# p.11 & p.14: on utilise le « i » minuscule pour numéroter les observations (et non un I majuscule)

Table 3.3

*Données brutes relatives à la taille de T-shirt d'enfants de 4ème primaires*

| i | Taille ($X_i$) | i | Taille ($X_i$) | i | Taille ($X_i$) | i | Taille ($X_i$) |
|---|---|---|---|---|---|---|---|
| 1 | S | 11 | M | 21 | M | 31 | L |
| 2 | M | 12 | S | 22 | M | 32 | L |
| 3 | M | 13 | S | 23 | M | 33 | L |
| 4 | M | 14 | M | 24 | M | 34 | L |
| 5 | S | 15 | S | 25 | S | 35 | S |
| 6 | XS | 16 | S | 26 | S | 36 | S |
| 7 | XS | 17 | M | 27 | M | 37 | S |
| 8 | XS | 18 | M | 28 | M | 38 | S |
| 9 | S | 19 | M | 29 | M | 39 | S |
| 10 | S | 20 | M | 30 | L | 40 | S |

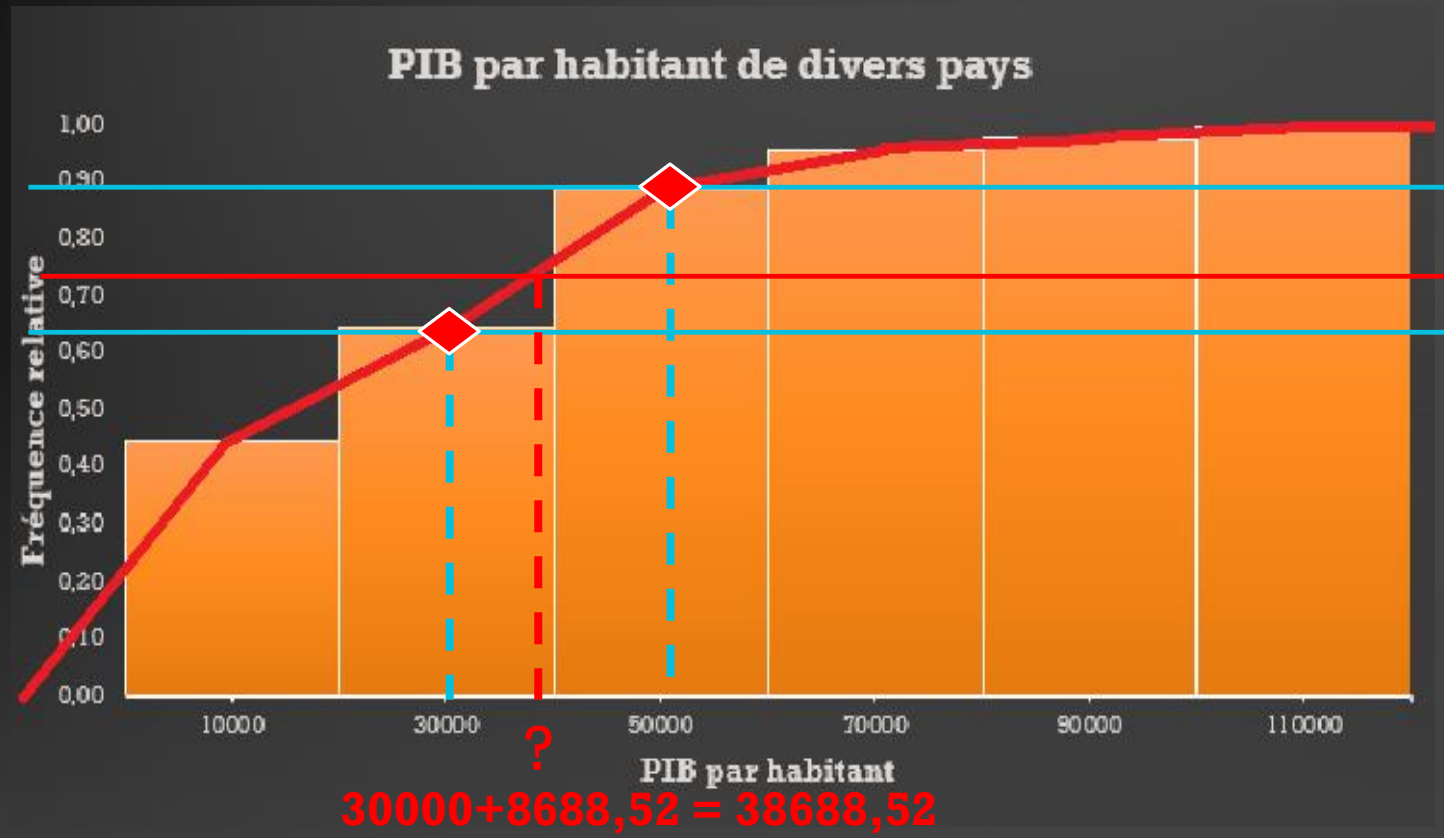
Table 3.6

*Données brutes relatives au nombre d'enfants par ménages, au sein de 40 ménages*

| i | Enfants ($X_i$) | i | Enfants ($X_i$) | i | Enfants ($X_i$) | i | Enfants ($X_i$) |
|---|---|---|---|---|---|---|---|
| 1 | 5 | 11 | 1 | 21 | 1 | 31 | 1 |
| 2 | 2 | 12 | 3 | 22 | 1 | 32 | 1 |
| 3 | 2 | 13 | 3 | 23 | 2 | 33 | 2 |
| 4 | 3 | 14 | 1 | 24 | 2 | 34 | 2 |
| 5 | 1 | 15 | 4 | 25 | 1 | 35 | 1 |
| 6 | 2 | 16 | 3 | 26 | 2 | 36 | 2 |
| 7 | 1 | 17 | 3 | 27 | 3 | 37 | 3 |
| 8 | 2 | 18 | 1 | 28 | 1 | 38 | 1 |
| 9 | 2 | 19 | 2 | 29 | 3 | 39 | 3 |
| 10 | 2 | 20 | 1 | 30 | 4 | 40 | 4 |

# MÉTHODE D'INTERPOLATION LINÉAIRE

- Déterminez le 3ème quartile.



