Statistiques descriptives

Livre p.269.

Objectifs :

- Connaître le vocabulaire de base des statistiques
- Passer des effectifs aux fréquences
- Calculer et savoir utiliser : la moyenne, la médiane, l'étendue, les 1er et 3ème quartiles, l'écart interquartile, l'écart-type.

Aperçu historique :

*On peut dire que la collecte des données a toujours existé. En effet, les premiers textes écrits retrouvés sont des recensements du bétail. On a des traces de recensements de population en Chine au XXIIIe siècle av. J.-C. ou en Égypte au XVIIIe siècle av. J.-C.. Ce n'est qu'au XVIIIe siècle que l'on voit apparaître le rôle prévisionnel des statistiques avec la construction des premières tables de mortalité, avec en 1746 "L'essai sur les probabilités de la durée de vie humaine" d'**Antoine Deparcieux**. C'est au XIXe siècle que les statistiques prennent leur essor et que l'on édicte des règles précises pour la collecte et l'interprétation des données. Leur première application industrielle a lieu lors du recensement américain de 1890, qui met en œuvre la carte perforée inventée par le statisticien Herman Hollerith. C'est avec l'avènement de l'informatique dans les années 1940 que les statistiques prendront la place prépondérante qu'elles occupent aujourd'hui.*

1. Vocabulaire et premiers outils

A. Vocabulaire

Une série statistique est une liste de nombres, par exemple la liste des notes obtenues par les élèves de 2nde au DS7 de mathématiques. Si deux élèves ont eu la même note, ce nombre apparaît deux fois dans la liste.

Exemple de série statistique : $6;6;6;8;8;10;10;10;10;10;10;10;10;10;10;12;12;12;12;12;12;12;12;14;14;16;16;16;16;16$.
On classe dès que possible ces nombres par ordre croissant, ce sera plus pratique ensuite.

Les différentes notes obtenues s'appellent les **valeurs** de la série statistiques.
S'il y a des élèves qui ont eu $12,5$, on dit que $12,5$ est une valeur de la série statistique.
Le nombre d'élèves qui ont eu $12,5$ s'appelle l'**effectif** de la valeur $12,5$.
Le pourcentage d'élèves qui ont eu $12,5$ s'appelle la **fréquence** de la valeur $12,5$.
Le nombre total d'élèves de 2nde s'appelle l'**effectif total**.
La "fréquence totale" est 100% = 1.

Il est souvent commode de présenter la série dans un tableau :

Notes (valeurs)	6	8	10	12	14	16	Effectif total :
Nombre d'élèves (effectif)	3	2	10	8	2	5	30

Nous travaillerons sur cet exemple tout au long du chapitre.

→Cas continu : Lorsque le caractère que l'on étudie peut prendre toutes sortes de valeurs (taille, superficie, salaire...) on dit que l'on se trouve dans le "cas **continu**". A contrario, lorsqu'on lance un dé à 6 faces, les seules valeurs possibles sont $\{1;2;3;4;5;6\}$, et on dira qu'on est dans le cas "**discret**".
Dans le cas continu, lorsque la série statistique comporte trop de valeurs pour être bien lisible, on peut ranger les données par intervalles appelés **classes**. Nous en verrons un exemple au paragraphe ci-dessous.

B. Effectifs et fréquences

a) Révisions sur les pourcentages

On peut écrire un pourcentage sous forme décimale : $18\% = \frac{18}{100} = 0,18$ (c'est 18 centièmes). C'est un nombre compris entre 0 et 1, on l'utilise pour définir les **fréquences**.

Pour calculer 18% de 368, on effectue : $0,18 \times 368$, où le mot "de" est traduit par l'opération "$\times$".
On peut aussi effectuer $\frac{18}{100} \times 368$, ce qui revient au même.

Pour savoir quel pourcentage de 535 représente 133,75, on effectue la division : $\frac{133,75}{535}$ ("le tout sur la partie").
On obtient 0,25, c'est-à-dire 25%. Ainsi, 133,75 représentent 25% de 535.

→Cas continu : Prenons l'exemple du temps d'attente dans un bureau de poste ; les valeurs sont regroupées par classes, et pour plus de lisibilité on a indiqué les fréquences plutôt que les effectifs :

Durée (en min)	[0;4[	[4;8[	[8;12[	[12;16[	[16;20[
Fréquences	0,22	0,34	0,21	0,16	0,07

b) Des effectifs aux fréquences

L'effectif traduit la présence d'une valeur dans la série statistique en "nombre de fois où elle apparaît", et la fréquence traduit cette présence en pourcentage. Ce pourcentage est en écriture décimale, il s'agit donc d'un nombre compris entre 0 et 1.

Méthode Convertir un effectif en fréquence :

$$\text{fréquence de la valeur} = \frac{\text{effectif de la valeur}}{\text{effectif total}}$$

Exemple : Dans notre série statistique, l'effectif de la valeur 12 est 8.
Calculons la fréquence de cette valeur :
fréquence de la valeur $= \frac{\text{effectif de la valeur}}{\text{effectif total}} = \frac{8}{30} \simeq 0,27$
(on travaille souvent avec des valeurs approchées en statistiques).

c) Des fréquences aux effectifs

La formule ci-dessous utilise une fréquence écrite en écriture décimale (nombre compris entre 0 et 1).

