Cours d‘Analyse des Données

Présenté par Monsieur
CHEKARAOU IDI

1. Introduction

- L'analyse des données est une technique relativement ancienne 1930 (PEARSON, SPEARMAN, HOTELLING). Elle a connu cependant des développements récents 1960-1970 du fait de l'expansion de l'informatique.
- L'analyse des données est une technique d'analyse statistique d'ensemble de données. Elle cherche à décrire des tableaux et à en exhiber des relations pertinentes. Elle se distingue de l’analyse exploratoire des données.
- L'objectif de la démarche statistique est de faire apparaître ces liaisons. Les deux types de relations fondamentales sont les relations d'équivalence et les relations d'ordre. Ainsi, une population peut-elle être décomposée en classes hiérarchisées.

2. But

- Synthétiser, structurer l'information contenue dans des données multidimensionnelles (n individus, p variables).

3. ELÉMENTS FONDAMENTAUX

3.1. Méthodes

- Algèbre linéaire:
  les données sont vues de manière abstraites comme un nuage de points dans un espace vectoriel. On utilise
  - Des matrices qui permettent de manipuler un ensemble de variables comme un objet mathématique unique ;
  - Des valeurs et vecteurs propres qui permettent de décrire la structure d'une matrice.
  - Des métriques : permettent de définir la distance entre deux points de l'espace vectoriel ; on utilise aussi des produits scalaires.
- Théorie des probabilités
  nécessaire en statistique inferentielle (estimation, tests, modélisation et prévision,...).

3.2. rappels de géométrie : produit scalaire

- produit scalaire

Le produit scalaire de deux vecteurs est le produit de la longueur de l'un par la projection de l'autre sur lui. (u.v.Cos(u,v))

- Propriétés

- Si les vecteurs sont orthogonaux le produit scalaire est nul.
- Si les vecteurs sont colinéaires le produit scalaire est ±(u.v)
- Si les vecteurs unitaires sont orthogonaux le produit scalaire est égal à la somme des produits des composantes correspondantes.

3.2. rappels de géométrie : projection

- projection

La projection d'un vecteur sur un axe est obtenue par le produit scalaire du vecteur par le vecteur unitaire de l'axe. Cela permet le changement d’axe de coordonnées.

3.2. rappels de géométrie: Distance

- distance

Dans l'espace des variables, un produit scalaire particulier, et donc une distance, s'impose.

Le choix d'une distance est toujours arbitraire dans l'espace des individus, car il est possible d'associer à chaque variable un coefficient de pondération.

3.2. rappels de géométrie : Métrique:

- **Métrique usuelle**
  M = I correspond au produit scalaire usuel et

$$I_g = Tr(V) = \sum_{j=1}^{p} s_i^2$$

- **Problèmes**
  - la distance entre individus dépend de l'unité de mesure.
  - la distance privilégie les variables les plus dispersées.
- **Métrique réduite**
  c'est la plus courante ;
  on prend la matrice diagonale des inverses des variances

$$M = D_{1/s^2} = \begin{bmatrix} \frac{1}{s_1^2} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{s_p^2} \end{bmatrix}$$

$$I_g = Tr(D_{1/s^2} V) = Tr(D_{1/s} V D_{1/s}) = Tr(R) = p$$

3.2. rappels de géométrie : Métrique et Tableau Transformé

- Utiliser la métrique M = T'T sur le tableau X est équivalent a travailler avec la métrique classique I sur le tableau transforme XT'.

- **Tableau transformé**
  Si on travaille sur le tableau transforme XT' (changement de variables) au lieu de X, alors les nouveaux individus seront de la forme Te_i et

$$\left\langle Te_{i_1}, Te_{i_2} \right\rangle = (Te_{i_1})'(Te_{i_2}) = e_{i_1}'T'Te_{i_2} = e_{i_1}'Me_{i_2} = \left\langle e_{i_1}, e_{i_2} \right\rangle_M$$

- **Réciproque**
  pour toute matrice symétrique positive M, il existe une matrice T (racine carrée de M) telle que

$$M = T'T$$

  et donc on peut ramener l'utilisation de la métrique a un changement de variables.

3.3. Rappels: notation matricielle

- **Matrice**
  tableau de données carre ou rectangulaire.

- **Vecteur**
  matrice a une seule colonne.

