Institut Supérieur d'Informatique et de Multimédia de Sfax

Auditoire: 2ème année LSI-ADBD

Chapitre 2:

Introduction aux méthodes multivariées

Souhir BOUAZIZ AFFES
souhir.bouaziz@isims.usf.tn

Amal ABBES
amal.abbes@isims.usf.tn

Plan

- Introduction
- La statistique
  - Définition, Vocabulaire, Méthodes, Types
- Statistique Multivariée
  - Définition, Généralités
- Paramètres Fondamentaux
  - Paramètres de position, de dispersion, et de relation
- Analyse statistique multivariée
  - Méthodes descriptives ou exploratoires
  - Méthodes explicatives et prédictives

Introduction

- L'**analyse multivariée** désigne un ensemble de **méthodes** et de **techniques** pour l'étude de tableaux de **plusieurs variables** décrivant plusieurs individus.
- Le but est de donner un panorama des méthodes pour aider au choix de méthodes adéquates en fonction du ***type de données*** ou de la ***problématique à étudier***.

Statistique

Statistique Multivariée

Analyse Multivariée

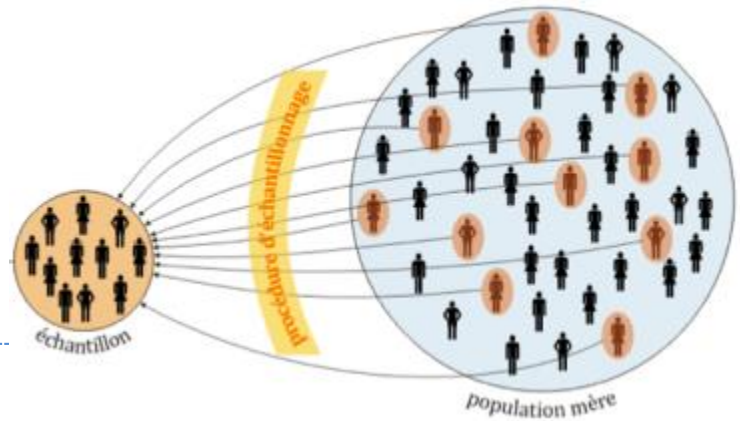
Analyse des données

La statistique : Définition

- La statistique est la discipline qui étudie des phénomènes à travers la collecte de données, leur traitement, leur analyse, l'interprétation des résultats et leur présentation afin de rendre ces données compréhensibles par tous.
- La statistique est une méthode scientifique qui consiste à observer et à étudier une/plusieurs propriété(s) commune(s) chez un groupe d'être, de choses, ou d'entités.
- La ***statistique*** est à différencier d'**une statistique**, qui est un nombre calculé à partir d'une population.

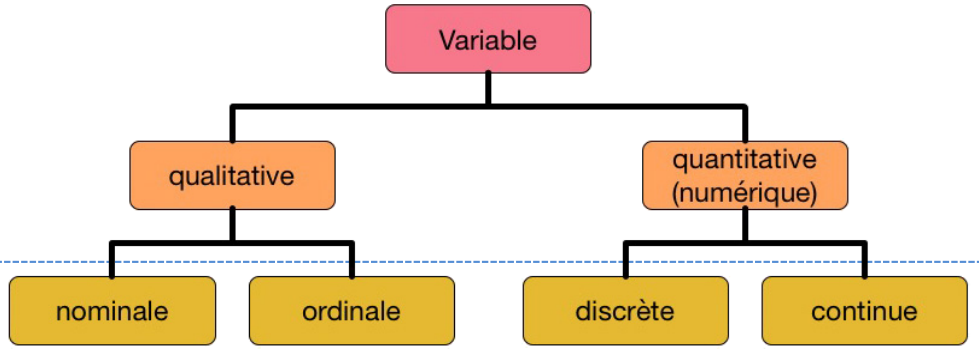
La statistique : Vocabulaire

- **Population (ou population statistique) :** la collection (d'être, de choses, ou d'entités) ayant des propriétés communes.
  - Terme hérité d'une des premières applications de la statistique, la démographie
  - **Exemple:** ensemble parcelles de terrain étudiées, population d'insectes, ensemble des plantes d'une espèce donnée, population d'humains, etc.
- **Individu (ou unité statistique) :** élément de la population
  - **Exemple:** une parcelle, un insecte, une plante, un humain, etc.
- **Échantillon :** individus de la population (sous-ensemble) sur laquelle les mesures ont été faites



La statistique : Vocabulaire

- **Variable (statistique) :** une des propriétés communes aux individus que l'on souhaite étudier. Peut-être :
  - **Variable quantitative :** numérique continue ou discrète
    - ses valeurs sont des ***nombres*** exprimant une ***quantité***, sur lesquels les opérations arithmétiques (somme, etc...) ont un sens.
  - **Variable qualitative :** variable ordinale ou nominal
    - ses valeurs sont des ***modalités*** exprimant une ***qualité***, sur lequel des opérations arithmétiques n'ont aucun sens. Il est possible de les répartir en classes ou catégories
    - Les modalités d'une variable sont l'ensemble des valeurs qu'elle prend dans les données. **Ex :** *Sexe* : féminin/ masculin; *Couleur* : bleu, verte,...



