STATISTIQUES

Pour étudier une série statistique, on a recourt à des indicateurs. Ils peuvent être à tendance centrale (comme la moyenne ou la médiane) ou de dispersion (comme l'étendue). Nous allons en étudier quelques uns, comprendre leur complémentarité de manière à pouvoir les interpréter.

On notera $\sum_{i=1}^{n} x_i$ la somme de i=1 à i=n des x_i soit $x_1+x_2+x_3+\ldots+x_n$. On remarque que Σ se prononce ***Somme.***

I. Moyenne et écart-type

i. Moyenne pondérée

Déf : Soit $\{x_1;\ldots;x_n\}$ une série de valeurs telle pour laquelle l'effectif d'un x_i est n_i ($i\in[0;n]$). On note $\bar{x}$ la moyenne de cette série qui est :

$$\bar{x}=\frac{\sum_{i=1}^{n} x_i n_i}{\sum_{i=1}^{n} n_i}$$

Ex : voici le classement des 18 candidats d'un concours :

Rang	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
Note	4	4.5	6	8	9	10	11	11.5	13	13.5	13.5	14	15	15.5	16	17	17	17.5

$$\bar{x}=\frac{4+4.5+6+8+9+10+11+11.5+13+13.5+13.5+14+15+15.5+16+17+17+17.5}{18}=12$$

ii. Moyenne par paquet

Prop : Soit $\{x_1;\ldots;x_n\}$ une série de taille n et de moyenne $\bar{x}$,obtenue en réunissant k séries de taille $n_1,n_2,..,n_k$ et de moyenne respective $\bar{y_1},...,\bar{y_k}$. Alors :

$$\bar{x}=\frac{\sum_{i=1}^{k} n_i \bar{y_i}}{\sum_{i=1}^{k} n_i}$$

Ex : Au lycée il y a trois classes de seconde. La première de 30 élèves a obtenue 11,5 de moyenne à un devoir de maths, la seconde de 20 élèves à obtenue 9,5 et la troisième de 35 élèves a obtenue 13.

La moyenne du devoir sur les trois classes est donc $\bar{x}=\frac{30\times11.5+20\times9.5+35\times13}{30+20+35}\approx11,6$

iii. Effet d'une application affine sur la moyenne

Prop :Soit a et b deux réels fixés. Soit $\{x_1;\ldots;x_n\}$ une série de taille n et de moyenne $\bar{x}$. Soit $\{y_1;\ldots;y_n\}$ la série telle que $y_i=ax_i+b$. Alors la moyenne $\bar{y}$ de cette dernière série est donnée par $\bar{y}=a\bar{x}+b$.

Exemple : calculer la moyenne $\bar{x}$ de {2,32 ; 2 ,39 ; 2,41 ; 2,35}

Soit la série {2 ; 9 ; 11 ; 5} de moyenne $\bar{y}$ déduite de la série précédente en enlevant 2 et en multipliant par 100.

On a : $\bar{x}=\bar{y}\times 100+2=\ldots$

iv. Moyenne des carrés des écarts

Prop : soit $\{x_1;\ldots;x_n\}$ une série statistique de taille n. La fonction f définie sur R par $f(x)=\frac{1}{n}\sum_{i=1}^{n}(x_i-x)^2$ est la moyenne des carrés des écarts de chaque x_i à x. f admet un minimum en $\bar{x}$.

Dém :

$(x_i-x)^2=x_i^2+x^2-2x_ix$ donc $\sum_{i=1}^{n}(x_i-x)^2=\sum_{i=1}^{n}(x_i^2+x^2-2x_ix)=\sum_{i=1}^{n}x_i^2+2x\sum_{i=1}^{n}x_i+nx^2$

Ainsi $f(x)=ax^2+bx+c$ avec : $\begin{cases} a=1 \\ b=\frac{2}{n}\sum_{i=1}^{n}x_i=2\bar{x} \\ c=\frac{1}{n}\sum_{i=1}^{n}x_i^2 \end{cases}$

F admet donc un minimum en $\frac{-b}{2a}=\bar{x}$. Ce minimum est $-\frac{\Delta}{4a}=-\frac{4\bar{x}^2-4\times 1\times c}{4}=-\bar{x}^2+\frac{1}{n}\sum_{i=1}^{n}x_i^2$

v. Variance et écart type

Déf :

- La variance (empirique) de $\{x_1;\ldots;x_n\}$, une série statistique de taille n, est :

$$V=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2=\frac{1}{n}\sum_{i=1}^{n}x_i^2-\bar{x}^2$$

- Soit $x_1;\ldots,x_k$ une série de valeurs telle que l'effectif de x_i soit n_i et leur fréquence f_i. La variance de cette série est alors :

$$V=\frac{1}{\sum_{i=1}^{k}n_i}\sum_{i=1}^{k}n_i(x_i-\bar{x})^2=\frac{1}{100}\sum_{i=1}^{k}f_i(x_i-\bar{x})^2$$

- On appelle écart type de la série $S=\sqrt{V}$.

