Faculté des sciences

1ère année Master Informatique (ISII)

Matière : Analyse de données

Année : 2023/2024

TD1

Corrigé type de la série 1 d'exercices

Exercice 1 :

Dans une petite localité, on a relevé une série de nombre de pièces par appartement

Nombre de pièces	1	2	3	4	5	6	7
Nombre d'appartements	48	72	96	64	39	25	3

1) Définir la population, la variable (le caractère) étudiée et quelle est sa nature.
   - La population représente la localité.
   - La variable étudiée est le nombre de pièces, de nature quantitative discrète.

2) Calculer les paramètres de tendance centrale et les paramètres de dispersion, sachant que l'effectif total de l'échantillon est 347.

Les paramètres de tendance centrale : moyenne, médiane et le mode.

Les paramètres de dispersion : variance, écart type, coefficient de variation, étendue, coefficient d'asymétrie, coefficient d'aplatissement, quantiles et écart inter-quantile.

Sachant que : x_i *est le* nombre de pièces et n_i *est* le nombre d'appartements.

- Calculer la moyenne de pièces par appartement :

$$\boldsymbol{moy} = \frac{1102}{347} \approx 3.18$$

- Calculer le mode : 3 (avec un effectif maximal de 96)
- Calculer la médiane : 4
- Calculer l'étendue de la série : E= 7-1= 6
- Calculer la variance : $\boldsymbol{var} = \frac{746}{347} \approx 2.15$
- Calculer l'écart type : $\boldsymbol{\sigma} = \sqrt{var} \approx 1.47$
- Calculer le coefficient de variation : $\boldsymbol{CV} = \frac{\sigma}{|moy|} = \frac{1.47}{|3.18|} \approx 0.46$
- Calculer le coefficient d'asymétrie :

$$\mu_3 = \frac{383}{347} \approx 1.10\ ,\ \sigma^3 = 1.47^3 = 3.18$$

$$\boldsymbol{\gamma_1} = \frac{\mu_3}{\sigma^3} = \frac{1.10}{3.18} \approx 0.35$$

- Calculer le coefficient d'aplatissement :

$$\mu_4 = \frac{3901}{347} \approx 11.24\ ,\ \sigma^4 = 1.47^4 = 4.67$$

$$\boldsymbol{\gamma_2} = \frac{\mu_4}{\sigma^4} = \frac{11.24}{4.67} \approx 2.41$$

x_i	n_i	$n_i * x_i$	$(x_i - \overline{X})$	$n_i(x_i - \overline{X})^2$	$n_i(x_i - \overline{X})^3$	$n_i(x_i - \overline{X})^4$	**Pourcentage (%)**	**Pourcentage cumulé**
1	48	48	-2,18	228	- 497	1084	13.83	13.83
2	72	144	-1,18	100	- 118	140	20.75	34 .58
3	96	288	-0 ,18	3	- 0.56	0.1	27.67	62.25
4	64	256	0,82	43	35	29	18.44	80.69
5	39	195	1,82	129	235	428	11 .24	91.93
6	25	150	2,82	199	560.5	1581	7.20	99.13
7	3	21	3,82	44	167	639	0.87	100
Total	347	1102		746	382.94 ≈ 383	3901	100	

- Calculer les quantiles :

$Q_1 = 2\ (25\%), Q_2 = 3\ (50\%), Q_3 = 4\ (75\%)$

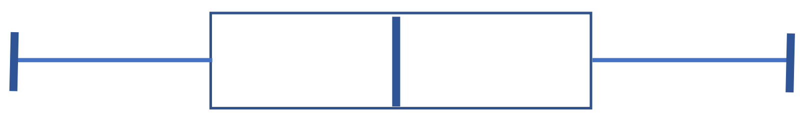
$D_1 = 1\ , D_2 = 2\ , D_3 = 2, D_4 = 3, D_5 = 3, D_6 = 3, D_7 = 4, D_8 = 4, D_9 = 5$