La statistique : Vocabulaire

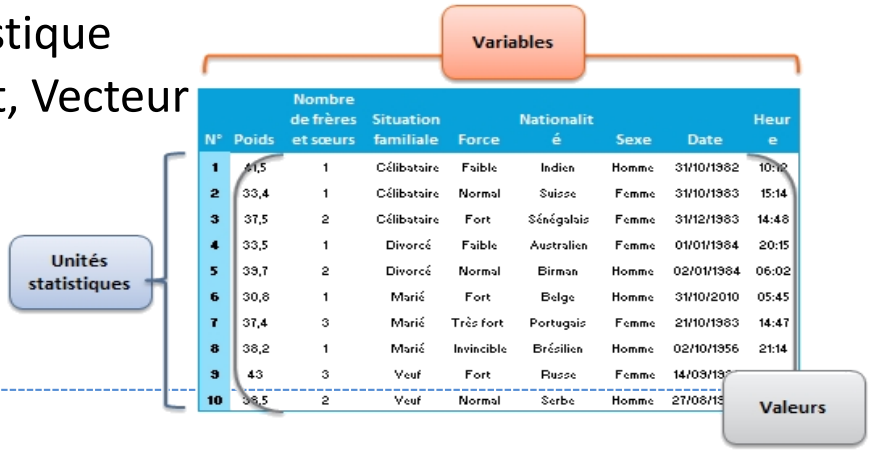
- **Variable quantitative :**
  - **Variable numérique continue :** peut prendre n'importe quelle valeur réelle : prendre un nombre infini ou non dénombrable de valeurs
    - **Ex :** le taux d'acidité du sol, la longueur de l'insecte, la longueur de la tige, l'indice de masse corporelle
  - **Variable numérique discrète :** prendre un nombre fini ou dénombrable de valeurs : dès qu'il y a un saut minimum obligatoire entre deux valeurs successives, **Ex.** les nombres entiers
    - **Ex :** la somme (sur tous les jours) du nombre de vaches présentes sur la parcelle, l'âge de l'insecte (en jours), le nombre de pétales sur la fleur, le nombre d'année d'études (réussies) depuis la petite école

La statistique : Vocabulaire

- **Variable qualitative :**
  - **Variable nominale :** quand ses valeurs sont des éléments d'une catégorie type nom **non hiérarchique**
    - ses éléments ne peuvent pas se ranger dans une gradation logique, selon une hiérarchie naturelle.
    - **Ex :** couleur des pétales {rouge, blanc, jaune}
  - **Variable ordinale :** quand ses valeurs sont des éléments d'une catégorie type nom **hiérarchique**
    - il sera possible de ranger dans une gradation logique, selon une hiérarchie naturelle, les individus de la population étudiée pour le caractère retenu.
    - **Ex :** appréciation de la parcelle {Mauvais, Passable, Bien, Très bien}
  - **Variable binaire ou variable dichotomique :** Il s'agit d'un type particulier de variables catégorielles. Elle ne peut pas être hiérarchisée
    - Elle ne possède que deux modalités (deux classes) possibles.
    - **Ex :** sexe {Masculin, Féminin}

La statistique : Vocabulaire

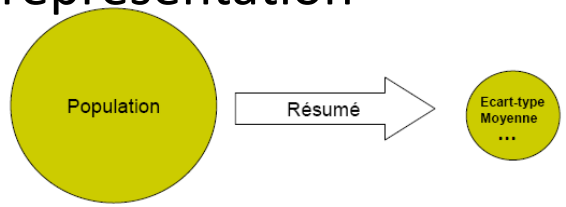
- **Données (statistiques) :** ensemble des individus observés (échantillon), des variables considérées, et des observations de ces variables sur ces individus.
  - Elles sont en général présentées sous forme de tableaux (**individus** en lignes et **variables** en colonnes)
  - Lorsqu'un tableau ne comporte que des **nombres** (valeurs des variables quantitatives ou codes associés aux variables qualitatives), il correspond à la notion mathématique de **matrice**.
  - Il existe des alternatives pour les colonnes, les lignes et les valeurs:
    - Colonne, Attribut, Variable, Caractéristique
    - Ligne, Enregistrement, Individu, Objet, Vecteur
    - Valeur, Observation, Donnée



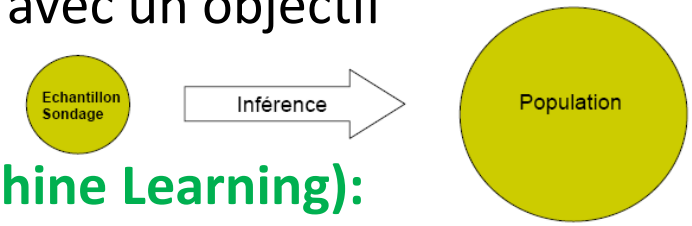
N°	Poids	Nombre de frères et sœurs	Situation familiale	Force	Nationalité	Sexe	Date	Heure
1	31,5	1	Célibataire	Faible	Indien	Homme	31/10/1982	10:12
2	33,4	1	Célibataire	Normal	Suisse	Femme	31/10/1983	15:14
3	37,5	2	Célibataire	Fort	Sénégalais	Femme	31/12/1983	14:48
4	33,5	1	Divorcé	Faible	Australien	Femme	01/01/1984	20:15
5	39,7	2	Divorcé	Normal	Birman	Homme	02/01/1984	06:02
6	30,8	1	Marié	Fort	Belge	Homme	31/10/2010	05:45
7	37,4	3	Marié	Très fort	Portugais	Femme	21/10/1983	14:47
8	38,2	1	Marié	Invincible	Brésilien	Homme	02/10/1956	21:14
9	43	3	Veuf	Fort	Russe	Femme	14/03/13	
10	38,5	2	Veuf	Normal	Serbe	Homme	27/08/19	

La statistique : Méthodes

- **La statistique descriptive:** décrire (résumer ou représenter) des données étudiées à travers leur présentation, leur représentation graphique, et le calcul de résumés numériques.



