Chapitre 12

Statistique à une variable

12.1 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.

12.2 Rappels

Dans ce chapitre, nous considèrerons une série de p observations ordonnées, notées $x_1, \ldots, x_p$, avec $p \in \mathbb{N}$. Par exemple, pour fixer les idées, on peut penser aux notes obtenues par une classe lors d'un devoir surveillé.

12.2.1 Vocabulaire

Voici quelques mots de vocabulaire à connaitre.

Définition 12.2.1. *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. 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.

Les exemples ci-dessous montrent que le caractère étudié peut-être de nature diverses.

Exemple 12.2.1. 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 » et « Ne se prononce pas »(il s'agit 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 un pas de $0,25$ (il s'agit donc d'un caractère quantitatif discret).

12.2.2 Moyenne et écart-type

Dans cette section, nous présentons deux valeurs associées une série statistique.

Définition 12.2.2. *Si chaque valeur $x_1, \ldots, x_p$ est, respectivement, associé à un **effectif** (coefficient) $n_1, \ldots n_p$ alors la moyenne pondérée $\bar{x}$ de la série statistique est donnée par :*

$$\bar{x} = \frac{x_1 n_1 + x_2 n_2 + \ldots + x_p n_p}{n_1 + \ldots + n_p}$$

Exemple 12.2.2.

1. Imaginons qu'un élève ait obtenu les notes suivantes durant un trimestre $6; 12; 7; 14; 10$ lesquelles sont affectées, de manière respective, des coefficients $2; 5; 3; 6; 4$, la moyenne pondérée de ses notes vaut alors

$$\bar{x} = \frac{2 \times 6 + 5 \times 12 + 3 \times 7 + 6 \times 14 + 4 \times 10}{2 + 5 + 3 + 6 + 4}$$

2. Voici un deuxième exemple

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.

La moyenne seule n'est qu'un outil limité ne tenant pas compte de la **dispersion** des valeurs de la série statistique. Pour palier ce manque, nous définissons les quantités suivantes.

Définition 12.2.3. *. La **variance** de la série statistique $(x_k; n_k)$, pour $1 \leq k \leq p$, est notée* V *est définie par*

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

***L'écart-type** σ de la série statistique vaut alors $\sigma = \sqrt{V}$.*

Remarque. Pour fixer les idées, l'écart type 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 réussi d'un joueur de basket. Plus σ sera petit, plus le sportif sera régulier et obtiendra des scores proches de son score moyen.

Il est essentiel de savoir utiliser sa calculatrice pour effectuer ce genre de calculs (variances, moyennes) à l'aide des listes. Il faut également avoir à l'esprit que **la moyenne et la variance sont sensibles aux valeurs extrêmes.**

Exercices à traiter : 39 page 165 ; 29, 30 page 163.

12.3 Quantiles

Dans cette section nous présentons d'autres paramètres qui peuvent être utilisés pour résumer une série statistique. Cette approche est complémentaire de celle reposant sur la moyenne et l'écart-type. Une partie de ce qui va être présentée a déjà été vu en classe de seconde.

12.3.1 Série statistique ordonnée et effectif cumulés croissant

Par la suite, il sera essentiel de ranger une série statistique par ordre croissant.

Exemple 12.3.1. 1. La série 1 ; 3 ; 2 ; 5 ; 7 ; 5 n'est pas rangée par ordre croissant.

2. La série 1 ; 2 ; 3 ; 5 ; 5 ; 7 est rangée par ordre croissant.
L'effectif total N vaut 6.

Lorsque qu'une série statistiques sera donnée, il faudra déterminer les **effectifs cumulés croissants** : il s'agit simplement d'additionner les uns après les autres les effectifs jusqu'à obtenir l'effectif total.

Exemple 12.3.2. 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

La deuxième valeur des effectifs cumulés croissant vaut 7 et a été obtenu en additionnant les effectifs 4 et 3 ; la troisième valeur 11 s'obtient à ajoutant l'effectif 4 à la valeur 7, et ainsi de suite.

12.3.2 Quantiles

Pour décrire une série statistique, dans un premier temps, nous allons nous focaliser sur la **médiane** ainsi que sur le **premier et dernier quartile**. Nous allons voir comment obtenir ces valeurs et ce qu'elles représentent. Dans un second temps, nous introduiront la notion de **décile**.