$C_1, \ldots, C_{13} = 1,\ C_{14} \ldots, C_{34} = 2,\ C_{35} \ldots, C_{62} = 3,\ C_{63} \ldots., C_{80} = 4,\ C_{81} \ldots, C_{91} = 5,$
$C_{92}, \ldots, C_{99} = 6$

- Calculer les écarts interquartiles :

$E = Q_3 - Q_1$=2, $E = D_9 - D_1$=4, $E = C_{99} - C_1$=5

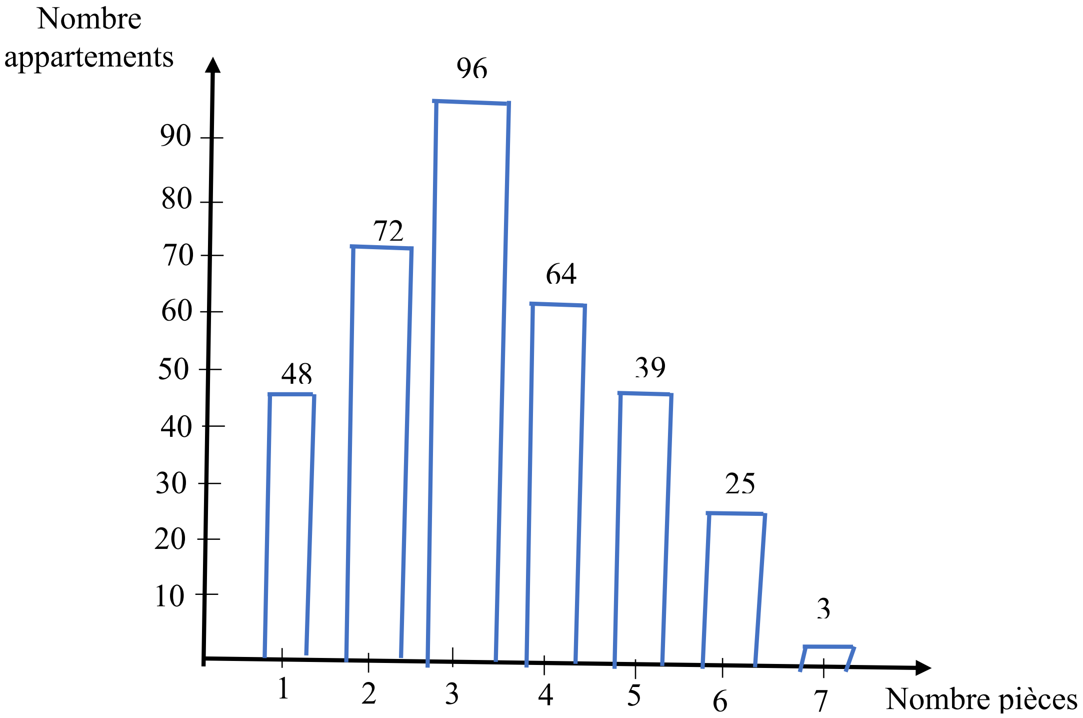
3) Représenter graphiquement dans la boite à moustache correspondante la médiane (Me) et les quartiles (Q_1 et Q_3).



