# STATISTIQUES

## 1. Vocabulaire et notations

Dans ce chapitre, on considère des séries à caractères quantitatifs discrètes (sous forme « ponctuelle ») ou continues (sous forme d'intervalles) (avec, dans le cas d'une série continue, l'hypothèse d'une répartition uniforme à l'intérieur de chaque classe).

On appelle population un ensemble de personnes ou d'objets étudiés. Un individu est un élément de la population. La particularité de la population étudiée lors d'une étude statistique est appelée caractère ou variable.

La valeur du caractère étudié se note $x_i$. L'effectif correspondant à cette valeur se note $n_i$. L'effectif total se note $N$. Et l'on a :

$$\sum_{i=1}^{p} n_i = n_1 + n_2 + ... + n_p = N \ .$$

Attention : Lorsque les caractères sont définis sous forme d'intervalles (*ie* sous forme de classes), les calculs sont réalisés avec le centre des intervalles.

La fréquence $f_i$ de la valeur $x_i$ est définie par : $f_i = \frac{n_i}{N}$ . On a :

$$\sum_{i=1}^{p} f_i = f_1 + f_2 + ... f_p = 1.$$

Remarque : la fréquence est un réel compris entre 0 et 1, mais il peut aussi s'exprimer sous forme de pourcentage.

## 2. Paramètres de position

### 2.1. Caractéristiques de position de tendance centrale

Ils permettent de dégager la tendance centrale de la série statistique étudiée.

**Définition 1** : La moyenne d'une série statistique se note $\overline{x}$ et vaut :

$$\overline{x} = \frac{n_1 x_1 + n_2 x_2 + ... + n_p x_p}{N} = \frac{1}{N} \sum_{i=1}^{p} n_i x_i \ .$$

On l'appelle moyenne pondérée.

Remarques :

i. La moyenne est le barycentre des nombres pondérés $(x_1;n_1)$, $(x_2;n_2)$, …, $(x_p;n_p)$.

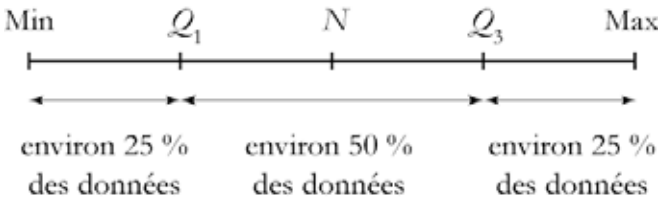
ii. Puisque la moyenne est un barycentre, on peut utiliser la propriété d'associativité du barycentre pour calculer la moyenne d'une série à partir des moyennes de séries partielles dont on connaît les effectifs.

**Définition 2** : Le mode est la valeur (ou les valeurs) du caractère de plus grand effectif. Pour un caractère continu, on parle de classe modale . Si les classes ont la même amplitude, la classe modale est la classe qui correspond au plus fort effectif.

**Définition 3** : La médiane est une valeur *M* du caractère qui partage la population ordonnée en deux sous-ensembles de même effectif.

## 2.1. Caractéristiques de position non centrale : les quartiles

**Définition 4** : Les quartiles sont les trois valeurs du caractère qui partagent les valeurs ordonnées du caractère en quatre sous-ensembles de même effectif. Le premier quartile, noté $Q_1$, est la plus petite valeur de la série statistique telle qu'au moins 25 % des valeurs de celle-ci lui sont inférieures ou égales ; le deuxième est la médiane *M* et le troisième, noté $Q_3$, est la plus petite valeur de la série statistique telle qu'au moins 75 % des valeurs de celle-ci lui sont inférieures ou égales. On peut représenter cette disposition de la manière suivante :




Remarques :

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

ii. 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.

iii. De la même façon, on peut définir des déciles d'une série statistique.

Exemples de détermination :

i. Dans le cas d'une série discrète, 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 la valeur qui dans cette liste occupe le rang immédiatement supérieur à $\frac{3N}{4}$. Par exemple : Le tableau suivant donne la répartition des notes de 31 élèves.

| Notes | 5 | 8 | 9 | 10 | 11 | 12 | 14 | 16 | 18 |
|---|---|---|---|---|---|---|---|---|---|
| Effectif | 1 | 2 | 6 | 7 | 5 | 4 | 3 | 2 | 1 |
| ECC | 1 | 3 | 9 | 16 | 21 | 25 | 28 | 30 | 31 |

