Chapitre 4 – Statistiques

I – Un symbole pour écrire une somme

En statistiques, on calcule souvent des sommes.

Pour ceci, on utilise un symbole afin d'écrire une somme en évitant les pointillés : par exemple, comment noter simplement $1+2+3+4+\ldots+20$?

Cette somme comporte 20 termes, on peut noter cette somme ainsi, avec un symbole *sigma* (Σ) :

$\sum_{i=1}^{20} i=1+2+3+\ldots+20$. Cela se lit « somme pour *i* allant de 1 à 20 des *i* ».

De la même manière, on a : $1^2+2^2+3^2+\ldots+100^2=\sum_{i=1}^{100} i^2$.

Cette écriture est bien appropriée pour des indices : $a_3+a_4+a_5+\ldots+a_{71}=\sum_{i=3}^{71} a_i$.

II – Indicateurs statistiques

On s'intéresse à un caractère prenant différentes valeurs. On suppose que le caractère prend p valeurs différentes. Les différentes valeurs du caractère sont x_1 , x_2 , ..., x_p d'effectifs respectifs n_1 , n_2 , …, n_p . n_i est donc l'effectif de la valeur x_i .

On peut donc résumer la série par un tableau statistique :

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

Effectif total : $N=n_1+n_2+...+n_p=\sum_{i=1}^{p} n_i$

Fréquence de la valeur x_i : $f_i=\frac{n_i}{N}$

<u>*Exemple :*</u> *Pour toute cette partie, on considère les âges d'un groupe de personnes.*

Âge (ans)	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Effectif	*1*	*2*	*1*	*3*	*5*	*6*	*7*	*4*	*1*	*2*	*2*
Effectif Cumulé Croissant	*1*	*3*	*4*	*7*	*12*	*18*	*25*	*29*	*30*	*32*	*34*

*L'**effectif total** est* $N=1+2+1+3+5+6+7+4+1+2+2=34$ *(c'est le dernier effectif cumulé croissant).*

a) Indicateurs de tendance centrale

Mode : C'est la valeur la plus fréquente.

Moyenne : $\bar{x}=\dfrac{n_1x_1+n_2x_2+...+n_px_p}{N}=\dfrac{\sum_{i=1}^{p} n_i x_i}{N}$

Théorème : On a aussi $\bar{x}=f_1x_1+f_2x_2+...+f_px_p=\sum_{i=1}^{p} f_i x_i$

Preuve : $\bar{x}=\dfrac{\sum_{i=1}^{p} n_i x_i}{N}=\sum_{i=1}^{p}\dfrac{n_i}{N}x_i$ or $f_i=\dfrac{n_i}{N}$ donc $\bar{x}=\sum_{i=1}^{p} f_i x_i$

Médiane : On suppose que les valeurs de la série d'effectif N sont rangées par ordre croissant (chacune d'elles étant répétée autant de fois que son effectif) : $x_1\leq x_2\leq ...\leq x_N$.

- Si N est impair, $Me=x_{\frac{N+1}{2}}$ (c'est le terme de rang $\dfrac{N+1}{2}$).
- Si N est pair, $Me=\dfrac{x_{\frac{N}{2}}+x_{\frac{N}{2}+1}}{2}$ (c'est la moyenne des termes de rang $\dfrac{N}{2}$ et $\dfrac{N}{2}+1$).

Exemple : Avec l'exemple précédent,

- *Le **mode** est 6 ans (car 7 personnes ont 6 ans).*
- *La **moyenne** est* $\bar{x}=\dfrac{0\times1+1\times2+2\times1+3\times3+4\times5+5\times6+6\times7+7\times4+8\times1+9\times2+10\times2}{34}\approx 5{,}26$ *ans.*
- *La **médiane** est la valeur qui sépare la série statistique en deux parties de même effectif. Ici, il y a 34 valeurs, donc la médiane est la moyenne de la 17ème et la 18ème valeur. Grâce aux effectifs cumulés croissant, la 17ème valeur est 5 ans, la 18ème valeur est 5 ans. La médiane est* $Me=\dfrac{5+5}{2}=5$ *ans.*

b) Indicateurs de position : Les quartiles

Les valeurs de la série d'effectif N sont rangées par ordre croissant (chacune d'elles étant répétée autant de fois que son effectif) : $x_1\leq x_2\leq ...\leq x_N$.