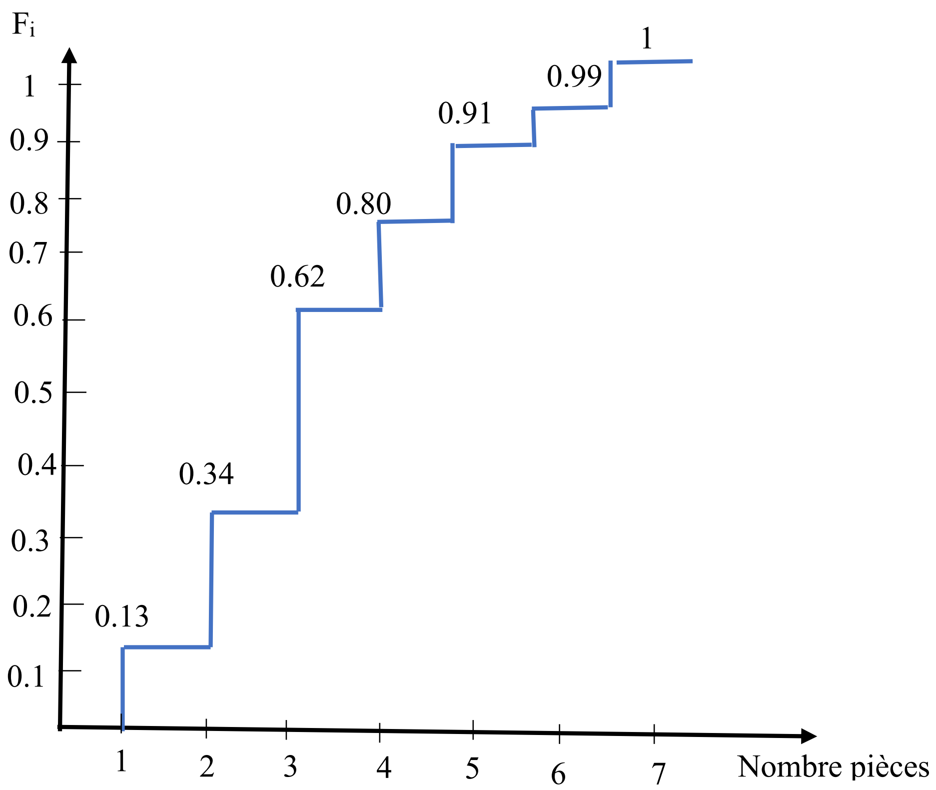
Min=1 Q_1=2 Q_2=3 Q_3=3 Max=7

4) Représenter graphiquement cette série par les graphiques correspondants.

✓ Histogramme

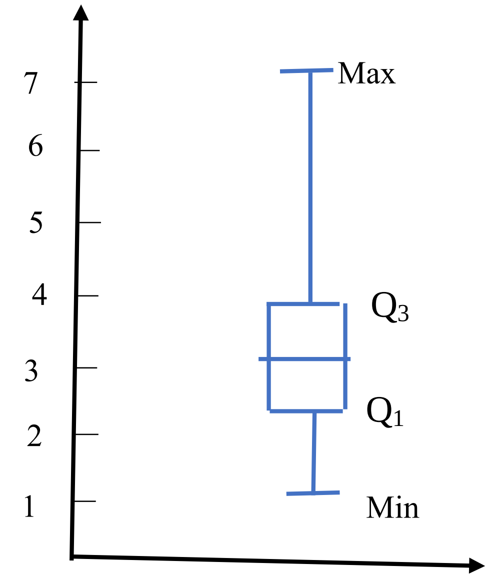


✓ Courbe Cumulative



- ✓ Boite à moustache

Nombre pièces



- ✓ Histogramme Quantile-Quantile

-Distribution normale $(0, 1^2)$

D_1	D_2	D_3	D_4	D_5	D_6	D_7
1	2	2	3	3	3	4

- Entrées : le pourcentage de quantiles et l'endroit du décile (gauche, bilatérale, à droite)

$\underbrace{D_1,\ D_2,\ D_3,\ D_4}_{\text{à gauche}}$, $\quad D_5$, $\quad \underbrace{D_6,\ D_7}_{\text{à droite}}$

