# Statistique Descriptive

4ème Chapitre : Les caractéristiques de dispersion (étendue, quantiles, variance et écart type).

# Introduction

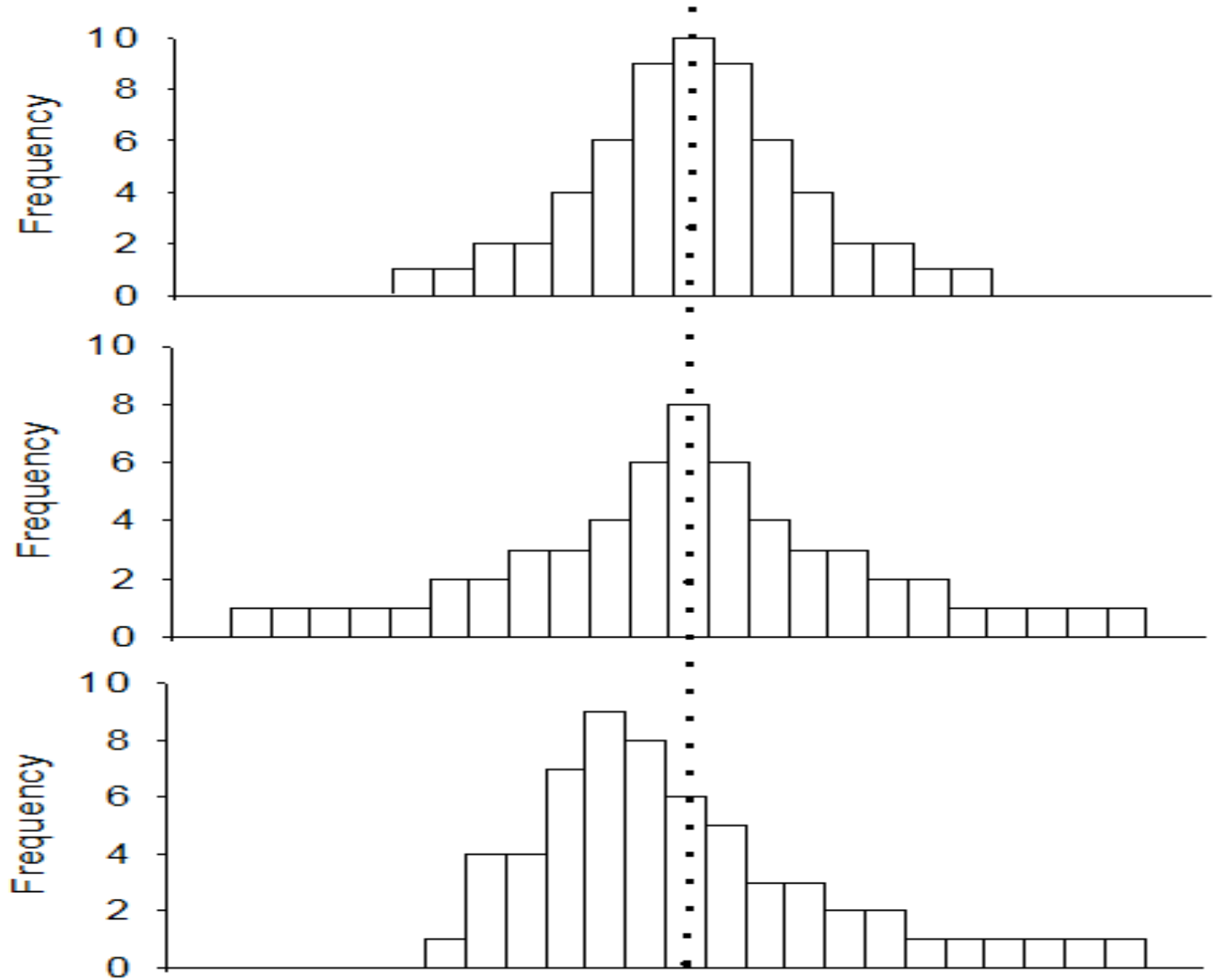
# Introduction

## QUEL EST L'INTERÊT DES PARAMÈTRES DE POSITION ?

**Les paramètres de position** (ou valeurs centrales) sont des valeurs numériques qui « résument » une série statistique en caractérisant l'ordre de grandeur des observations. Ils permettent de situer la position de plusieurs séries comparables.

