

STATISTIQUES

La statistique est l'activité consistant à recueillir, exploiter et interpréter un ensemble de données, d'informations. Le traitement des données permet de produire des statistiques. L'objectif est de rendre lisible des données difficiles à interpréter à première vue.

I Caractéristiques de position

Leur objectif est de résumer toute la série en un seul nombre, le plus représentatif possible de la série. On parle aussi de « mesures centrales ».

1° Moyenne

Définition : La moyenne pondérée – ou moyenne **coefficientée** – d'une série statistique est notée $\bar{x}$ et

définie par $\bar{x} = \frac{\text{somme de toutes les valeurs coefficientées}}{\text{somme des coefficients}}$

Plus précisément, chaque valeur de la série est **multipliée par son coefficient / son effectif**, et on fait la somme de toutes ces valeurs. On divise alors le résultat par la somme de tous les coefficients.

Ex 1 : Calculer la moyenne trimestrielle de cet élève :

Note	14	10	19	11
Coeff.	2	2	1	3

$$\begin{aligned}\bar{x} &= \frac{14 \times 2 + 10 \times 2 + 19 \times 1 + 11 \times 3}{2 + 2 + 1 + 3} \\ &= \frac{100}{8} = 12,5\end{aligned}$$

La moyenne de cet élève est **12,5/20**.

Ex 2 : Calculer la moyenne des notes du contrôle ci-dessous :

Note	5	8	9	10	11	12	14	18	19
Effectif	1	4	5	9	6	3	4	2	1

$$\begin{aligned}\bar{x} &= \frac{5 \times 1 + 8 \times 4 + 9 \times 5 + 10 \times 9 + 11 \times 6 + 12 \times 3 + 14 \times 4 + 18 \times 2 + 19 \times 1}{1 + 4 + 5 + 9 + 6 + 3 + 4 + 2 + 1} \\ &= \frac{385}{35} = 11\end{aligned}$$

La moyenne des notes lors de ce contrôle est **11/20**.

⇒ Comment interpréter la moyenne ?

Le fait que la moyenne soit 11 signifie que **si tous les élèves avaient obtenu la même note, ils auraient tous obtenu 11/20**.

Parfois, les valeurs d'une série sont données sous forme de **classes** (c'est-à-dire d'intervalles). Pour calculer la moyenne, on prendra alors **les centres** de ces classes comme valeurs de la série.

Et dans le cas où une série est donnée sous forme de **pourcentages** (et non d'effectifs), on calcule la moyenne en **multipliant chaque valeur par son pourcentage et en additionnant le tout, et en divisant par 100 à la fin...**

Ex : Le tableau ci-dessous regroupe les tailles des élèves de seconde en cm par classes :

Tailles	[140 ; 160[	[160 ; 170[	[170 ; 180[	[180 ; 190[	[190 ; 210]
Pourcentages	6	18	51	22	3

Calculer la taille moyenne d'un élève de seconde.

⇒ On remplace la classe [140 ; 160[par **150**.

⇒ On remplace la classe [160 ; 170[par **165**.

$$\begin{aligned}\bar{x} &= (150 \times 6 + 165 \times 18 + 175 \times 51 + 185 \times 22 + 200 \times 3) : 100 \\ &= 17\,465 : 100 \\ &= 174,65\end{aligned}$$

En moyenne, un élève de seconde mesure **174,65 cm**.

Notation : si on note $x_1, x_2, \dots, x_k$ les valeurs de la série et $n_1, n_2, \dots, n_k$ les coefficients /effectifs correspondants, on a alors $\bar{x} = \frac{x_1 \times n_1 + x_2 \times n_2 + \dots + x_k \times n_k}{n_1 + n_2 + \dots + n_k}$

Propriété : Lorsqu'on multiplie toutes les valeurs d'une série statistique par un même nombre a , la moyenne de la nouvelle série **est aussi multipliée par a** .

Et lorsqu'on ajoute à toutes les valeurs d'une série statistique le même nombre b , la moyenne de la nouvelle série **est aussi augmentée de b** .

