Université Mohammed V
Faculté des Sciences de Rabat
Département de Mathématiques

Année Universitaire **2022–2023**
Module **Probabilités et Statistique**
Filière – **SMI (S3)**

TD n° 1

Statistique descriptive univariée et bivariée

– *Un corrigé*

Questions. Parmi ces assertions, préciser celles qui sont vraies, celles qui sont fausses :

1. On appelle variable, une caractéristique que l'on étudie
2. La tâche de la statistique descriptive est de recueillir des données.
3. La tâche de la statistique descriptive est de présenter les données sous forme de tableaux, de graphiques et d'indicateurs statistiques.
4. Les valeurs des variables sont aussi appelées modalités.
5. Pour une variable qualitative, chaque individu statistique ne peut avoir qu'une seule modalité.
6. Pour faire des traitements statistiques, il arrive qu'on transforme une variable quantitative en variable qualitative.

Solution –

1) Vrai
2) Faux
3) Vrai
4) Vrai
5) Vrai
6) Vrai

Exercice 1. Donner la nature de chacune des variables suivantes :

(*i*) Degré de satisfaction concernant le séjour estival dans un centre de vacances (Très élevé ; Élevé ; Modéré ; Bas ; Nul).
(*ii*) La moyenne semestrielle d'un étudiant à la faculté de pharmacie.
(*iii*) Le temps consacré par un étudiant de S2 pour suivre les informations sur un journal écrit ou télévisé.
(*iv*) La section (A,B,C,...) du cours de biologie à laquelle est inscrit un étudiant.
(*v*) Le nombre d'appels téléphoniques reçus en une journée par un étudiant.

Solution –

(*i*) *Degré de satisfaction concernant le séjour estival dans un centre de vacances (Très élevé ; Élevé ; Modéré ; Bas ; Nul) :*
C'est une var. qualitative ordinale. Une hiérarchie est établie entre les degrés de satisfaction.
(*ii*) *La moyenne semestrielle d'un étudiant à la faculté de pharmacie :*
C'est une var. quantitative continue prenant ses valeurs dans un intervalle de $\mathbb{R}^+$: $[0, 20]$.
(*iii*) *Le temps consacré par un étudiant de S2 pour suivre les informations sur un journal écrit ou télévisé :*
C'est une var. quantitative continue prenant ses valeurs dans un intervalle $[0, t]$, $t > 0$.
(*iv*) *La section (A,B,C,...) du cours de biologie à laquelle est inscrit un étudiant :*
C'est une var. qualitative nominale.
(*v*) *Le nombre d'appels téléphoniques reçus en une journée par un étudiant :*
C'est une var. quantitative discrète prenant ses valeurs dans $\{0, 1, 2, \cdots, n\}$, $n \in \mathbf{N}^*$.

Exercice 2. On a demandé à 300 jeunes collégiens quel était leur fruit préféré parmi les six fruits les plus consommés au Maroc : banane, nectarine, orange, pêche, poire, pomme. Voici les résultats obtenus :

Fruit	Banane	Nectarine	Orange	Pêche	Poire	Pomme
Effectif	72	33	30	36	45	84

1) Identifier la variable et préciser sa nature.
2) Dresser le tableau des fréquences.
3) Existe-t-il un mode ? si oui, donner le.
4) Faire deux représentations graphiques de cette variable.

Solution –

1. *Il s'agit de la variable "le fruit préféré des 300 collégiens". C'est une var. qualitative nominale.*
2. *Tableau des fréquences :*

Fruit	*Eff* n_i	*Fréquence* f_i
Banane	72	0.24
Nectarine	33	0.11
Orange	30	0.10
Pêche	36	0.12
Poivre	45	0.15
Pomme	84	0.28
Total	300	1

3. *"La pomme" dont l'**effectif est le plus élevé** est la valeur modale (ou mode) de cette série statistique.*
4. *(i) On trace le Diagramme en tuyaux d'orgue des effectifs (ou des fréquences). (appelé aussi Diagramme en barres).*
   (ii) On trace le Diagramme en Secteur : $\alpha_i = f_i \times 360$ *est le nombre de degrés mesurant le secteur angulaire de la modalité* $i = 86;\ ;40;\ 36;\ 54$ *et* 101.

Exercice 3. Une enquête en vue de la réduction du montant des allocations familiales, a été réalisée auprès d'un échantillon de 100 femmes de 40 ans. Cette enquête a donné les résultats suivants :

Nombre d'enfants (x_i)	*Nombre de femmes* (n_i)
0	10
1	20
2	20
3	30
4	20

1) Caractériser la distribution *(la population et sa taille, l'individu, les modalités, le caractère (la variable) et son type).*
2) Tracer le diagramme correspondant.
3) Définir et représenter la courbe cumulative croissante.
4) Donner la proportion (fréquence) des femmes ayant moins de 4 enfants.
5) Donner la fréquence des femmes ayant au moins 2 enfants.

Solution *–*

1) *Caractériser la distribution* (la population et sa taille, l'individu, les modalités, le caractère (la variable) et son type).

la population : "femmes de 40ans" ; sa taille : $n = 50$*, l'individu : "une femme de 40ans", les modalités : "*$0, 1, 2, 3, 4$*", le caractère (la variable) :* X *: "Nombre de femmes" ; son type : "quantitatif discret.*

1) *Tracer le diagramme correspondant : On peut tracer soit le digramme en bâtons des effectif soit le diagramme en bâtons des fréquences.*

3) *Définir et représenter la courbe cumulative croissante.*

La représentation de la fonction cumulative croissante (appelée aussi fonction de répartition) est réalisée au moyen des fréquences cumulées. Cette fonction est définie de $\mathbb{R}$ *dans* $[0, 1]$ *et vaut*

$$F(x) = \begin{cases} 0 & si\ x < 0 \\ 0.1 & si\ 0 \leq x < 1 \\ 0.3 & si\ 1 \leq x < 2 \\ 0.5 & si\ 2 \leq x < 3 \\ 0.8 & si\ 3 \leq x < 4 \\ 1 & si\ x \geq 4 \end{cases}$$

voir tableau :

Nombre d'enfants (x_i)	*Nombre de femmes* (n_i)	f_i	F_i
0	10	0.1	0.1
1	20	0.2	0.3
2	20	0.2	0.5
3	30	0.3	0.8
4	20	0.2	1
Σ	100	1	//