- **La statistique inférentielle:** préciser un phénomène sur une population globale, à partir de son observation sur une partie restreinte de cette population, **l'échantillon**. Il s'agit donc d'induire (ou encore d'inférer) du particulier au général avec un objectif principalement explicatif.



- **L'apprentissage statistique (Statistique + Machine Learning):** construire un modèle statistique traditionnel ou algorithmique, en privilégiant soit la description des données, ou la prévision d'une variable qualitative (discrimination ou classification supervisée) ou quantitative (régression).

La statistique : Types

- **La statistique univariée:**
  - à chaque individu de la population correspond une seule variable statistique.
  - **Exemple:** *Population* : iris. *Variable* : longueur des pétales.

- **La statistique multivariée:**
  - à chaque individu de la population correspond au moins deux variables statistiques.
  - Dans le cas de deux variables, ça sera notée ***statistique bivariée***
  - **Exemple:** *Population* : iris. *Variable 1* : longueur des pétales.
    Variable 2 : largeur des pétales.

Statistique Multivariée : Définition

- La statistique **multivariée** ou **multidimensionnelle** est l'ensemble des méthodes de la statistique permettant de traiter simultanément un nombre quelconque de variables
  - il s'agit d'aller au-delà de l'étude d'une seule (cas de la statistique univariée) ou de deux variables (cas de la statistique bivariée)

- **Analyse des données ou l'analyse multivariée**
  - traitement de données en masse : grand nombre de variables et d'individus
  - vision globale **multidimensionnelle** des individus et des variables
  - représentations géométriques, création de nouvelles variables

Statistique Multivariée : Généralités

- Avant de pouvoir analyser les données multidimensionnelles, il faut un moyen pour les répertorier.
- L'outil naturel est d'utiliser une **matrice** X, appelée **matrice des données** ou encore tableau de données.

-Tableau individus x variables comportant des variables numériques et une variable dichotomique

	Age	Etat-Civil	Feministe	Frequence	Agressivite	Harcelement
1	13	1	102	2	4	**0**
2	45	2	101	3	6	**0**
3	19	2	102	2	7	**1**
4	42	2	102	1	2	**1**
5	27	1	77	1	1	**0**
6	19	1	98	0	6	**1**
7	37	1	96	1	6	**0**

Questions à réponses fermées : sexe (2 modalités), niveau de revenu (2 modalités), préférence (3 modalités)

	1 Sexe	2 Revenu	3 Preference
s1	F	M	A
s2	F	M	A
s3	F	E	B
s4	F	E	C
s5	F	E	C
s6	H	E	C
s7	H	E	B
s8	H	M	B
s9	H	M	B
s10	H	M	A

Statistique Multivariée : Généralités

- **Matrice des données:**
  - On suppose que l'on a n **individus**, et que pour chacun de ces individus, on observe p **variables**. Alors, les données sont répertoriées comme suit :

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

Lorsque n et p deviennent grands, le nombre de données np est grand → Techniques pour résumer et analyser ces données.

  - x_{ij}: l'observation de la $j^{ème}$ variable pour l'individu i.
  - X_i^t: la $i^{ème}$ ligne de X représentant les données de toutes les variables pour le $i^{ème}$ individu $\quad X_i^t = (x_{i1}, \cdots, x_{ip})$
  - $X_{(j)}$: la $j^{ème}$ colonne de X représentant les données de la $i^{ème}$ variable pour tous les individus. Ainsi,

$$X_{(j)} = \begin{pmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{pmatrix}$$

Statistique Multivariée : Généralités

- La matrice de données peut être considérée de deux points de vue :
  - si on compare des colonnes, alors on étudie la relation entre les variables correspondantes. (comme la **matrice de contingence**: croisement de 2 variables ***qualitatives***)
  - si on compare des lignes, on étudie la relation entre des individus. (comme l'étude des **paramètres fondamentaux** dans le cas des variables ***quantitatives***)
- Différents types de matrices de données:
  - **Tableau (matrice) de contingence :** croisement de 2 variables qualitatives
  - **Matrices de préférences entre objets :** les variables sont les objets et chaque individu range ces objets par ordre de préférence décroissante.
  - **Matrices de distances :** tableaux des nxn distances entre individus
  - **Autres types de matrices :** matrices de présence/absence, matrices de notes, matrices de pourcentage, etc.

