BURHAN

Résumé : Statistique

Tronc Commun · Sciences · Mathématiques

Version en ligne : burhan.ma/tc/sciences/mathematiques/statistique/resume

Chapitre	Semestre
Statistique	Semestre 2

Séries, effectifs et fréquences

À retenir : Ce que tu sais déjà

Tu l'as vu au collège : l'effectif d'une valeur compte les individus qui la prennent. La fréquence rapporte cet effectif au nombre total d'individus.

Au lycée, on garde ces deux mots et on les note : n_i pour un effectif, N pour le total, f_i pour une fréquence.

Avant tout calcul, dis qui est étudié et ce que l'on observe.

Définition : Population, individu, caractère

Une **population** est l'ensemble des personnes ou des objets sur lesquels porte une étude statistique ; chacun d'eux est un **individu**. Le **caractère** (ou variable statistique) est la propriété étudiée sur chaque individu.

- Le caractère est **qualitatif** lorsque ses modalités ne sont pas des nombres (groupe sanguin, couleur, moyen de transport).
- Il est **quantitatif** lorsque ses valeurs sont des nombres. Il est **discret** s'il ne prend que des valeurs isolées (nombre de frères et sœurs, pointure) ; il est **continu** s'il peut prendre toutes les valeurs d'un intervalle (taille, masse, durée).

La **série statistique** est la liste des valeurs, ou des modalités, du caractère relevées sur les individus de la population.

Définition : Effectif et fréquence

On considère une série statistique dont le caractère prend les valeurs (ou modalités) distinctes x_1, x_2, …, x_p.

- L'**effectif** n_i de la valeur x_i est le nombre d'individus pour lesquels le caractère prend la valeur x_i.
- L'**effectif total** est le nombre d'individus de la population : $N = n_1 + n_2 + \cdots + n_p$.
- La **fréquence** de la valeur x_i est le nombre $f_i = \dfrac{n_i}{N}$. On l'écrit sous forme de fraction ou de nombre décimal ; multipliée par 100, elle donne un pourcentage.

Propriété : Somme des fréquences

Avec les notations de la définition de l'effectif et de la fréquence, pour tout i de 1 à p, la fréquence f_i est comprise entre 0 et 1 :

$$0 \leq f_i \leq 1.$$

De plus, la somme des fréquences vaut 1 :

$$f_1 + f_2 + \cdots + f_p = 1.$$

Méthode : dresser le tableau d'une série

1. **Nomme** la population, puis le caractère, et dis s'il est qualitatif, quantitatif discret ou quantitatif continu.
2. **Range** les valeurs distinctes, par ordre croissant pour un caractère quantitatif, et compte combien de fois chacune apparaît.
3. **Vérifie** que le total des effectifs redonne N.
4. **Divise** chaque effectif par N : la somme des fréquences obtenues doit valoir 1, aux arrondis près.

Exemple

Énoncé. On a relevé le nombre de livres empruntés en un mois par 20 lecteurs d'une bibliothèque. Dresse le tableau des effectifs et des fréquences.

Nombre de livres x_i	0	1	2	3
Effectif n_i	3	7	6	4
Fréquence f_i	$0{,}15$	$0{,}35$	$0{,}3$	$0{,}2$

Contrôle. $3 + 7 + 6 + 4 = 20$, et $0{,}15 + 0{,}35 + 0{,}3 + 0{,}2 = 1$.

Nature. Le caractère étudié est quantitatif discret.

Attention : Une fréquence n'est pas un effectif

Dans le tableau des livres, la fréquence de la valeur 1 n'est pas 7 : c'est $\dfrac{7}{20} = 0{,}35$.

Une fréquence se calcule toujours sur l'effectif total ; elle est donc comprise entre 0 et 1. Si tu arrondis les fréquences, leur somme peut s'écarter un peu de 1.