Cependant, comme le montre le schéma qui suit, ces paramètres ne suffisent pas pour « résumer », pour « décrire » (de façon synthétique) une distribution. En effet, ces paramètres permettent de **situer** la gamme de valeur où la série se situe, mais, pour des paramètres de position très proches, on peut rencontrer des courbes dont la dispersion (l'étalement) est très différente. C'est ici qu'interviennent les paramètres de dispersion.

# Introduction



# Introduction

Une information supplémentaire à la tendance centrale est alors nécessaire pour pouvoir distinguer entre ces différentes formes de distribution.

**Les Mesures de dispersion** : Ces paramètres permettent de mesurer l'étalement de la série statistique autour de sa tendance centrale, et de comparer les étendues des distributions entre elles.

Une autre signification de la mesure de dispersion est l'information qu'elle fournit visant à préciser la position relative d'une observation par rapport aux autres.

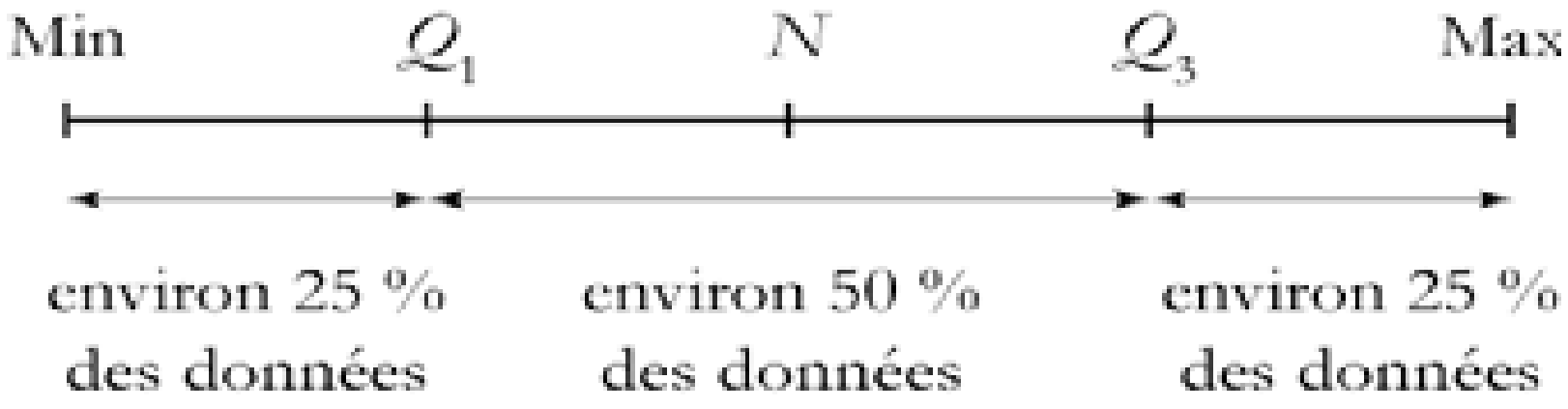
La dispersion d'une série statistique peut être mesurée par les fluctuations des valeurs de la série autour de la moyenne, c'est-à-dire par les différences $x_i - \overline{x}$ .

