Statistiques

I. Présentation et Vocabulaire de base

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 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 où 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 par exemple 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).
Ainsi, le mode d'une série (valeur qui a le plus grand effectif de la série) est un indicateur de position.
L'étendue de cette série (différence entre la plus grande et la plus petite valeur) est un indicateur de dispersion.
La moyenne et la médiane sont des indicateurs de position.

II. Médiane

Définition
Soit une série statistique d'effectif total n, rangée par ordre croissant.
On appelle médiane la valeur "du milieu". On dit qu'elle partage la série en deux moitiés : il y a autant de valeurs en dessous qu'au-dessus.
Pour déterminer son rang, il y a 2 cas :

- si n est impair : la médiane est la valeur de rang $\frac{n+1}{2}$
- si n est pair : nous prendrons la demi-somme des deux valeurs dont les rangs entourent le nombre $\frac{n+1}{2}$

Méthode 1 : Déterminer une médiane (avec une série de valeurs)
Déterminer la médiane pour les deux cas suivants :
Série 1 : 62, 44 , 46 , 18, 43 Série 2 : 18, 43, 44, 46, 62, 95.

Méthode 2 : Déterminer une médiane (avec une série compilée dans un tableau)
Déterminer la médiane pour la série suivante :

valeurs	1	2	3	4	5	6
effectifs	6	11	25	19	15	5
effectifs cumulés						

Remarque :
Si les données ont été regroupées en classes, on ne peut déterminer la valeur exacte de la médiane. En revanche, on appellera classe médiane, la classe qui la contient (et permet donc d'en donner un encadrement).

Exemples
Avec des données 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 strictement inférieures à 4
Et 93% des valeurs sont strictement inférieures à 6
La classe médiane est donc la classe [4 ; 6[
On peut donc en déduire l'encadrement suivant $4 < \text{méd} < 6$
On verra plus tard comment tirer une approximation de la médiane

III. Quartiles

Définitions :
Le premier quartile Q_1 est la plus petite des valeurs de la série telle qu'au moins 25% de la population ait sa valeur inférieur ou égale à Q_1.
Le troisième quartile Q_3 est la plus petite des valeurs de la série telle qu'au moins 75% de la population ait sa valeur inférieure ou égale à Q_3.

Méthode 3 : **Méthode pour trouver Q_1 et Q_3.**
Déterminer le premier et de troisième quartile de la série suivante :

valeurs	1	2	3	4	5	6
effectifs	6	11	25	19	15	5
effectifs cumulés	6	17	42	61	76	81

Exercice 1
Soit la série statistique donnée par le tableau suivant :

x_i	0	1	2	3	4	5	6	7
n_i	7	10	3	11	8	5	2	2

1) Déterminer la médiane Me et les quartiles Q_1 et Q_3.
2) Construire le diagramme en boîte.

Exercice 2
Déterminer les médianes et les quartiles des séries suivantes

valeurs	**5**	**6**	**7**	**9**	**13**
effectifs	**3**	**5**	**1**	**2**	**4**
effectifs cumulés croissants					

valeurs	**5**	**6**	**9**	**10**	**13**
effectifs	10	5	6	4	5
effectifs cumulés croissants					

IV. Moyenne

Exemple :

Soit la série statistique ci-contre :

valeurs	0	1	2	3	4
effectifs	1	2	1	4	2

La moyenne est : $\bar{x} = \dfrac{0 + 1 + 1 + 2 + 3 + 3 + 3 + 3 + 4 + 4}{1 + 2 + 1 + 4 + 2} = \dfrac{24}{10} = 2{,}4$

On préférera écrire : $\bar{x} = \dfrac{0 + 2\times1 + 2 + 4\times3 + 2\times4}{1 + 2 + 1 + 4 + 2} = \dfrac{24}{10} = 2{,}4$

Notation.

Le symbole Σ (sigma) signifie que l'on ajoute les éléments, par exemple

$$\sum_{i=1}^{5} x_i = x_1 + x_2 + x_3 + x_4 + x_5$$

Formule pour le calcul de la moyenne

Soit la série statistique ci-contre :

valeurs	x_1	x_2	…	x_p
effectifs	n_1	n_2	…	n_p

La <u>moyenne</u> est : $\bar{x} = \frac{n_1x_1+n_2x_2+\ldots+n_{N-1}x_{N-1}+n_Nx_N}{n_1+n_2+\ldots+n_{N-1}+n_N} = \frac{\sum_{i=1}^{N} n_ix_i}{\sum_{i=1}^{N} n_i}$

Remarque :

Si les données ont été regroupées en classes, on ne peut calculer la valeur exacte de la moyenne. On peut toutefois en déterminer une bonne approximation en remplaçant chaque classe par son milieu.

