Département MIDO
Module Analyse de Données
2014 / 2015

Travaux Dirigés n° 1 : Liens entre variables quantitatives

Objectifs : comprendre le lien entre deux variables quantitatives. Savoir calculer et interpréter la covariance et la corrélation entre deux variables d'un tableau de données.

1 Partie 1

Soit $\mathbf{x}$ une variable prenant les valeurs suivantes : $\{1, 3, 5, 7, 8\}$.

1.1 Exercice 1

Considérons une variable $\mathbf{y}$ définie à partir de $\mathbf{x}$ comme suit :

$$\begin{aligned} \mathbf{y} &= f(\mathbf{x}) \\ &= 2\mathbf{x} + 1 \end{aligned}$$

- Représenter graphiquement le nuage de points correspondant.
- Quelle est la nature de la dépendance entre $\mathbf{x}$ et $\mathbf{y}$ (la nature de la fonction f) ?
- Calculer la covariance et le coefficient de corrélation correspondants. La corrélation est-elle un bon indicateur de dépendance dans ce cas ?

1.2 Exercice 2

Considérons désormais une variable $\mathbf{y}$ construite à partir de $\mathbf{x}$ comme suit :

$$y_i = f(x_i) + \varepsilon_i \, , \forall i \in \{1, \cdots, 5\}$$

où ε_i est une variable aléatoire identiquement et indépendamment distribuée suivant une loi normale $\mathcal{N}(0, 2)$ et $f(x_i) = 2x_i + 1$.

- Donner un exemple de réalisation pour $\mathbf{y}$.
- Représenter graphiquement le nuage de points correspondant.
- Calculer le coefficient de corrélation entre $\mathbf{x}$ et $\mathbf{y}$. Celui-ci reste t-il valide comme indicateur de dépendance ?

Si l'on suppose que ε_i suit une loi normale $\mathcal{N}(0, \sigma_i)$ où σ_i est un écart-type variable pour chaque valeur x_i, la corrélation reste t-elle un bon indicateur de dépendance ? Expliquer en donnant un exemple de réalisation pour $\mathbf{y}$.

1.3 Exercice 3

Soit maintenant f définie comme suit :

$$f(x) = \sin(x)$$

— Quelle est la nature de la dépendance entre $\mathbf{x}$ et $\mathbf{y}$ (la nature de la fonction f) ?
— Calculer la covariance et le coefficient de corrélation correspondants. La corrélation est-elle un bon indicateur de dépendance dans ce cas ?

2 Partie 2

Soit le tableau de données suivant :

$\mathbf{x}^1$	$\mathbf{x}^2$
1	0
2	-1
-2	1
-1	0
-1	-2
1	2

2.1 Exercice 1

1. Représenter graphiquement le nuage de points correspondant.
2. Calculez la moyenne, la variance et l'écart-type de $\mathbf{x}^1$ et $\mathbf{x}^2$.
3. Calculez la covariance et la corrélation linéaire entre $\mathbf{x}^1$ et $\mathbf{x}^2$. Commentez.
4. Si on ajoute un 7ème individu $\begin{pmatrix} \mathbf{x}^1 & \mathbf{x}^2 \\ 10 & 10 \end{pmatrix}$, peut-on raisonner avec le coefficient de corrélation linéaire ? (refaire les calculs dans 2 et 3).

2.2 Exercice 2

Soient $\mathbf{z}^1$ et $\mathbf{z}^2$ deux variables obtenues par centrage et réduction de $\mathbf{x}^1$ et $\mathbf{x}^2$. Donnez l'expression de $\rho_{\mathbf{z}^1\mathbf{z}^2}$ en fonction de $\rho_{\mathbf{x}^1\mathbf{x}^2}$.

2.3 Exercice 3

Soit $Z = [z_{ij}]$ la matrice de données centrées réduites : $z_{ij} = \frac{x_{ij} - \overline{\mathbf{x}}^j}{\sigma_{\mathbf{x}^j}}$.
Donnez une expression de la matrice Z en termes de la matrice X des données brutes.

2.4 Exercice 4

On s'intéresse aux prix de 3 articles a_1, a_2, a_3 (exemple : lait, pain, fromage) dans 2 magasins m_1 et m_2. On dispose du tableau suivant :

	m_1	m_2
a_1	3	4
a_2	1.6	2
a_3	5	6

Calculez

1. la matrice Y des données centrées,
2. la matrice des variances-covariances,
3. la matrice de corrélations.