12.3.3 Médiane

Définition 12.3.1. *Soit $(x_1, \ldots, x_p)$ une série statistique (à caractère quantitatif) ordonnée. La médiane de la série statistique est une valeur (notée* Med*) telle que :*

1. 50% *des valeurs de la série sont inférieures ou égales à* Med *;*
2. 50% *des valeurs de la série sont supérieures ou égales à* Med

Remarque. 1. Il est important d'observer que la médiane est un paramètre de position : sa valeur ne dépend que de sa place (centrale) dans la série statistique ordonnée. En particulier, **la médiane est insensible aux valeurs extrêmes** ; ce qui n'était pas le cas de la moyenne. Dis plus simplement, en considérant les notes obtenues par des élèves à un devoir, les notes extrêmes (0 ou 20) affectent beaucoup la valeur de la moyenne mais n'ont pas d'impact sur la médiane.

2. En pratique, **lorsque les valeurs sont rangées par ordre croissant**, nous avons deux cas de figures :
   (a) si l'effectif total N est impair la médiane est la $\frac{N+1}{2}$ valeur (correspondant à la valeur centrale) ;
   (b) si N est pair la médiane est la moyenne de la $\frac{N}{2}$-ième valeur et de la $\frac{N+1}{2}$ valeur (correspondant aux deux valeurs centrales).

Voyons sur un exemple.

Exemple 12.3.3. 1. Le 1er juillet 2018, la limitation de vitesse à 80 km/h est entrée en vigueur sur certaines routes secondaires. Le tableau suivant donne les vitesses mesurées pendant 10 minutes consécutives sur l'une de ces routes.

Vitesse (en mk/h)	76	77	79	80	82	87	97
Effectifs	11	8	11	13	5	6	1
Effectifs cumulés	11	19	30	43	48	54	55

L'effectif total 55 est impair. Puisque $\frac{55}{2} = 27,5$, la médiane correspond à la 28 valeur. La ligne des effectifs cumulés croissant montre que la vitesse associée à cette 28ième valeur vaut 79 km/h. Autrement dit,

$$\text{Med} = 79.$$

Un simple dessin montre qu'il y a bien 50% des effectifs (i.e 27 personnes) qui roulaient à une vitesse supérieure à 79 km/h et 50% des effectifs qui roulaient à une vitesse inférieure à 79 km/h.

2. La série statistique suivante correspond à l'ensemble des notes (sur 10) obtenues par un groupe d'élève.

Notes	3	4	4	6	7	7	8	9
Effectifs	1	1	1	1	1	1	1	1
Effectifs cumulés	1	2	3	4	5	6	7	8

Ici, l'effectif total 8 est pair et $\frac{8}{2} = 4$ nous devons donc faire la moyenne de la 4ième et 5ième valeur. Ainsi,

$$\text{Med} = \frac{6+7}{2} = 6,5$$

En résumé, la moitié des élèves ont eu plus de $6,5/10$ et la moitié des élèves ont eu moins de $6,5/10$.

12.3.4 Quartiles

La médiane permet de partager une série statistique en 2 partie de taille égale, nous allons maintenant voir comment partager à nouveau ces deux parties en deux. Il s'agit de la notion de quartiles (qui partage donc la série en 4 parts égales).

Définition 12.3.2. *Soit $(x_1, \ldots, x_p)$ une série statistique (à caractère quantitatif) ordonnée. Voici la définition de premier et troisième quartiles :*

- *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.*

  Au niveau de la terminologie, la quantité $Q_3 - Q_1$ est appelé « écart interquartile » ; notons que l'intervalle $[Q_1; Q_3]$ contient 50% de la population étudiée.

Remarque.
1. **L'écart interquartile** est un **paramètre de dispersion** associé à la médiane. Il joue un rôle analogue à l'écart-type (qui lui est associé à la moyenne). Plus il est grand, plus les valeurs sont dispersées autour de la médiane.
2. Voici comment calculer ces valeurs en pratique :
   (a) il suffit de calculer $\frac{N}{4}$, le rang de Q_1 est le nombre entier immédiatement supérieur à $\frac{N}{4}$.
   (b) il suffit de calculer $\frac{3N}{4}$, le rang de Q_3 est le nombre entier immédiatement supérieur à $\frac{3N}{4}$.

Le graphique suivant permet de visualiser plus facilement ces informations.

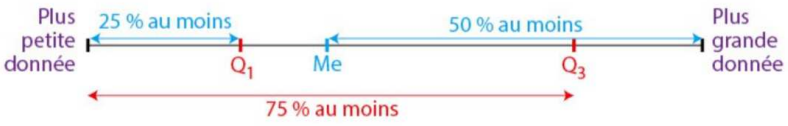


FIGURE 12.1: Résumé graphique

Voyons sur un exemple.

