Chapitre 5

Statistiques

5.1 Graphiques

5.1.1 Vocabulaire

Définition 5.1
– l'*effectif* d'une classe (ou « catégorie ») est le nombre d'éléments de la classe ;
– la *fréquence* d'une classe est le quotient de l'effectif de la classe par l'effectif total :

$$f_i = \text{fréquence de } x_i = \frac{\text{effectif de } x_i}{\text{effectif total}} = \frac{n_i}{N}$$

Exemple 5.1
On donne ci-dessous le tableau récapitulatif des niveaux de pollutions atteints au cours d'une année dans une grande ville. Calculer les fréquences :

Niveau de pollution	0	1	2	3	4
Nombre de jours	5	81	143	100	36
fréquence fréquence en %	0,014 1,4	0,222 22,2	0,392 39,2	0,274 27,4	0,099 9,9

Remarque 5.1
La somme des fréquences vaut 1.

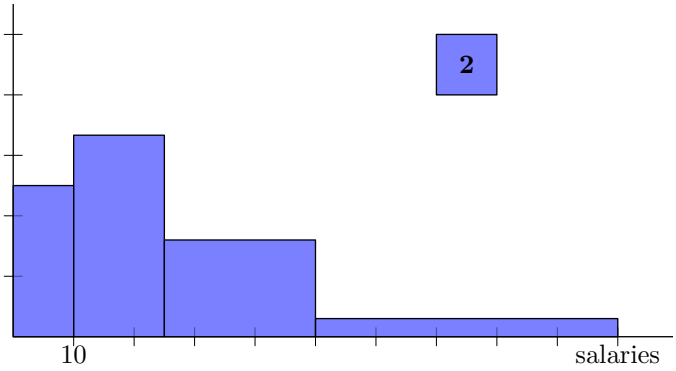
5.1.2 Histogramme

Si on représente une série statistique par un histogramme, chaque classe correspond à un rectangle dont l'*aire* est proportionnelle à l'effectif de la classe, et la largeur est proportionnelle à l'amplitude de la classe. On l'utilise pour représenter une série dont le caractère est quantitatif.

Exemple 5.2
Le tableau suivant donne l'effectif des entreprises d'une zone industrielle suivant le nombre d'employés :

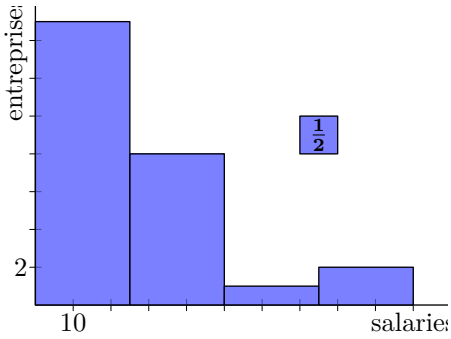
Nombre d'employés N	$N < 10$	$10 \leq N < 25$	$25 \leq N < 50$	$50 \leq N < 100$
Nombre d'entreprises	5	10	8	3

La représentation de ce tableau en histogramme donne :



Remarque 5.2
Si les classes ont toutes la même amplitude, les hauteurs des rectangles sont proportionnelles aux effectifs.

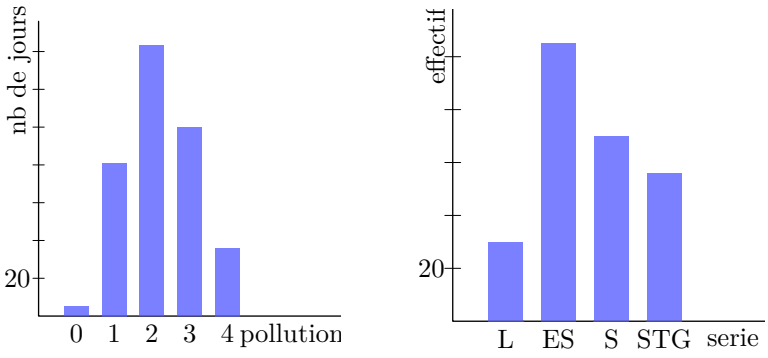
Exemple 5.3
Dans l'exemple 5.2 si on regroupe les deux premières classes et qu'on sait de plus que les entreprises ayant plus de 75 salariés sont au nombre de 2, on obtient :



5.1.3 Diagramme en bâtons

Dans un diagramme en bâtons, la hauteur de chaque bâton est proportionnelle à l'effectif de la classe. On l'utilise pour représenter une série dont le caractère est qualitatif.

Exemple 5.4
Les deux diagrammes suivants sont des exemples de diagrammes en bâtons ; le premier représente les données de l'exemple 5.1, le second représente le nombre d'élèves de première dans chacune des séries d'un lycée :



5.2 Paramètres de position

5.2.1 Le mode

Définition 5.2
Le mode ou valeur modale est la valeur de la variable statistique qui est le plus souvent observée. C'est à dire la valeur du caractère ou la classe qui a le plus grand effectif.