Statistique Multivariée : Généralités

- **Exemple de tableau de contingence:**
  - $n = 4$ individus, $p = 3$ variables qualitatives à 3, 3 et 2 modalités
  - Tableau brut de données :

$$\begin{matrix} & \boldsymbol{X_1} & \boldsymbol{X_2} & \boldsymbol{X_3} \\ \begin{matrix} A \\ B \\ B \\ C \end{matrix} & \begin{matrix} 3 \\ 1 \\ 2 \\ 1 \end{matrix} & \begin{matrix} 1 \\ 0 \\ 1 \\ 1 \end{matrix} \end{matrix}$$

  - Transformation → tableau **disjonctif complet**

$$X = \left[\begin{array}{ccc|ccc|cc|c} 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 3 & p \\ 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 3 & p \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 3 & p \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 3 & p \\ 1 & 2 & 1 & 2 & 1 & 1 & 1 & 3 & 12 & np \end{array}\right]$$

$$X = (X_1 \mid X_2 \mid \ldots \mid X_p)$$

indicatrices des modalités

Tableau de **contingence** $\boldsymbol{X_1 * X_3}$

$$\begin{bmatrix} 0 & 1 \\ 1 & 1 \\ 0 & 1 \end{bmatrix}$$

Statistique Multivariée : Généralités

- Pour résumer l'ensemble de données, nous recherchons une mesure qui peut caractériser l'ensemble de données : **mesure descriptive**
  - calculée à partir des données de population: ***un paramètre***
  - calculée à partir des données de l'échantillon: ***une statistique***
- Trois types de paramètres fondamentaux dans l'approche multivariée des différences individuelles:
  - **Paramètres de position ou de tendance centrale:** *Quelle est la valeur typique de chaque variable ?* → **Moyenne arithmétique, Médiane**
  - **Paramètres de dispersion :** *Quelle est la distance entre les observations individuelles et la valeur centrale d'une variable donnée ?* → **Variance, Écart type, Variables centrées- réduites**
  - **Paramètres de relation:** *Lorsque plusieurs variables sont étudiées ensemble, comment chaque variable est-elle liée aux autres variables ? Comment les variables sont-elles simultanément liées les unes aux autres ? Sont-elles liées positivement ou négativement ?* → **Covariance, Corrélation**

Paramètres Fondamentaux

Paramètres de position

- **Moyenne arithmétique:** la somme des valeurs d'une variable divisée par le nombre total de valeurs
  - $\overline{X_{(j)}}$: La moyenne arithmétique des données $X_{(j)}$ de la $j^{ème}$ variable :

$$\overline{X_{(j)}} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$$

  - On peut représenter les p moyennes arithmétiques des données des p variables sous la forme ***du vecteur ligne des moyennes arithmétiques***, noté $\bar{X}^t$:

$$\overline{\mathbf{x}}^t = (\overline{X_{(1)}}, \cdots, \overline{X_{(p)}})$$

  - **Exemple:**

$$X = \begin{pmatrix} 11 & 13.5 \\ 12 & 13.5 \\ 13 & 13.5 \\ 14 & 13.5 \\ 15 & 13.5 \\ 16 & 13.5 \end{pmatrix} \qquad \overline{\mathbf{x}}^t = \left(\frac{11+\cdots+16}{6}, \frac{13.5+\cdots+13.5}{6}\right) = (13.5, 13.5)$$

Paramètres Fondamentaux

- **Paramètres de position ...**
  - **Médiane:** la valeur médiane d'une variable
    - On suppose que les valeurs des données $X_{(j)}$ de la $j^{ème}$ variable sont classées en ordre croissant.
    - Si n est impair, la médiane, notée $m_{(j)}$, est l'*élément du milieu:*

$$m_{(j)} = x_{\frac{n+1}{2}j}$$

    - Si n est pair, on prendra par convention :

$$m_{(j)} = \frac{x_{\frac{n}{2},j} + x_{\frac{n}{2}+1,j}}{2}$$

    - Le ***vecteur ligne des médianes***, noté m^t

$$\mathbf{m}^t = (m_{(1)}, \cdots, m_{(p)})$$

    - **Exemple:** Le vecteur ligne des médianes pour l'exemple des notes est :

$$\mathbf{m}^t = \left(\frac{13+14}{2}, \frac{13.5+13.5}{2}\right) = (13.5, 13.5)$$

Paramètres Fondamentaux

Paramètres de dispersion

- La moyenne ne donne qu'une information partielle.
- Il est aussi important de pouvoir mesurer combien ces données sont dispersées autour de la moyenne.
- Revenons à l'exemple des notes, les données des deux variables ont la même moyenne, mais vous sentez bien qu'elles sont de nature différente.
- Il existe plusieurs manières de mesurer la dispersion des données:
  - Étendue
  - Variance et écart type
  - Variables centrées-réduites

Paramètres Fondamentaux

Paramètres de dispersion ...

Étendue

- $w_{(j)}$: l'étendue des données $X_{(j)}$ de la $j^{ème}$ variable : la différence entre la donnée la plus grande pour cette variable, et la plus petite

$$X_{(j)}^{\max} = \max_{i \in \{1,\cdots,n\}} x_{ij}$$

$$X_{(j)}^{\min} = \min_{i \in \{1,\cdots,n\}} x_{ij}$$

Alors:

$$w_{(j)} = X_{(j)}^{\max} - X_{(j)}^{\min}$$

- On peut représenter les p étendues sous la forme d'un vecteur ligne: ***vecteur ligne des étendues***, et noté w^t:

$$\mathbf{w}^t = (w_{(1)}, \cdots, w_{(p)})$$

- **Exemple:** Le vecteur ligne des étendues pour l'exemple des notes est :

$$\mathbf{w}^t = (5, 0)$$

Paramètres Fondamentaux

- **Paramètres de dispersion …**
  - **Étendue …**

> - L'étendue est un indicateur instable étant donné qu'il ne dépend que des valeurs extrêmes.
> - Vous pouvez avoir un grand nombre de données qui sont similaires, mais qui ont une plus grande et plus petite valeur qui sont très différentes,
>   - auront alors une étendue très différente,
>   - mais cela ne représente pas bien la réalité des données.

