Éléments de Statistiques

I/Vocabulaire

Une étude statistique porte sur un ensemble (de personnes, d'animaux, d'objets, …) appelé **population.**

Chaque élément de la population est un **individu.**

L'aspect étudié est nommé **caractère** ou **variable.**

Les résultats obtenus après observation donnent une **série statistique**.

Il existe des séries à une ou plusieurs variables.

Lorsque les variables prennent des valeurs numériques (exemple : notes, tailles, âges, …), les variables sont dites **quantitatives** (si la variable prend n'importe quelle valeur dans un intervalle donné, la variable est dite **continue** , si elle prend des valeurs isolées, la variable est dite **discrète**).

Dans le cas contraire, les variables sont dites **qualitatives** (nationalité, couleurs, …) ; les différentes possibilités du caractère sont appelées **modalités** (la commune de résidence pour des élèves de première fréquentant un lycée …).

Caractéristiques de position

Le mode

Soit (x_i, n_i) une série statistique, on appelle **mode** de cette série toute valeur de la variable correspondant à l'effectif maximal.

La moyenne arithmétique

Soit (x_i, n_i) une série statistique, on appelle **moyenne arithmétique** le quotient de la somme, pondérée par les effectifs, par l'effectif total $\bar{x} = \dfrac{\sum n_i x_i}{\sum n_i}$

La médiane

Soit (x_i) une série statistique **triée dans l'ordre croissant**, on appelle **médiane la plus petite valeur de la variable** telle qu'au moins 50% des valeurs de la série lui soient inférieures ou égales.

Ainsi la série statistique suivante : 2 – 2 – 3 – 3 – 3 – 3 – **4** – 4 – 5 – 6 – 7 – 8 – 8 – 8

- Est constituée de 14 valeurs : son effectif total est **14**.
- 50% de l'effectif correspondent à **7** individus.
- La 7ème valeur est égale à **4**. La médiane de la série est donc égale à **4**

Exemple : Considérons la série statistique suivante :

x_i	1	3	4	5	8	9
n_i	5	10	30	20	7	3

L'effectif maximal est 30, le mode de la série est donc 4.

La moyenne de la série précédente est :

$$\bar{x} = \frac{5\times1+10\times3+30\times4+20\times5+7\times8+3\times9}{5+10+30+20+7+3} = \frac{338}{75} \approx 4,5$$

On détermine la médiane en utilisant les effectifs cumulés de la manière suivante :

x_i	n_i	n_i croissants
1	5	5
3	10	15
4	30	45
5	20	65
8	7	72
9	3	75

- L'effectif total est 75

- 50% de l'effectif correspond à 37,5 individus – donc au moins 50% signifie 38 individus.
- La 38ème valeur de la série est égale à 4 puisque 15 valeurs sont inférieures ou égales à 3 et 45 sont inférieures ou égales à 4.
- La médiane de cette série est par conséquent égale à 4.

Quelques remarques au sujet de la médiane.

Soit $(x_1, x_2, x_3, ..., x_n)$ la série ordonnée des valeurs de la variable. Ces valeurs ne sont pas nécessairement toutes distinctes…

Donc les échantillons B (valeurs inférieures ou égales à la médiane) et $\bar{A}$ (valeurs supérieures ou égales à la médiane) contiennent tous les deux au moins 50% de la population totale.

Nous pouvons donc affirmer que :

> Au moins 50% des valeurs sont inférieures ou égales à la médiane
> Au moins 50% des valeurs sont supérieures ou égales à la médiane

Moyenne et médiane

Moyenne et médiane ne sont pas liée. Voici quelques exemples parfois extrêmes ou sans intérêt qui vous montrent que tout est possible…

Série	Médiane et moyenne	Commentaires
1 – 1 – 1 – 3 – 3	Me=1	La médiane est la plus petite valeur de la série
1 – 1 – 3 – 3 – 3	Me=3	La médiane est la plus grande valeur de la série
1 - 1 – 2 – 3 – 3	Me=2 $\bar{x} = 2$	La médiane est égale à la moyenne.
1 – 1 – 2 – 6 – 6	Me=2 $\bar{x} = 3,2$	La médiane est inférieure à la moyenne.
1 – 1 – 3 – 4 – 4	Me=3 $\bar{x} = 2,4$	La médiane est supérieure à la moyenne.

