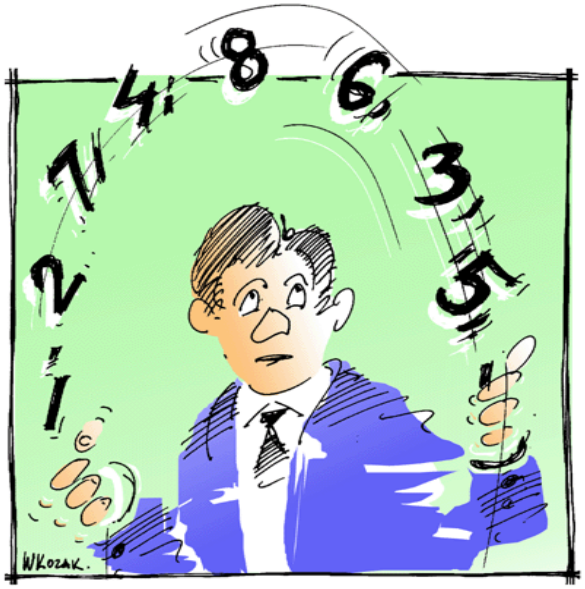
Statistique descriptive, analyse de données



(moyenne ; écart type)

ou

(médiane ; écart interquartile)

???

Effets de structure

La classe de seconde est l'occasion:

- d'une part de consolider l'utilisation des fonctions statistiques des calculatrices
- d'autre part de traiter, à l'aide d'un tableur, des séries statistiques riches et variées comportant un grand nombre de données brutes, en lien avec des situations réelles ou avec d'autres disciplines …

- La détermination de la médiane nécessite un tri des données.

- Dans la mesure du possible, il faut éviter de calculer une moyenne ou une médiane après un regroupement des données en classes.

(moyenne, écart type) / (médiane, écart interquartile)

« Il n'existe pas de règle, au sens mathématique, indiquant quel type d'indicateur utiliser par rapport à une situation donnée. Le choix des indicateurs dépend de ce qu'on veut en faire et de la réalité de la situation. On peut juste proposer quelques remarques qui permettent de privilégier tel couple plutôt que tel autre.

Deux séries statistiques de même écart type, même moyenne, même médiane, peuvent avoir des distributions très différentes ; dans ce cas, un graphique peut être plus parlant qu'un simple résumé numérique. »

On se donne une série statistique de type quantitatif, que l'on veut résumer par un couple d'indicateurs : un **indicateur de position** et un **indicateur de dispersion**.

L'indicateur de position réalise la « plus courte distance » à la série, l'indicateur de dispersion associé est cette distance minimale.

Mais quelle distance ?

On appelle $x_1, x_2, \ldots, x_n$ les valeurs de la série.

x est un réel quelconque.

On peut considérer les distances d_1, d_2, d_∞ définies par :

$$d_1(x) = \frac{1}{n}\sum_{i=1}^{n} |x_i - x| \quad ; \quad d_2(x) = \frac{1}{n}\sum_{i=1}^{n} (x_i - x)^2$$

$$d_\infty(x) = \max_{1 \leq i \leq n} |x_i - x|$$

On montre facilement que, pour tout réel *x* :

$$d_2(\bar{x}) \leq d_2(x)$$

(ou on étudie les variations de la fonction).

La distance minimale est donc atteinte en $x = \bar{x}$

et le minimum correspondant est la variance.

La moyenne est donc « naturellement » associée à l'écart type.

euler – Ress. 69 (outil) euler.ac-versailles.fr/wm3/pi2/moyenne/arithmetique1.jsp

Complétez le tableau d'effectifs de la série statistique S.

Modalité				
Effectif				

(Tableau à compléter par l'utilisateur)

Soit S la série statistique dont le tableau des effectifs est le suivant.