Ex : voici le classement des 18 candidats d'un concours :

Rang	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
Note	4	4.5	6	8	9	10	11	11.5	13	13.5	13.5	14	15	15.5	16	17	17	17.5

On a vu au I. 1. $\bar{x}=12$,

$\frac{1}{n}\sum_{i=1}^{n} x_i^2 = \frac{1}{18}(4^2+4.5^2+6^2+8^2+9^2+10^2+11^2+11.5^2+13^2+13.5^2+13.5^2+14^2+15^2+15.5^2+16^2+17^2+17^2+17.5^2) \approx 161,4$

$V=161.4-12^2=\underline{17.4}$

$S=\sqrt{17,4} \approx \underline{4,2}$

vi. Interprétation

La moyenne donne la tendance centrale d'une série : c'est un indicateur à **tendance centrale**.

L'écart type tient compte de la répartition des valeurs autour de la moyenne : plus il est grand plus la série est dispersée et inversement : l'écart type est une **mesure de dispersion.**

Tous les deux étant liés, ils ont le même défaut : la sensibilité aux valeurs extrêmes.

vii. Effet d'une application affine

Prop :Soit a et b deux réels fixés. Soit $\{x_1;\dots;x_n\}$ une série de taille n et de variance V_x et d'écart-type σ_x Soit $\{y_1;\dots;y_n\}$ la série telle que $y_i=ax_i+b$. Alors la variance V_y et l'écart-type σ_y de cette dernière série sont donnés par $V_y=a^2\ V_x$, *et* $\sigma_y=|a|\sigma_x$.

Dém : On sait $\bar{y}=a\bar{x}+b$ est on remplace dans la formule de la variance.

II. Quartiles et intervalle interquartile.

1. Définitions

Déf : Soit $\{x_1;\dots;x_n\}$ une série statistique de taille n rangée par ordre croissant.

- Le **premier quartile Q_1** est la plus petite valeur x_i tel que 25% au moins des données soit inférieure ou égale à x_i.
- Le **deuxième quartile Me**, aussi appelé ***médiane***, est la plus petite valeur x_i tel que 50% au moins des données soit inférieure ou égale à x_i.
- Le **troisième quartile Q_3** est la plus petite valeur x_i tel que 75% au moins des données soit inférieure ou égale à x_i.

Ex : Revoici le classement des 18 candidats d'un concours :

Rang	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
Note	4	4.5	6	8	9	10	11	11.5	13	13.5	13.5	14	15	15.5	16	17	17	17.5

L'effectif est de 18 personnes.

Q_1 est 9 car c'est la plus petite note telle qu'au moins 25% des candidats (soit au moins 4,5) ont moins de 9.
Me est 13 car c'est la plus petite note telle qu'au moins 50% des candidats (soit au moins 9) ont moins de 13.
Q_3 est 15.5 car c'est la plus petite note telle qu'au moins 75% des candidats (soit au moins 13.5) ont moins de 15.5

Rem :

On définit les ***déciles*** comme les quartiles : le premier décile D_1 est le plus petit x_i tel que 10% au moins des valeurs soient plus petite ou égale à xi et le neuvième décile D_9 qui est le plus petit xi tel que 90% au moins des valeurs soient plus petite ou égale à xi .

Ex : En reprenant la série du 1. D_1=4.5 (au moins 1,8 candidats au moins au pareil) et D_9=17 (au moins 16.2 personnes au moins ou pareil).

2. L'intervalle interquartile.

Déf : On appelle **intervalle interquartile** l'intervalle [Q1 ;Q3] et **l'écart interquartile** est Q3-Q1.

Rem :

Cet intervalle est, comme l'écart-type, une mesure de dispersion. Cependant, comme les quartiles, il n'est pas sensible aux valeurs extrêmes, tout comme la médiane : si on remplace la note du premier rang par 0 et celle du dernier par 20, ni l'un ni l'autre ne seront affectés. Ainsi la moyenne et l'écart-type sont très bien complétés par les quartiles et l'intervalle interquartile.

3. Diagramme en boîte.

On utilise un diagramme en boîte aussi appelé diagramme du Tuckey ou diagramme à pattes ou à moustaches pour représenter une série statistique dont les valeurs extrêmes ne sont pas importantes.

Le long d'une échelle représentant l'étendue de la série , On forme une boîte limité par le premier et neuvième décile ainsi que par les premier et troisième quartile. On y indique la médiane.