Chapitre 10 : Statististiques

Dans tout ce chapitre, on considérera les 3 séries statistiques suivantes :

Série A :
Notes obtenues à un contrôle dans une classe de 30 élèves :

$2-3-3-4-5-6-6-7-7-7-8-8-8-8-8-9-9-9-9-9-9-10-10-11-11-11-13-13-15-16$

Série B :
Salaires en euros des employés d'une entreprise :

Salaires	[900; 1200[	[1200; 1400[	[1400; 1600[	[1600; 1800[	[1800; 2000[	[2000; 2400[	TOTAL
Effectif	30	30	60	80	40	40	280

Série C :
Proportion d'adhérents à un club sportif dans différentes sections :

- ➔ 17% jouent au handball,
- ➔ 25% jouent au rugby,
- ➔ 58% jouent au tennis.

I. Méthodes de représentation

1) Vocabulaire

La **population** est l'ensemble des individus sur lesquels portent l'étude statistique. *(Par exemple la section d'ECO1, la population féminine, les fonctionnaires ...)* dont chaque élément est appelé **individu**.

Un **échantillon** est une partie de la population considérée.

Le **caractère** (ou **variable**) d'une série statistique est une propriété étudiée sur chaque individu :

- ➩ Lorsque le caractère ne prend que des valeurs (ou **modalités**) numériques, il est **quantitatif** :
  - ⋆ **discret** s'il ne peut prendre que des valeurs isolées (notes, âge ...)
  - ⋆ **continu** dans le cas contraire *(poids, taille ...)*. Dans ce cas on effectue souvent un regroupement des valeurs par **classes**.
- ➩ Sinon, on dit qu'il est **qualitatif** *(couleur des yeux, sport pratiqué ...)* : les modalités ne sont pas des nombres.

A chaque valeur (ou classe) est associée un **effectif** n : c'est le nombre d'individus associés à cette valeur.

2) Tableaux

Définition. On considère une série statistique X à caractère quantitatif, dont les p valeurs sont données par $x_1, x_2, \ldots, x_p$ d'effectifs associés $n_1, n_2, \ldots, n_p$ avec $n_1 + n_2 + ... + n_p = N$.

- ➤ A chaque valeur (ou classe) est associée une fréquence f_i : c'est la proportion d'individus associés à cette valeur.
- ➤ $f_i = \dfrac{n_i}{N}$ est un nombre compris entre 0 et 1, que l'on peut écrire sous forme de pourcentage.
- ➤ L'ensemble des fréquences de toutes les valeurs du caractère s'appelle la distribution des fréquences de la série statistique.

Exemple 1. On peut représenter la **série A** par un tableau d'effectifs, et le compléter par la distribution des fréquences :

Notes	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Eff.	0	1	2	1	1	2	3	5	6	2	3	0	2	0	1	1	0	0	0
Fréq. en %	0	3	7	3	3	7	10	17	20	7	10	0	7	0	3	3	0	0	0

On peut aussi faire un regroupement par classe, ce qui rend l'étude moins précise, mais qui permet d'avoir une vision plus globale.

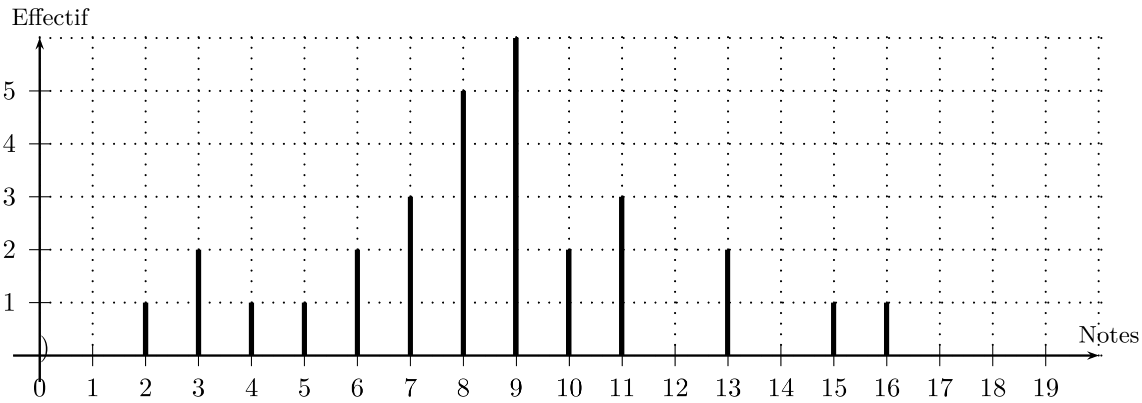
Exemple 2. Toujours pour la **série A**, si on regroupe les données par classes d'amplitude 5 points, on obtient :

