Chapitre 5

Statistiques (1ère partie)

5.1 Introduction

La Statistique (l'étude de données statistiques) 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.

5.2 Vocabulaire

Dans ce chapitre, nous considèrerons une série de n observations ordonnées, notées $x_1, \ldots, x_N$, 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 5.2.1. 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 5.2.1. *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 5.2.2. 1. Reprenons l'exemple 5.2.1. 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*.

Exercices à traiter : à partir des énoncés des exercices 14 page 286 et 28 page 287, indiquer à chaque fois quelle est la population étudiée et quel est le caractère mis en jeu. *Il n'est pas demandé de traiter les questions (du livre) indiquées dans ces énoncés.*

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 5.2.3. 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).

Exercice à traiter : Donner un exemple de caractère qualitatif et un exemple d'un caractère quantitatif discret. *Bien entendu, il ne faut pas reprendre les exemples du cours mais en proposer de nouveaux.*

5.3 Effectifs et fréquences

Le document suivant concerne les langues parlées dans le monde :

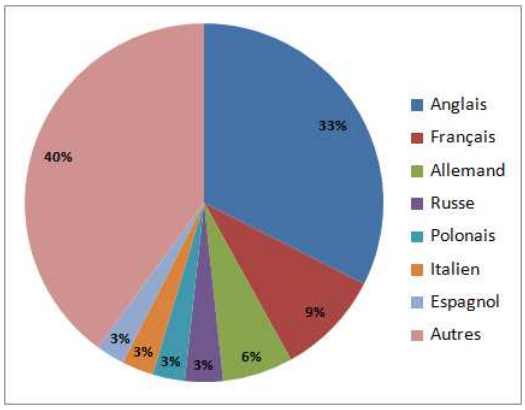

FIGURE 5.1: Fréquences des langues parlées dans le monde

Cela permet de constater que 33% de la population mondiale (1 personne sur 3 environ) parle anglais. Comment avons-nous obtenu ce **pourcentage** ? Pour déterminer ces valeurs, il est nécessaire d'introduire la notion de **fréquence**.

Définition 5.3.1. *Etant donné une série statistique*

$$x_1, \ldots, x_N \quad \textit{d'effectif total} \quad N \in \mathbb{N},$$

il est possible de déterminer la fréquence d'apparition d'une des valeurs obtenues $(x_1, x_2, \ldots)$ *par la formule suivante :*

$$\textit{fréquence d'une valeur} = \frac{\textit{effectif de la valeur}}{\textit{effectif total}}.$$

Remarque. Cette formule est similaire à ce que nous ferons en **probabilité** dans une situation **d'équiprobabilité** avec le quotient

$$\frac{\text{nombre de cas favorables}}{\text{nombre de cas total}}.$$

Voyons ce qu'il faut faire en pratique.

Exemple 5.3.1. Sur un parking, nous étudions la couleur des voitures. Le caractère étudié est dit **qualitatif** (les valeurs ne sont pas numériques). La distribution des effectifs est donnée dans le tableau ci-dessous :

Couleur	**grise**	blanche	bleue	rouge
Effectif	18	7	5	2

Ainsi, l'effectif total vaut $N = 18 + 7 + 5 + 2 = 32$, l'effectif de la valeur **grise** vaut 18. En conséquence, la fréquence de la valeur **grise** vaut

$$\frac{18}{32} = \frac{9}{16} = 0,5625 = 56,25\%.$$

Autrement dit, plus de la moitié (56,25% pour être exact) des voitures du parking sont grises.

Voyons un deuxième exemple.

Exemple 5.3.2. Dans un village, nous avons dénombré les foyers selon leur nombre d'enfants. Voici les données obtenues

Nombre d'enfants par foyers	0	1	2	3	4
Effectif (nombre de foyers)	18	14	8	7	3

L'effectif total vaut $N = 18+14+8+7+3 = 47$, il s'agit du nombre total de foyers interrogés. La fréquence de la valeur 2 *enfants par foyers* vaut alors :

$$\frac{8}{47} \approx 0,17 = 17\%.$$

Remarque. Imaginons que la question suivante nous soit posée : quelle est la fréquence de la valeur *au moins* 2 *enfants par foyer* ? Cela signifie qu'il faut tenir compte des foyers avec 2 enfants (cela représente 8 foyers), ceux avec 3 enfants (cela représente 7 foyers) et ceux avec 4 enfants (cela représente 3 foyers). La fréquence voulue s'obtient donc en calculant

$$\frac{8+7+3}{47} = \frac{18}{47} \approx 0.38 = 38\%.$$

Autrement dit, environ 38% des foyers étudiés possèdent au moins 2 enfants.

Exercices à traiter : en utilisant le tableau (ne pas faire les questions de l'énoncé du livre) de l'exercice 28 page 287 :

1. Déterminer l'effectif total.
2. Quelle est la fréquence associée à la valeur 3 ?
3. Quelle est la fréquence associée à *le foyer possède* ***au moins*** 3 *véhicules* ?

Reprendre ces questions avec le tableau de l'exercice 12 page 286. La troisième question est modifiée en : Quelle est la fréquence associée à la valeur *le texte contient* ***au plus*** 2 *fautes de frappe* ?

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 5.4).

