Statistique descriptive

1 Vocabulaire des statistiques

Une étude statistique se fait sur **une population** (animaux, végétaux, objets, *etc.*), composés d'**individus**, sur laquelle on étudie de manière systématique un même **caractère** (masse, taille, couleur, *etc.*). L'ensemble des nombres obtenus est appelé **la série statistique**. Un caractère peut être **quantitatif** quand on lui associe un nombre (masse, vitesse, *etc.*) ; sinon on dit qu'il est **qualitatif** (couleur des yeux, *etc.*).
Un caractère est partagé en **modalités** ou en **classes**. Pour chaque modalité ou classe, **l'effectif** de cette modalité/classe est le nombre d'individus présentant cette modalité ou appartenant à cette classe.
L'effectif total est l'effectif de la population.
Pour chaque modalité ou classe, **la fréquence** est le quotient de son effectif par l'effectif total. La fréquence est un nombre compris entre 0 et 1 et est souvent exprimée en pourcentage (%).
La fréquence cumulée (croissante) de cette modalité/classe est la fréquence d'individus ayant une valeur dans la série inférieure ou égale à cette valeur. On peut représenter une série statistique par un graphique : nuage de points, diagramme en bâtons, diagramme circulaire (« camembert »), histogramme, *etc.* .

2 E.C.C. & F.C.C.

Définition 1

L'effectif cumulé (croissant) (E.C.C.) d'une modalité ou classe est le nombre d'individus ayant une valeur dans la série inférieure ou égale à cette valeur.

Exemple 1

La série statistique suivante indique le nombre de cadeaux reçus par enfant pour Noël dans un quartier.

Nombre de cadeaux	0	1	2	3	4
Effectif	14	16	30	12	2
E.C.C.	14	30	60	72	74

Quand la série est rangée par ordre croissant, le premier E.C.C. est égal au premier effectif ; les autres E.C.C. s'obtiennent en sommant l'E.C.C. précédent avec l'effectif de la modalité.

On remarque que le dernier E.C.C. est égal à l'effectif total.

Définition 2

La fréquence cumulée (croissante) (F.C.C.) d'une modalité ou classe est la fréquence d'individus ayant une valeur dans la série inférieure ou égale à cette valeur.

Exemple 2

Nombre de cadeaux	0	1	2	3	4
Effectif	14	16	30	12	2
E.C.C.	14	30	60	72	74
E.C.C.	19 %	41 %	81 %	97 %	100 %

On remarque que les F.C.C. s'obtiennent en faisant le quotient de l'E.C.C. de la modalité par l'effectif total ; la dernière F.C.C. vaut 100 %.

3 Indicateurs de centre

3.1 Le mode

Définition 3

Le **mode** est la modalité dont l'effectif est le plus élevé d'une série statistique.

3.2 La moyenne

Définition 4

On considère une série statistique d'effectif total N :

Caractère	x_1	x_2	...	x_p
Effectif	n_1	n_2	...	n_p

La moyenne (arithmétique) de la série est le nombre

$$\overline{x} = \frac{x_1 \times n_1 + x_2 \times n_2 + \cdots + x_p \times n_p}{n_1 + n_2 + \cdots + n_p} = \frac{x_1 \times n_1 + x_2 \times n_2 + \cdots + x_p \times n_p}{\mathrm{N}}$$

Remarque :
Par abus de langage, on omet souvent le mot « arithmétique » mais il existe d'autres moyennes : géométrique, quadratique, harmonique, énergétique, *etc.* .

Exemple 3

En reprenant la série

Nombre de cadeaux	0	1	2	3	4
Effectif	14	16	30	12	2

L'effectif total vaut $14+16+30+12+2=74$ et la moyenne vaut
$\overline{x} = \dfrac{0\times 14 + 1 \times 16 + 2\times 30 + 3 \times 12 + 4 \times 8}{74} = \dfrac{120}{74} = \dfrac{60}{37}$.
La moyenne vaut $\frac{60}{37}$; une valeur approchée est 1,6.

3.3 La médiane

Définition 5

On considère une série statistique d'effectif total N et rangée dans l'ordre croissant : $x_1 ; x_2 ; \ldots ; x_{\mathrm{N}}$.

