CHAPITRE 1 : Séries statistiques à une seule variable

2. Paramètres de dispersion

Nous avons vu comment représenter une série statistique par un tableau statistique et par un graphique. Nous avons aussi traité les paramètres de tendance centrale. Nous allons découvrir dans cette section les paramètres de dispersion qui complètent les paramètres de tendance centrale. En effet, on peut avoir des séries ayant mêmes paramètres de tendance centrale et pourtant sont en réalité différentes.

Les paramètres de position caractérisent l'ordre de grandeur des observations alors que les paramètres de dispersion caractérisent l'étalement des valeurs autour d'un paramètre de position.

Position du problème

Considérons 4 groupes de 5 jeunes personnes (A,B,C et D) auxquelles on demande leur âge. On obtient les résultats suivants :

groupe	A	B	C	D
âge (en années)	10	8	6	4
	10	9	8	7
	10	10	10	10
	10	11	12	13
	10	12	14	16
moyenne	10	10	10	10
médiane	10	10	10	10

Les quatre séries A, B, C et D ont même Me et même moyenne arithmétique et pourtant sont différentes :

- il n'a y a pas de dispersion dans A ;
- les valeurs sont plus proches dans B que dans C ;
- il existe deux valeurs extrêmes dans D (4 et 16).

Il est donc clair que les paramètres de position tels que la moyenne et la médiane ne suffisent pas pour décrire l'hétérogénéité de ces séries. Or, ce qui les distingue c'est l'étalement des âges observés.

En d'autres termes, il nous manque un paramètre (ou des paramètres) pour caractériser la dispersion des observations autour d'un paramètre de position.

Définition de la dispersion : C'est la variabilité ou l'étendue des différentes valeurs que peut prendre une variable statistique. Si la différence entre les valeurs est grande la dispersion est grande et si la différence entre les valeurs est petite la dispersion est petite.

2.1 L'étendue statistique

L'étendue statistique est le plus simple paramètre de dispersion notée **E** . C'est la différence entre la valeur maximale et la valeur minimale du caractère statistique.

Étendue = $x_{\max} - x_{\min}$

Si l'étendue est petite cela signifie que les valeurs de la série statistique sont proches l'une de l'autre. Donc ces valeurs sont peu dispersées.

Exemple :

B : 29 26 24 21 20 avec **E = 29 – 20 = 9**

C : 52 33 24 8 3 avec **E = 52 – 3 = 49**

Exemple 2 :

L'étendue de cette série ?

salaire en euros	ni
[500,1000[	60
[1000,1500[	70
[1500,2500[	70
[2500,3000[	200
total	400

E = 3000 - 500 = 2 500 euros

2.2 Les quartiles

Soit une série de N données rangées par ordre croissant.

- Le premier quartile est la plus petite donnée $\mathbf{Q_1}$ de la série telle qu'au moins un quart des données (25%) de la série soit inférieure ou égale à $\mathbf{Q_1}$.
- Le troisième quartile est la plus petite donnée $\mathbf{Q_3}$ de la série telle qu'au moins les trois quarts des données (75%) de la série sont inférieures ou égales à $\mathbf{Q_3}$.

  Remarque : calcul pratique des quartiles pour une série à caractère discret :
  - Pour $\mathbf{Q_1}$ on calcule N/4 puis on détermine le premier entier p supérieur ou égal à N/4 ;
  Cet entier p est le rang de $\mathbf{Q_1}$ que l'on peut alors déterminer.
  - Pour $\mathbf{Q_3}$, on fait de même en remplaçant N/4 par 3N/4 .

  Remarque : Pour le calcul des quartiles d'une série à caractère continu on utilisera le polygone des fréquences cumulées comme pour la médiane.

  - Dans le cas du polygone des fréquences cumulées croissantes le premier quartile est l'abscisse du point du polygone qui a pour ordonnée 0,25 et le troisième quartile est l'abscisse du point du polygone qui a pour ordonnée 0,75.
  - Dans le cas du polygone des fréquences cumulées décroissantes le premier quartile est l'abscisse du point du polygone qui a pour ordonnée 0,75 et le troisième quartile est l'abscisse du point du polygone qui a pour ordonnée 0,25.

EXEMPLE

Cas discret :

C : 52 33 24 8 3

On range cette série par ordre croissant : 3 8 24 33 52

N = 5 et N/4 = 1,25 3N/4 = 3,75

Q1 correspond la valeur du rang $p \geq N/4$ donc 2 donc Q1 = 8

Q3 correspond la valeur du rang $p \geq 3N/4$ donc 4 donc Q3 = 33

Cas continu :

On commence par déterminer la classe $[e_i, e_{i+1}[$ qui contient Q1. $Q1 \in [e_i, e_{i+1}[$ avec $i \in \{0, \dots, k-1\}$

Le rang de Q1 est N/4 ou $F(Q1) = 0{,}25$ et $F(e_i) \leq F(Q1) < F(e_{i+1})$

Puis on calcule Q1 par la relation suivante d'une façon analogue au calcul de la Me :

$$Q1 = (e_{i+1} - e_i).\frac{0,25-F(ei)}{F(e_{i+1})-F(e_i)} + e_i$$

$$Q1 = (e_{i+1} - e_i).\frac{\frac{N}{4}-\vec{N}(ei)}{\vec{N}(e_{i+1})-\vec{N}(e_i)} + e_i$$

De la même façon, on calcule Q3 :