Modalité	1	2	3	4
Effectif	4	3	2	1

Soit f la fonction donnant la somme des carrés des distances d'un nombre à chacun des termes de S.
f est définie pour tout $x \in [1;4]$ par

$$f(x) = (x-4)^2 + 2(x-3)^2 + 3(x-2)^2 + 4(x-1)^2$$

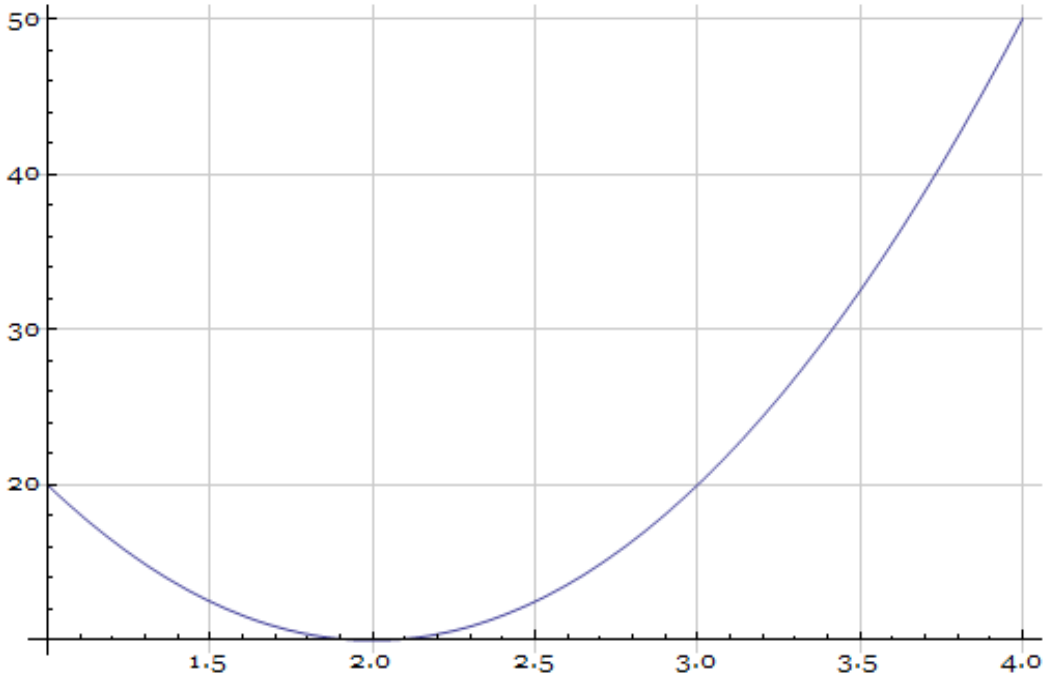
et, pour tout $x \in [1;4]$,

$$f(x) = 10x^2 - 40x + 50$$

$$f(x) = 10(x-2)^2 + 10$$

Le minimum de f est atteint en 2 qui représente la moyenne arithmétique de S ; le minimum de f, égal ici à 10, divisé par l'effectif total 10, soit 1 est appelé *variance* de S.

Une représentation graphique de f sur $[1;4]$ est donnée ci-contre.



Pour la distance d_1 : on commence par ordonner les valeurs de la série.

On supposera $x_1 \leq x_2 \leq \ldots \leq x_n$

On distingue deux cas :

- si n est impair, $n = 2p+1$, le minimum est atteint en x_{p+1}
- si n est pair, $n = 2p$, le minimum est atteint en tout point de l'intervalle $[x_p ; x_{p+1}]$

$\rightarrow$ définition de la médiane

Définition lexique euler

Définition

Soit a une série statistique quantitative discrète à une variable de taille $n \in \mathbb{N}^*$ définie par $a = \{a_i\}_{1 \leq i \leq n}$, *ordonnée dans l'ordre croissant*.

- Si n est impair avec $n = 2p + 1 (p \in \mathbb{N})$, alors une **médiane** de a est égale à a_{p+1} ;
- si n est pair avec $n = 2p (p \in \mathbb{N})$, alors une **médiane** de a est égale à $\frac{a_p + a_{p+1}}{2}$.