4) *Donner la proportion (fréquence) des femmes ayant moins de* 4 *enfants :* 0.8
5) *Donner la fréquence des femmes ayant au moins* 2 *enfants :* $0.2 + 0.3 + 0.2 = 0.7$

Exercice 4. On a relevé la taille (en cm) de 50 étudiantes de la filière **SMI**, les résultats sont regroupés dans le tableaux suivant

Classe	$[151.5, 155.5[$	$[155.5, 159.5[$	$[159,5; 163,5[$	$[163,5; 167,5[$	$[167,5; 171,5[$
Effectif	10	12	11	7	10

1. Caractériser la distribution *(la population et sa taille, l'individu, la variable et son type).*
2. Dresser le tableau statistique complet (calculer les fréquences, les fréquences cumulées et les effectifs cumulés)
3. Tracer le diagramme correspondant.
4. Quelle est la classe modale ?
5. Définir et représenter la courbe cumulative croissante.
6. Calculer la moyenne et la variance.
7. Calculer le coefficient de variation. Interpréter le résultat.
8. Calculer la médiane ainsi que le premier et le troisième quantile.
9. Quelle est la fréquence des étudiantes ayant au moins $165cm$?

Solution –

(1) Caractériser la distribution (la population et sa taille, l'individu, la variable et son type).
Population étudiée : Les étudiantes de la filière SMI ; Taille : 50 *;*
L'individu : une étudiante de la filière SMI ;
Variable : "taille en cm des étudiantes" ; Type : Quantitative continue.

(2) Le tableau statistique est le suivant :

Classe	n_i	f_i	F_i	N_i
$[151.5, 155.5[$	10	0.20	0.20	10
$[155.5, 159.5[$	12	0.24	0.44	22
$[159,5; 163,5[$	11	0.22	0.66	33
$[163,5; 167,5[$	7	0.14	0.80	40
$[167,5; 171,5[$	10	0.20	1.00	50
Σ	50	1.00	//	//

(3) Le diagramme correspondant : Puisque la variable est quantitative continue, on trace l'histogramme des effectif ou des fréquence. Et puisque les classes sont d'amplitudes égales alors on trace directement l'histogramme.

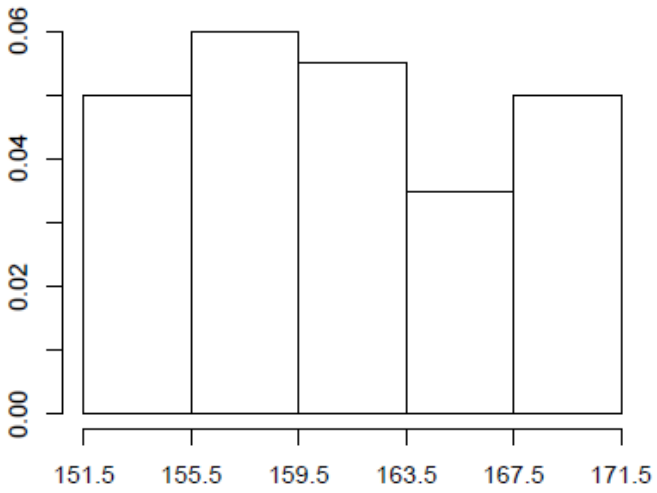


FIGURE 1 – Histogramme des fréquences

(4) Puisque les classes sont d'amplitudes égales alors on retrouve directement la classe qui contient l'effectif (ou la fréquence) le plus élevé(e) : il s'agit de la classe des taille entre 155.5 *et* 159.5 *centimètre.*

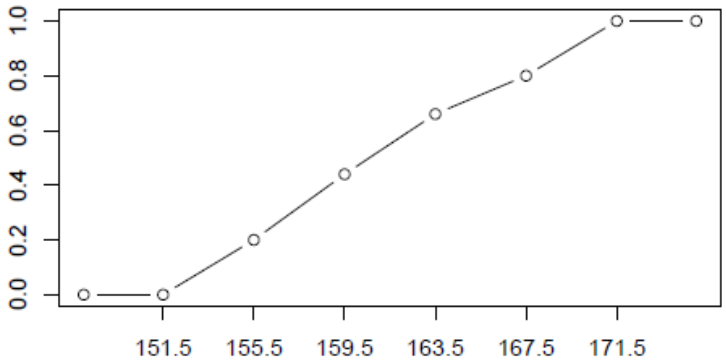


FIGURE 2 – Fonction cumulative (fonction de répartition)

(5) La courbe cumulative croissante (fonction de répartition) est définie par les points $A_i(x_{i+1}, F_i)$ *donnés dans le tableau statistique.*

(6) La moyenne : $\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i c_i = \sum_{i=1}^{k} f_i c_i$, *avec* $c_i = \frac{x_i + x_{i+1}}{2}$ *est le centre de la classe* $[x_i, x_{i+1}[$.

$$\begin{aligned}\bar{x} &= \frac{10 \times 153.5 + 12 \times 157.5 + 11 \times 161.5 + 7 \times 165.5 + 10 \times 169.5}{50} \\ &= 0.20 \times 153.5 + 0.24 \times 157.5 + 0.22 \times 161.5 + 0.14 \times 165.5 + 0.20 \times 169.5 \\ &= 161.1cm.\end{aligned}$$

(6) La variance : $S^2 = \frac{1}{n}\sum_{i=1}^{k} n_i c_i^2 - (\bar{x})^2 = \sum_{i=1}^{k} f_i c_i^2 - (\bar{x})^2$.

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{k} n_i c_i^2 &= \frac{10 \times 153.5^2 + 12 \times 157.5^2 + 11 \times 161.5^2 + 7 \times 165.5^2 + 10 \times 169.5^2}{50} \\ &= 25984.73cm^2\end{aligned}$$

$$\sum_{i=1}^{k} f_i c_i^2 = 0.20 \times 153.5^2 + 0.24 \times 157.5^2 + 0.22 \times 161.5^2 + 0.14 \times 165.5^2 + 0.20 \times 169.5^2.$$

$$S^2 = 25984.73 - 161.1^2 = 31.52cm^2.$$

(7) Le coefficient de variation

$$CV = \frac{S}{\bar{x}} \times 100 = \frac{\sqrt{31.52}}{161.1} \times 100 = 03.49\%.$$

Interprétation : la série est très homogène.