La médiane partage la population totale en deux échantillons de même effectif. En partageant la population en 4 on obtient les quartiles d'où les définitions :

> Soit (x_i) une série statistique **triée dans l'ordre croissant** :
> - On appelle **premier quartile et on note Q1 la plus petite valeur de la variable** telle qu'au moins 25% des valeurs lui soient **inférieures** ou égales.
> - On appelle troisième quartile **et on note Q3 la plus petite valeur de la variable** telle qu'au moins 75% des valeurs lui soient **inférieures** ou égales.

Exemple :

x_i	n_i	n_i croissants
1	5	5
3	10	15
4	30	45
5	20	65
8	7	72
9	3	75

En reprenant l'exemple précédent, on obtient :
25% de la population représentent 18,75 donc 19 individus.
75% de la population représentent 56,25 donc 57 individus.
Ainsi : **Q_1= 4 ; Q_3=5**

Caractéristiques de dispersion

Étendue d'une série

> L'étendue d'une série statistique est la différence entre la plus grande et la plus petite des valeurs de la série.

Intervalles interquartiles

> L'intervalle interquartile est la différence entre le troisième et le premier quartile : **Q_3-Q_1**

Fréquence Lorsqu'on a à comparer des séries statistiques d'effectifs différents, on peut s'intéresser aux

fréquences plutôt qu'aux effectifs.
La fréquence associée à une valeur du caractère étudiée est le quotient de son effectif par l'effectif total.

Variance et écart-type

- On appelle variance de la série (x_i, n_i) le réel positif : $V = \dfrac{\sum n_i (x_i - \overline{x})^2}{\sum n_i}$
- On appelle écart-type de la série (x_i, n_i) le réel positif : $\sigma = \sqrt{V}$

La formule de calcul de la variance présentée ci-dessus n'est guère adaptée au calcul. On lui préfère la formule suivante obtenue en développant la relation précédente :

$$V = \frac{\sum n_i x_i^2}{\sum n_i} - \overline{x}^2$$

Cette formule signifie que la variance est égale à « la moyenne des carrés » moins le « carré de la moyenne »/

Exemple :

x_i	n_i	$n_i\, x_i$	$n_i\, x_i^2$
1	5	5	8
3	10	30	90
4	30	120	480
5	20	100	500
8	7	56	448
9	3	27	243
Totaux	75	338	1 769

$\overline{x} = \dfrac{338}{75} \approx 4{,}5$

$V = \dfrac{1769}{75} - \left(\dfrac{338}{75}\right)^2 \approx 3{,}28$

$\sigma = \sqrt{V} \approx 1{,}81$

En moyenne, les données s'écartent de 1,81 de la valeur moyenne 4,5

II/Série statistique à deux variables

1.Exemple

Les cadres d'une entreprise ont reçu des primes différentes selon leur ancienneté.
Sept d'entre eux comparent le montant de leur prime. Leurs observations sont reportées dans le tableau ci-dessous, où l'ancienneté est exprimée en années et la prime en dinars.

Cadre	n°1	n°2	n°3	n°4	n°5	n°6	n°7
Ancienneté (x_i)	2	8	11	17	20	22	26
Prime (y_i)	270	390	460	590	650	710	830

Quand on étudie deux caractères statistiques sur une même population, on obtient une série statistique double.
Si les valeurs prises par le premier caractère sont $x_1, x_2, \ldots, x_n$ et celles prises par le second caractère sont $y_1, y_2, \ldots, y_n$, alors les valeurs prises par cette série sont les couples ($x_1 ; y_1$), ($x_2 ; y_2$), …, ($x_n ; y_n$).

2.Nuage de points

L'ensemble des points $\{M_1, M_2, \ldots, M_n\}$ où M_1 a pour coordonnées ($x_1 ; y_1$), M_2 ($x_2 ; y_2$), …, M_n ($x_n ; y_n$) dans un repère du plan, est le nuage de points de la série.

3.Point moyen

Le point moyen du nuage $\{M_1, M_2, \ldots, M_n\}$ est le point G de coordonnées ($\overline{x}$; $\overline{y}$), où $\overline{x}$ est la moyenne de la série (x_i) et $\overline{y}$ est la moyenne de la série (y_i).

Dans l'exemple, $\overline{x}$ = …………..
$\overline{y}$ = …………..

