1. Introduction

L'analyse des données est un sous domaine des statistiques qui se préoccupe de la description de données conjointes. On cherche par ces méthodes à donner les liens pouvant exister entre les différentes données et à en tirer une information statistique qui permet de décrire de façon plus succincte les principales informations contenues dans ces données. On peut également chercher à classer les données en différents sous-groupes plus homogènes.

Le but de ces méthodes est de synthétiser les grands tableaux pour en fournir une présentation simplifiée.

L'analyse des données est utilisée dans tous les secteurs de l'activité humaine. Elle est un ensemble plus ou moins défini de méthodes statistiques. Ces méthodes permettent de collecter, organiser, résumer, présenter et étudier des données pour permettre d'en tirer des conclusions et de prendre des décisions.

2. C'est quoi l'analyse des données ?

L'Analyse est une méthode qui s'oppose à la synthèse. Elle vise à comprendre un objet en le décomposant en ses constituants.

L'analyse : L'analyse est une étude minutieuse, précise faite pour dégager les éléments qui constituent un ensemble, pour l'expliquer, l'éclairer : Faire l'analyse de la situation.

Les données : La donnée est l'élément fondamental, indispensable à tout raisonnement pour extraire de l'information nécessaire à la compréhension des phénomènes.

3. Les types des données

Il y a principalement des données qualitatives et des données quantitatives :

Les données collectées de type quantitatif sont d'ordre numérique. Elles concernent les échelles de mesure d'intervalle ou de rapport.

Les données collectées de type qualitatif sont dites catégorielles car elles sous-tendent des groupes (échelle de mesure nominale ou des classes hiérarchisées .

En d'autre terme, les données qualitatives se réfèrent à la qualité : La description d'une couleur, de textures et l'aspect d'un objet, la description d'une expérience sont toutes des données qualitatives. Par contre, les données quantitatives sont des données qui se réfèrent aux chiffres. Ex : Le nombre de balles de golf, la taille, le prix, le résultat d'un test, etc.

4. Les Tableaux des données

Les données sont les mesures effectuées sur n unités (x_{i1}, x_{i2},….., x_{ip}). Les p variables qui représentent ces mesures sont $v1, v2 \ldots\ldots vj, \ldots . vp$

Le tableau des données brutes à partir duquel on va faire l'analyse est noté x et a la forme suivante

$$x = \begin{bmatrix} x_{11} & x_{12} \cdots x_{1p} \\ x_{21} & x_{22} \cdots x_{2p} \\ x_{n1} & x_{n2} \cdots x_{np} \end{bmatrix}$$

Chaque unité xi peut être représentée par le vecteur

$$x_i = \left(x_{i1}, x_{i2}, \ldots . , x_{ip}\right), x_i \epsilon \mathbb{R}^p$$

De façon analogue, on peut représenter chaque variable par un vecteur de $\mathbb{R}^n$ dont les composantes sont les valeurs de la variable pour les n unités :

$$v_j = \begin{bmatrix} v_{1j} \\ \vdots \\ v_{nj} \end{bmatrix}$$

Pour avoir une image de l'ensemble des unités, on se place dans un espace affine en choisissant comme origine un vecteur particulier de $\mathbb{R}^p$, par example le vecteur dont toutes les cordonnée sont nulles. Alors, chaque unité sera représentée par un point dans cet espace. L'ensemble des points qui représentent les unités est appelé traditionnellement « nuage des individus ».

En faisant de même dans $\mathbb{R}^n$. Chaque variable pourra être représentée par un point de l'espace affine correspondant. L'ensemble des points qui représentent les variables est appelé « nuage des variables ».

Remarque : On appelle données multidimensionnelles, l'ensemble des valeurs d'un certain nombre des variables statistiques sur un individu d'une population donnée.

4.1. Tableau individu X caractères quantitatifs :

Ce type de tableau l'un le plus simple et le plus répandu. En générale $x_j(i)$ est un nombre réel représentent la mesure de la variable x_j sur l'individu i.

Example : Soit le tableau suivant

	Couleur	Poids
X_1	Blanc	5
X_2	Bleu	5
X_3	Rouge	5
X_4	Blanc	10
X_5	Rouge	10

4.2. Tableaux logiques

L'on définit, pour les variables quantitatives, une répartition en classe, l'ensemble des résultats de l'observation peut être présenté sous la forme d'un tableau logique composé de 0 et de 1. On présente de façon analogue les tableaux x des caractères qualitatifs.

Les données qualitatives peuvent être nominales ou ordinales. Dans l'exemple suivant on peut considérer la variable couleur comme nominale et la variable taille comme ordinale.