On parle de linéarité de la moyenne : la moyenne de la série de valeurs « $ax_i + b$ » est **$a\bar{x} + b$** .

Ex : Le salaire moyen dans une entreprise est de 1 300 € par mois. À la suite de forts bénéfices, le patron décide d'augmenter tous les salaires de 5% et d'ajouter une prime mensuelle permanente de 100 €. Quelle sera la moyenne des salaires après cette opération ?

Une augmentation de 5% revient à **multiplier par 1,05**.

La nouvelle moyenne sera alors $\bar{x} = 1,05 \times 1\,300 + 100 = 1\,465$ €

⇒ Quel est le défaut de la moyenne ?

La moyenne est **sensible aux valeurs extrêmes** qui peuvent en fausser le sens.

Le salaire moyen est alors $\bar{x} = \frac{1\,300 \times 90 + 8\,500 \times 9 + 16\,500 \times 1}{100} = 2\,100$

Lorsque les valeurs extrêmes sont douteuses ou excessivement éloignées des autres valeurs de la série, on peut décider de les retirer de l'étude : on parle alors de « moyenne élaguée ».

- au moins la moitié des valeurs de la série sont inférieures à la médiane
- au moins la moitié des valeurs de la série sont supérieures à la médiane

$$9 - 10 - 10 - 11 - 12 - 14 - 17$$

$m = 11$

9 - 10 - 11 - 12 - 14 - 17

$m = 11,5$

$$\left\lceil \frac{n}{2} \right\rceil + 1.$$

rang $\frac{n}{2}$ et la suivante.

$146:2 = 73$ donc la médiane est la moyenne entre la 73^{ème} et la 74^{ème} valeur.



médiane vaut donc $\frac{20 + 23}{2} = 21,5$.

⇒ Comment interpréter la médiane d'une série ?

Dire que la médiane des notes d'un contrôle est 11/20 signifie que **la moitié des élèves (au moins) a eu plus de 11/20 alors que l'autre moitié des élèves a eu moins de 11/20.**

Remarque : lorsqu'on dispose des fréquences (et non des effectifs), il faut les additionner jusqu'à atteindre 0,5 (ou 50% si elles sont en pourcentages).

La médiane est alors **la première valeur pour laquelle la « fréquence cumulée croissante » dépasse 0,5.**

3° Moyenne ou médiane : que choisir ?

Les deux ont leur intérêt, et leur comparaison va donner des informations sur la « forme » de la série statistique.

Si la moyenne et la médiane ont des valeurs équivalentes, on peut en déduire que la répartition des valeurs est a priori plutôt **symétrique**.

Ex : En France, le salaire moyen est de 2 238 € et le salaire médian est de 1 789 €. Que peut-on en déduire ?

Si **tous les français recevaient le même salaire, ils recevraient 2 238 € par mois.**

Or, **la moitié d'entre eux gagne moins de 1 789 €.**

Cet écart s'explique par **la présence de très hauts revenus qui tirent la moyenne vers le haut. Il y a donc des inégalités en haut de l'échelle.**

4° Avec la calculatrice

La calculatrice permet de vérifier l'ensemble des calculs des indicateurs précédents (et bien d'autres !)

1. Entrer la série dans la calculatrice :



- Aller dans « stats »
- Sélectionner « 1 : Modifier... »
- Entrer dans L₁ les valeurs de la série et dans L₂ les coefficients / effectifs.
 - ⇒ Pour effacer les valeurs de L₁ par exemple, aller sur L₁ (avec la flèche du haut) et appuyer sur « annul ».
- Quitter en faisant « 2nde » et « mode ».

2. Lancer les calculs des indicateurs :



- Aller dans « stats »
- Aller dans le menu « CALC » (flèche vers la droite)
- Sélectionner « 1 : Stats 1 Var ».
- Indiquer L₁ dans « Xliste : » (en faisant « 2nde » et la touche « 1 ») et L₂ dans « ListeFréq : »
- Valider « Calculer ».