Exemple 5.5
Dans l'exemple 5.1, le mode est le niveau de pollution 2.
Dans l'exemple 5.4 (le deuxième diagramme), le mode est le bac ES.

5.2.2 La médiane

Définition 5.3
La médiane d'une série statistique est la valeur de la variable qui partage la population en deux groupes de même effectif :
- ceux qui ont une valeur du caractère inférieure à la médiane,
- ceux qui ont une valeur du caractère supérieure à la médiane,

Remarque 5.3
Deux cas sont possibles :
- s'il y a un nombre impair d'observations : $N = 2k + 1$, où $k \in \mathbf{N}$, alors la médiane est la $k + 1^{\text{e}}$ valeur du caractère (les valeurs étant rangées par ordre croissant).
- s'il y a un nombre pair d'observations : $N = 2k$, où $k \in \mathbf{N}$, alors on convient de prendre comme médiane la moyenne des k^{e} et $k + 1^{\text{e}}$ valeurs du caractère (les valeurs étant rangées par ordre croissant).

Exemple 5.6 (nombre impair d'observations)
On donne la série statistique suivante :

valeur	3	4	6	7
effectif	1	3	2	1

On a ici un effectif total de 7. La médiane est donc la 4^{e} valeur lorsqu'elles sont rangées par ordre croissant :
3 ; 4 ; 4 ; 4 ; 6 ; 6 ; 7. La médiane vaut 4.

Exemple 5.7 (nombre pair d'observations)
On donne la série statistique suivante :

valeur	3	4	6	7
effectif	2	3	1	4

On a ici un effectif total de 10. La médiane est donc la moyenne de la 5^{e} et de la 6^{e} valeurs lorsqu'elles sont rangées par ordre croissant :
3 ; 3 ; 4 ; 4 ; 4 ; 6 ; 7 ; 7 ; 7 ; 7. La médiane vaut $\frac{4+6}{2} = 5$.

5.2.3 La moyenne

Définition 5.4
La moyenne d'une série statistique est le quotient de la somme de toutes les valeurs de la série (comptées autant de fois que leur effectif) par l'effectif total. En considérant une série statistique de N observations où la variable x prend p valeurs notées x_1, x_2, ..., x_p, chacune ayant un effectif noté n_i, on a :

$$\overline{x} = \frac{\sum_{i=1}^{p} n_i x_i}{N}, \text{ où } \sum_{i=1}^{p} n_i x_i = n_1 x_1 + n_2 x_2 + \cdots + n_p x_p$$

Propriété 5.1
x est une série statistique prenant p valeurs x_i ($1 \leq i \leq p$), chacune ayant une fréquence f_i.

$$\overline{x} = \sum_{i=1}^{p} f_i x_i$$

Exemple 5.8
En reprenant l'exemple 5.1, on peut calculer le niveau de pollution moyen de la ville étudiée de deux façons différentes :

$$\overline{x} = \frac{5 \times 0 + 81 \times 1 + 143 \times 2 + 100 \times 3 + 4 \times 36}{365} \approx 2{,}2$$

$$\overline{x} = 0{,}014 \times 0 + 0{,}222 \times 1 + 0{,}392 \times 2 + 0{,}274 \times 3 + 0{,}099 \times 4 \approx 2{,}2$$

Propriété 5.2 (linéarité de la moyenne)
- si on multiplie chaque valeur d'une série statistique par un réel a, alors la moyenne est multipliée par a.
- si on ajoute à chaque valeur d'une série statistique un nombre réel b, alors la moyenne est augmentée de b (si $b < 0$, il s'agira d'une « augmentation négative »).

Propriété 5.3 (conséquence)
Soit x une série statistique de moyenne $\overline{x}$. a et b sont deux réels. On considère y la série statistique définie par $y_i = ax_i + b$ pour tout i. Alors la moyenne de la série y est :

$$\overline{y} = a\overline{x} + b$$

Propriété 5.4 (écarts à la moyenne)
La moyenne des écarts à la moyenne est nulle : Soit x une série statistique de moyenne $\overline{x}$. La série statistique y définie par $y_i = x_i - \overline{x}$ a une moyenne nulle.

Propriété 5.5 (moyennes partielles)
Soit x et y deux séries statistiques d'effectifs respectifs N et P. On considère la série statistique z constituée du regroupement des séries x et y. On a :

$$\overline{z} = \frac{N\overline{x} + P\overline{y}}{N + P}$$

Exemple 5.9
Lors d'un contrôle, la moyenne des 15 filles d'une classe était de 12/20. La moyenne des 10 garçons de 11/20. La moyenne de la classe est :

$$M = \frac{15 \times 12 + 10 \times 11}{15 + 10} = 11{,}6$$

5.3 Paramètres de dispersion

Les paramètres de positions sont insuffisants pour étudier correctement une série statistique : deux séries ayant les mêmes paramètres peuvent être très différentes.