Notes	$[\,0\ ;\ 5\,[$	$[\,5\ ;\ 10\,[$	$[\,10\ ;\ 15\,[$	$[\,15\ ;\ 20\,[$	total
Effectif	4	17	7	2	30
Fréquence	$0,13$	$0,57$	$0,23$	$0,07$	1

3) Graphiques

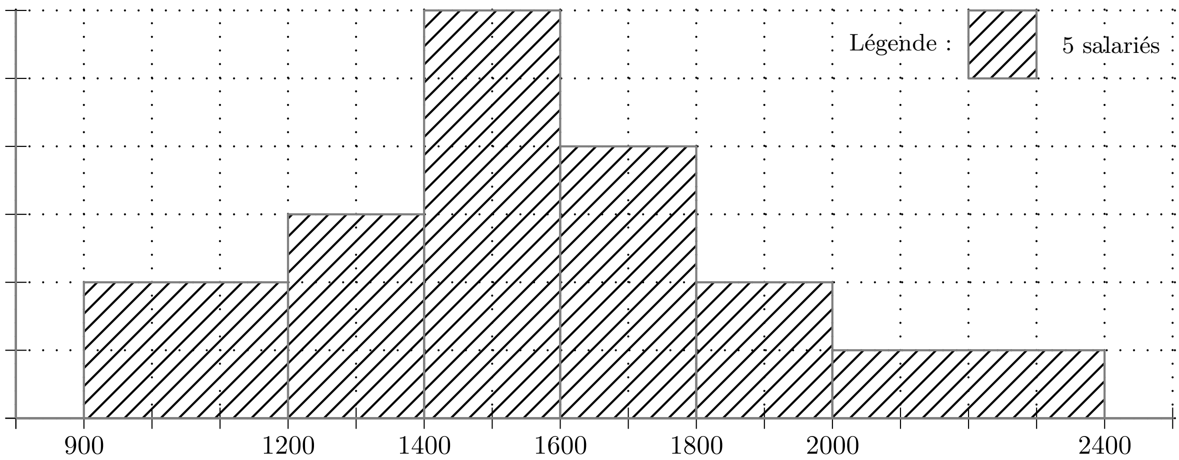
Lorsque le caractère étudié est **quantitatif et discret**, on peut représenter la série statistique étudiée par un **diagramme en bâtons** : la hauteur de chaque bâton est alors proportionnelle à l'effectif (ou à la fréquence) associé à chaque valeur.

Exemple 3. Voici le diagramme en bâtons représentant la série des notes de la **série A** :



Lorsque le caractère étudié est **quantitatif et continu**, et lorsque les modalités sont regroupées en classes, on peut représenter la série par un **histogramme** : l'aire de chaque rectangle est alors proportionnelle à l'effectif (ou à la fréquence) associée à chaque classe.
Lorsque les classes ont la même **amplitude**, c'est la hauteur qui est proportionnelle à l'effectif.

Exemple 4. Pour la **série B**, on obtient par exemple l'histogramme suivant :



Lorsque le caractère est **qualitatif**, on peut représenter la série par :

• **Un diagramme circulaire** (« camemberts ») :
La mesure de chaque secteur angulaire est proportionnelle à l'effectif associé.

• **Un diagramme en tuyaux d'orgue :**
Chaque classe est représentée par un rectangle de même largeur et de longueur proportionnelle à l'effectif, donc à la fréquence.

• **un diagramme en bandes :**
Chaque classe est représentée par un rectangle de même largeur et de longueur proportionnelle à l'effectif, donc à la fréquence.