Premier quartile : Le premier quartile Q_1 de la série est la valeur x_i dont l'indice i est le plus petit entier supérieur ou égal à $\dfrac{N}{4}$.

Troisième quartile : Le troisième quartile Q_3 de la série est la valeur x_i dont l'indice i est le plus petit entier supérieur ou égal à $\dfrac{3N}{4}$.

Remarque : On prend souvent comme deuxième quartile la médiane.

Exemple : *Avec l'exemple précédent,*

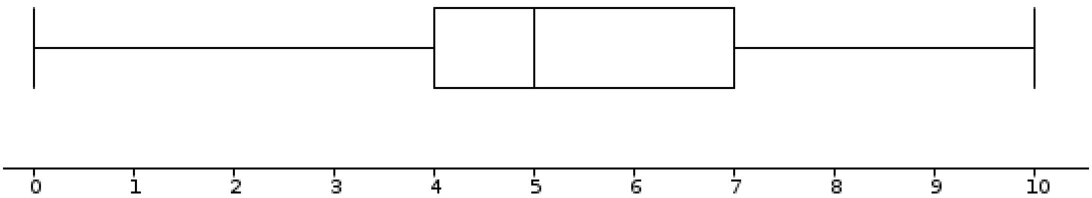
- $\frac{N}{4}=\frac{34}{4}=8,5$ *, donc* Q_1 *est la 9ème valeur :* $Q_1=4$ *ans.*
- $\frac{3N}{4}=\frac{3\times 34}{4}=25,5$ *, donc* Q_3 *est la 26ème valeur :* $Q_3=7$ *ans.*

c) Boîtes-à-moustaches

Pour résumer notre série statistique, on construit un diagramme en boîte.

- Les valeurs du caractère sont résumées sur un axe.
- On construit un rectangle (la boîte), parallèlement à l'axe, dont la longueur est l'intervalle interquartile $[Q_1;Q_3]$.
- Un trait symbolise la médiane Me .
- On place les moustaches au niveau des valeurs extrêmes.

Exemple : *Avec l'exemple précédent, on a cette boîte-à-moustaches pour les âges du groupe de personnes :*



d) Indicateurs de dispersion

Variance : $V=\frac{n_1(x_1-\overline{x})^2+n_2(x_2-\overline{x})^2+...+n_p(x_p-\overline{x})^2}{N}=\frac{\sum_{i=1}^{p} n_i(x_i-\overline{x})^2}{N}$

Théorème admis : On a aussi $V=\frac{n_1x_1^2+n_2x_2^2+...+n_px_p^2}{N}-\overline{x}^2=\frac{\sum_{i=1}^{p} n_i x_i^2}{N}-\overline{x}^2$

Écart-type : $\sigma=\sqrt{V}$

Écart interquartile : $E_i=Q_3-Q_1$

Étendue : C'est la différence entre les valeurs extrêmes de la série : la plus grande moins la plus petite.

<u>Exemple :</u> Avec l'exemple précédent, comme $\bar{x} \approx 5{,}26$ *on a :*

- $V = \dfrac{1(0-\bar{x})^2+2(1-\bar{x})^2+1(2-\bar{x})^2+3(3-\bar{x})^2+\ldots+1(8-\bar{x})^2+2(9-\bar{x})^2+2(10-\bar{x})^2}{34} \approx 5{,}72$.

  On a également $V = \dfrac{1\times 0^2+2\times 1^2+1\times 2^2+3\times 3^2+\ldots 1\times 8^5+2\times 9^2+2\times 10^2}{34} - \bar{x}^2 \approx 5{,}72$.
- $\sigma = \sqrt{V} \approx 2{,}39$ *ans.*
- $E_i = Q_3 - Q_1 = 7-4=3$ *ans.*
- *Étendue :* $10-0=10$ *ans.*

<u>e) Résumer une série statistique</u>

Résumer une série statistique, c'est indiquer la répartition des données en utilisant différents indicateurs, notamment un indicateur de tendance centrale et un indicateur de dispersion.

Paramètre de tendance centrale	**Paramètre de dispersion**	**Propriété**
Médiane : Me	Écart interquartile : E_i	Peu sensible aux valeurs extrêmes
Moyenne : $\bar{x}$	Écart-type : σ	Sensible aux valeurs extrêmes