Paramètres Fondamentaux

- **Paramètres de dispersion …**
  - **Variance et l'écart type**
    - $Var(X_{(j)})$: la variance des données $X_{(j)}$ de la $j^{ème}$ variable

$$\mathrm{Var}(X_{(j)}) = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{X_{(j)}})^2 = \frac{1}{n}\left[(x_{1j} - \overline{X_{(j)}})^2 + \cdots + (x_{nj} - \overline{X_{(j)}})^2\right]$$

    - Pour compenser le fait que l'on prenne des carrés, on peut reprendre la racine, et on obtient alors l'écart-type :

$$\sigma(X_{(j)}) = \sqrt{\mathrm{Var}(X_{(j)})} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{X_{(j)}})^2}.$$

    - **Exemple:** calcul des variances et des écart-types pour l'exemple des notes:

$$\mathrm{Var}(X_{(1)}) = \frac{1}{6}\left((11 - 13.5)^2 + (12 - 13.5)^2 + \cdots + (16 - 13.5)^2\right) = 2.917$$
$$\sigma(X_{(1)}) = 1.708$$
$$\mathrm{Var}(X_{(2)}) = \frac{1}{6}\left(6(13.5 - 13.5)^2\right) = 0$$
$$\sigma(X_{(2)}) = 0.$$

Paramètres Fondamentaux

Paramètres de dispersion …

Variance et l'écart type : Notation matricielle …

- Considérant les notions suivantes:
  - La ***norme*** d'un vecteur x, notée $\|x\|$ et $< \mathrm{x}, \mathrm{x} >$ est le ***produit scalaire*** :

$$\|\mathrm{x}\| = \sqrt{x_1^2 + \cdots + x_n^2} = \sqrt{< \mathrm{x}, \mathrm{x} >} \qquad \text{tel que: } < \mathrm{x}, \mathrm{x} >= (x_1 \ldots x_n) \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = x^t x$$

  - La ***matrice des moyennes arithmétiques***

$$\overline{X} = \begin{pmatrix} \overline{X_{(1)}} & \cdots & \overline{X_{(p)}} \\ \vdots & \vdots & \vdots \\ \overline{X_{(1)}} & \cdots & \overline{X_{(p)}} \end{pmatrix} \Rightarrow X - \overline{X} = \begin{pmatrix} x_{11} - \overline{X_{(1)}} & \cdots & x_{1p} - \overline{X_{(p)}} \\ \vdots & \vdots & \vdots \\ x_{n1} - \overline{X_{(1)}} & \cdots & x_{np} - \overline{X_{(p)}} \end{pmatrix}$$

- la variance peut s'écrit comme la ***norme d'un vecteur***:

$$\mathrm{Var}(X_{(j)}) = \frac{1}{n} < (X - \overline{X})_{(j)}, (X - \overline{X})_{(j)} >= \frac{1}{n}((X - \overline{X})_{(j)})^t (X - \overline{X})_{(j)} = \frac{1}{n}\|(X - \overline{X})_{(j)}\|^2$$

Paramètres Fondamentaux

- **Paramètres de dispersion …**
  - **Variance et l'écart type : Notation matricielle ….**
    - L'écart type peut s'écrit aussi comme la ***norme d'un vecteur***:

$$\sigma(X_{(j)}) = \frac{1}{\sqrt{n}}\|(X - \overline{X})_{(j)}\|$$

    - **Exemple:** Réécrivons la variance pour l'exemple des notes en notation matricielle

$$\overline{X} = \begin{pmatrix} 13.5 & 13.5 \\ 13.5 & 13.5 \\ 13.5 & 13.5 \\ 13.5 & 13.5 \\ 13.5 & 13.5 \\ 13.5 & 13.5 \end{pmatrix}, \text{ et } \quad X - \overline{X} = \begin{pmatrix} -2.5 & 0 \\ -1.5 & 0 \\ -0.5 & 0 \\ 0.5 & 0 \\ 1.5 & 0 \\ 2.5 & 0 \end{pmatrix}$$

$$\mathrm{Var}(X_{(1)}) = \frac{1}{6}\left\langle \begin{pmatrix} -2.5 \\ -1.5 \\ -0.5 \\ 0.5 \\ 1.5 \\ 2.5 \end{pmatrix}, \begin{pmatrix} -2.5 \\ -1.5 \\ -0.5 \\ 0.5 \\ 1.5 \\ 2.5 \end{pmatrix} \right\rangle = \frac{1}{6}\left\| \begin{pmatrix} -2.5 \\ -1.5 \\ -0.5 \\ 0.5 \\ 1.5 \\ 2.5 \end{pmatrix} \right\|^2 = 2.917$$

Paramètres Fondamentaux

Paramètres de dispersion …

Variables centrées-réduites

- Les données d'une variable sont dites ***centrées*** si on leur soustrait leur moyenne. Elles sont dites ***centrées réduites*** si elles sont centrées et divisées par leur écart-type.
- Elles sont utiles car elles n'ont plus **d'unité**, et des données de variables différentes deviennent ainsi comparables.
- Z la matrice des données centrées réduites

$$(Z)_{ij} = z_{ij} = \frac{x_{ij} - \overline{X_{(j)}}}{\sigma(X_{(j)})}$$

Si $\sigma(X_{(j)})$=0, on a aussi $x_{ij} - \overline{X_{(j)}} = 0 \rightarrow z_{ij} = 0$