- Si N est impair, **la médiane** de la série est « la valeur du milieu », soit $x_{\frac{\mathrm{N}+1}{2}}$.
- Si N est pair, **la médiane** de la série est la moyenne des deux termes centraux : $\dfrac{x_{\frac{\mathrm{N}}{2}} + x_{\frac{\mathrm{N}}{2}+1}}{2}$.

Exemple 4

En reprenant la série

Nombre de cadeaux	0	1	2	3	4
Effectif	14	16	30	12	2
E.C.C.	14	30	60	72	74

l'effectif total valant 74, la médiane est la moyenne des 37^e et 38^e valeurs.
D'après les E.C.C., les données de rang 30 à 59 valent 2 : la médiane vaut donc 2.

Propriété 1

- Au moins 50 % des valeurs de la série sont inférieures à la médiane.
- Au moins 50 % des valeurs de la série sont supérieures à la médiane.

4 Indicateurs de dispersion

4.1 L'étendue

Définition 6

On considère une série statistique d'effectif total N et rangés dans l'ordre croissant : $x_1\,;x_2\,;\ldots\,;x_N$.
L'étendue est la différence entre la plus grande valeur et la plus petite valeur : $x_N - x_1$.

Exemple 5

On considère la série statistique : 22 ; 23 ; 23 ; 25 ; 25 ; 25 ; 25 ; 28 ; 30 ; 32.
Son étendue vaut $32 - 22 = 10$.

4.2 L'écart type

Définition 7

Une série statistique d'effectif total N, de moyenne $\overline{x}$ est notée

Caractère	x_1	x_2	...	x_p
Effectif	n_1	n_2	...	n_p

.

L'écart type de la série est le nombre

$$\sigma = \sqrt{\frac{1}{N}\left[n_1\left(x_1-\overline{x}\right)^2 + \cdots + n_p\left(x_p-\overline{x}\right)^2\right]}$$

Propriété 2

L'écart type permet de mesurer l'écart moyen à la moyenne.

Exemple 6

On considère la série statistique : 22 ; 23 ; 23 ; 25 ; 25 ; 25 ; 25 ; 28 ; 30 ; 32.
Son écart-type vaut $\sigma = \sqrt{10,4} \approx 3,22$.

4.3 Quartiles & écart interquartile

Définition 8

On considère une série statistique d'effectif total N et rangée dans l'ordre croissant : $x_1\,;x_2\,;\ldots\,;x_N$.

- **Le premier quartile** noté Q_1 est la plus petite valeur de la série telle qu'au moins 25 % des valeurs de la série lui soient inférieures ou égales.
- **Le troisième quartile** noté Q_3 est la plus petite valeur de la série telle qu'au moins 75 % des valeurs de la série lui soient inférieures ou égales.

Définition 9

On considère une série statistique de premier quartile Q_1 et de troisième quartile Q_3.
L'écart inter-quartile est la différence entre le troisième et le premier quartile : $Q_3 - Q_1$.

Exemple 7

On considère la série statistique : 22 ; 23 ; 23 ; 25 ; 25 ; 25 ; 25 ; 28 ; 30 ; 32.
Pour le premier quartile, son rang vaut $10 \times \frac{1}{4} = 2,5 \approx 3$ par excès.
Q_1 est la $3^{\text{ème}}$ valeur de la série, soit $Q_1 = 23$.
Pour le troisième quartile, son rang vaut $10 \times \frac{3}{4} = 7,5 \approx 8$ par excès.
Q_3 est la $8^{\text{ème}}$ valeur de la série, soit $Q_3 = 28$.
L'écart inter-quartile vaut $Q_3 - Q_1 = 28 - 23 = 5$.

5 Représentation de données statistiques

5.1 Nuage de points

Définition 10

On considère une série statistique d'effectif total N :

Caractère	x_1	x_2	...	x_p
Effectif	n_1	n_2	...	n_p

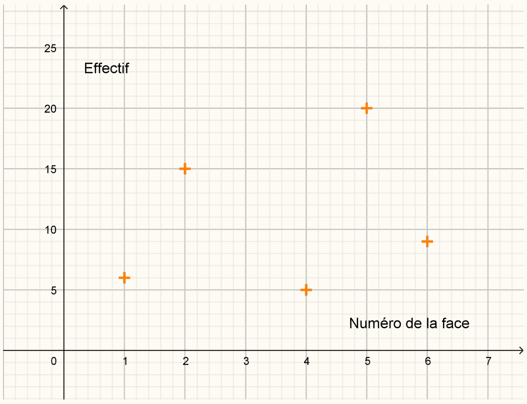
On peut représenter la série par un **nuage de points** constitués des points $(x_1\,;n_1)\,;\cdots;\,(x_p\,;n_p)$.