Classe	F_i	N_i
$[151.5, 155.5[$	0.20	10
$[155.5, 159.5[$	0.44	22
$[159,5; 163,5[$	0.66	33
$[163,5; 167,5[$	0.80	40
$[167,5; 171,5[$	1.00	50

(8) La médiane : $M_e \in]159,5; 163,5[$ *:*

$$M_e = 159.5 + \frac{0.50 - 0.44}{0.66 - 0.44} \times (163.5 - 159.5) \simeq 160.59cm$$

Le premier quartile : $Q_1 \in]155,5; 159,5[$ *:*

$$Q_1 = 155.5 + \frac{0.25 - 0.20}{0.44 - 0.20} \times (159.5 - 155.5) \simeq 156.33cm$$

Le troisième quartile : $Q_3 \in]163,5; 167,5[$ *:*

$$Q_3 = 163.5 + \frac{0.75 - 0.66}{0.80 - 0.66} \times (167.5 - 163.5) \simeq 166.07cm$$

$$\Rightarrow \qquad EIQ = Q_3 - Q_1 \simeq 9.74cm$$

(9) Quelle est la fréquence des étudiantes ayant au moins $165cm$ *?*
Par interpolation, on cherche d'abord la fréquence f *des étudiantes ayant moins de* $165cm$ *: puisque* $165 \in]163,5; 167,5[$*, alors par interpolation linéaire on a :*

$$\frac{f - 0.66}{0.80 - 0.66} = \frac{165 - 163.5}{167.5 - 163.5}$$

qui donne $f = 0.66 + \frac{165-163.5}{167.5-163.5} \times (0.80 - 0.66) = 0.7125$
Donc la proportion (fréquence) des étudiantes ayant au moins $165cm$ *est égale à* $1 - 0.7125 = 0.2875$

Exercice 5. Les données suivantes sont les frais d'électricité (en DH) durant le mois de mars pour un échantillon de 50 petits appartements dans une grande ville :

80	90	95	96	102	108	109	111	114	116
119	123	127	128	129	130	130	135	137	139
141	143	144	147	148	149	149	150	151	153
154	157	158	163	165	166	167	168	171	172
175	178	183	185	187	191	197	202	206	220

1) Calculer les quartiles Q_1, Q_2 et Q_3 de cet échantillon.
2) Regrouper les données en classes d'amplitudes égales puis construire le tableau des fréquences et fréquences cumulées.
3) Tracer l'histogramme et le polygone des fréquences.
Solution *–*

1) En utilisant la procédure de calcul,vu dans le cours, les données sont ordonnée :
– *Calcul de* Q_1 *:* $np = 50 \times 0.25 = 12.5$ *n'est pas un entier, donc :*

$$Q_1 = x_{(\lceil 12.5 \rceil)} = x_{(13)}$$

– *Calcul de* $M_e = Q_2$ *:* $np = 50 \times 0.5 = 25$ *est un entier, donc :*

$$M_e = \frac{x_{(np)} + x_{(np+1)}}{2} = \frac{x_{(25)} + x_{(26)}}{2} = \frac{148 + 149}{2} = 148.5$$

– *Calcul de* Q_3 *:* $np = 50 \times 0.75 = 37.5$ *n'est pas un entier, donc :*

$$Q_3 = x_{(\lceil 37.5 \rceil)} = x_{(38)} = 168$$

2) La formule de ***Sturges*** *donne le nombre de classes :*

$$k = 1 + 3.33 \log_{10} N \simeq 1 + 3.33 \log_{10}(50) \simeq 7 \text{ classes.}$$

On calcule l'étendue : $e = x_{max} - x_{min} = 220 - 80 = 140$.
On calcule le pas (l'amplitude) de chaque classe : $A = e/140 = 20$
On obtient le tableau :

$Classe_i$	n_i	f_i	F_i
$[80, 100[$	4	0.08	0.08
$[100, 120[$	7	0.14	0.22
$[120, 140[$	9	0.18	0.40
$[140, 160[$	13	0.26	0.66
$[160, 180[$	9	0.18	0.84
$[180, 200[$	5	0.10	0.94
$[200, 220[$	3	0.06	1
Total	50	1	//

3) Construire l'histogramme, puis après le polygone des fréquence (sur la même figure)

Exercice 6. Dans une enquête, menée auprès des étudiants de l'université Mohammed V, l'enquêteur relevait le temps (en minutes) mis par chaque répondant pour se rendre à l'université. Le tableau suivant résume les temps observés.

Classe	[21; 22[	[22; 23[	[23; 24[	[24; 26[	[26; 30[
Effectif	50	90	70	60	40

1) Dresser le tableau des fréquences cumulées.
2) Représenter ces données à l'aide d'un histogramme et tracer le polygone des fréquences.
3) Calculer la moyenne et la variance.
4) Calculer les quartiles.
5) Cette distribution est-elle symétrique ou asymétrique ?

Solution –

1) *On obtient le tableau de fréquence suivant,*

Classes	*Eff* n_i	*Fréq* f_i	*Fréq cumu* $F(x)$	*Fréq corri* f_i^c	c_i	$f_i c_i$	$n_i(c_i - \overline{x})^2$
[21,22[	*50*	*0.16*	*0.16*	*0.16*	*21.5*	*3.44*	*255.38*
[22,23[	*90*	*0.29*	*0.45*	*0.29*	*22.5*	*6.525*	*142.884*
[23,24[	*70*	*0.23*	*0.68*	*0.23*	*23.5*	*5.405*	*4.732*
[24,26[	*60*	*0.19*	*0.87*	*0.09*	*25*	*4.75*	*92.256*
[26,30[	*40*	*0.13*	*1*	*0.03*	*28*	*3.64*	*719.104*
Total	*310*	*1*	-	-	-	*23.76*	*1214.356*

où la fréquence corrigée de la classe i est noté f_i^c est définie par,

$$f_i^c = fi \times \frac{a_0}{a_i}.$$

— a_0 *étant l'amplitude de base (ici 1).*

— a_i *est l'amplitude de la classe i.*

— *l'amplitude de la classe* $[x_i, x_{i+1}[$ *est* $a_i = x_{i+1} - x_i$.