Si M est « la » médiane, le minimum de d_1 est :

$$\frac{1}{n}\sum_{i=1}^{n}|x_i - M|$$

C'est l'écart moyen à la médiane.

Mais en pratique, l'indicateur de dispersion que l'on associe à la médiane est l'écart interquartile (une justification possible : la valeur absolue se prête mal aux calculs algébriques).

euler – Ress. 59 (outil)

euler.ac-versailles.fr/wm3/pi2/mediane/mediane4.jsp

Soit S la série statistique dont le tableau des effectifs est le suivant.

Modalité	1	2	3	4
Effectif	4	3	2	1

Soit f la fonction donnant la somme des distances d'un nombre à chacun des termes de S.
f est définie pour tout $x \in [1; 4]$ par

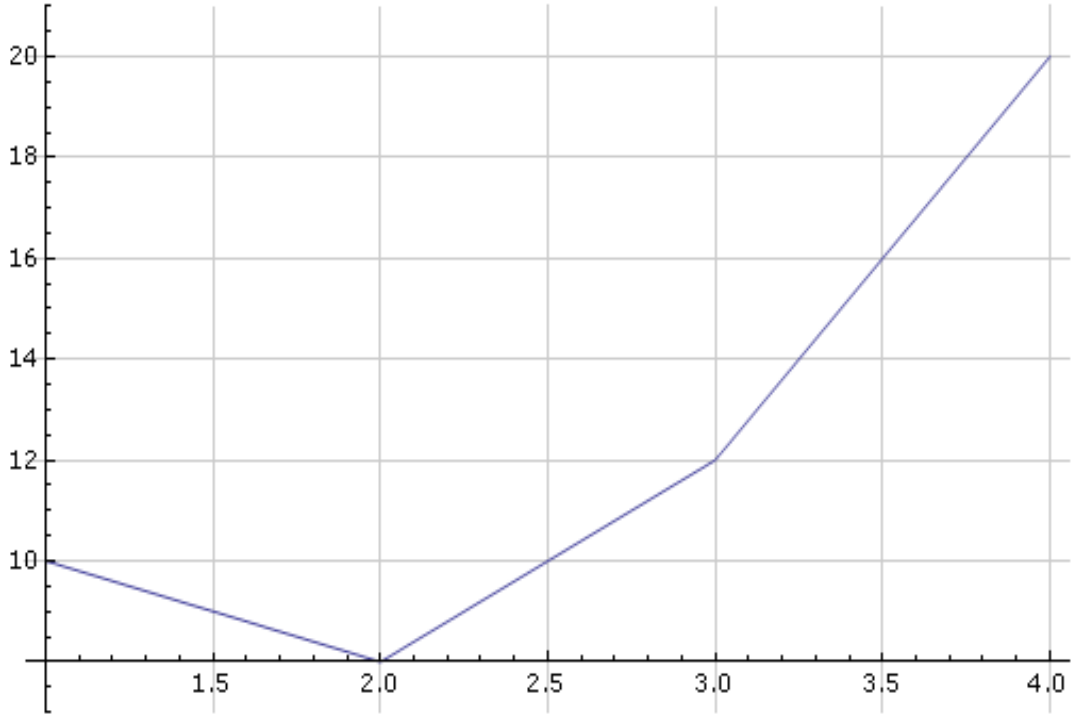
$$f(x) = |x-4| + 2|x-3| + 3|x-2| + 4|x-1|$$

et, pour tout $x \in [1; 4]$,

$$\begin{cases} f(x) = -2(x-6) & \text{si } x \in [1; 2] \\ f(x) = 4x & \text{si } x \in [2; 3] \\ f(x) = 8x-12 & \text{si } x \in [3; 4] \end{cases}$$

Le minimum de f est atteint en 2 qui représente la médiane de S.

Une représentation graphique de f sur $[1; 4]$ est donnée ci-contre.



Soit S la série statistique dont le tableau des effectifs est le suivant.

