Statistiques descriptives. Analyse de données.

I. Moyenne

a) Moyenne et moyenne pondérée

DÉFINITION

Soit une série statistique de p valeurs, $x_1, x_2, \ldots, x_p$ d'effectifs respectifs $n_1, n_2, \ldots, n_p$ donnés dans le tableau ci-dessous.

Valeur	x_1	x_2	...	x_p
Effectif	n_1	n_2	...	n_p

La **moyenne pondérée** de la série statistique ci-contre est le nombre réel, noté $\bar{x}$, tel que :

$$\bar{x} = \frac{n_1x_1 + n_2x_2 + \cdots + n_px_p}{n_1 + n_2 + \cdots + n_p}$$

Exemple

Audrey prend souvent le train venant de Rouen en direction de Paris. En rentrant dans le wagon, elle compte le nombre de places assises disponibles.
Après 20 trajets, elle obtient les résultats ci-contre.

Valeur	0	1	2	5	6	7	10
Effectif	5	1	3	1	5	4	1

Le nombre moyen de places assises disponibles sur ces 20 trajets est :

$$\frac{5 \times 0 + 1 \times 1 + 3 \times 2 + 1 \times 5 + 5 \times 6 + 4 \times 7 + 1 \times 10}{5 + 1 + 3 + 1 + 5 + 4 + 1} = 4$$

Remarque

Dans la propriété précédente, en posant $N = n_1 + n_2 + \cdots + n_p$, on obtient la fréquence des x_i : $f_1 = \frac{effectif\ de\ x_1}{effectif\ total} = \frac{n_1}{N}$; celle de x_2 : $f_2 = \frac{n_2}{N}$, etc.
On en déduit donc la formule

$$m = \frac{n_1}{N} \times x_1 + \frac{n_2}{N} \times x_2 + \cdots + \frac{n_p}{N} \times x_p = f_1 \times x_1 + f_2 \times x_2 + \cdots + f_p \times x_p$$

Propriété - Moyenne pondérée

On considère une série statistique constituée de p valeurs $x_1, x_2, \ldots, x_p$ affectées de p coefficients (ou poids) $c_1, c_2, \ldots, c_p$.
La moyenne pondérée de cette série est

$$m = \frac{c_1x_1 + c_2x_2 + \cdots + c_px_p}{c_1 + c_2 + \cdots + c_p}$$

Remarque

La formule est la même que pour la moyenne d'une série donnée sous forme de tableau d'effectifs, mais il est important de comprendre que les séries sont de natures différentes :

- dans le 1er cas (avec un tableau d'effectifs) la série est constituée de $n_1 + n_2 + \cdots + n_p$ valeurs, précisément : $\overbrace{x_1\ ;\ x_1\ ;\ \ldots\ ;\ x_1}^{n_1\ fois}\ ;\ \overbrace{x_2\ ;\ x_2\ ;\ \ldots\ ;\ x_2}^{n_2\ fois}\ ;\ \ldots\ ;\ \overbrace{x_p\ ;\ x_p\ ;\ \ldots\ ;\ x_p}^{n_p\ fois}$

- dans le 2ème cas (avec les coefficients ou poids) la série est constituée de p valeurs $x_1, x_2, \ldots, x_p$ (éventuellement identiques) auxquelles on attribue un coefficient (ou poids) $c_1, c_2, \ldots, c_p$ qui correspond à l'importance de la valeur.

Exemple

Ce trimestre, Émilia a eu quatre contrôles de mathématiques (notés sur 20) de coefficients 1 ; 1,5 ; 4 et 0,5 auxquels elle a obtenu respectivement les notes 8 ; 9 ; 20 et 5.

Sa moyenne en mathématiques est donc

$$m = \frac{1 \times 8 + 1{,}5 \times 9 + 4 \times 20 + 0{,}5 \times 5}{1 + 1{,}5 + 4 + 0{,}5} \approx 14{,}9$$

Remarque

On constate que, bien qu'elle ait eu 3 notes sur 4 en dessous de 10, sa moyenne est bonne : ceci est dû au fait qu'elle ait eu une très bonne note (20) à un devoir ayant un grand « poids » (donc une plus grande importance) par rapport aux autres.

Exercice résolu 1 page 294

↪ Exercices 17 à 22 p. 296

b) Linéarité de la moyenne

Propriété – Linéarité de la moyenne

Soit a et b deux nombres réels et x_1 ; x_2 ; … ; x_n une série statistique de moyenne m.

- Si on multiplie par a toutes les valeurs de la série, on obtient la moyenne de la nouvelle série en multipliant par a la moyenne de la série de départ.
  Autrement dit, la moyenne de la série $ax_1, ax_2, \ldots, ax_n$ est am.
- Si on ajoute b à toutes les valeurs de la série, on obtient la moyenne de la nouvelle série en ajoutant b à la moyenne de la série de départ.
  Autrement dit, la moyenne de la série $x_1 + b, x_2 + b, \ldots, x_n + b$ est $m + b$.
- Les deux points précédents assurent également que la moyenne de la série $ax_1 + b, ax_2 + b, \ldots, ax_n + b$ est $am + b$.