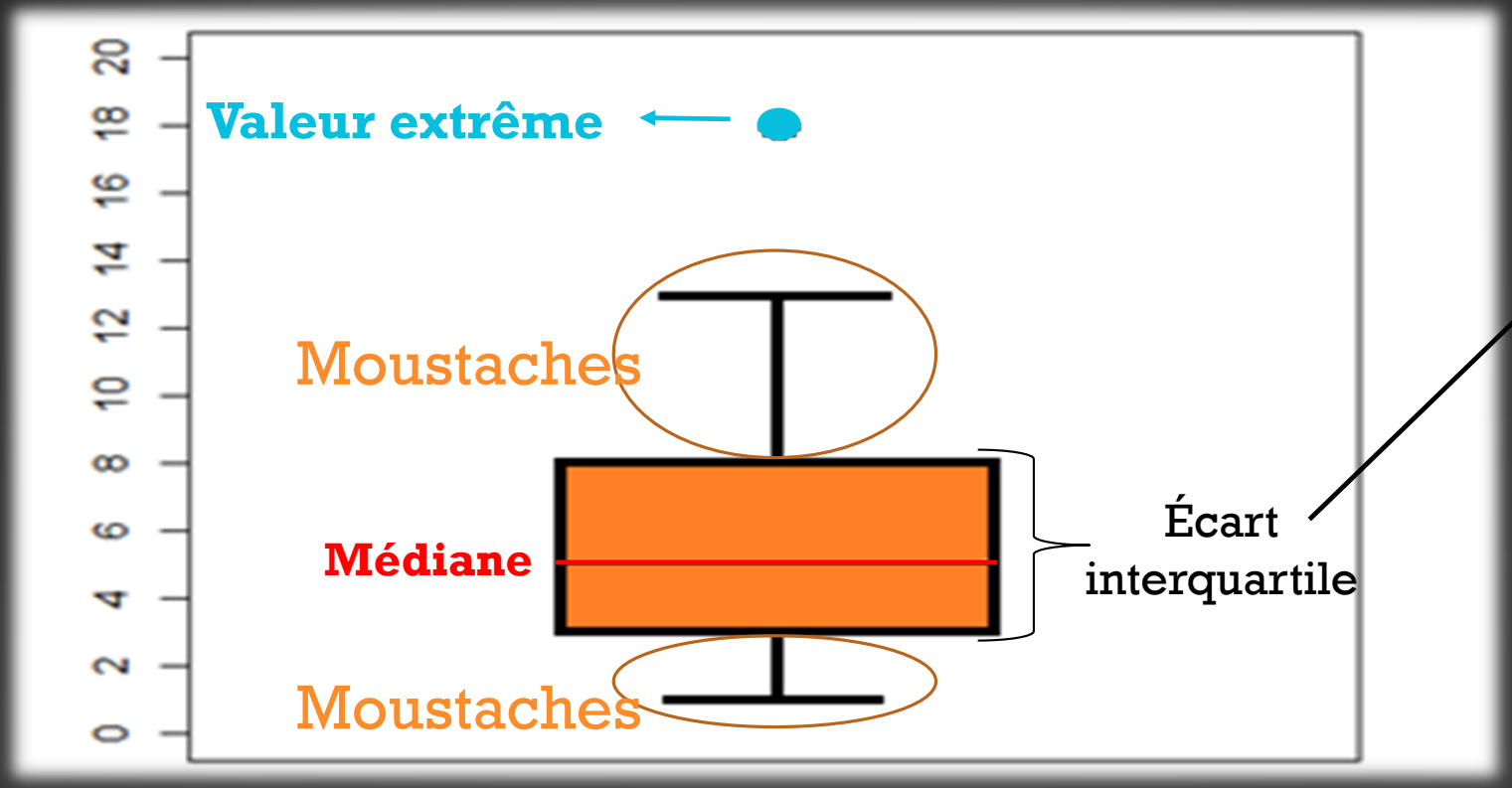

| Classe | Centre de classe | $n_j$ | $f_j$ ( en %) | $N_j$ | $F_j$ |
|---|---|---|---|---|---|
| [0-20000[ | 10000 | 20 | 0.444 | 20 | **0.444** |
| [20000-40000[ | 30000 | 9 | 0.200 | 29 | **0.644** |
| [40000-60000[ | 50000 | 11 | 0.244 | 40 | **0.888** |
| [60000-80000[ | 70000 | 3 | 0.067 | 43 | **0.955** |
| [80000-100000[ | 90000 | 1 | 0.022 | 44 | **0.977** |
| [100000-120000[ | **110000** | **1** | **0.022** | **45** | **1** |

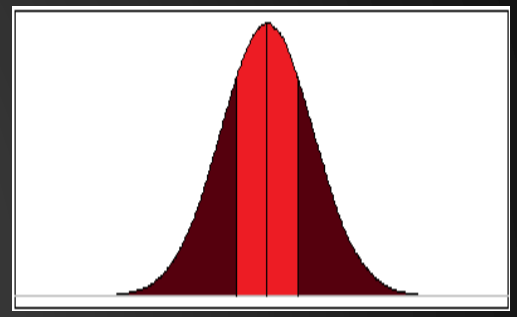
.244 = 20000

.001 = 81,97

.106 = 8688,52

# BOÎTES À MOUSTACHE



Q3 – Q1

- **Les moustaches:**
  - **Limites = les barrières (à distance de 1,5 x écart interquartile)**

