Ecole Supérieure
d'Informatique et du Numérique
COLLEGE OF ENGINEERING & ARCHITECTURE

COMPTE RENDU

Analyse de données - TP2 Summarizing Data
3e année Cybersécurité - École Supérieure d'Informatique et du Numérique (ESIN)
Collège d'Ingénierie & d'Architecture (CIA)

Étudiant : HATHOUTI Mohammed Taha

Filière : Cybersecurité

Année : 2025/2026

Enseignants : M.MOUKAFIH

Date : 29 septembre 2025

1 Exercice 1 : Exploring the Students Performance Dataset

1.1 Q1 : Analyse de l'histogramme des notes en mathématiques

a) Estimation de la médiane

Pour estimer la médiane, on cherche la valeur qui sépare la distribution en deux parties égales. Nous avons 1000 étudiants, la médiane sera la 500ème valeur. On calcule les fréquences cumulées jusqu'à atteindre 0.5 (soit 500 étudiants sur 1000).

Intervalle	**Fréquence relative**	**Fréquence. cumulée**
[0,10]	0.002	0.002
[10,20]	0.002	0.004
[20,30]	0.01	0.014
[30,40]	0.026	0.040
[40,50]	0.095	0.135
[50,60]	0.188	0.323
[60,70]	**0.268**	**0.591**
[70,80]	0.216	0.807
[80,90]	0.135	0.942
[90,100]	0.058	1

La médiane se situe dans l'intervalle [60,70].

$$\text{Médiane} \approx \boxed{65} \tag{1}$$

b) Estimation de Q1, Q3 et IQR

Premier quartile (Q1) : On cherche la valeur pour laquelle la fréquence cumulée atteint 0.25.

Intervalle	**Fréquence relative**	**Fréquence. cumulée**
[0,10]	0.002	0.002
[10,20]	0.002	0.004
[20,30]	0.01	0.014
[40,50]	0.095	0.135
[50,60]	**0.188**	**0.323**
[...,...]	...	...

$$Q_1 \approx \boxed{55} \tag{2}$$

Troisième quartile (Q3) : On cherche la valeur pour laquelle la fréquence cumulée atteint 0.75.

Intervalle	**Fréquence relative**	**Fréquence. cumulée**
[0,10]	0.002	0.002
[10,20]	0.002	0.004
[20,30]	0.01	0.014
[30,40]	0.026	0.040
[40,50]	0.095	0.135
[50,60]	0.188	0.323
[60,70]	0.268	0.591
[70,80]	**0.216**	**0.807**
[...,...]	...	...

$$Q_3 \approx \boxed{75} \tag{3}$$

Intervalle interquartile (IQR) :

$$IQR = Q_3 - Q_1 = 75 - 55 = \boxed{20} \tag{4}$$

c) Estimation de la moyenne
La moyenne est calculée comme l'espérance mathématique :

$$\bar{X} = E[X] = \sum_i x_i \cdot P(X = x_i) \tag{5}$$

où x_i est le centre de chaque intervalle et $P(X = x_i)$ est la fréquence correspondante.

Intervalle	**Centre (x_i)**	$P(X = x_i)$	$x_i \cdot P(X = x_i)$
[0,10]	5	0.002	0.010
[10,20]	15	0.002	0.030
[20,30]	25	0.010	0.250
[30,40]	35	0.026	0.910
[40,50]	45	0.095	4.275
[50,60]	55	0.188	10.340
[60,70]	65	0.268	17.420
[70,80]	75	0.216	16.200
[80,90]	85	0.135	11.475
[90,100]	95	0.058	5.51
		Total :	**66.42**

$$Moyenne = \boxed{66.42} \tag{6}$$

1.2 Q2 : Analyse du boxplot des scores en mathématiques

a) Extraction des statistiques des boxplots

Genre	Médiane	Q1	Q3	Lower Whisker	Upper Whisker
Femmes	74	≈ 64	≈ 82	≈ 37	100
Hommes	64	≈ 52	≈ 74	≈ 22	100

b) Présence d'anomalies

On remarque bien des anomalies dans le diagramme surtout chez les femmes.

c) Comparaison des distributions

On remarque que le 1er quartile chez les femmes est suppérieur à celui des hommes, idem pour le 3ème quartile et la médiane, on en deduit donc que les femmes ont de meilleurs notes en expression écrites que les hommes.

1.3 Q3 : Comparaison histogramme vs boxplot

Caractéristiques apparentes dans l'histogramme mais pas dans le boxplot :

1. Distribution et la répartition des notes (fréquences) ;
2. Analyse de la forme (left-skewed, right-skewed, symétrique, bimodale, etc) ;

Caractéristiques apparentes dans le boxplot mais pas dans l'histogramme :

1. Identification claire des anomalies (ouliers) ;
2. Statistiques de positions (1er quartile, 3ème quartile, médiane, moustaches) ;

2 Exercice 2 : Infant Mortality

2.1 Q1 : Différence entre histogramme de fréquence et histogramme de fréquence relative

1. Un histogramme de fréquence : Affiche le nombre absolu d'observations dans chacun des intervalles (ex : histogramme de l'exercice 1) ;

2. Un histogramme de fréquence relative : Affiche la proportion ou le pourcentage d'observations dans chacun des intervalles (ex : histogramme de l'exercice courant 2) ;

3. La somme des fréquences est un entier tandis que la somme des fréquences relatives = 1 (ou 100%) ;

2.2 Q2 : Comparaison moyenne vs médiane

L'histograme montre une distribution de la mortalité infantile asymétrique à droite (right-skewed) avec une majorité des pays ayant un taux de mortalité infantile très faible (à gauche) et quelques pays ayant un taux de mortalité infantile très haut.