Le point moyen du nuage est le point G (……………… ; ………………)

4.Ajustement affine (Droite de MAYER)

Quand les points d'un nuage sont sensiblement alignés, on peut construire une droite passant au plus près de ces points ; on dit que cette droite réalise un ajustement affine du nuage de points.

On peut tracer cette droite au jugé, ou alors utiliser des méthodes particulières.(Exemple : droite de Mayer)
On choisit en général une droite passant par le point moyen du nuage.

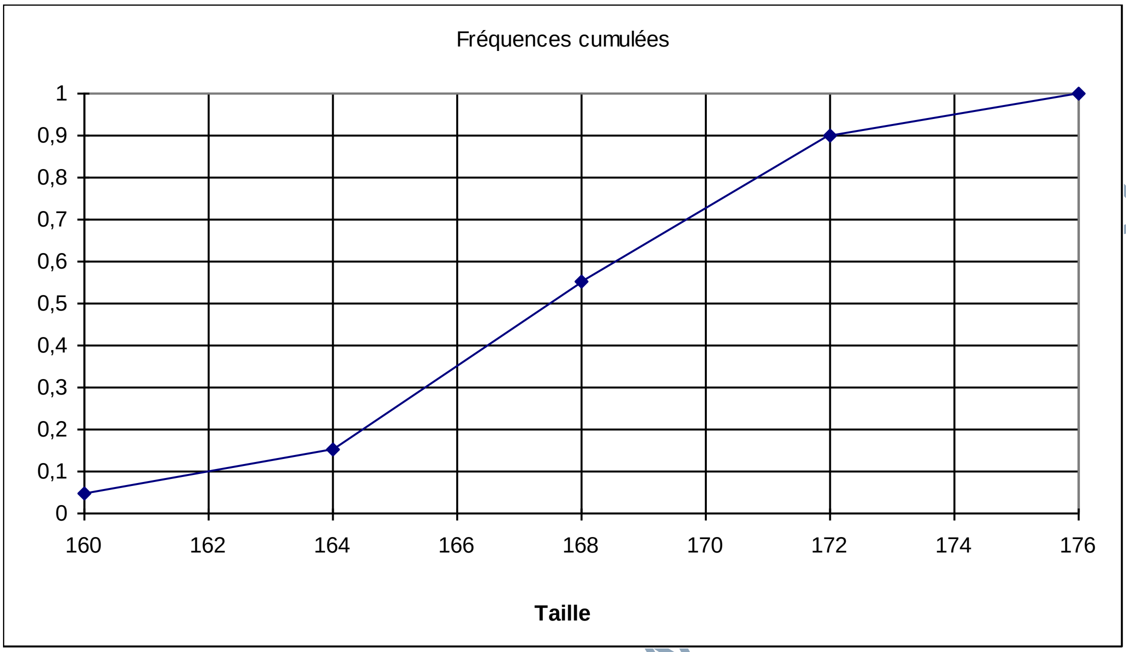
Exercice 1 :

Le tableau suivant donne la répartition de la taille x de 500 élèves d'un établissement scolaire .

x_i : Tailles en cm	**160**	**164**	**168**	**172**	**176**
n_i : Effectif	**25**	**50**	**200**	**175**	**50**

1- Calculer la taille moyenne $\bar{x}$ des élèves de l'établissement et déterminer le mode M_o des tailles .

2- En vous servant du polygône des fréquences cumulées croissantes indiqué ci-dessous , déterminer le premier quartile , la médiane de et le troisième quartile de cette série (x_i , n_i) .



Exercice 2 :

On a relevé pour chacun des dix élèves d'un groupe ses moyennes en math au $2^{ème}$ et au $3^{ème}$ trimestre.

Elève	1	2	3	4	5	6	7	8	9	10
X :moyenne au $2^{ème}$ Trimestre	3,5	9,5	1	7,5	5	8,5	15,5	13,5	13	17,5
Y : :moyenne au $3^{ème}$ Trimestre	7,5	9,5	6,5	12	5,5	13,5	14	15	15,5	19

A/ 1- Calculer la moyenne $\bar{X}$ et l'écart-type σ_X de cette série.

2- Calculer la moyenne $\bar{Y}$et l'écart-type σ_Y de cette série.

3- Laquelle des deux variables X et Y est plus dispersée ?

B/ 1- Représenter dans un repère orthogonal le nuage de points M_i (x_i , y_i) de cette série .

