BIOSTATISTIQUE

(Méthode statistique en Biologie et en Médecine)

Première partie (statistique descriptive)

1. *Introduction*

La statistique est une méthode scientifique qui consiste à réunir des données chiffrées sur des ensembles nombreux, puis à analyser, à commenter et à critiquer ces données.

La statistique a envahi tous les domaines de la vie scientifique .Le Biologiste dans son laboratoire, comme l'ingénieur agronome dans ses champs relève les résultats de leurs expériences, les sociologues dans des groupes ethniques notent leurs observations....

***Remarque** : il ne faut pas confondre la statistique qui est la science qui vient d'être définie et une statistique qui est un ensemble de données chiffrées sur un sujet précis.*

*Comme toute science, la statistique fait appel à un **vocabulaire** spécialisé.*

- *Les ensembles sont appelés **populations.** Comme un ensemble, une population statistique doit être clairement définie.*

  *Une population peut être finie ou infinie. Elle est finie si elle comporte un nombre déterminé d'individus ou d'objets. **Exemple** : population d'une ville à un instant donné.*

  *Elle est infinie si elle comporte un nombre fini d'individus ou d'objets. **Exemple** : l'ensemble des résultats (face et pile) lors de parties successives de pile ou face avec une même pièce de monnaie.*
- *Au lieu d'examiner l'ensemble qu'on appelle population, on examine un nombre restreint qu'on appelle **échantillon**.*

- *Les éléments de la population sont appelés **individus** ou **unités statistiques.***
- *Chaque individu peut être étudié relativement à un ou plusieurs **caractères**.*
- *Un caractère permet de déterminer une partition de la population selon ses divers types de **modalités**.*
- *Lorsque les modalités du caractère sont des nombres, le caractère est dit **quantitatif** ; on lui donne souvent alors le nom de **variable statistique.***
- *Une variable statistique peut être **discrète** si elle ne prend que des valeurs isolées, ou **continue** si elle peut prendre n'importe quelle valeur intermédiaire entre deux valeurs données.*
- *Lorsque les modalités du caractère ne sont pas mesurables, le caractère et dit **qualitatif,** les modalités d'un caractère qualitatif peuvent faire l'objet d'une nomenclature ou énumération.*
- *Suivant le nombre de variables étudiées simultanément, on peut parler de **la statistique descriptive** à **une dimension**, à **deux dimension**, à **trois dimension** et **multidimensionnelle**.*

2. Statistique descriptive à une dimension

Le but de simplification de la statistique descriptive peut être atteint en condensant les données sous trois formes distinctes :

***2.1. Les tableaux statistiques** : qui permettent de présenter les données sous la forme numérique de distribution de **fréquences**.*

***2.2. Les diagrammes** : qui permettent de présenter **graphiquement** ces distributions sous forme de **diagramme en bâtons**, **histogramme** ou **diagramme circulaire**.*

2.3. Les paramètres de réduction** : les données peuvent également être condensées sous la forme de quelques **paramètres :

a. ***Les paramètres de tendance centrale*** *sont des mesures qui localisent le centre d'une distribution. Les plus utilisés sont :* ***la moyenne arithmétique, la médiane et le mode.***

b. ***Les paramètres de position*** *qui occupent un certain rang dans la série des valeurs d'une série statistique, valeurs classées selon un ordre croissant. Ces valeurs sont aussi appelées* ***quantiles****. Ceux-ci se subdivisent en* ***quartiles****,* ***déciles*** *et* ***percentiles****.*

c. ***Les paramètres de dispersion*** *: les caractéristiques de tendance centrale et de position sont insuffisantes pour caractériser complètement une série statistique. Nous avons donc besoins de paramètres de dispersion qui vont permettre d'estimer dans quelle mesure les observations s'écartent de la tendance centrale. Les plus connus de ces paramètres sont* ***l'écart type*** *et la* ***variance****.*

2.3.1. La moyenne arithmétique : $(\bar{x})$ la *moyenne arithmétique d'une série de valeurs d'une variable statistique est égale à la somme de ces valeurs divisée par leur nombre.*

La formule générale est, pour n observations $x_1, x_2, \ldots\ldots x_n$

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Comme chaque valeur x_i *doit être prise en considération autant de fois qu'elle a été observée, cette expression devienne, dans le cas de distribution de fréquences :*

