Analyse exploratoire des données

PST

2 - Analyse exploratoire des données

Résumé du document

Definition

Table des matières

1. Statistiques élémentaires 2
1.1. Tendance 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. Boite à moustache 4
3. Visualisation pour deux variables 5
3.1. Covariance 5
3.2. Coefficiant de corrélation 5

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 médianne ne permettent pas de représenter précisement les détails des valeurs calculée.

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

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

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 représentée 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.

$$x_{n \mid n} - x_{1 \mid n}$$

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 $(n-1)$:

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

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.

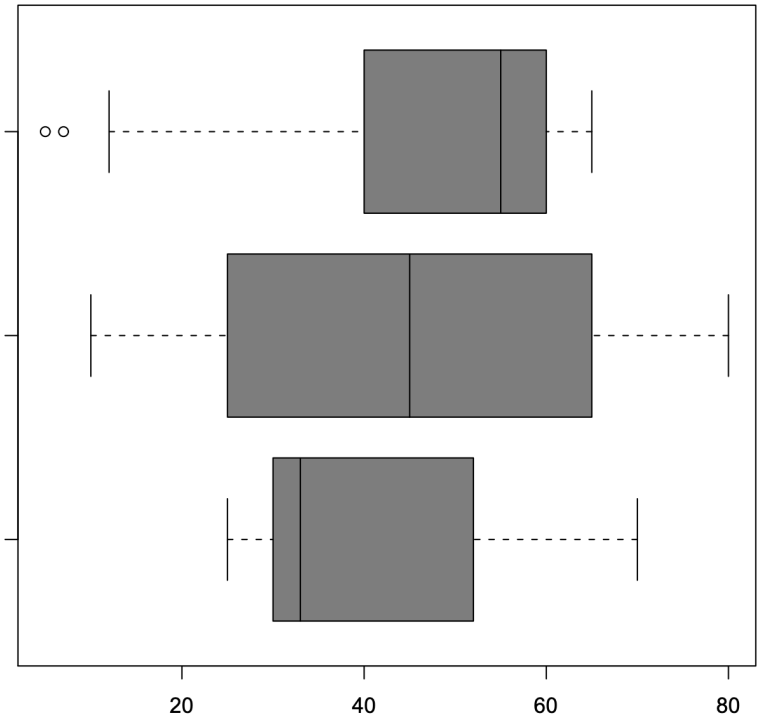
$$\text{CV} = \frac{s}{\overline{x}} \text{ avec } \overline{x} \neq 0$$

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. Boite à moustache

La boite à moustache permet de représenter visuellement les données des 3 quartiles ($q1, q2, q3$).



Pour construire la boite 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 longeur de la boite
2. multiplier cette longeur par 1.5
3. se placer à la distance calculée des deux extrémités de la boite puis se rapprocher de la boite jusqu'à tomber sur une valeur de l'ensemble de donnée

3. Visualisation pour deux variables

3.1. Covariance

En prenant 2 variables X, Y dont on dispose de n observations présentées sous la forme de couple de nombre:

$$\binom{x_1}{y_1}, \binom{x_2}{y_2}, ..., \binom{x_n}{y_n}$$

Pour quantifier le lien existatn entre deux variables, on utilise la **covariance** définie par:

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}) * (y_i - \overline{y})$$

avec $\overline{x}$ vaut la moyenne arithmétique de x et $\overline{y}$ vaut la moyenne arithmétique de y.

3.2. Coefficiant de corrélation

Le degré de dépendance linéaire existant entre deux grandeurs X, Y noté r est défini par

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x}) * (y_i - \overline{y})}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - \overline{x}^2\right) * \left(\sum_{i=1}^{n} y_i^2 - \overline{y}^2\right)}}$$