La moyenne est sensible aux valeurs extrêmes ce qui fait que elle sera vite influencées par la minorité des pays ayant un taux de mortalité infantile très haut. Tandis que la moyenne ne sera pas affectée par ces valeurs.

Dans une distribution right-skewed :

$$Moyenne > \text{Médiane} \tag{7}$$

3 Exercice 3 : Stats Scores

Données : 78, 81, 94, 81, 73, 72, 69, 66, 57, 71, 89, 88, 82, 83, 83, 77, 78, 74, 79, 79

3.1 Q1 : Calcul des statistiques

Calcul de la moyenne :

$$Moyenne = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{20} \times (1554) = \boxed{77.7} \tag{8}$$

Calcul de la médiane :
Données triées : 57, 66, 69, 71, 72, 73, 74, 77, 78, 78, 79, 79, 81, 91, 82, 83, 83, 88, 89, 94

Pour $n = 20$ (pair), la médiane est la moyenne des deux valeurs centrales (positions 10 et 11).

$$\text{Médiane} = \frac{x_{10} + x_{11}}{2} = \frac{78 + 79}{2} = \boxed{78.5} \tag{9}$$

Premier quartile (Q1) :

$$\text{Q1} = \frac{x_5 + x_6}{2} = \frac{72 + 73}{2} = \boxed{72.5} \tag{10}$$

$$Q_1 = \boxed{72.5} \tag{11}$$

Troisième quartile (Q3) :

$$\text{Q3} = \frac{x_{15} + x_{16}}{2} = \frac{82 + 3}{2} = \boxed{82.5} \tag{12}$$

$$Q_3 = \boxed{82.5} \tag{13}$$

Lower Whisker :

$$\text{Lower Whisker} = Q1 - 1.5 \times IQR = 72.5 - 1.5 \times 10 = 72.5 - 15 = \boxed{57.5} \tag{14}$$

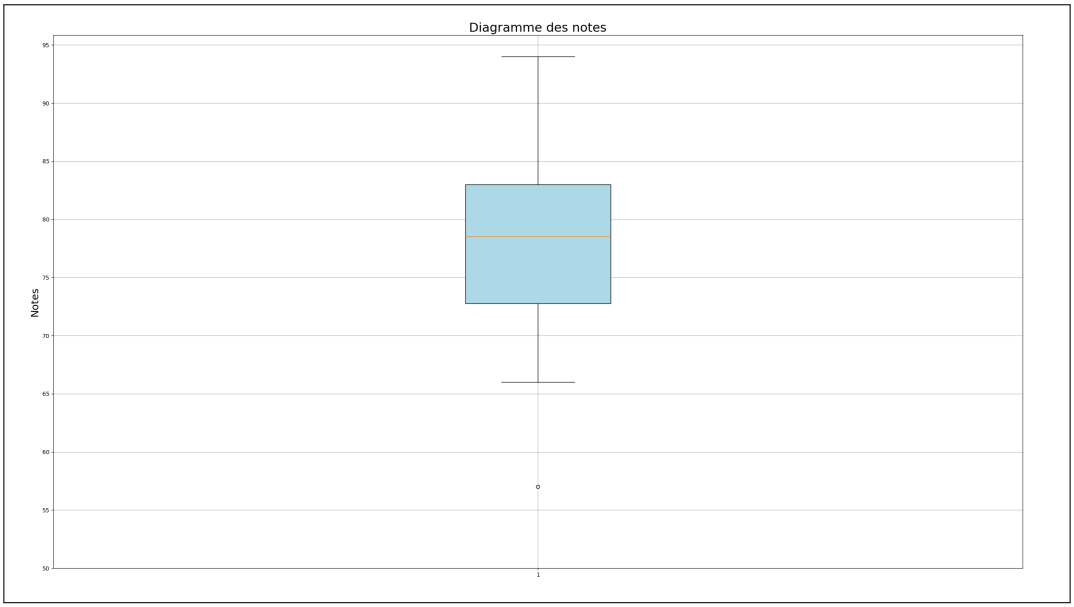
$$\text{Lower Whisker} = \boxed{57.5} \tag{15}$$

Upper Whisker :

$$\text{Lower Whisker} = Q3 + 1.5 \times IQR = 82.5 + 1.5 \times 10 = 82.5 + 15 = \boxed{97.5} \tag{16}$$

$$\text{Upper Whisker} = \boxed{97.5} \tag{17}$$

3.2 Q2 : Boxplot de la distribution



3.3 Q3 : Impact de l'ajout d'un score de 28

Nouvelle moyenne :

$$Moyenne_{\text{nouvelle}} = \frac{1}{21}\sum_{i=1}^{21} x_i = \frac{1}{21} \times (1582) \approx \boxed{75.33} \tag{18}$$

Variation de la moyenne :

$$\Delta Moyenne = Moyenne_{\text{nouvelle}} - Moyenne_{\text{ancienne}} \approx \boxed{-2.37} \tag{19}$$

Nouvelle médiane :

Pour $n = 21$ (impair), la médiane est la valeur en position 11.

$$\text{Médiane}_{\text{nouvelle}} = \boxed{78} \tag{20}$$

Variation de la médiane :

$$\Delta\text{Médiane} = \text{Médiane}_{\text{nouvelle}} - \text{Médiane}_{\text{ancienne}} = \boxed{-0.5} \tag{21}$$

Analyse :

La moyenne peut être affectée par des anomalies (-2.37) tandis que rajouter une note ne modifie pas voir affecte très peu la médiane.