Chapitre 11

Statistiques

11.1 Introduction historique

> *La loi déterminée pour l'homme moyen sera celle qui exprimera le mieux ce qui a eu lieu dans la société.*
>
> Adolphe Quetelet - *Sur l'homme et le développement de ses facultés* (1835)

Une série de mille nombres est illisible : personne ne peut la comprendre en la regardant. Toute l'affaire des statistiques est là — résumer une masse de données par quelques valeurs qui en disent l'essentiel, et par quelques images qui les rendent lisibles. Encore faut-il choisir ces résumés avec soin, car un même ensemble de données peut, selon l'indicateur retenu, raconter deux histoires très différentes.

11.1.1 Compter les hommes

Le mot lui-même trahit son origine : la *statistique* fut d'abord la science de l'**État**, celle qui recensait les habitants, les terres et les richesses afin de lever l'impôt et les armées. Les recensements sont attestés en Égypte, en Chine et à Rome, où le *census* donne son nom à l'opération.

Ces relevés ne servaient qu'à compter. Il faut attendre 1662 pour qu'un marchand londonien, **John Graunt**, ait l'idée d'y chercher des *régularités* : en dépouillant les registres de décès de Londres, il découvre que la mortalité, imprévisible individuellement, obéit à des lois très stables lorsqu'on la considère en masse.

11.1.2 De la description à la science

C'est le belge **Adolphe Quetelet** qui, en 1835, franchit le pas décisif en appliquant à l'homme les méthodes de l'astronomie. Il forge la notion d'*homme moyen*, valeur autour de laquelle se répartissent les mesures d'une population, et fait de la moyenne un véritable objet d'étude.

L'idée sera abondamment critiquée — cet homme moyen n'existe évidemment pas —, mais elle fonde la statistique moderne, aujourd'hui présente partout : sondages, contrôle qualité, épidémiologie, sport, économie. Ce chapitre en présente les outils de base.



11.2 Vocabulaire et représentations

Définition : Une **série statistique** est l'ensemble des données recueillies sur une **population**, dont les éléments sont appelés les **individus**.

La propriété étudiée s'appelle le **caractère**. Il est **quantitatif** lorsqu'il se mesure par un nombre (une taille, un nombre de frères et sœurs), et **qualitatif** sinon (une couleur, une matière préférée).

Définition : Soit une série statistique portant sur une population de N individus, et x_i l'une des valeurs prises par le caractère.

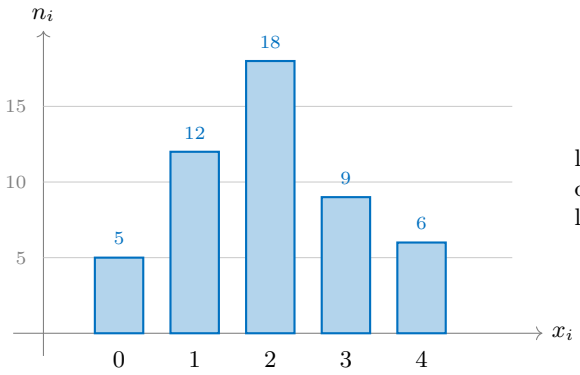
- L'**effectif** n_i de la valeur x_i est le nombre d'individus qui la prennent.
- La **fréquence** de x_i est le quotient $f_i = \dfrac{n_i}{N}$, souvent exprimé en pourcentage.

Remarque : La somme de tous les effectifs vaut N, et celle de toutes les fréquences vaut donc 1, c'est-à-dire 100 %. C'est un contrôle systématique à effectuer sur tout tableau : si la somme des fréquences ne fait pas 1, il y a une erreur.