Effectifs et fréquences cumulés

Définition : Effectifs et fréquences cumulés

On considère une série statistique d'un caractère quantitatif de valeurs x_1, x_2, ..., x_p, d'effectifs n_1, n_2, ..., n_p et d'effectif total N. Les valeurs sont rangées dans l'ordre croissant : $x_1 < x_2 < \cdots < x_p$. On note f_1, f_2, ..., f_p les fréquences.

- L'**effectif cumulé croissant** de la valeur x_i est $n_1 + n_2 + \cdots + n_i$: c'est le nombre d'individus pour lesquels le caractère est **inférieur ou égal** à x_i.
- L'**effectif cumulé décroissant** de la valeur x_i est $n_i + n_{i+1} + \cdots + n_p$: c'est le nombre d'individus pour lesquels le caractère est **supérieur ou égal** à x_i.
- On définit de même la **fréquence cumulée croissante** $f_1 + f_2 + \cdots + f_i$ et la **fréquence cumulée décroissante** $f_i + f_{i+1} + \cdots + f_p$ de la valeur x_i.

Méthode : calculer et lire des cumuls

1. **Ordonne** les valeurs, ou les classes, de la plus petite à la plus grande.
2. **Cumule** : chaque effectif cumulé croissant est le cumul précédent augmenté de l'effectif de la ligne.
3. **Contrôle** le dernier cumul croissant : il vaut N.
4. **Choisis la colonne** : une question « au plus » se lit sur les cumuls croissants, une question « au moins » sur les cumuls décroissants.

Exemple

Énoncé. Complète le tableau des livres empruntés par les cumuls.

Nombre de livres x_i	0	1	2	3
Effectif n_i	3	7	6	4
Effectif cumulé croissant	3	10	16	20
Effectif cumulé décroissant	20	17	10	4

Lecture. 16 lecteurs ont emprunté au plus 2 livres ; 10 en ont emprunté au moins 2.

La fréquence cumulée croissante de la valeur 1 vaut $\dfrac{10}{20} = 0{,}5$.

Remarque : Un contrôle gratuit

Le cumul croissant de la plus grande valeur vaut toujours N. Le cumul décroissant de la plus petite valeur aussi. Pour les fréquences, ces deux cumuls valent 1. Si ce n'est pas le cas, une addition est fausse.

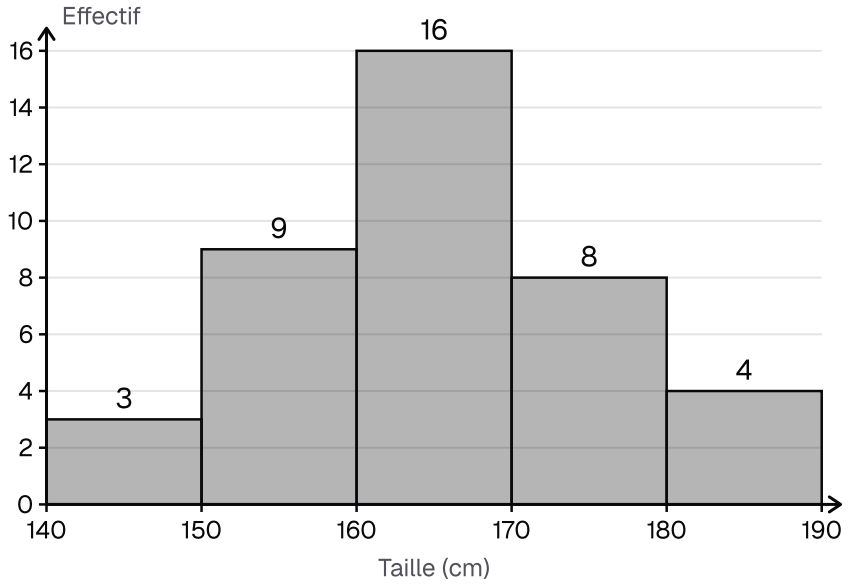
Classes, histogramme et polygone

Définition : Classes et histogramme

Lorsque le caractère est continu, ou prend de nombreuses valeurs, on regroupe ses valeurs en **classes** : des intervalles consécutifs $[a_0; a_1[$, $[a_1; a_2[$, ..., $[a_{p-1}; a_p[$, sans valeur commune. L'effectif d'une classe est le nombre d'individus dont la valeur appartient à cette classe.