$N = 31$ ; $\frac{N}{4} = 7,75$ donc $Q_1 = 9$ ; $\frac{N}{2} = 15,5$ donc $M = 10$ et $\frac{3N}{4} = 23,25$ donc $Q_3 = 12$. Le mode est 10 et la moyenne vaut :

$$\bar{x} = \frac{1\times 5 + 2\times 8 + ... + 1\times 18}{31} = \frac{340}{31} \simeq 11 .$$

ii. Dans le cas d'une série continue (valeurs regroupées par classe), le premier quartile $Q_1$ est la valeur correspondant à la fréquence cumulée croissante égale à 0,25. De la même manière, $Q_3$ correspond à la fréquence cumulée croisssante égale à 0,75. Par exemple : Une enquête est effectuée pour étudier le temps (en minutes) consacré au sport, chaque semaine, par les 1312 employés d'une usine. Les résultats, regroupés en classes, sont indiqués dans le tableau suivant :

| Temps (min) | Effectifs | Fréquence (%) | FCC (%) |
|---|---|---|---|
| [0 ; 30[ | 175 | 13 | 13 |
| [30 ; 60[ | 392 | 30 | 43 |
| [60 ; 90[ | 267 | 21 | 64 |
| [90 ; 120[ | 127 | 9 | 73 |
| [120 ; 150[ | 168 | 13 | 86 |
| [150 ; 180[ | 120 | 9 | 95 |
| [180 ; 240[ | 63 | 5 | 100 |

On place, dans le repère orthogonal d'échelle 1 cm pour 30 minutes en abscisses et 1 cm pour 10 % en ordonnées, les points de coordonnées (0 ; 0), (30 ; 13), …, (240 , 100), puis on les relie. On trace les droites d'équations $y = 50$ pour trouver $M \approx 70$, $y = 25$ pour trouver $Q_1 \approx 40$ et $y = 75$ pour trouver $Q_3 \approx 120$.

## 3. Paramètres de dispersion

Ces paramètres permettent de mesurer l'étalement (*ie* la dispersion) de la série statistique autour de sa tendance centrale.

### 3.1. Ecart interquartile

**Définition 5 :** L'intervalle $[Q_1 ; Q_3]$ est appelé intervalle interquartile.

**Définition 6** : Le réel $Q_3 - Q_1$ est appelé écart interquartile.

**Définition 7** : On appelle étendue la différence entre les deux valeurs extrêmes prises par le caractère étudié.

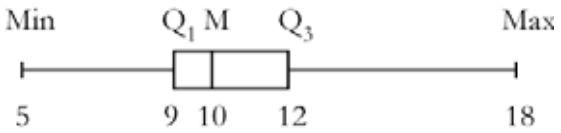
Remarques :
i. L'écart interquartile mesure la dispersion des valeurs autour de la médiane ; plus l'écart est petit, plus les valeurs de la série appartenant à l'intervalle interquartile sont concentrées autour de la médiane.
ii. Contrairement à l'étendue qui mesure l'écart entre la plus grande et la 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.
iii. On peut correctement résumer une série statistique par le couple (médiane ; intervalle interquartile) car il est peu sensible aux variations des valeurs extrêmes de la série, contrairement au couple (moyenne ; écart type) vu plus tard…

Exemple : Suite du i. précédent… L'intervalle interquartile est [9 ; 12] ; l'écart interquartile est 12 – 9 = 3.

**Définition 8** : Le diagramme en boîtes (ou à pattes ou boîte à moustaches ou diagramme de Tukey) permet de visualiser l'étendue, la médiane et les quartiles d'une série statistique.

Méthode : Pour l'obtenir, on trace un axe horizontal (ou vertical) sur lequel on place les valeurs de $Q_1$, $M$ et $Q_3$. 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 (ou verticaux) : l'un joignant $Q_1$ au minimum de la série et l'autre joignant $Q_3$ au maximum de la série.

Exemple : Suite de l'exemple 1 de paragraphe 1.2.




### 3.2. Variance et écart type

**Définition 9** : On appelle variance, notée $V(x)$, la moyenne des carrés des écarts entre chaque valeur $x_i$ et la moyenne $\bar{x}$ :

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

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

Remarque : une autre formule de la variance est :

$$V(x) = \frac{n_1 x_1^{\,2} + n_2 x_2^{\,2} + \ldots + n_p x_p^{\,2}}{N} - \bar{x}^2$$
$$= \frac{1}{N}\sum_{i=1}^{p} n_i x_i^{\,2} - \bar{x}^2.$$

Preuve :

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

**Définition 10** : On appelle écart type, noté $s$, la racine carrée de la variance.
$s = \sqrt{V(x)}$

Remarques :
i. L'écart type est un paramètre plus fin que l'étendue, car il tient compte de la répartition des valeurs.
ii. L'écart type est exprimé dans la même unité que la variable.
iii. L'écart type mesure la dispersion des valeurs autour de la moyenne. Plus la variance est grande, plus les valeurs du caractère étudié sont dispersées autour de la moyenne.
iv. On peut correctement résumer une série statistique par le couple (moyenne ; écart type).

