STATISTIQUES

I. Un peu de vocabulaire

Toute étude statistique s'appuie sur des données. Dans le cas où ces données sont numériques, on distingue les données discrètes (qui prennent un nombre fini de valeurs : par ex, le nombre d'enfants par famille en France) et des données continues (qui prennent des valeurs quelconques : par ex, la taille des hommes).

- Dans le cas d'une série discrète, le nombre de fois ou l'on retrouve la même valeur s'appelle l'effectif de cette valeur. Si cet effectif est exprimé en pourcentage, on parle alors de fréquence de cette valeur.
- Dans le cas d'une série continue, on répartit souvent les données par classes.

Le but des statistiques est d'analyser les données dont on dispose. Pour cela, on peut s'aider

- d'un graphique : Nous verrons notamment cette année les diagrammes circulaire, en bâtons, en boites, et les histogrammes.
- On peut aussi chercher à déterminer la moyenne ou la médiane de la série.
  De tels nombres permettent notamment de comparer plusieurs séries entre elles. On les appelle indicateurs statistiques ou paramètres statistiques.
  On distingue les indicateurs de position (qui proposent une valeur "centrale" de la série) et les indicateurs de dispersion (qui indiquent si la série est très regroupée autour de son "centre" ou non).

II. Representations graphiques

1. Diagramme circulaire

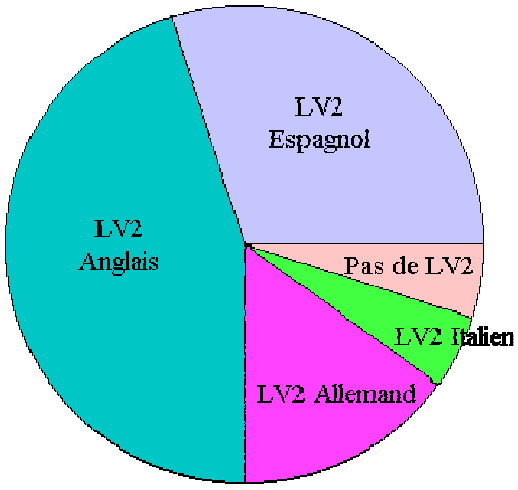
Pour construire un diagramme circulaire, il suffit de faire correspondre à chaque effectifs, pourcentage ou fréquence un angle de mesure proportionnelle.

Exemple :
On s'intéresse à la deuxième langue vivante choisie par les 500 élèves d'un lycée : 150 élèves font de l'espagnol en première langue, 225 des élèves font de l'anglais, 75 de l'allemand , 25 de l'italien et enfin 25 aucune deuxième langue.

On construit un tableau :

Effectif	500	150	225	75	25	25
Angle	360°	108°	162°	54°	18°	18°

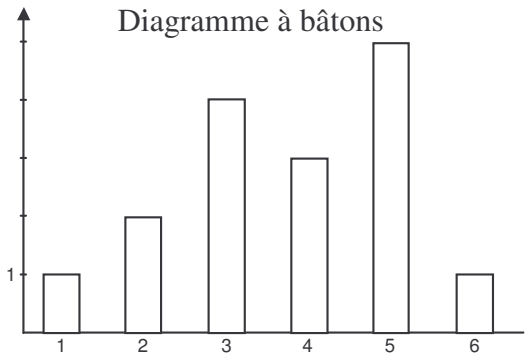
on obtient le diagramme circulaire suivant :

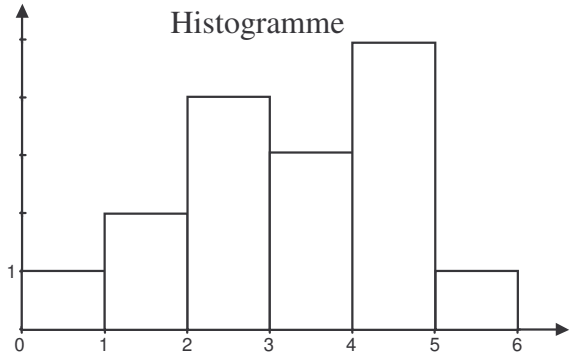


2. ***Diagramme en bâtons et histogramme***

Le diagramme en bâtons (ou en barres) est formé de bâtons (de barres) dont l'abscisse est la valeur x_i et la hauteur l'effectif n_i (ou la fréquence f_i).

L'histogramme est formé de rectangles dont la surface est proportionnelle à l'effectif (ou la fréquence) de la modalité.