Exemple 12.3.4. Observons la série statistique suivante, obtenue à partir de lancers de javelots :

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 $\frac{26}{4} = 6,5$, Q_1 est la 7ème longueur, à savoir : $Q_1 = 39m$.
3. Ce raisonnement permet aussi 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 vaut donc à 5.

Définition 12.3.3. *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 noterons cette quantité e.*

Exemple 12.3.5. En reprenant l'exemple du lancer de javelots, nous constatons que $e = 48 - 37 = 11$.

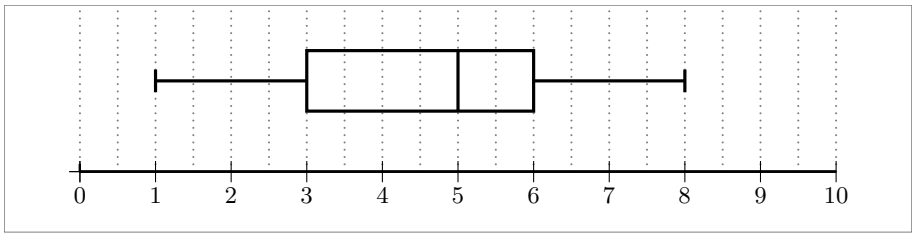
12.3.5 Diagramme en boite

Il est possible de préciser le graphique précédent (cf. figure 8.1) afin d'obtenir une description plus complète de la série $(x_1, \ldots, x_p)$.

Définition 12.3.4. *Un diagramme de Tukey (aussi appelé « boîte à moustache ») résume de manière graphique, sur un axe gradué, les valeurs définies plus tôt (médiane, quartiles, maximum et minumum). Ce diagramme est constitué*

- *d'une boîte (dont la hauteur est prise de manière arbitraire) délimitée par Q_1 et Q_3 (le premier et troisième quartile). Cette même boite est ensuite partagée par la médiane* Med.
- *de « moustaches » qui relient les quartiles aux valeurs extrêmes (maximum et minimum) de la série.*

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



Exemple 12.3.6. 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.

Remarque. Comme nous le verrons en exercice, la comparaison des diagrammes en boite permet facilement de comparer deux séries statistiques en utilisant uniquement les quantiles.

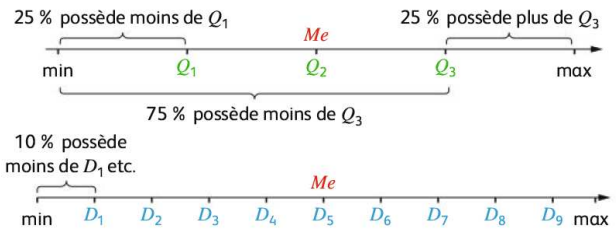
Exercice à traiter : 40 page 165 et 32 page 164.

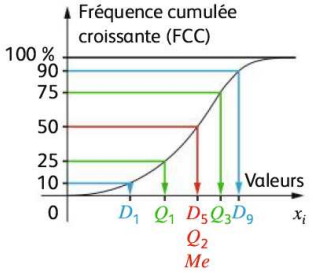
12.3.6 Déciles

Cette partie est nouvelle mais, comme nous allons le voir, les notions abordées sont relativement simples et ressemblent beaucoup à ce qui a déjà été vu pour les quartiles et la médiane. Cette fois-ci, au lieu de découper la série statistiques en quart nous allons la **découper en dixième**.

Définition 12.3.5. *Pour k allant de 1 à 9, le $k$ème **décile** D_k est la plus petite valeur de la série telle que $k \times 10\%$ des valeurs de la série lui soient inférieures ou égales.*

Remarque. Le rapport interdécile $\frac{D_9}{D_1}$ met en évidence le rapport entre le « haut »et le « bas » de la série statistique, il peut alors servir d'indicateur d'inégalité de répartition. Comme pour les quartiles, il est possible de représenter les déciles sur un diagramme en boite.





Exercices à traiter : 1,2 page 157

12.4 Valeur moyenne et indice de Gini

Dans cette section nous mettons une nouvelle approche en lumière, celle-ci s'appuie sur des outils issus de l'analyse.

12.4.1 Valeur moyenne

Parfois, certains phénomènes sont modélisés par des fonctions plutôt qu'une série statistique.