bilatérale

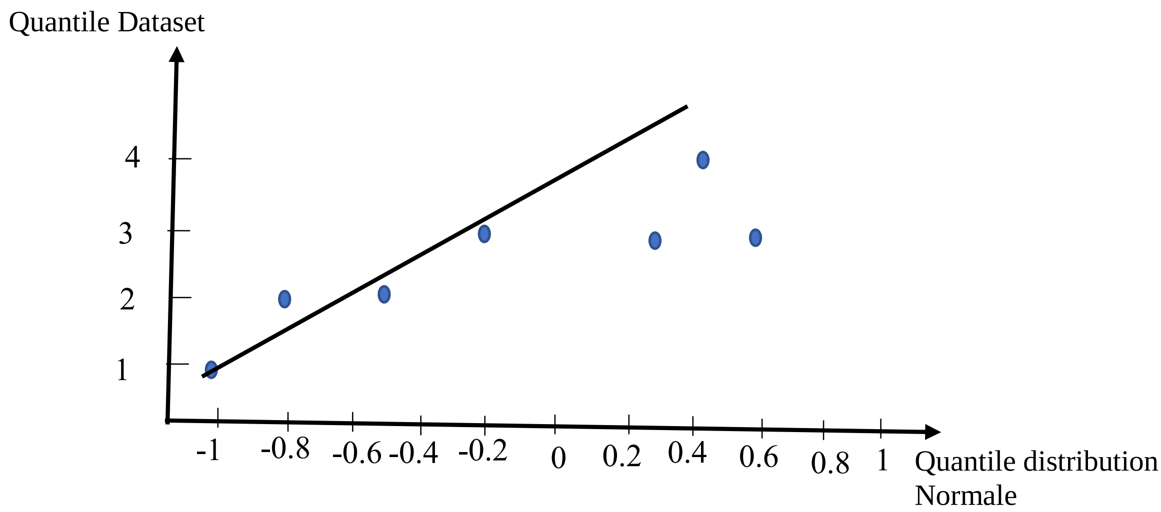
Calculer le score Z_1 telle que :

$P(Z < Z_1)=0.1$ (D_1) ; Probabilité d'obtenir $Z_1 \leq 0.1$

$P(Z < Z_1)=0.2$ (D_2), $P(Z < Z_1)=0.3$ (D_3), $P(Z < Z_1)=0.4$ (D_4),

$P(Z < Z_1)=0.5$ (D_5), $P(Z < Z_1)=0.6$ (D_6), $P(Z < Z_1)=0.7$ (D_7)

-1.28	-0.84	-0.52	-0.25	0.67	0.25	0.52
1	2	2	3	3	3	4



Les données ne sont pas à la distribution normale (car on a 3 points sur 7 qui ne sont pas alignés).

Exercice 2 :

- La population : Œufs

- Variable : Masse des Œufs (variable continu : intervalle)

Masse x_i	**Nombre des Œufs** n_i	C_i	**Effectif Cumulé** N_i	**Fréquence cumulé**	$(C_i - \overline{X})^2 * n_i$	**Pourcentage (%)**	**Pourcentage cumulé (%)**
[28-38[	4	33	4	0.011	607.62*4	1.14	1.14
[38-48[	55	43	59	0.16	214.62* 55	15.71	16.85
[48-58[	78	53	137	0.39	21.62*78	22.29	39.14 (Q_1)
[58-62[	112	60	249	0.71	5.52*112	32	71.14 (Q_2)
[62-72[	95	67	344	0.98	87,42*95	27.14	98.28 (Q_3, C_{85})
[72-38[	6	77	350	1	374.42*6	1.72	100
Total	350						

- Calculer l'étendue : E=82-28= 54, $E_1 = 4, E_4 = 10$

$\bar{X} = \frac{4*33+55*43+78*53+112*60+95*67+6*77+132+2365+4134+6720+6365+462}{350} = 57{,}65$

- Médiane (Me)

$\frac{N}{2} = 175$

Chercher la première classe médiane qui a un effectif cumulatif qui dépasse 175 : [58-62[

$$Me = 58 + 4 * \frac{(175 - 137)}{112} = 59{,}36 \in [58{,}62[$$

- Mode (Mo)

Chercher la première classe modale ayant un effectif maximal: [58-62[

$$Mo = 58 + 4 * \frac{34}{34 + 17} = 60{,}66 \in [58{,}62[$$

- C_{85}

Chercher la première classe centile qui un pourcentage cumulé qui dépasse 85%: [62-72[

$$C_{85} = 62 + 10 * \frac{\left(\frac{85}{100} * 350 - 249\right)}{95} = 62 + 10 * \frac{297.5 - 249}{95} = 67.10$$