# L'Écart interquartile & L'Étendue

# L'Écart interquartile & L'Étendue

Rappelons que :

- **Le premier quartile** d'une série statistique est la plus petite valeur $Q_1$ telle qu'au moins 25% des valeurs sont inférieures ou égales à $Q_1$.
- **Le deuxième quartile** d'une série statistique c'est la **médiane**.
- **Le troisième quartile** d'une série statistique est la plus petite valeur $Q_3$ telle qu'au moins 75% des valeurs sont inférieures ou égales à $Q_3$.



# L'Écart interquartile & L'ÉTENDUE

## Définitions

- L'intervalle $[Q_1; Q_3]$ est appelé intervalle interquartile.
- Le réel $Q_3 - Q_1$ est appelé écart interquartile.
- On appelle étendue d'une série statistique la **différence** entre la plus grande valeur de la série et la plus petite :

$$E = V_{\text{max}} - V_{\text{min}}.$$

**Remarque :** La connaissance de l'étendue permet de mieux cerner la dispersion autour des valeurs de position. Ainsi, une étendue élevée par rapport à la moyenne arithmétique renseigne sur une importante dispersion.

# L'Écart interquartile & L'ÉTENDUE

**Remarques :**

1- 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.

2- 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.

# L'Écart interquartile & L'ÉTENDUE

## Exemple 1 : Cas Discret

Le tableau suivant donne la répartition des notes de 31 élèves.

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

On a $N = 31 \Longrightarrow N/2 = \mathbf{15.5}$

Donc Médiane = Me = 10

On a $N/4 = 7.75$ et $3N/4 = 23.25$

Donc $Q_1 = 9$ et $Q_3 = 12$ Intervalle interquartile = [9 , 12]

L'écart interquartile = IQ = 12- 9 = 3

L'**étendue** de la série = E = 18-5 =13

**Moyenne arithmétique** simple $= \overline{x} = \frac{1\times5+2\times8+6\times9+\cdots+1\times18}{31} = 11$

Le **mode** de la série = Mo = 10 (c'est la modalité ayant l'effectif le plus élevé)

# L'Écart interquartile & L'ÉTENDUE

## Exemple 2 : Variable Continue

Une enquête est effectuée pour étudier le temps (en minutes) consacré au sport, semaine, par les 1312 employés d'une usine. Les résultats, regroupés en classes, indiqués dans le tableau suivant :

| Temps (min) | $C_i$ | Effectifs | Fréquence 100 $f_i$ % | Fréq cumul croissante % | $c_i f_i$ |
|---|---|---|---|---|---|
| [0 ; 30[ | 15 | 175 | 13 | 13 | 1,95 |
| [30 ; 60[ | 45 | 392 | 30 | 43 | 13,5 |
| [60 ; 90[ | 75 | 267 | 21 | 64 | 15,75 |
| [90 ; 120[ | 105 | 127 | 9 | 73 | 9,45 |
| [120 ; 150[ | 135 | 168 | 13 | 86 | 17,55 |
| [150 ; 180[ | 165 | 120 | 9 | 95 | 14,85 |
| [180 ; 240[ | 210 | 63 | 5 | 100 | 10,5 |
| | | | | Total | 83,55 |

$$\textbf{Moyenne arithmétique} = \bar{x} = (13\times 15 + 30\times 45 + \ldots\ldots + 5\times 210)/100 = 83{,}55$$

$$\textbf{Mode} = \textbf{Mo} = 30 + \frac{(30-13)}{(30-13)+(30-21)} \times (60-30) = 49.6$$

$$\textbf{Premier Quartile} = Q_1 = 30 + \frac{100\times 0.25 - 13}{30} \times (60-30) = \textbf{42}$$

$$\textbf{Mediane} = Me = Q_2 = 60 + \frac{100\times 0.5 - 43}{21} \times (90-60) = 70$$

$$\text{Troisième Quartile} = Q_3 = 120 + \frac{100\times 0.75 - 73}{13} \times (150-120) = 124{,}62$$

$$\textbf{L'étendue} = 240 - 0 = 240$$

# Variance & Écart type

# Variance & l'Écart type

La variance est un indicateur de la dispersion d'une série par rapport à sa moyenne. Elle représente la somme des carrés des écarts à la moyenne divisée par le nombre d'observations.