Modalité	12	17	18	21	29
Effectif	4	9	17	16	14

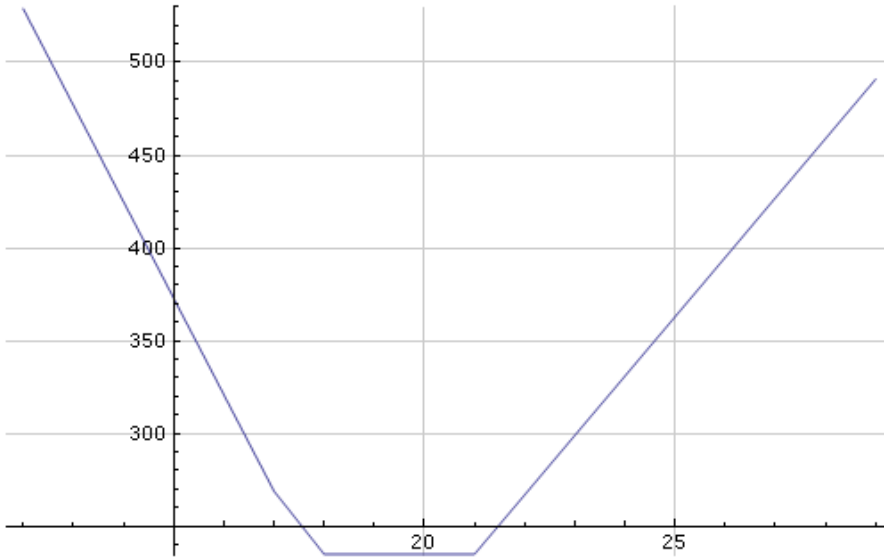
Soit f la fonction donnant la somme des distances d'un nombre à chacun des termes de S.
f est définie pour tout $x \in [12; 29]$ par

$$f(x) = 14|x - 29| + 16|x - 21| + 17|x - 18| + 9|x - 17| + 4|x - 12|$$

et, pour tout $x \in [12; 29]$,

$$\begin{cases} f(x) = 1153 - 52x & \text{si } x \in [12; 17] \\ f(x) = 847 - 34x & \text{si } x \in [17; 18] \\ f(x) = 235 & \text{si } x \in [18; 21] \\ f(x) = 32x - 437 & \text{si } x \in [21; 29] \end{cases}$$

Le minimum de f est atteint en tout nombre de $[18; 21]$ et on convient qu'une médiane de S est $\frac{39}{2}$.



(moyenne ; écart type)/(médiane ; écart interquartile)

Théorie :

- La moyenne (arithmétique) minimise la somme des carrés des distances à chacun des termes de la série

- La médiane minimise la somme des distances à chacun des termes de la série

- La médiane est généralement plus appropriée pour l’analyse des petits groupes, car la moyenne peut être influencée par des valeurs extrêmes peu représentatives.
- Dans les grands groupes, de plusieurs dizaines ou centaines de données, la médiane et la moyenne tendent à se confondre. Les dispersions vers le haut et vers le bas tendent à se compenser, et les cas extrêmes isolés ont peu d'influence sur la moyenne.

Dans les grands groupes, la médiane peut donner une image peu représentative de la réalité, par exemple

pour des groupes formés de sous-groupes eux-mêmes assez homogènes, mais bien distincts les uns des autres.

Prenons un cas simple : deux groupes de même effectif.

Dans chacun, il y a un sous-groupe de salaires de 2000 € et un sous-groupe de salaires de 3000 €.

On suppose la proportion des deux sous-groupes différente entre les deux groupes : dans le premier, il y a 60% de « bas salaires » et 40% de « hauts salaires ». Dans le second groupe, la proportion est inversée.

- Sur l’ensemble des 2 groupes, la moyenne et la médiane sont égales à 2500 €.

- Pour le groupe 1, la médiane est 2000 €, la moyenne est 2400 €.