	Couleur	Taille
X_1	Blanc	Petit
X_2	Vert	Grand
X_3	Blanc	Très grand
X_4	Blanc	Petit

Petit < Grand < Très grand

- **Codage de variable qualitatives nominales :**
  Soit la fonction : $\mathcal{N}: v \to \{0,1\}^k$ tel que
  k est le nombre maximum de valeurs (modalités) que peut prendre la variable nominale.

  $N(x_{ij}) = (0, \dots, 0, 1, 0, \dots, 0)$ → Rang de x_{ij} $\begin{cases} 1 \; au \; rang \; x_{ij} \\ \\ 0 \; ailleurs \end{cases}$

Couleur (blanc, vert)
Codage du blanc est (1,0) et codage du vert est (0,1)

- **Codage des données qualitatives ordinales :**
  $\boldsymbol{O}: v \to \{0,1\}^k$
  $$\boldsymbol{O}(x_{ij}) = (\mathbf{1},..,\mathbf{1},\mathbf{0},..,\mathbf{0}) \begin{cases} 1\ jusq'au\ rang\ x_{ij} \\ \\ 0\ apres \end{cases}$$

Exp :

Taille (petit, grand, très grand)

$\boldsymbol{O}(\boldsymbol{petit}) = (\mathbf{1},\mathbf{0},\mathbf{0})$

$\boldsymbol{O}(\boldsymbol{grand}) = (\mathbf{1},\mathbf{1},\mathbf{0})$

$\boldsymbol{O}(\boldsymbol{T.grand}) = (\mathbf{1},\mathbf{1},\mathbf{1})$

A partir de l'exemple précèdent on peut coder les variables qualitatives comme suit :

	Couleur		Taille		
X_1	1	0	1	0	0
X_2	0	1	1	1	0
X_3	1	0	1	1	1
X_4	1	0	1	0	0

(Codage disjonctif)

RQ : Si aucune des variables qualitatives n'est ordinales le tableau de codage obtenu est un tableau sous forme disjonctive complète.

4.3. Tableaux de contingence

Soit P, une population examinée suivant deux caractères. L'un prenant les valeurs sur un ensemble I. de n modalités i, et l'autre sur un ensemble J, de p modalités j. le tableau de contingence K (ou **tableau croisé**). Associé à ces données est le tableau de dimensions $n \times p$ et de terme général k_{ij} ; dont k_{ij} le nombre d'invidus présentant simultanément la modalité i pour le premier caractere. Et j pour le second.

Exemple :

Couleur \ Poids	5	10
Blanc	1	1
Bleu	1	0
Rouge	1	1

Signifie qu'il y a 1 seul individu de couleur rouge et poids 5

4.4. Tableaux des fréquences

Les fréquences sont calculées par : $f_{ij} = \frac{k_{ij}}{n}$

Poids / Couleur	5	10	Fréquence Marginale ($f_{.j}$)
Blanc	1/5	1/5	2/5
Bleu	1/5	0	1/5
Rouge	1/5	1/5	2/5
Fréquence Marginale ($f_{i.}$)	3/5	2/5	1

Tableau des fréquences lignes : $fl_{ij} = \frac{f_{ij}}{f_{i.}}$

Tableau de profil lignes :

1/2	1/2	1
1	0	1
1/2	1/2	1

Tableau des fréquences colonnes : $fc_{ij} = \frac{f_{ij}}{f_{.j}}$

Tableau de profil colonnes :

1/3	1/2
1/3	0
1/3	1/2
1	1

4.5. Tableau de Proximité

Considérons un ensemble I d'objets, on dispose d'une mesure de ressemblance ou de dissemblance entre tous les éléments de I pris deux a deux.

$$i \epsilon I, i' \epsilon I : d(i, i') \geq 0$$

- **Choix de la distance entre individus**

Soit un tableau à double entrée des individus i $i \epsilon l = \{1, \dots, p\}$ et des variables numériques j, $j \epsilon k = \{1, \dots, k\}$ avec généralement chaque ligne i est un vecteur définé dans R^k ; on évalue ensuite la ressemblance entre individu en calculant la distance euclidienne entre points pris deux à deux, la distance euclidienne entre les individus :

$e_i = \left(x_i^1, x_i^2, \dots x_i^p\right)$ et $e_j = \left(x_j^1, x_j^2, \dots x_j^p\right)$ est définie par :

$$d^2 = \left(e_i, e_j\right) = \left(x_i^1 - x_j^1\right)^2 + \left(x_i^2 - x_j^2\right)^2 + \cdots + \left(x_i^p - x_j^p\right)^2$$

En **analyse des données**, les **distances** et les **mesures de similarité** sont des outils essentiels pour comparer des objets, des points ou des vecteurs, en particulier dans le cadre du **clustering**, de la **classification** ou de la **réduction de dimensions**. Ces mesures permettent d'évaluer dans quelle mesure deux objets sont proches ou similaires.