2) *Histogramme et polygone de fréquence (voir la figure ci-dessous)*

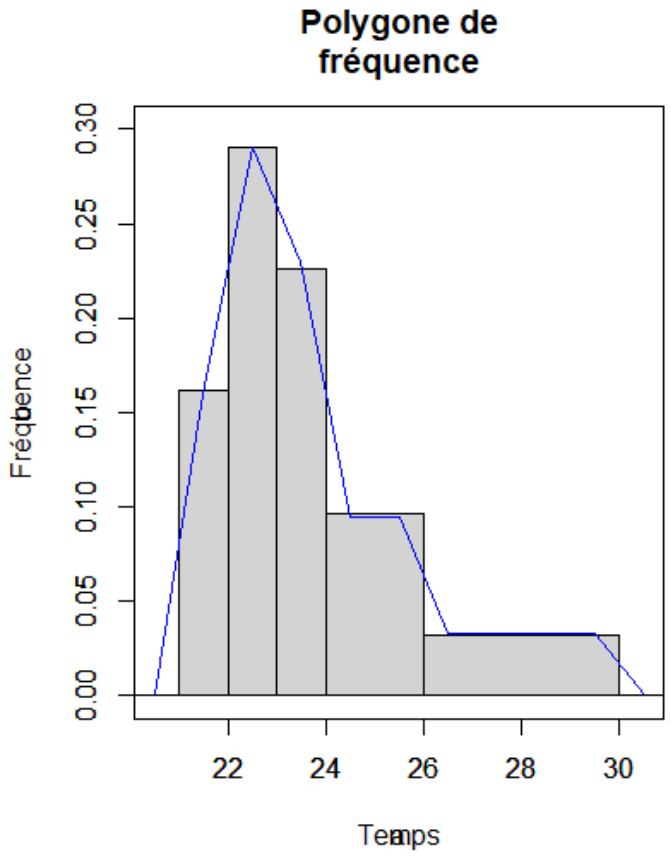


FIGURE 3 – Histogramme et polygone des fréquences de la distribution

3) *La moyenne est donnée par*

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{5} n_i c_i = \sum_{i=1}^{5} f_i c_i = 23.76 \ min$$

et la variance par

$$s^2 = \frac{1}{n}\sum_{i=1}^{5} n_i (c_i - \overline{x})^2 = \frac{1214.356}{310} \simeq 3.92 \ min^2$$

4) *Pour calculer les quartiles, on peut obtenir les formules pour* $j = 1, 2, 3$

$$Q_j = x_i + (x_{i+1} - x_i) \times \frac{F(Qj) - F(x_i)}{F(x_{i+1}) - F(x_i)}; \qquad Q_j \in [x_i, x_{i+1}[$$

Dans ce cas,

-
$$Q_1 = 22 + (23 - 22) \times \frac{0.25 - 0.16}{0.45 - 0.16} = 22.31,$$

-
$$Q_2 = 23 + (24 - 23) \times \frac{0.5 - 0.45}{0.68 - 0.45} = 23.22,$$

-
$$Q_3 = 24 + (26 - 24) \times \frac{0.75 - 0.68}{0.87 - 0.68} = 24.75.$$

5) *Le coefficient d'asymétrie de* ***Pearson*** *est :*

$$\gamma_3 = \frac{3(\bar{x} - Q_2)}{s} = \frac{3(23.76?23.22)}{\sqrt{3,92}} \simeq 0.82.$$

γ_3 *est positif donc la série présente une asymétrie à droite,*

Exercice 7. Une étude de budget a donné les résultats suivants :

Budget	$[800, 1000[$	$[1000, 1400[$	$[1400, 1600[$	$[1600, y[$	$[y, 2400[$	$[2400, x[$
Fréq. cumulée	0.08	0.18	0.34	0.64	0,73	1

PARTIE 1 : Certaines données sont manquantes.

1) Calculer la borne manquante x sachant que l'étendue de la série est égale à 3200 euros.

2) Calculer la borne manquante y dans les deux cas suivants

a) le budget moyen est égal à 1995 euros,

b) le budget médian est égal à 1920 euros.

PARTIE 2 : Considérons maintenant que la borne manquante y est égale à 2000 euros.

3) Donner une représentation graphique de la distribution des budgets.

4) Calculer le budget moyen et médian.

5) Retrouver les effectifs n_i correspondant à chacune des tranches de budgets ainsi que l'effectif total n, sachant que :

$$\sum_{i=1}^{n} n_i c_i^2 = 4741200000 \qquad \text{et} \qquad \mathbb{V}(X) = 604044.$$

Solution *–*

PARTIE 1 : *Certaines données sont manquantes.*

1. *La borne manquante* $x = x_{\max}$ *sachant que l'étendue e de la série est égale à* 3200 *euros :*
   On sait que l'étendue $e = x_{\max} - x_{\min}$
   et donc
   $3200 = x_{\max} - 800 \Longrightarrow x = x_{\max} = 4000.$
2. *La borne manquante y dans les deux cas suivants :*

(a) le budget moyen est égal à 1995 *euros :*

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i c_i = \sum_{i=1}^{k} f_i c_i$$

Il faut donc au préalable calculer les fréquences à partir des fréquences cumulées dans le tableau précédent.

Classes	$[800, 1000[$	$[1000, 1400[$	$[1400, 1600[$	$[1600, y[$	$[y, 2400[$	$[2400, 4000[$
F_i	0.08	0.18	0.34	0.64	$0,73$	1
f_i	0.08	0.1	0.16	0.3	$0,09$	0.27

Donc
$\bar{x} = \sum_{i=1}^{k} f_i c_i = 1995$
c-à-d,
$0.08 \times 900 + 0.1 \times 1200 + 0.16 \times 1500 + 0,3 \times \frac{1600+y}{2} + 0.09 \times \frac{y+2400}{2} + 0,27 \times 3200 = 1995$
On trouve : $y = 1800$.

(b) le budget médian est égal à 1920 *euros.*
Le budget médian est égal à 1920 *euros. Il faut raisonner par interpolation linéaire sur l'intervalle* $[1600, y[$ *:*

$$\frac{1920 - 1600}{y - 1600} = \frac{0.5 - 0.34}{0.64 - 0.34}.$$

On trouve : $y = 2200$.