<u>Méthode 4 :</u> **Méthode pour trouver la moyenne avec des classes ou des fréquences.**

Déterminer les moyennes des séries suivantes :

Cas : a) Tableau de fréquences

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

Cas b) Données réparties en classes

Milieu de classe				
classes	[0 ; 5[	[5 ; 10[	[10 ; 15[	[15 ; 20]
effectifs	7	12	14	2

V. données regroupées en classe : histogrammes et polygones d'effectifs cumulés

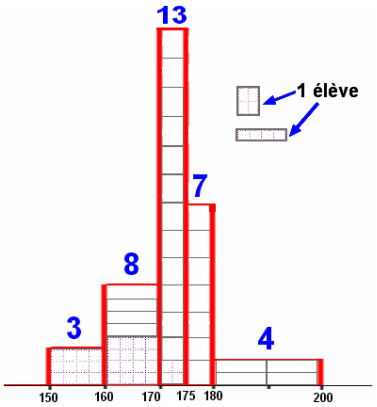
Un histogramme est la représentation graphique d'une variable continue.
A chaque classe de la variable, correspond la surface d'un rectangle qui a pour base l'amplitude de cette classe.
L'amplitude est la différence entre la borne supérieure et la borne inférieure de la classe.
Deux cas sont à considérer :
Les amplitudes sont égales, alors les hauteurs des rectangles sont proportionnelles aux effectifs ou aux fréquences. La construction est sensiblement la même que pour un diagramme en bâtons. (vu en troisième)
Les amplitudes sont inégales, il faudra alors corriger la hauteur des rectangles de manière à ce que leur surface corresponde bien aux effectifs ou aux fréquences.
Choisir une échelle simple et déterminer une graduation tenant compte des valeurs des classes.
Porter ces classes en abscisse et sur chacune d'elles prises comme base construire un rectangle dont l'aire (et non pas la hauteur) est proportionnelle à l'effectif ou à la fréquence de la classe considérée.
Il ne doit donc pas y avoir de graduations verticales mais une unité d'aire
On ne part pas forcément de l'origine

Exemple : Étude de la taille en cm des élèves dans une classe de seconde :
Ici les amplitudes ne sont pas égales.
Par exemple, l'amplitude de la classe [170 – 175[est 5 et celle de la classe [180 – 200[est 20.
Comme c'est la surface des rectangles qui représente l'ampleur du phénomène étudié, il faudra corriger leur hauteur :
Nous allons calculer le ratio "effectif / amplitude" afin de déterminer les hauteurs h des différents rectangles.
En fonction des données, ces hauteurs pourront être traduites en carreaux (cx) ou en unité de longueur (cm ou mm)

Tailles	Effectif
[150 – 160[	3
[160 – 170[	8
[170 – 175[	13
[175 – 180[	7
[180 – 200[	4

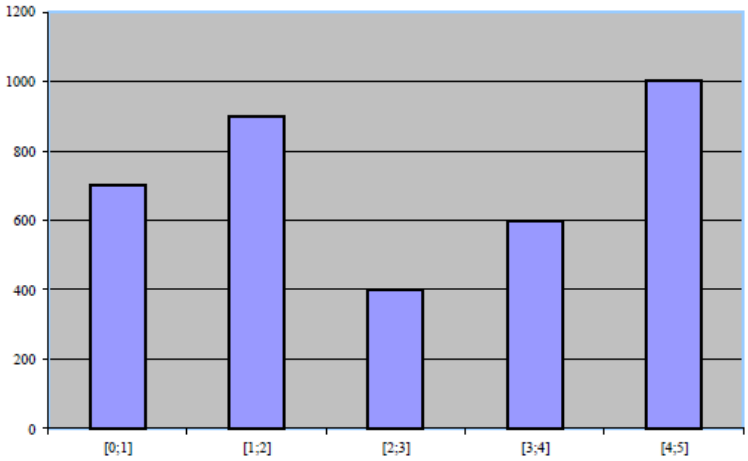
Traitement :

Tailles	Effectif	amplitude	eff / ampl	h en cx ou cm
[150 – 160[	3	10	3/10 = 0,3	3
[160 – 170[	8	10	8/10 = 0,8	8
[170 – 175[	13	5	13/5 = 2,6	26
[175 – 180[	7	5	7/5 = 1,4	14
[180 – 200[	4	20	4/20 = 0,2	2



Remarques :
1) h est une HAUTEUR en carreaux et NON un nombre de carreaux par amplitude.
2) Quand le nombre de carreaux risque de ne pas être entier, on préférera des hauteurs h en cm ou mm.