Signification : On admet que pour des séries statistiques de caractère continu et relativement symétriques : au moins 68 % des valeurs étudiées se situent dans l'intervalle $\left[\bar{x} - s; \bar{x} + s\right]$ ; au moins 75 % des valeurs étudiées se situent dans l'intervalle $\left[\bar{x} - 2s; \bar{x} + 2s\right]$ ; au moins 88 % des valeurs étudiées se situent dans l'intervalle $\left[\bar{x} - 3s; \bar{x} + 3s\right]$. Bien entendu, si l'écart type est faible, les intervalles précédents sont de faibles amplitudes. Dans ces conditions, la moyenne est fiable et fournit un bon renseignement sur le caractère étudié.

Exemple : On a relevé les salaires mensuels, en euros, dans une entreprise.

| Salaire | [800 ; 1000[ | [1000 ; 1200[ | [1200 ; 1500[ | [1500 ; 2000[ |
|---|---|---|---|---|
| Effectif | 20 | 15 | 10 | 5 |

D'où $\bar{x} = \frac{56750}{50} = 1135$ et $V(x) = \frac{20 \times 900^2 + \ldots + 5 \times 1750^2}{50} - 1135^2$

On en déduit $V(x) = 69525$ et $s = \sqrt{69525} \simeq 264$. On peut ainsi dire qu'au moins 68 % des salaires sont compris entre 871 € et 1399 €.

On a choisit de calculer la moyenne des carrés des écarts par rapport à la moyenne. Ce choix se justifie par le résultat suivant :

**Théorème 1** : Soit la série statistique ($x_1$, $x_2$, …, $x_p$) de moyenne $\bar{x}$.

$x$ étant une variable réelle, la somme définie par $f(x) = \frac{1}{N}\sum_{i=1}^{p} n_i (x_i - x)^2$ est minimale lorsque $x$ prend la valeur $\bar{x}$. Ce minimum est égal à $V$.

Preuve :

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

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

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

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

$f(x)$ est de la forme $ax^2 + bx + c$ avec $a = 1$, $b = -2\bar{x}$ et $c = \frac{1}{N}\sum_{i=1}^{p} n_i {x_i}^2$. $f(x)$ admet son minimum pour $x = -\frac{b}{2a}$.

Or $-\frac{b}{2a} = -\frac{2\bar{x}}{2} = \bar{x}$. De plus :

$$f(\bar{x}) = \bar{x}^2 - 2\bar{x}\bar{x} + \frac{1}{N}\sum_{i=1}^{n} n_i {x_i}^2 = \frac{1}{N}\sum_{i=1}^{n} n_i {x_i}^2 - \bar{x}^2 = V(x).$$

## 4. Influence d'une transformation affine de données

Parfois, une série statistique est modifiée par un changement d'échelle. Par exemple, des prix peuvent être convertis d'euros en dollars, des températures de degrés Celsius en degrés Fahrenheit.

On appelle image d'une série statistique $S$ de valeurs $\left(x_1, x_2, ..., x_p\right)$ ayant pour effectifs respectifs $\left(n_1, n_2, ..., n_p\right)$ par une fonction $f$, la série statistique $S$' de valeurs $\left(f(x_1), f(x_2), ..., f(x_p)\right)$ ayant pour effectifs respectifs $\left(n_1, n_2, ..., n_p\right)$.

**Théorème 2 :** Soit S une série statistique, $f : x \mapsto ax + b$ (avec $a \neq 0$) une application affine et $S$' l'image de $S$ par $f$. On a, avec les notations usuelles :

**i.** $\bar{x}' = a\bar{x} + b$ ;

**ii.** Si $a > 0$, $Q_3' - Q_1' = a\left(Q_3 - Q_1\right)$ ;

**iii.** $V' = a^2 V$ ;

**iv.** $s' = |a| s$ .

Preuve :

i. $\bar{x}' = \frac{1}{N}\sum_{i=1}^{p} n_i \left(ax_i + b\right) = a \times \frac{1}{N}\sum_{i=1}^{p} n_i x_i + b \times \frac{1}{N}\sum_{i=1}^{p} n_i$

$= a\bar{x} + b$ .

ii. Si $a > 0$, $Q_1' = aQ_1 + b$ et $Q_3' = aQ_3 + b$ . On a alors

$Q_3' - Q_1' = aQ_3 + b - aQ_1 - b = a\left(Q_3 - Q_1\right)$.

iii. $V' = \frac{1}{N}\sum_{i=1}^{p} n_i \left(x_i' - \bar{x}'\right)^2 = \frac{1}{N}\sum_{i=1}^{p} n_i \left(ax_i + b - a\bar{x} - b\right)^2$

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

iv. $s' = \sqrt{a^2 V} = |a| s$ .

Remarque : Si $f : x \mapsto x + b$, alors la variance, l'écart type et l'écart interquartile sont inchangés.