- **Exemple:** matrice des données centrées réduites de l'exemple des notes

$$\sigma(X_{(1)}) = 1.708 \quad \sigma(X_{(2)}) = 0$$
$$\overline{X_{(1)}} = 13.5 \quad \overline{X_{(2)}} = 13.5$$

$$Z = \begin{pmatrix} \frac{11-13.5}{1.708} & 0 \\ \frac{12-13.5}{1.708} & 0 \\ \frac{13-13.5}{1.708} & 0 \\ \frac{14-13.5}{1.708} & 0 \\ \frac{15-13.5}{1.708} & 0 \\ \frac{16-13.5}{1.708} & 0 \end{pmatrix} = \begin{pmatrix} -1.464 & 0 \\ -0.878 & 0 \\ -0.293 & 0 \\ 0.293 & 0 \\ 0.878 & 0 \\ 1.464 & 0 \end{pmatrix}$$

Paramètres Fondamentaux

Paramètres de relation entre deux variables

Covariance

- Pour tout i et j compris entre 1 et p, on définit la covariance entre les données $X_{(i)}$ et $X_{(j)}$ des $i^{ème}$ et $j^{ème}$ variables par :

$$\mathrm{Cov}(X_{(i)}, X_{(j)}) = \frac{1}{n} < (X - \overline{X})_{(i)}, (X - \overline{X})_{(j)} >= \frac{1}{n}((X - \overline{X})_{(i)})^t (X - \overline{X})_{(j)}$$

- **Théorème 1 (Köning-Huygens)**

$$\mathrm{Cov}(X_{(i)}, X_{(j)}) = \left(\frac{1}{n} < X_{(i)}, X_{(j)} >\right) - \overline{X_{(i)}}\,\overline{X_{(j)}}.$$

- **Exemple:** la covariance entre les données des 1ère et 2ème variables de l'exemple des notes, en utilisant le Théorème de König-Huygens

$$\mathrm{Cov}(X_{(1)}, X_{(2)}) = \frac{1}{6}[11 \cdot 13.5 + 12 \cdot 13.5 + 13 \cdot 13.5 + \cdots + 16 \cdot 13.5] - 13.5^2 = 0$$

Paramètres Fondamentaux

- **Paramètres de relation entre deux variables …**
  - **Covariance …**
    - ***Remarques:***
      - $\mathrm{Cov}\left(X_{(i)}, X_{(j)}\right) = \frac{1}{n}\left((X - \bar{X})^t (X - \bar{X})\right)_{ij}$: est le coefficient (i, j) de la matrice $\frac{1}{n}(X - \bar{X})^t (X - \bar{X})$
      - $\mathrm{Cov}\left(X_{(i)}, X_{(i)}\right) = \mathrm{Var}\left(X_{(i)}\right)$
      - La covariance est symétrique, i.e. : $\mathrm{Cov}\left(X_{(i)}, X_{(j)}\right) = \mathrm{Cov}\left(X_{(j)}, X_{(i)}\right)$
      - Dans le cas de la variance, le Théorème de Köning-Huygens s'écrit :

$$\mathrm{Var}(X_{(j)}) = \left(\frac{1}{n}\|X_{(j)}\|^2\right) - \overline{X_{(j)}}^2$$

Paramètres Fondamentaux

- **Paramètres de relation entre deux variables …**
  - **Covariance : matrice de covariance …**
    - Les covariances sont naturellement répertoriées dans *la matrice de covariance des données* X, de taille $p \times p$, notée $C(X)$, définie par :

$$C(X) = \frac{1}{n}(X - \bar{X})^t (X - \bar{X})$$

$$\mathrm{Cov}(X_{(i)}, X_{(j)}) = (C(X))_{ij}$$

    - Remarquer que les coefficients sur la diagonale de la matrice $C(X)$ donnent les variances.
    - **Exemple:** la matrice de covariance pour l'exemple des notes:

$$C(X) = \frac{1}{6}(X - \overline{X})^t (X - \overline{X})$$

$$= \frac{1}{6}\begin{pmatrix} -2.5 & -1.5 & -0.5 & 0.5 & 1.5 & 2.5 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} -2.5 & 0 \\ -1.5 & 0 \\ -0.5 & 0 \\ 0.5 & 0 \\ 1.5 & 0 \\ 2.5 & 0 \end{pmatrix} = \begin{pmatrix} 2.91667 & 0 \\ 0 & 0 \end{pmatrix}$$

Paramètres Fondamentaux

- **Paramètres de relation entre deux variables …**
  - **Covariance : matrice de covariance …**
    - La ***variabilité totale*** de la matrice des données X est par définition :

$$\mathrm{Tr}(C(X)) = \sum_{i=1}^{p} \mathrm{Var}\left(X_{(i)}\right)$$

    - Cette quantité est importante car elle donne en quelque sorte la **quantité d'information** qui est contenue dans la matrice X.
    - Elle joue un rôle clé dans l'ACP.