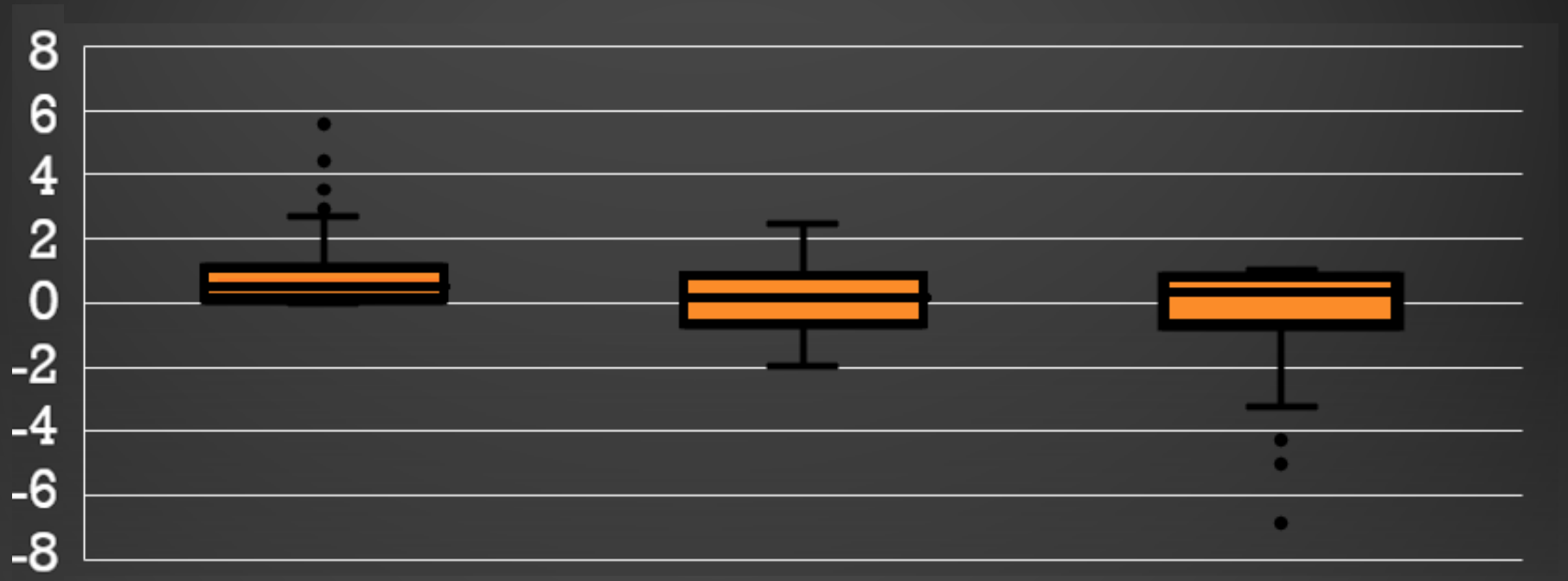
***Exemple.:*** *$Q_1 = 8$; $Q_3 = 13$*

➔**Écart interquartile = 5➔ 5 x 1,5 = 7,5!!**

➔**Barrière inférieure = 8 – 7,5 = 0,5**

➔**Barrière supérieure = 13 + 7,5 = 20;5**

# DECRIRE CHAQUE BOXPLOT

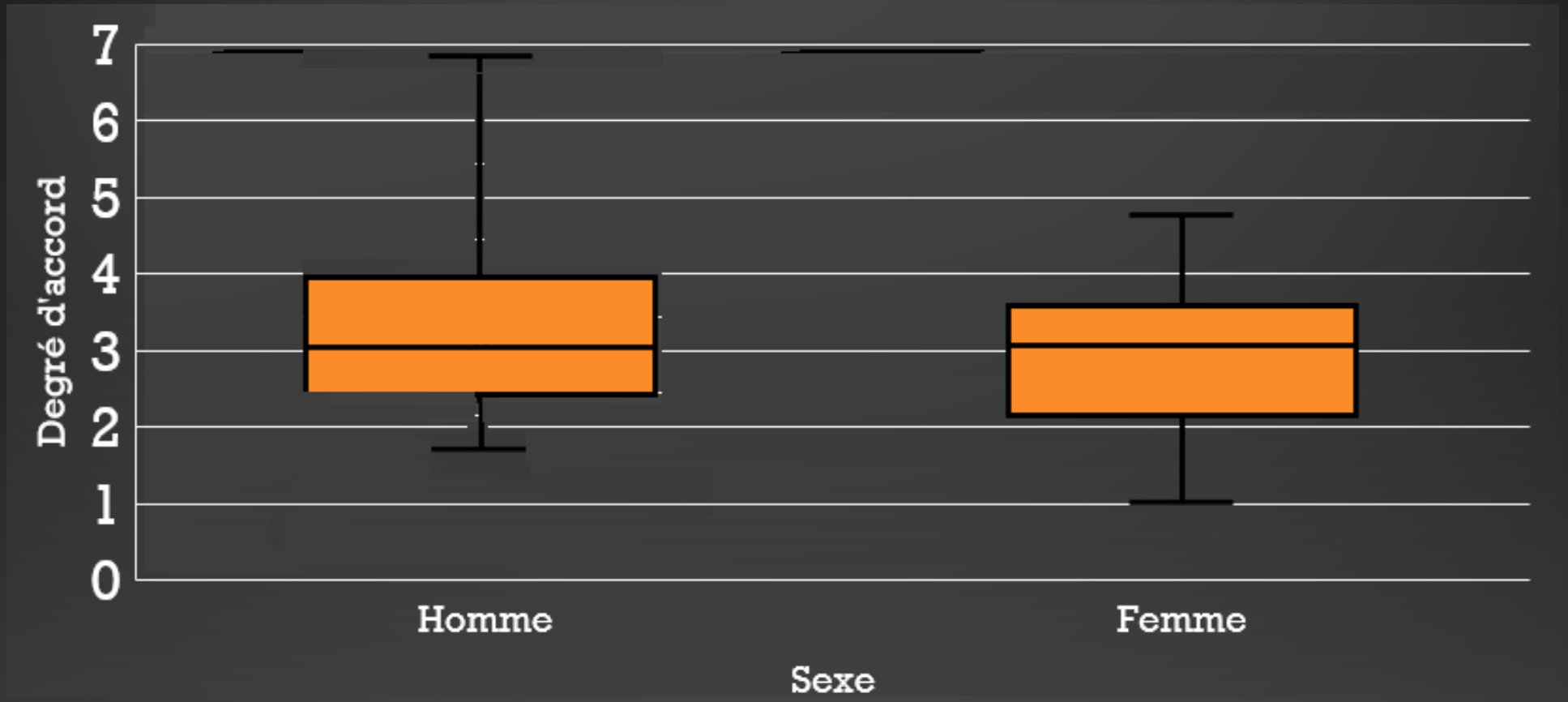



**Asymétrie positive**

**symétrie**

**Asymétrie négative**

Décrivez chaque boite à moustache. Dans quel groupe le score est-il le plus élevé « globalement »?



# CHAPITRE 4

## EXPLORATION ALGÉBRIQUE DES DONNÉES À UNE DIMENSION

# LES INDICATEURS ALGÉBRIQUE PRINCIPAUX

1. Les mesures de la tendance centrale
2. Les mesures de la dispersion
3. La détermination algébrique de la symétrie et de l'aplatissement

# 1. **MESURE DE LA TENDANCE CENTRALE**

- LE MODE
- LA MOYENNE
- LA MÉDIANE

**(a) Série**

| i | Ancienneté ($Y_i$) |
|---|---|
| 1 | 1 |
| 2 | 5 |
| 3 | 2 |
| 4 | 8 |
| 5 | 8 |
| 6 | 9 |
| 7 | 9 |
| 8 | 1 |
| 9 | 1 |
| 10 | 1 |
| 11 | 8 |
| 12 | 8 |
| 13 | 7 |
| 14 | 10 |
| 15 | 8 |

**(b) Distribution de fréquence (= tableau de transnumérisation)**

| $y_j$ | $n_j$ fréquence absolue) | $f_j$ (fréquence relative) | Fréquences relatives (en %) |
|---|---|---|---|
| 1 | 4 | 0.27 | 27% |
| 2 | 1 | 0.07 | 7% |
| 3 | 0 | 0.00 | 0% |
| 4 | 0 | 0.00 | 0% |
| 5 | 1 | 0.07 | 7% |
| 6 | 0 | 0.00 | 0% |
| 7 | 1 | 0.07 | 7% |
| 8 | 5 | 0.33 | 33% |
| 9 | 2 | 0.13 | 13% |
| 10 | 1 | 0.07 | 7% |
| TOTAL | 9 | 1,00 | 100% |

- **MODE**: classe/valeur la plus représentée.

