TP5 : Analyse bivariée
Croisement Qualitatif-Quantitatif

Durée : 2h30

L'objectif de ce TP est d'étudier un lien éventuel entre deux variables, l'une qualitative et l'autre quantitative, au travers deux séries de données.

Exercice 1 — *Décomposition de la variance*

Dans une population Ω de taille n, on observe deux variables :

- une qualitative, $x=\{x_k\}_{k=1,\dots,n}$, à p modalités notées, $m_1,\dots,m_p$
- une quantitative continue $y=\{y_k\}_{k=1,\dots,n}$ de moyenne $\bar{y}$ et de variance s_y^2 .

On suppose que les modalités de la série x définissent des sous-populations

$$\Omega = \Omega_1 \cup \dots \cup \Omega_p \quad \text{où } \Omega_i \cap \Omega_j = \varnothing ,$$

de tailles respectives $n_1,\dots,n_p$.

On peut alors considérer les restrictions de la caractéristiques y sur chacune des sous-populations et calculer les indicateurs numériques usuels pour chaque modalités de x,

- moyennes : $\bar{y}_i$, i=1,…,p
- variances : s_i^2, i=1,…,p

Montrer que

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{p} n_i \bar{y}_i$$

$$s_y^2 = \frac{1}{n}\sum_{i=1}^{p} n_i (\bar{y}_i - \bar{y})^2 + \frac{1}{n}\sum_{i=1}^{p} n_i s_i^2 = s_E^2 + s_R^2$$

A quoi correspondent les termes s_E^2 et s_R^2 ?

On définit un indice de liaison entre les deux caractéristiques x et y par le rapport de corrélation

$$s_{y/x} = \sqrt{\frac{s_E^2}{s_y^2}}$$

Donner un encadrement de $S_{y/x}$. A quoi correspondent les cas $S_{y/x}=0$ et $S_{y/x}=1$?

Exercice 2 — *Données : SalairesData.csv*

Le fichier présente les salariés d'une entreprise ayant 3 sites (A, B et C). On y indique leur sexe, leur salaire annuel (millier d'euros), leur catégorie (CS : cadre supérieur, CM : cadre moyen, OE : ouvrier employé), leur âge et leur site.

1) Sur un même graphique, représenter les boites de Tuckey pour chaque catégorie. Commenter

2) Faire le même graphique mais par site. Sur quel site vaut-il mieux travailler à votre avis ?

3) Pouvez-vous justifier votre réponse à l'aide d'un indicateur numérique ?

Exercice 3 — *Données : EnsSuperieurData.csv*

Le fichier EnsSuperieur.csv comptabilise le nombre d'étudiants par sexe dans l'enseignement supérieur de premier et deuxième cycles. Il s'agit chiffres relevés par Eurostat en 2008.

Illustrer et commenter ces chiffres en travaillant dans un premier temps sur le nombre d'étudiants et ensuite sur le taux d'étudiants pour 1000 habitants.