- L'**amplitude** de la classe $[a; b[$ est le nombre $b - a$; son **centre** est le nombre $\frac{a+b}{2}$.
- L'**histogramme** de la série est formé, pour chaque classe $[a; b[$, d'un rectangle qui a pour base le segment $[a; b]$ de l'axe des abscisses et dont l'**aire est proportionnelle à l'effectif** de la classe.

Lorsque toutes les classes ont la même amplitude, on peut prendre pour hauteurs des rectangles les effectifs eux-mêmes.



Cas des classes de même amplitude, ici 10 cm pour les tailles de 40 élèves : la hauteur de chaque rectangle est l'effectif.

Méthode : construire un histogramme

1. **Calcule** l'amplitude $b - a$ de chaque classe $[a; b[$.
2. **Compare** les amplitudes. Si elles sont égales, chaque rectangle a pour hauteur l'effectif de sa classe.
3. **Sinon**, choisis une amplitude de référence. Divise l'effectif de chaque classe par le nombre d'amplitudes de référence qu'elle contient.
4. **Dessine** des rectangles accolés, chacun posé sur sa classe, et indique ce que représente une unité d'aire.

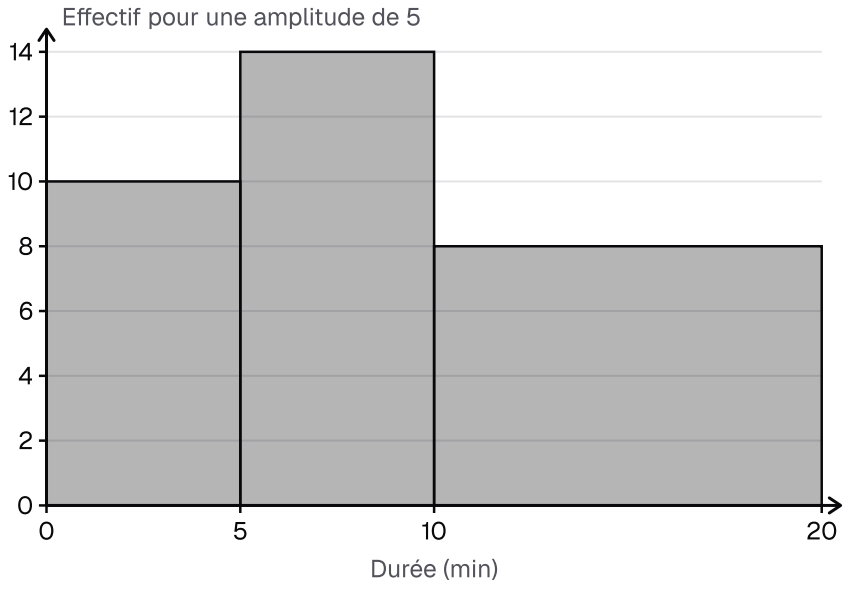
Exemple

Énoncé. Les durées de 40 appels reçus par un standard, en minutes, sont regroupées en trois classes. Calcule les hauteurs des rectangles de l'histogramme.

Durée (min)	$[0;5[$	$[5;10[$	$[10;20[$
Effectif	10	14	16
Amplitude	5	5	10
Hauteur	10	14	8

Amplitude de référence. On choisit 5 minutes. Unité d'aire : un rectangle de base 5 min et de hauteur 1 représente un appel.

Explication. La dernière classe contient deux amplitudes de référence : la hauteur de son rectangle est $\frac{16}{2} = 8$. Son aire représente bien 16 appels.



Les trois rectangles des durées d'appels, avec 5 minutes pour amplitude de référence : le dernier, deux fois plus large, ne monte qu'à 8.