- **AVANTAGE**: insensible aux valeurs aberrantes (>< moyenne)
- **INCONVENIENT**: insensible à TOUTES les autres valeurs de la distribution

**(a) Série**

| i | Ancienneté ($Y_i$) |
|---|---|
| 1 | 1 |
| 2 | 5 |
| 3 | 2 |
| 4 | 8 |
| 5 | 8 |
| 6 | 9 |
| 7 | 9 |
| 8 | 1 |
| 9 | 1 |
| 10 | 1 |
| 11 | 8 |
| 12 | 8 |
| 13 | 7 |
| 14 | 10 |
| 15 | 8 |

**(b) Distribution de fréquence (= tableau de transnumérisation)**

| $y_j$ | $n_j$ (valeur absolue) | $f_j$ (fréquence relative) | Fréquences relatives (en %) |
|---|---|---|---|
| 1 | $4 = n_1$ | 0.27 | 27% |
| 2 | $1 = n_2$ | 0.07 | 7% |
| 3 | 0 | 0.00 | 0% |
| 4 | 0 | 0.00 | 0% |
| 5 | 1 | 0.07 | 7% |
| 6 | 0 | 0.00 | 0% |
| 7 | 1 | 0.07 | 7% |
| 8 | 5 | 0.33 | 33% |
| 9 | 2 | 0.13 | 13% |
| 10 | 1 | 0.07 | 7% |
| TOTAL | 15 | 1,00 | 100% |

**NOTATION**:

- Indice i dans les séries statistiques
- Indice j dans les distributions de fréquence

# ATTENTION

Ne pas confondre le mode et la fréquence associée au mode!

**MOYENNE calculée à partir d'une série statistique**: = somme des valeurs divisées par le nombre de valeurs de la somme

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{somme\ des\ observations}{n}$$

n = effectif total

$X_i$ = les valeurs que peut prendre la variable

i = numéro de la donnée

$\bar{X}$ = moyenne de l'échantillon

| (a) Série | |
|---|---|
| **i** | **Ancienneté ($Y_i$)** |
| 1 | 1 |
| 2 | 5 |
| 3 | 2 |
| 4 | 8 |
| 5 | 8 |
| 6 | 9 |
| 7 | 9 |
| 8 | 1 |
| 9 | 1 |
| 10 | 1 |
| 11 | 8 |
| 12 | 8 |
| 13 | 7 |
| 14 | 10 |
| 15 | 8 |

$$\bar{Y} = \frac{1+5+2+8+8+9+9+1+1+1+8+8+7+10+8}{15}$$

# ERRATUM

- P.34: La variable « ancienneté » s'appelle « $Y_i$ » et non « $X_i$ » (cf. Table 1.1)
- P.35: Dans le tableau de fréquence de l'âge (variable $X_i$), la notation requise est « $x_j$ » et non « $y_j$ ».

**MOYENNE calculée à partir de données présentées sous forme de distribution de fréquences**:

= lorsque très grand nombre de données

- Calcul sur base de fréquences absolues ($n_j$)

$$\bar{X} = \frac{\sum_{j=1}^{J} n_j x_j}{n}$$

- Calcul sur base de fréquences relatives ($f_j$)

$$\bar{X} = \sum_{j=1}^{J} f_j x_j$$

$$\bar{Y} = \frac{1+5+2+8+8+9+9+1+1+1+8+8+7+10+8}{15}$$

$$\Leftrightarrow \bar{Y} = \frac{1+1+1+1+2+5+7+8+8+8+8+8+9+9+10}{15}$$

$$\bar{X} = \frac{\sum_{j=1}^{J} n_j x_j}{n} \Leftarrow \Leftrightarrow \bar{Y} = \frac{4\times 1+1\times 2+1\times 5+1\times 7+5\times 8+2\times 9+1\times 10}{15}$$

$$\bar{X} = \sum_{j=1}^{J} f_j x_j \Leftarrow \Leftrightarrow \bar{Y} = \frac{4}{15}\times 1+\frac{1}{15}\times 2+\frac{1}{15}\times 5+\frac{1}{15}\times 7+\frac{5}{15}\times 8+\frac{2}{15}\times 9+\frac{1}{15}\times 10$$

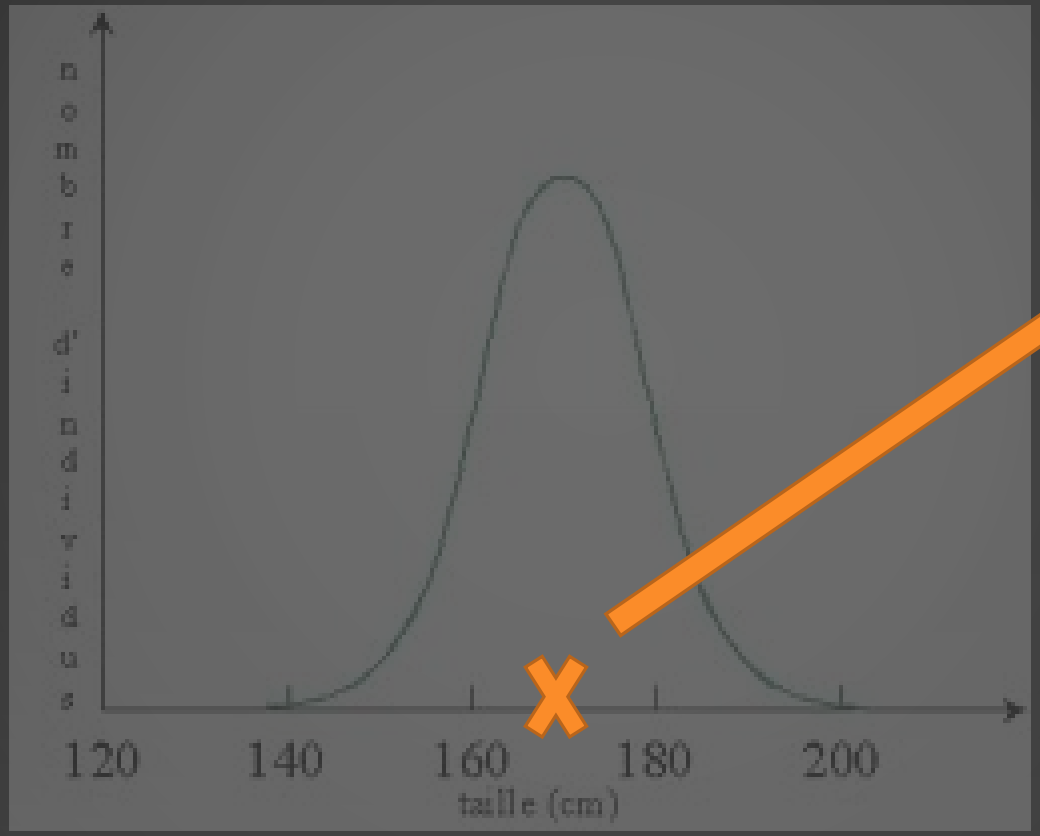
# INCONVÉNIENTS DE LA MOYENNE

- Sensible aux valeurs aberrantes
- Peu représentatif d'une distribution **non symétrique** ou **multimodale**