Exemple : On interroge 50 élèves sur leur nombre de frères et sœurs. Les résultats sont consignés ci-dessous, avec la ligne des fréquences :

Nombre x_i	0	1	2	3	4	**Total**
Effectif n_i	5	12	18	9	6	50
Fréquence f_i	0,10	0,24	0,36	0,18	0,12	1

On lit par exemple que 36 % des élèves interrogés ont exactement deux frères et sœurs.



le **diagramme en bâtons** donne à voir d'un coup la répartition

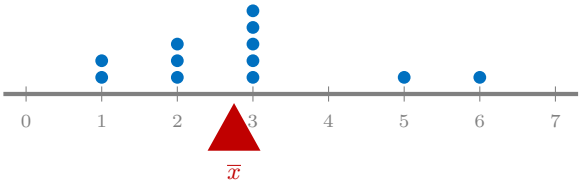
11.3 Indicateurs de position

11.3.1 Moyenne

Définition : Soit une série statistique de valeurs x_1, …, x_p d'effectifs respectifs n_1, …, n_p, et N l'effectif total. La **moyenne** de cette série est le nombre :

$$\overline{x} = \frac{n_1x_1 + n_2x_2 + \cdots + n_px_p}{N}$$

Remarque : La moyenne est le **point d'équilibre** de la série : si l'on place les données sur une règle, en imaginant un poids à chaque valeur, la règle tient en équilibre exactement sur la moyenne. C'est pourquoi une valeur très éloignée des autres, même isolée, la déplace fortement.



la série tient en équilibre
sur sa moyenne

Exemple : Reprenons la série des frères et sœurs. Sa moyenne vaut :

$$\overline{x} = \frac{5\times 0 + 12\times 1 + 18\times 2 + 9\times 3 + 6\times 4}{50} = \frac{99}{50} = 1{,}98$$

Les élèves interrogés ont donc en moyenne un peu moins de deux frères et sœurs — ce qui ne correspond au cas d'aucun élève en particulier.

Définition : Soient $x_1, \ldots, x_p$ des valeurs affectées de **coefficients** respectifs $c_1, \ldots, c_p$. On appelle **moyenne pondérée** de ces valeurs le nombre :

$$\overline{x} = \frac{c_1x_1 + c_2x_2 + \cdots + c_px_p}{c_1 + c_2 + \cdots + c_p}$$

Remarque : C'est exactement la formule précédente, où les effectifs jouent le rôle de coefficients : une valeur qui apparaît 18 fois pèse dans la moyenne comme une valeur de coefficient 18. Les deux situations n'en font donc qu'une, et il suffit de retenir un seul calcul.

Attention en revanche au dénominateur : ce n'est pas le nombre de valeurs, mais la **somme des coefficients**.

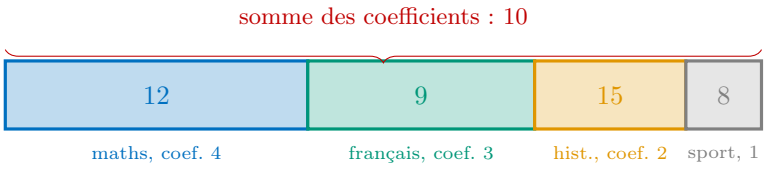
Exemple : Un élève obtient les notes suivantes, affectées des coefficients de chaque matière :

Matière	Maths	Français	Histoire	Sport	**Total**
Note x_i	12	9	15	8	
Coefficient c_i	4	3	2	1	10

Sa moyenne pondérée vaut :

$$\overline{x} = \frac{4\times 12 + 3\times 9 + 2\times 15 + 1\times 8}{4+3+2+1} = \frac{48+27+30+8}{10} = \frac{113}{10} = 11{,}3$$

Sans les coefficients, la moyenne des quatre notes vaudrait $\frac{12+9+15+8}{4} = 11$. La pondération a donc relevé la moyenne, car les meilleures notes portent ici les plus forts coefficients.



la largeur mesure
le *poids* de chaque note

La largeur de chaque bloc représente le coefficient : plus une note est « large », plus elle tire la moyenne vers elle.