2- On partage l'ensemble des dix points du nuage en deux parties A et B :

La partie A correspondant aux élèves de 1 à 5 et la partie B aux élèves de 6 à 10 .

a) Déterminer les coordonnées du point moyen G_1 de la partie A puis les coordonnées du point moyen G_2 de la partie B .

b) Tracer la droite (G_1G_2) .

Exercice 1 :

1- $\bar{x} = \frac{\sum_1^5 n_i x_i}{500} =$ 169,4 M_o = 168 .

2-

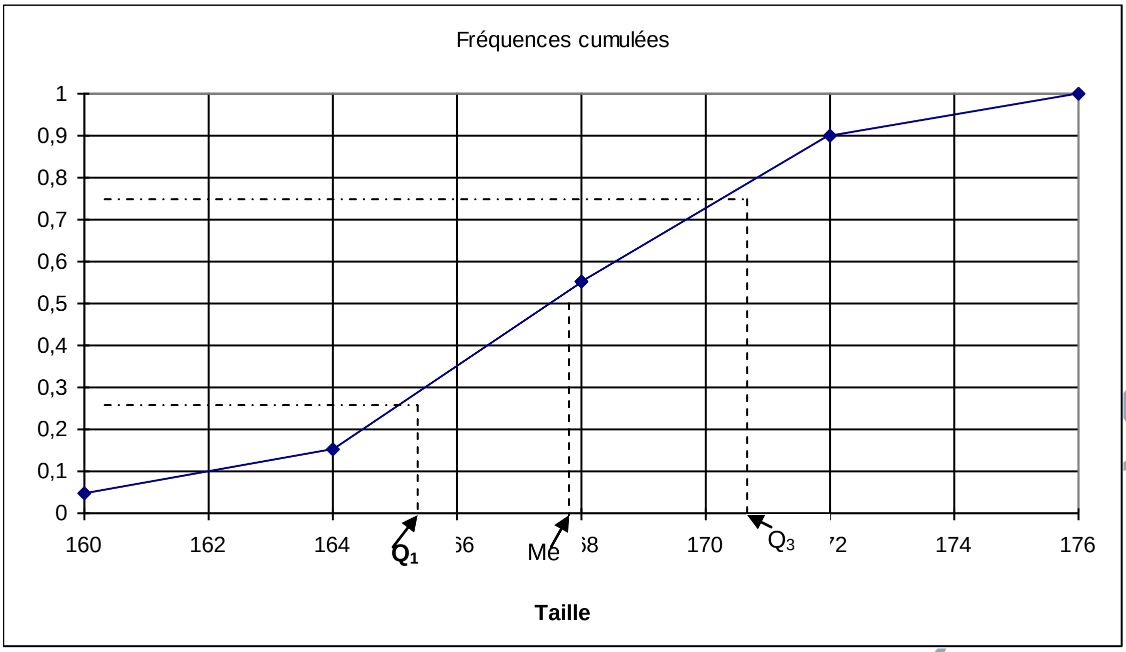


Figure 1

$Q_1 = 165$	;	$Me = 167,5$	;	$Q_3 = 170,33$

Exercice 2 :

A/

1- $\overline{X} = 9,45$; $\sigma_X = 5,115$.

2- $\overline{Y} = 11,8$; $\sigma_Y = 4,184$.

3- $\sigma_X > \sigma_Y$ donc X est plus dispersée que Y .

B/

1- Le nuage de points est formé par les points M_i (x_i , y_i) de cette série ($1 \leq i \leq 10$) .

2- La partie A est formée par les points M_i (x_i , y_i) de cette série ($1 \leq i \leq 5$) et la partie B est formée par les points M_i (x_i , y_i) de cette série ($6 \leq i \leq 10$) .

a) G_1 (5,3 ; 8,2) , G_2 (13,6 ; 15,4) .

b) Voir figure 2 .

3- Montrer que la droite (G_1G_2) a pour équation $y = (0,87)x + (3,60)$.

Car $\dfrac{15,4 - 8,2}{13,6 - 5,3} \approx 0,87 = a$ et $b = 15,4 - (0,87)\,13,6 \approx 3,60$.

4-Pour $x = 10$, on a $y = (0,87)10 + (3,60) = 12,3$ donc une estimation de sa moyenne en maths ade troisième trimestre est égale à 12,25 .