# AVANTAGES DE LA MOYENNE

- Représente parfaitement une distribution **normale** (uni-modale et symétrique)

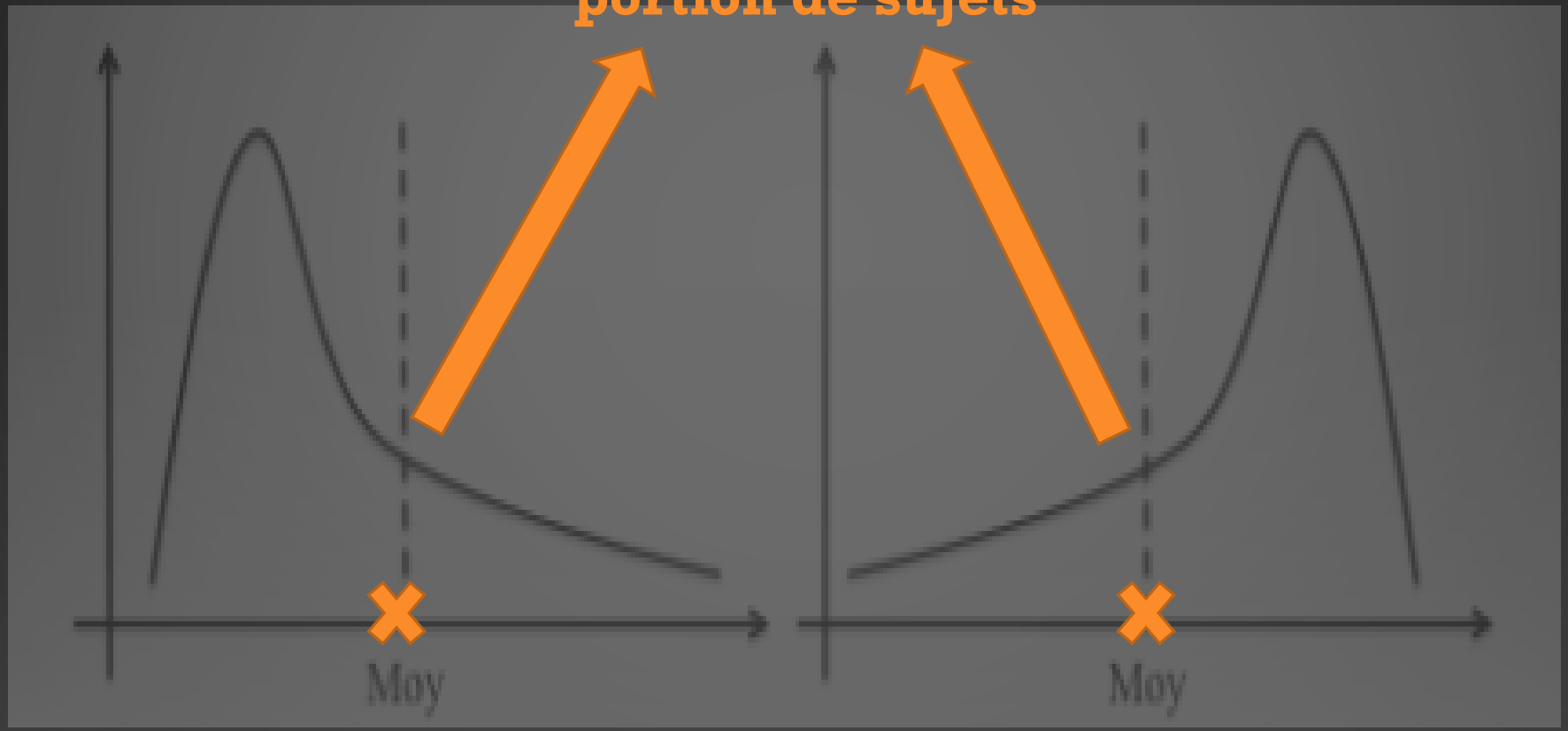
# LA DISTRIBUTION NORMALE



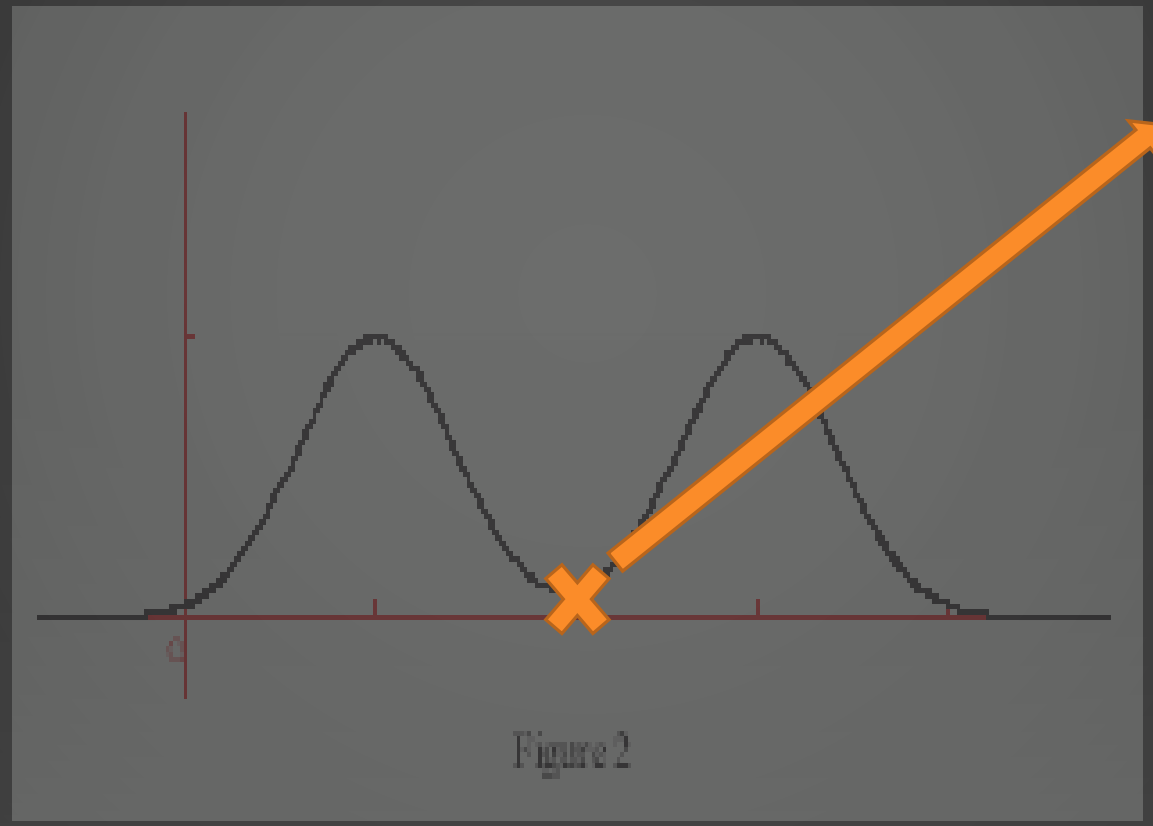

**La moyenne correspond à la valeur la plus représentée**

# DISTRIBUTION ASYMÉTRIQUE

**La moyenne ne correspond qu'à une très faible portion de sujets**



# DISTRIBUTION SYMÉTRIQUE BIMODALE




**La moyenne ne correspond qu'à une très faible portion de sujets**

# SYNTHÈSE

| Distribution normale: | Distribution multimodale ou asymétrique: |
| --- | --- |
| La moyenne représentera parfaitement la distribution | Moyenne très peu représentative |

Importance des indices de symétrie!