Exemple 8

On lance un dé à 6 faces et on relève le numéro de la face supérieure.
On obtient les résultats suivants.

Numéro de face	1	2	3	4	5	6
Effectif	6	15	32	5	20	9

On obtient le nuage de points ci-contre.



5.2 Courbe de fréquences cumulées croissantes

Définition 11

On peut représenter une série statistique au caractère quantitatif continu par **la courbe des fréquences cumulées croissantes**.
Les classes sont représentés en abscisse et la fréquence cumulée en ordonnée.

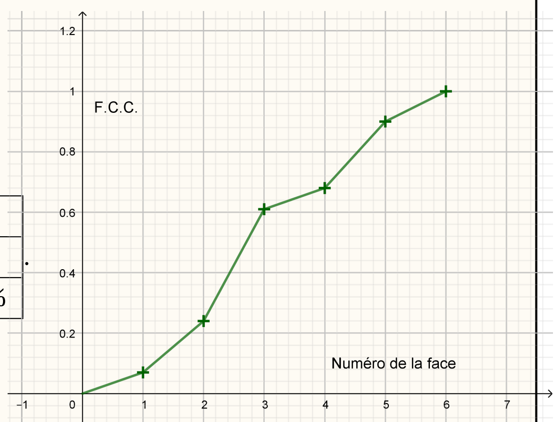
Exemple 9

On mesure la taille de 50 personnes et on regroupe les résultats dans le tableau suivant :

On lance un dé à 6 faces et on relève le numéro de la face supérieure.
On obtient les résultats suivants.

Numéro de face	1	2	3	4	5	6
Effectif	6	15	32	5	20	9
F.C.C.	7 %	24 %	61 %	68 %	90 %	100 %

On réalise la courbe des F.C.C. (fréquences cumulées croissantes) ci-contre.
Estimer la médiane et les quartiles à partir de cette courbe.



6 Choix de résumé statistique

♯ Propriété 3

La moyenne et la médiane permettent de visualiser le « centre » de la série.

↪ Exemple 10

À une table, 49 personnes ont 1 000 € et 1 a 1 000 000 €. Déterminer la moyenne et la médiane.
La médiane vaut 1 000 tandis que la moyenne vaut $\frac{1\ 000 \times 49 + 1\ 000\ 000 \times 1}{50} = 20\ 980$€.
Ici la moyenne n'est pas vraiment significative.

♯ Propriété 4

La moyenne est « sensible aux valeurs extrêmes », contrairement à la médiane.
La moyenne a des propriétés de calcul et théoriques que n'a pas la médiane.

♯ Propriété 5

L'écart-type et l'écart inter-quartile sont des indicateurs de dispersion.

- L'écart-type indique si les valeurs de la série sont « rassemblées » ou « dispersées » autour de la moyenne.
- L'écart inter-quartile indique si les valeurs de la série sont « rassemblées » ou « dispersées » autour de la médiane.

♯ Propriété 6

L'écart-type, comme la moyenne, est « sensible aux valeurs extrêmes ».
L'écart-type a des propriétés de calcul et théoriques que n'a pas l'écart inter-quartile.

Pour résumer une série statistique, on présente un indicateur de centre et un indicateur de dispersion, on a donc deux couples possibles :

- le couple (moyenne ; écart-type) qui est un puissant outil théorique, mais qui est sensible aux valeurs extrêmes ; il est cependant bien adapté aux séries dont l'effectif total est « grand » et dont les valeurs sont réparties sous forme de « cloche » ;
- le couple (médiane ; écart inter-quartile) a moins de propriétés théoriques, mais permet de situer les répartitions des valeurs de la série et, en outre, il n'est pas sensible aux valeurs extrêmes.