Voici les principales distances et mesures de similarité utilisées dans ce contexte :

1. Mesures de distance

Les mesures de distance quantifient la dissimilarité entre des objets dans un espace de données. Voici quelques-unes des plus courantes :

a. Distance Euclidienne

La distance euclidienne est la distance "à vol d'oiseau" entre deux points dans un espace à plusieurs dimensions.

Formule pour deux points $x = (x_1, x_2, \cdots, x_n)$ et $y = (y_1, y_2, \cdots, y_n)$

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Cette distance est utilisée dans les algorithmes comme **k-means**, où l'objectif est de minimiser la somme des distances euclidiennes entre les points et leurs centres.

b. Distance de Manhattan (ou distance L1)

La distance de Manhattan, aussi appelée distance des blocs ou distance L1, est la somme des différences absolues entre les coordonnées des deux points.

Formule : $d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$

Elle est adaptée lorsque les distances sont prises dans des chemins en forme de grille, comme dans certaines applications géographiques.

c. Distance de Minkowski

La distance de Minkowski généralise les distances euclidiennes et de Manhattan. Elle est définie par un paramètre p , qui permet d'ajuster le type de distance.

Formule :

$$d(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}}$$

- p = 1 : correspond à la distance de Manhattan.
- p = 2 : correspond à la distance euclidienne.

d. Distance de Chebyshev

La distance de Chebyshev mesure la plus grande différence absolue entre les coordonnées correspondantes de deux points. Elle est utilisée dans des contextes où les différences maximales sont critiques.

Formule :

$$d(x, y) = \max_i |x_i - y_i|$$

e. Distance de Mahalanobis

La distance de Mahalanobis prend en compte la corrélation entre les variables. Elle est particulièrement utile dans des contextes multivariés où les dimensions des données sont corrélées.

Formule :

$$d(x, y) = \sqrt{(x - y)^T \Sigma^{-1} (x - y)}$$

Où :

Σ est la matrice de covariance des données. Elle est utilisée notamment dans des algorithmes de classification supervisée.

f. Distance de Hamming

La distance de Hamming mesure le nombre de positions différentes entre deux chaînes de caractères ou vecteurs binaires. Elle est utile dans le traitement de données discrètes.

2. Mesures de similarité

Les mesures de similarité quantifient la proximité ou la ressemblance entre deux objets ou vecteurs. Elles sont souvent utilisées en apprentissage automatique ou en analyse de texte.

a. Coefficient de corrélation de Pearson

Le coefficient de corrélation de Pearson mesure la linéarité de la relation entre deux variables.

Formule :

$$r(x, y) \frac{\sum_{i=1}^{n} (x_i \bar{x})(y_i \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i \bar{y})^2}}$$

$\bar{x}\ et\ \bar{y}$ sont les **moyennes** des variables $x\ et\ y$.

Ce coefficient varie de -1 (corrélation négative) à 1 (corrélation positive).

Une valeur proche de 0 indique une absence de corrélation.

b. Cosinus de similitude

La similarité cosinus mesure l'angle entre deux vecteurs, ce qui permet de capturer la similarité de direction plutôt que de magnitude. Elle est souvent utilisée en analyse de texte, comme pour la modélisation des documents.

Formule :

$$similarité\ cosinus(x, y) = \frac{x\ .\ y}{\|x\|\|y\|}$$

La valeur varie entre -1 (complètement opposé) et 1 (parfaitement similaire), avec 0 représentant une orthogonalité, c'est-à-dire une absence de relation.

- $x\ .\ y$ est le produit scalaire des deux vecteurs.

- $\|x\|\ et\ \|y\|$ sont les normes des vecteurs.

c. Coefficient de Jaccard

Le coefficient de Jaccard est utilisé pour mesurer la similarité entre deux ensembles, c'est-à-dire le rapport entre la taille de l'intersection des ensembles et celle de leur union.

Formule :

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Cette mesure est particulièrement utilisée dans la comparaison d'ensembles binaires, comme en analyse de données catégorielles ou dans des tâches de clustering.

d. Mesure de Dice (ou indice de Dice)

La mesure de Dice est une alternative au coefficient de Jaccard. Elle met davantage l'accent sur l'intersection des ensembles.

Formule :

$$D(A,B) = \frac{2|A \cap B|}{|A| + |B|}$$

Elle est utilisée dans des domaines comme la bio-informatique ou l'analyse d'images pour comparer des ensembles de données.

e. Indice de Bray-Curtis

L'indice de Bray-Curtis mesure la dissimilarité entre deux ensembles basés sur la somme des différences absolues des abondances de chaque espèce.