Une démonstration de ce résultat se trouve à la page 291 (Sésamath)

Exemple

La semaine dernière, pour se préparer le matin, Juan a mis 20 minutes en moyenne :

- S'il avait mis deux minutes de plus chaque jour, il aurait mis en moyenne $20 + 2 = 22$ minutes pour se préparer ;
- S'il avait mis 5 % de temps en plus chaque jour, c'est-à-dire si son temps de préparation avait été multiplié par $1 + \frac{5}{100} = 1{,}05$ chaque jour, alors il aurait mis en moyenne $20 \times 1{,}05 = 21$ minutes pour se préparer.

Remarque

La propriété de linéarité de la moyenne reste vraie lorsque :

- on soustrait un même nombre à toutes les valeurs de la série, puisque soustraire un nombre est équivalent à ajouter son opposé ;

- on divise toutes les valeurs de la série par un même nombre, puisque diviser par un nombre est équivalent à le multiplier par son inverse.

Exemples

① Si l'on soustrait 5 à toutes les valeurs d'une série statistique de moyenne m_1 alors cela revient à ajouter -5 à toutes ses valeurs. Ainsi, la moyenne de la nouvelle série est $m_1 + (-5) = m_1 - 5$.

② Si l'on divise par 4 toutes les valeurs d'une série statistique de moyenne m_2 alors cela revient à multiplier par $\frac{1}{4}$ toutes ses valeurs. Ainsi, la moyenne de la nouvelle série est $m_2 \times \frac{1}{4} = \frac{m_2}{4}$.

II. Écart-type

Définition - Écart-type

L'écart-type s d'une série statistique est un **indicateur de dispersion** de cette série statistique autour de la moyenne. Concrètement il donne une certaine mesure de l'écart entre les valeurs de la série et la moyenne de celle-ci :

- plus l'écart-type s d'une série est petit, plus les valeurs de la série sont concentrées autour de la moyenne, donc plus la série est homogène ;
- plus l'écart-type s d'une série est grand, plus les valeurs de la série sont éloignées de la moyenne, donc moins la série est homogène.

Remarque

Nous utiliserons la calculatrice pour déterminer l'écart-type (voir Tuto Vidéo et le TP1) mais il existe des formules pour le calculer (pour info : livre page 292).

Exemple

On considère deux entreprises de 10 employés :

- l'entreprise 1 dans laquelle 5 employés gagnent 2 500 € et 5 employés gagnent 3 500 € par mois ;
- l'entreprise 2 dans laquelle 9 employés gagnent 1 200 € et 1 employé gagne 19 200 € par mois.

Le salaire moyen dans l'entreprise 1 est de $\frac{5\times2500+5\times3500}{10} = 3000$ € et le salaire moyen dans l'entreprise 2 est de $\frac{9\times2500+5\times3500}{10} = 3000$ € également.

On comprend pourtant bien que la répartition des salaires dans les deux entreprises est extrêmement différente : la moyenne seule ne fournit pas une information suffisante pour résumer la série de manière satisfaisante.

On utilise alors un indicateur permettant de mesurer l'homogénéité des salaires dans les deux entreprises : l'écart-type.

Avec la calculatrice, on obtient :

- l'écart-type de la série des salaires de l'entreprise 1 , qui est de 500 € ;
- l'écart-type de la série des salaires de l'entreprise 2, qui est de 5 400 €.

On constate donc que, bien que le salaire moyen soit le même dans les deux entreprises, 3 000 € :

- dans l'entreprise 1, les employés ont globalement des salaires proches de cette moyenne ;
- dans l'entreprise 2, les employés ont globalement des salaires éloignés de cette moyenne.

Remarques

- On peut utiliser le couple (moyenne ; écart-type) pour résumer une série et en comparer plusieurs.
- Les valeurs éloignées de la moyenne ont de l'influence sur l'écart-type, plus précisément elles le font augmenter.

- Pour les séries dont le diagramme en bâtons (ou l'histogramme) est en forme de cloche, on peut s'attendre à trouver l'essentiel des valeurs de la série (environ 95 %) dans l'intervalle $[m - 2s\ ;\ m + 2s]$ où m désigne la moyenne et s l'écart-type de la série (voir le TP2)

Exercice résolu 2 page 294

↪ Exercices 27 à 32 p. 297

III. Quartiles et écart interquartile

DÉFINITIONS - Quartiles

Les valeurs d'une série statistique étant **rangées par ordre croissant**

- **le premier quartile** est la plus petite valeur Q_1 de la série telle qu'**au moins 25 %** des valeurs de la série sont **inférieures ou égales à Q_1** ,
- **le troisième quartile** est la plus petite valeur Q_3 de la série telle qu'**au moins 75 %** des valeurs de la série sont **inférieures ou égales à Q_3**.