Attention : Classe plus large, rectangle moins haut

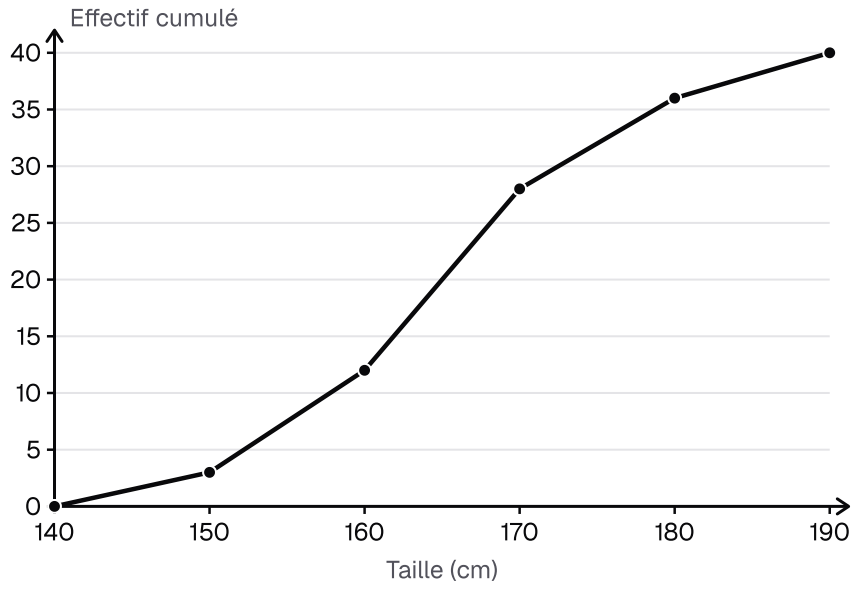
Pour les appels, un rectangle de hauteur 16 posé sur $[10;20[$ aurait une aire double : on croirait lire 32 appels. Quand une classe est plus large, sa hauteur diminue dans la même proportion. Avec des amplitudes inégales, on lit l'aire des rectangles, jamais leur seule hauteur.

Définition : Polygone des effectifs cumulés croissants

On considère une série regroupée en classes $[a_0; a_1[$, $[a_1; a_2[$, ..., $[a_{p-1}; a_p[$, d'effectif total N. Pour chaque i, on note N_i l'effectif cumulé croissant de la classe $[a_{i-1}; a_i[$: c'est le nombre d'individus dont la valeur est strictement inférieure à a_i.

Le **polygone des effectifs cumulés croissants** est la ligne brisée qui joint, dans l'ordre, les points de coordonnées $(a_0, 0)$, (a_1, N_1), (a_2, N_2), ..., (a_p, N_p), avec $N_p = N$.

En remplaçant les effectifs cumulés par les fréquences cumulées, on obtient le **polygone des fréquences cumulées croissantes**, qui se termine au point $(a_p, 1)$.



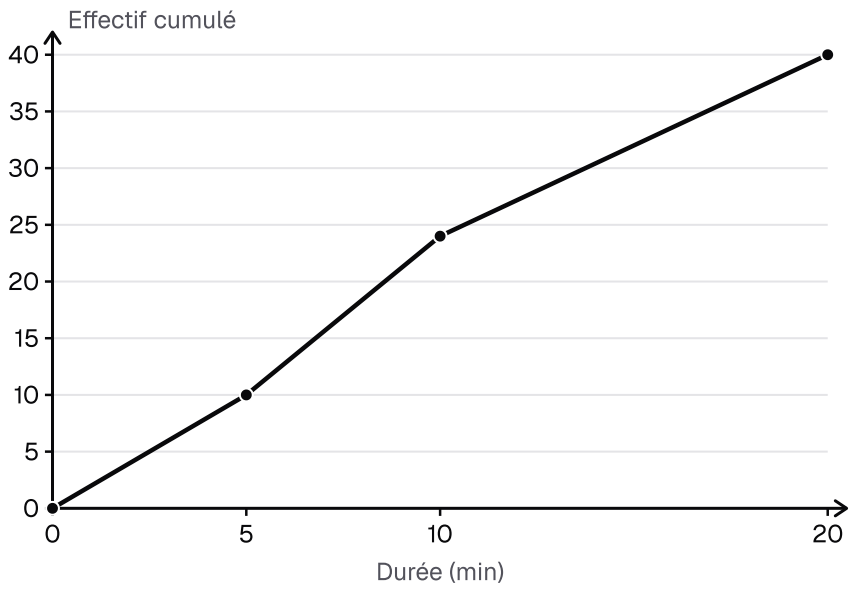
Ligne brisée des cumuls croissants pour les 40 tailles : elle part de la première borne au niveau 0 et finit à l'effectif total.

