

I. Introduction à l'analyse de données

A. Rappels

1. Matrice de données

$$X_{np} = \begin{bmatrix} x_{11} & x_{1p} \\ \vdots & \vdots \\ x_{n1} & x_{np} \end{bmatrix} = [x_{ij}]$$

avec n lignes et p colonnes.
 x_i : vecteur ligne (individus).
 x^j : vecteur colonne (variables).

2. Table disjonctive

$$Z = \begin{bmatrix} z_{11} & z_{1q} \\ \vdots & \vdots \\ z_{n1} & z_{nq} \end{bmatrix}$$

avec n individus et q modalités.

3. Tableau de contingence

		variable 2						Σ
		1					q	
1	n_{11}							$n_{1.}$
p							n_{pq}	$n_{p.}$
Σ	$n_{.1}$						$n_{.q}$	$n_{..}$

avec n_{ij} : le nombre d'individus ayant la modalité i de la variable 1 et j de la variable 2.

4. Statistiques de base

Moyenne : $\bar{X} = \frac{1}{n} \sum_{i=1}^n x_i$

$$\text{Variance : } s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

$$\text{Matrice variance-covariance : } S = \begin{bmatrix} s_{11} & s_{1p} \\ s_{p1} & s_{pp} \end{bmatrix} = [s_{ij}] \text{ où } s_{ij} = \frac{1}{n} \sum_{i=1}^n (x_{ti} - \bar{x}_i)(x_{tj} - \bar{x}_j).$$

→ dépend de l'unité choisie.

→ matrice symétrique.

5. Matrice de centrage

$$M_n = I_n - \frac{1}{n} = \begin{bmatrix} 1 - \frac{1}{n} & -\frac{1}{n} \\ -\frac{1}{n} & 1 - \frac{1}{n} \end{bmatrix}$$

$$\rightarrow M = M^t = M^2$$

$$\rightarrow S = \frac{1}{n} X^t M^t M X = \frac{1}{n} X^t M X$$

6. Matrice de corrélation

$$R = [r_{ij}]$$

$$r_{ij} = \frac{s_{ij}}{s_i s_j}$$

7. Transformation linéaire d'une variable

Si A est une matrice p lignes q colonnes et $Y_{nq} = X_{np} A_{pq}$, Y est une transformation linéaire de X. Dans ce cas, les relations suivantes sont vérifiées :

$$\rightarrow \bar{Y} = A^t \bar{X}$$

$$\rightarrow S_y = A^t S_x A$$

$$\rightarrow S_{XY} = S_x A$$

8. Mesure d'association dans un tableau de contingence

Si deux variables étudiées dans une table de contingence sont indépendantes, la probabilité p_{ij} d'être dans la cellule (ij) est $p_i \times p_j$.

→ Hypothèse : $p_{ij} - (p_i \times p_j) = 0$ où $p_{ij} = \frac{n_{ij}}{n}$, $p_i = \frac{n_{i.}}{n}$, $p_j = \frac{n_{.j}}{n}$, ce qui est équivalent à

$$\frac{n_{ij}}{n} - \frac{n_{i.}n_{.j}}{n^2} = 0.$$

Cette hypothèse est testée par la statistique du $\chi^2 = \sum_{ij} \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}} = 0$ qui n'est rien d'autre

que la somme des écarts au carré par rapport à l'hypothèse d'indépendance.