Le rang de Q3 est 3N/4 ou F(Q1) = 0,75 et $F(e_i) \leq F(Q3) < F(e_{i+1})$

Puis on calcule Q3 par la relation suivante d'une façon analogue au calcul de la Me :

$$Q3 = (e_{i+1} - e_i).\frac{0,75-F(ei)}{F(e_{i+1})-F(e_i)} + e_i$$

$$Q3 = (e_{i+1} - e_i).\frac{\frac{3N}{4}-\vec{N}(ei)}{\vec{N}(e_{i+1})-\vec{N}(e_i)} + e_i$$

salaire en euros	ni	fi	FI
[500,1000[	60	0,15	0,15
[1000,1500[	70	0,175	0,325
[1500,2500[	70	0,175	0,5
[2500,3000[	200	0,5	1
total	400	1	

En se servant de Fi, Q1 = F(0,25) ϵ [1000,1500[puisque 0,15 correspond à la borne 1000 et 0,325 correspond à la borne 1500.

En appliquant les formules d'interpolation linéaire

$$Q1 = (e_{i+1} - e_i).\frac{0,25-F(ei)}{F(e_{i+1})-F(e_i)} + e_i = (1500 - 1000).\frac{0,25-0,15}{0,325-0,15} + 1000$$

Q3 = F(0,75) ϵ [2500,3000[puisque 0,5 correspond à la borne 2500 et 1 correspond à la borne 3000.

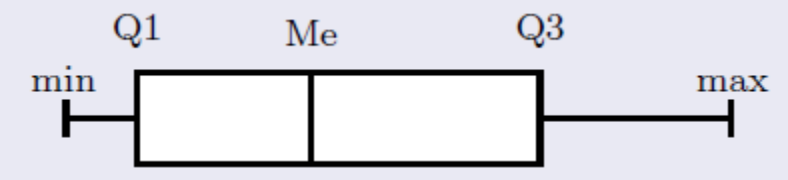
$$Q3 = (e_{i+1} - e_i).\frac{0,75-F(ei)}{F(e_{i+1})-F(e_i)} + e_i = (3000 - 2500).\frac{0,75-0,5}{1-0,5} + 2500$$

2.3 L'écart interquartile

L'écart interquartile est la différence entre le troisième et le premier quartile.

Écart interquartile = Q3 – Q1

L'écart interquartile correspond à l'étendue de la série statistique après élimination de 25 % des valeurs les plus faibles et de 25 % des valeurs les plus fortes. Cette mesure est plus robuste que l'étendue, qui est sensible aux valeurs extrêmes.



2.4 La variance

L'utilisation des valeurs absolues est souvent une impasse en mathématique (parce que la fonction valeur absolue n'est pas dérivable). Pour *rendre positifs* les écarts, un autre outil est à notre disposition : la mise au carré. On ne va donc pas calculer la moyenne des écarts mais la moyenne des carrés des écarts. C'est ce qu'on appelle la variance :

dans le cas d'une série discrète :

$$V = \frac{1}{N}\sum_{i=1}^{k}(xi - \bar{x})^2$$

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

dans le cas d'une série continue.

$$V = \frac{1}{N}\sum_{i=1}^{k}(ci - \bar{x})^2$$

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

La disparition des valeurs absolues permet des calculs plus simples. On démontre que la variance peut se calculer plus simplement par les formules suivantes :

dans le cas d'une série discrète :

$$V = \frac{1}{N}\sum_{i=1}^{k} x_i^2 \; - \; \bar{x}^2$$

$$V = \frac{\sum_{i=1}^{k} n_i x_i^2}{N} - \; \bar{x}^2 \; = \sum_{i=1}^{k} f_i x_i^2 \; - \; \bar{x}^2$$

dans le cas d'une série continue.

$$V = \frac{1}{N}\sum_{i=1}^{k} c_i^2 \; - \; \bar{x}^2$$

$$V = \frac{\sum_{i=1}^{k} n_i c_i^2}{N} - \; \bar{x}^2 \; = \sum_{i=1}^{k} f_i c_i^2 \; - \; \bar{x}^2$$

Ces formules étaient surtout utiles dans le cadre de calculs à la main ; l'usage des ordinateurs les rend un peu obsolètes...

2.5 L'écart type

De par la mise au carré des écarts, l'unité de la variance est le carré de celle du caractère (si le caractère est en kg, sa moyenne est en kg mais sa variance est en kg^2) d'où l'impossibilité d'additionner la moyenne et la variance. On a donc défini l'écart type noté σ. L'écart type est la racine de la variance (son unité est donc la même que celle de la moyenne). Cela a l'air anecdotique mais la possibilité d'additionner moyenne et écart type est fondamentale, en particulier pour le calcul d'intervalle de confiance (voir plus bas).

- dans le cas d'une série discrète :

$$\sigma = \sqrt{\frac{\sum_{i=1}^{k} (xi - \bar{x})^2}{N}}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{k} n_i(xi - \bar{x})^2}{\sum_{i=1}^{k} n_i}} = \sqrt{\sum_{i=1}^{k} f_i(xi - \bar{x})^2}$$

- dans le cas d'une série continue.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{k} n_i(ci - \bar{x})^2}{\sum_{i=1}^{k} n_i}} = \sqrt{\sum_{i=1}^{k} f_i(ci - \bar{x})^2}$$

$\sigma^2 = V$

2.6 Coefficient de variation

2.7 L'écart absolu par rapport à la moyenne arithmétique

2.8 L'écart absolu par rapport à la médiane

2.9 Les moments