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 matrices de dispersion (*scatter plots*) (entre deux attributs) ou des cartes thermiques (*heatmaps*) (plusieurs attributs) pour représenter les relations entre les variables.

1. Formule de la moyenne :

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

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

Formule de la mediane

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

Puisque N est impair $= 27$:

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

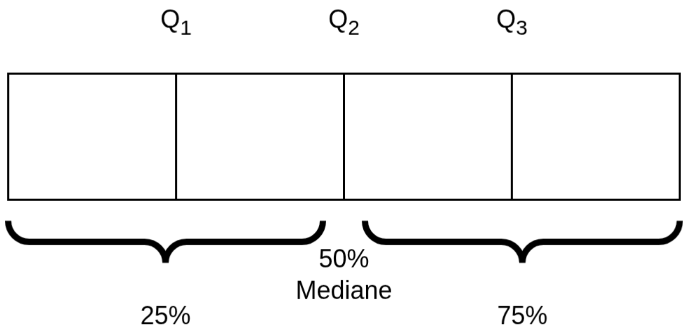
2. Le *mode* représente la ou les valeurs dont la fréquence est la plus élevée (le nombre de répétitions).
   Le *type de modalité* fait référence au nombre de valeurs distinctes correspondant au mode (*bimodale* : 2 valeurs, *trimodale* : 3 valeurs, etc.).

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 :

   $$\boxed{\text{Donc, le } \textit{mode} \text{ est 25 et 35, et la modalité est dite } \textit{bimodale.}}$$

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



   Les formules :

$$Q_1 = \begin{cases} X\left[\dfrac{N+1}{4}\right] & \text{si } N \text{ est impair,} \\[2ex] X\left[\dfrac{N}{4}\right] & \text{si } N \text{ est pair} \end{cases}$$

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

$$Q_3 = \begin{cases} X\left[\dfrac{3\cdot(N+1)}{4}\right] & \text{si } N \text{ est impair,} \\[2ex] X\left[\dfrac{3\cdot N}{4}\right] & \text{si } N \text{ est pair} \end{cases}$$

Puisque N est impaire $= 27$:

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

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

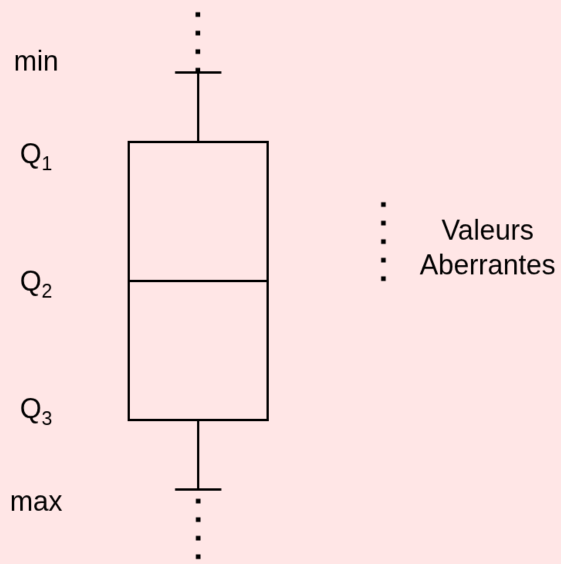
4. les cinq nombre de donne sont definie par : min Q_1 Q_2 Q_3 max
   le min ettant 13 *ans* et le max 70 *ans*.

Résumé

On remarque que les âges vont de 13 *ans* à 70 *ans*, donc les données peuvent provenir d'un collège, par exemple.

5. boxPlot

BoxPlot



Elle nous permet de repérer les valeurs aberrantes. Nous verrons le boxplot en détail prochainement.

6. L'interpretation de la BoxPlot on verra ca prochainment

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 $= 18$:

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

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
$(y_i-\overline{x})^2$	5.856	6.864	16.974	21.344	28.302	33.872	47.886	154.256	188.238

Puisque N représente un échantillon :

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