Propriété 1 : Si toutes les valeurs d'une série sont multipliées par un même nombre a, puis augmentées d'un même nombre b, alors la moyenne subit exactement la même transformation :

$$\text{moyenne de } (ax+b) = a\overline{x} + b$$

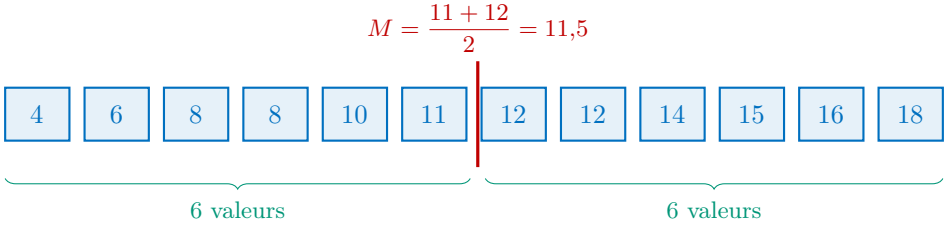
Exemple : Un professeur ajoute deux points à toutes les copies d'un devoir dont la moyenne était 9,5. La nouvelle moyenne vaut $9{,}5 + 2 = 11{,}5$, sans qu'il soit nécessaire de recalculer quoi que ce soit.

11.3.2 Médiane

Définition : La **médiane** d'une série statistique est une valeur M telle qu'au moins la moitié des données lui soient inférieures ou égales, et au moins la moitié supérieures ou égales.

Pour la déterminer, on range les N valeurs dans l'ordre croissant :

- si N est **impair**, la médiane est la valeur du milieu, celle de rang $\frac{N+1}{2}$;
- si N est **pair**, la médiane est la demi-somme des deux valeurs centrales.



Exemple : Considérons les douze notes ci-dessus, déjà rangées dans l'ordre croissant. L'effectif $N = 12$ étant pair, la médiane est la demi-somme des sixième et septième valeurs :

$$M = \frac{11+12}{2} = 11{,}5$$

La moyenne de cette série vaut quant à elle $\frac{134}{12} \approx 11{,}17$: les deux indicateurs sont proches, mais différents.

Remarque : La médiane a un avantage décisif sur la moyenne : elle est **insensible aux valeurs extrêmes**. Si la meilleure note passait de 18 à 100, la moyenne bondirait à plus de 18, alors que la médiane resterait à 11,5.

C'est la raison pour laquelle on préfère la médiane à la moyenne pour décrire les salaires ou les prix des logements : quelques valeurs très élevées suffiraient à rendre la moyenne trompeuse.

11.4 Indicateurs de dispersion

Les indicateurs précédents résument la série par une valeur centrale, mais ne disent rien de son étalement. Deux classes peuvent avoir la même moyenne, l'une avec des notes toutes proches de 10, l'autre avec des 2 et des 18. Il faut donc mesurer aussi la **dispersion**.

Définition : Soit une série statistique rangée dans l'ordre croissant.

- L'**étendue** est la différence entre la plus grande et la plus petite valeur.
- Le **premier quartile** Q_1 est la plus petite valeur de la série telle qu'au moins 25 % des données lui soient inférieures ou égales.
- Le **troisième quartile** Q_3 est la plus petite valeur telle qu'au moins 75 % des données lui soient inférieures ou égales.
- L'**écart interquartile** est la différence $Q_3 - Q_1$.

Méthode : Pour déterminer les quartiles d'une série de N valeurs rangées dans l'ordre croissant :

1. On calcule $\frac{N}{4}$. Si le résultat est entier, Q_1 est la valeur de ce rang ; sinon, on arrondit à l'entier supérieur.

