Travaux Dirigés
Module : L'Analyse de Données
Filières : Intelligence Artificielles et Génie Informatique.
la Série I :

Rappel :

Définition générale

L'analyse de données est un ensemble de méthodes statistiques et informatiques permettant d'explorer, de résumer et d'interpréter des ensembles de données afin d'en extraire de l'information utile à la prise de décision. Elle constitue une étape essentielle dans tout processus scientifique ou économique basé sur les données.

Typologie des variables

Les variables décrivent les caractéristiques mesurées sur les individus d'un échantillon.

- **Variables qualitatives** : prennent des modalités non numériques.
  - *Nominales* : sans ordre (ex. sexe, couleur, pays).
  - *Ordinales* : avec un ordre (ex. niveau d'étude, satisfaction).
- **Variables quantitatives (métriques)** : valeurs numériques permettant des calculs.
  - *Discrètes* : valeurs entières (ex. nombre d'enfants).
  - *Continues* : valeurs réelles (ex. taille, poids, revenu).

Structure des données

Les données sont généralement organisées dans une **matrice** $X_{n\times p}$:

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

où :

- n : nombre d'individus (observations) ;
- p : nombre de variables (caractéristiques mesurées).

Chaque ligne correspond à un individu, chaque colonne à une variable.

Étapes d'une analyse de données

1. **Collecte et préparation** : acquisition, nettoyage, codage, traitement des valeurs manquantes.
2. **Analyse descriptive** :
   - statistiques univariées (moyenne, variance, médiane, écart-type) ;
   - représentations graphiques (histogrammes, boîtes à moustaches, nuages de points).
3. **Analyse exploratoire** :
   - recherche de liens ou de structures cachées (corrélations, regroupements) ;
   - réduction de dimension (ACP, AFC, ACM).
4. **Analyse explicative et modélisation** :
   - régression linéaire, logistique, modèles de classification, etc.
5. **Interprétation et visualisation** : synthèse, graphiques, restitution des résultats.

Notions statistiques essentielles

- **Moyenne :** $\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$
- **Variance :** $s^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- **Écart-type :** $s = \sqrt{s^2}$
- **Covariance :** $s_{xy} = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$
- **Corrélation linéaire (de Pearson) :**

$$r_{xy} = \frac{s_{xy}}{s_x s_y}, \qquad -1 \leq r_{xy} \leq 1$$

  - $r_{xy} > 0$: relation linéaire croissante ;
  - $r_{xy} < 0$: relation décroissante ;
  - $r_{xy} \approx 0$: pas de lien linéaire.

Types de méthodes en analyse de données

- **Méthodes descriptives** : visent à résumer les données (moyenne, variance, ACP, AFC, ACM).
- **Méthodes explicatives** : cherchent à modéliser ou prédire une variable (régression, ANOVA, modèles linéaires).
- **Méthodes exploratoires** : visent à découvrir des structures cachées (classification, clustering, analyse de correspondances).

Objectifs de l'analyse de données

- Décrire et résumer l'information ;
- Mettre en évidence des relations entre variables ;
- Réduire la dimension des données tout en conservant l'essentiel de l'information ;
- Identifier des profils ou des groupes homogènes ;
- Aider à la prise de décision à partir des résultats.

Exercice 1. *On dispose des informations suivantes concernant cinq étudiants d'une même classe :*

Étudiant	*Âge (années)*	*Note en Mathématiques*	*Note en Statistiques*
A	20	14	13
B	22	16	17
C	21	10	11
D	23	18	19
E	20	12	14

1. *Identifier le type (qualitative/quantitative) de chaque variable.*
2. *Calculer la moyenne et l'écart-type des notes de Mathématiques.*
3. *Calculer la covariance et la corrélation entre les notes de Mathématiques et de Statistiques.*
4. *Interpréter le signe et la valeur du coefficient de corrélation obtenu.*

Exercice 2. *On dispose du tableau suivant représentant les ventes (en milliers d'unités) réalisées par quatre commerciaux au cours de trois mois :*

Commercial	*Janvier*	*Février*	*Mars*
A	12	15	14
B	10	11	13
C	8	9	9
D	15	18	20

1. *Identifier les variables et leur type.*
2. *Calculer la moyenne et l'écart-type des ventes pour chaque mois.*
3. *Quelle est la corrélation entre les ventes de janvier et de mars ?*
4. *Interpréter le résultat obtenu.*

Exercice 3. *Soit cinq individus mesurés sur trois variables* X_1, X_2, X_3 *:*

$$X = \begin{pmatrix} 4 & 2 & 0 \\ 2 & 0 & 1 \\ 3 & 1 & 2 \\ 5 & 3 & 4 \\ 4 & 2 & 3 \end{pmatrix}.$$

1. *Calculer les moyennes des variables et la matrice centrée* X_c*.*
2. *Calculer la matrice de covariance empirique* $S = \dfrac{1}{n-1} X_c^\top X_c$*.*
3. *Déterminer les valeurs propres de* S *et calculer la proportion de variance expliquée (PVE) par chaque composante.*
4. *Donner (sans normalisation) un vecteur propre associé à la plus grande valeur propre et interpréter brièvement.*

Exercice 4. *On étudie les résultats d'un test de performance pour deux groupes d'étudiants : le groupe A (méthode classique) et le groupe B (méthode innovante).*

Groupe A	10	12	13	9	11	10
Groupe B	14	15	13	16	15	17

1. *Identifier la nature et le type des variables observées.*
2. *Calculer la moyenne et l'écart-type pour chaque groupe.*
3. *Comparer les deux groupes et commenter les résultats.*
4. *Représenter graphiquement les distributions (expliquer le choix du graphique).*