TD N° 1

Exercice 1

Les Statistiques

Les statistiques en data mining nous aident à **mieux connaître la base de données** et interviennent dans le **prétraitement**.

On a comme données sur l'attribut *âge* :
13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

Remarque Sur Les Données

- Un seul attribut *âge*.
- Les données sont croissantes ce qui facilite les calcules.

Données à plusieurs attributs

- On a tendance à comparer les attributs d'une même base de données afin de découvrir des relations entre eux (des connaissances).
- Pour visualiser une base de données comportant plusieurs attributs (dimensions), on peut utiliser des graphiques comme les *scatter plots* (entre deux attributs) ou
des cartes thermiques *heatmaps* (plusieurs attributs) pour représenter les relations entre les variables.

1. Formule de la moyenne :

$$\boxed{\overline{X} = \frac{\sum x_i}{N}} \quad \text{où } N \text{ est l'effectif, et } x_i \text{ la } i^{\text{ème}} \text{ valeur.}$$

$$\begin{aligned}\overline{x} &= \frac{13 + 15 + 16 \cdot 2 + 19 + 20 \cdot 2 + 21 + 22 \cdot 2 + 25 \cdot 4 + 30 + 33 \cdot 2 + 35 \cdot 4 + 36 + 40 + 45 + 46 + 52 + 70}{27} \\ &= \frac{869}{27} \\ &= \boxed{29.96 \ ans}\end{aligned}$$

Formule de la mediane

$$\text{Med}(X) = \begin{cases} X\left[\frac{N+1}{2}\right] & \text{si } N \text{ est impair,} \\ \frac{X\left[\frac{N}{2}\right] + X\left[\frac{N}{2}+1\right]}{2} & \text{si } N \text{ est pair.} \end{cases}$$

Puisque N est impair $= 27$:

$$\begin{aligned}\text{Med}(X) &= X\left[\frac{27+1}{2}\right] \\ &= X[14] \\ &= \boxed{25 \ ans}\end{aligned}$$

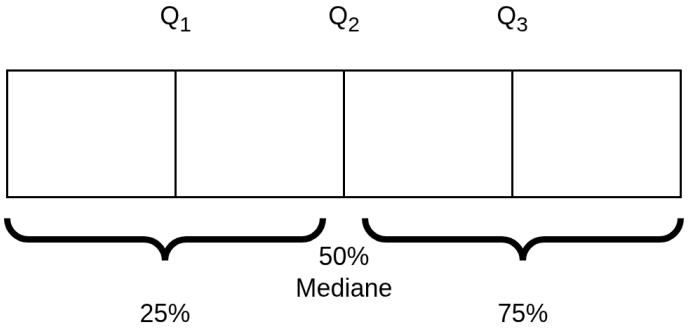
2. Le *mode* représente la ou les valeurs dont la fréquence est la plus élevée (le nombre de répétitions).
   Le *type de modalité* fait référence au nombre de valeurs distinctes correspondant au mode (*bimodale* : 2 valeurs, *trimodale* : 3 valeurs, etc.).

Valeur	Frequence
13	1
16	2
19	1
20	2
21	1
22	2
25	**4**
30	1
33	2
35	**4**
36	1
40	1
45	1
46	1
52	1
70	1

   On remarque que les valeurs 25 et 35 ont la plus haute fréquence, répétées 4 fois :

   Donc, le *mode* est 25 et 35, et la modalité est dite *bimodale*.

3. Les quartiles permettent de diviser nos données en quatre parties égales :



   Les formules :

$$Q_1 = \begin{cases} X\left[\dfrac{N+1}{4}\right] & \text{si } N \text{ est impair,} \\ X\left[\dfrac{N}{4}\right] & \text{si } N \text{ est pair} \end{cases}$$

$$Q_2 = \text{Med}(X)$$

$$Q_3 = \begin{cases} X\left[\dfrac{3\cdot(N+1)}{4}\right] & \text{si } N \text{ est impair,} \\ X\left[\dfrac{3\cdot N}{4}\right] & \text{si } N \text{ est pair} \end{cases}$$

Puisque N est impaire $= 27$:

$$\begin{aligned} Q_1 &= X\left[\frac{27+1}{4}\right] \\ &= X[7] \\ &= \boxed{20\ ans} \end{aligned}$$