II Caractéristiques de dispersion

Leur objectif est d'indiquer comment sont dispersées les données de la série. Autrement dit, on souhaite savoir si la série est **homogène** ou **hétérogène**.

1° Etendue

Définition : l'étendue d'une série statistique est la différence entre la plus grande valeur et la plus petite valeur de la série.

L'étendue donne donc l'écart maximal entre des valeurs de la série : lorsque l'étendue est faible, la série est **homogène**, et lorsque l'étendue est importante, la série est **hétérogène**.

Ex : l'étendue de la série 04 / 05 / 09 / 10 / 12 / 12 / 15 / 16 est $16 - 4 = 12$.

Il y a 12 points d'écarts entre la meilleure note et la moins bonne.

Ex : si l'étendue d'un contrôle est 7, on peut considérer que les résultats sont **relativement homogènes**.

Les notes peuvent toutefois aller de 0 à 8 ou de 12 à 20 donc...

2° Quartiles

Il s'agit de valeurs précisant la dispersion : là où la médiane coupe la série en **deux séries**, les quartiles la coupent en **quatre**.

Définition : le 1^{er} quartile noté Q_1 est une valeur telle que **au moins un quart des valeurs de la série y sont inférieures et au moins trois quarts y sont supérieures**.

Le 3^{ème} quartile noté Q_3 est une valeur telle que **au moins trois quarts des valeurs de la série y sont inférieures et au moins un quart y sont supérieures**.

Remarques : 1. Q_2 est la médiane !

2. Il y a environ **25%** des valeurs entre la plus petite valeur et Q_1 , **25%** des valeurs entre Q_1 et la médiane, **25%** des valeurs entre la médiane et Q_3 et **25%** des valeurs entre Q_3 et la plus grande valeur.

↪ Si n n'est pas un multiple de 4, Q_1 est la valeur de la série de rang $\left\lceil \frac{n}{4} \right\rceil + 1$.

↪ Si n est un multiple de 4, Q_1 est la moyenne entre la valeur de la série de rang $\frac{n}{4}$ et la valeur suivante.

Même chose pour Q_3 mais en prenant $3 \times \frac{n}{4}$ au lieu de $\frac{n}{4}$

Ex : 1. Déterminer le rang de Q_1 et de Q_3 dans une série de 35 notes.

$$\frac{35}{4} = 8,75 \text{ donc } Q_1 \text{ est la } 9^{\text{ème}} \text{ note.}$$

$$3 \times \frac{35}{4} = 26,25 \text{ donc } Q_3 \text{ est la } 27^{\text{ème}} \text{ note.}$$

2. Déterminer le rang de Q_1 et de Q_3 dans une série de 36 notes.

$$\frac{36}{4} = 9 \text{ donc } Q_1 \text{ est entre la } 9^{\text{ème}} \text{ et la } 10^{\text{ème}} \text{ note.}$$

$$3 \times \frac{36}{4} = 27 \text{ donc } Q_3 \text{ est entre la } 27^{\text{ème}} \text{ et la } 28^{\text{ème}} \text{ note.}$$

Définition : L'écart interquartile est la différence $Q_3 - Q_1$.

On parle aussi d'intervalle interquartile pour désigner l'intervalle $[Q_1 ; Q_3]$. Cet intervalle contient 50% des valeurs les plus centrales de la série.

Remarque : lorsqu'on ne dispose que des fréquences, Q_1 correspond à la première fréquence cumulée croissante qui dépasse 0,25 et Q_3 à la première qui dépasse 0,75.

3° Ecart-type

Définition : considérons une série statistique dont les valeurs sont $x_1, x_2, \dots, x_k$ avec pour coefficients / effectifs respectifs $n_1, n_2, \dots, n_k$.

La variance V de cette série est définie par :

$$V = \frac{n_1 \times (x_1 - \bar{x})^2 + n_2 \times (x_2 - \bar{x})^2 + \dots + n_k \times (x_k - \bar{x})^2}{n_1 + n_2 + \dots + n_k}$$

L'écart-type σ d'une série statistique de variance V est défini par $\sigma = \sqrt{V}$.