Paramètres Fondamentaux

Paramètres de relation entre deux variables …

Corrélation de Bravais-Pearson

- On définit la corrélation de Bravais-Pearson entre les données $X_{(i)}$ et $X_{(j)}$ des $i^{ème}$ et $j^{ème}$ variables par :

$$r(X_{(i)}, X_{(j)}) = \frac{\mathrm{Cov}(X_{(i)}, X_{(j)})}{\sigma(X_{(i)})\,\sigma(X_{(j)})} = \frac{< (X - \bar{X})_{(i)} , (X - \bar{X})_{(j)} >}{\|(X - \bar{X})_{(i)}\| . \|(X - \bar{X})_{(j)}\|}$$

$$= \cos(\widehat{(X - \bar{X})_{(i)} , (X - \bar{X})_{(j)}})$$

Car on a:
$< x, y > = \|x\| . \|y\| \cos(\widehat{x, y})$

- Elle satisfait les propriétés :
  - $r(X_{(i)}, X_{(i)}) = 1$
  - $|r(X_{(i)}, X_{(j)})| \leq 1$
  - $|r(X_{(i)}, X_{(j)})| = 1$, *si et seulement si il existe un nbre* $a \in \mathbb{R}$, *tel que*:

$$(X - \overline{X})_{(j)} = a(X - \overline{X})_{(i)}$$

⇒ une dépendance linéaire entre les données $X_{(i)}$ *et* $X_{(j)}$

Paramètres Fondamentaux

- **Paramètres de relation entre deux variables …**
  - **Corrélation de Bravais-Pearson : matrice de corrélation …**
    - On définit *la matrice de corrélation des données* X, de taille $p \times p$, notée $R(X)$, par :

$$(R(X))_{ij} = r\left(X_{(i)}, X_{(j)}\right)$$

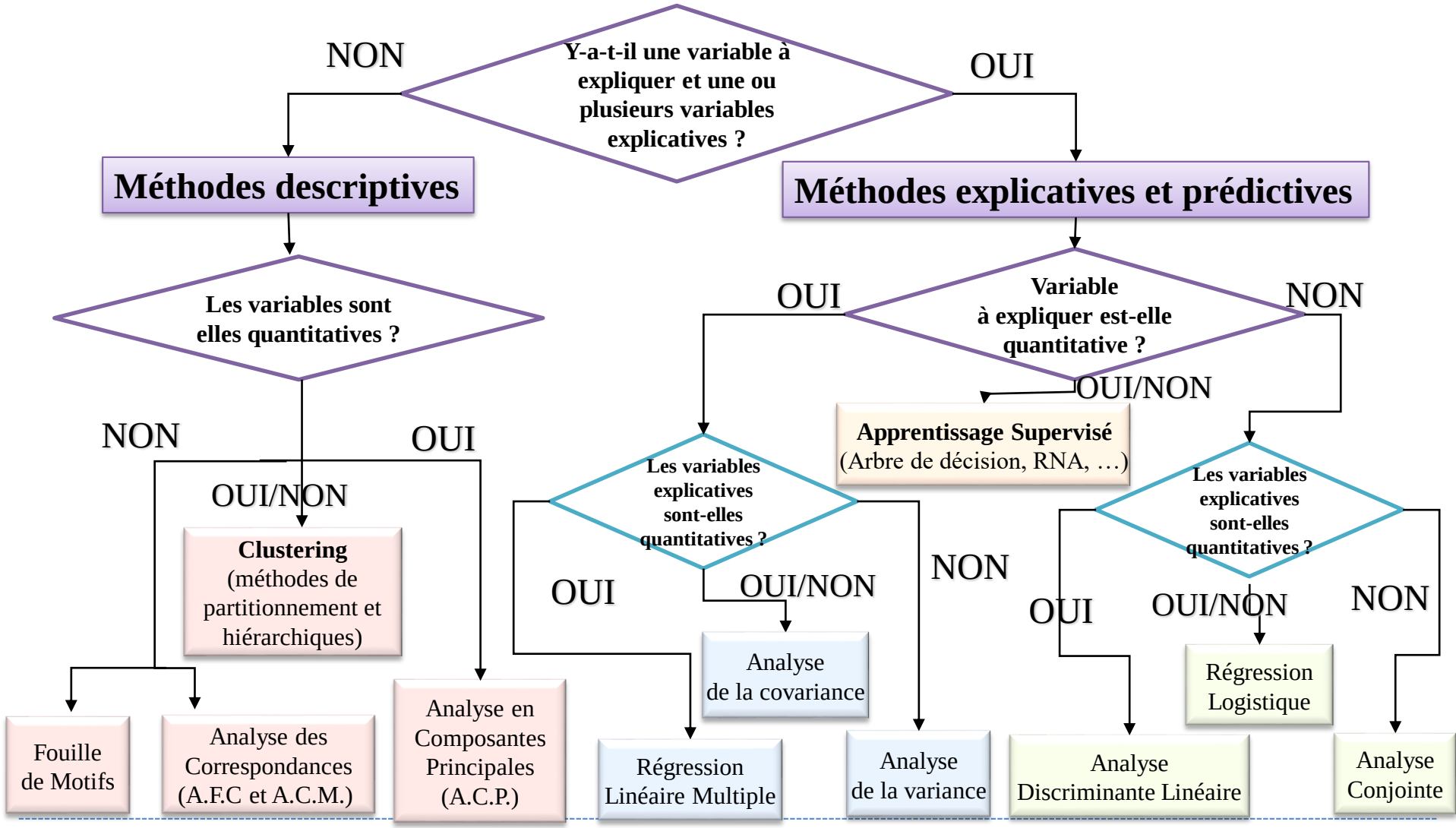
    - Remarquer que les éléments diagonaux de cette matrice sont tous égaux à 1.
    - **Exemple :** La matrice de corrélation de l'exemple des notes est :

$$R(X) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Analyse statistique multivariée

