Statistiques
2nd

Sacha Darthenucq

Prérequis:

- **Vocabulaire ensembliste et logique (2nd)**
- **Évolutions de quantités (2nd)**

Intro: brefs rappels de collège

Définition: On appelle population l'ensemble des individus concernés par l'étude statistique, et caractère la propriété étudiée sur chacun d'entre eux.

Définition: Une série statistique est la suite des valeurs que prends un caractère au sein d'une population.

Définition: L'effectif d'une valeur correspond au nombre d'apparition de cette valeur dans la série.

Soit une série statistique représentée sous forme d'un tableau:

valeur	x_1	x_2	...	x_p
effectif	n_1	n_2	...	n_p

L'effectif total de la série vaut $N = n_1 + ... + n_p$, l'effectif de la valeur x_i vaut n_i.

Définition: La fréquence d'un caractère x_i correspond au quotient de l'effectif n_i de ce caractère par l'effectif total N: $f_i = \frac{n_i}{N}$.

1 Indicateurs de tendance d'une série statistique

1.1 Moyenne

Définition: La moyenne pondérée d'une série statistique (comme définit dans le tableau précédent) vaut $\bar{x} = \frac{x_1 n_1 + ... + x_p n_p}{N}$.

Propriété: $\bar{x} = f_1 x_1 + ... + f_p x_p$.

Démo: $\bar{x} = \frac{x_1 n_1 + ... + x_p n_p}{N} = x_1 \frac{n_1}{N} + ... + x_p \frac{n_p}{N} = x_1 f_1 + ... + x_p f_p$.

Propriété: La moyenne pondérée est linéaire, c'est à dire que si l'on multiplie par un réel a toutes les valeurs de la série, la moyenne est multipliée par a, et si l'on rajoute la valeur b à toutes les valeurs de la série, la moyenne augment de b.

Démo:

- $\dfrac{n_1 \times (ax_1) + ... + n_p \times (ax_p)}{N} = a \times \dfrac{n_1 x_1 + ... + n_p x_p}{N} = a\bar{x},$
- $\dfrac{n_1(x_1 + b) + ... + n_p(x_p + b)}{N} = \dfrac{n_1 x_1 + ... + n_p x_p}{N} + \dfrac{n_1 b + ... + n_p b}{N} = \bar{x} + b$

1.2 Médiane

Définition: La médiane d'une série statistique est une valeur m qui permet de couper la série en 2 parties égales: la première correspondant à l'ensemble des valeurs inférieures ou égales à la médiane, la seconde correspondant à l'ensemble des valeurs qui lui sont supérieurs ou égales.

Méthode: Calcul de la médiane

On trie la série par ordre croissant des valeurs:

valeur	x_1	x_2	...	x_p
effectif	n_1	n_2	...	n_p

on a donc $x_i \leq x_{i+1}$.

- Si la série statistique a un effectif total N impaire, alors la médiane m correspond à la valeur du terme de rang $\dfrac{N+1}{2}$,
- Si la série statistique a un effectif total N paire, alors la médiane m vaut correspond a la moyenne des valeurs des termes de rang $\dfrac{N}{2}$ et $\dfrac{N}{2}+1$.

Exemple: Considérons la série statistique:

notes	12	13	14	15	16	17
nb d'élèves	7	9	3	14	4	1

Elle est triée par ordre croissant des valeurs.
L'effectif total est $N = 7 + 9 + 3 + 14 + 4 + 1 = 38$. Il est paire.
La médiane correspond à la moyenne des valeurs de rang $\dfrac{N}{2} = 19$ et $\dfrac{N}{2} + 1 = 20$.
La valeur de rang 19 est 14, la valeur de rang 20 est 15, d'où $m = \dfrac{14 + 15}{2} = 14.5$
La médiane de la série statistiques est $m = 14.5$.

Remarque: Utilité de la médiane
On utilise souvent la moyenne pour dégager une tendance globale d'une série statistique, par exemple une moyenne de notes pour savoir ou l'on se place par rapport à l'ensemble de la classe.
Le problème de la moyenne est que quelques valeurs très hautes ou très basses peuvent vite brouiller son utilité.
Par exemple en moyenne le salaire net d'un français est de 2230 €/mois, mais le salaire médian est lui de 1710 €/mois, une grosse différence !
Ainsi la moitié des gens en France gagnent moins de 1710 €/mois, malgré une moyenne des salaires à 2230 €/mois.
On peut donc dire de manière paradoxale: "la majorité des français gagne moins que la moyenne des français".
La médiane permet donc de positionner plus précisément une valeur par rapport à une série statistique.

1.3 Quartiles

Définition:

- Le premier quartile, noté Q_1, correspond à la plus petite valeur de la série telle qu'au moins $\frac{1}{4}$ (soit 25%) des valeurs de la série lui soient inférieures ou égales,
- Le troisième quartile, noté Q_3, correspond à la plus petite valeur de la série telle qu'au moins $\frac{3}{4}$ (soit 75%) des valeurs de la série lui soient inférieures ou égales.

Méthode: Calcul des quartiles

On trie la série par ordre croissant des valeurs:

valeur	x_1	x_2	...	x_p
effectif	n_1	n_2	...	n_p

on a donc $x_i \leq x_{i+1}$.

- Le premier quartile correspond à la valeur du premier entier supérieur ou égal à $\frac{N}{4}$,
- Le troisième quartile correspond à la valeur du premier entier supérieur ou égal à $\frac{3N}{4}$.

Exemple: En reprenant l'exemple précédent:

$\frac{N}{4} = \frac{38}{4} = 9.5$, le premier quartile correspond à la $10^{\text{ème}}$ valeur de la série soit $Q_1 = 13$.

$\frac{3N}{4} = \frac{3 \times 38}{4} = 28.5$, le troisième quartile correspond à la $29^{\text{ème}}$ valeur de la série soit $Q_3 = 15$.

Remarque: Utilité des quartiles

De la même manière que la médiane, les quartiles permettent de situer une valeur par rapport aux autres valeurs de la série. Ainsi dans l'exemple précédent, je sais qu'avec un 16 je fais partie du 1^{er} quart de la classe.

2 Indicateur de dispersion d'une série statistique

2.1 Variance et écart-type

Définition: La variance d'une série statistique, noté V est définie par $V = \frac{n_1(x_1 - \bar{x})^2 + ... + n_p(x_p - \bar{x})^2}{N}$

Remarque: La variance mesure la moyenne des carrés des écarts à la moyenne, elle mesure donc à quel point les valeurs de la série sont dispersés.

Si toutes les valeurs de la série sont égales à la moyenne alors la série n'est pas du tout dispersée et la variance vaut 0.

Définition: L'écart-type d'une série statistique, noté σ, est définit par $\sigma = \sqrt{V}$.

Remarque: L'écart-type mesure la dispersion des valeurs de la série par rapport à la moyenne. Plus l'écart-type est grand, plus les valeurs de la série sont dispersées.

2.2 Écart interquartile

Définition: Soit une série statistique de premier quartile Q_1 et de troisième quartile Q_3. L'écart interquartile correspond à la différence $Q_3 - Q_1$.

Remarque: Utilité de l'écart interquartile
De part la définition des quartiles, environ la moitié des valeurs de la série se situent dans l'intervalle $[Q_1; Q_3]$. L'écart interquartile mesure la longueur de cet intervalle. Plus l'écart interquartile est grand plus les valeurs de la série sont dispersées.
Un autre point intéressant de l'écart interquartile est qu'il mesure la dipsersion de la série sans prendre ne compte les valeurs extrêmes (contrairement à l'écart-type) qui peuvent fausser les interprétations.