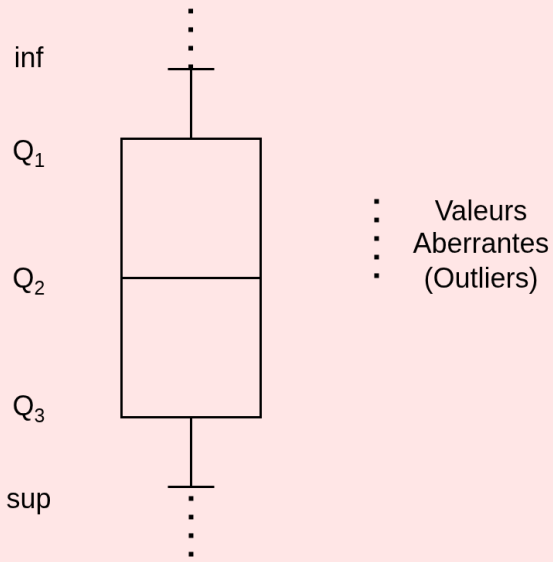
$$\begin{aligned} Q_3 &= X\left[\frac{3\cdot(27+1)}{4}\right] \\ &= X[21] \\ &= \boxed{35\ ans} \end{aligned}$$

4. les cinq nombre de donne sont definie par : min Q_1 Q_2 Q_3 max

$$\min = 13\ ans\ ,\ Q_1 = 20\ ans\ ,\ Q_2 = 25\ ans\ ,\ Q_3 = 35\ ans$$

5. boxPlot

BoxPlot



- Elle nous permet de repérer les valeurs aberrantes.
- La médiane n'est pas forcément située au milieu de la boxplot.
- Les formules :

$$\boxed{\lim_{inf} = Q_1 - (1.5 \times EIQ)}$$

$$\boxed{\lim_{sup} = Q_3 + (1.5 \times EIQ)}$$

$$\boxed{EIQ(\text{Écart interquartile}) = Q_3 - Q_1}$$

- Si $\lim_{inf}$ ou $\lim_{sup}$ ont des valeurs incohérentes ou trop différentes du min ou max, alors on prendra comme limites min et max respectivement.

On a :

$$\begin{cases} X_{min} & = 13\ ans \\ Q_1 & = 20\ ans \\ Q_2 & = 25\ ans \\ Q_3 & = 35\ ans \\ X_{max} & = 70\ ans \end{cases}$$

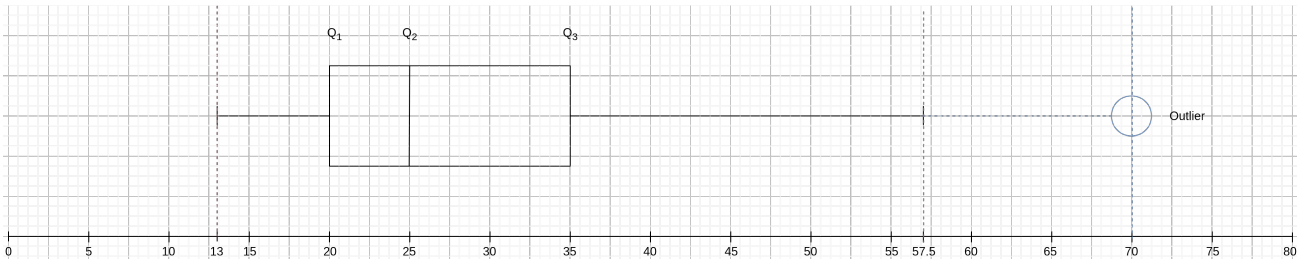
On calcule les limites :

$$EIQ = Q_3 - Q_1 = 35 - 20 = \boxed{15\ ans}$$
$$\lim_{inf} = Q_1 - (1.5 \times EIQ) = 20 - (1.5 \times 15) = \boxed{-2.5\ ans}$$
$$\lim_{sup} = Q_3 + (1.5 \times EIQ) = 35 + (1.5 \times 15) = \boxed{57.5\ ans}$$

Limites

- On a $\lim_{inf} = -2.5$ ans, ce qui est une valeur incohérente dans le contexte âge $\Rightarrow$ on prend comme limite le min = 13 ans.
- On a $\lim_{sup} = 57$ ans, valeur normale et pas loin du max $\Rightarrow$ on la prend comme limite.

