Chapitre 6

Statistiques

6.1 Introduction

La Statistique (l'étude de données statistique) est relativement récente (bien qu'il existe de nombreuse traces, dans l'Histoire, de listes d'objets ou de nombres) et fait partie des mathématiques traitant les évènements aléatoires.

A COMPLETER

Quant à elle, la boite à « moustaches »a été inventée par Tukey en 1977.

6.2 Rappels

Dans ce chapitre, nous considèrerons une série de n observations ordonnées, notées $x_1, \ldots, x_n$, avec $n \in \mathbb{N}$.

Voici quelques mots de vocabulaire à connaitre .

Définition 6.2.1. *Une série d'observations, ou série statistique, se définit à partir de deux paramètres :*

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

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

- 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.

Exemple 6.2.1. Supposons que nous ayons un sondage à disposition. Celui-ci a été réalisé auprès de n 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 » et « Ne se prononce pas »(il s'agit caractère qualitatif).

Exemple 6.2.2. 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 un pas de $0,25$ (il s'agit donc d'un caractère quantitatif discret).

Dans toute la suite, nous essayerons de décrire une série statistique à caractère quantitatif à partir de certains indicateurs. Certains d'entres eux ont déjà été vus au collège ou en classe de seconde.

6.3 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 6.3.1. *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 correspond 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 6.3.1. 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 $\text{Med} = 41,5m$. 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 $\frac{26}{4} = 6,5$, Q_1 est la 7ème longueur, à savoir : $Q_1 = 39m$.

3. Ce qui permet d'obtenir le troisième quartile : Q_3 est la 20ème longueur, puisque $3 \times 6,5 = 19,5$, à savoir : $Q_3 = 44m$.

4. Enfin, l'écart interquartile correspond donc à $[39; 44]$.

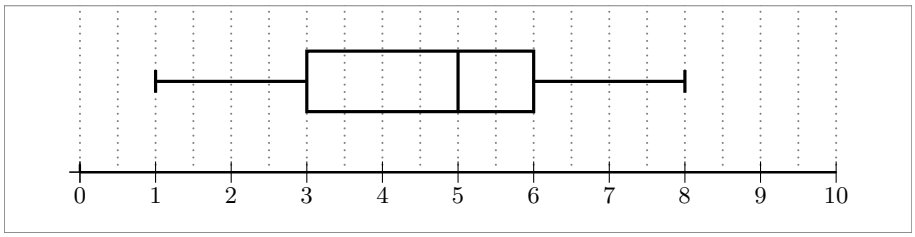
L'ensemble de ces informations donne une première description de la série $(x_1, \ldots, x_n)$. Elles sont résumées dans la représentation graphique suivante.

Définition 6.3.2. *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 6.3.2. Sur le diagramme précédent, nous lisons donc :

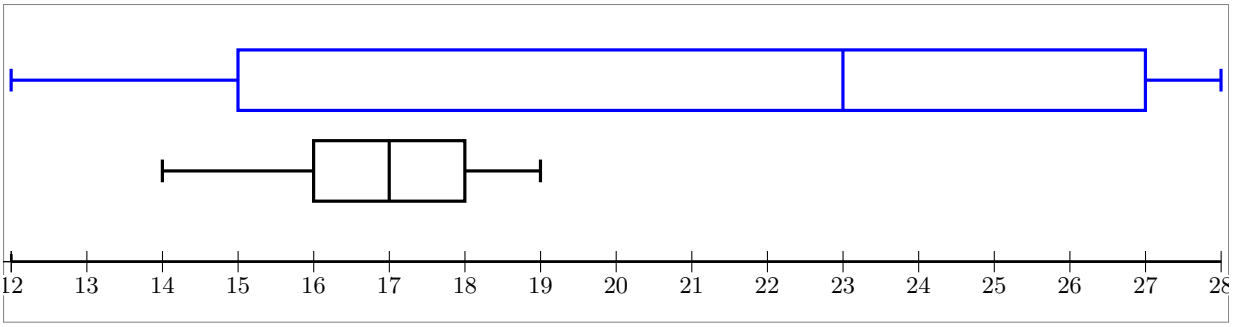
- La médiane Med vaut 5. Ainsi la moitié des élèves ont eu plus de la moyenne.

- 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.