PARTIE 2 : *Considérons maintenant que la borne manquante* y *est égale à* 2000 *euros.*

3. *Donner une représentation graphique de la distribution des budgets.*
On trace l'histogramme : *Pour donner la représentation graphique correcte de la distribution,, il faut au préalable corriger les fréquences puisque les amplitudes des classes ne sont pas égales :*

$$f_i^c = \frac{f_i}{a_i} \times \alpha$$

α *est le correcteur de l'échelle, est égale à la valeur de la plus petite amplitude ou la valeur de l'amplitude qui se répète. Ici on prend* $\alpha = 400$

Classes	$[800, 1000[$	$[1000, 1400[$	$[1400, 1600[$	$[1600, 2000[$	$[2000, 2400[$	$[2400, 4000[$
a_i	200	400	200	400	400	1600
f_i	0.08	0.1	0.16	0.3	$0,09$	0.27
f_i^c	0.16	0.1	0.32	0.3	$0,09$	0.0675

4. *Calculer le budget moyen et médian.*
$\bar{x} = \sum_{i=1}^{k} f_i c_i = 0.08 \times 900 + 0.1 \times 1200 + 0.16 \times 1500 + 0.3 \times 1800 + 0.09 \times 2200 + 0.27 \times 3200 = 2034$
Le budget médian est dans l'intervalle $[1600, 2000[$ *. Il faut raisonner par interpolation linéaire :*

$$\frac{M_e - 1600}{2000 - 1600} = \frac{0.5 - 0.34}{0.64 - 0.34}.$$

On trouve : $M_e = 1813$.

5. *Retrouver les effectifs* n_i *correspondant à chacune des tranches de budgets ainsi que l'effectif total* n*, sachant que :*

$$\sum_{i=1}^{n} n_i c_i^2 = 4741200000 \qquad et \qquad \mathbb{V}(X) = 604044.$$

On a :

$$V(X) = 604044 = \frac{1}{n}\sum_{i=1}^{k} n_i (c_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{k} n_i c_i^2 - (\bar{x})^2$$

et

$$\sum_{i=1}^{n} n_i c_i^2 = 4741200000 \qquad et \quad \bar{x} = 2034.$$

Donc

$$604044 = \frac{1}{n} \times 4741200000 - (2034)^2$$

$$\Longrightarrow n = 1000$$

et pour calculer les effectifs des classes on applique la formule : $n_i = f_i \times n$

Exercice 8. On considère les statistiques (des *"frais d'électricité (en DH)"* durant le mois de mars pour un échantillon de 50 petits appartements dans une grande ville) données dans l'**Exercice 5**.

1) Construire le diagramme en boite.

2) Interpréter les résultats.

Solution *–*

Les paramètres de la boite à moustache : *(voir* ***Exercice 5.****)*

– $Q_1 = 127$

– $M_e = 148.5$

– $Q_3 = 168$

– $EIQ = Q_3 - Q_1 = 168 - 127 = 41$

– $a = \max(Q_1 - 1.5 \times EIQ; x_{\min}) = \max(65.5; 80) = 80$

– $b = \min(Q_3 + 1.5 \times EIQ; x_{\max}) = \min(229.5; 220) = 220$

Cette distribution ne présente aucune valeur aberrante.

Exercice 9. Une voiture roule pendant 200 kilomètres à $50km/h$, puis pendant 100 kilomètres à $100km/h$.

Quelle est sa vitesse moyenne sur son trajet ?

Solution *–*

Cette vitesse moyenne sera égale au rapport entre la distance parcourue et le temps de trajet. Soit $200 + 100 = 300$ *kilomètres parcourus.*

$$\overline{x}_H = \frac{300}{\frac{200}{50} + \frac{100}{100}} = 60km/h$$

Soit :

– $200km$ *à* $50km/h$ *durent* $4heures$ *;*

– $100km$ *à* $100km/h$ *durent* $1heure$.

Le trajet dure donc 5 *heures pour 300 kilomètres parcourus : la vitesse moyenne est bien de* $60km/h$.

Exercice 10. Soit le tableau suivant donnant la distribution du couple (X, Y).

X \ Y	0	1
$[0.5, 1.5[$	21	8
$[1.5, 2.5[$	23	15
$[2.5, 3.5[$	10	23

1) Quelles sont les distributions marginales de X et de Y ?

2) Calculer les moyennes et les variances marginales de X et de Y.

3) Calculer le coefficient de variation marginale de Y. Interpréter.

4) Les variables X et Y sont elles indépendantes ?

5) Calculer la moyenne et la variance de la variable $Z = 0.165X + 0.13Y$.

Solution *–*

1. *La distribution marginale de* X *est donnée dans le tableau suivant :*

X	*effectif*
$[0.5, 1.5[$	29
$[1.5, 2.5[$	38
$[2.5, 3.5[$	33
Σ	100

La distribution marginale de Y est donnée dans le tableau suivant :

Y	*effectif*
0	54
1	46
Σ	100

2. On trouve :

$$\overline{x} = \frac{1}{100}\sum_{i=1}^{3} n_{i.}c_i = \frac{29\times 1 + 38\times 2 + 33\times 3}{100} = 2.04,$$

$$\overline{y} = \frac{1}{100}\sum_{j=1}^{2} n_{.j}y_j = \frac{54\times 0 + 46\times 1}{100} = 0.46,$$

$$V(X) = s_x^2 = \left(\frac{1}{100}\sum_{i=1}^{3} n_{i.}c_i^2\right) - (\overline{x})^2 = 4.78 - 2.04^2 = 0.6184,$$

$$V(Y) = s_y^2 = \left(\frac{1}{100}\sum_{j=1}^{2} n_{.j}y_j^2\right) - (\overline{y})^2 = 0.2484.$$

3. $CV_Y = \frac{s_y}{\overline{y}} = \frac{\sqrt{0.2484}}{0.46} = 1.083473 \simeq 108\%$. *la distribution de Y est hétérogène.*

4. Rappelons que les variables X et Y sont indépendantes si et seulement si

$$n_{ij} = \frac{n_{i.}\times n_{.j}}{n}, \forall i = 1,2,3 \quad et \quad j = 1,2.$$

X \ Y	0	1	Σ
$[0.5, 1.5[$	21	8	29
$[1.5, 2.5[$	23	15	38
$[2.5, 3.5[$	10	23	33
Σ	54	46	100

Or, on a (contre exemple)

$$n_{21} = 23 \neq \frac{n_{2.}\times n_{.1}}{n} = \frac{38\times 54}{100} = 20.52,$$

donc les variables X et Y sont liées.