Méthode Convertir une fréquence en effectif :

$$\text{effectif de la valeur} = \text{fréquence de la valeur} \times \text{effectif total}$$

Exemple : D'après ce qui précède, la fréquence de la valeur 12 est 0,27.
Essayons de retrouver l'effectif de cette valeur :
effectif de la valeur = fréquence de la valeur × effectif total $= 0,27 \times 30 = 8,1 \simeq 8$.
Comme il s'agit d'un nombre d'élèves, on obtient nécessairement un nombre entier, d'où l'arrondi à l'entier le plus proche.

2. Paramètres de position

A. Moyenne

Pour calculer la moyenne d'une série statistique, on multiplie chaque valeur par son effectif, et on additionne les nombres obtenus. On divise ensuite ce résultat par l'effectif total.

Définition 8.1 Soient $x_1, x_2, x_3...$ les valeurs d'une série statistique, et $n_1, n_2, n_3...$ les effectifs respectifs de ces valeurs. Soit $N = n_1 + n_2 + n_3...$ l'effectif total de la série. La moyenne de la série est le nombre :

$$\overline{x} = \frac{n_1 x_1 + n_2 x_2 + n_3 x_3...}{N}$$

.

Exemple : Toujours avec notre série de notes :

Notes (valeurs)	$x_1 = 6$	$x_2 = 8$	$x_2 = 10$	$x_3 = 12$	$x_4 = 14$	$x_5 = 16$	Effectif total :
Nombre d'élèves (effectif)	$n_1 = 3$	$n_2 = 2$	$n_3 = 10$	$n_4 = 8$	$n_5 = 2$	$n_6 = 5$	$N = 30$

D'où la moyenne : $\overline{x} = \frac{3\times6+2\times8+10\times10+8\times12+2\times14+5\times16}{30} \simeq 11,27$
Il s'agit d'une "moyenne pondérée" : la pondération signifie qu'il y a des coefficients ; ici, les coefficients sont les effectifs des valeurs. dans un calcul de moyenne scolaire, les notes sont souvent pondérées en fonction du types de devoir (DS, DM, etc...).

La moyenne est une **caractéristique de position** : elle signifie que les valeurs se trouvent "autour" de la valeur 11,27. La moyenne est **très sensible aux valeurs extrêmes** : s'il y a quelques notes extrêmement faibles ou extrêmement élevées, cela change beaucoup la valeur de la moyenne. Dans notre exemple, si l'on ajoute la note 20 à la série statistique (qui comprend alors 31 valeurs), la moyenne devient 11,55 : elle a augmenté de presque 3%.

→Cas continu : Comme on considère que la répartition des valeurs dans les classes est régulière, on prend comme "valeur" de chaque classe le milieu de cette classe, afin de calculer la moyenne. Dans l'exemple des temps d'attente, on obtient : $\overline{x} = \frac{22\times2+34\times6+21\times10+16\times14+7\times18}{100} = 8,08$.

Propriété 8.1 Soient a et b deux nombres réels.
Si une série de valeurs $\{x_i\}_{1\leqslant i\leqslant p}$ a pour moyenne $\overline{x}$, alors la série de valeurs $\{ax_i+b\}_{1\leqslant i\leqslant p}$ a pour moyenne $a\overline{x}+b$.

Exemple Si, dans notre premier exemple, on décide de multiplier par 2 toutes les valeurs, puis de leur ajouter 1 (par exemple pour ramener les notes sur 40 et donner un point de "bonus" à tout le monde), on obtient la série des $\{2x_i+1\}_{1\leqslant i\leqslant p}$, dont la moyenne sera $2\overline{x}+1 = 2\times 11.27+1 = 23.54$ (cette moyenne est pour des notes sur 40).

B. Médiane

Définition 8.2 Après avoir rangé les nombres d'une série statistique dans l'ordre croissant, on appelle médiane de cette série un nombre qui partage la série en deux ensembles de même effectif : les nombres qui précèdent la médiane et les nombres qui suivent la médiane.

Il y a autant de valeurs plus petites que la médiane que de valeurs plus grandes que la médiane.

Exemple : Les joueurs de l'équipe de hand-ball mesurent en mètres : 1,58 - 1,65 - 1,65 - 1,68 - 1,75 - 1,78 - 1,80 – 1,90. Ceci est une série statistique.
On peut prendre comme médiane de cette série n'importe quel nombre compris entre 1,68 et 1,75 (la médiane n'est pas forcément une valeur de la série statistique). Par exemple une médiane de cette série est 1,70 m : il y a autant de joueurs plus grands que 1,70 m que de joueurs plus petits.

→Cas continu : On peut lire la médiane sur le polygone des fréquences cumulées : c'est la valeur qui correspond à la fréquence cumulée "0,5".