- Q_1 (25%)=51.65 → [48-58[

- Q_3 (75%) → [62-72[

$$Q_3 = 62 + 10 * \frac{\left(\frac{3}{4} * 350 - 249\right)}{95} = 63.42$$

- Q_{25}

Chercher la première classe quantile contenant 25% de données (qui un pourcentage cumulé qui dépasse 25%): → $[48 - 58[$

$$Q_{25} = 48 + 10 * \frac{\left(\frac{25}{100} * 350 - 59\right)}{78} = 51.65$$

Interprétation : 25% des œufs ayant une masse moins de 51.65

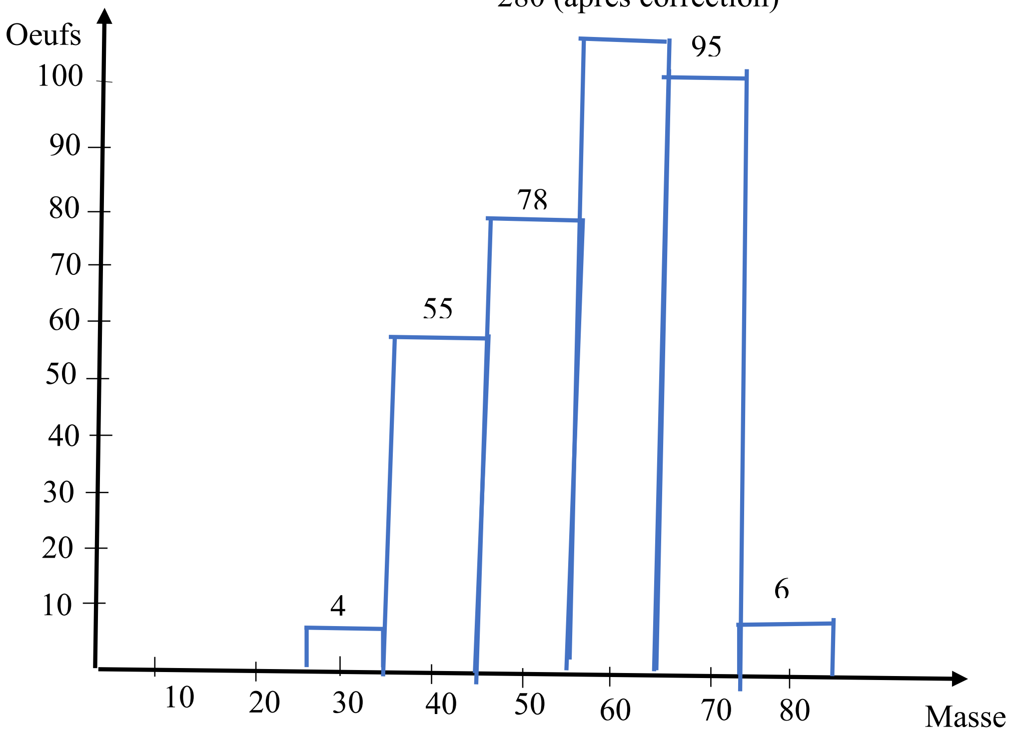
- Ecart interquartile : $Q_3 - Q_1 = 11.77$

- Ecart interdécile : $D_9 - D_1 = 68.94 - 43.63 = 25.31$

- Calculer la variance : $var = \frac{27091}{350} \approx 77.40$
- Calculer l'écart type : $\sigma = \sqrt{var} \approx 8.79$

- $[49{,}21,\ Q_3[$ → pourcentage d'individus appartenant [48.86, 63,42[= 22.29+32+27.14=81.44