Exemple 5.10
On donne les résultats de deux groupes d'élèves à un même contrôle :

Groupe 1 :

note x	3	5	6	7	8	9	10	13	14	18	20
effectif	1	1	2	2	4	2	1	2	3	1	1

Groupe 2 :

note y	1	2	3	4	13	14	18	19	20
effectif	3	2	2	4	1	2	4	2	2

Ces deux séries ont pour moyenne $\overline{x} = \overline{y} = 10$ et pour médiane $\text{Med}_x = \text{Med}_y = 8{,}5$.

5.3.1 L'étendue

Définition 5.5
L'*étendue* d'une série statistique est la différence entre les deux valeurs extrêmes observées.

Exemple 5.11
Dans l'exemple 5.10, l'étendue du groupe 1 vaut $20 - 3 = 17$, et l'étendue du groupe 2 vaut $20 - 1 = 19$.

5.3.2 Les quartiles

Définition 5.6
Soit x une série statistique avec $x_1 \leq x_2 \leq \cdots \leq x_n$.
Les *quartiles* (au nombre de 3 : Q_1, Q_2 et Q_3) partagent la population classée par ordre croissant de valeur du caractère en quatre sous ensembles, en respectant les règles ci-dessous :
- Q_1 est la valeur x_i dont l'indice i est le plus petit entier supérieur ou égal à $n/4$.
- Q_2 est la valeur x_i dont l'indice i est le plus petit entier supérieur ou égal à $n/2$.
- Q_3 est la valeur x_i dont l'indice i est le plus petit entier supérieur ou égal à $3n/4$.

Définition 5.7
L'*écart interquartile* est la différence entre le 3$^{\text{e}}$ et le 1$^{\text{er}}$ quartile. Au moins 50% des observations ont une valeur du caractère comprise entre Q_1 et Q_3.

Exemple 5.12
Calculer les trois quartiles de chacune des séries de l'exemple 5.10 :

série x : $n = 20$ donc $n/4 = 5$, $n/2 = 10$, $3n/4 = 15$,
on a donc : $Q_1 = x_5 = 7$, $Q_2 = x_{10} = 8$, $Q_3 = x_{15} = 13$;

série y : $n = 22$ donc $n/4 = 5{,}5$, $n/2 = 11$, $3n/4 = 16{,}5$,
on a donc $Q_1 = y_6 = 3$, $Q_2 = y_{11} = 14$, et $Q_3 = y_{17} = 18$.

Remarque 5.4
De la même façon que les quartiles partagent la population en quatre groupes d'effectifs « proches », on peut aussi définir les *déciles* qui partagent la population en dix groupes d'effectifs comparables. En fait, on utilise surtout les premier et neuvième déciles qui sont définis comme ceci :
- le premier décile D_1 est la valeur x_i de la série dont l'indice i est le plus petit entier supérieur ou égal à $\frac{n}{10}$ lorsqu'elles sont rangées par ordre croissant ;
- le neuvième décile D_9 est la valeur x_i de la série dont l'indice i est le plus petit entier supérieur ou égal à $\frac{9n}{10}$ lorsqu'elles sont rangées par ordre croissant.

Exemple 5.13
En reprenant les données de l'exemple 5.10, on a :

série x : $n = 20$ donc $n/10 = 2$; $9n/10 = 18$ donc $D_1 = x_2 = 5$ et $D_9 = x_{18} = 14$;

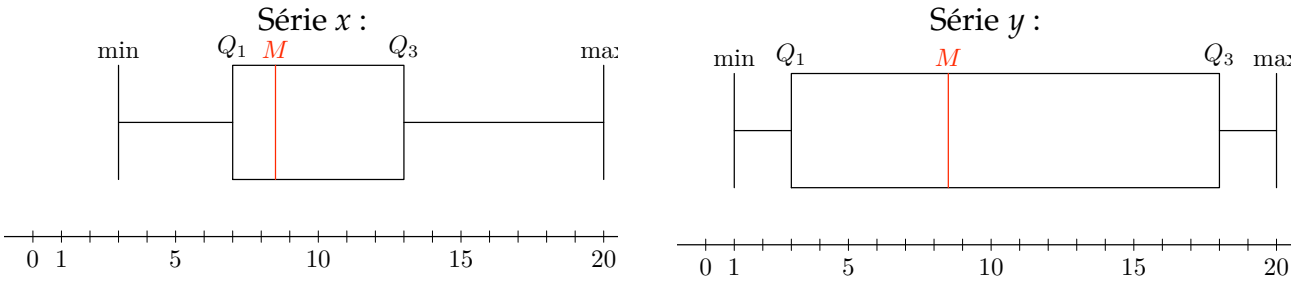
série y $n = 22$ donc $n/10 = 2{,}2$; $9n/10 = 19{,}8$ donc $D_1 = y_3 = 1$ et $D_9 = y_{20} = 19$.