  Pour le groupe 2, la médiane est 3000 €, la moyenne est 2600 €.

- L’écart des moyennes est 200 €, soit 8% de la moyenne globale.

  L’écart des médianes est 1000 €, soit 40 % de la médiane globale.

- Médiane et écart interquartile sont peu sensibles aux valeurs extrêmes …

- Le couple (moyenne ; écart type) trouve sa pertinence, entre autres, en liaison avec les probabilités, en particulier avec la loi normale (utilisée dans de nombreuses situations …) ; on étudie la proportion de valeurs dans $[m - 2\sigma ; m + 2\sigma]$ …

Un autre exemple….

Le salaire annuel d'une personne est 32 000 €.

Elle apprend que la moyenne et la médiane des salaires (pour le même type de poste) sont toutes deux égales à 44 000 €.

- La comparaison à la **moyenne** ne permet pas de conclure que la personne est « mal payée » ; en effet, peut-être que la plupart des salariés ont un salaire du même ordre que le sien, et que quelques privilégiés font « basculer » la moyenne.
- La valeur de la **médiane** indique qu'au moins la moitié des salariés gagnent 44 000 € ou davantage ; la personne peut considérer qu'elle est « mal payée » …

Comparaison moyenne / médiane

- Si la moyenne est beaucoup plus élevée que la médiane, cela signifie que quelques valeurs du caractère sont beaucoup plus hautes que l'ensemble des autres.
- Si la moyenne est beaucoup plus basse que la médiane, cela signifie que quelques valeurs du caractère sont beaucoup plus basses que l'ensemble des autres.
- Il est possible que la plupart des valeurs (éventuellement toutes sauf une) soient inférieures à la moyenne (ou supérieures à la moyenne).

Effet de structure
(d'après DNB Polynésie juin 2010)

Deux entreprises emploient chacune 100 personnes et publient les informations suivantes :

Salaire moyen net (en €) (effectif)	Entreprise A	Entreprise B
Hommes	1680 (50)	1800 (20)
Femmes	1200 (50)	1320 (80)

Dans l'entreprise A, la moyenne des salaires est 1440 €, dans l'entreprise B elle est égale à 1416 €.

Autres exemples …

Dans un collège, on étudie les résultats au DNB des deux classes de troisième :

	Classe de 3e A		
	effectif	nbre de reçus	% de reçus
G	20	8	40 %
F	5	1	20 %

	Classe de 3e B		
	effectif	nbre de reçus	% de reçus
G	8	6	75 %
F	25	17	68 %

Dans ce collège, il semblerait que les garçons réussissent mieux que les filles …

Pourtant, globalement :

- 28 garçons dont 14 sont reçus, soit 50 %
- 30 filles dont 18 sont reçues, soit 60 % …

- En 2000 : 3 cadres, 7 ouvriers

  Salaire mensuel moyen d'un cadre : 10 000 €, d'un ouvrier : 1000 €
- En 2001 : 2 cadres, 8 ouvriers

  Salaire mensuel moyen d'un cadre : 10 100 €, d'un ouvrier : 1100 €
- Evolution globale des salaires

  Salaire moyen d'un salarié en 2000 : 3700 €

  Salaire moyen d'un salarié en 2001 : 2900 €

Population active à l'ouest et à l'est d'un pays :

	Agriculteurs	Ouvriers
Ouest	2200	1800
Est	1400	2800

Nombre de chômeurs :

	Agriculteurs	Ouvriers
Ouest	500	600
Est	300	900

Pourcentages de chômeurs :

	Agriculteurs	Ouvriers
Ouest	$22,7\%$	$33,3\%$
Est	$21,4\%$	$32,1\%$

Le taux de chômage, dans chacune des deux catégories, est plus important à l'ouest qu'à l'est. Et pourtant, globalement ...

À l'ouest, 1100 chômeurs sur 4000 personnes, soit $27,5\%$.

À l'est, 1200 chômeurs sur 4200 personnes, soit $28,57\%$.