Chapitre 8 : Statistiques

I Vocabulaire

1) Généralités

L'ensemble sur lequel porte l'étude statistique s'appelle *la population.*

Un élément de cet ensemble est un *individu*.

L'aspect étudié s'appelle le *caractère*.

Un *échantillon* est une partie de la population.

Exemple: Au cours d'une enquête portant sur les cahiers vendus dans une papeterie en 2010, on s'intéresse à leurs couleurs, leurs mois de vente, leurs poids ainsi qu'au nombre de pages. Voici les résultats obtenus :

couleur	bleu	rouge	jaune	vert
nombre de cahiers vendus	50	25	75	50

mois	janv.	fév.	mars	avril	mai	juin	juillet	aout	sept.	oct.	nov.	déc.
nb cahiers vendus	15	20	15	5	5	15	15	40	20	20	20	10

poids en grammes	[75 ; 100 [	[100 ; 125 [	[125 ; 150 [	[150 ; 175 [	[175 ; 200 [
nb cahiers vendus	20	40	80	30	30

nombre de pages	80	100	120	180	200
nb cahiers vendus	20	100	20	20	40

• La population étudiée est : ………………………………

• Un individu est ……………………………………………

• Un des caractères étudié est : …………………………………

2) Qualitatif ou quantitatif

Si le caractère prend des valeurs numériques, on dit qu'il est quantitatif (on parle de variable).
Sinon, on dit qu'il est qualitatif (on parle de modalité).

Exemple : Dans l'étude statistique précédente,
• les caractères quantitatifs sont : ……………………………………………………………
• les caractères qualitatifs sont : ………………………………………………………………

3) Discret ou continu

Un caractère quantitatif peut être discret ou continu.
On dit qu'il est discret s'il prend des valeurs isolées (ex: 0 ; 1 ; 2 ; …).
On dit qu'il est continu s'il peut prendre toutes les valeurs d'un intervalle.

Exemple : Dans l'étude statistique du 1),
• le caractère quantitatif discret est : ……………………………………………………
• le caractère quantitatif continu est : ……………………………………………………

4) Effectif et fréquence

Pour une valeur (ou une modalité), l'effectif est le nombre d'individus de la population ayant cette valeur (ou cette modalité).

Pour une valeur la fréquence est le quotient de l'effectif de la valeur par l'effectif total (une fréquence est un nombre toujours compris entre 0 et 1).

Activité 1 :

1) Voici une étude statistique portant sur la taille des élèves d'une classe. Compléter le tableau suivant.

taille en centimètres	[140 ; 150 [	[150 ; 160 [	[160 ; 170 [	[170 ; 180 [	[180 ; 190 [
effectif	2	6	12	4	1
fréquence					
effectifs cumulés croissants					
fréquences cumulées croissantes					

2) Construire le polygone des fréquences cumulées croissantes.

II Représentations graphiques

1) Diagramme en bâtons

Il est utilisé pour représenter graphiquement une série statistique dont le caractère est discret.

On représente sur l'axe des abscisses les différentes valeurs du caractère et, sur l'axe des ordonnées, les effectifs.

Activité 2 : Dans le tableau ci-dessous donne les salaires dans une entreprise.

Salaire (euros)	4 000	2 500	2 000	1 750	1 500	1 250	1 000
Effectif	2	4	7	5	10	12	4

Effectuer une représentation de cette série à l'aide d'un diagramme en bâtons.

2) Diagramme circulaire

Activité 3 : On connaît le mode de vie des personnes de 25 à 29 ans.

Représenter ces deux séries par un diagramme circulaire.

	Hommes	Femmes
chez les parents	29,1 %	15 %
Seuls	18 %	15,4 %
En couple sans enfants	25,6 %	27,3 %
En couple avec enfants	19,6 %	32,7 %
autres	7,7 %	9,6 %

3) Diagramme en barres (histogramme)

Ils sont utilisés pour représenter des séries continues où les données ont été réparties en classes.

Activité 4 :

Représenter à l'aide de diagrammes en barres les deux séries suivantes

Masse (en g)	[98 ; 98,5 [	[98,5 ; 99 [	[99 ; 99,5 [	[99,5 ; 100 [	[100 ; 100,5 [	[100,5 ; 101 [
Effectif	3	5	9	8	7	2

Distance (km)	[100 ; 120[	[120 ; 160[	[160 ; 180[	[180 ; 220[
Effectif	1000	1500	500	500

III Moyenne d'une série statistique

Activité 5 : Déterminer la moyenne des deux séries suivantes

1)

Valeurs (x_i)	3	6	8	9
Effectif (n_i)	2	1	3	4

2)

Classe	[380 ; 460[	[460 ; 480 [	[480 ; 500 [	[500 ; 520[
Effectif	6	9	7	4

Ce que je retiens :
On appelle moyenne d'une série statistique (notée souvent $\bar{x}$) le quotient de la somme des valeurs du caractère par l'effectif total.
Pour la série statistique :

