

Chapitre 12 : Statistiques descriptives

I. La moyenne et l'écart-type

La moyenne est une manière de dire où se situe le « centre de la série » et l'écart-type est une mesure de la dispersion de la série autour de la moyenne.

Moyenne



La moyenne de la série statistique est le réel, notée $\bar{x}$ ou m , tel

$$\text{que : } \bar{x} = \frac{n_1 x_1 + n_2 x_2 + \dots + n_p x_p}{n_1 + n_2 + \dots + n_p}$$

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

Effectif total : $N = n_1 + n_2 + \dots + n_p$

Remarque : on a bien : $\text{moyenne} = \frac{\text{somme de toutes les valeurs}}{\text{effectif total}}$



A partir des fréquences, la moyenne se calcule par :

$$\bar{x} = f_1 x_1 + f_2 x_2 + \dots + f_p x_p$$

Valeur	x_1	x_2	...	x_p
Fréquence	f_1	f_2	...	f_p

Rappel : $f_i = \frac{n_i}{N} = \frac{\text{nombre d'apparition de } x_i}{\text{effectif total}}$

Equivalence des 2 formules :

$$\bar{x} = \frac{n_1 x_1 + n_2 x_2 + \dots + n_p x_p}{n_1 + n_2 + \dots + n_p} = \frac{n_1}{N} x_1 + \frac{n_2}{N} x_2 + \dots + \frac{n_p}{N} x_p = f_1 x_1 + f_2 x_2 + \dots + f_p x_p$$

Linéarité de la moyenne :

(piste rouge)



a et b désignent deux nombres réels et (x_i) une série statistique

Si (x_i) a pour moyenne m alors la série $(ax_i + b)$ a pour moyenne $am + b$

Preuve (piste noire):

$$\begin{aligned} M &= \frac{n_1(ax_1+b) + n_2(ax_2+b) + \dots + n_p(ax_p+b)}{n_1 + n_2 + \dots + n_p} = \frac{n_1ax_1 + n_1b + n_2ax_2 + n_2b + \dots + n_pax_p + n_pb}{n_1 + n_2 + \dots + n_p} \\ &= \frac{(n_1x_1 + n_2x_2 + \dots + n_px_p)a}{n_1 + n_2 + \dots + n_p} + \frac{(n_1 + n_2 + \dots + n_p)b}{n_1 + n_2 + \dots + n_p} = am + b \end{aligned}$$

Exemple :

On a planté des graines et on a mesuré la hauteur des plants après vingt jours. La hauteur moyenne était de 5,1 cm. On ajoute alors un engrais, la hauteur de chaque plante devrait croître encore de 25 % avant la mise en pot de chaque plant dans un pot de 6 cm de haut.

Quelle sera la hauteur moyenne de l'ensemble pot plus plant ?

Augmenter la hauteur de 25% revient à la multiplier par 1,25. D'après la linéarité de la moyenne :

$$m_{\text{finale}} = 1,25 \times 5,1 + 6 = 12,375 \text{ cm}$$

Variance et écart-type Les calculs, sauf cas simple explicitement demandé, se font à la calculatrice !

- La variance de la série statistique

A partir des effectifs $V = \frac{n_1(x_1 - \bar{x})^2 + n_2(x_2 - \bar{x})^2 + \dots + n_p(x_p - \bar{x})^2}{N}$

A partir des fréquences $V = f_1(x_1 - \bar{x})^2 + f_2(x_2 - \bar{x})^2 + \dots + f_p(x_p - \bar{x})^2$

- L'écart-type est $\sigma = \sqrt{V}$ l'écart type dans la même unité que les valeurs de la série.

Exemple : Comparaison de la performance en attaque de deux joueurs de handball d'Aix (début de saison 2019-2020)

Baptiste Bonnefond

Nombre de buts	1	2	3	4	5	7	8
Nombre de matchs	2	2	2	1	3	1	1

$$m_B = \frac{2 \times 1 + 2 \times 2 + \dots + 1 \times 8}{2 + 2 + \dots + 1} \approx 3,83 \text{ buts}$$

$$s_B \approx \sqrt{\frac{2 \times (1 - 3,83)^2 + 2 \times (2 - 3,83)^2 + \dots + 1 \times (8 - 3,83)^2}{12}} \approx 2,15 \text{ buts}$$

