Thématique 4 - Analyse bivariée Croisement Qualitatif-Quantitatif

Data Sciences Revision

Mention Mathématiques et Informatique

Parcours HPDA

18/09/2024

Rédigé par :
PAULY Alexandre
alexandre.pauly@cy-tech.fr

L'objectif d'une analyse bivariée est de faire lien entre deux variables (ici qualitatives et quantitatives). Il est donc possible de faire le lien entre le salaire d'une personne et sa catégrie socio-professionnelle par exemple.

1 Indicateurs importants

Lors de la phase préparatoire à l'analyse des données, l'une des premières étapes, et probablement la plus importante avant de commencer, consiste à calculer certains indicateurs permettant de mettre en lumière la répartition du jeu de données : effectif, moyenne, variance, médiane, Q1 et Q3.

Ces indicateurs peuvent être restitués sur une boîte à moustache pour offrir une meilleure représentation d'une variable.

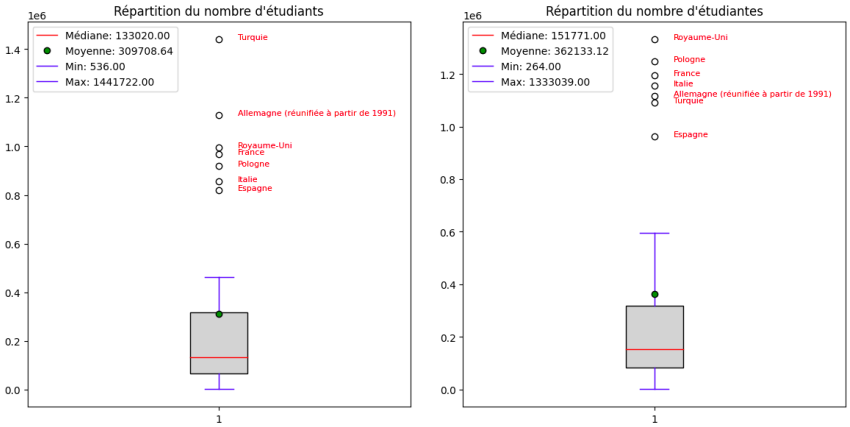

Figure 1: Boîtes à moustaches

Mais il est également intéressant de représenter sur un même graphique les différentes modalités d'une variable qualitative.

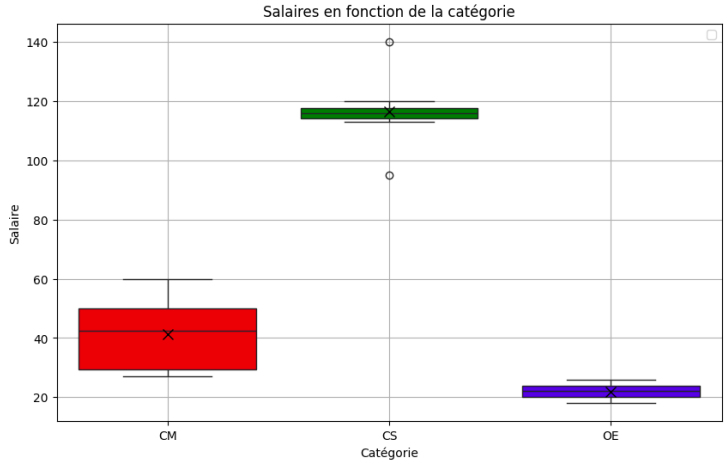

Figure 2: Boîtes à moustaches pour plusieurs modalités

Bien que moins précis, les diagrammes circulaires peuvent également montrer certaines répartitions dans leur globalité. Comme la part d'hommes et femmes au sein d'une population, etc.

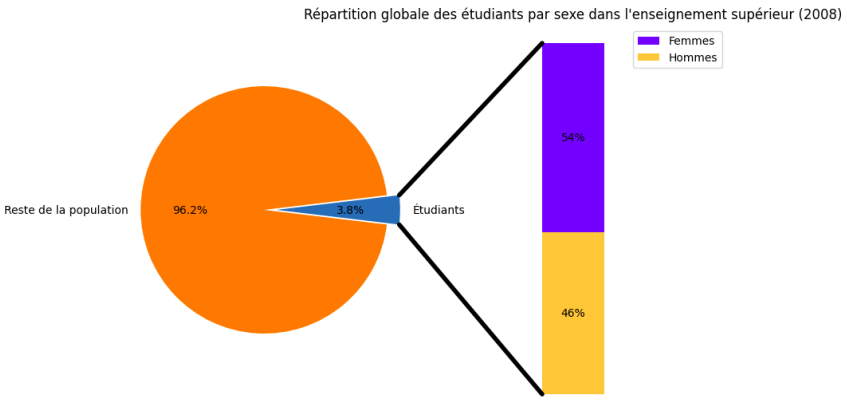

Figure 3: Diagramme circulaire

2 Décomposition de la variance

2.1 Définition

La décomposition de la variance est une technique statistique qui permet d'analyser comment la variance totale d'une variable dépend de plusieurs sources de variabilité et impacte les autres.

2.2 Formule générale

Soit Y une variable aléatoire. La variance de Y, notée $\mathrm{Var}(Y)$, peut être décomposée en deux parties :

$$\mathrm{Var}(Y) = \mathrm{Var}(E[Y|X]) + E[\mathrm{Var}(Y|X)]$$

où :

- $\mathrm{Var}(E[Y|X])$ est la variance expliquée par X (variance inter-groupes),
- $E[\mathrm{Var}(Y|X)]$ est la variance non expliquée ou résiduelle (variance intra-groupes).

2.3 Objectifs

- **Quantifier les contributions** : La décomposition permet de déterminer combien chaque facteur contribue à la variance totale.
- **Identifier les sources de variabilité** : Comprendre quelles sources influencent la variance et où se trouve l'incertitude.
- **Évaluer l'importance des facteurs** : Identifier les facteurs qui expliquent le mieux la variance, pour améliorer les modèles ou expériences.