Exemple

Énoncé. Donne les points du polygone des effectifs cumulés croissants des durées d'appels.

Cumuls. 10 appels durent moins de 5 minutes, $10 + 14 = 24$ moins de 10 minutes, et $24 + 16 = 40$ moins de 20 minutes.

Points. On joint, dans l'ordre, $(0, 0)$, $(5, 10)$, $(10, 24)$ et $(20, 40)$. Le dernier point a pour ordonnée l'effectif total 40.



Polygone des appels cumulés : quatre sommets, de la durée 0 jusqu'à 20 minutes, où l'on retrouve les 40 appels.

La moyenne

Définition : Moyenne arithmétique

La **moyenne** d'une série statistique d'un caractère quantitatif de valeurs x_1, x_2, ..., x_p, d'effectifs n_1, n_2, ..., n_p et d'effectif total N est le nombre :

$$\overline{x} = \frac{n_1 x_1 + n_2 x_2 + \cdots + n_p x_p}{N}.$$

Lorsque la série est regroupée en classes, on calcule la moyenne en remplaçant chaque classe par son centre : on obtient une **valeur approchée** de la moyenne de la série.

Propriété : Moyenne et fréquences

Avec les notations de la définition de la moyenne, si f_1, f_2, ..., f_p sont les fréquences des valeurs x_1, x_2, ..., x_p, alors :

$$\overline{x} = f_1 x_1 + f_2 x_2 + \cdots + f_p x_p.$$

Propriété : Moyenne et changement affine des valeurs

Soit une série statistique d'un caractère quantitatif de valeurs x_1, x_2, ..., x_p, d'effectifs n_1, n_2, ..., n_p et d'effectif total N, de moyenne $\overline{x}$, et soit a et b deux réels.

On considère la série obtenue en remplaçant chaque valeur x_i par $y_i = a x_i + b$, avec les mêmes effectifs. Sa moyenne est :

$$\overline{y} = a\,\overline{x} + b.$$

Si $a = 0$, toutes les valeurs y_i sont égales à b : chacune compte encore avec l'effectif n_i, et $\overline{y} = b$.

Méthode : calculer une moyenne

1. **Ajoute** au tableau une ligne des produits $n_i x_i$; pour une série en classes, x_i désigne le centre de la classe.
2. **Fais la somme** de ces produits, puis divise-la par N.
3. **Avec de grandes valeurs**, pose d'abord $y_i = ax_i + b$ pour travailler sur de petits nombres. Calcule $\overline{y}$, puis retrouve $\overline{x}$ par $\overline{y} = a\,\overline{x} + b$.
4. **Contrôle** : une moyenne ne dépasse jamais la plus grande valeur et n'est jamais inférieure à la plus petite.

Exemple

Énoncé. Calcule le nombre moyen de livres empruntés par les 20 lecteurs.

Produits. On additionne les produits $n_i x_i$:

$$\begin{aligned} &3 \times 0 + 7 \times 1 \\ &+ 6 \times 2 + 4 \times 3 = 31. \end{aligned}$$

Moyenne. D'après la définition de la moyenne, $\overline{x} = \dfrac{31}{20} = 1{,}55$ livre.