Nicolas Claire

Nombre de buts	1	2	3	4	5	8	10
Nombre de matchs	3	2	1	2	2	1	1

$$m_C = \frac{1 \times 3 + 2 \times 2 + \dots + 1 \times 10}{3 + 2 + \dots + 1} \approx 3,83 \text{ buts}$$

$$s_C \approx \sqrt{\frac{3 \times (1 - 3,83)^2 + 2 \times (2 - 3,83)^2 + \dots + 1 \times (10 - 3,83)^2}{12}} \approx 2,73 \text{ buts}$$

Baptiste Bonnefond et Nicolas Claire ont été aussi performants l'un que l'autre (même moyenne de buts).

Baptiste Bonnefond a été un peu plus régulier (écart-type plus faible).

Vous avez ci-dessus une rédaction de DS ... on utilise la calculatrice pour les calculs donc il faut impérativement savoir s'en servir.

II. La médiane et l'écart-interquartile

La médiane est une autre manière de dire où se situe le « centre de la série » et l'écart interquartile est la mesure de dispersion de la série autour de la médiane.

Médiane La liste des données est rangée par ordre croissant.



- Si la série est de taille **impaire** ($\text{effectif total} = 2k + 1$), la *médiane* est la donnée de rang $k + 1$ (la seule donnée centrale)
- Si la série est de taille **paire** ($\text{effectif total} = 2k$), la *médiane* est la demi-somme des données de rang k et de rang $k + 1$ (les deux données centrales)

Les quartiles La liste des données est rangée par ordre croissant.



- Le *premier quartile* est la plus petite donnée Q_1 de la liste telle qu'au moins un quart des données de la liste sont inférieures ou égales à Q_1
- Le *troisième quartile* est la plus petite donnée Q_3 de la liste telle qu'au moins trois quarts des données de la liste sont inférieures ou égales à Q_3

Intervalle interquartile : $[Q_1 ; Q_3]$ cet intervalle contient au moins 50% des valeurs de la série

Ecart interquartile : $Q_3 - Q_1$

Savoir déterminer Q_1, M_e, Q_3 , intervalle interquartile, écart interquartile

1) A partir des effectifs (effectif total impair)

- Premier quartile : $\frac{1}{4} \times \text{effectif total} = \frac{1}{4} \times 17 = 4,25$
On cherche à l'aide des effectifs cumulés croissant la valeur de rang 5 $Q_1 = 6$

Valeurs	5	6	8	10	11	12
Effectifs	3	2	5	4	2	1
Effectifs cumulés croissants	3	5	10	14	16	17

- Médiane : $\text{effectif total} = 17 = 2 \times 8 + 1$ il y a une seule valeur centrale, la valeur de rang 9
On cherche à l'aide des effectifs cumulés croissant la valeur de rang 9 $M_e = 8$
- Troisième quartile : $\frac{3}{4} \times \text{effectif total} = \frac{3}{4} \times 17 = 12,75$
On cherche à l'aide des effectifs cumulés croissant la valeur de rang 13 $Q_3 = 10$
- Intervalle interquartile : $[Q_1 ; Q_3] = [6 ; 10]$ il contient au moins 50% des valeurs de la série
- Ecart interquartile : $Q_3 - Q_1 = 10 - 6 = 4$

2) A partir des effectifs (effectif total pair)

- Premier quartile : $\frac{1}{4} \times \text{effectif total} = \frac{1}{4} \times 20 = 5$
On cherche à l'aide des effectifs cumulés croissant la valeur de rang 5 $Q_1 = 14$

Valeurs	12	14	15	16	17	18
Effectifs	2	3	2	3	4	6
Effectifs cumulés croissants	2	5	7	10	14	20

- Médiane : $\text{effectif total} = 20 = 2 \times 10$ les deux valeurs centrales sont 16 et 17 (de rang 10 et de rang 11)
 $M_e = \frac{16+17}{2} = 16,5$
- Troisième quartile : $\frac{3}{4} \times \text{effectif total} = \frac{3}{4} \times 20 = 15$
On cherche à l'aide des effectifs cumulés croissant la valeur de rang 15 $Q_3 = 18$
- Intervalle interquartile : $[Q_1 ; Q_3] = [14 ; 18]$ il contient au moins 50% des valeurs de la série
- Ecart interquartile : $Q_3 - Q_1 = 18 - 14 = 4$ Cette série est plus dispersée que celle de la question 1 !