5.4 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 5.4.1. *La moyenne de la série statistique $x_1, \ldots, x_N$ est le nombre $\bar{x}$ défini par :*

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_N}{N}$$

Voyons cette formule en application sur un exemple.

Exemple 5.4.1. Imaginons que Raphaël ait obtenu les notes suivantes durant un trimestre 6 ; 12 ; 7 ; 14 ; 10. La moyenne de celui-ci vaut alors

$$\bar{x} = \frac{6 + 12 + 7 + 14 + 10}{5} = \frac{49}{5} = 9,8.$$

L'élève a donc une moyenne de $9,8/20$.

Parfois les notes sont affublées d'un coefficient (pour distinguer les DS des interrogations, par exemple). Voyons comment modifier la formule de la moyenne pour prendre en compte cela.

Définition 5.4.2 (Moyennes pondérées)**.** *Supposons que nous ayons à disposition la série statistique suivante :*

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

Ce tableau signifie que les valeurs $x_1, \ldots, x_N$ sont respectivement affectées de coefficients $n_1, \ldots, n_N$. 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.

Reprenons l'exemple de Raphaël.

Exemple 5.4.2. Cette fois-ci les notes de Raphaël sont prises en compte avec des coefficients :

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 5.4.3. 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 sera présenté dans la section suivante ; les quartiles Q_1, Med, Q_3 seront étudiés dans un chapitre ultérieur.

Exercices à traiter : 11 et 16 page 286.

La proposition suivante explique une propriété importante de la moyenne qui permet d'éviter certains calculs.

Proposition 20 (Linéarité de la moyenne)**.** *Soient $a, b \in \mathbb{R}$ et $x_1, \dots, x_N$ une série statistique de moyenne $\bar{x}$. Si $y_1, \dots, y_N$ est une nouvelle série statistique obtenue par la formule*

$$y_i = ax_i + b \quad \textit{pour tout} \quad i = 1, \dots, N$$

alors sa moyenne vaut

$$\bar{y} = a\bar{x} + b. \tag{5.4.1}$$

Voyons cela sur deux exemples.

Exemple 5.4.4.

1. Imaginons qu'un enseignant ait corrigé les devoirs maisons d'une de ses classes et que la moyenne obtenue (à partir des notes des élèves) est de $\bar{x} = 18/20$. Par la suite, il se rend compte que ses élèves ont tous triché et l'enseignant souhaite diviser toutes les notes par 2, quelle est la nouvelle moyenne de la classe $\bar{y}$?

   Au lieu de remplir un nouveau tableau avec les notes divisées par 2 des élèves pour ensuite calculer de nouveau la moyenne, il est préférable d'utiliser la formule (5.4.1) (avec les valeurs $a = \frac{1}{2}$ et $b = 0$). Nous trouvons alors que la nouvelle moyenne vaut

   $$\bar{y} = \frac{1}{2}\bar{x} + 0 = \frac{18}{2} = 9.$$

   La nouvelle moyenne de la classe est donc de 9/20.

2. Un enseignant corrige les copies d'une de ses classes à un DS et obtient la moyenne (obtenue à partir des notes des élèves) de $\bar{x} = 8/20$. L'enseignant se rend compte qu'il y avait une erreur dans l'énoncé et qu'il doit ajouter 3 points à tous les élèves. Quelle est la nouvelle moyenne $\bar{y}$?

Au lieu de remplir un nouveau tableau avec les nouvelles notes des élèves pour ensuite calculer de nouveau la moyenne, il est préférable d'utiliser la formule (5.4.1) (avec les valeurs $a = 1$ et $b = 3$). Nous trouvons alors que la nouvelle moyenne vaut

$$\bar{y} = 1 \times \bar{x} + 3 = 11.$$

La nouvelle moyenne de la classe est donc de 11/20.

Exercices à traiter : 20 page 286.

5.5 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 5.5.1. 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 5.5.1. *La variance de la série statistique*

Valeurs	x_1	x_2	$\ldots$	x_N
Effectifs	n_1	n_2	$\ldots$	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 5.5.2. 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.

Exercices à traiter : 39 page 288 et 40 page 289 ; (facultatif) faire l'exercice 45 page 289.

Exercice. Voici les notes obtenues, par une classe de seconde, après un interrogation de mathématiques (notée sur 10).

2nde 8 (1er groupe) : $7, 8, 3, 8, 9, 9, 4, 9, 3, 5, 7, 8, 7, 9, 6, 4, 10, 9, 10, 8, 7, 7, 6, 8, 7, 6, 9, 10, 8, 6$.

1. Reproduire et compléter le tableau suivant :

Note sur 10	3	4	5	6	7	8	9	10
Effectifs								

2. Calculer la moyenne $\bar{x}_1$ et l'écart-type σ_1 de ce groupe.
3. Voici les données qui ont été obtenu pour le deuxième groupe : $\bar{x}_2 = 7,233$ et $\sigma_2 = 1,41$. Quelle groupe a un niveau plus homogène que l'autre ? Justifier votre réponse.

5.6 Bilan du chapitre

Voici les savoirs faire à acquérir dans ce chapitre :

- Identifier la population interrogée et le caractère étudié.
- Calculer des fréquences, des effectifs,....
- Calcul de moyennes et de variances avec la calculatrice.
- Utiliser ces indicateurs (moyenne et variance) pour comparer des séries statistiques.