



# Analyse bivariable Quantitatif-Qualitatif

Durée : 3h

L'objectif de ce TP est d'étudier un lien éventuel entre deux variables, l'une qualitative et l'autre quantitative, à l'aide de la formule de décomposition de la variance.

## Exercice 1

## Décomposition de la variance

Dans une population  $\Omega$  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

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

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

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

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

1) Montrez que

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

2) Montrer que

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

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

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

3) Donnez un encadrement de  $s_{y/x}$ . A quoi correspondent les cas  $s_{y/x}=0$  et  $s_{y/x}=1$  ?

4) Que se passe-t-il si  $n=p$  ?

**Exercice 2***Données : Salaires.txt*

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

- 1) Sur un même graphique, représentez les boîtes de Tuckey du salaire en fonction des catégories (avec R)
- 2) Y-a-t-il une différence de salaire entre les catégories ? Justifiez votre réponse à l'aide d'un indicateur numérique (avec un tableur).
- 3) De la même façon, étudiez le lien entre l'âge et les catégories et entre l'âge et les établissements.
- 4) Sur un même graphique, représentez les boîtes de Tuckey du salaire en fonction des établissements. Pensez-vous pertinent de calculer le rapport de corrélation pour ces séries ?

**Exercice 3***Données : EnsSuperieur.txt*

Le fichier EnsSuperieur.xls 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.

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