Avec les fréquences. On retrouve le même nombre :

$$\begin{aligned} &0{,}15 \times 0 + 0{,}35 \times 1 \\ &+ 0{,}3 \times 2 + 0{,}2 \times 3 \\ &= 1{,}55. \end{aligned}$$

Exemple

Énoncé. Une machine remplit des bouteilles. Sur 10 bouteilles, on mesure en millilitres : 995 deux fois, 1000 cinq fois et 1005 trois fois. Calcule la contenance moyenne.

Changement. On pose $y_i = x_i - 1000$: les valeurs deviennent -5, 0 et 5. Donc :

$$\begin{aligned} \overline{y} &= \frac{2 \times (-5) + 5 \times 0 + 3 \times 5}{10} \\ &= \frac{5}{10} = 0{,}5. \end{aligned}$$

Retour. Ici $a = 1$ et $b = -1000$. Par la propriété du changement affine, $\overline{y} = \overline{x} - 1000$, donc $\overline{x} = 1000{,}5$ mL.

Attention : Chaque individu compte

Pour les livres, la moyenne des quatre valeurs distinctes vaut $\frac{0+1+2+3}{4} = 1{,}5$: ce n'est pas la moyenne de la série. Chacun des 20 lecteurs doit compter une fois. On multiplie donc chaque valeur par son effectif, et on trouve $1{,}55$.

Médiane et mode

Définition : Médiane

On range les N valeurs d'une série statistique, chacune répétée autant de fois que son effectif, dans l'ordre croissant.

- Si N est impair, la **médiane** M_e est la valeur de rang $\frac{N+1}{2}$.
- Si N est pair, la **médiane** M_e est la moyenne des valeurs de rangs $\frac{N}{2}$ et $\frac{N}{2}+1$.

La médiane partage ainsi la série ordonnée en deux groupes de même effectif.

Lorsque la série est regroupée en classes, on prend pour **valeur approchée** de la médiane l'abscisse du point du polygone des effectifs cumulés croissants dont l'ordonnée est $\frac{N}{2}$.

Méthode : déterminer une médiane

1. **Écris** les effectifs cumulés croissants et regarde si N est pair ou impair.
2. **Série de valeurs** : le rang du milieu est $\frac{N+1}{2}$ si N est impair ; les deux rangs du milieu sont $\frac{N}{2}$ et $\frac{N}{2}+1$ si N est pair. À chaque rang correspond la première valeur dont le cumul vaut au moins ce rang. Si N est pair, M_e est la moyenne des deux valeurs ainsi lues.
3. **Données groupées en classes** : note N_a et N_b les cumuls croissants aux bornes a et b, et trouve la classe $[a;b[$ telle que $N_a < \frac{N}{2} \leq N_b$. Calcule d'abord la part de la classe à parcourir, $t = \left(\frac{N}{2} - N_a\right) \div (N_b - N_a)$, puis $M_e = a + (b-a) \times t$.
4. **Vérifie** l'encadrement $a < M_e \leq b$ pour la classe trouvée.

Exemple

Effectif impair. Sept élèves ont lu ce soir-là 12, 15, 9, 20, 15, 11 et 18 pages.

Rangées : 9, 11, 12, 15, 15, 18, 20. Comme $N = 7$, la médiane est la valeur de rang 4 : $M_e = 15$ pages.

Effectif pair. Pour les 20 lecteurs, les rangs du milieu sont 10 et 11. Le cumul croissant vaut 10 à la valeur 1 et 16 à la valeur 2. La valeur de rang 10 est donc 1, celle de rang 11 est 2, et $M_e = \frac{1+2}{2} = 1{,}5$.

Exemple

Énoncé. Estime la durée médiane des 40 appels.

Classe. $\frac{N}{2} = 20$. Sur la classe $[5;10[$, les cumuls aux bornes vérifient $10 < 20 \leq 24$.

Part parcourue. $t = \frac{20-10}{24-10} = \frac{5}{7}$.