Exemple 5. Diagrammes de la **série C**

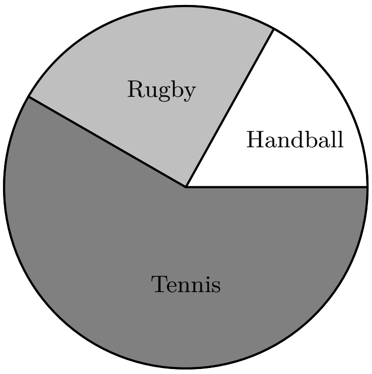


Diagramme circulaire

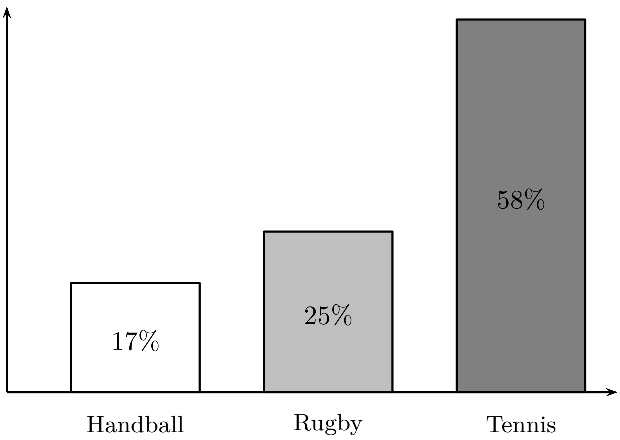


Diagramme en tuyau d'orgue

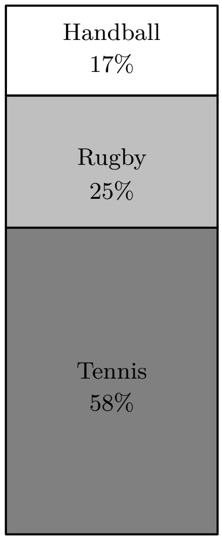


Diagramme en bandes

II. Caractéristiques de position

1) Moyenne

Définition. Soit une série statistique à caractère quantitatif, dont les p valeurs sont données par $x_1, x_2, \ldots, x_p$ d'effectifs associés $n_1, n_2, \ldots, n_p$ avec $n_1 + n_2 + ... + n_p = N$.
La moyenne pondérée de cette série est le nombre noté $\overline{x}$ qui vaut

$$\overline{x} = \frac{n_1x_1 + n_2x_2 + ... + n_px_p}{n_1 + n_2 + ... + n_p} = \frac{1}{N}\sum_{i=1}^{p} n_ix_i.$$

Remarque . Lorsque la série est regroupée en classes, on calcule la moyenne en prenant pour valeurs x_i le **centre de chaque classe** ; ce centre est obtenu en faisant la moyenne des deux extrémités de la classe.

Exemple 6. ➔ Dans la **série A**, la moyenne du contrôle est égale à $\overline{m} = \dfrac{254}{30} \approx 8,47$.

➔ Dans la **série B**, une estimation du salaire moyen est donné par : $\overline{S} = \dfrac{460500}{280} \approx 1644,64$.

Remarque . On peut aussi calculer une moyenne à partir de la distribution de fréquences :

$$\overline{x} = f_1x_1 + f_2x_2 + \cdots + f_px_p = \sum_{i=1}^{p} f_ix_i.$$

Proposition 1.LINÉARITÉ DE LA MOYENNE

- Si on ajoute (ou soustrait) un même nombre k à toutes les valeurs d'une série, alors la moyenne de cette série se trouve augmentée (resp. diminuée) de k.
- Si on multiplie (ou divise) par un même nombre non nul k toutes les valeurs d'une série, alors la moyenne de cette série se trouve multipliée (resp. divisée) par k.

Exemple 7. On considère la **série A** :

- ➔ Si on ajoute $1,5$ points à chaque note du contrôle, alors la moyenne de classe devient $\overline{m} = 8,47 + 1,5 = 9,97$.
- ➔ Si on augmente chaque note de 10%, cela revient à multiplier chaque note par $1,1$, ce qui donne $\overline{m} = 8,47 \times 1,1 = 9,32$.

Proposition 2.MOYENNE PAR SOUS-GROUPES

Soit une série statistique, d'effectif total N, de moyenne $\overline{x}$.
Si on divise cette série en deux sous-groupes **disjoints** d'effectifs respectifs p et q (avec $p + q = N$) de moyennes respectives $\overline{x}_1$ et $\overline{x}_2$, alors on a :

$$\overline{x} = \frac{p}{N} \times \overline{x}_1 + \frac{q}{N} \times \overline{x}_2.$$

Exemple 8. On suppose par exemple que les 12 garçons de la classe de la **série A** ont obtenu une moyenne globale de 8 sur 20.