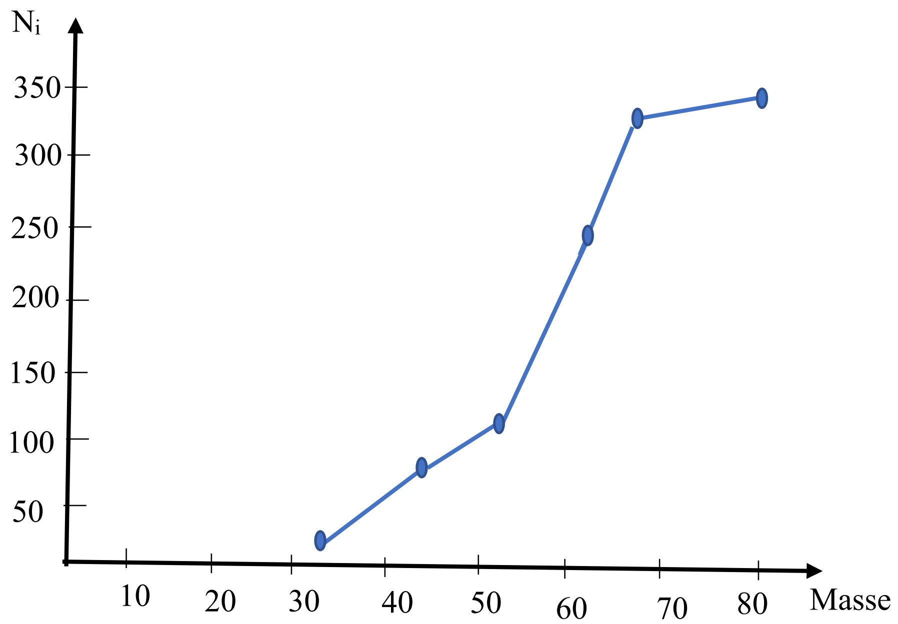
✓ Histogramme



On a un intervalle qui a une longueur 2, on doit corriger son effectif.

$a_{base}=10$, $a_i=4$, $n_{corr}=\frac{10}{4}*112=280$ de l'intervalle [58-62[

✓ Courbe Cumulative



Exercice 3 :

Série 1 (variable quantitative continue) : [18-20 [, [20-22 [, [22-24 [

Série 1 (variable quantitative discrète) : 1, 2, 3 , 4, 5

-Calculer la moyenne et l'écart type de chaque série

y \ x	[18-20[	[20-22[	[22-24[	$\boldsymbol{n_{j.}}$	$\boldsymbol{n_{j.}y_j}$	$(\boldsymbol{y_j}-\overline{\boldsymbol{y}})^2$	$\boldsymbol{n_{j.}}(\boldsymbol{y_j}-\overline{\boldsymbol{y}})^2$
1	19	56	11	86	86	5.06	435.37
2	71	182	23	276	552	1.56	431.25
3	149	391	57	597	1791	0.0625	37.31
4	36	187	51	274	1096	0.5625	154.125
5	24	181	79	284	1420	3.062	869.60
$\boldsymbol{n_{.i}}$	299	997	221	1517			1926.75

Moyenne marginale de y :

$$\bar{y} = \sum_{i=1}^{n} \frac{\boldsymbol{n_{j.}y_j}}{N}$$

$$\bar{y} = \frac{86 + 552 + 1791 + 1096 + 1420}{1517} = 3.25$$

Variance de y :

$$var_y = \sum_{i=1}^{n} \frac{\boldsymbol{n_{j.}}(\boldsymbol{y_j}-\overline{\boldsymbol{y}})^2}{N} = 1.27$$

$$\sigma_y = 1.12$$

Age (x)	$\boldsymbol{n_{.i}}$	$\boldsymbol{c_i}$	$\boldsymbol{n_{.i}}$	$(\boldsymbol{c_i}-\overline{\boldsymbol{x}})^2$	$\boldsymbol{n_{.i}}(\boldsymbol{c_i}-\overline{\boldsymbol{x}})^2$
[18-20[	299	19	5681	3.53	1079.32
[20-22[	997	21	20937	0.0529	9.97
[22-24[	221	23	5083	4.97	974.61

Moyenne marginale de x :

$$\bar{x} = \frac{5681 + 20937 + 5083}{1509} = 20.9$$

$$var_x = \sum_{i=1}^{n} \frac{\boldsymbol{n}_{.i}(\boldsymbol{y}_j - \overline{\boldsymbol{y}})^2}{N} = 1.36$$