Interpolation. $M_e = 5 + 5 \times \frac{5}{7} = 5 + \frac{25}{7}$, soit environ $8{,}6$ minutes, arrondi au dixième. C'est une valeur approchée, lue sur le polygone des effectifs cumulés croissants.

Attention : Ranger d'abord

Dans la liste des pages lues, non rangée, la valeur placée au milieu est 20 : ce n'est pas la médiane, qui vaut 15. De même, $\frac{9+20}{2} = 14{,}5$ ne donne pas la médiane. Range toujours les valeurs avant de compter les rangs.

Définition : Mode et classe modale

Le **mode** d'une série statistique est une valeur (ou une modalité) du caractère dont l'effectif est le plus grand. Une série peut avoir plusieurs modes.

Lorsque la série est regroupée en classes, une **classe modale** est une classe dont le rectangle est le plus haut dans l'histogramme. Si toutes les classes ont la même amplitude, c'est une classe de plus grand effectif.

Exemple

Pour les livres, le mode est 1 : c'est la valeur de plus grand effectif, 7. Pour les appels, la classe modale est $[5;10[$, dont le rectangle a la hauteur 14. La classe $[10;20[$ a un effectif plus grand, 16, mais son rectangle est moins haut : elle n'est pas modale.

Étendue, écart moyen, variance, écart type

Définition : Étendue, écart moyen, variance, écart type

On considère une série statistique d'un caractère quantitatif de valeurs x_1, x_2, ..., x_p, d'effectifs n_1, n_2, ..., n_p et d'effectif total N, de moyenne $\overline{x}$.

- L'**étendue** est la différence entre la plus grande et la plus petite valeur de la série.
- L'**écart moyen** est la moyenne des distances des valeurs à la moyenne :

$$e = \frac{n_1|x_1 - \overline{x}| + \cdots + n_p|x_p - \overline{x}|}{N}.$$

- La **variance** est la moyenne des carrés des écarts à la moyenne :

$$V = \frac{\begin{matrix} n_1(x_1 - \overline{x})^2 + \dots \\ + n_p(x_p - \overline{x})^2 \end{matrix}}{N}.$$

- L'**écart type** est le nombre $\sigma = \sqrt{V}$; il s'exprime dans la même unité que les valeurs.

Propriété : Variance : moyenne des carrés moins carré de la moyenne

Avec les notations de la définition de la variance :

$$V = \frac{n_1 x_1^2 + \dots + n_p x_p^2}{N} - \overline{x}^{\,2}.$$

Autrement dit, la variance est la moyenne des carrés des valeurs, diminuée du carré de la moyenne.

Exemple

Énoncé. Cinq clients ont attendu à un guichet 4, 6, 7, 9 et 14 minutes. Calcule les paramètres de dispersion.

Moyenne et étendue. $\overline{x} = \dfrac{4+6+7+9+14}{5} = 8$ min ; l'étendue vaut $14 - 4 = 10$ min.

Écart moyen. Les distances à la moyenne sont 4, 2, 1, 1 et 6 : $e = \dfrac{14}{5} = 2{,}8$ min.

Variance et écart type. Par la définition :

$$\begin{aligned} V &= \frac{16 + 4 + 1 + 1 + 36}{5} \\ &= \frac{58}{5} = 11{,}6. \end{aligned}$$

Donc $\sigma = \sqrt{11{,}6} \approx 3{,}4$ min, arrondi au dixième.

Méthode : calculer une variance et un écart type

1. **Commence** par la moyenne $\overline{x}$.
2. **Complète** le tableau par une ligne des produits $n_i x_i^2$, puis fais-en la somme.
3. **Calcule** V : la moyenne des carrés, moins le carré de la moyenne. Si $\overline{x}$ est un nombre simple, la définition convient aussi.
4. **Termine** par $\sigma = \sqrt{V}$, dans l'unité des valeurs. Une variance négative signale une erreur.