Remarques :

- Dans le diagramme à bâtons, l'axe des abscisses n'est pas gradué et la largeur des bâtons ne signifie rien.
- Dans l'histogramme, l'axe des abscisses est gradué et les bâtons sont donc "collés" les uns aux autres. L'histogramme est donc surtout utilisé pour représenter graphiquement des séries continues où les données ont été réparties en classes.

Exemple :

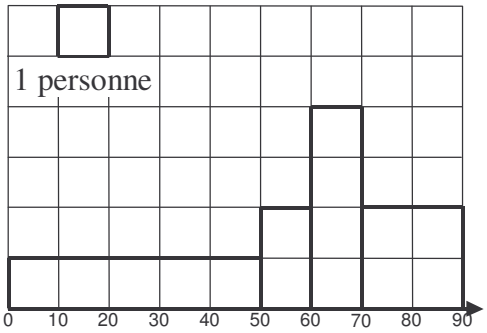
On résume dans un tableau le poids des personnes d'une chorale :

poids (Kg)	[0 ; 50[	[50 ; 60[	[60 ; 70[	[70 ; 90]
nbre de personnes	5	2	4	4

Pour construire cet histogramme, on réalise le tableau ci-dessous :

classe	[0 ; 50[	[50 ; 60[	[60 ; 70[	[70 ; 90]
effectif	5	2	4	4
amplitude	50	10	10	20
effectif/amplitude	0.1	0.2	0.4	0.2

On obtient :



On ne trace pas l'axe des ordonnées, car il aurait fallu le graduer en nombre de personnes par kilo
Par contre, pour permettre la lecture du graphique, on indique en légende la signification de l'unité d'aire.

III. PARAMETRES STATISTIQUES

1. Mode, étendue

Si les données d'une série sont discrètes, le mode est la ou les valeurs qui ont le plus grand effectif.
Si les données ont été réparties en classes, on parle alors plutôt de classe modale.
L'étendue d'une série est la différence entre la plus grande valeur et la plus petite.

Exemple de données discrètes : 9, 11, 8, 10, 13, 12, 10, 11, 10

Faisons le tableau des effectifs :

valeur	8	9	10	11	12	13
effectif	1	1	3	2	1	1

Le mode est la valeur qui a le plus gros effectif, c'est à dire 10
13 − 8 = 5 donc l'étendue de cette série est 5

Exemple de données réparties par classes :

classe	[0 ; 5[	[5 ; 10[	[10 ; 15[	[15 ; 20]
effectif	0	5	14	2

La classe modale est la classe qui a le plus gros effectif, c'est à dire la classe [10 ; 15 [
20 – 5 = 15 donc l'étendue de cette série est inférieure ou égale à 15

Remarque :
Par simplification, on dira souvent que l'étendue est 15 mais c'est un abus de langage ! En effet, dans le tableau des données ci dessus, rien ne permet d'affirmer que les valeurs extrêmes sont 5 et 20 !

2. Moyenne, écart-type

Soit la série statistique ci-contre :

valeurs	x_1	x_2	…	x_p
effectifs	n_1	n_2	…	n_p

La moyenne est : $\bar{x} = \dfrac{n_1x_1 + n_2x_2 + \ldots + n_px_p}{n_1 + n_2 + \ldots + n_p}$

L'écart type est : $\sigma = \sqrt{\dfrac{n_1(x_1 - \bar{x})^2 + n_2(x_2 - \bar{x})^2 + \ldots + n_p(x_p - \bar{x})^2}{n_1 + n_2 + \ldots + n_p}}$

Remarques :
- L'écart type mesure la dispersion de la série autour de sa moyenne.
- On parle aussi de variance de la série. Il s'agit en fait de σ^2 . L'avantage de l'écart type sur la variance est qu'il s'exprime, comme la moyenne, dans la même unité que les données.
- Dans le cas de données regroupées en classes, on ne peut calculer la valeur exacte de la moyenne ou de l'écart type. On peut toutefois en déterminer une bonne approximation en remplaçant chaque classe par son milieu dans les formules ci-dessus.

Exemple avec un tableau des fréquences :

valeurs	12	13	14	15	16
fréquences	0,05	0,17	0,43	0,30	0,05

$\bar{x} = 14{,}13$ et $\sigma = 0{,}92$

Exemple avec un tableau des fréquences :

classes	[0 ; 5[	[5 ; 10[	[10 ; 15[	[15 ; 20[
effectifs	7	12	14	2

$\bar{x} = 9{,}07$ et $\sigma = 4{,}27$

IV. MEDIANE, QUARTILES, DECILES

1. Définitions

Soit une série rangée par ordre croissant. Appelons *n* l'effectif total de la série.

