STATISTIQUES

I) UN PEU DE VOCABULAIRE

Toute étude statistique s'appuie sur des données. Dans le cas ou ces données sont numériques (99% des cas), on distingue les données discrètes (qui prennent un nombre fini de valeurs : par ex, le nombre de voitures par famille en France) des données continues (qui prennent des valeurs quelconques : par ex, la taille des animaux d'un zoo).

- 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. (cf 17 p82)
- Dans le cas d'une série continue, on répartit souvent les données par classes. (cf 13 p82)

Dans les exercices, les données se présenteront donc ainsi :

données numériques	discrètes	données "en vrac"
		tableau des effectifs ou des fréquences
	continues	données "en vrac"
		données réparties par classes

Le but des statistiques est d'analyser les données dont on dispose :

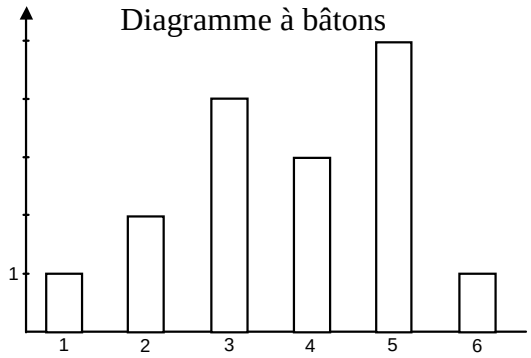
- Pour cela, on peut s'aider d'un graphique : Nous verrons notamment cette année les diagrammes à bâtons, les histogrammes et les diagrammes en boîtes (ou à moustaches).
- On peut aussi chercher à déterminer la moyenne ou la médiane de la série. De tels nombres permettent notamment de comparer plusieurs 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).
  Nous étudierons cette année les indicateurs statistiques suivants :

Indicateurs de position :		**Indicateurs de dispersion :**
mode, classe modale		étendue
médiane, classe médiane	quartiles, déciles	écart interquartile
moyenne		écart type

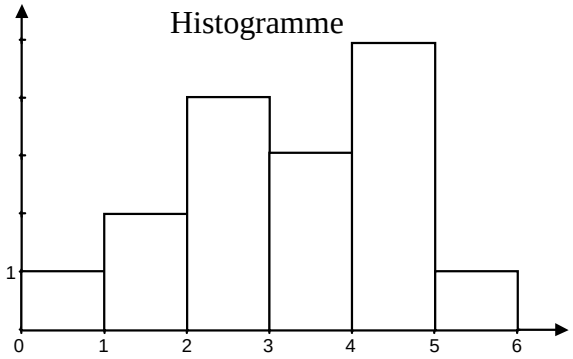
II) DIAGRAMMES A BATONS OU HISTOGRAMMES

1) Quelles différences voyez-vous entre les deux graphiques ci-dessous ?

Diagramme à bâtons



Histogramme



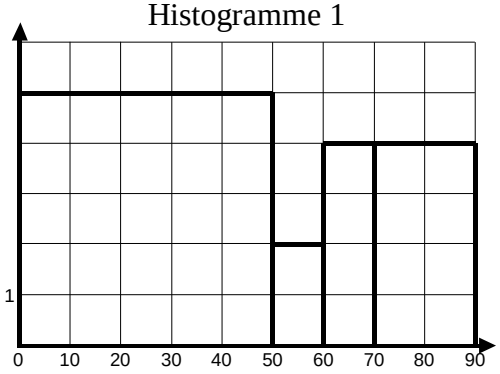
- Dans le diagramme à bâtons, l'axe des abscisses n'est pas gradué et la largeur des bâtons ne signifie rien.
- Dans l'histogramme, l'axe des abscisses est gradué et les bâtons sont donc "collés" les uns aux autres. L'histogramme est donc surtout utilisé pour représenter graphiquement des séries continues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

2) Le cas des classes d'amplitudes différentes

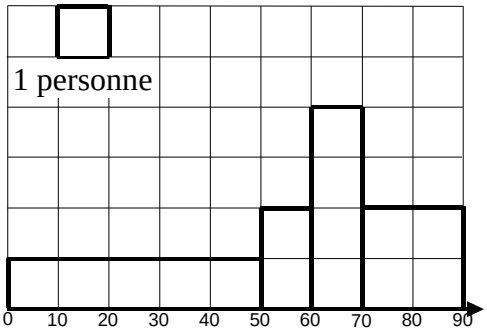
Pour représenter la série ci-contre, quel est le graphique le plus équitable ?