5. $V(Z) = V(0.165X + 0.13Y) = 0.165^2V(X) + 0.13^2V(Y) + 2\times 0.165\times 0.13\ cov(X,Y)$,
avec, la covariance entre X et Y :

$$s_{xy} = cov(X,Y) = \left(\frac{1}{100}\sum_{i=1}^{3}\sum_{j=1}^{2} n_{ij}c_iy_j\right) - \overline{x}\times\overline{y} = 0.1316$$

Exercice 11. Un responsable bancaire aimerait savoir s'il existe une relation entre le revenu annuel X et le montant d'argent Y consacré à l'épargne. Pour un échantillon de 10 familles, il a obtenu les résultats suivants (en 10^4DH)

X	12	15	13	10	10	14	16	18	16	14
Y	0,2	1,2	1	0,7	0,3	1	1,6	1,4	1,2	0,7

1) Calculer les moyennes $\bar{x}, \bar{y}$ et les écart-types s_x, s_y :
2) Calculer le coefficient de corrélation et interpréter ce résultat.
3) Chercher l'expression de la droite de régression de Y en X.
4) Estimer le montant d'argent consacré à l'épargne par une famille ayant un revenu de 110000DH.

Solution –

1. *On trouve :*

$$\overline{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = \frac{138}{10} = 13.8,$$

$$\overline{y} = \frac{1}{10}\sum_{i=1}^{10} y_i = \frac{9.3}{10} = 0.93,$$

$$s_x = \sqrt{\frac{1}{10}\sum_{i=1}^{10}(x_i - \overline{x})^2} \simeq 2.4855,$$

$$s_y = \sqrt{\frac{1}{10}\sum_{i=1}^{10}(y_i - \overline{y})^2} \simeq 0.4269.$$

2. *Le coefficient de corrélation est* $\rho = \frac{s_{xy}}{s_x s_y}$, *où*

$$s_{xy} = \frac{1}{10}\sum_{i=1}^{10}(x_i - \overline{x})(y_i - \overline{y}) = \frac{8.86}{10} = 0.886.$$

   Donc $\rho \simeq 0.835$, *il y a une forte liaison entre* Y *et* X.

3. *Puisque* $\rho > 0.8$, *on peut chercher à déterminer la droite de régression de* Y *en* X. *D'autre on peut définir par l'équation,* $y = ax + b$ *où* $a = \frac{s_{xy}}{s_x^2}$ *et* $b = \overline{y} - a\overline{x}$.
   Donc $a \simeq 0.1434$ *et* $b \simeq -1.0489$. *d'où*

$$y = 0.1434 \times x - 1.0489.$$

4. *si* $x = 11$, *alors* $y \simeq 0.5285$ *soit environ* 5285 DH *comme montant épargné par cette famille.*

Exercices supplémentaires

Exercice 12. Les téléspectateurs sont invités à évaluer une émission en envoyant un message contenant l'une des lettres A, B, C ou D qui représentent respectivement "très bonne émission", "bonne émission", "mauvaise émission" et "très mauvaise émission" ; çi après les évaluations de 32 spectateurs :

B, B, A, C, A, D, A, A, B, C, D, D, C, A, B, B, C, A, D, C, A, A, B, A, C, D, B, B, C, D, B, A

1) Caractériser la variable.
2) Dresser le tableau de distribution des effectifs et des fréquences.
3) Tracer une représentation graphique associée.

Exercice 13. Les durées, en minutes et arrondies à l'entier le plus proche, enregistrées pour 22 communications, dans un centre d'appel, sont données dans le tableau suivant :

10	12	14	14	15	15	16	16	17	17	17
18	18	18	19	19	20	20	21	22	23	24

1) Établir le tableau des fréquences et fréquences cumulées, en utilisant cinq classes.
2) Construire l'histogramme de cette série statistique.
3) Représenter graphiquement la courbe cumulée croissante.
4) A quelle valeur sont inférieures 25% des durées observées ?
5) Calculer la moyenne et l'écart-type de cette distribution.
6) Calculer le coefficient de variation. Cette distribution statistique est-elle homogène ?

Solution –

1) *l'amplitude commune à ces 5 classes est donnée par* $a \sim \frac{e}{k} = \frac{24-10}{5} = 2,8$ *(qu'on peut arrondir à 3)*

où $k = 1 + 3,322 log_{10} 22 = 1 + 3,322 \times 1,342 = 5,459$ *qu'on arrondit à 5 et on construit donc les classes. On obtient le tableau de fréquences suivant :*

$Classe_i$	n_i	f_i	F_i
$[10, 13[$	2	0.09	0.09
$[13, 16[$	4	0.18	0.27
$[16, 19[$	8	0.36	0.63
$[19, 22[$	5	0.23	0.86
$[22, 25[$	3	0.14	1
$Total$	22	1	1

2) *On peut tracer soit l'histogramme des effectifs ou des fréquences ;*

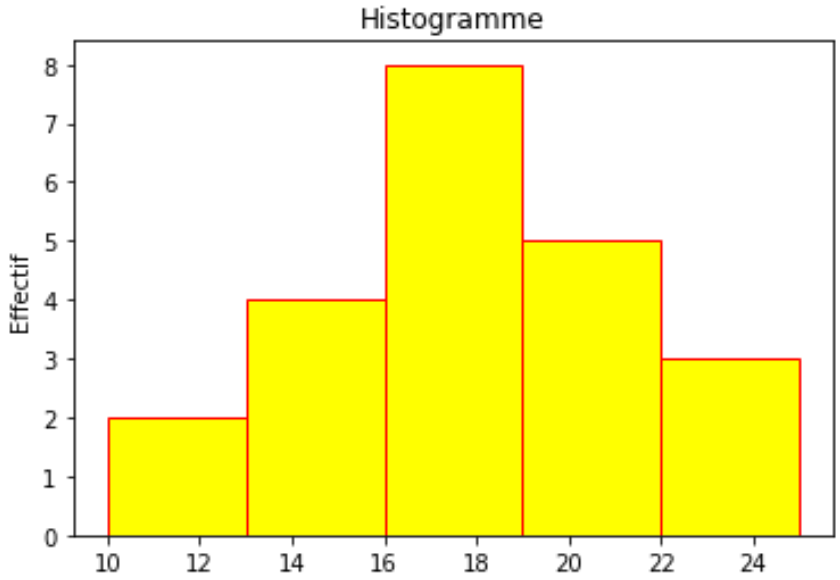


FIGURE 4 – Histogramme des effectifs

3) *Les points particuliers de la courbe cumulée croissante sont :*

Absc	10	13	16	19	22	25
Ord (en %)	0	9	27	63	56	100

La fonction de répartition (fonction cumulée croissante) est représentée comme suit :