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} n_i \, x_i$$

2.3.2. La moyenne géométrique ;($\bar{x}_g$)

La moyenne géométrique $\bar{x}_g$ *d'une série statistique composée de n valeurs positives* $x_1, x_2, \ldots .. x_n \ldots\ldots x_p$ *et par la définition, la racine* $n^{ème}$ *du produit de ces n valeurs .*

$$\bar{x}_g = \sqrt[n]{x_1 \,.\, x_2 \ \ldots\ldots\ldots x_n}$$

Pour une distribution de fréquences, la moyenne géométrique peut être définie comme suit :

$$\bar{x}g = \sqrt[n]{x_1^{n_1} . x_2^{n_2} \ldots\ldots\ldots . x_p^{n_p}}$$

2.3.3. La moyenne harmonique ($\bar{x}_h$) : *la moyenne harmonique est égale à l'inverse de la moyenne arithmétique des inverses.*

$$\bar{x}_h = \frac{1}{\frac{1}{n}(\frac{1}{x_1}+\frac{1}{x_2}+\cdots\ldots+\frac{1}{x_n})} \qquad \bar{x}_h = \frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}}$$

Par extension, pour les distributions de fréquences, la moyenne harmonique peut être définie comme suit :

$$\bar{x}_h = \frac{n}{\sum_{i=1}^{p}\frac{n_i}{x_i}}$$

2.3.4. La moyenne quadratique : ($\bar{x}_q$)

La moyenne quadratique est la racine carrée de la moyenne arithmétique des carrées des observations ;

$$\bar{x}_q = \sqrt{1/n \sum_{i=1}^{n} x_i^2}$$

$$\bar{x}_q = \sqrt{1/n \sum_{i=1}^{p} n_i x_i^2}$$ *(Dans le cas d'une distribution de fréquences)*

Remarque :

La moyenne arithmétique est certainement la plus utilisée.

2.3.5. *La médiane (med) :* *la médiane est la valeur de la variable qui correspond à la fréquence cumulée* ***50%*** *ou à l'effectif cumulé n/2.*

$$Med = x_{(n+1)\ /2}$$ *(si n est impair)*

$$Med = 1/2[x_{n\ /2} + (x_{(\left(\frac{n}{2}\right)+1)}]$$ *(si n est pair)*

2.3.6. *Le mode ou la valeur modale(Mo)* *: est la valeur que la variable statistique prend le plus fréquemment*

2.3.7. *Quartiles :* *la médiane partage la série en deus groupes de même effectif. les quartiles (Q1, Q2, Q3) partage la série en quatre groupes de même effectif. Le second quartile se confond avec la médiane. Q2=med*

2.3.8. *La variance S^2 :* *est la moyenne arithmétique des carrées des écarts des valeurs observées par rapport à la moyenne.*

$$V = S^2 = 1/n \sum_{i=1}^{n} (x_i - \bar{x})^2$$

L'écart-type S *: L'écart-type est la racine carrée de la variance.* $\boldsymbol{S = \sqrt{S^2}}$

La variance ou L'écart-type nous renseigne sur la dispersion des valeurs observées autour de la moyenne ; plus La variance ou l L'écart-type est petit plus les valeurs observées sont proche de la moyenne.

3 .Statistique descriptive à deux dimensions :

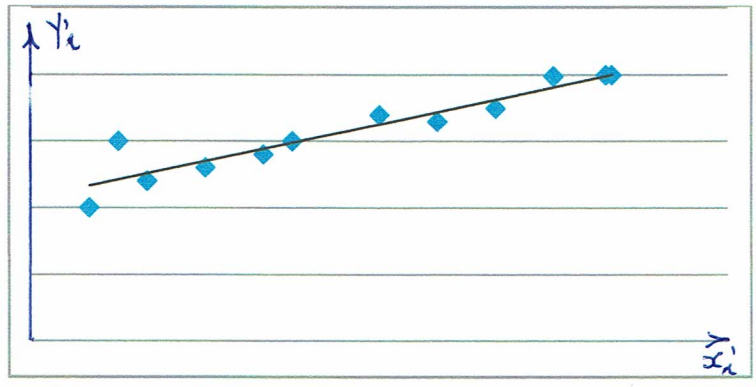
Supposons que l'étude statistique d'une population porte simultanément sur deux caractères quantitatifs. Si nous considérons par exemple le poids (x_i) *et la taille* (y_i) *ou une caractéristique morphologique quelconque d'un individu et son âge.*

Le problème se pose de déterminer s'il existe une liaison ou une corrélation entre les grandeurs xi et yi de ces caractères pour un même individu.