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

Histogramme 1



Histogramme 2



- L'histogramme 1 est inadapté car il laisse entendre que la majorité des gens pèsent moins de 50 kg !
- L'histogramme 2 est équitable car on a pondéré la hauteur de chaque bâton en tenant compte de l'amplitude de la classe. Pour construire ce deuxième histogramme, on réalise le tableau ci-dessous :

classe	[0 ; 50[	[50 ; 60[	[60 ; 70[	[70 ; 90]
effectif	5	2	4	4
amplitude				
effectif/amplitude				

Remarques :

- Dans l'histogramme 1, c'est la hauteur des bâtons qui permet de lire l'effectif.
  Dans l'histogramme 2, c'est l'aire des bâtons qui permet de lire l'effectif.
- Dans l'histogramme 2, nous n'avons pas tracé l'axe des ordonnées, car il aurait fallu le graduer en nombre de personnes par kilo ! Par contre, pour permettre la lecture du graphique, nous avons indiqué en légende la signification de l'unité d'aire.
- Dans les exercices, quand les classes ont toutes la même amplitude, on fait un histogramme de type 1, quand les classes ont des amplitudes différentes, ont fait un histogramme de type 2.

1L-exo-statistiques.doc : A, B, C, D, E, F, G, H

III) MODE, ETENDUE

1) Définitions

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.

2) Dans les exercices :

a) Données discrètes 9, 11, 8, 10, 13, 12, 10, 11, 10

Faisons le tableau des effectifs :

valeur	8	9	10	11	12	13
effectif						

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

Remarque :
Ici, vu le petit nombre de données, faire un tableau des effectifs est un peu artificiel. Par contre, dès que l'on travaille sur un nombre important de données, il devient vite très utile pour mettre en évidence le mode et l'étendue de la série.

b) 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
- 20 – 5 = 15 donc l'étendue de cette série est inférieure ou égale à

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 !

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 <u>médiane</u> • C'est la valeur "centrale" de la série. On dit qu'elle partage la série en deux moitiés	• si *n* est impair : la médiane est la valeur de rang • si *n* est pair : nous prendrons la moyenne des deux valeurs qui sont au centre de la série, c'est à dire dont les rangs entourent le nombre
Les <u>quartiles</u> (partagent la série en 4 : il y en a donc)	
• Le 1er quartile Q1 est la plus petite valeur telle que 25% des données lui soit inférieures ou égales.	• Q1 est la valeur dont le rang est le premier entier supérieur ou égal à
• Le 3ème quartile Q3 est la plus petite valeur telle que 75% des données lui soit inférieures ou égales.	• Q3 est la valeur dont le rang est le premier entier supérieur ou égal à
Les <u>déciles</u> (partagent la série en 10 : il y en a donc)	
• Le 1er décile D1 est la plus petite valeur telle que 10% des données lui soit inférieures ou égales.	• D1 est la valeur dont le rang est le premier entier supérieur ou égal à
• Le 9ème décile D9 est la plus petite valeur telle que 90% des données lui soit inférieures ou égales.	• D9 est la valeur dont le rang est le premier entier supérieur ou égal à

<u>Remarques :</u>

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

2) Dans les exercices :

a) 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 ème et ème termes : méd = $\frac{+}{2}$ =
- $\frac{14}{4} = 3{,}5$. Le 1er quartile est donc le ème terme : Q1 =
- $\frac{3\times14}{4} = 10{,}5$. Le 3ème quartile est donc le ème terme : Q3 =
- $8 - 5 = 3$. L'écart interquartile est donc

b) Tableau d'effectifs

valeur	1	2	3	4	5	6
effectif	6	11	25	19	15	5
effectif cumulé						

⚠ Bien interpréter la dernière ligne !
La valeur 3 va du rang au rang

L'effectif total est de

- $\frac{81+1}{2} = 41$. La médiane est donc le ème terme : méd =
- $\frac{81}{10} = 8{,}1$. Le 1er décile est donc le ème terme : D1 =
- $\frac{81}{4} = 20{,}25$. Le 1er quartile est donc le ème terme : Q1 =
- $\frac{3\times81}{4} = 60{,}75$. Le 3ème quartile est donc le ème terme : Q3 =
- $\frac{9\times81}{10} = 72{,}9$. Le 9ème décile est donc le ème terme : D9 =