- L'exploration et l'analyse de données est l'étape cœur du processus de l'ECD, en vue d'en tirer l'information pertinente pour la compréhension du phénomène étudié et d'en aider à la prise de décision à travers plusieurs **méthodes** et **techniques**
- Ces différentes méthodes peuvent être classées selon l'objectif poursuivi :
  - **description :** but est de comprendre au mieux les données grâce à une description simplifiée aussi proche que possible de la réalité. (On étudie le tableau entier)
  - **explication et prévision :** but est d'expliquer et de prévoir une ou plusieurs variables du tableau en fonction d'autres variables. (tableau partitionné en 2)
    - **variables explicatives ou variables indépendantes:** dont on se sert pour expliquer et prédire le phénomène à étudier.
    - **variables à expliquer ou variables dépendantes:** dont on veut expliquer et prédire la variation dans une recherche

Analyse statistique multivariée



Analyse statistique multivariée

- **Méthodes descriptives ou exploratoires :**
  - **Méthodes factorielles**

    réduction du nombre de variables en les résumant par un petit nombre de composantes synthétiques appelés **facteurs** :
    - **Analyse en Composantes Principales (ACP):** pour les variables quantitatives
      - cherche à représenter dans un espace de dimension faible ($\ll p$) un nuage de points représentant n individus, décrits par p variables quantitatives (donc de dimension p) en utilisant les corrélations existant entre ces variables.

Analyse statistique multivariée

- **Méthodes descriptives ou exploratoires …**
  - **Méthodes factorielles …**

    réduction du nombre de variables en les résumant par un petit nombre de composantes synthétiques appelés **facteurs** :
    - **Analyse Factorielle des Correspondances simples (AFC):** pour 2 variables qualitatives
    - **Analyse des Correspondances Multiples (ACM):** pour plusieurs variables qualitatives
      - L'analyse des correspondances (AFC ou ACM) étudie les proximités entre individus décrits par deux ou plusieurs variables qualitatives ainsi que les proximités entre les modalités de ces variables.

Analyse statistique multivariée

- **Méthodes descriptives ou exploratoires …**
  - **Méthodes de regroupement**

    réduction du nombre d'individus par la formation de groupes homogènes :

    - **Méthodes de partitionnement:** en un nombre fixé de classes a priori: partitionnement des objets et évaluation des partitions : méthode des centres mobiles, K-means, nuées dynamiques
    - **Méthodes hiérarchiques:** suite de partitions emboîtées: décomposition hiérarchique d'ensembles d'objets: méthodes de classification ascendante hiérarchique (CAH)…

Analyse statistique multivariée

- **Méthodes descriptives ou exploratoires …**
  - **Méthodes d'association**

    déterminer l'ensemble de descripteurs (attributs) qui sont les plus corrélés :

    - **Association simple ou les règles d'association:** Recherche de relations « *stables* » existant entre les attributs d'un individu

    - **Association séquentiel:** recherche de corrélations entre attributs mais en prenant en compte le temps entre attributs => comportement

Analyse statistique multivariée

- **Méthodes explicatives et prédictives :**
  - **Modèle linéaire général :** recherche d'une relation entre une variable quantitative et plusieurs autres :
    - **Régression Linéaire Multiple:** variables quantitatives
    - **Analyse de la Variance (ANOVA):** variables qualitatives
    - **Analyse de la Covariance (ANCOVA):** variables mixtes (qualitatives et/ou quantitatives)
  - **Analyse discriminante prédictive:** prédiction d'une variable qualitative à l'aide de plusieurs prédicteurs en général quantitatives
    - **Analyse Discriminante Linéaire :** expliquer et prédire l'appartenance d'un individu à une classe (groupe) prédéfinie à partir de ses caractéristiques mesurées à l'aide de variables prédictives

Analyse statistique multivariée

- **Méthodes explicatives et prédictives …**
  - **Régression Logistique:** vise à construire un modèle permettant de prédire / expliquer les valeurs prises par une variable cible qualitative à partir d'un ensemble de variables explicatives quantitatives ou qualitatives (un codage est nécessaire dans ce cas)
    - **Régression logistique binaire:** si la variable à expliquer est binaire;
    - **Régression logistique polytomique:** si la variable à expliquer possède plus de 2 modalités
  - **Analyse conjointe:** permettant de recueillir les préférences des consommateurs sur les attributs d'un produit ou d'un service
    - la variable à expliquer reflète un choix ou une situation de compromis
    - les variables explicatives sont qualitatives, par conséquent codées comme des nombres binaires (0 et 1)

Analyse statistique multivariée

Méthodes explicatives et prédictives …

- **Méthodes d'Apprentissage Supervisée:** étudient la prévision d'une variable qualitative ou quantitative dépendante par une combinaison de variables explicatives. Les exemples annotés constituent **une base d'apprentissage**, et la fonction de prédiction apprise est appelée aussi « *hypothèse* » ou « *modèle* ».
  - **Arbre de décision:** le modèle d'apprentissage induit est un **arbre** qui peut être traduit sous forme de règle de décision pour :
    - prédire une étiquette (**arbre de classification**), ou
    - prédire une quantité réelle (**arbre de régression**)
  - **Réseaux de Neurones Artificiel :** visent à établir par apprentissage un modèle susceptible d'affecter à un jeu de test (différent du jeu d'apprentissage):
    - une **qualité** (reconnaissance d'image), une **valeur** (estimation numérique), appartenance à un groupe, ...