3.1. Nuage de point : *le nuage de point représente les couples de valeurs* x_i *et* y_i*. La liaison statistique est d'autant plus serrée que le nuage de points est plus mince.*

3.2. Ajustement linéaire : Droite de régression

L'ajustement linéaire consiste à remplacer le nuage de points par une droite de régression.une telle approximation est d'autant meilleur que le nuage de points s'apparente à celui de la figure suivante :



3.3. ***Équation de la droite de régression :*** *la droite de régression de y par rapport à x noté Dy(x)=ax+b passe par le point moyen du nuage* M_0*, appelé centre de gravité du nuage et ayant pour coordonnées* ***(x,y)***Tapez une équation ici.

$$a = \left(\sum x_i y_i - \bar{y}\sum x_i\right)/\left(\sum x_i^2 - \bar{x}\sum x_i\right)$$

$$b = \bar{y} - a\bar{x}$$

3.4. Coefficient de corrélation :

$$r = \left(\sum x_i y_i - \bar{y}\sum x_i\right)/\left(\sqrt{\sum x_i^2 - \bar{x}\sum x_i}\right)\left(\sqrt{\sum y_i^2 - \bar{y}\sum y_i}\right)$$

Le coefficient de corrélation est toujours compris entre -1 et+1

$/r/\in]0-0.3]$ *Il ya une faible corrélation*

$/r/\in]0.3-0.6]$ *Il ya une corrélation moyenne*

$/r/\in]0.6-1]$ *Il ya une forte corrélation*

Exemple :

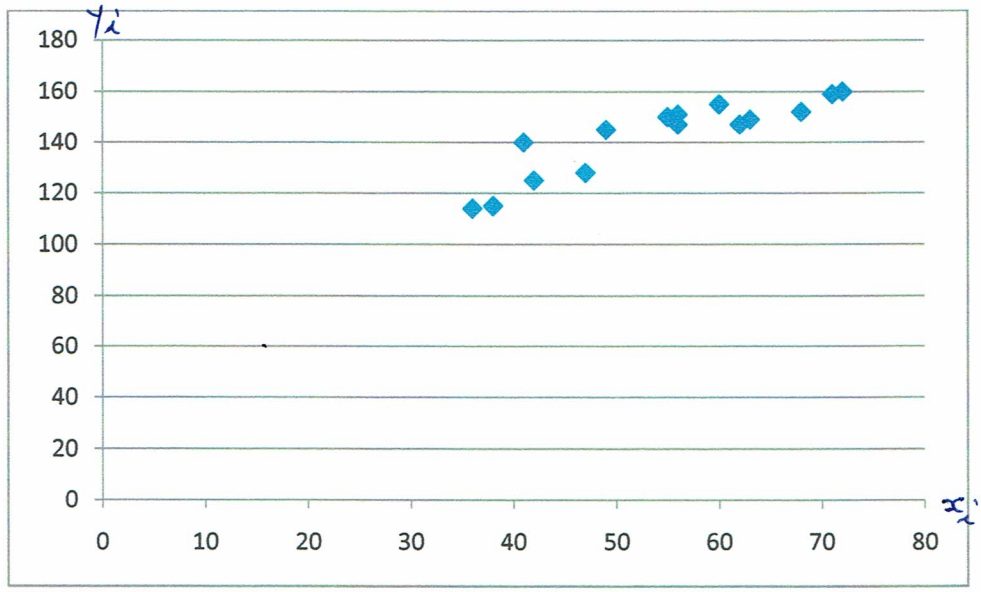
Afin d'étudier la relation qui pourrait exister entre l'âge et la pression sanguine, un médecin mesure sur 15 femme d'âge (x_i) *différents la pression sanguine systolique* (y_i).

x_i (ans)	56	42	72	36	63	47	55	49	38	41	68	60	56	62	71
y_i (mmHg)	147	125	160	114	149	128	150	145	115	140	152	155	151	147	159

- 1 .Représenter le nuage de points correspondant à la série (x_i, y_i) sur un repère orthogonal.