- **Cas particuliers : matrice identité**

$$I = \begin{bmatrix} 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1 \end{bmatrix} \qquad 1 = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

- **Transposition de matrice**
  échange des lignes et des colonnes d'une matrice ; on note M’ la transposée de M.

3.3. La matrice des poids

- **Pourquoi**
  utile quand les individus n'ont pas la même importance

- **Comment**
  on associe aux individus un poids pi tel que
  $$p_1 + p_2 + ... + p_n = 1$$
  et on représente ces poids dans la matrice diagonale de taille n

$$D = \begin{bmatrix} p_1 & & \dots & 0 \\ & p_2 & & \\ \vdots & & \ddots & \vdots \\ 0 & & \dots & p_n \end{bmatrix}$$

- **Cas uniforme**
  tous les individus ont le même poids $p_i = 1 / n$ et $D = I / n$

3.3. Point moyen et tableau centre

- **Point moyen**
  c'est le vecteur g des moyennes arithmétiques de chaque variable :

$$g' = (\bar{x}^1 \quad \dots \quad \bar{x}^p)$$

ou $$\bar{x}^j = \sum_{i=1}^{n} p_i x_i^j$$

On peut aussi écrire $$g = X' D 1$$

Tableau centré
il est obtenu en centrant les variables autour de leur moyenne

$$y_i^j = x_i^j - \bar{x}^j$$

ou, en notation matricielle,

$$Y = X - 1g' = (I - 11'D)X$$

3.3. Matrice de variance covariance

- **Définition**
  c'est une matrice carrée de dimension p

$$V = \begin{bmatrix} s_1^1 & s_1^2 & \dots & s_1^p \\ s_2^1 & s_2^2 & & \\ \vdots & & \ddots & \vdots \\ s_p^1 & & \dots & s_p^p \end{bmatrix}$$

  ou s_{kl} est la covariance des variables x^k et x^l et s^2_j est la variance de la variable x^j

- **Formule matricielle**

$$V = X'DX - gg' = Y'DY$$

3.3. Matrice de corrélation

- **Définition**
  Si l'on note

$$r_{kl} = \frac{s_{kl}}{s_k s_l} \qquad R = \begin{bmatrix} 1 & r_1^2 & \dots & r_1^p \\ r_2^1 & 1 & & \\ \vdots & & \ddots & \vdots \\ s_p^1 & & \dots & 1 \end{bmatrix}$$

- **Formule matricielle**

$$R = D_{1/s} V D_{1/s} \qquad D_{1/s} = \begin{bmatrix} \frac{1}{s_1} & & & 0 \\ \vdots & \frac{1}{s_2} & & \vdots \\ & & \ddots & \\ 0 & & & \frac{1}{s_p} \end{bmatrix}$$

3.3. rappels sur les matrices

- Trace

La trace d'une matrice est la somme des termes de la diagonale principale.

- Valeur propre

λ est valeur propre de A <=> Det(A - λ I) = 0

- Vecteur propre

V est vecteur propre de f si f(V) = λ V

3.3. rappels sur les matrices

- matrice diagonale

Une matrice diagonale est une matrice dont tous les termes appartiennent à la diagonale principale.

3.3. Valeurs et vecteurs propres

- **Définition**
  un vecteur v de taille p est un vecteur propre d'une matrice A de taille p x p s'il existe λ Є C telle que

$$Av = \lambda v$$

  est une valeur propre de A associée à v.

- **Domaine**
  En général, les vecteurs propres et valeurs propres sont complexes; dans tous les cas qui nous intéressent, ils seront réels.

- **Interprétation des vecteurs propres**
  ce sont les directions dans lesquelles la matrice agit.

- **Interprétation des valeurs propres**
  c'est le facteur multiplicatif associe a une direction donnée.

3.3. Exemple: valeurs et vecteurs propres

La matrice

$$\begin{pmatrix} 5 & 1 & -1 \\ 2 & 4 & -2 \\ 1 & -1 & 3 \end{pmatrix}$$

a pour vecteurs propres

$$v_1 = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \qquad v_2 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \qquad v_3 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$$

On vérifie facilement que les valeurs propres associées sont

$$\lambda_1 = 2 \qquad \lambda_2 = 4 \qquad \lambda_3 = 6$$

3.3. Cas particuliers: Valeurs et vecteurs propres

- **Matrice nulle**
  sa seule valeur propre est 0, et tout vecteur est vecteur propre.

- **Matrice identité**
  tout vecteur est vecteur propre de I avec valeur propre 1, puisque Iv = v.