Formule :

$$BC(A,B) = \frac{\sum_{i=1}^{n} |x_i - y_i|}{\sum_{i=1}^{n} (x_i + y_i)}$$

Elle est utilisée dans des contextes écologiques ou dans des études de biodiversité.

3. Choix des mesures en fonction des applications

- **Clustering (comme K-means)** : on utilise souvent la distance euclidienne, bien que des distances comme celle de Manhattan soient aussi utilisées en fonction des données.
- **Analyse de texte :** la similarité cosinus est fréquemment utilisée pour mesurer la similarité entre documents ou vecteurs de mots.
- **Données catégoriques** : des mesures comme le coefficient de Jaccard ou la distance de Hamming sont adaptées pour les ensembles ou les vecteurs binaires.
- **Données corrélées** : la distance de Mahalanobis est particulièrement utile pour tenir compte des dépendances entre les variables.

Le choix de la distance ou de la mesure de similarité dépend des caractéristiques des données et de l'application. Les distances euclidienne et cosinus de similarité sont très populaires pour des tâches de clustering ou de classification. Cependant, pour des données spécifiques (binaires, corrélées,

discrètes), des mesures plus spécialisées comme la distance de Hamming ou de Mahalanobis peuvent s'avérer plus pertinentes.

Rappel statistique

- **La moyenne**

La moyenne est la somme des valeurs divisée par le nombre d'éléments :

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{N} \qquad = \frac{1}{N}\left(x_1 + x_2 + \ldots + x_n\right)$$

D'où

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{n} x_i$$

- **Variance et Ecart-type**

L'écart-type est une valeur très importante car il nous donne une idée de la dispersion d'une variable autour de sa moyenne arithmétique. L'écart-type est la racine carrée de la variance, elle-même définie comme la moyenne du carré des écarts à la moyenne :

Variance = $$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

L'écart type = $$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

Exemple :

Une entreprise multinationale, le résultat économique qu'elle a eu au cours des cinq dernières années est connu, dans la plupart elle a obtenu des bénéfices mais une année elle a présenté des pertes considérables : 11,5, 2, -9, 7 millions de Dinars. Calculez la variance de cet ensemble de données.

Calculer sa moyenne arithmétique :

$$\bar{x} = \frac{11 + 5 + 2 + (-9) + 7}{5} = 3.2$$

$$Var(X) = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n}$$

Nous substituons les données fournies par la déclaration d'éxercice dans la formule :

$$Var(X) = \frac{(11 - 3.2)^2 + (5 - 3.2)^2 + (2 - 3.2)^2 + (-9 - 3.2)^2 + (7 - 3.2)^2}{5}$$

Finalement, il ne reste plus qu'à résoudre les opérations pour cqlculer la varianceM

$$Var(X) = \frac{7.8^2 + 1.8^2 + (-1.2)^2 + (-12.2)^2 + 3.88^2}{5}$$

$$= \frac{60.84 + \ 3{,}24 + 1.44 + 148.84 + 14.44}{5}$$

$$= \frac{228.8}{5}$$

$$45.76\ milliones\ de\ \ euros^2$$

Notez que les unités de variances sont les mêmes unités des données statistiques mais élevées au carré, pour cette raison la variance de ce groupe de données et de $45.76\ milliones\ de\ \ euros^2$

- **Covariance**

La covariance est la moyenne du produit des écarts à la moyenne.

En statistique, la covariance est une valeur qui indique le degré de variation conjointe de deux variables aléatoires. Autrement dit, la covariance est utilisée pour analyser la dépendance entre deux variables.

La covariance est égale à la somme des produits des différences entre les données des deux variables et leurs moyennes respectives divisée par le nombre total de données.

$$Cov(X,Y) = \frac{1}{N}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)$$

- **Le coefficient de corrélation**

Ce coefficient permet de détecter la présence ou l'absence d'une relation linéaire entre deux caractères quantitatifs continus. Pour calculer ce coefficient il faut tout d'abord calculer la covariance. Le coefficient de corrélation linéaire de deux caractères X et Y est égal à la covariance de X et Y divisée par le produit des écarts-types de X et Y

$$r(X,Y) = \frac{Cov(X,Y)}{\sigma_x \,.\, \sigma_y}$$

Propriétés et interprétation de $r(X,Y)$

On peut démontrer que ce coefficient varie entre -1 et +1. Son interprétation est la suivante :

- si *r* est proche de 0, il n'y a pas de relation linéaire entre X et Y
- si *r* est proche de -1, il existe une forte relation linéaire négative entre X et Y
- si *r* est proche de 1, il existe une forte relation linéaire positive entre X et Y

Le **signe** de *r* indique donc le sens de la relation tandis que la valeur absolue de *r* indique l'**intensité** de la relation c'est-à-dire la capacité à prédire les valeurs de Y en fonction de celles de X.