- 2. Déterminer le mode, la médiane et les quartiles de chaque distribution.
- 3. Calculer la moyenne, la variance et l'écart-type de chaque distribution.
- 4. Déterminer les coordonnées de point moyenne G de ce nuage. Le placer sur le graphique.
- 5. Déterminer une équation de la droite (D) de y en x.
- 6. Représenter (D) sur le graphique.
- *7. Calculer le coefficient de corrélation linéaire r.*

1. Nuage de points correspondant à la série (x_i, y_i)



2. Mode, médiane et quartiles de chaque distribution.

	Mo	*Méd*	*Q1*	*Q2*	*Q3*
x_i (ans)	*56*	*56*	*42*	*56*	*63*
y_i (mmHg	*147*	*147*	*128*	*147*	*152*

3. Moyenne, variance et écart-type de chaque distribution.

	Moyenne	*Variance*	*Ecart-type*
x_i (ans)	*54.4*	*130.91*	*11.44*
y_i (mmHg	*142.47*	*208.92*	*14.45*

4. *coordonnées de point moyenne G de ce nuage* $(\overline{x_i}, \overline{y_i})$

$$\overline{x} = 54.4$$

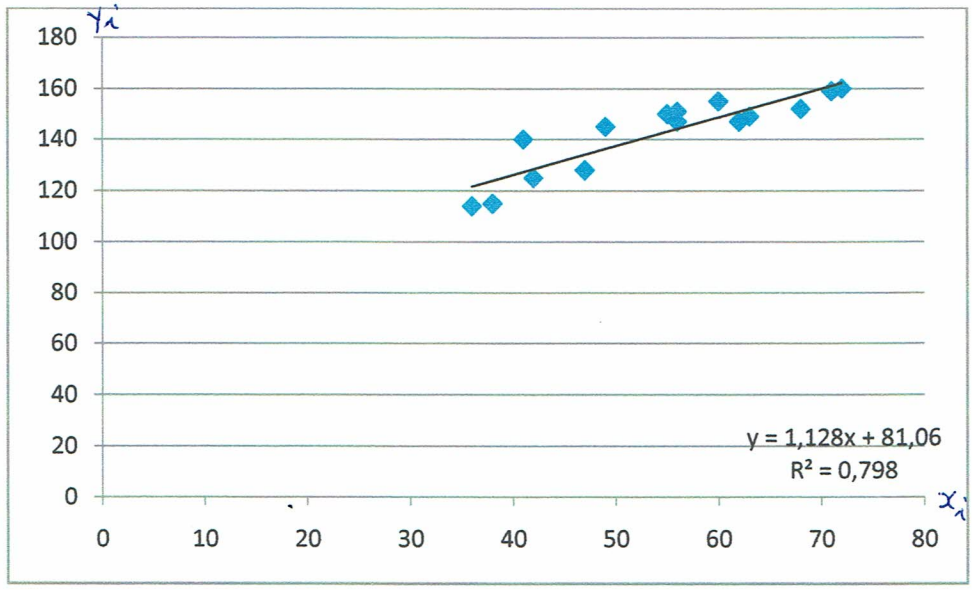
$$\overline{y} = 142.47$$

5. *équation de la droite (D) de y en x*

y= ax+b

y=1.13x+81.06

6. *Représentation de (D) sur le graphique*



7. *coefficient de corrélation linéaire r.*

r =0.89

Exercice 1

Les grenouilles hébergent divers parasites, en particulier des vers trématodes.

On prélève au hasard des grenouilles dans un étang et on compte les trématodes que chacune d'elle héberge. On obtenu les résultats suivants :

Nombre de trématode par grenouille	0	1	2	3	4	5	6
Nombre de grenouilles correspondantes	11	22	45	40	19	11	2

1. Quel sont la population étudiée ? le caractère ? la nature du caractère ? les modalités du caractère ?
2. Représenter graphiquement cette distribution de fréquence.
3. Calculer la moyenne, la variance et l'écart-type de cette distribution.

Exercice 2

Dans le cadre de l'étude de la population de gélinottes huppées (Bonasa umbellus), les valeurs de la longueur de la rectrice principale peuvent être réparties de la façon suivante :

158 152 171 163 140 157 162 171 158 164 163 159 153
160 149 158 152 165 156 162 150 154 155 162 155 164
164 157 159 158 159 153 163 158 174 162 156 151 160
158 162 166 162 164 158 153 165 158 150 160

1. quelle est l'étendue de cette série statistique ?
2. donner la répartition en classe de cette série.
3. établir le tableau de distribution des fréquences de cette série.