- ➔ La moyenne du groupe formé par les filles de la classe vérifie : $9,47 = \dfrac{12}{30} \times 8 + \dfrac{18}{30} \times \overline{m}_f$.
- ➔ Soit $\overline{m}_f = \dfrac{30}{18}\left(9,47 - \dfrac{12}{30} \times 8\right) = 10,45$.

2) Médiane

Définition. Soit une série statistique ordonnée dont les n valeurs sont $x_1 \leqslant x_2 \leqslant x_3 \leqslant \cdots \leqslant x_n$.
La médiane est un nombre M qui permet de diviser cette série en deux sous-groupes de même effectif.

- ➤ Si n est impair, n est la valeur de cette série qui est située au milieu, à savoir la valeur dont le rang est $\dfrac{n+1}{2}$, notée $x_{\frac{n+1}{2}}$.
- ➤ Si n est pair, n est le centre l'intervalle médian, qui est l'intervalle formé par les deux nombres situés « au milieu » de la série, à savoir $x_{\frac{n}{2}}$ et $x_{\frac{n}{2}+1}$.

Exemple 9. ➔ La médiane de la série « $2 - 5 - 6 - 8 - 9 - 9 - 10$ » est 8.

- ➔ La médiane de la série « $2 - 5 - 6 - 8 - 9 - 9$ » est 7.
- ➔ La médiane de la série « $2 - 5 - 6 - 6 - 9 - 10$ » est 6.

Exemple 10. On souhaite calculer la médiane de la **série A**.

- ➔ Pour cela, on commence par remplir le tableau des effectifs cumulés croissants :

Notes	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20
Eff.	0	1	2	1	1	2	3	5	6	2	3	0	2	0	1	1	0	0	0	0
ECC.	0	1	3	4	5	7	10	$\boxed{15}$	$\boxed{21}$	23	26	26	28	28	29	30	30	30	30	30

➔ Ensuite, l'effectif étant de 30, on chosit la moyenne entre la $15^{ième}$ et la $16^{ième}$ note.
On obtient $Med = \dfrac{8+9}{2} = 8,5$.

➔ Ce qui signifie que la moitié des notes est inférieure ou égale à $8,5$, et que l'autre moitié des notes est supérieure
ou égale à $8,5$.

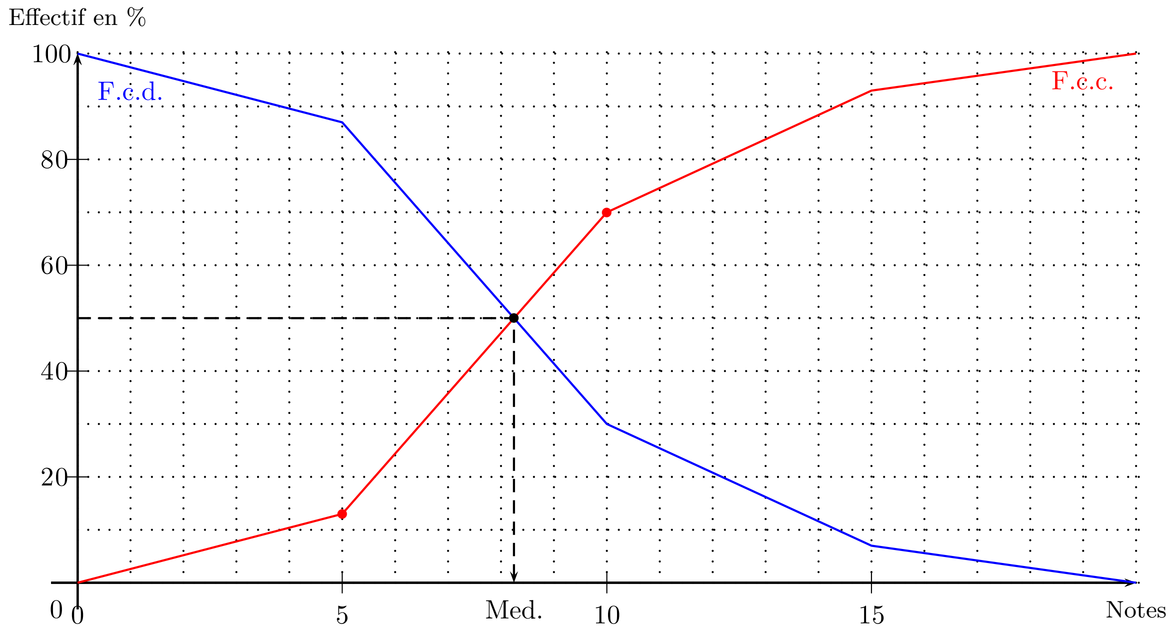
Dans le cas de répartition par classes, la médiane peut être évaluée soit graphiquement, soit par interpolation affine à l'aide d'un polygone des effectifs cumulés.