On prend comme unité :
8 carreaux $\rightarrow$ 5 ans



Conclusion

70 est un outlier

Exercice 2

On dispose des données de l'âge et du taux de graisse de 18 adultes dans un hôpital, sélectionnés au hasard :

Âge	23	23	27	27	39	41	47	49	50
Fat %	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
Âge	52	54	54	56	57	58	58	60	61
Fat %	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

Remarques sur les données

- Deux attributs : *âge* et *graisse*.
- Les données sur l'*âge* sont croissantes, ce qui facilite les calculs.
- Les données sur la *graisse* ne sont pas croissantes, donc on doit les ordonner.

1. Formule de l'écart-type :

$$\sigma = \begin{cases} \sqrt{\dfrac{\sum(x_i - \overline{x})^2}{N-1}} & \text{si } N \text{ représente un échantillon} \\ \sqrt{\dfrac{\sum(x_i - \overline{x})^2}{N}} & \text{si } N \text{ représente la population complète} \end{cases}$$

Âge :

$$\begin{aligned} \overline{x} &= \frac{23 \cdot 2 + 27 \cdot 2 + 39 + 40 + 49 + 50 + 52 + 54 \cdot 2 + 56 + 57 + 58 + 60 + 61}{18} \\ &= \frac{836}{18} \\ &= \boxed{46.44 \ ans} \end{aligned}$$

Puisque N est pair $= 18$:

$$\begin{aligned} \text{Med}(X) &= \frac{X\left[\frac{18}{2}\right] + X\left[\frac{18}{2} + 1\right]}{2} \\ &= \frac{X[9] + X[10]}{2} \\ &= \frac{50 + 52}{2} \\ &= \boxed{51 \ ans} \end{aligned}$$

x_i	23	27	39	41	47	49	50
$(x_i - \overline{x})^2$	549.433	377.913	55.353	29.593	0.313	6.553	12.673
x_i	52	54	56	57	58	60	61
$(x_i - \overline{x})^2$	30.913	57.153	91.393	111.513	133.633	183.873	211.993

Puisque N représente un échantillon :

$$\begin{aligned} \sigma &= \sqrt{\frac{\sum(x_i - \overline{x})^2}{18-1}} \\ &= \sqrt{\frac{2970.434}{17}} \\ &= \boxed{13.218 \ ans} \end{aligned}$$

Graisse :

7.8	9.5	17.8	25.9	26.5	27.2	27.4	28.8	30.2
31.2	31.4	32.9	33.4	34.1	34.6	35.7	41.2	42.5

$$\begin{aligned}\overline{y} &= \frac{\substack{7.8+9.5+17.8+25.9+26.5+27.2+27.4+28.8+30.2+31.2+31.4+32.9+33.4+34.1+34.6+35.7\\+41.2+42.5}}{18}\\ &= \frac{518.1}{18}\\ &= \boxed{28.78\ \%}\end{aligned}$$

Puisque N est pair $= 18$:

$$\begin{aligned}\text{Med}(Y) &= \frac{Y\left[\frac{18}{2}\right] + Y\left[\frac{18}{2}+1\right]}{2}\\ &= \frac{Y[9]+Y[10]}{2}\\ &= \frac{30.2+31.2}{2}\\ &= \boxed{30.7\ \%}\end{aligned}$$

y_i	7.8	9.5	17.8	25.9	26.5	27.2	27.4	28.8	30.2
$(y_i-\overline{y})^2$	440.160	371.718	120.560	8.294	5.198	2.496	1.904	0.0001	2.016
y_i	31.2	31.4	32.9	33.4	34.1	34.6	35.7	41.2	42.5
$(y_i-\overline{y})^2$	5.856	6.864	16.974	21.344	28.302	33.872	47.886	154.256	188.238