Exemple 6.3.3. Il est souvent utile de comparer des diagrammes de ce genre pour émettre des hypothèses. Par exemple, imaginons que nous ayons relevé, à différents intervalles (toutes les heures, pendant quatre jours), les températures dans une forêt (en noir) et dans un champ (en bleu) proche

de cette même forêt. Ces mesures ont donné lieu aux diagrammes suivants.



Quelle semble être l'influence des arbres sur la température à l'intérieur de la forêt ?

- Ces diagrammes montrent que les températures sont beaucoup plus dispersées dans les champs. Il est possible de supposer que les arbres permettent de maintenir une température plus stable.

- Ces diagrammes montrent aussi que les températures sont globalement plus basses en forêt, il est possible émettre l'hypothèse que les arbres aident à conserver la fraîcheur.

6.4 Description par moyennes

Dans cette section, nous présentons une autre manière de décrire une série statistique. Plaçons nous à présent dans le cadre d'une série statistique $(x_k; n_k)$, $k = 1, \ldots, l$, $r \in \mathbb{N}$ où les valeurs distinctes $x_1, \ldots, x_l$ ont pour effectif respectif $n_1, \ldots, n_l$ (ou pour fréquence $f_1, \ldots, f_r$). L'effectif total de la série est noté N où $N = n_1 + \ldots + n_l$.

Définition 6.4.1. *La moyenne de la série statistique $(x_k; n_k)$ où $1 \leq k \leq l$ est le nombre m (aussi noté $\bar{x}$) défini par :*

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{l} n_i x_i = \sum_{i=1}^{r} f_i x_i$$

Exemple 6.4.1.

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

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

Similairement au cas de la médiane, la moyenne seule n'est qu'un outil limité ne tenant pas compte de la dispersion. Pour palier ce manque, nous définissons les quantités suivantes.

Définition 6.4.2. *. La variance de la série statistique $(x_k; n_k)$, pour $1 \leq k \leq l$, est notée* V *est définie par*

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

Remarque. Observons que la variance est simplement la moyenne des écarts à la moyenne au carré. Autrement dit, la moyenne des $(\bar{x} - x_i)^2$. Par souci d'homogénéité, nous utiliserons plus souvent la racine carrée de la variance notée $\sigma = \sqrt{V}$ et appelée écart type de la série.

Exemple 6.4.2. Comme nous allons le voir sur l'exemple suivant, la variance permet d'en apprendre plus sur une série statistique et complète l'information apportée par la moyenne.

Considérons les deux séries statistiques suivantes :

$$S_1 : \{9; 9; 11; 11\} \quad \text{et} \quad S_2 : \{1; 1; 19; 19\}.$$

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

1. Pour la première série S_1, nous avons

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

2. pour la deuxième série S_2, nous obtenons

$$V2 = \frac{2(1-10)^2 + 2(19-10)^2}{4} = 81 \quad \text{et} \quad \sigma_2 = \sqrt{V_2} = 9$$

6.5 Comparaison de séries

La description d'une série passe souvent par le choix d'un paramètre dit de position (moyenne ou médiane) qui donne une première information. Cette information, partielle, est complétée par un paramètre de dispersion correspondant (écart interquartile ou écart type). Ceci donne un couple de paramètres qui offre une première synthèse des observations et facilite la comparaison de séries. En effet,

1. Le couple $(\text{Med}; Q_3 - Q_1)$ donne une indication de la tendance centrale de la série (médiane) et de la concentration de la moitié des données autour de cette dernière (écart interquantile). Il est peu sensible aux valeurs extrêmes. En revanche, l'écart interquartile ne prend en compte que la moitié de la population et peut ne pas être représentatif.

2. Le couple $(m; \sigma)$ donne une indication de la tendance moyenne de la série (moyenne) et mesure le carré des écarts à cette dernière (écart type). Il est très sensible aux valeurs extrêmes. En revanche, il s'avère très efficace dans le cas de séries symétriques autour de la moyenne (phénomène rencontré par exemple dans les sondages).

Remarque. Comme nous le verrons, la calculatrice permet facilement d'obtenir ces différentes valeurs associées à une série statistique.

6.6 Bilan

- Déterminer la nature d'une série statistique (population, caractère...).

- Déterminer la boîte à moustache correspondant à une série statistique.
- Savoir calculer la moyenne, la variance et l'écart type d'une série statistique.
- Comparer deux séries et commenter.