La médiane est aussi une **caractéristique de position** : elle signifie que les valeurs se trouvent "autour" de la valeur 1,70. Mais contrairement à la moyenne, la médiane **ne dépend pas des valeurs extrêmes** (la plus grande et la plus petite). Cela peut être intéressant par exemple dans une expérience de physique, où les valeurs extrêmes sont souvent dues à des erreurs de manipulation.

C. Quartiles

Définition 8.3 Soit une série statistique.
On appelle premier quartile la plus petite valeur de la série, notée Q_1, telle qu'au moins 25% des valeurs de la série soient inférieures ou égales à Q_1.
On appelle troisième quartile la plus petite valeur de la série, notée Q_3, telle qu'au moins 75% des valeurs de la série soient inférieures ou égales à Q_3.

Ainsi, Q_1 est "la première valeur qui suit le premier quart", et Q_3 est "la première valeur qui suit le troisième quart". Contrairement à la médiane, Q_1 et Q_3 sont obligatoirement des valeurs de la série statistique.

Premier exemple (Hand-ball) : La série est $1,58-1,65-1,65-1,68-1,75-1,78-1,80–1,90$. Elle comprend 8 valeurs, donc $Q_1 = 1,65$ (3ème valeur), et $Q_3 = 1,80$ (7ème valeur).

Deuxième exemple (notes) : La série est $6;6;6;8;8;10;10;10;10;10;10;10;10;10;10;12;12;12;12;12;12;12;12;14;14;16;16;16;16;16$. Elle comprend 30 valeurs, donc $Q_1 = 10$ (8ème valeur), et $Q_3 = 12$ (23ème valeur).

→Cas continu : De même que pour la médiane, on peut lire les quartiles sur le polygone des fréquences cumulées : elles correspondent respectivement à "0,25" et "0,75".

Remarque : Le second quartile est la médiane : au moins 50% des valeurs de la série statistique sont inférieures ou égales à la médiane.

3. Paramètres de dispersion

A. Étendue

Définition 8.4 Après avoir rangé les nombres d'une série statistique *dans l'ordre croissant*, on calcule l'étendue de cette série en effectuant la soustraction : (plus **grande** valeur) – (plus **petite** valeur).

Premier exemple (Hand-ball) : L'étendue de la série est $E = 1,90 - 1,58 = 0,32$.

Deuxième exemple (notes) : L'étendue de la série est $E = 16 - 6 = 10$.

L'étendue est une **caractéristique de dispersion** : elle montre à quel point les valeurs sont "rapprochées" ou "éloignées" entre elles. L'étendue est **très sensible aux valeurs extrêmes** : c'est avec ces "valeurs extrêmes" qu'on la calcule !!!

B. Écart interquartile

Définition 8.5 La différence $Q_3 - Q_1$ s'appelle écart interquartile.

Premier exemple (Hand-ball) : L'écart interquartile de la série est $Q_3 - Q_1 = 1,80 - 1,65 = 0,15$.

Deuxième exemple (notes) : L'écart interquartile de la série est $Q_3 - Q_1 = 12 - 10 = 2$.

L'écart interquartile est une **caractéristique de dispersion** : elle montre à quel point les valeurs sont "rapprochées" ou "éloignées" entre elles, autour de la médiane. Mais l'écart interquartile **ne dépend pas des valeurs extrêmes** : on n'en a pas tenu compte pour la calculer.
Ainsi, pour la série de notes (deuxième exemple), les valeurs semblent très dispersées si on regarde l'étendue ($E = 10$), mais si on regarde l'écart interquartile, on se rend compte qu'elles ne le sont pas du tout ($Q_3 - Q_1 = 2$) : en fait, presque tous les élèves ont eu 10 ou 12, et la médiane est 11.

C. Écart-type

Définition 8.6 L'écart-type d'une série de valeurs $\{x_i\}_{1\leqslant i\leqslant p}$ est le nombre positif, noté σ, défini par :

$$\sigma = \sqrt{\frac{n_1(x_1-\overline{x})^2 + n_2(x_2-\overline{x})^2 + \ldots + n_p(x_p-\overline{x})^2}{N}}$$

où N désigne l'effectif total

Exemple Avec notre exemple, il vient : $\sigma = \sqrt{\frac{3(6-11.27)^2+2(8-11.27)^2+10(10-11.27)^2+8(12-11.27)^2+2(14-11.27)^2+5(16-11.27)^2}{30}}$
L'écart-type peut être vu comme "la moyenne des écarts à la moyenne".

4. Synthèse sur les statistiques

	Paramètre de position	**Paramètre de dispersion**
Dépend des valeurs extrêmes	**Moyenne** $\overline{x} = \frac{n_1x_1+n_2x_2+n_3x_3\ldots}{N}$	**Étendue** (plus **grande** valeur) – (plus **petite** valeur) **Écart-type** $\sigma = \sqrt{\frac{n_1(x_1-\overline{x})^2+n_2(x_2-\overline{x})^2+\ldots+n_p(x_p-\overline{x})^2}{N}}$
Ne dépend pas des valeurs extrêmes	**Médiane**, Q_1 premier quartile, Q_3 troisième quartile	**Écart inter-quartile** $Q_3 - Q_1$