x})^2 + \ldots + n_r(x_r - \bar{x})^2] = f_1(x_1 - \bar{x})^2 + \ldots + f_r(x_r - \bar{x})^2$
est appelé **variance** de la série statistique.

- La racine carrée de la variance, $s = \sqrt{V}$ est appelée **écart type** de la série.

La variance et l'écart type permettent de mesurer la dispersion des valeurs par rapport à la valeur moyenne.
Lorsque les valeurs de la série sont des mesures d'une grandeur dans une unité donnée, par exemple, de longueur en mètres, l'écart type exprime la dispersion dans la même unité.

**D'une manière générale on utilise les formules suivantes :**

Cas n°1 : la population est donnée par la liste de ses n éléments $x_i$.
Cas n°2 : la population est donnée par le tableau des effectifs $n_i$ des $p$ classes $x_i$.
Cas n°3 : la population est donnée par une liste d'intervalles. On utilise alors le centre des classes.
Cas n°4 : La population est donnée par une liste de fréquences.

| | |
|---|---|
| **Cas n°1 Des valeurs isolées** | $V = \frac{(x_1 - \bar{x})^2 + \cdots \ldots + (x_n - \bar{x})^2}{n} = \frac{{x_1}^2 + {x_2}^2 + \cdots . + {x_n}^2}{n} - \bar{x}^2$ |
| **Cas n°2 Des valeurs** | $V = \frac{n_1(x_1 - \bar{x})^2 + \cdots \ldots + n_p(x_p - \bar{x})^2}{n} = \frac{n_1{x_1}^2 + n_2{x_2}^2 + \cdots . + n_p{x_p}^2}{n} - \bar{x}^2$ |

| | |
|---|---|
| **en paquets** | |
| **Cas n°3 Des lasses de valeurs** | $V = \frac{n_1(c_1 - \bar{x})^2 + \cdots \ldots . + n_p(c_p - \bar{x})^2}{n} = \frac{n_1 {c_1}^2 + n_2 {c_2}^2 + \cdots . + n_p {c_p}^2}{n} - \bar{x}^2$<br>Avec $c_i$ le centre des classes. |
| **Cas n°4 Avec des fréquences.** | $V = f_1(x_1 - \bar{x})^2 + \cdots \ldots . + f_p(x_p - \bar{x})^2 = f_1 {x_1}^2 + f_2 {x_2}^2 + \cdots . + f_p {x_p}^2 - \bar{x}^2$ |

## III Choisir un résumé d'une série statistique

Si l'on souhaite retenir un seul nombre pour résumer une série statistique, on choisit un indicateur de position : la médiane ou la moyenne.

Si on souhaite aussi rendre compte de la répartition des valeurs autour de cette « valeur centrale », on lui associe un indicateur de dispersion : l'écart interquartile ou l'écart type.

On obtient ainsi deux couples (Me, $Q_3 - Q_1$) et $(\overline{x}, s)$ qui sont deux résumés d'une série statistique. Ils ne prétendent pas restituer toute l'information de la série statistique mais ils permettent d'en synthétiser l'essentiel et de faciliter la comparaison de plusieurs séries.

### 1) Le couple (médiane, écart interquartile)

Ce couple donne à la fois :
– un indicateur de la tendance centrale de la série : la médiane ;
– un indicateur de dispersion : la longueur de l'intervalle interquartile qui contient la moitié centrale des valeurs de la série.
Plus l'écart interquartile $Q_3 - Q_1$ est petit, plus les valeurs centrales de la série se concentrent autour de la médiane Me.

### 2) Le couple (moyenne, écart type)

Ce couple donne à la fois :
– un indicateur de la tendance centrale de la série : la moyenne ;
– un indicateur de dispersion faisant intervenir les carrés des écarts à la moyenne de toutes les valeurs de la série.
Plus l'écart type $s$ est petit, plus les valeurs se concentrent autour de la moyenne.

## 3) Choix du couple

• **Le couple (médiane, écart interquartile)**
Il est assez facile à interpréter.
Il a l'avantage d'être très peu sensible aux valeurs extrêmes (parfois suspectes).
En revanche, et contrairement à la moyenne, il ne se calcule pas par paquets : connaissant la médiane de sous-séries disjointes constituant une série, on ne peut pas en déduire la médiane de cette série.

• **Le couple (moyenne, écart type)**
Il joue un grand rôle en statistique théorique (par exemple appliquée aux sondages).
Les définitions algébriques de la moyenne et de l'écart type se prêtent à des calculs algébriques permettant d'établir différentes propriétés.
Cependant l'écart type tient compte des écarts de toutes les valeurs à la moyenne. Il donne ainsi beaucoup de poids aux valeurs extrêmes, et son choix n'est pertinent que lorsque le diagramme qui représente la série est assez symétrique et évoque la forme d'une « courbe en cloche » comme ci-contre.