Définitions	Pour déterminer le rang
La médiane C'est la valeur "centrale" de la série. Elle partage la série en deux moitiés.	si *n* est impair : la médiane est la valeur de rang $\frac{n+1}{2}$ si *n* est pair : On prend la moyenne des deux valeurs qui sont au centre de la série, c'est à dire dont les rangs entourent le nombre $\frac{n+1}{2}$
Les quartiles (partagent la série en 4 : il y en a donc 3) Le 1er quartile Q1 est la plus petite valeur telle que 25% des données lui soit inférieures ou égales. Le 3ème quartile Q3 est la plus petite valeur telle que 75% des données lui soit inférieures ou égales.	Q1 est la valeur dont le rang est le premier entier supérieur ou égal à $\frac{n}{4}$ Q3 est la valeur dont le rang est le premier entier supérieur ou égal à $\frac{3n}{4}$
Les déciles (partagent la série en 10 : il y en a donc 9) Le 1er décile D1 est la plus petite valeur telle que 10% des données lui soit inférieures ou égales. Le 9ème décile D9 est la plus petite valeur telle que 90% des données lui soit inférieures ou égales.	D1 est la valeur dont le rang est le premier entier supérieur ou égal à $\frac{n}{10}$ D9 est la valeur dont le rang est le premier entier supérieur ou égal à $\frac{9n}{10}$

Remarques :

- Les trois nombres Q1, Q2 = méd, Q3 partagent la série en 4 parts égales (à une unité près)
- Si les données ont été réparties en classes, on ne peut déterminer la médiane exacte. En revanche, on appellera classe médiane, la classe qui la contient (et permet donc d'en donner un encadrement).
- L'intervalle [Q1 ; Q3] s'appelle l'intervalle interquartile.
- Le nombre Q3 – Q1 s'appelle l'écart interquartile.

Exemples avec des données discrètes "en vrac" : 21, 25, 28, 30, 27, 24, 31, 21, 28, 30, 25, 28, 26, 25

Ordonnons la série par ordre croissant :
21, 21, 24, 25, 25, 25, 26, 27, 28, 28, 28, 30, 30, 31

Il y a 14 termes :
$\frac{14+1}{2}$ *= 7,5. La médiane est donc la demi somme des 7 ème et 8 ème termes : méd =* $\frac{26+27}{2}$ *=26,5*
$\frac{14}{4}$ *= 3,5. Le 1er quartile est donc le 4 ème terme : Q1 =25*
$\frac{3\times14}{4}$ *= 10,5. Le 3ème quartile est donc le 11 ème terme : Q3 =28*
8 − 5 = 3. L'écart interquartile est donc3

Exemples avec un tableau d'effectifs

valeur	*1*	*2*	*3*	*4*	*5*	*6*
effectif	*6*	*11*	*25*	*19*	*15*	*5*
effectif cumulé	*6*	*17*	*42*	*61*	*76*	*81*

L'effectif total est de 81

$\frac{81+1}{2}$ *= 41. La médiane est donc le 41* ^{ème} *terme : méd =3*

$\frac{81}{10}$ *= 8,1. Le 1*^{er} *décile est donc le 9* ^{ème} *terme : D1 =2*

$\frac{81}{4}$ *= 20,25. Le 1*^{er} *quartile est donc le 21* ^{ème} *terme : Q1 =3*

$\frac{3\times81}{4}$ *= 60,75. Le 3*^{ème} *quartile est donc le 61* ^{ème} *terme : Q3 =4*

$\frac{9\times81}{10}$ *= 72,9. Le 9*^{ème} *décile est donc le 73* ^{ème} *terme : D9 =5*

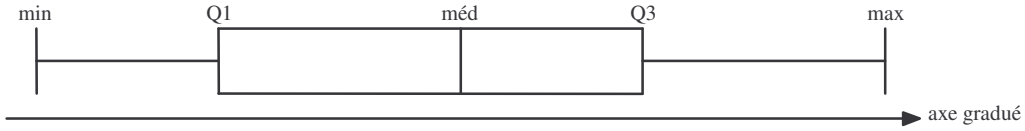
Exemples avec un tableau de fréquences réparties par classes :

classe	*[0 ; 2[*	*[2 ; 4[*	*[4 ; 6[*	*[6 ; 8]*
fréquence	*10%*	*38%*	*45%*	*7%*
fréquence cumulée	*10*	*48*	*93*	*100*