2. On procède de même avec $\frac{3N}{4}$ pour obtenir Q_3.

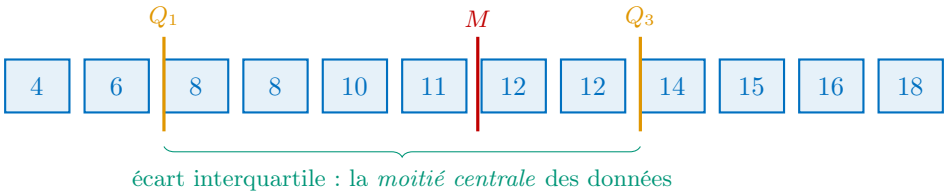
Exemple : Reprenons les douze notes. Ici $N = 12$, donc :

$$\frac{N}{4} = 3 \qquad \text{et} \qquad \frac{3N}{4} = 9$$

Ces deux résultats étant entiers, Q_1 est la troisième valeur et Q_3 la neuvième :

$$Q_1 = 8 \qquad \text{et} \qquad Q_3 = 14$$

L'étendue vaut $18 - 4 = 14$, et l'écart interquartile $14 - 8 = 6$.



Les trois repères Q_1, M et Q_3 découpent la série en quatre parts d'effectifs égaux, d'où leur nom. L'écart interquartile mesure l'étalement de la moitié centrale, en ignorant les valeurs extrêmes.

11.4.1 Lecture sur un tableau d'effectifs cumulés

Ranger les valeurs une à une n'est possible que sur de petites séries. Dès que l'effectif dépasse la vingtaine, on travaille sur un tableau, et l'on y ajoute une ligne qui rend la lecture immédiate.

> **Définition :** L'**effectif cumulé croissant** d'une valeur est le nombre d'individus dont la valeur du caractère lui est *inférieure ou égale*. On l'obtient en additionnant les effectifs de proche en proche, de la gauche vers la droite.

Méthode : Pour déterminer la médiane et les quartiles à partir d'un tableau :

1. On range les valeurs dans l'ordre croissant, et l'on ajoute une ligne d'effectifs cumulés croissants. Le dernier de ces cumuls doit valoir N : c'est le contrôle du calcul.

2. On détermine les rangs cherchés : $\frac{N}{4}$ pour Q_1, $\frac{N}{2}$ pour la médiane et $\frac{3N}{4}$ pour Q_3, en arrondissant à l'entier supérieur lorsque le quotient n'est pas entier.

3. On lit dans la ligne des cumuls la première valeur dont l'effectif cumulé *atteint ou dépasse* ce rang : c'est l'indicateur cherché.

Exemple : Reprenons la série des frères et sœurs des 50 élèves, et complétons le tableau par la ligne des effectifs cumulés croissants :

Valeur x_i	0	1	2	3	4
Effectif n_i	5	12	18	9	6
Effectif cumulé	5	17	35	44	50

La ligne des cumuls se lit ainsi : 17 élèves ont au plus un frère ou une sœur, 35 en ont au plus deux, et le dernier cumul vaut bien 50.

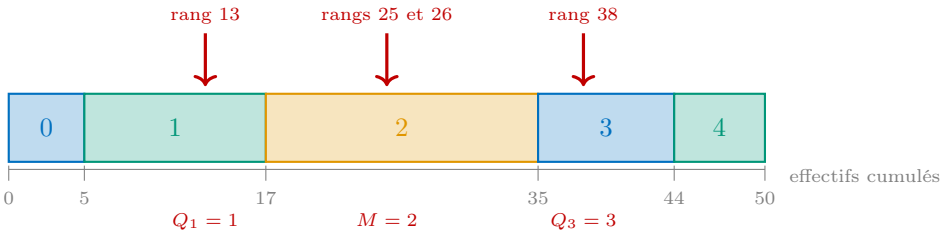
Déterminons maintenant les rangs cherchés :

$$\frac{N}{4} = 12{,}5 \longrightarrow \text{rang } 13 \quad ; \quad \frac{N}{2} = 25 \quad ; \quad \frac{3N}{4} = 37{,}5 \longrightarrow \text{rang } 38$$

Il ne reste qu'à les situer dans la ligne des cumuls :

- le rang 13 est atteint dès le cumul 17, donc $Q_1 = 1$;
- les rangs 25 et 26 sont atteints au cumul 35, donc $M = 2$;
- le rang 38 est atteint au cumul 44, donc $Q_3 = 3$.