c) Données réparties par classes

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

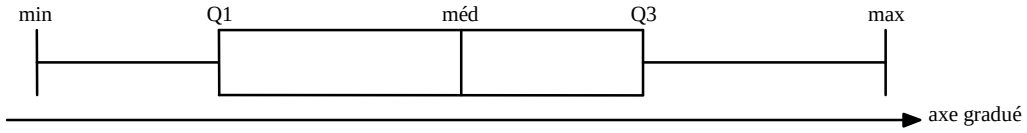
⚠ Bien interpréter ce tableau !
45% des valeurs sont comprises entre
93% des valeurs sont

48% des valeurs sont
Et 93% des valeurs sont
La classe médiane est donc la classe
On peut donc en déduire l'encadrement suivant ⩽ méd <

p80: 3, 4
1L-exo-statistiques.doc : I, J, K, L
vidéo-projecteur : 1L-cmp-quartiles.xls

3) Diagrammes en boîtes

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.
- Vous devez légender votre 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,…

Ex :
Deux classes de 1L 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 1L1	0	3	4	4	5	7	3	4	2	1	0	0
effectifs 1L2	2	4	3	3	3	4	3	2	2	3	1	2

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

Pour la 1L1 : L'effectif total est 3+4+4+…+1 = 33

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

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

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

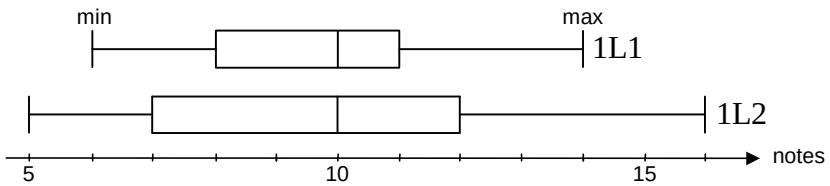
Pour la 1L2 : L'effectif total est 2+4+3+…+2 = 32

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

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

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

Diagrammes en boîtes :



Bilan : Le graphique ci-dessus met bien en évidence que l'écart interquartile et l'étendue sont plus resserrés en 1L1 qu'en 1L2 donc les élèves de 1L1 ont globalement un niveau plus homogène que ceux de 1L2.

p81: 7
p82: 12
1L-exo-statistiques.doc : M, N, O, P
vidéo-projecteur : 1L-outil-boites-a-moustaches.xls

V) MOYENNE, ECART TYPE, DONNEES GAUSSIENNES

1) Définitions

AP sur l'écart type puis reprendre les données, les présenter dans un tableau d'effectif et en déduire les 2 formules ci-dessous

Soit la série statistique ci-contre :

valeurs	x_1	x_2	…	x_p
effectifs	n_1	n_2	…	n_p

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

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

Remarques :

- L'écart type mesure la dispersion de la série autour de sa moyenne.
- Vous entendrez aussi parler de variance de la série. Il s'agit en fait de σ^2

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

  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.

2) Dans les exercices :

a) Tableau des fréquences

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

$\overline{x}$ =

σ =

b) Données réparties en classes

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

Remplaçons chaque classe par son milieu :

$\overline{x} \approx$

$\sigma \approx$

p82: 13, 14
1L-exo-statistiques.doc : Q, R, S
Salle info : 1L-cmp-moyenne-ecart-type.xls

3) Propriétés

a) Addition ou Multiplication de toutes les données par un même nombre :

Ex Soit la série : 10, 12, 14. $\overline{x}$ = et σ =
Ajoutons 2 : la nouvelle série est : 12, 14, 16. $\overline{x}$ = et σ =
Divisons par 2 : la nouvelle série est : 6, 7, 8. $\overline{x}$ = et σ =

Cas général : Soit α un réel quelconque :

- Si l'on ajoute α à toutes les données, la moyenne augmente d'α
  l'écart type ne change pas
- Si on multiplie toutes les données par α, la moyenne est multipliée par α
  l'écart type est multipliée par α