- **Matrice diagonale**
  si D_λ est une matrice diagonale avec les coefficients $\lambda_1, \lambda_2, \ldots \lambda_p$, alors le i-eme vecteur coordonnée est vecteur propre de D_λ associe à la valeur propre λ_i.
  L'action d'une matrice diagonale est de multiplier chacune des coordonnées d'un vecteur par la valeur propre correspondante.

- **Matrice diagonalisable**
  c'est une matrice dont les vecteurs propres forment une base de l'espace vectoriel : tout vecteur peut être représenté de manière unique comme combinaison linéaire des vecteurs propres. Une matrice de taille p x p qui a p valeurs propres réelles distinctes est diagonalisable dans R.

3.3. Quelques matrices diagonalisables

- **Matrice symétrique**
  une matrice symétrique réelle (A' = A) possède une base de vecteurs propres orthogonaux et ses valeurs propres sont réelles

$$\langle v_i, v_j \rangle = 0 \quad si \quad i \neq j \quad et \quad \lambda_i \in \Re$$

- **Matrice M-symetrique**
  une matrice M-symetrique réelle (A'M = MA) possède une base de vecteurs propres M-orthogonaux et ses valeurs propres sont positives ou nulles

$$\langle v_i, v_j \rangle_M = 0 \quad si \quad i \neq j \quad et \quad \lambda_i \in \Re$$

- **Matrice définie positive**
  c'est une matrice symétrique dont les valeurs propres sont strictement positives et donc

$$\langle v_i, v_j \rangle = 0 \quad si \quad i \neq j \quad et \quad \lambda_i > 0$$

3.3. Analyse de la matrice notée:VM

- **Valeurs propres**
  la matrice VM est M-symetrique: elle est donc diagonalisable et ses valeurs propres λ_1, λ_2, λ_p sont réelles.

- **Vecteurs propres**
  il existe donc p vecteurs $a_{1, ..., } a_p$ tels que

$$VMa_i = \lambda a_i \qquad avec \quad \left\langle a_i, a_j \right\rangle_M = \begin{cases} 1 & \text{si} \quad i = j \\ 0 & \text{sinon} \end{cases}$$

  Les a_i sont les axes principaux d'inertie de VM. Ils sont M-orthonormaux.

- **Signe des valeurs propres**
  les valeurs propres de VM sont positives et on peut les classer par ordre décroissant

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$$

- **Idée du lien avec l'inertie**
  on sait que .

$$Tr(VM) = \lambda_1 + \lambda_2 + \ldots + \lambda_p$$

  Si on ne garde que les données relatives a $a_{1, ..., } a_p$ on gardera l'inertie $\lambda_1 + \lambda_2 + \ldots + \lambda_p$, et c'est le mieux qu'on puisse faire.

3.4. ***rappels de mécanique***

- centre de gravité

Le centre de gravité d'un solide, ou barycentre, correspond à la notion statistique de moyenne.

- inertie

L'inertie d'un solide correspond à la notion de variance

Un corps a d'autant plus d'inertie qu'il faut d'énergie pour le mettre en rotation autour d'un axe.

3.4. Inertie

l'inertie totale est égale à l'inertie expliquée par l'axe et l'inertie autour de l'axe. les 3 valeurs propres de la Matrice V sont les inerties expliquées par les 3 axes du nuage. leur somme est égale à la trace de V, soit à l'inertie du nuage.

3.4. Inertie

- **Définition**
  l'inertie en un point a du nuage de points est

$$I_a = \sum_{i=1}^{n} p_i \left\| e_i - a \right\|_M^2 = \sum_{i=1}^{n} p_i (e_i - a)' M (e_i - a)$$

- **Autres relations**
  l'inertie totale Ig est la moitie de la moyenne des carres des distances entre les individus

$$2I_g = \sum_{i=1}^{n} \sum_{j=1}^{n} p_i p_j \left\| e_i - e_j \right\|_M^2$$

- L'inertie totale est aussi donnée par la trace de la matrice MV (la trace d'une matrice étant la somme de ses éléments diagonaux).

$$I_g = Tr(MV)$$

3.5. rappels de statistique descriptive

- **La Statistique Descriptive** est l'ensemble des méthodes et techniques permettant de présenter, de décrire, de résumer, des données nombreuses et variées.
- **population statistique** est l'ensemble étudié dont les éléments sont des individus ou unités statistiques.
- **Recensement**
  étude de tous les individus d'une population donnée.
- **Sondage**
  étude d'une partie seulement d'une population appelée échantillon.
- **Echantillon** est un ensemble d'individus extraits d'une population initiale de manière aléatoire de façon à ce qu'il soit représentatif de cette population
- **Caractère** est l'aspect des individus que l'on étudie
- .