Exemple 11. On choisit la répartition par classes de la **série A** :

➔ On crée le tableau des fréquences cumulées croissantes et décroissantes.

Notes	[0 ; 5 [	[5 ; 10 [	[10 ; 15 [	[15 ; 20 [
Fréq. en %	13	57	23	7
F.c.c.	13	70	93	100
F.c.d.	87	43	7	0

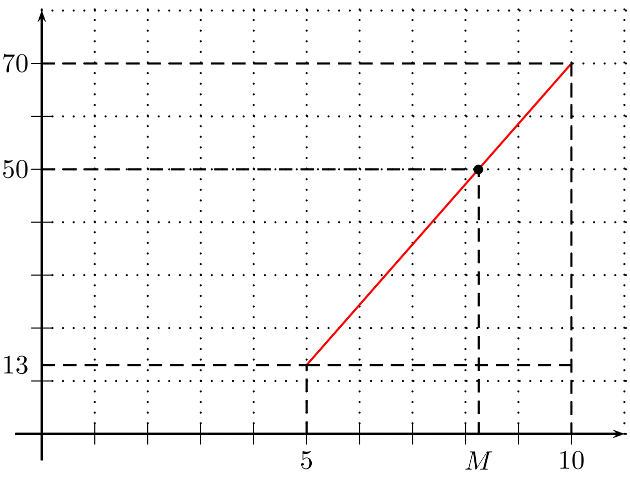
➔ Puis on place les points correspondants aux extrémités de chaque classe sur un graphique :



➔ On détermine le point du polygone d'ordonnée 50% et on trouver environ $8,2$.

➔ Pour trouver la médiane, on peut utiliser le polygone des fréquences cumulées décroissantes et lire l'abscisse ou bien du point de concours des deux polygones. On trouve aussi $8,2$.

➔ Enfin, par le calcul, 50% se situe dans l'intervalle $[\ 5\ ;\ 10\ [$.
On fait l'hypothèse que les longueurs des axes sont uniformément réparties dans cette classe.
On peut alors procéder à une interpolation linéaire d'après le théorème de Thalès :



$$\frac{M-5}{10-5} = \frac{50-13}{70-13} \iff \frac{M-5}{5} = \frac{37}{57} \iff M = 5 \times \frac{37}{57} + 5 = \frac{470}{57} \approx 8,25.$$

3) Quartiles, déciles ...

Définition. Soit une série statistique.

- On appelle quartiles de la série un triplet de réels (Q_1 ; Q_2 ; Q_3) qui sépare la série en quatre groupes de même effectif.
- On appelle déciles de la série un 9-uplet de réels (D_1 ; D_2 ; ... ; D_9) qui sépare la série en dix groupes de même effectif.

Remarque . Par définition, si X est une série statistique, $Q_2 = D_5 = Med(X)$.

Le calcul des valeurs des quartiles ou des déciles se fait en général à partir des graphiques des effectifs (ou fréquences) cumulés croissants, par interpolation linéaire.
La calculatrice donne les valeurs de Q_1, Med et Q_3.