Puisque N représente un échantillon :

$$\begin{aligned}\sigma &= \sqrt{\frac{\sum(y_i-\overline{y})^2}{18-1}}\\ &= \sqrt{\frac{1455.938}{17}}\\ &= \boxed{9.254\ \%}\end{aligned}$$

2. boxPlots :

   Âge

   On a :

$$\begin{cases} X_{min} & = 23\ ans \\ Q_1 & = X\left[\frac{N}{4}\right] = X\left[\frac{18}{4}\right] = X[4.5] = X[5] = 39\ ans \\ Q_2 & = 51\ ans \\ Q_3 & = X\left[\frac{3\cdot N}{4}\right] = X\left[\frac{3\cdot 18}{4}\right] = X[13.5] = X[14] = 57\ ans \\ X_{max} & = 61\ ans \end{cases}$$

   On calcule les limits :

$$EIQ = Q_3 - Q_1 = 57 - 39 = \boxed{18\ ans}$$

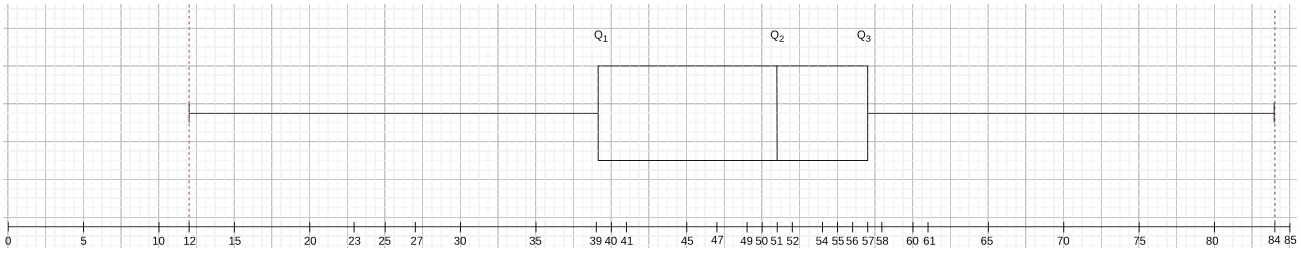
$$\text{lim}_{inf} = 39 - (1.5 \times EIQ) = 20 - (1.5 \times 18) = \boxed{12\ ans}$$

$$\text{lim}_{sup} = 57 + (1.5 \times EIQ) = 35 + (1.5 \times 18) = \boxed{84\ ans}$$

Limits

- On a $\lim_{inf} = 12$ ans, valeur normale et pas loin du min $\Rightarrow$ on la prend comme limite.
- On a $\lim_{sup} = 84$ ans, valeur normale et pas loin du max $\Rightarrow$ on la prend comme limite.

On prend comme unité :
8 carreaux $\rightarrow$ 5 ans



Conclusion

aucun outlier

Graisse
On a :

$$\begin{cases} Y_{min} & = 7.8\ \% \\ Q_1 & = Y\left[\frac{N}{4}\right] = Y\left[\frac{18}{4}\right] = Y[4.5] = Y[5] = 26.5\ \% \\ Q_2 & = 30.7\ \% \\ Q_3 & = Y\left[\frac{3 \cdot N}{4}\right] = Y\left[\frac{3 \cdot 18}{4}\right] = Y[13.5] = X[14] = 34.1\ \% \\ Y_{max} & = 42.5\ \% \end{cases}$$

On calcule les limits :

$$EIQ = Q_3 - Q_1 = 34.1 - 26.5 = \boxed{7.6\ \%}$$

$$\lim_{inf} = 26.5 - (1.5 \times EIQ) = 26.5 - (1.5 \times 7.6) = \boxed{15.1\ \%}$$