48% des valeurs sont inférieures à 4
93% des valeurs sont supérieures ou égales à4
La classe médiane est donc la classe [4 ; 6 [
On peut donc en déduire l'encadrement suivant 4 < méd < 6

2. ***Diagramme en boites***

Le diagramme en boîte d'une série à l'allure suivante :



Remarques :

- Lorsque la série est trop importante, que l'on ne connaît pas les valeurs extrêmes ou qu'on les considère comme non significatives, on raccourci souvent les moustaches au déciles D1 et D9.
- La boîte centrale représente l'intervalle interquartile et contient donc la moitié des données.
- Il faut légender le diagramme (min, max, nom de la série) et graduer l'axe.
- On emploie surtout ce type de diagramme pour comparer plusieurs séries entre elles.
- Ces diagrammes ont reçu beaucoup de noms différents : boîtes à pattes, diagrammes à moustaches,…

Exemple
Deux classes de 1STG comparent leurs résultats du trimestre et déclarent : "nos classes ont le même profil puisque dans les deux cas la médiane des résultats est 10". Qu'en pensez-vous ?

notes	5	6	7	8	9	10	11	12	13	14	15	16
effectifs 1STG1	0	3	4	4	5	7	3	4	2	1	0	0
effectifs 1STG2	2	4	3	3	3	4	3	2	2	3	1	2

Vérifier que les deux médianes valent 10 et déterminer les quartiles de chaque série
Tracer côte à côte les diagrammes en boites de ces deux séries.

Pour la 1STG1 *: L'effectif total est 3+4+4+…+1 = 33*

$\frac{33+1}{2}$ *= 17 donc la médiane est le 17*^{ème} *terme de la série : Méd = 10*

$\frac{33}{4}$ *= 8,25 donc le 1*^{er} *quartile est le 9*^{ème} *terme de la série : Q1 = 8*

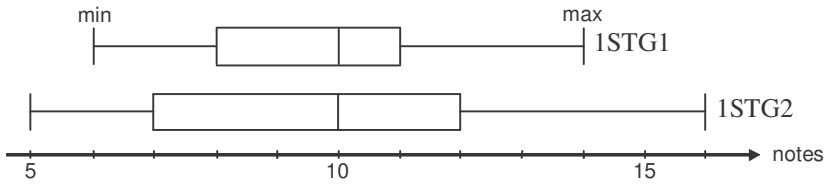
$\frac{3\times33}{4}$ *= 24,75 donc le 3*^{ème} *quartile est le 25*^{ème} *terme de la série : Q3 = 11*

Pour la 1STG2 *: L'effectif total est 2+4+3+…+2 = 32*

$\frac{32+1}{2}$ *= 16,5 donc la médiane est la moyenne des* $16^{ème}$ *et* $17^{ème}$ *terme de la série : Méd =* $\frac{10+10}{2} = 10$

$\frac{32}{4}$ *= 8 donc le* 1^{er} *quartile est le* $8^{ème}$ *terme de la série : Q1 = 7*

$\frac{3\times32}{4}$ *= 24 donc le* $3^{ème}$ *quartile est le* $24^{ème}$ *terme de la série : Q3 = 12*



Bilan *: Le graphique ci-dessus met bien en évidence que l'écart interquartile et l'étendue sont plus resserrés en 1STG1 qu'en 1STG2 donc, les élèves de 1STG1 ont globalement un niveau plus homogène que ceux de 1STG2.*

V. ETUDE SIMULTANEE DE DEUX SERIES

Sur une même population, on considère deux caractères A et B ayant chacun deux modalités.
On représente les effectifs pour étudier simultanément les deux caractères A et B.

Exemple :
En observant les résultats de 60 candidats l'oral du bac de français, Eric affirme : 65% des élèves ayant la moyenne sont des filles.
Ophélie quant à elle dit : 78% des garçons ont la moyenne contre seulement 62% des filles.

	Filles	Garçons	Total
Notes ⩾ 10	26	14	40
Notes < 10	16	4	20
Total	42	18	60

Vérifier les calculs d'Eric et d'Ophélie
Compléter les tableaux suivants des fréquences conditionnelles en pourcentages :

	Filles	Garçons	Total
Notes ⩾ 10	65	35	100
Notes < 10	80	20	100

	Filles	Garçons
Notes ⩾ 10	62	78
Notes < 10	38	22
Total	100	100