# 2. **MESURE DE LA DISPERSION**

- ÉTENDUE
-ÉCART INTERQUARTILE
- ÉCART MOYEN ABSOLU
- VARIANCE ET ÉCART-TYPE

# ÉTENDUE DES DONNÉES

| (a) Série | |
|---|---|
| **i** | **math ($X_i$)** |
| 1 | 8.5 |
| 2 | 4 |
| 3 | 4.5 |
| 4 | 6 |
| 5 | 3.5 |
| 6 | 6.5 |
| 7 | 7 |
| 8 | 5 |
| 9 | 7.5 |
| 10 | 7.5 |

Valeur maximale observée – valeur minimale observée

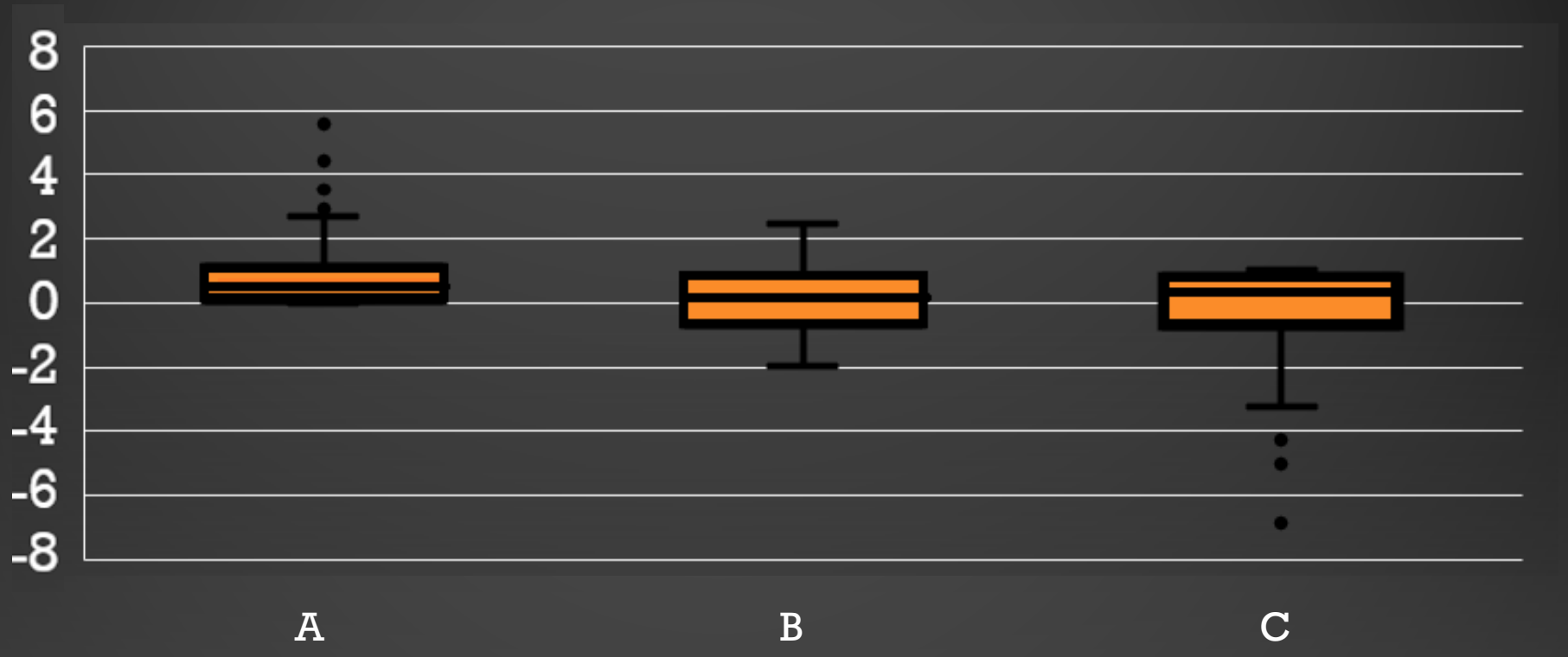
→ Dans l'exemple

Max = 8.5

Min = 3.5

E = ***8.5 – 3.5= 5***

# EXERCICE:
# QUELLE EST L'ÉTENDUE DES SÉRIES A, B ET C?
# (PRÉCISION À LA DEMI UNITÉ PRÈS)



# ÉTENDUE DES DONNÉES

**Avantage:**

**- Très facile et rapide à calculer**

**Problème**:

- Ne dépend que de deux valeurs, donc très peu représentatif de la distribution!

- Est très sensible aux valeurs extrêmes

# ÉCART INTER-QUARTILE

Correspond à la boite centrale des boîtes à moustaches

Q3 – Q1

# ÉCART INTER-QUARTILE

| (a) Série | |
|---|---|
| i | math ($X_i$) |
| 1 | 8.5 |
| 2 | 4 |
| 3 | 4.5 |
| 4 | 6 |
| 5 | 3.5 |
| 6 | 6.5 |
| 7 | 7 |
| 8 | 5 |
| 9 | 7.5 |
| 10 | 7.5 |