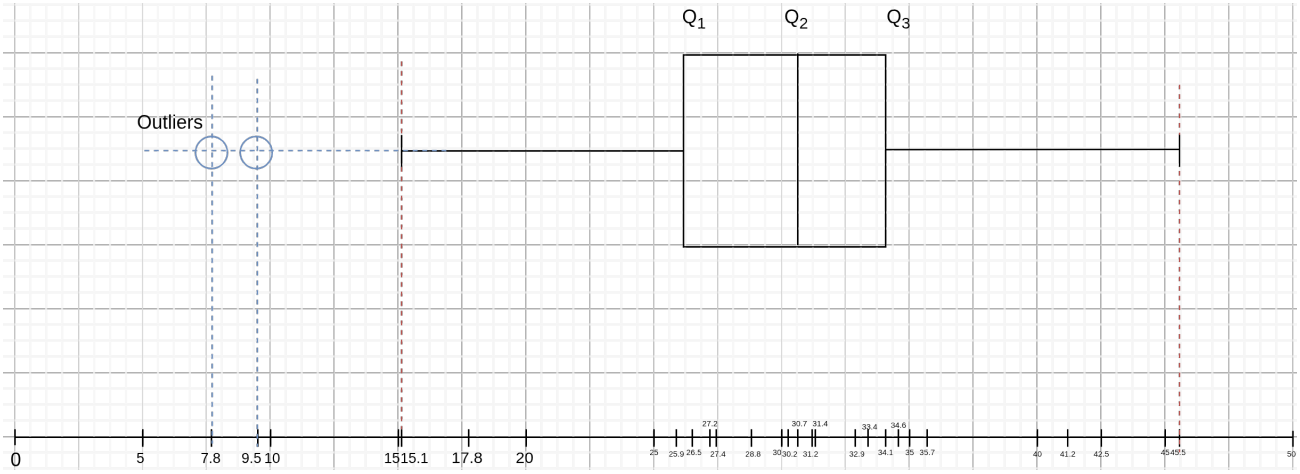
$$\lim_{sup} = 34.1 + (1.5 \times EIQ) = 34.1 + (1.5 \times 7.6) = \boxed{45.5\%}$$

Limits

- On a $\lim_{inf} = 15.1\ \%$, valeur normale et pas loin du min $\Rightarrow$ on la prend comme limite.
- On a $\lim_{sup} = 45.5\ \%$, valeur normale et pas loin du max $\Rightarrow$ on la prend comme limite.

On prend comme unité :

8 carreaux → 5 %



Conclusion

9.5 et 7.8 sont des outliers

3. scatter plot

Corrélation

La corrélation dans un scatter plot, c'est la relation entre l'attribut de l'axe X et l'axe Y :

- Corrélation positive : quand X augmente, Y augmente ; quand X diminue, Y diminue aussi.
- Corrélation négative : quand X augmente, Y diminue ; quand X diminue, Y augmente.
- Pas de corrélation : aucune relation.

4. Normalization Z-score:

Z-score

$$Z = \frac{x_i - \overline{x}}{\sigma}$$

Les valeurs sont entre $[0, 1]$. On normalise parce que les algorithmes (modèles) sont plus performants dans cet intervalle.

Âge

x_i	23	27	39	41	47	49	50
Z	-1.77	-1.47	-0.56	-0.41	0.04	0.19	0.26
x_i	52	54	56	57	58	60	61
Z	0.42	0.57	0.73	0.79	0.87	1.02	1.10

Graisse

y_i	7.8	9.5	17.8	25.9	26.5	27.2	27.4	28.8	30.2
Z	-2.26	-2.08	-1.18	-0.31	-0.24	-0.17	-0.14	0.002	0.15
y_i	31.2	31.4	32.9	33.4	34.1	34.6	35.7	41.2	42.5
Z	0.26	0.28	0.44	0.49	0.57	0.62	0.74	1.34	1.48

5. coefficient de corrélation:

$$r = \begin{cases} \dfrac{\sum(x_i - \overline{x}) \cdot (y_i - \overline{y})}{(N-1) \cdot (\sigma_x \cdot \sigma_y)} & \text{si } N \text{ représente un échantillon} \\ \dfrac{\sum(x_i - \overline{x}) \cdot (y_i - \overline{y})}{N \cdot (\sigma_x \cdot \sigma_y)} & \text{si } N \text{ représente la population complète} \end{cases}$$

Interprétation de $r \in [-1, 1]$:

$$\begin{cases} r = -1 & \text{corrélation négative linéaire parfaite} \\ -1 < r < 0 & \text{corrélation négative} \\ r = 0 & \text{aucune corrélation} \\ 0 < r < 1 & \text{corrélation positive} \\ r = 1 & \text{corrélation positive linéaire parfaite} \end{cases}$$