4) *La fonction cumulée* F *étant croissante on a* $0.09 \leq 0.25 \leq 0.27$

Soit $F(13) \leq F(x) \leq F(16)$ *donc* $13 \leq x \leq 16$, *en notant* x *la valeur cherchée. Par interpolation linéaire :*

$$\frac{x-13}{16-13} = \frac{0.25-0.09}{0.27-0.09} \Rightarrow x = 13 + (16-13) \times \frac{0.16}{0.18} \approx 15.67 \text{ minutes}$$

Donc 25% *des durées sont inférieurs à* $15.67mm$

5) *La moyenne de ces durées est donnée par :*

$$\bar{x} = \frac{1}{22} \sum_{i=1}^{22} x_i = \frac{1}{22}(10 + 12 + ... + 23 + 24) = \frac{385}{22} = 17.5 \text{ minutes.}$$

et la variance $s^2 = \frac{1}{22} \sum_{i=1}^{22} x_i^2 - (17.5)^2$ *où* $\sum_{i=1}^{22} x_i^2 = 6989$

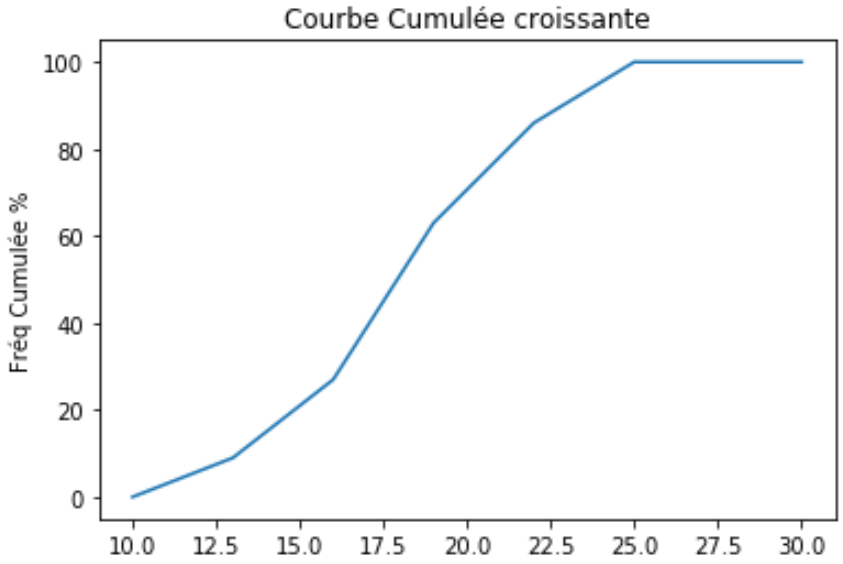


FIGURE 5 – Fonction cumulée croissante

Donc $s^2 \approx 11.43$ *minutes*^{2}

d'où l'écart-type $s = \sqrt{11.43} \approx 3.38$ *minutes.*

6) Le coefficient, noté CV, est défini par :

$CV = \frac{s}{\bar{x}}$. *Dans notre cas* $CV = \frac{3.38}{17.5} \approx 0.1931 = 19.31\%$

Cette distribution statistique est homogène.

Exercice 14. Dans une usine spécialisée dans la conception d'appareils médicaux, la répartition des employés classés d'après leur âge, en années, est présentée dans le tableau suivant :

Age (en année)	$[20, 25[$	$[25, 30[$	$[30, 35[$	$[35; 40[$	$[40; 45[$	$[45; 50[$	$[50; 55[$
Nombre d'employés	18	54	72	84	36	22	14

1) Quelle est la nature de cette variable ?
2) Construire le tableau des fréquences.
3) Calculer l'âge médian des employés de l'usine.
4) Sachant que 70% de ces employés dépasse un âge x, en années, quelle est alors la valeur de x ?
5) Quelle est la moyenne et l'écart-type de cette distribution statistique.
6) Cette distribution est-elle asymétrique ? justifier votre réponse.

Exercice 15. L'analyse du taux de calcium X du sérum humain sur 100 personnes a donné les résultats suivants :

Taux de calcium en mg/l	$[430, 440[$	$[440, 450[$	$[450, 460[$	$[460, 470[$	$[470, 480[$	$[480, 490]$
effectif	11	25	35	19	7	3

1) Dresser le tableau des fréquences et des fréquences cumulées.
2) Construire l'histogramme et le polygone des fréquences.
3) Déterminer la médiane M_e.
4) Calculer le taux de calcium moyen $\bar{x}$ et l'écart type s.
5) En déduire le mode M_o à partir de la relation suivante : $\bar{x} - M_o = 3(\bar{x} - M_e)$.

Exercice 16. L'objectif de cet exercice est d'étudier le degré d'un certain type de polluants X, Sur 400 endroits différents, on a mesuré le degré (en $10ppm$) de ce polluant. Les résultats sont regroupés dans le tableau suivant :

$Classe_i$	c_i	n_i	N_i	f_i	F_i
$[0, 10[$			55		
$[10, 20[$			120		
$[20, 30[$			220		
$[30, 40[$			300		
$[40, 50[$			360		
$[50, 60[$			400		
$\mathbf{\Sigma}$	–		–		–

– c_i : centre de la classe i – n_i : effectif – N_i : effectif cumulé – f_i : fréquence – F_i : fréquence cumulée.

1) Préciser la variable étudiée sa nature.

2) Recopier et compléter le tableau.

3) Construire l'histogramme de la distribution.

4) Quel est le degré modal de ce type de polluants ?