On pourra se référer à l'annexe A de la page 59 pour l'obtention des paramètres statistiques à l'aide de la calculatrice.

5.3.3 Application : les diagrammes en boîtes

La représentation graphique de la dispersion d'une série statistique se fait à l'aide de graphiques appelés *diagrammes en boites*, *boites à moustaches*, ou *box plot*, voire *diagramme de Tuckey*. On les trace comme ceci :
- on construit en face d'un axe gradué, permettant de repérer les valeurs extrêmes de la série étudiée, un rectangle dont la longueur est égale à l'écart interquartile et dans lequel on représente la médiane par un trait ;
- deux traits repèrent les valeurs extrêmes de la série.

Exemple 5.14
On reprend les deux séries de l'exemple 5.10 :

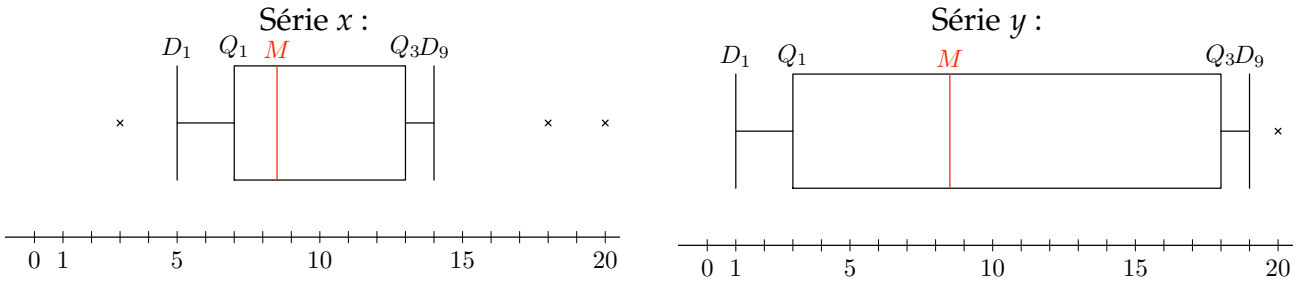


On pourra se référer à l'annexe A de la page 59 pour l'obtention des boites à moustaches à l'aide de la calculatrice.

Remarque 5.5
Parfois, les « moustaches » sont placées aux premier et neuvième déciles. Les valeurs extrêmes

ou les valeurs inférieures à D_1 et supérieures à D_9 sont alors indiquées par une croix.

Exemple 5.15
En reprenant les séries de l'exemple 5.10, on obtient les boîtes à moustaches (avec déciles) suivantes :



5.3.4 Variance et écart type

Définition 5.8
La *variance* est la moyenne des carrés des écarts à la moyenne. C'est un nombre positif.

$$V = \frac{\sum_{i=1}^{p} n_i (x_i - \overline{x})^2}{N} = \frac{n_1 (x_1 - \overline{x})^2 + \cdots + n_p \left(x_p - \overline{x}\right)^2}{N}$$

Remarque 5.6
On a aussi (voir la prop 5.1) :

$$V = \sum_{i=1}^{p} f_i (x_i - \overline{x})^2$$

Propriété 5.6
Autre méthode de calcul de la variance :

$$V(x) = \frac{\sum_{i=1}^{n} n_i x_i^2}{N} - \overline{x}^2 = \frac{n_1 x_1^2 + \cdots + n_p x_p^2}{N} - \overline{x}^2$$

Exemple 5.16
En reprenant le groupe 1 de l'exemple 5.10, on peut calculer la variance de deux façons différentes :

avec la définition 5.8 :

$$V = \frac{1 \times (3-10)^2 + \cdots + 1 \times (20-10)^2}{20} = \frac{372}{20} = 18{,}6$$

avec la propriété 5.6 :

$$V = \frac{1 \times 3^2 + \cdots + 1 \times 20^2}{20} - 10^2 = \frac{2372}{20} - 10^2 = 118{,}6 - 100 = 18{,}6$$

Définition 5.9
L'*écart type* d'une série statistique est égal à la racine carrée de la variance.

Remarque 5.7
L'écart type permet de mesurer la dispersion d'une série statistique ; à moyenne égale, plus il est important, plus les valeurs observées sont dispersées. Son avantage par rapport à la variance est qu'il est exprimé dans la même unité que les valeurs de la série.