Exercice 3

On donne la répartition d'un groupe d'enfants par taille :

Taille (cm)	effectif
80 à moins de 90cm	3
90 " 95	15
95 " 100	22
100 " 105	18
105 " 110	12
110 " 120	5

1. Quel sont la population étudiée ? le caractère ? la nature du caractère ? les modalités du caractère ?
2. Tracer l'histogramme de cette répartition.
3. Calculer la moyenne, la variance et l'écart-type de la distribution des enfants par taille.
4. Calculer les effectifs cumulés et les représenter graphiquement.
5. Déterminer la médiane de cette série d'observation.
6. Quel est le mode de cette distribution ?

Exercice 4

On donne la répartition des femmes âgées de 50 à 54 ans d'après le nombre de leurs enfants née vivants :

Nombre d'enfants	Fréquences(%)
0	19
1	25
2	23
3	14
4 et plus	19
total	100

1. Quel sont la population étudiée ? le caractère ? la nature du caractère ? les modalités du caractère ?
2. Représenter cette série par le diagramme qui convient.
3. Calculer les fréquences cumulées et tracer la courbe cumulative.
4. Déterminer le mode, la médiane et les quartiles de distribution.
5. Calculer la moyenne et l'écart-type de la distribution.

Exercice 5

L'analyse du sang de 100 individus a donné les résultats suivants :

Groupe sanguin	effectifs
O	40
A	43
B	12
AB	5

1. Quelle est la nature du caractère étudié ?
2. Déterminer les fréquences et les pourcentages.
3. Représenter graphiquement la série.

Exercice 6

Le taux de glucose sanguin (glycémie) déterminé chez 32 sujets est donné ci-dessous en (g/l)

0.85 0.87 0.90 0.93 0.94 0.94 0.95 0.97 0.97 0.98 0.98 0.98 0.99 1.00 1.01 1.03 1.03
1.03 1.04 1.06 1.07 1.08 1.08 1.10 1.10 1.11 1.13 1.14 1.15 1.17 1.19 1.20

1. quelle est l'étendue de cette série statistique ?
2. donner la répartition en classe de cette série.
3. établir le tableau de distribution des fréquences de cette série.

Exercice 7

Les valeurs de la longueur totale de 90 individus de l'espèce de crevette « Aristeus antennatus » exprimée en (mm) et rangées dans l'ordre croissant

102 109 114 117 119 120 123 123 123 125 125 126 126 128 129 131 131 131 131 132 132
133 133 134 135 135 135 135 136 137 137 137 138 139 139 139 140 140 140 140 140 140
142 142 142 143 144 144 144 145 146 146 146 147 148 148 148 149 149 149 150 151 151
151 151 152 152 152 152 153 154 155 157 157 157 157 158 159 165 166 167 158 168 168
168 170 170 170 173 176

1. Déterminer le mode et la médiane de cette série.
2. Regrouper ces valeurs en classes. Justifiez en quelques mots l'intervalle de classe choisi.
3. Tracer l'histogramme correspondant.
4. Calculer la moyenne, la variance et l'écart-type de cette distribution.

Exercice 8

L'étude du poids de 50 personnes a donné les résultats suivants (en kg) présentés sous la forme d'une série statistique ordonnée.

37 43 47 50 52 54 55 56 58 58 61 62 63 63 64 65 66 66 67 68 68 69 69 70 71 72 72
72 73 73 74 74 75 76 76 77 79 79 80 82 82 84 86 87 88 90 92 93 91 98

1. Regrouper ces classes de poids.
2. Construire l'histogramme correspondant.
3. Calculer la moyenne, la variance et l'écart-type de cette distribution.

Exercice 9

Dans le cadre d'une étude du laboratoire de biochimie, on a dosé la quantité de créatinine en mg/100 cm^3 d'urine chez 80 hommes normaux. Les résultats ont été représentés dans le tableau suivant :

Quantité de créatinine (mg/100 cm^3)	Nombre de personnes
[2.5 ; 3.5 [	2
[3.5 ; 4.5 [	11
[4.5 ; 5.5 [	20
[5.5 ; 6.5 [	30
[6.5 ; 7.5 [	14
[7.5 ; 8.5 [	3

1. Comment appelle-t-on ce type de tableau ?
2. Quelle est la variable étudiée et quelle est sa nature ?
3. Calculer la moyenne et l'écart-type.