Exemple

Même attente, par la formule. Les carrés des valeurs ont pour somme $16 + 36 + 49 + 81 + 196 = 378$. Par la formule de la variance :

$$\begin{aligned} V &= \frac{378}{5} - 8^2 \\ &= 75{,}6 - 64 = 11{,}6. \end{aligned}$$

On retrouve la valeur de la définition.

Série en tableau. Pour les livres, la somme des produits $n_i x_i^2$ vaut :

$$\begin{aligned} &3 \times 0^2 + 7 \times 1^2 \\ &+ 6 \times 2^2 + 4 \times 3^2 = 67. \end{aligned}$$

Donc :

$$\begin{aligned} V &= \frac{67}{20} - 1{,}55^2 \\ &= 3{,}35 - 2{,}4025 \\ &= 0{,}9475, \end{aligned}$$

et $\sigma \approx 0{,}97$ livre, arrondi au centième.

Attention : Le bon ordre, puis la racine

Pour les livres, calculer $1{,}55^2 - 3{,}35$ donne $-0{,}9475$: ce signe négatif trahit l'ordre inversé. La formule part de la moyenne des carrés et lui enlève le carré de la moyenne. N'oublie pas ensuite la racine carrée : $0{,}9475$ est la variance, et l'écart type vaut environ $0{,}97$ livre.

Comparer et choisir le bon paramètre

Méthode : comparer deux séries

1. **Position** : compare les moyennes, puis les médianes, exprimées dans la même unité.
2. **Dispersion** : compare les écarts types ; à défaut, les étendues ou les écarts moyens.
3. **Conclusion** : réponds dans les mots de la situation. Dis quelle série a les plus grandes valeurs, et laquelle est la plus régulière.

Exemple

Énoncé. Deux archers tirent chacun dix flèches. Les scores du premier ont pour moyenne $7{,}2$ points et pour écart type $0{,}8$ point. Ceux du second ont pour moyenne $7{,}2$ points et pour écart type $2{,}1$ points. Compare les deux séries.

Position. Les moyennes sont égales : en moyenne, les deux archers se valent.

Dispersion. $0{,}8 < 2{,}1$: les scores du premier sont moins dispersés.

Conclusion. Le premier archer est le plus régulier.

Méthode : Quel paramètre pour quelle question ?

- **Combien d'individus prennent une valeur ?** L'effectif, ou la fréquence pour une part du total.
- **Combien ont au plus, ou au moins, une valeur ?** Les effectifs cumulés.
- **Quelle valeur typique ?** La moyenne, la médiane ou le mode.
- **Quelle valeur partage la série en deux moitiés ?** La médiane.
- **Les valeurs sont-elles groupées ou étalées ?** L'étendue, l'écart moyen, l'écart type.
- **Quelle série est la plus régulière ?** Celle qui a le plus petit écart type.

Fréquence

$$f_i = \frac{n_i}{N}$$

Moyenne

$$\overline{x} = \frac{n_1 x_1 + n_2 x_2 + \cdots + n_p x_p}{N}$$

Moyenne par les fréquences

$$\overline{x} = f_1 x_1 + f_2 x_2 + \cdots + f_p x_p$$

Changement affine des valeurs, y = ax + b

$$\overline{y} = a\,\overline{x} + b$$

Part de la classe [a ; b[à parcourir, avec N_a < N/2 et N/2 au plus égal à N_b

$$t = \left(\frac{N}{2} - N_a\right) \div (N_b - N_a)$$

Médiane d'une série en classes, par interpolation

$$M_e = a + (b - a) \times t$$

Variance, par la formule

$$V = \frac{n_1 x_1^2 + \cdots + n_p x_p^2}{N} - \overline{x}^{\,2}$$

Écart type

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

À retenir : Rédiger en devoir

Nomme la définition ou la propriété que tu emploies : « d'après la définition de la médiane », « par la formule de la variance ». Annonce chaque arrondi. Précise qu'une valeur lue sur un graphique est approchée. Termine par une phrase dans les mots de la situation, avec l'unité.