3) A partir des fréquences

- Premier quartile : $\frac{1}{4} = 0,25$

On cherche à l'aide des fréquences cumulées

croissantes la valeur de rang pour laquelle on atteint ou dépasse 0,25 (soit 25 % de l'effectif total) $Q_1 = 14$

- Médiane : on cherche à l'aide des fréquences cumulées croissantes la valeur de rang pour laquelle on atteint ou dépasse 0,5 (soit 50 % de l'effectif total) $M_e = 15$
- Troisième quartile : on cherche à l'aide des fréquences cumulées croissantes la valeur de rang pour laquelle on atteint ou dépasse 0,75 (soit 75 % de l'effectif total) $Q_3 = 16$
- Intervalle interquartile : $[Q_1 ; Q_3] = [14 ; 16]$ il contient au moins 50% des valeurs de la série
- Ecart interquartile : $Q_3 - Q_1 = 16 - 14 = 2$

Valeurs	12	14	15	16	17	18
Fréquences	0,12	0,19	0,27	0,31	0,08	0,04
Effectifs cumulés croissants	0,12	0,31	0,58	0,88	0,96	1

Savoir interpréter Q_1, M_e, Q_3 Pour une série statistique, $Q_1 = 4$ $M_e = 6$ $Q_3 = 8$. Interpréter ces nombres.

- Interprétation de Q_1

Au moins $\frac{1}{4}$ (soit 25 %) des valeurs sont inférieures ou égales à 4
 Au moins $\frac{3}{4}$ (soit 75 %) des valeurs sont supérieures ou égales à 4
- Interprétation de M_e

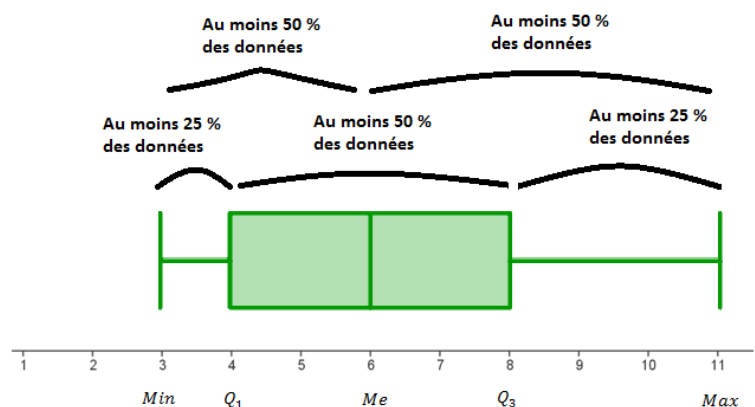
Au moins $\frac{1}{2}$ (soit 50 %) des valeurs sont inférieures ou égales à 6
 Au moins $\frac{1}{2}$ (soit 50 %) des valeurs sont supérieures ou égales à 6
- Interprétation de Q_3

Au moins $\frac{3}{4}$ (soit 75 %) des valeurs sont inférieures ou égales à 8
 Au moins $\frac{1}{4}$ (soit 25 %) des valeurs sont supérieures ou égales à 8

Diagramme en boîtes

(piste noire)

Les interprétations sont faites en incluant les bornes d'où « au moins »



Comparer la dispersion / concentration des séries avec les écarts interquartiles

Quel pays a sa population la mieux répartie dans sa pyramide d'âge ? Justifier.

Réponse :
 Ecart interquartile pour la France : $59 - 20 = 39$ ans
 Ecart interquartile pour l'Azerbaïdjan : $46 - 16 = 30$ ans
 Ecart interquartile pour le Japon : $65 - 27 = 38$ ans

	France	Azerbaïdjan	Japon
Premier quartile	20	16	27
Médiane	40	30	46
Troisième quartile	59	46	65

Répartition de la population par âge
 (au 1er janvier 2014)

L'Azerbaïdjan a un l'écart interquartile le plus petit donc sa population est la plus concentrée autour de la médiane.

La France et le Japon ont à peu près la même dispersion. Attention, les répartitions autour de la médiane se valent, mais la comparaison des médianes de la France et du Japon fait apparaître que les Japonais sont globalement plus vieux que les français. (D'ailleurs, en 2020, l'âge moyen au Japon est de 48 ans contre 42 ans en France.)