## Définition

La variance d'une série statistiques est donnée par la formule :

$$V(x) \overset{def}{=} \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 . \quad (1)$$

Avec

$\{x_1, \ldots, x_n\}$ sont les observations ;

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ est la moyenne arithmétique simple de la série.

**Remarque :** Il existe une formule plus simple que (1), qui nécessite moins de calcul ; surtout quand $\bar{x}$ est un décimal.

# Variance & l'Écart type

**Théorème de Koenig :**

La variance peut aussi s'écrire

$$V(x) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 \qquad (2)$$

**Démonstration :**

$$V(x) \overset{def}{=} \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \overset{id.rem}{=} \frac{1}{n}\sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2)$$

$$\overset{distr}{=} \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)\bar{x} + \bar{x}^2\frac{1}{n}\sum_{i=1}^{n} 1$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\bar{x}^2 + \bar{x}^2 \quad \text{car } \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ et } \sum_{i=1}^{n} 1 = n$$

$$V(x) = \frac{1}{n}\sum x_i^2 - x^2$$

# Variance & l'Écart type

**Exemple :** Considérons la série suivante

$$\{11; 14; 24; 8; 32; 9; 10; 17\}$$

On a

$$\overline{x} = \frac{11 + 14 + 24 + 8 + 32 + 9 + 10 + 17}{8} = 15.625$$

En appliquant la définition :

$$V(x) = \frac{(11 - 15.625)^2 + (14 - 15.625)^2 + \ldots + (17 - 15.625)^2}{8}$$

$$\approx 62.225$$

En appliquant la propriété :

$$V(x) = \frac{11^2 + 14^2 + \ldots + 17^2}{8} - 15.625^2 = \frac{2451}{8} - 244.15 \approx 62.225$$

# Variance & l'Écart type

**Remarques :**

* Dans l'expression (2), la moyenne n'intervient qu'une seule fois.

* Si une variable quantitative discrète X, pouvant prendre $k$ valeurs distinctes $x_1, \ldots, x_k$ avec des effectifs $n_1, \ldots, n_k$ alors

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

i.e ; $k$ est le nombre de répétition d'une valeur de la modalité.

# Variance & l'Écart type

## Exemple : Calcul de la variance à partir de la définition

*Répartition de 10 notes obtenues par un élève en MATHS :*

| Note= $x_i$ | Effectif $n_i$ | $x_i - \bar{x}$ | $(x_i - \bar{x})^2$ | $n_i(x_i - \bar{x})^2$ |
|---|---|---|---|---|
| 6 | 1 | -5.5 | 30.25 | 30.25 |
| 9 | 1 | -2.5 | 6.25 | 6.25 |
| 10 | 2 | -1.5 | 2.25 | 4.5 |
| 11 | 1 | -0.5 | 0.25 | 0.25 |
| 12 | 2 | 0.5 | 0.25 | 0.5 |
| 13 | 1 | 1.5 | 2.25 | 2.25 |
| 16 | 2 | 4.5 | 20.25 | 40.50 |
| **Total** | 10 | | | **84.50** |

D'abord, on calcul **la moyenne arithmétique** de cette série :

$$\bar{x} = \frac{1\times6+1\times9+2\times10+\cdots+2\times16}{10} = 11.5$$

$$\text{La variance } V = \frac{\sum n_i (x_i - \bar{x})^2}{10} = 8.45$$

# Variance & l'Écart type

## Exemple : Calcul de la variance à partir de la propriété

*Répartition de 10 notes obtenues par un élève en MATHS :*

| Note = $x_i$ | Effectif $n_i$ | $x_i^2$ | $n_i x_i^2$ |
|---|---|---|---|
| 6 | 1 | 36 | 36 |
| 9 | 1 | 81 | 81 |
| 10 | 2 | 100 | 200 |
| 11 | 1 | 121 | 121 |
| 12 | 2 | 144 | 288 |
| 13 | 1 | 169 | 169 |
| 16 | 2 | 256 | 512 |
| **Total** | 10 | | 1407 |