Valeur	x_1	x_2	...	x_k
Effectif	n_1	n_2	...	n_k

$$\bar{x} = \frac{x_1n_1 + x_2n_2 + \dots + x_kn_k}{n_1 + n_2 + \dots + n_k} = \frac{\sum_{i=1}^{k} x_in_i}{\sum_{i=1}^{k} n_i}$$

IV Mesures d'une série quantitative

Activité 6 :
Treize étudiants en mathématiques jouent au bowling.
- Voici la série des scores obtenus par les sept joueurs de l'équipe des Verts : 105 – 120 – 104 – 121 – 99 – 127 – 108
- Voici la série des scores obtenus par les six joueurs de l'équipe des Bleus : 93 – 181 – 89 – 98 – 117 – 94

Proposer une méthode permettant de désigner l'équipe qui a réalisé la meilleure performance.

Définition : La différence des valeurs extrêmes du caractère s'appelle l'étendue.

Remarque : L'étendue d'une série permet de caractériser la dispersion des valeurs.
Plus l'étendue est grande, plus les valeurs sont dispersées.

Définition : Le mode pour un caractère discret est la valeur du caractère qui a le plus grand effectif.
Dans le cas d'un caractère continu, dont les valeurs sont regroupées en classes de même amplitude, on parle de classe modale.

Exemple : Voici les notes obtenues en une année en français par Thomas.

13 – 5 – 4 – 12 – 17 – 18 – 5 - 14 – 16 - 3 – 15 – 7 – 5 – 16 – 15 – 14 – 13

Déterminer le mode et l'étendue de cette série.

Définition : La médiane d'une série statistique partage cette série en deux parties de telle sorte que :
- au moins la moitié des données sont inférieures ou égales à la médiane ;
- au moins la moitié des données sont supérieures ou égales à la médiane.

Méthode : Pour déterminer la médiane d'une série statistique à caractère quantitatif discret, on classe les données dans l'ordre croissant. Si la série contient n données rangées dans l'ordre croissant :

• si n est impair, on prend la $\frac{n+1}{2}$ ième donnée pour médiane.

• si n est pair, on prend pour médiane la demi-somme entre la $\frac{n}{2}$ ième donnée et la $\frac{n}{2}+1$ ième donnée.

Activité 7 : Recherche de la médiane dans le cas d'une série à caractère continu
Utiliser le polygone des fréquences cumulées obtenu à l'activité 1 pour déterminer la taille médiane des élèves de la classe. A quelle classe appartient cette taille médiane ?

Activité 8 : Voici, en années, les âges de quinze personnes d'un club de théâtre :

12 – 21 – 27 – 11 – 13 – 12 – 13 – 20 – 22 – 22 – 12 – 25 – 18 – 19 – 21

1) Déterminer la médiane de cette série.
2) Montrer qu'au moins 25 % des âges sont inférieurs ou égaux à 12.
Montrer qu'au moins 75 % des âges sont inférieurs ou égaux à 22.
3) Onze personnes supplémentaires s'inscrivent au club théâtre.
Voici leurs âges : 11 – 13 – 18 – 18 – 19 – 11 – 13 – 21 – 25 – 21 – 18

a) Compléter le tableau ci-dessous qui indique la répartition des membres selon leur âge.

Age	11	12	13	18	19	20	21	22	25	27
Effectif										
Effectif cumulé croissant										
Fréquences cumulées croissantes (en %)										

b) Déterminer la médiane de cette série.
c) Vérifier qu'au moins 25 % des âges sont inférieurs ou égaux à 13.
13 est la plus petite valeur de la série qui satisfasse à cette propriété.
On exprime ceci en disant que 13 est le premier quartile de la série, noté Q_1.
d) Déterminer le troisième quartile de cette série.
e) Vérifier qu'au moins 10 % des âges sont inférieurs ou égaux à 11.
11 est le plus petit nombre possédant cette propriété. On dit que 11 est le premier décile, noté D_1 .
f) Déterminer le neuvième décile de cette série.

Ce que je retiens :
On considère une série de n valeurs rangées dans l'ordre croissant.
• Le premier quartile Q_1 est la plus petite valeur de la série telle qu'au moins 25% des données soient inférieures ou égales à ce nombre Q_1.
• Le troisième quartile Q_3 est la plus petite valeur de la série telle qu'au moins 75% des données soient inférieures ou égales à ce nombre Q_3.
• Le premier décile D_1 est la plus petite valeur de la série telle qu'au moins 10% des données soient inférieures ou égales à ce nombre D_1.
• Le neuvième décile D_9 est la plus petite valeur de la série telle qu'au moins 90% des données soient inférieures ou égales à ce nombre D_9.

