Chapitre 5

Statistiques descriptives à 1 variables

5.1 Vocabulaire

Dans ce chapitre, nous considèrerons une série de n observations ordonnées, notées $x_1, \dots, x_N$, avec $N \in \mathbb{N}$. Par exemple, pour fixer les idées, il est utile de penser aux notes obtenues par un élève lors d'un trimestre.

Exemple 5.1.1. Imaginons que Fanny ait eut les notes suivantes en mathématiques : 12 ; 8 ; 15 ; 13. Dans ce cas, $N = 4$ (car il y a 4 notes) et ceci pourrait se noter :

$$x_1 = 12 \quad ; \quad x_2 = 8 \quad ; \quad x_3 = 15 \quad ; \quad x_4 = 13.$$

Dans ce qui précède, nous avons utiliser les notes d'un élève comme exemple mais nous aurions pu étudier d'autres quantités : les températures (relevées à 10h dans la cour du lycée) durant 1 mois, la taille des élèves d'une classe, ... Tout cela pousse à introduire quelques mots de vocabulaire à connaitre.

Définition 5.1.1. *Une série d'observations ou série statistique se définit à partir de deux paramètres :*

1. *Une* ***population*** *: l'ensemble des individus (ou objets) observés.*
2. *Un* ***caractère*** *qui est la qualité étudiée dans la population.*

Voici un moyen mnémotechnique pour ne pas confondre caractère et population. La population désigne **l'ensemble** des personnes qui vont être **interrogées** ; le caractère désigne la **question** qui va être **posée** à l'un des membres de la population.

Exemple 5.1.2. 1. Reprenons l'exemple 5.1.1. La population est Fanny (c'est elle qui est interrogé), le caractère étudié correspond aux notes obtenues par Fanny durant un trimestre (Fanny doit indiquer qu'elle note elle a eu à ses DS).

2. Si nous nos intéressons à la taille des élèves d'une classe, la population est *la classe* et le caractère étudié est *la taille*.

Remarque. Observons que le caractère étudié peut-être de nature diverse :

- **qualitatif** lorsqu'il n'est pas numérique.
- **quantitatif discret** lorsqu'il peut prendre un nombre **fini de valeurs numériques**.
- quantitatif continu lorsqu'il peut prendre un nombre infini de valeurs réelles. (cette notion ne sera pas abordée en classe de seconde)

Exemple 5.1.3. 1. Supposons que nous ayons un sondage à disposition. Celui-ci a été réalisé auprès de 1000 personnes (composant la population étudiée) pour connaître leur intention de vote au second tour d'une élection (il s'agit du caractère étudié). Les réponses possibles de ce sondage sont :

   - « oui »
   - « non »
   - « ne se prononce pas ».

   Il s'agit donc d'un caractère **qualitatif**.

2. Un professeur reporte les notes de son dernier contrôle sur son ordinateur. Pour chaque copie (l'ensemble des copies correspond à la population), il a attribué une note (correspondant au caractère étudié) pouvant aller de 0 à 20 (avec une précision allant jusqu'au demi-point : $12,5/20$ par exemple). Il s'agit donc d'un **caractère quantitatif discret**.

Cette année, nous allons principalement nous focaliser sur des caractères **quantitatifs discrets** (des notes par exemple).

Dans les deux sections qui vont suivre nous allons chercher à calculer des paramètres permettant de « résumer » une série statistique. Les formules peuvent sembler compliquées mais nous allons observer que la **calculatrice va s'occuper des calculs pénibles à notre place** (cf. fin de la section 5.2).

5.2 Moyenne

A la fin d'un trimestre, l'enseignant d'une matière calcule la moyenne de vos notes pour se faire une idée de votre niveau. Nous allons voir comment faire ce genre de calcul à partir de n'importe quelle série statistiques.

Définition 5.2.1 (Moyennes pondérées)**.** *Supposons que nous ayons à disposition la série statistique suivante :*

Valeurs	x_1	x_2	$\dots$	x_N
Effectifs	n_1	n_2	$\dots$	n_N

Ce tableau signifie que les valeurs $x_1, \dots, x_N$ sont respectivement affectées de coefficients $n_1, \dots, n_N$. Dans ce cas, la moyenne pondérée est donnée par

$$\bar{x} = \frac{x_1 \times n_1 + x_2 \times n_2 + \ldots + x_N \times n_N}{n_1 + \ldots + n_N}$$

et la somme des effectifs $n_1 + \ldots + n_N$ *correspond à l'effectif total* N *de la série.*

Considérons l'exemple de Raphaël.

Exemple 5.2.1. Raphaël a obtenu les notes suivantes :

Valeurs (notes)	6	12	7	14	10
Effectifs (coefficients)	2	5	3	6	4

Dans ce cas, la moyenne pondérée de ses notes vaut alors

$$\bar{x} = \frac{6 \times 2 + 12 \times 5 + 7 \times 3 + 14 \times 6 + 10 \times 4}{2 + 5 + 3 + 6 + 4} = \frac{217}{20} = 10,85.$$

La moyenne (pondérée) de Raphaël vaut donc $10,85/20$.