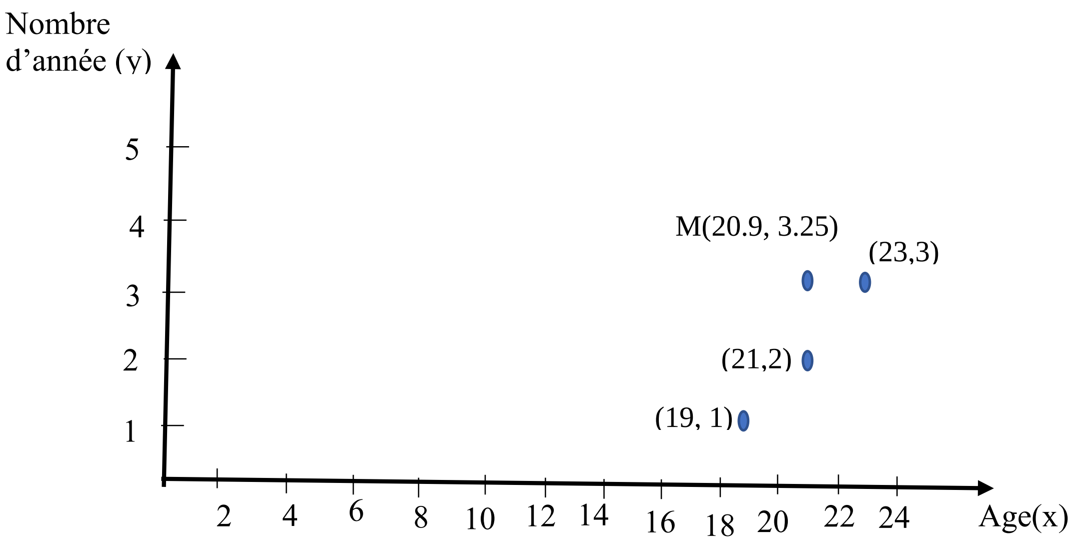
$$\sigma_{x} = 1.16$$

-Calculer le point moyen du nuage de point

$$\bar{x} = \frac{19 + 21 + 23}{3} = 21$$

$$\bar{y} = \frac{1 + 2 + 3 + 4 + 5}{5} = 3$$

Le point moyen du nuage M(21,3)



Exercice 4 :

X	Y	$(\boldsymbol{x_i} - \overline{\boldsymbol{x}})$	$(\boldsymbol{y_i} - \overline{\boldsymbol{y}})$	$(\boldsymbol{x_i} - \overline{\boldsymbol{x}})^2$	$(\boldsymbol{y_i} - \overline{\boldsymbol{y}})^2$
4	510	0	-642.22	0	412446.52
3	590	-1	-562,22	1	316091.328
2	900	-2	-252.22	4	63614.92
1	1420	-3	267.78	9	71796.13
0	2000	-4	847.78	16	718730.93
5	600	1	-552.22	1	304946.93
6	850	2	-302.22	4	91336.93
7	1300	3	147.78	9	21838.93
8	2200	4	1047.78	16	1097842.9
Total				60	3098645.53

$\bar{x} = 4, \qquad var_x = \frac{60}{9} = 6.66$

$\sigma_x = 2.21$

$\bar{y} = 1152.22\,,var_y = \frac{3098645.53}{9} = 3498645.53$

$\sigma_y = 586.8$

$Cov(x,y) = \frac{1}{9}(4 * 510 + 3 * 590 + \cdots + 8 * 2200) - 4 * 1152 \; = \frac{1}{9}(41830) - 4 * 1152 = 38.9$

$r_{x,y} = \frac{38.9}{2.58 * 586.8} = 0.02$ → Pas de corrélation entre le nombre d'enfants et le salaire.