L'écart-type exprime la dispersion des valeurs d'une série statistique autour de sa moyenne : plus il est important, et plus la série est hétérogène.

Remarque : 1. Très précisément, la variance est la moyenne des carrés des écarts entre les valeurs de la série et la moyenne $\bar{x}$: le fait de mettre les écarts au carré les rend positifs et accentue l'impact des valeurs extrêmes, ce qui permet de mieux faire ressortir l'éventuelle hétérogénéité.

2. Si on étudie une série composée de mesures en mètres par exemple, la variance a pour unité le mètre carré, ce qui n'est pas très cohérent... L'écart-type permet alors de revenir dans l'unité de départ car la racine carrée vient « annuler » les carrés de la variance.

Ex : Voilà la répartition des moyennes d'un groupe d'élèves :

Moyenne	[0 ; 6[	[6 ; 9[	[9 ; 11[	[11 ; 13[	[13 ; 16[	[16 ; 20]
Effectif	2	6	12	7	6	2

Calculer l'écart-type de ce groupe.

$$\bar{x} = \frac{3 \times 2 + 7,5 \times 6 + 10 \times 12 + 12 \times 7 + 14,5 \times 6 + 18 \times 2}{2 + 6 + 12 + 7 + 6 + 2}$$

$$= \frac{378}{35} = 10,8$$

$$V = (2 \times (3 - 10,8)^2 + 6 \times (7,5 - 10,8)^2 + 12 \times (10 - 10,8)^2 + 7 \times (12 - 10,8)^2 +$$

$$6 \times (14,5 - 10,8)^2 + 2 \times (18 - 10,8)^2) : 35$$

$$= \frac{390,6}{35} = 11,16$$

$$\sigma = \sqrt{11,16} \approx 3,34$$

L'écart-type est de 3,34 points.

Remarque : il serait intéressant de comparer cet écart-type avec un autre groupe d'élèves de manière à avoir une interprétation plus intéressante...

Ici, on peut partir du principe que les élèves entre $\bar{x} - \sigma$ et $\bar{x} + \sigma$ sont « dans la norme ». Autrement dit, les élèves entre 7,46 et 14,14 ont une moyenne « normale ».

Pour avoir un ordre de grandeur, on peut considérer que 68% des élèves sont dans l'intervalle $\bar{x} - \sigma$ et $\bar{x} + \sigma$. Et de la même manière, on peut considérer que 95% des élèves sont dans l'intervalle $\bar{x} - 2\sigma$ et $\bar{x} + 2\sigma$ (autrement dit les élèves ayant une moyenne entre 4,12 et 17,48)

Pour un laboratoire devant doser avec précision la molécule d'un médicament, on pourra considérer que les médicaments dont le dosage est compris entre $\bar{x} - \sigma$ et $\bar{x} + \sigma$ sont totalement conformes. Ceux dont le dosage ira au maximum jusqu'à $\bar{x} - 2\sigma$ ou $\bar{x} + 2\sigma$ seront acceptables ; mais ceux allant jusqu'à $\bar{x} - 3\sigma$ ou $\bar{x} + 3\sigma$ déclencheront une alerte qualité ; au-delà, les médicaments seront non conformes et donc détruits.

III Prudence, prudence...

Lors de l'interprétation des indicateurs d'une série statistique, il convient de rester prudent...

Considérons les trois séries suivantes :

S_1 : 2 6 6 6 6 10 10 10 10 10 10 10 10 10 14 14 14 14 18
 S_2 : 2 3 4 5 6 7 8 8 9 10 11 12 12 13 14 15 16 17 18
 S_3 : 2 2 2 2 6 6 6 6 6 10 14 14 14 14 14 18 18 18 18

	$\bar{x}$	Min	Q_1	Méd.	Q_3	Max	Etendue	σ
S_1	10	2	6	10	14	18	16	3,67
S_2	10	2	6	10	14	18	16	4,68
S_3	10	2	6	10	14	18	16	5,95

Morale de l'histoire : l'interprétation des statistiques doit toujours rester prudente !