Voici un autre exemple dans lequel nous calculons une moyenne pondérée.

Exemple 5.2.2. Imaginons que nous ayons à disposition ce tableau résumant une série de dons.

Don (en euros)	10	15	20	30	50	Total
Effectif	12	17	10	11	5	55

C'est pourquoi, le don moyen vaut $\bar{x} = \frac{12\times10+17\times15+10\times20+11\times30+5\times50}{55} = 21$ euros.

Manipulation calculatrice : les lien suivants expliquent (pour les TI) comment utiliser sa calculatrice pour déterminer une moyenne :

- (sans pondération) https://www.youtube.com/watch?v=_q7MKnLOFe4&feature=youtu.be
- (avec pondération) https://www.youtube.com/watch?v=JPTDZtSrd2o&feature=youtu.be

Remarque. L'écart-type σ et les quartiles Q_1, Med, Q_3 seront étudiés dans le reste du chapitre.

5.3 Variance et écart-type

La moyenne seule n'est qu'un outil limité ne tenant pas compte de certains aspects d'une série statistiques. Observons cela sur un exemple.

Exemple 5.3.1. Imaginons que Ioana ait obtenu les notes suivantes

$$9\,;\,9\,;\,11\,;\,11$$

tandis que Sofiane a obtenu les notes

$$1\,;\,1\,;\,19\,;\,19$$

Il n'est pas difficile de montrer que ces deux séries ont la même moyenne : $10/20$, pourtant les deux séries statistiques semblent vraiment différentes.

L'exemple précédent nous pousse à introduire une nouvelle quantité, la variance et l'écart-type. Ces deux nouvelles quantités apportent de nouvelles informations sur une série statistiques. Les formules suivante semblent peu simples, nous rappelons qu'**en pratique ces valeurs sont obtenues grâce à la calculatrice**.

Définition 5.3.1. *La variance de la série statistique*

Valeurs	x_1	x_2	...	x_N
Effectifs	n_1	n_2	...	n_N

est donnée par la formule suivante :

$$\mathrm{V} = \frac{n_1(\bar{x} - x_1)^2 + n_2(\bar{x} - x_2)^2 + \ldots + n_N(\bar{x} - x_N)^2}{n_1 + n_2 + \ldots + n_N}.$$

L'écart-type σ *(lire* sigma*) de la série est ensuite donné par*

$$\sigma = \sqrt{V}.$$

Remarque. Voici quelques mots permettant de mieux comprendre ce que signifie ces nouvelles quantités.

Pour fixer les idées, l'écart type σ permet de quantifier de quelle manière les valeurs se répartissent autour de la moyenne. Prenons l'exemple suivant pour illustrer ceci.

Imaginons qu'une classe ait obtenu une moyenne de 11 à un devoir. L'enseignant décide alors de calculer l'écart-type (associé à la série statistique des notes du devoir) pour obtenir plus d'information. Si σ est **grand** ($\sigma = 6$ par exemple), grossièrement cela signifie que certains élèves ont au 6 points de plus par rapport à la moyenne tandis que d'autres ont eu 6 points de moins par rapport à la moyenne. Il est possible de montrer qu'une large partie de la classe a donc ses notes comprises entre $[\bar{x} - \sigma; \bar{x} + \sigma] = [5; 17]$. Cela signifie que la classe a un **niveau plutôt hétérogène**.

Au contraire, si σ est **petit** ($\sigma = 1,5$ par exemple). La majorité des notes sera comprise entre $[9,5; 12,5]$ attestant que la classe a un **niveau homogène**.

σ peut aussi s'interpréter comme une **mesure de précision**, plus celui-ci est petit plus les valeurs de la série vont rester proche de la moyenne. Cela peut notamment s'utiliser en sport si nous décidions de faire des statistiques sur les tirs réussis d'un joueur de basket. Plus σ sera petit, plus le sportif sera régulier et obtiendra des scores proches de son score moyen.

Voyons si l'écart-type permet de différencier les notes obtenues par Ioana de celles obtenues par Sofiane.

Exemple 5.3.2. Les deux séries statistiques étaient :

$$\text{Ioana :} \quad 9\,;\,9\,;\,11\,;\,11 \quad \text{et} \quad \text{Sofiane :} \quad 1\,;\,1\,;\,19\,;\,19.$$

Ces deux séries ont la même moyenne : 10/20. Calculons la variance et l'écart-type de ces deux séries :

1. Pour la première série (celle de Ioana), nous avons

$$V_1 = \frac{2(10-9)^2 + 2(10-11)^2}{4} = 1 \quad \text{et} \quad \sigma_1 = \sqrt{V_1} = 1$$

   et donc $\sigma_1 = \sqrt{1} = 1$.
2. pour la deuxième série (celle de Sofiane), nous obtenons

$$V_2 = \frac{2(10-1)^2 + 2(10-19)^2}{4} = 81 \quad \text{et} \quad \sigma_2 = \sqrt{V_2} = 9$$

   et donc $\sigma_2 = \sqrt{9} = 3$.