3.5,4,4.5, 5, 6 | 6.5,7, 7.5,7.5,8.5

3.5,4,4.5,5,6

=Q1

6.5,7,7.5,7.5,8.5

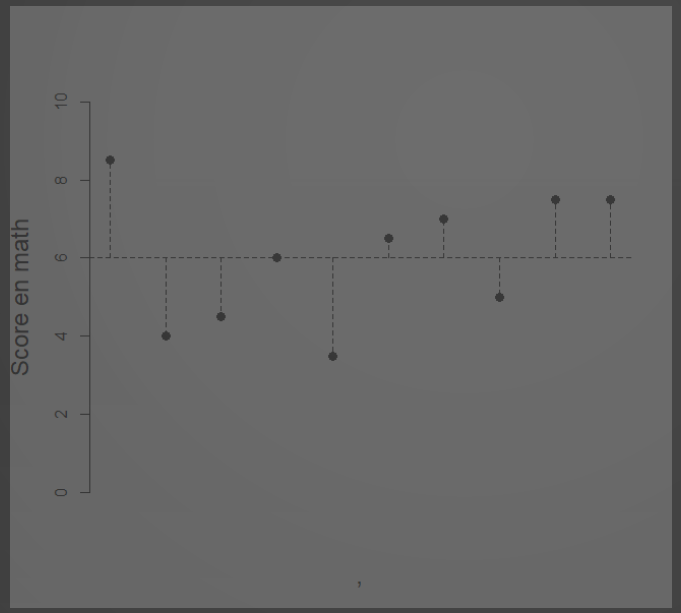
=Q3

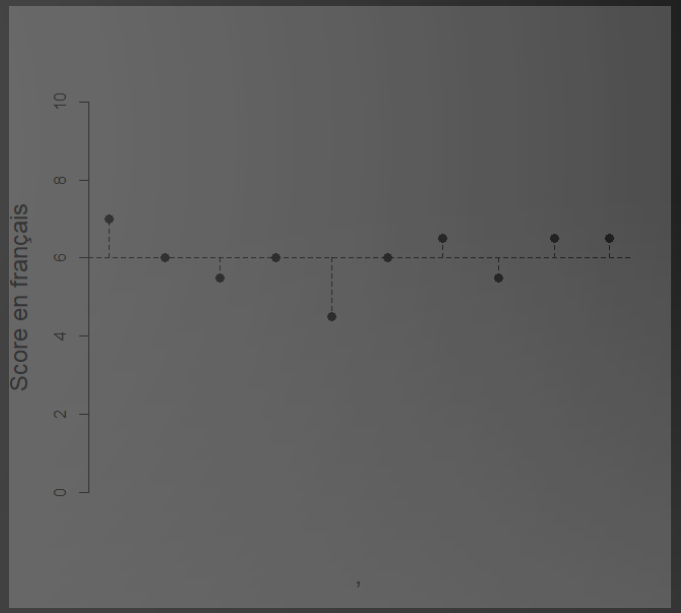
Q3 – Q1 = 7.5 – 4.5 = 3

# EMA ET VARIANCE

| (a) Série | |
|---|---|
| **i** | **math ($X_i$)** |
| 1 | 8.5 |
| 2 | 4 |
| 3 | 4.5 |
| 4 | 6 |
| 5 | 3.5 |
| 6 | 6.5 |
| 7 | 7 |
| 8 | 5 |
| 9 | 7.5 |
| 10 | 7.5 |

Sert à déterminer à quel point les données tendent à s'éloigner plus ou moins fort de la moyenne



# ETAPE 1 & 2

- Calculer la moyenne de la série
- Calculer l'écart de chaque observation par rapport à la moyenne.

| (a) Série | | |
|---|---|---|
| **i** | $X_i$ | $X_i$-$\bar{X}$ |
| 1 | 8.5 | 2.5 |
| 2 | 4 | -2 |
| 3 | 4.5 | -1.5 |
| 4 | 6 | 0 |
| 5 | 3.5 | -2.5 |
| 6 | 6.5 | 0.5 |
| 7 | 7 | 1 |
| 8 | 5 | -1 |
| 9 | 7.5 | 1.5 |
| 10 | 7.5 | 1.5 |

# ETAPE 3

Puisqu'on cherche à déterminer de combien les sujets s'éloignent de la tendance centrale « en moyenne », il semblerait logique de calculer la moyenne des écarts $(X_i\text{-}\bar{X})$

Mais….

# ETAPE 3

| (a) Série | | |
|---|---|---|
| **i** | $X_i$ | $X_i$-$\bar{X}$ |
| 1 | 8.5 | 2.5 |
| 2 | 4 | -2 |
| 3 | 4.5 | -1.5 |
| 4 | 6 | 0 |
| 5 | 3.5 | -2.5 |
| 6 | 6.5 | 0.5 |
| 7 | 7 | 1 |
| 8 | 5 | -1 |
| 9 | 7.5 | 1.5 |
| 10 | 7.5 | 1.5 |

Moyenne des $(X_i\text{-}\bar{X}) = 0$

Les « + » et les « - » s'annulent!

# SOLUTION 1: EMA: ÉCART MOYEN ABSOLU

= Moyenne des écarts (pris en valeur absolue) de chaque donnée par rapport au paramètre estimé (***cf.**: la moyenne*)

$$EMA = \frac{\sum_{i=1}^{n} |X_i - \bar{X}|}{n}$$

| (a) Série | | | |
|---|---|---|---|
| **i** | $X_i$ | $X_i$-$\bar{X}$ | $\lvert X_i$-$\bar{X}\rvert$ |
| 1 | 8.5 | 2.5 | 2.5 |
| 2 | 4 | -2 | 2 |
| 3 | 4.5 | -1.5 | 1.5 |
| 4 | 6 | 0 | 0 |
| 5 | 3.5 | -2.5 | 2.5 |
| 6 | 6.5 | 0.5 | 0.5 |
| 7 | 7 | 1 | 1 |
| 8 | 5 | -1 | 1 |
| 9 | 7.5 | 1.5 | 1.5 |
| 10 | 7.5 | 1.5 | 1.5 |

$$\text{EMA} = \frac{2.5+2+1.5+0+2.5+0.5+1+1+1.5+1.5}{10} = 1.4$$

**Ccl**: en moyenne, les sujets s'écartent de 1.4 points de la moyenne

# EMA: ÉCART MOYEN ABSOLU

- Très bonne représentation de la dispersion

MAIS

Supplantée par l'écart-type, à cause des propriétés mathématiques de la variance dont il est dérivé

# SOLUTION 2: VARIANCE ($S^2$) ET ÉCART-TYPE (S)

= Moyenne des écarts (élevés au carré) de chaque donnée par rapport au paramètre estimé (***cf.**: la moyenne*)

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}$$

| (a) Série | | | |
|---|---|---|---|
| **i** | $X_i$ | $X_i$-$X$ | $(X_i-\bar{X})^2$ |
| 1 | 8.5 | 2.5 | 6.25 |
| 2 | 4 | -2 | 4 |
| 3 | 4.5 | -1.5 | 2.25 |
| 4 | 6 | 0 | 0 |
| 5 | 3.5 | -2.5 | 6.25 |
| 6 | 6.5 | 0.5 | 0.25 |
| 7 | 7 | 1 | 1 |
| 8 | 5 | -1 | 1 |
| 9 | 7.5 | 1.5 | 2.25 |
| 10 | 7.5 | 1.5 | 2.25 |

$$S^2 = \frac{6,25+4+2,25+0+6,25+0,25+1+1+2,25+2,25}{10} = 2.55$$

**Ccl**: en moyenne, les sujets s'écartent de 2.55 « points au carré » de la moyenne

Difficile à interpréter, car exprimé dans une unité « au carré »

**Solution = écart-type!**

# VARIANCE ET ÉCART-TYPE

- Écart-type = racine carré de la variance

$$S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}}$$

$$S = \sqrt{2{,}55} = 1{,}597$$

- Remarque:

CME > S: 1,597 > 1,4

➔ Surestimation de l'erreur par l'écart-type (par rapport à l'EMA).