Exercice 10

On veut étudier la relation entre la dose d'une substance toxique et le temps de survie, les résultats observés sur un lot de 20 rats figurent sur le tableau suivant :

Dose x_i (g/l)	2	4	6	8
Temps de survie y_i (minutes)	360	180	40	60
	420	240	90	45
	300	240	90	25
	480	120	60	10
	540	80	55	15

1. Ajuster graphiquement cette distribution à deux variables.
2. Calculer le temps de survie lorsque la dose est de 10g/l et x_i lorsque y_i est de 600 minutes.
3. Calculer le coefficient de corrélation linéaire.
4. Ajuster analytiquement (méthode analytique) cette distribution.
5. Recalculer ce qui est demandé à la question 2.

Exercice 11

Le tableau ci-après donne les résultats de 7 déterminations de la distance nécessaire à l'arrêt d'une voiture automobile(y) suivant sa vitesse (x).

Numéro de voiture	Vitesse (km/h)(x)	(distance(m) (y)
1	33	5.30
2	49	14.45
3	65	20.21
4	33	6.50
5	79	38.45
6	49	11.23
7	93	50.42

1. Représenter graphiquement les 7 points (x, y). quelle est la forme de la courbe représentant la moyenne de y en fonction de x ?
2. Substituer à y sa racine carrée $z=\sqrt{y}$.
3. Représenter les 7 points x, z. quelle serait la forme de la courbe représentant z en fonction dex ? vérifier la validité de l'ajustement.
4. Ajuster la droite $z^* = a+bx$ par la méthode des moindres carrés. Utiliser cette équation pour déterminer la distance nécessaire à l'arrêt d'une voiture lancée à 85km/h.

Exercice 12

En recherchant un ajustement linéaire par la méthode des moindres carrés pour des séries statistiques doubles formées de la taille et du poids d'un certain nombre d'individus, on trouvé les résultats suivants :

- Pour 100 enfants, coefficient de corrélation 0.90
- Pour 200 adultes, coefficient de corrélation 0.55
- Pour 200 adultes, coefficient de corrélation 0.40

1. Un ajustement linéaire est-il possible, et pour quels individus ?
2. Ya-t-il une relation- vraie en moyenne- entre la taille et le poids de certains individus observés ?

06 page 158

Exercice 13

On a relevé l'âge x et le poids y de 492 enfants.

	Age(années)					
Poids (kg)	5 - 6	7 - 8	9 - 10	11 - 12	13 - 14	totale
63 - 67					1	1
59 - 63					8	8
55 - 59				2	6	8
51 55				1	18	19
47 - 51				10	30	40
43 - 47			1	17	41	59
39 - 43			3	31	31	65
35 - 39		1	21	26	13	61
31 - 35		6	33	28	10	77
27 - 31	5	17	26	10		58
23 - 27	8	32	14	1		55
19 - 23	16	13		1		30
15 - 19	9	1				10
11 - 15	1					1
total	39	70	98	127	158	492

1°. (Lecture du tableau carré)

a) Combien d'enfants ont 7 ou 8 ans et un poids de 27 à 31kg ?
b) Quelle est la proportion d'enfants pesant moins de 35 kg ?
c) Quelle est la proportion d'enfants de 13 ans ou plus ?

2° .calculer le coefficient de corrélation linéaire entre l'âge et le poids des enfants.

On donne :

$$\sum n_i (y_i - y)^2 = 49\,034$$

$$\sum n_j (x_j - \bar{x})^2 = 3\,232$$

$$\sum n_{ij} (x_j - \bar{x})\ (y_i - \bar{y}) = 10\,360$$

Indiquer rapidement comment cette dernière somme aurait pu être calculée à partir des données du tableau.

3° a)Calculer pour chaque classe d'âge, le poids moyen.

b) Représenter graphiquement les moyennes retrouvées (poids moyen en fonction de l'âge).

4°. Calculer par la méthode des moindres carrés, les coefficients de l'équation de la droite d'ajustement du poids en fonction de l'âge : $\hat{y} = \bar{y} + a(x - \bar{x})$