3.5. rappels de statistique descriptive

- **Nature du caractère**
  - **quantitatives**: nombres sur lesquels les opérations usuelles (somme, moyenne,...) ont un sens ; elles peuvent être discrètes (ex : nombre d'éléments dans un ensemble) ou continues (ex: prix, taille) ;

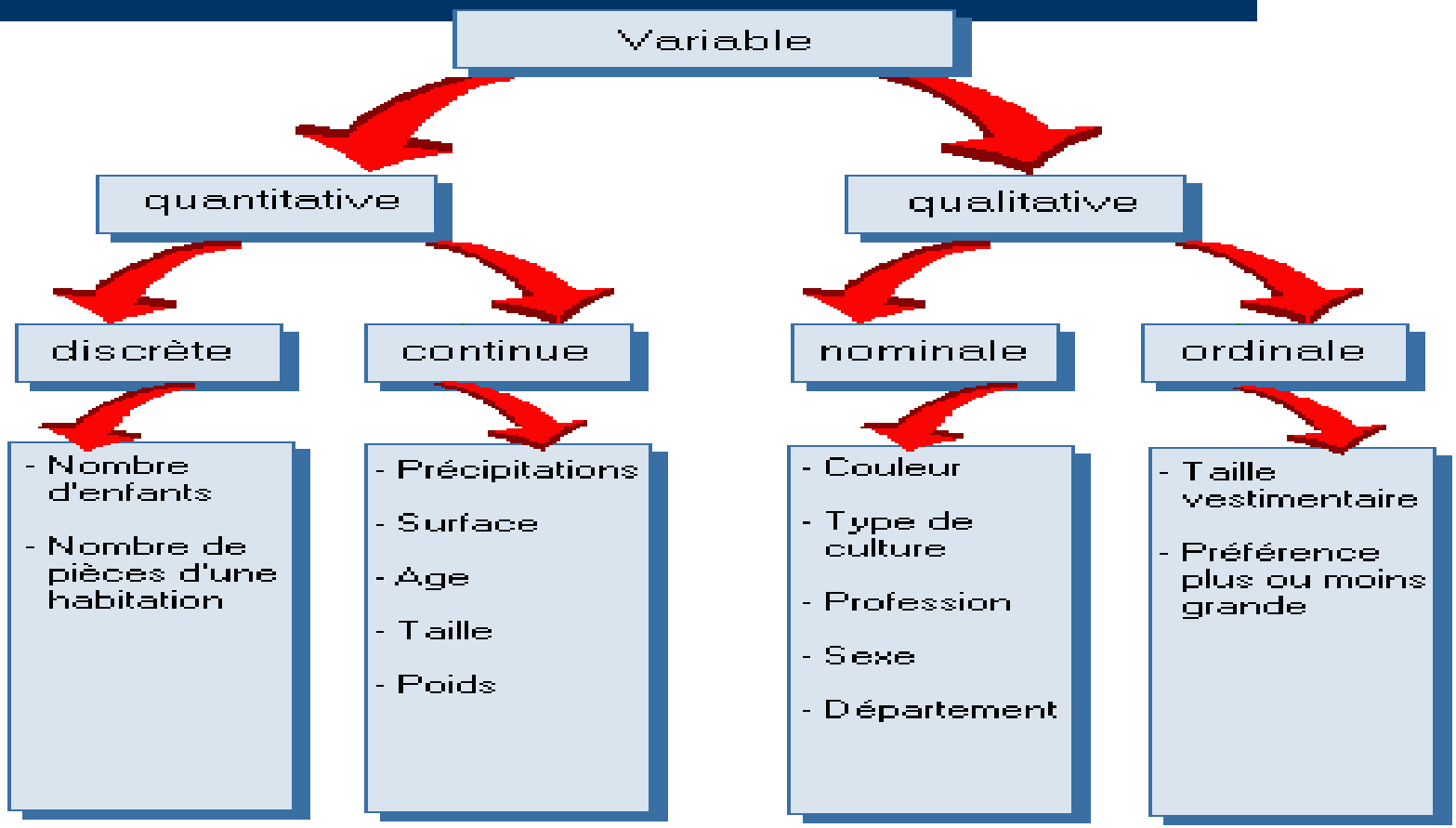
    La variable peut alors être discrète ou continue selon la nature de l'ensemble des valeurs qu'elle est susceptible de prendre (valeurs isolées ou intervalle).