Exercice 5 :

- Calculer les moyennes marginales $\bar{X}$et $\bar{Y}$ et les écarts-types marginaux $\sigma_x$et σ_y.

y \ x	[1-3[	[3-11[	[11-19[	[19-31[	[31-59[	$n_{i.}$
[5-7[	0	0	2	9	29	40
[7-9[	0	3	8	26	15	52
[9-11[	2	12	35	22	6	77
[11-15[	36	26	16	3	0	81
$n_{.j}$	38	41	61	60	50	250

x_i	**$n_{i.}$**	$\boldsymbol{c_i}$	$\boldsymbol{n_{i.}c_i}$	$(\boldsymbol{c_i - \bar{x}})$	$(\boldsymbol{c_i - \bar{x}})^2$	$\boldsymbol{n_{i.}(c_i - \bar{x})^2}$
[5-7[	40	6	240	-3.92	15.37	614.8
[7-9[	52	8	416	-1.92	3.69	191.88
[9-11[	77	10	770	0.08	0.0064	0.4928
[11-15[	81	13	1053	3.08	9.49	768.69
Total	250		2479			1575.86

y_i	$\boldsymbol{n_{.j}}$	$\boldsymbol{c_j}$	$\boldsymbol{n_{.j}c_j}$	$(\boldsymbol{c_j - \bar{y}})$	$(\boldsymbol{c_j - \bar{y}})^2$	$\boldsymbol{n_{.j}(c_i - \bar{x})^2}$
[1-3[	38	2	76	-18.11	327.97	12462.86
[3-11[	41	7	287	-13.11	171.87	7046.67
[11-19[	61	15	915	-5.11	26.11	1592.71
[19-31[	60	25	1500	4.89	23.91	1436.6
[31-59[	50	45	2250	24.89	619.51	30975.5
Total			5028		1169.37	53514.34

-Déterminer la covariance et le coefficient de corrélation linéaire.

$$r_{x,y} = \frac{cov(x,y)}{\sigma_x \ \sigma_y}$$

$$\boldsymbol{C}ov(x,y) = \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\boldsymbol{n_{ij}}\,(\boldsymbol{c_i}-\overline{\boldsymbol{x}})(\boldsymbol{c_j}-\overline{\boldsymbol{y}})}{N} = (\frac{\boldsymbol{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}\sum_{\boldsymbol{j=1}}^{\boldsymbol{m}}\boldsymbol{c_i c_j}\,\boldsymbol{n_{ij}}) - \overline{\boldsymbol{x}}\,\overline{\boldsymbol{y}}$$

$$\bar{x} = \sum_{i=1}^{n}\frac{\boldsymbol{n_{i.}}(\boldsymbol{c_i})}{N} = \frac{2479}{250} = 9.92$$

$$var_x = \sum_{i=1}^{n}\frac{\boldsymbol{n_{i.}}(\boldsymbol{c_i}-\overline{\boldsymbol{x}})^{\boldsymbol{2}}}{N} = 6.30$$

$$\sigma_x = 2.51$$

$$\bar{y} = \frac{5028}{250} = 20,11$$

$$var_y = \sum_{i=1}^{n}\frac{\boldsymbol{n_{.i}}\,(\boldsymbol{c_i}-\overline{\boldsymbol{y}})^{\boldsymbol{2}}}{N} = 214.05$$

$$\sigma_y = 14.63$$

$$Cov(x,y) = \frac{1}{250}(6*2*0+6*7*0+\cdots+13*45*0) - 9.92*20.11$$
$$= \frac{1}{250}(42815) - 9.92*20.11 = 173.54 - 199.49 = -28.23$$

c_i	c_j	n_{ij}	$c_i c_j n_{ij}$
6	2	0	0
6	7	0	0
6	15	2	180
6	25	9	1350
6	45	29	7830
8	2	0	0
8	7	3	168
8	15	8	960
8	25	26	5200
8	45	15	5400
10	2	2	40
10	7	12	840
10	15	35	5250
10	25	22	5500
10	45	6	2700
13	2	36	936
13	7	26	2360
13	15	16	3120
13	25	3	975

13	45	0	0

$$r_{x,y} = \frac{-28.23}{2.51 * 14.63} = -0.76$$