

Analyse exploratoire des données

PST

2 - Analyse exploratoire des données

Résumé du document

Definition

Table des matières

- 1. Statistiques élémentaires 2
 - 1.1. Tendances centrale 2
 - 1.1.1. Mauvaise représentation 2
 - 1.1.2. Moyenne 2
 - 1.1.3. Médiane 2
 - 1.1.4. Mode 2
 - 1.2. Indicateurs de dispersion 2
 - 1.2.1. Etendue 2
 - 1.2.2. Ecart-type 2
 - 1.2.3. Etendue interquartile 2
 - 1.2.4. Etendue interquartile 2
 - 1.2.5. Coefficient de variation 2
 - 1.3. Indicateurs de forme 3
- 2. Boîte à moustache 4

1. Statistiques élémentaires

1.1. Tendance centrale

Les indicateurs de tendance centrale informent sur le milieu d'une distribution grâce à la moyenne, la médiane et accessoirement le mode.

1.1.1. Mauvaise représentation

La moyenne et la médiane ne permettent pas de représenter précisément les détails des valeurs calculées.

Note: 4.5 4.5 4.5 → moyenne = 4.5 → médiane = 4.5 → écart-type = 0
 Note: 4 4.5 5 → moyenne = 4.5 → médiane = 4.5 → écart-type > 0

1.1.2. Moyenne

$$\bar{x} = \frac{1}{n} * \sum x_i$$

1.1.3. Médiane

$$\text{med}(x)$$

Valeur observée donc valeur représentant la moitié de l'ensemble de donnée.

1.1.4. Mode

Le mode aussi dit **valeur dominante** et consiste à trouver le/les valeurs les plus représentées dans notre échantillon.

1.2. Indicateurs de dispersion

Les indicateurs de dispersion informent sur la variabilité de la distribution grâce à l'étendue, l'écart-type, l'étendue interquartile et le coefficient de variation.

1.2.1. Etendue

L'étendue notée R et représente la plus grande valeur et la plus petite de l'échantillon.

$$x_n - x_1$$

1.2.2. Ecart-type

L'écart-type noté s d'une liste de données est la racine carrée de la somme des carrés des écarts à la moyenne divisée par $(n - 1)$:

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n - 1}}$$

1.2.3. Etendue interquartile

La médiane permet de diviser l'échantillon ordonné en deux parties de la même grandeur. Pour calculer les différents quartile nous feront:

- le **premier quartile** est noté $q(25\%)$ et sera la médiane des 50% des valeurs les plus petites
- le **second quartile** est la médiane
- le **troisième quartile** est noté $q(75\%)$ et sera la médiane des 50% des valeurs les plus grandes

1.2.4. Etendue interquartile

Un troisième indicateur de dispersion (le moins sensibles aux données dites atypiques) est l'**étendue interquartile** définie par:

$$q(75\%) - q(25\%)$$

1.2.5. Coefficient de variation

Le coefficient de variation, noté, CV est obtenu en divisant l'écart-type s par la moyenne arithmétique.

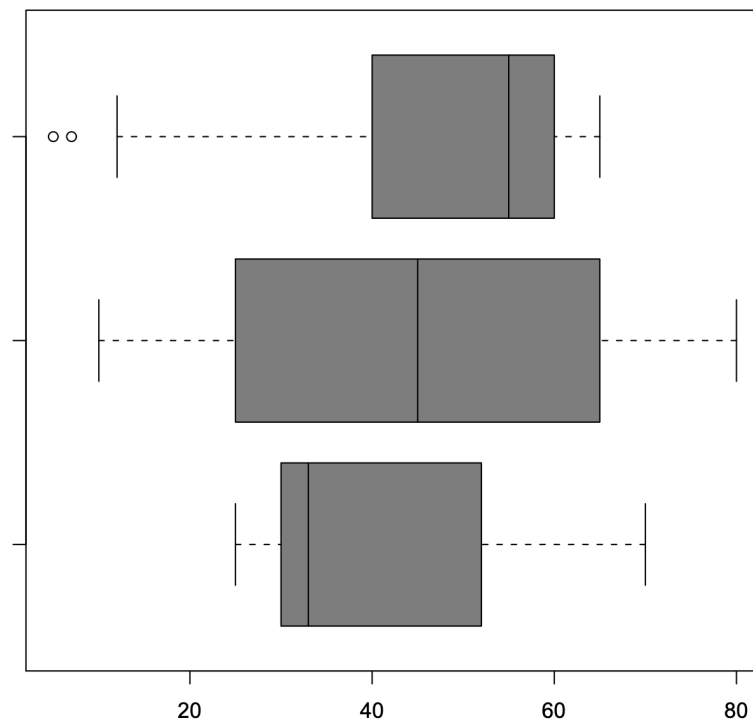
$$CV = \frac{s}{\bar{x}} \text{ avec } \bar{x} \neq 0$$

1.3. Indicateurs de forme

Les indicateurs de forme qui mesurent le degré d'asymétrie ou d'aplatissement d'une distribution grâce à l'asymétrie et l'aplatissement.

2. Boîte à moustache

La boîte à moustache permet de représenter visuellement les données des 3 quartiles (q_1 , q_2 , q_3).



Pour construire la boîte nous aurons:

- la barre gauche vaudra $q(25\%)$
- la barre centrale vaudra $\text{med}(x)$ donc la médiane
- la barre de droite vaudra $q(75\%)$

Pour construire les moustache nous devons:

1. mesurer la longueur de la boîte
2. multiplier cette longueur par 1.5
3. se placer à la distance calculée des deux extrémités de la boîte puis se rapprocher de la boîte jusqu'à tomber sur une valeur de l'ensemble de donnée