  - **qualitatives**: appartenance a une catégorie donnée ; elles peuvent être nominales (ex : sexe, CSP) ou ordinales quand les catégories sont ordonnées (ex : très résistant, assez résistant, peu résistant)

    On distingue des variables qualitatives ordinales ou nominales, selon que les modalités peuvent être naturellement ordonnées ou pas.

    Une variable est ordinale si l'ensemble des catégories est munie d'un ordre total si non elle est nominale

3.5. rappels de statistique descriptive



3.5. SD: paramètres de position et dispersion

- **Introduction**

  on dispose d'une série d'indicateurs qui ne donne qu'une vue partielle des données : effectif, moyenne, médiane, variance, écart type, minimum, maximum, étendue, 1er quartile, 3eme quartile, ...

  Ces indicateurs mesurent principalement la tendance centrale et la dispersion. On utilisera principalement la moyenne, la variance et l'écart type.

3.5. SD: paramètres de position :Moyenne arithmétique

- **Définition** : **La moyenne arithmétique** d'une série brute numérique x_1 , x_2 , ... , x_n est le quotient de la somme des observations par leur nombre
- On note

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

ou pour des données pondérées

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

- **Propriétés**
  la moyenne arithmétique est une mesure de tendance centrale qui dépend de toutes les observations et est sensible aux valeurs extrêmes. Elle est très utilisée a cause de ses bonnes propriétés mathématiques.

3.5. SD: paramètres de dispersion: Variance et ecart-type

- **Définition** : calculés généralement en complément de la moyenne, pour mesurer la plus ou moins grande dispersion autour de celle-ci

la variance de x est définie par

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad ou \qquad s_x^2 = \sum_{i=1}^{n} p_i (x_i - \bar{x})^2$$

L'écart type s_x est la racine carrée de la variance.

- **Propriétés**

  La variance satisfait la formule suivante

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n} p_i x_i^2 - (\bar{x})^2$$

La variance est « la moyenne des carres moins le carre de la moyenne ». L'ecart-type, qui a la même unité que x, est une mesure de dispersion.

3.5. Distribution statistique à deux variables: Mesure de liaison entre deux variables

- Relations entre deux caractères quantitatifs

- Covariance

- Coefficient de corrélation linéaire de BRAVAIS-PEARSON

- relations entre deux caractères qualitatifs

- Le khi-deux

- relations entre caractères quantitatifs et qualitatifs

- Le rapport de corrélation théorique

- Le rapport de corrélation empirique

3.5. Distribution statistique à deux variables: Mesure de liaison entre deux variables

- Définitions: la **covariance** observée entre deux variables x et y est

$$s_{xy} = \sum_{i=1}^{n} p_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} p_i x_i y_i - \overline{xy}$$

et le **cœfficient** de r de Bravais-Pearson ou coefficient de corrélation est donnée par

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n} p_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} p_i (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} p_i (y_i - \bar{y})^2}}$$

3.5. Propriétés du coefficient de corrélation

- **La covariance est positive** si X et Y ont tendance à varier dans le même sens, et négative si elles ont tendance à varier en sens contraire
- **La covariance** ne dépend pas de l'origine choisie pour X et Y, mais dépend des unités de mesure. C'est pourquoi, pour mesurer l'aspect plus ou moins "allongé" du nuage dans une direction, par un coefficient sans unité : C'est le **coefficient de corrélation** linéaire
- Ce coefficient, symétrique en X et Y, indépendant des unités choisies pour X et Y, et de l'origine, est toujours **compris entre - 1 et 1**.

  - $|r_{xy}| = 1$ si et seulement si x et y sont **linéairement liées**
  En particulier, $r_{xx} = 1$.

  - si $r_{xy} = 0$, on dit que les variables sont **de -corrélées ou indépendants**.

3.5. Corrélation et liaison significative

- **Problème**
  A partir de quelle valeur de r_{xy} peut-on considérer que les variables x et y sont liées?

- **Domaine d'application**
  on se place dans le cas ou le nombre d'individus est n > 30.