5) Calculer la moyenne et la variance. En déduire l'écart-type.

6) Tracer la courbe des fréquences cumulées et déterminer graphiquement les quantiles (Q_1, Me, Q_3)

7) Calculer la médiane et les quartiles Q_1 et Q_3

8) Calculer le coefficient de variation et conclure ?

9) Les responsable sanitaires ont décidé de soumettre indépendamment les 400 endroits à un type de désinfectants Y qui permettra de dégrader le degré du polluant. Sachant que la moyenne de Y vaut $\overline{y} = -5.25 \;\; 10ppm$ et sa variance est égale à $S_Y^2 = 7.5 \;\; (10ppm)^2$, calculer la moyenne et l'écart-type de la nouvelle variable $Z = X + 2Y$.

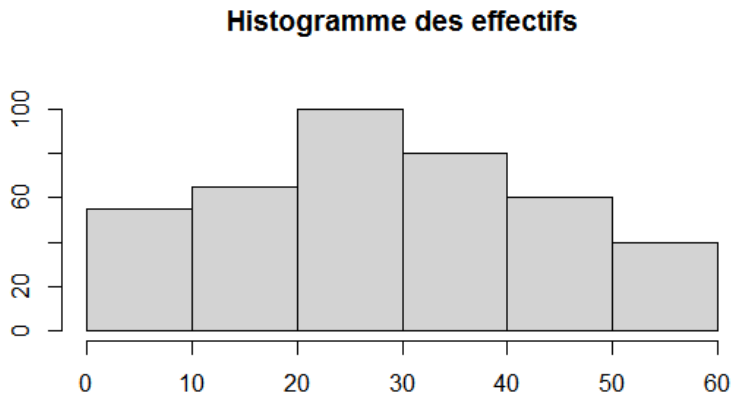
Solution –

1. *Préciser la variable étudiée sa nature :*
   On note par X la variable représentant ***"le degré d'influence d'un certain type de polluants"****.*
   Il s'agit d'une variable ***quantitative continue****.*

2. *Recopier et compléter le tableau :*

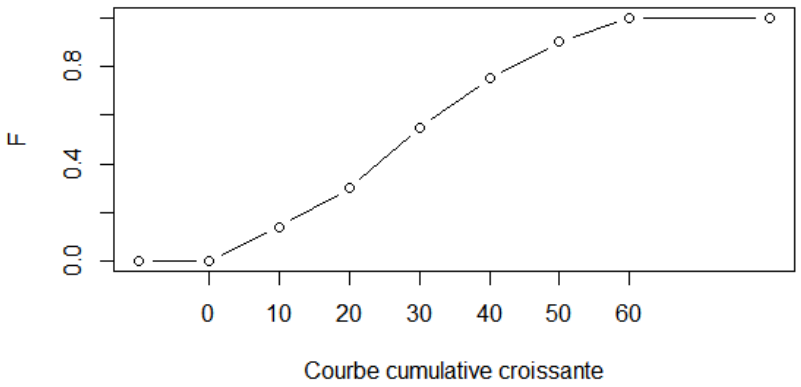
$Classe_i$	c_i	n_i	N_i	f_i	F_i
$[0, 10[$	5	55	55	0.1375	0.1375
$[10, 20[$	15	65	120	0.1625	0.3
$[20, 30[$	25	100	220	0.25	0.55
$[30, 40[$	35	80	300	0.2	0.75
$[40, 50[$	45	60	360	0.15	0.9
$[50, 60[$	55	40	400	0.1	1
$\mathbf{\Sigma}$	–	400	–	1	–

3. *Construire l'histogramme de la distribution : Histogramme des effectifs*



4. *Quel est le degré modal de ce type de polluants ?* ***Mo = 25***

5. *Calculer la moyenne et la variance. En déduire l'écart-type.*
   $\overline{x} = 28.625$,
   $s^2 = \overline{x^2} - (\overline{x})^2 = 1047.5 - (28.625)^2 = 228.1094$
   $s = \sqrt{228.1094} = 15.10329$
6. *Tracer la courbe des fréquences cumulées et déterminer graphiquement les quantiles* (D_1, Q_1, Me, Q_3, D_9) *:*



7. *Calculer la médiane et les quartiles* Q_1 *et* Q_3 *:*
   Directement $Q_3 = 40$.
   Par interpolation on trouve :
   $M_e = 20 + (30 - 20) \times \frac{(0.5-0.3)}{(0.55-0.3)} = 28$
   $Q_1 = 10 + (20 - 10) \times \frac{(0.25-0.1375)}{(0.3-0.1375)} \simeq 16.923$
8. *Calculer le coefficient de variation et conclure ?*
   $CV = \frac{s}{\overline{x}} = \frac{15.10329}{28.625} = 0.5276259 \simeq 52.76\%$
9. *Les responsable sanitaires ont décidé de soumettre indépendamment les* 400 *endroits à un type de désinfectants* Y *qui permettra de dégrader le degré du polluant. Les caractéristiques de* Y *sont telle que :* $\overline{Y} = -5.25$ *et* $\mathbb{V}(Y) = 7.5$.
   Calculer la moyenne et l'écart-type de la nouvelle variable $Z = X + 2Y$.
   $\overline{z} = \overline{x + 2y} = \overline{x} + 2\overline{y} = 28.625 - 2 \times 5.25 = 18.125$
   $s_Z^2 = s_{X+2Y}^2 = s_X^2 + 2^2 s_Y^2 = 228.1094 + 4 \times 7.5 = 258.1094$
   $s_Z \simeq 16.0658$

Exercice 17. Dans le but d'évaluer la relation entre la *densité des grains semés* et le *rendement*, on a procédé à une série d'essais sur différentes parcelles d'une céréale. L'expérimentation a donné les résultats suivants :

x_i	150	250	350	450
z_i	57.06	60.73	62.73	63.48
y_i				

avec, x_i désigne le *nombre de grain semés* par m^2 et z_i désigne le *rendement* par hectare.

1) Calculer les nombres $y_i = \ln(64 - z_i)$, pour $i = 1, 2, 3, 4$.

2) Calculer le coefficient de corrélation linéaire entre x et y. Interpréter.

3) Déterminer l'équation de la droite de régression de y en x. En déduire une expression de z en fonction de x.