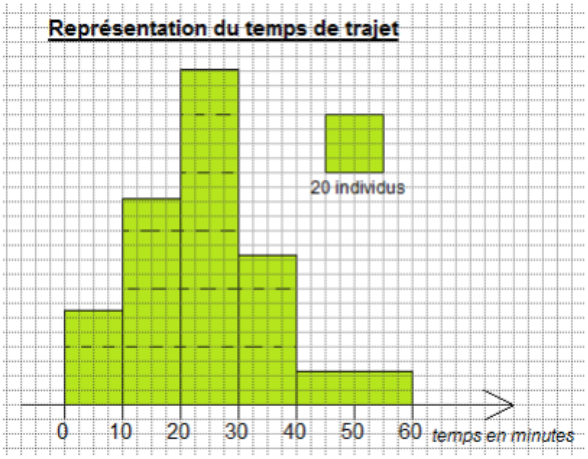
Exercice 3:
Pendant 5 heures, on a recensé les véhicules qui se sont présentés à un péage d'autoroute, et on a présenté les résultats sur l'histogramme ci-contre.
On admet que la répartition est uniforme dans chaque classe.
Construire le diagramme en boîte représentant la répartition des véhicules



Méthode 5 : **Méthode pour trouver la médiane et les quartiles d'une série en classes.**
Déterminer la médiane et les quartiles de la série suivante en utilisant un polygone de fréquences cumulées croissantes :

	F.C.C. en %
Moins de 5 ha	5,5
De 5 à moins de 20 ha	8,3
De 20 à moins de 50 ha	20,6
De 50 à moins de 75 ha	46
De 75 à moins de 100 ha	66,4
De 100 à moins de 200 ha	94,1
200 ha et plus	100

Exercice 4
En utilisant l'histogramme suivant remplir le tableau, puis déterminer la médiane et les quartiles de la série en utilisant un polygone de fréquences cumulées croissantes :



Correction

Temps de trajet en minutes	[0 ;10[	[10 ;20[	[20 ;30[	[30 ;40[	[40 ;60[
Centre de classes x_i					
Effectif n_i	34	72	116	53	25
$x_i\ n_i$					
Fréquences en % f_i					
Fréquence cumulée croissante					

Exercice 5
Compléter le tableau, puis déterminer la médiane et les quartiles de la série en utilisant un polygone de fréquences cumulées croissantes :

Classes	[0 ;20[	[20 ;40[	[40 ;60[	[60 ;80[	[80 ;100]
Centre de classes x_i					
Effectif n_i	1000	5000	10 000	30 000	59 000
Fréquences en % f_i					
Fréquence cumulée croissante					

Exercice 6
Déterminer la médiane et les quartiles en utilisant un polygone d'effectif cumulé.

Note	7	8	9	11	12	12,5	14	17
Effectif	3	6	5	2	3	1	1	1
Effectif cumulé croissant	3	9	14	16	19	20	21	22

VI. Propriétés avancées

a) Addition ou Multiplication de toutes les données par un même nombre :
Ex : Soit la série : 10, 12, 14. $\overline{x}$ =
Ajoutons 2 : la nouvelle série est : 12, 14, 16. $\overline{x}$ =
Multiplions par $\frac{1}{2}$: la nouvelle série est : 6, 7, 8. $\overline{x}$ =

Théorème de linéarité de la moyenne.
Si toutes les valeurs d'une série sont multipliées par un nombre *k* alors la moyenne est aussi multipliée par *k*.
Si on ajoute un nombre *k* à toutes les valeurs d'une série alors la moyenne est augmentée de *k*.