Définition 12.4.1. *Soit f une fonction continue et positive sur $[a;b]$. Sa valeur moyenne μ est donnée par*

$$\mu = \frac{1}{b-a}\int_a^b f(x)dx$$

Remarque. Cette nouvelle quantité s'interprète comme $\bar{x}$.

Exemple 12.4.1. Une chaine de télévision étudie son audience journalière depuis son lancement en 2 000. On admet que le nombre journalier de téléspectateurs (en milliers), est modélisé sur $[0;+\infty[$ par la fonction $f(t) = (20x^2 - 80x + 460)e^{-0,1x}$.

L'audience journalière moyenne entre le 1er janvier 2018 et le 1er janvier 2 020 est donnée par

$$\mu = \frac{1}{20-18}\int_{18}^{20} f(x)dx \approx 920,486.$$

Exercices à traiter : 34, 35, 37, 38 page 164.

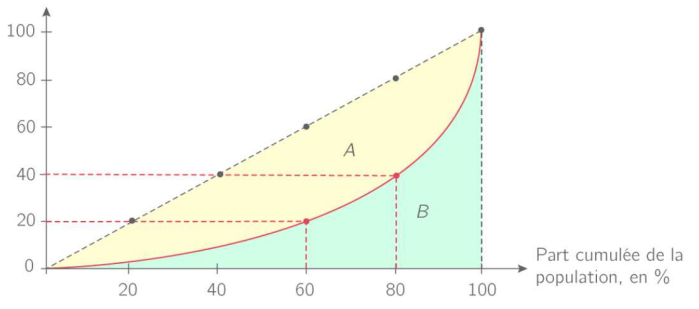
12.4.2 Courbe de Lorenz et indice de Gini

Dans ce chapitre, nous avons rencontré plusieurs objets permettant de décrire la dispersion des valeurs d'une série statistiques :

- l'écart-type σ.
- l'écart interquartile $Q_3 - Q_1$
- le rapport $\frac{D_9}{D_1}$.

La plupart du temps, en interprétant, ces objets sont également utiles pour mesurer des inégalités de répartition. Dans notre nouveau contexte, utilisant des intégrales, nous allons devoir introduire une nouvelle manière de mesurer les inégalités. A cet effet, débutons par un exemple afin d'introduire la notion de **courbe de Lorenz.**

Exemple 12.4.2. Considérons le graphique ci-dessous :



Courbe de Lorenz de la distribution du revenu disponible des ménages en 2006

Nous constatons alors que :

- Les 60% de la population les plus pauvres représentent 20% des revenus du pays.
- 40% des revenus du pays sont détenus par 80% de la population les plus pauvres. Ainsi, 60% des revenus appartiennent au 20% de la population les plus riches.

Plus formellement, une courbe de Lorenz se définit comme suit.

Définition 12.4.2. ***Une courbe de Lorenz*** *est une représentation graphique qui met en relation la proportion x (en %) d'une population détentrice d'une proportion y (en %) de la grandeur étudiée (revenus,...).*

Cette courbe est un **moyen de quantifier les inégalités**. En effet, la répartition serait **équitable si la courbe de Lorenz** (en rouge sur la figure ci-dessous) **correspondait à la droite** $y = x$ (en vert sur la figure ci-dessous). Autrement dit, x% de la population détient x% de la richesse du pays. L'aire $\mathcal{A}$ permet de quantifier cet écart. Il s'agit de **l'indice de Gini**.

Définition 12.4.3. ***L'indice de Gini*** *(noté G) est égal au double de l'aire $\mathcal{A}$ de la partie délimitée par la courbe rouge et le segment $[OA]$:*

$$G = 2 \times \mathcal{A}$$

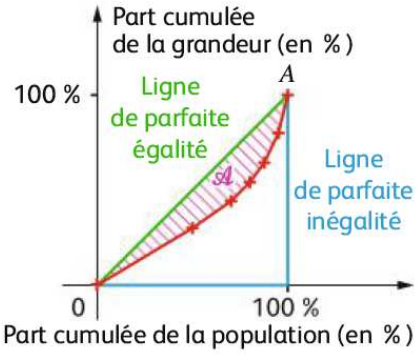


FIGURE 12.3: Indice de Gini

Remarque. G est un nombre compris entre 0 (égalité parfaite) et 1 (inégalité maximale). Plus les inégalités sont importants pour la courbe de Lorenz s'éloigne du segment $[OA]$ et plus l'indice de Gini est élevé.

Exercices à traiter : 42, 43 , 47, 49 page 165-166 ; 50 page 168 en DM.