**La moyenne arithmétique** de cette série c'est :

$$\bar{x} = \frac{1\times6+1\times9+2\times10+\cdots+2\times16}{10} = 11.5$$

**La variance** est donnée par :

$$V(x) = \frac{1407}{10} - 11.5^2 = 140.7 - 132.25 = 8.45$$

# Variance & l'Écart type

## Cas de variable continue

Dans le cas continu, nous remplaçons $x_i$ dans (1) par $c_i$ le centre de la classe afin de calculer la variance.

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

| Salaire | Effectif | centre | $n_i c_i$ | $n_i c_i^2$ |
|---|---|---|---|---|
| [800 ;1000[ | 20 | 900 | 18 000 | 16 200 000 |
| [1000 ;1200[ | 15 | 1100 | 16 500 | 18 150 000 |
| [1200 ;1500[ | 10 | 1350 | 13 500 | 18 225 000 |
| [1500 ;2000[ | 5 | 1750 | 8 750 | 15 312 500 |
| Total | 50 | | 56 750 | 67 887 500 |

$$\overline{x} = \frac{56750}{50} = 1135 \text{ et } V(x) = \frac{67887500}{50} - 1135^2 = 69525$$

# Variance & l'Écart type

**Définition :**

L' Écart type est la racine carrée de la variance :

$$\sigma_x = \sqrt{v(x)}$$

**Remarques :**

1- L'écart type est un paramètre plus fin que l'étendue, car il tient compte de la répartition des valeurs.

2- L'écart type est exprimé dans la même unité que la variable.

3- 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.

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

# L’indice de dispersion : coefficient de variation

# L'indice de dispersion : coefficient de variation

## coefficient de variation

Le coefficient de variation est égal au **rapport** de l'écart type par la moyenne de la distribution :

$$C_X = \frac{\sigma_X}{\overline{x}}$$

**Intérêt de l'utilisation du coefficient de variation (ou de dispersion)**

Lorsque deux séries (ou plusieurs) sont exprimées en unités différentes, l'analyse de la dispersion doit se faire par le biais du coefficient de variation plutôt que part le seul écart type ou écart absolu moyen (à voir plus loin).

# L'indice de dispersion : coefficient de variation

**Exemple :**

Les deux séries suivantes correspondent aux salaires mensuels perçus par les ouvriers d'une entreprise exprimés une première fois en **DH** et une seconde fois en **centimes**.

*Série 1* : 1 150, 1 200, 1 600, 1 850, 2 150, 2 200, 2 350, 2 400, 3 000.
*Série 2* : 115 000, 120 000, 160 000, 185 000, 215 000, 220 000, 235 000, 240 000, 300 000.

La variance des deux séries est obtenue à partir de la formule:

$$V = \frac{1}{n}\sum_{i=1}^{9} x_i^2 - \bar{x}^2$$

L'écart type pour chacune des deux séries est obtenu à partir de la formule:

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

Nous trouvons **$\sigma_1$=566,55** et **$\sigma_2$=56 655.**

Le fait que $\sigma_2$ est plus élevée que $\sigma_1$ ne révèle pas une grande dispersion de la second par rapport à la première : c'est seulement **l'effet des unités de mesure** (1 dh = 100 centimes).

Les **coefficients de variation** (CV) associés à chaque série :

$$CV_1 = \frac{\sigma_1}{\bar{x}_1} = \frac{566{,}55}{1988{,}88} = 0{,}28 \quad \textbf{et} \quad CV_2 = \frac{\sigma_2}{\bar{x}_2} = \frac{56655}{198888} = 0{,}28.$$

**Conclusion :** Les deux séries sont bien de même dispersion.