

Statistiques à une variable

Prélude - Historique et contexte

Le mot « statistique » vient de l'allemand *Statistik*, la « science de l'État » : dès l'Antiquité, les dirigeants recensaient population et récoltes. Au XIX^e siècle, le Belge Adolphe Quetelet étudie « l'homme moyen » et Florence Nightingale invente des diagrammes pour convaincre l'armée britannique d'améliorer l'hygiène des hôpitaux. Aujourd'hui, résumer une série de données par quelques nombres bien choisis est indispensable, mais il faut savoir ce que chacun d'eux mesure vraiment.

I Indicateurs de position

Définition - Moyenne pondérée

Pour une série de valeurs $x_1, \dots, x_p$ d'effectifs $n_1, \dots, n_p$ et d'effectif total $N = n_1 + \dots + n_p$:

$$\bar{x} = \frac{n_1 x_1 + n_2 x_2 + \dots + n_p x_p}{N} = \frac{1}{N} \sum_{i=1}^p n_i x_i$$

Si les données sont regroupées en classes, on remplace chaque classe par son **centre**.

Exemple

Note	8	10	12	15	18
Effectif	3	5	7	4	1

$$\bar{x} = \frac{3 \times 8 + 5 \times 10 + 7 \times 12 + 4 \times 15 + 1 \times 18}{20} = \frac{236}{20} = 11,8.$$

Propriété - Linéarité de la moyenne

Si on transforme chaque valeur x_i en $ax_i + b$, la moyenne devient $a\bar{x} + b$.

Exemple : si le professeur ajoute 1 point à toutes les notes puis multiplie par 1,1, la nouvelle moyenne est $1,1 \times (11,8 + 1) = 14,08$.

Définition - Médiane et quartiles

On range les N valeurs dans l'ordre croissant.

- La **médiane** Me partage la série en deux groupes de même effectif : au moins 50 % des valeurs sont inférieures ou égales à Me et au moins 50 % sont supérieures ou égales.
- Le **premier quartile** Q_1 est la plus petite valeur telle qu'au moins 25 % des valeurs lui soient inférieures ou égales.
- Le **troisième quartile** Q_3 est la plus petite valeur telle qu'au moins 75 % des valeurs lui soient inférieures ou égales.

Exemple

Avec les 20 notes précédentes : Me est la moyenne des 10^e et 11^e valeurs : $Me = 12$. $\frac{20}{4} = 5$ donc Q_1 est la 5^e valeur : $Q_1 = 10$. $\frac{3 \times 20}{4} = 15$ donc Q_3 est la 15^e valeur : $Q_3 = 12$.

II Indicateurs de dispersion

Définition

- L'**étendue** est la différence entre la plus grande et la plus petite valeur.
- L'**écart interquartile** est $Q_3 - Q_1$: il mesure la dispersion de la moitié « centrale » des valeurs.
- L'**écart type** s mesure la dispersion des valeurs autour de la moyenne :

$$s = \sqrt{\frac{n_1(x_1 - \bar{x})^2 + \dots + n_p(x_p - \bar{x})^2}{N}}$$

Remarque

- Plus l'écart type est grand, plus les valeurs sont dispersées autour de la moyenne. On le calcule en pratique à la calculatrice ou au tableur.
- La moyenne et l'écart type sont sensibles aux valeurs extrêmes. La médiane et l'écart interquartile le sont beaucoup moins : ajouter une valeur très grande modifie fortement la moyenne mais peu la médiane.
- Pour beaucoup de séries, environ 95 % des valeurs sont dans l'intervalle $[\bar{x} - 2s ; \bar{x} + 2s]$.

Remarque - Python

```
from math import sqrt
def indicateurs(L):
    n = len(L)
    m = sum(L) / n
    s = sqrt(sum((x - m)**2 for x in L) / n)
    p = len([x for x in L if m - 2*s <= x <= m + 2*s]) / n
    return m, s, p
```

III Séries continues et représentations graphiques

Définition

Lorsque le caractère étudié est continu (taille, durée, masse...), on regroupe les valeurs en **classes** $[a ; b[$ de même amplitude.

- On représente la série par un **histogramme** : chaque classe est un rectangle dont l'aire est proportionnelle à l'effectif.
- Le **polygone des fréquences cumulées croissantes** relie les points ayant pour abscisse la borne supérieure de chaque classe et pour ordonnée la fréquence cumulée.
- La **classe médiane** est la première classe pour laquelle la fréquence cumulée atteint 50 %. En supposant les valeurs réparties uniformément dans cette classe, on estime la médiane par lecture graphique sur le polygone (abscisse du point d'ordonnée 0,5).

Exemple

Temps de trajet (en min) de 50 élèves :

Durée	[0 ; 10[	[10 ; 20[	[20 ; 30[	[30 ; 40[
Effectif	12	18	14	6
Fréq. cumulée	0,24	0,60	0,88	1

La classe médiane est [10 ; 20[. Sur le segment reliant (10 ; 0,24) à (20 ; 0,60), l'ordonnée 0,5 est atteinte pour $x = 10 + 10 \times \frac{0,5 - 0,24}{0,60 - 0,24} \approx 17,2$. La médiane est estimée à environ 17 min.

Moyenne (avec les centres) : $\bar{x} = \frac{12 \times 5 + 18 \times 15 + 14 \times 25 + 6 \times 35}{50} = \frac{890}{50} = 17,8$ min.