Méthode : Si la série contient n données rangées dans l'ordre croissant :

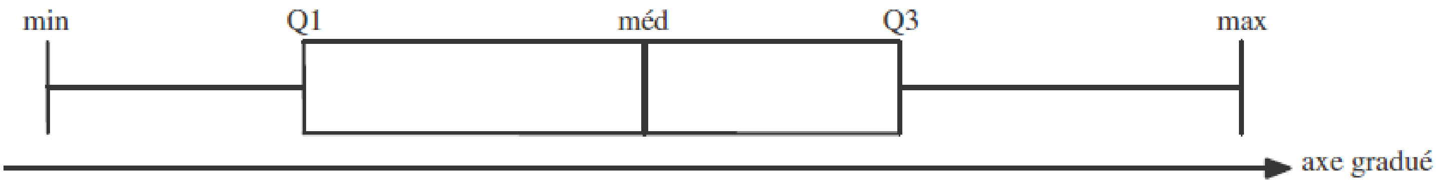
• Q_1 est la donnée de la série dont le rang est le premier entier supérieur ou égal à $\frac{n}{4}$

• Q_3 est la donnée de la série dont le rang est le premier entier supérieur ou égal à $\frac{3n}{4}$

• D_1 est la donnée de la série dont le rang est le premier entier supérieur ou égal à $\frac{n}{10}$

• D_9 est la donnée de la série dont le rang est le premier entier supérieur ou égal à $\frac{9n}{10}$

On peut effectuer une représentation graphique appelée diagramme en boîte :



On place sur un axe gradué : le minimum, le maximum, le premier quartile, le troisième quartile et la médiane.
On construit alors une boîte rectangulaire de largeur arbitraire dont les extrémités sont Q_1 et Q_3.
Un trait dans la boîte représente la médiane.
Le nombre $Q_3 - Q_1$ est appelé écart interquartile. L'intervalle $[Q_1 ; Q_3]$ est appelé intervalle interquartile.

Remarques :
• Ce diagramme est aussi appelé boîte à « moustaches ».
• Si l'étendue de la série est trop importante ou si on ne considère pas les valeurs extrêmes comme significatives on peut raccourcir les « moustaches » aux déciles D_1 et D_9.
• La boîte centrale représente l'intervalle interquartile et contient donc au moins 50% des données.
• On emploie ce type de diagramme pour comparer plusieurs séries entre elles.

V Variance et écart-type

Activité 9 :
Le patron d'une concession automobile veut récompenser le meilleur de ses trois vendeurs en lui offrant une prime. Pour cela, il étudie le nombre de voitures vendues lors de chacun des neuf premiers mois de l'année :

mois	Janvier	Février	Mars	Avril	Mai	Juin	Juillet	Aout	Sept.
Aurélien	5	4	5	11	7	11	4	3	4
Bérénice	3	11	7	4	4	5	5	9	6
Clément	4	3	11	5	11	7	3	5	5

Proposer un raisonnement permettant au patron de la concession de déterminer le « meilleur » de ses vendeurs.

Définitions :

• La variance d'une série statistique dont les valeurs du caractère sont $x_1, x_2, \dots, x_p$ d'effectifs correspondants $n_1, n_2, \dots, n_p$ et de moyenne $\overline{x}$ est égale à :

$$V = \frac{n_1\left(x_1 - \overline{x}\right)^2 + n_2\left(x_2 - \overline{x}\right)^2 + \dots + n_p\left(x_p - \overline{x}\right)^2}{N} = \frac{1}{N}\sum_{i=1}^{i=p} n_i\left(x_i - \overline{x}\right)^2$$

où $N = n_1 + n_2 + \dots + n_p = \sum_{i=1}^{i=p} n_i$ est l'effectif total

La variance est la moyenne des carrés des écarts entre les valeurs de la série et la valeur moyenne.

• L'écart-type d'une série statistique, noté s, est égal à la racine carré de la variance : $s = \sqrt{V}$

Remarques :

1) Parfois (notamment en probabilités et sur la calculatrice) l'écart type est noté σ.

2) **L'écart-type** s'exprime dans les **mêmes unités que les observations** et il est donc plus facile à interpréter que la variance.

3) L'écart type d'une série **rend compte de la répartition des données autour de la moyenne**.

Le **couple (Moyenne , écart type)** est un indicateur de la tendance centrale de la série : **plus l'écart type est proche de 0, plus les valeurs se concentrent autour de la moyenne.**

Ce couple donne beaucoup de poids aux valeurs extrêmes.

Malgré cela interpréter la valeur d'un écart-type n'est pas toujours une tâche simple : la construction de certains intervalles remarquables $[\overline{x} - s ; \overline{x} + s]$, $[\overline{x} - 2s ; \overline{x} + 2s]$ ou $[\overline{x} - 3s ; \overline{x} + 3s]$ permet de la faciliter

4) On peut aussi calculer l'écart moyen absolu d'une série mais l'outil utilisé (la valeur absolue) est peu maniable et cette mesure de dispersion est peu employée.