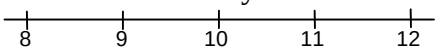
b) Moyennes partielles

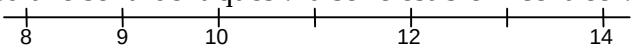
Ex : Sur les 5 premières interros, Paul a eu 12,5 de moyenne. Il vient d'avoir 15,5 à la 6ème interro.
Les notes ayant toutes le même coefficient, quelle est sa nouvelle moyenne ?
La somme des notes des 5 premières interros est : $12,5 \times 5$
La somme des notes des 6 interros est donc : $12,5 \times 5 + 15,5$
La nouvelle moyenne est donc : $\overline{x} = \dfrac{12,5 \times 5 + 15,5}{6} = 13$

Cas général : Si on réunis deux groupes disjoints ayant respectivement pour moyennes et effectifs, $\overline{x}_1$ et n_1 d'une part, $\overline{x}_2$ et n_2 d'autre part, la moyenne de l'ensemble sera alors :

$$\overline{x} = \frac{n_1 \times \overline{x}_1 + n_2 \times \overline{x}_2}{n_1 + n_2}$$

4) Moyenne et médiane

- Quand on modifie les valeurs extrêmes d'une série, la moyenne change contrairement à la médiane qui ne change pas. On dit que la moyenne est "sensible aux valeurs extrêmes".
  Il arrive que certaines de ces valeurs extrêmes soient douteuses ou influent de façon exagérée sur la moyenne. On peut alors, soit calculer une moyenne élaguée (c'est à dire recalculer la moyenne sans ces valeurs gênantes), soit utiliser la médiane.
- Comment interpréter un écart entre la moyenne et la médiane ?
  Soit la série suivante : 

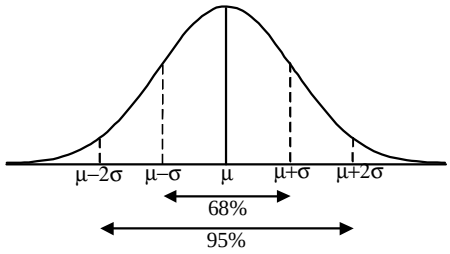
  Ici la moyenne et la médiane sont identiques : la série est bien "centrée".
  Soit la nouvelle série : 

  Ici la moyenne est plus importante que la médiane : la série est plus "étalée vers la droite".

p82: 16, 17
1L-exo-statistiques.doc : T, U, V

5) Données Gaussiennes

Dans de très nombreuses situations (issues de la biologie, géographie, sociologie, économie…) les données se présentent graphiquement sous la forme de courbes "en cloche" dites de Gauss.
Le comportement de ces séries est modélisable par une loi mathématique appelé loi normale ou loi de Gauss qui donne une grande importance à la moyenne μ et à l'écart type σ :



- Ces séries sont à peu près symétriques autour de μ
- Environ 68% des données sont dans l'intervalle $[\mu - \sigma ; \mu + \sigma]$
- Environ 95% des données sont dans l'intervalle $[\mu - 2\sigma ; \mu + 2\sigma]$
- Environ 99% des données sont dans l'intervalle $[\mu - 3\sigma ; \mu + 3\sigma]$

Les intervalles ci-dessus sont appelés plages de normalité pour les niveaux de confiance 0,68 ; 0,95 ; 0,99

Remarque : les observations ci-dessus n'ont aucun sens pour :
Les séries qui traduisent des phénomènes non gaussiens
Les séries gaussiennes pour lesquelles l'échantillon est trop petit.

VI) QUELS INDICATEURS STATISTIQUES UTILISER ?

Dans la pratique :

- On utilise très peu le mode et l'étendue (faciles à déterminer mais simplistes !)
- On utilise la médiane, quartiles, déciles et écart interquartile surtout pour les séries à grands effectifs (pas de calculs, il suffit d'ordonner la série ; peu sensible aux valeurs douteuses)
- On utilise souvent la moyenne et l'écart type pour des séries de tailles intermédiaires ou des séries gaussiennes (la moyenne reste l'indicateur le plus intuitif ; intérêt des plages de normalité)

p81: 10
p82: 15
p83: 19
1L-exo-statistiques.doc : W, X, Y, Z
Salle info : 1L-cmp-donnees-gaussiennes.xls