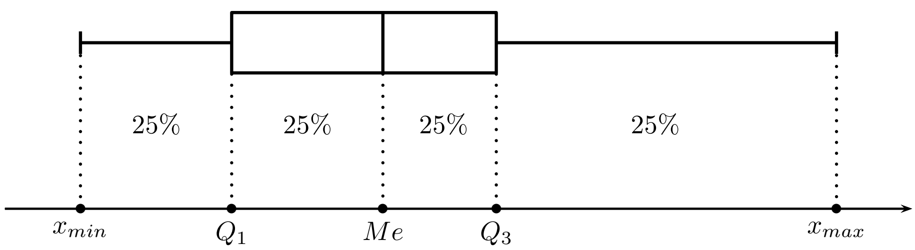
Exemple 12. ➔ Pour la **série A**, la calculatrice nous donne $Q_1 = 7$, $Méd = 8,5$ et $Q_3 = 10$.

➔ Graphiquement, on trouve $D_1 \approx 3,8$ et $D_9 \approx 14,2$.

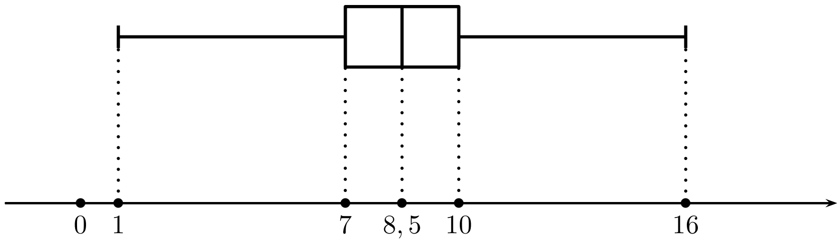
➔ Pour la **série B**, on trouve $Q_1 = 1500$, $Med = 1700$ et $Q_3 = 1900$.

4) Diagramme en boîtes :

Le diagramme en boîte affiche les quartiles et le valeurs extrémales comme suit :



Exemple 13. Pour la **série A** on otient :



III. Caractéristiques de dispersion

1) Étendue

Il s'agit de la première mesure de la dispersion d'une série statistique. Son principal mérite a longtemps été d'exister, et de fournir une information sur la dispersion très simple à obtenir.

> **Définition**. Soit X une série statistique discrète. On appelle étendue de la série le réel, défini par $Etd(X) = \max(X) - \min(X)$.

Exemple 14. L'étendue de la **série A** est de $16 - 2 = 14$.

2) Intervalle interquartille

> **Définition**. On appelle intervalle inter-quartiles l'intervalle $[\ Q_1\ ;\ Q_3\ ]$.
> L'amplitude de cet intervalle est appelée écart inter-quartiles.

Exemple 15. ➔ Dans la **série A**, l'intervalle interquartile est l'intervalle $[\ 7\ ;\ 10\ ]$ dont l'écart vaut $10 - 7 = 3$.

➔ Cet intervalle comprend donc la moitié des notes de la série située au centre de celle-ci.

3) Variance d'une série statistique

> **Définition**. La variance d'une série statistique est le nombre noté $V(x)$ obtenu comme moyenne des carrés des écarts constatés par rapport à la moyenne de la série :
>
> $$V(X) = \frac{n_1(x_1 - \bar{x})^2 + n_2(x_2 - \bar{x})^2 + ... + n_p(x_p - \bar{x})^2}{n_1 + n_2 + ... + n_p} = \frac{1}{N}\sum_{i=1}^{p} n_i(x_i - \bar{x})^2.$$

Remarque . Cette formule s'applique bien sûr au cas d'une série statistique sans coefficients : on est ramené à une série pour laquelle tous les coefficients valent 1.

Exemple 16. La variance de la **série B** vaut :
$V(X) = \dfrac{30(1050 - 1645)^2 + 30(1300 - 1645)^2 + ... + 40(2200 - 1645)^2}{280} \approx 109346.$

Proposition 3.

On utilise aussi la formule :

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

4) Écart-type d'une série statistique

Définition. L'écart-type d'une série statistique X, noté $\sigma(X)$, est la racine carrée de la variance de cette série :

$$\sigma(X) = \sqrt{V(X)}.$$

Exemple 17. L'écart-type de la **série B** vaut : $\sigma(X) = \sqrt{109561} = 331$.

Proposition 4.

La variance et l'écart-type présentent les propriétés suivantes :

- La variance et l'écart-type sont des nombres positifs ou nuls,
- Une variance nulle ou un écart-type nul signifient que toutes les valeurs de la série son égales à sa moyenne,
- Plus la variance (ou l'écart-type) d'une série est grande, plus cette série est dispersée autour de sa moyenne,