Puisque $\sigma_1 < \sigma_2$, nous constatons bien que **Ioana est plus régulière** dans ses résultats que Sofiane.

A nouveau, il est essentiel de savoir **utiliser sa calculatrice** pour effectuer ce genre de calculs (variances, moyennes, écarts-types). Des liens vers des tutoriels ont été donné plus haut dans le cours.

5.4 Description par quantiles

Pour décrire une série statistique nous allons déterminer des valeurs associées : il s'agit de quantile particuliers. Nous allons nous focalisé sur la médiane ainsi que sur le premier et dernier quartile permettant de mesurer la dispersion de la série autour de sa médiane.

Définition 5.4.1. *Soit $(x_1, \ldots, x_N)$ une série statistique à caractère quantitatif. Voici la définition de certains quantiles.*

- *La médiane est :*
  1. *la valeur* ***centrale*** *lorsque n est* ***impair****.*
  2. *la* ***moyenne des deux valeurs centrales*** *lorsque n est* ***pair****.*

  Dans tous les cas, nous noterons la médiane par Med.
- *Le premier quartile Q_1 est la plus petite valeur de la série telle que 25% des valeurs de la série lui soient inférieures ou égales.*
- *Le troisième quartile Q_3 est la plus petite valeur de la série telle que 75% des valeurs de la série lui soient inférieures ou égales.*

  La quantité $Q_3 - Q_1$ est appelé ***écart interquartile*** *et l'intervalle associée $[Q_1; Q_3]$ correspond alors à 50% de la population étudiée.*

Remarque. Rappelons que « l'étendue » d'une série statistiques est la différence entre la plus grande valeur de la série (son maximum) et la plus petite (son minimum). Nous la noterons e.

Exemple 5.4.1. Observons la série statistique suivante :

Longueur (en m)	37	39	40	41	42	43	44	48
Effectif	4	3	4	2	2	4	5	2
Effectif cumulés	4	7	11	13	15	19	24	26

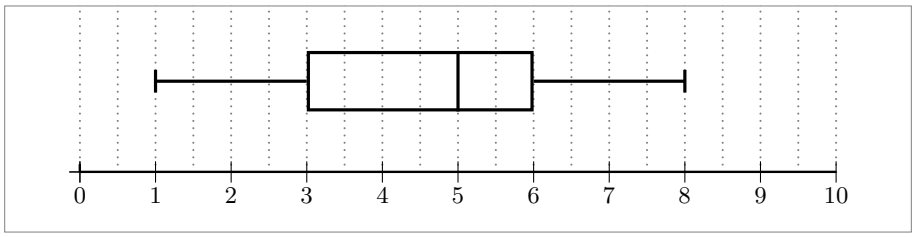
1. La longueur médiane des lancers de javelot présentés dans ce tableau est Med $= 41,5m$. En effet, l'effectif total est **pair** (ici $N = 26$) donc la médiane est la moyenne des 13ème et 14ème longueurs ; lesquelles sont égales à $41m$ et $42m$.
2. Il est facile de déterminer le 1er quartile, puisque $\frac{26}{4} = 6,5$, le premier quartile Q_1 est la 7ème longueur, à savoir : $Q_1 = 39m$.
3. Le calcul précédent permet aussi d'obtenir facilement le troisième quartile : Q_3 est la 20ème longueur, puisque $3 \times 6,5 = 19,5$, à savoir : $Q_3 = 44m$.
4. Enfin, l'intervalle $[Q_1; Q_3] = [39; 44]$ contient 50% de la population.

L'ensemble de ces informations donne une nouvelle description (en complément de celle fournie par la moyenne $\bar{x}$ et l'écart-type σ) de la série $(x_1, \ldots, x_N)$. Elles sont résumées dans la représentation graphique suivante.

Définition 5.4.2. *Un diagramme de Tukey (aussi appelé « boîte à moustache ») est un résumé, sur un axe gradué, des quantiles définis ci-dessus. Ce diagramme est constitué*

- *d'une boîte (dont la hauteur est prise de manière arbitraire) délimitée par le* 1*er et* 3*ème quartiles. Cette même boite est ensuite partagée par la médiane.*
- *de « moustaches » qui relient les quartiles aux valeurs extrêmes de la série.*

Un exemple est donné ci-dessous correspondant aux notes (sur 10) d'étudiants.



Exemple 5.4.2. Sur le diagramme précédent, nous lisons donc :

- La médiane Med vaut 5. Ainsi la moitié des élèves ont eu plus de la moyenne.
- Le premier quartile Q_1 vaut 3 et le troisième quartile Q_3 vaut 6.
- Les valeurs extrêmes valent 1 pour le minimum et 8 pour le maximum.