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 : 5× 12,5
La somme des notes des 6 interros est donc : 5×12,5 + 15,5
La nouvelle moyenne est donc : $\bar{x} = \frac{5\times12,5+15,5}{6} = 13$

Cas général : Si on réunit 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 :

$$\bar{x} = \frac{n_1 \times \overline{x_1} + n_2 \times \overline{x_2}}{n_1 + n_2}$$

Exercice 7
Dans un club sportif, le groupe des sauteurs à la perche est composé de 27 athlètes ; l'entraîneur a partagé ce groupe en deux :

- le groupe A composé des 10 athlètes qui sautent 5 mètres et plus,
- et le groupe B composé des 17 autres qui sautent moins de 5 mètres

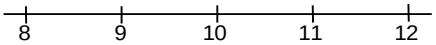
1) La moyenne des performances du groupe A est 5,40 m, celle du groupe B est 4,60 m. Quelle est la moyenne des performances de l'ensemble des athlètes ?
2) Après quelques entraînements, la moyenne des performances du groupe A augmente de 10 cm, celle du groupe B augmente de 20 cm. De combien la moyenne de l'ensemble est-elle modifiée ?

Exercice 8
Dans l'entreprise A, il y a 100 ouvriers et 100 cadres et dans l'entreprise B, il y a 180 ouvriers et 20 cadres.
Dans l'entreprise A (respectivement dans l'entreprise B) , le total des salaires mensuels des ouvriers est égal à 700 000F (respectivement 1 350 000F) et celui des cadres est 1 3000 000F (respectivement 280 000F).
1) Vérifier que les salaires mensuels moyens des ouvriers comme des cadres sont plus élevés dans l'entreprise B que dans l'entreprise A.
2) Comparer le salaire mensuel moyen de l'ensemble des employés dans les entreprises A et B.
Remarque : cet exercice donne un exemple d'« effet de structure »

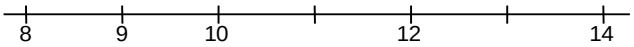
c) Lien entre la moyenne et la médiane

- 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 (10) et la médiane (10) sont identiques : la série est bien "centrée".

  Soit la nouvelle série : 

  Ici la moyenne (10,6) est plus importante que la médiane (10) : la série est plus "étalée à droite".

d) Variance et écart type

<u>Définitions :</u> - La **variance** V d'une série statistique de moyenne $\bar{x}$ dont les valeurs du caractère sont $x_1, x_2, x_3, \ldots, x_k$ et les effectifs correspondants sont $n_1, n_2, n_3, \ldots, n_k$ est égale à :

$$V = \frac{n_1 \times (x_1 - \bar{x})^2 + n_2 \times (x_2 - \bar{x})^2 + \cdots + n_k \times (x_k - \bar{x})^2}{n_1 + n_2 + \cdots + n_k}.$$

- L'**écart-type** σ d'une série statistique de variance V est égal à : $\sigma = \sqrt{V}$.

Ainsi en reprenant l'exemple des tailles de la 2nde5, la variance est égale à :

$$V = \frac{2\times(152-164{,}8)^2 + 4\times(157-164{,}8)^2 + 7\times(162-164{,}8)^2 + 8\times(167-164{,}8)^2 + 3\times(172-164{,}8)^2 + 3\times(177-164{,}8)^2}{27}$$

$\approx 46{,}914$

$\sigma \approx \sqrt{46{,}914} \approx 6{,}85$

L'écart-type possède la même unité que les valeurs de la série.
Ainsi pour la série étudiée, l'écart-type est environ égal à 6,85 cm.

Remarque :
L'écart-type exprime la dispersion des valeurs d'une série statistique autour de sa moyenne. Les valeurs extrêmes influencent l'écart-type.

Exercice 9 : On a relevé, pour les élèves d'une classe de 1ère , la distance du lycée à leurs domiciles et on a obtenu les résultats suivants (arrondis au km) :

Distance	0	1	2	3	4	5
Effectif	5	6	5	3	4	4

1) Calculer la moyenne et l'écart-type de cette série.
2) Quel est le mode de cette série ?
3) Un élève arrive en cours d'année, il habite à 250 km du lycée ; calculer les nouvelles valeurs des paramètres précédents
Conseil : le mode est le caractère ayant le plus grand effectif

Exercice 10 :
Voici la liste des gains (en francs) et le nombre de gagnants enregistrés pour le mercredi 23 mai 2001 au loto :

Gains	3 175 250	75 380	3 890	210	115	25	12
Gagnants	5	16	1040	10 540	48 220	45 360	820 285

1) On veut estimer le nombre de grilles jouées ; la Française des jeux indique que 55% des mises des joueurs sont redistribuées aux gagnants ; sachant que la mise est de 2F, donner une estimation du nombre de grilles enregistrées.
2) Calculer la moyenne et l'écart-type de cette série.
3) Déterminer sa médiane et ses deux quartiles.