- **Méthode**
  si x et y sont deux variables gaussiennes indépendantes, alors on peut montrer que

$$\frac{(n-2)r_{xy}^2}{1-r_{xy}^2}$$

  suit une loi de Fischer-Snedecor F(1; n-2). Le résultat est valable dans le cas non gaussien pour n > 30.

3.5. Le test

- on se fixe un risque d'erreur (0,01 ou 0,05 en général) et on calcule la probabilité

$$P(F(1,n-2) > \frac{(n-2)r_{xy}^2}{1-r_{xy}^2}) = \pi$$

- Si $\pi < \alpha$ on considère que l'événement est trop improbable et que donc que l'hypothèse originale d'indépendance doit être rejetée au seuil . On trouvera en général ces valeurs dans une table pré-calculée de la loi F.

3.6. les tableaux

les populations comprennent des individus distingués selon un certain nombre de variables. ces informations sont rassemblées dans des tableaux de base croisant individus et variables. ces tableaux peuvent s'interpréter de deux façons, un nuage d'individus dans un ensemble de variables ou un nuage de variables dans un ensemble d'individus.

.

3.6. exemple :Tableau de données

- Pour n individus et p variables, on a le tableau
  X est une matrice rectangulaire a n lignes et p colonnes

$$X = (x^1, \ldots, x^p) = \begin{bmatrix} x_1^1 & x_1^2 & \cdots & x_1^p \\ x_2^1 & x_2^2 & & \\ & \ddots & & \\ \vdots & & x_i^j & \vdots \\ & & & \ddots \\ x_n^1 & & \cdots & x_n^p \end{bmatrix}$$

3.6. Vecteurs variable et individu

- **Variable**
  Une colonne du tableau

$$x^j = \begin{bmatrix} x_1^j \\ x_2^j \\ \\ \\ x_n^j \end{bmatrix}$$

- **Individu**
  Une ligne du tableau

$$e_i' = (x_i^1 \quad x_i^2 \quad \quad x_i^p)$$

3.6. les tableaux

- Tableaux individus x variables quantitatives
- Tableaux logiques ou booléens ou binaires

- Tableaux disjonctifs complet : individu x variable
à chaque modalité, placée en colonne, correspond une variable indicatrice. c'est la juxtaposition de plusieurs tableaux logiques.
x'x est une matrice diagonale dont les éléments sont les effectifs de chaque modalité.

.

3.6. les tableaux

-TABLEAUX PRÉSENCE ABSENCE

-TABLEAUX DE DONNÉES ORDINALES OU DE PRÉFÉRENCES
INDIVIDUS X OBJETS À CLASSER. UNE CASE CORRESPOND À UNE NOTE VARIANT DE 1 AU NOMBRE D'OBJETS À CLASSER

-TABLEAU DE DISTANCES OU DE PROXIMITÉS : INDIVIDUS X INDIVIDUS

IL PRÉSENTE LES DISTANCES ENTRE LES INDIVIDUS. CES TABLEAUX SONT SYMÉTRIQUE AUTOUR DE LA DIAGONALE PRINCIPALE.
TABLEAUX DE CONTINGENCE : VARIABLE X VARIABLE
IL CROISE LES MODALITÉS DE DEUX VARIABLES QUALITATIVES

-TABLEAUX DE BURT : IL CROISE LES MODALITÉS DE PLUS DE 2 VARIABLES QUALITATIVES. IL EST SYMÉTRIQUE.

-TABLEAUX DES RANGS

-TABLEAUX HÉTÉROGÈNES OU MIXTES
INDIVIDUS X VARIABLES LES VARIABLES SONT DE DIFFÉRENTES NATURES

SOIT LES VARIABLES SONT DÉJÀ DES CLASSEMENTS, SOIT POUR LES VARIABLES QUANTITATIVES ON REMPLACE LES VALEURS PAR LEUR RANG.

4. Les analyses factorielles

4.1 L'analyse en composantes principales (ACP)

4.2 L'analyse factorielle des correspondances (AFC)

4.3 L'analyse des correspondances multiples ACM

4.4 L'analyse factorielle des similarités (ou de dissimilarités) et des préférences

4.5 L'analyse discriminante (AFD)

4.6 L'analyse des mesures conjointes

4.7 L'analyse canonique

5 Les méthodes de classification, de typologie ou de taxinomie

5.1 L'analyse non hiérarchique

5.2 L'ANALYSE HIÉRARCHIQUE

.

BONNE COMPRÉHENSION