# INCONVÉNIENT DE LA VARIANCE (ET ÉCART-TYPE)

- Sensible aux valeurs extrêmes

➔Car calcul des écart par rapport à la **moyenne** (elle-même très sensible)

➔erreur élevée au carré → sensibilité exacerbée!

# REPRISE DE NOTRE EXEMPLE

| (a) Série | |
|---|---|
| Num | Cotes ($X_i$) |
| 1 | 8.5 |
| 2 | 4 |
| 3 | 4.5 |
| 4 | 6 |
| 5 | 3.5 |
| 6 | 6.5 |
| 7 | 7 |
| 8 | 5 |
| 9 | 7.5 |

$S^2 = 2,55$

| (a) Série | |
|---|---|
| Num | Cotes ($X_i$) |
| 1 | 8.5 |
| 2 | 4 |
| 3 | 4.5 |
| 4 | **16** |
| 5 | 3.5 |
| 6 | 6.5 |
| 7 | 7 |
| 8 | 5 |
| 9 | 7.5 |

$S^2 = 11,55$

| (a) Série | |
|---|---|
| Num | Cotes ($X_i$) |
| 1 | 8.5 |
| 2 | 4 |
| 3 | 4.5 |
| 4 | **66** |
| 5 | 3.5 |
| 6 | 6.5 |
| 7 | 7 |
| 8 | 5 |
| 9 | 7.5 |

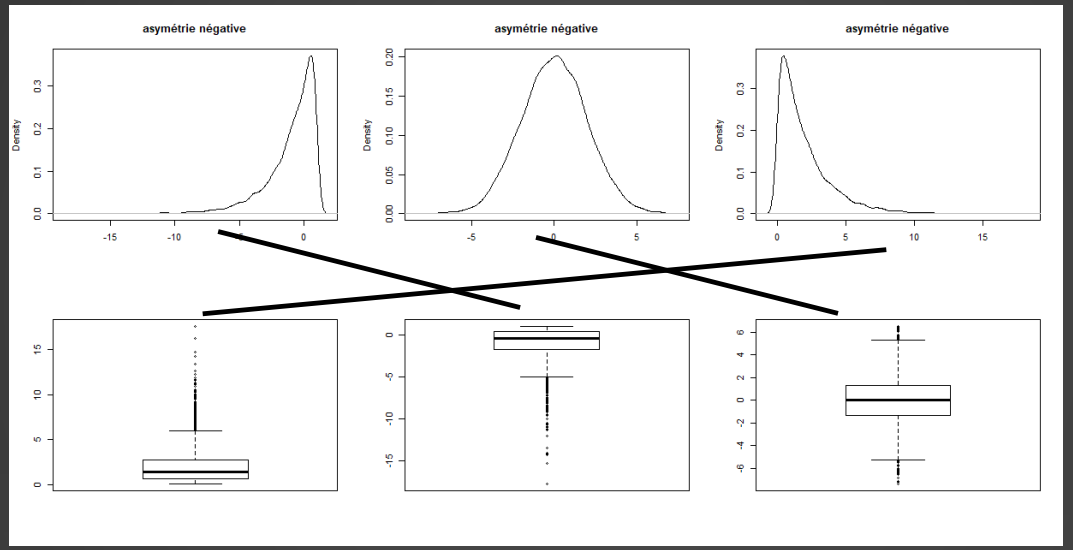
$S^2 = 326,55$

# 3. **LA DÉTERMINATION ALGÉBRIQUE DE LA SYMÉTRIE ET DE L'APLATISSEMENT**

- LES MOMENTS
- COEFFICIENT $G_1$ DE FISHER (ASYMÉTRIE)

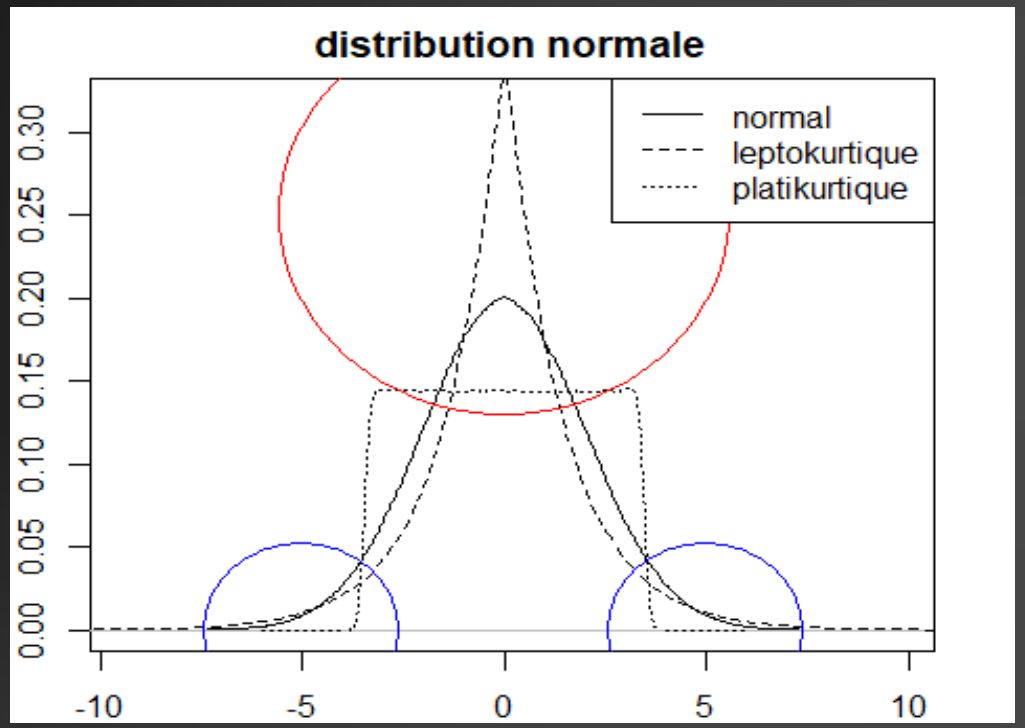
COEFFICIENT G2 DE FISHER (APLATISSEMENT)

# INDICE D'ASYMÉTRIE (SKEWNESS) OU COEFFICIENT G1 DE FISHER




$G_1 = 0$ si la distribution est symétrique
$G_1 > 0$ si asymétrie positive
$G_1 < 0$ si asymétrie négative

# INDICE D'APLATISSEMENT (KURTOSIS) OU COEFFICIENT G2 DE FISHER




$G_2 = 0$ si la distribution a l'aplatissement la normale
$G_2 > 0$ si plus pointue que la normale (leptokurtique)
$G_2 < 0$ si aplatie que la normale (platikurtique)