L'écart interquartile vaut donc $3 - 1 = 2$.



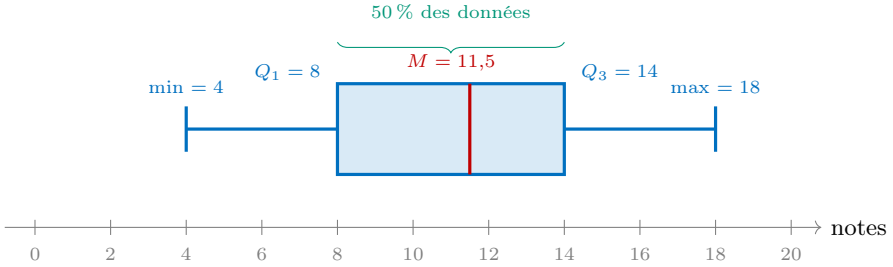
Chaque bloc représente une valeur, et sa largeur son effectif : repérer un rang revient à voir dans quel bloc il tombe.

Remarque : Cette lecture évite de réécrire les 50 données une à une. Elle donne en outre la bonne intuition : les effectifs cumulés indiquent, pour chaque valeur, quelle *part* de la population a déjà été atteinte lorsqu'on parcourt la série dans l'ordre croissant.

11.5 Diagramme en boîte

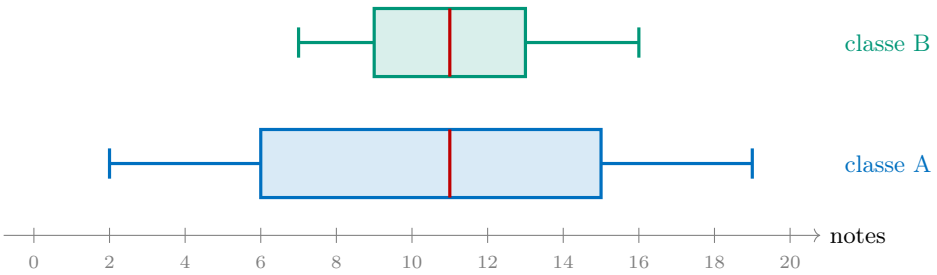
Définition : Le **diagramme en boîte**, aussi appelé **boîte à moustaches**, est une représentation graphique résumant une série par cinq nombres : le **minimum**, Q_1, la **médiane**, Q_3 et le **maximum**.

La boîte s'étend de Q_1 à Q_3 et contient donc la moitié centrale des données ; les moustaches rejoignent les valeurs extrêmes.



Remarque : Ce diagramme n'affiche que cinq nombres, mais il permet de comparer deux séries d'un seul coup d'œil, ce qu'aucun tableau ne permet. On les représente alors l'un au-dessus de l'autre, sur un même axe gradué.

Exemple : Deux classes ont composé le même devoir. Leurs diagrammes sont représentés ci-dessous.



Les deux classes ont la *même médiane*, 11. Pourtant, la boîte de la classe A est bien plus large : ses résultats sont beaucoup plus dispersés, avec des élèves très faibles et d'autres très forts. La classe B est homogène. Deux séries de même position peuvent donc différer radicalement par leur dispersion, et c'est précisément ce que ce diagramme met en évidence.

Remarque : Pour résumer une série, il faut donc toujours donner *au moins deux* indicateurs : un de position et un de dispersion. Annoncer une moyenne sans rien dire de l'étalement, c'est ne livrer que la moitié de l'information — et c'est un procédé courant lorsqu'on veut faire dire à des chiffres ce qu'ils ne disent pas.