Exemple

On considère la série ordonnée de 9 valeurs 1 ; 3 ; 7 ; 8 ; 10 ; 11 ; 12 ; 12 ; 58. On a alors :
Plus de 25 % des valeurs : $25\% \times 9 = 2{,}25$ donc 3 valeurs.
Moins de 25 % des valeurs : $25\% \times 9 = 2{,}25$ donc 2 valeurs.

$$1\ ;\ \underbrace{3}_{2ème}\ ;\ \overbrace{7}^{3ème}\ ;8\ ;10\ ;11\ ;12\ ;12\ ;58\ donc\ Q_1 = 7$$

Plus de 75 % des valeurs : $75\% \times 9 = 6{,}75$ donc 7 valeurs.
Moins de 75 % des valeurs : $75\% \times 9 = 6{,}75$ donc 6 valeurs.

$$1\ ;\ 3\ ;7\ ;8\ ;10\ ;\ \underbrace{11}_{6ème}\ ;\ \overbrace{12}^{7ème}\ ;12\ ;58\ donc\ Q_3 = 12$$

Propriété - Rang des quartiles

Pour une série ordonnée d'effectif n, Q_1 (resp. Q_3) est la k-ième valeur où k est le plus petit entier supérieur ou égal à $\frac{n}{4}$ (resp. $\frac{3n}{4}$).

Exemple

On reprend la série du nombre de places disponibles dans un train sur 20 trajets du paragraphe 1. a.

Valeur	0	1	2	5	6	7	10
Effectif	5	1	3	1	5	4	1

- Pour trouver Q_1, on calcule $\frac{n}{4} = \frac{20}{4} = 5$ donc Q_1 est la 5ème valeur, c'est-à-dire $Q_1 = 0$ (car la série est $\underbrace{0\ ;\ 0\ ;0\ ;0\ ;0}_{5\ valeurs}\ ;1\ ;2\ ;2\ ;2\ ;etc$)
- Pour trouver Q_3, on calcule $\frac{3n}{4} = \frac{3\times20}{4} = 15$ donc Q_3 est la 15ème valeur, c'est-à-dire $Q_3 = 6$.

Propriété - Médiane

Pour une série ordonnée d'effectif n, la médiane est :

- la valeur de rang $\frac{n}{2} + 0{,}5$ si n est impair ;

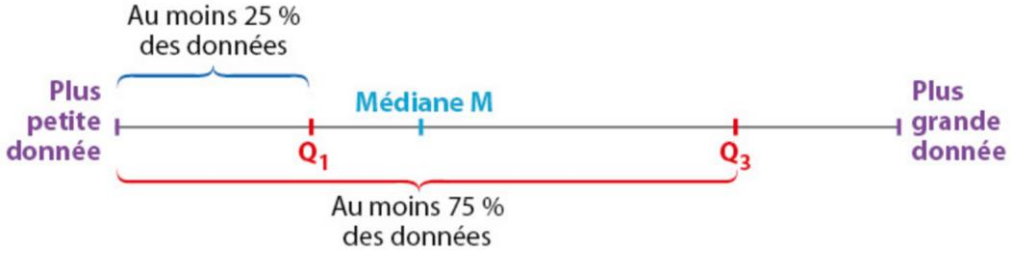
- la moyenne des valeurs de rang $\frac{n}{2}$ et $\frac{n}{2}+1$ si n est pair.

Exemple

Dans l'exemple précédent, la série a pour effectif $n = 20$ qui est pair. $\frac{20}{2} = 10$ et $\frac{20}{2}+1 = 11$ donc la médiane est la moyenne des 10ème et 11ème valeurs, c'est-à-dire $\frac{5+6}{2} = 5{,}5$

Remarque

Pour une série statistique de valeur minimale $x_{\min}$ et de valeur maximale x_{max}, chacun des intervalles $[x_{\min}\ ;\ Q_1]$, $[Q_1\ ;\ médiane]$, $[médiane\ ;\ Q_2]$ et $[Q_2\ ;\ x_{max}]$ contient au moins 25 % des valeurs de la série (et environ 25 % si la série est de grand effectif et constituée essentiellement de valeurs différentes).



Définition - Écart interquartile

L'écart interquartile d'une série statistique est $Q_3 - Q_1$. Il s'agit d'un indicateur de dispersion.

Exercice résolu 3 page 295

↪ Exercices 33 à 37 p. 297

Remarques

- Plus l'écart interquartile est petit, plus les valeurs « centrales » de la série (celles dans l'intervalle $[Q_1\ ;\ Q_3]$) sont proches les unes des autres. Les valeurs supérieures à Q_3 ou inférieures à Q_1 n'ont pas d'influence sur l'écart interquartile.
- On peut utiliser le couple (médiane ; écart interquartile) pour résumer une série et en comparer plusieurs.