

Analyse des données

- 1 Introduction à l'analyse des données
- 2 Analyse univariée
- 3 Analyse bivariée
- 4 Ajustement linéaire
- 5 Analyse en composante principale (ACP)

Chapitre I :

I-Introduction
à l'analyse
des données

1-
Introduction :

2. Définitions

3. Les types
de variables

4. Les
échelles de
mesure

I-Introduction à l'analyse des données

1- Introduction :

I-Introduction
à l'analyse
des données

1-
Introduction :

2. Définitions

3. Les types
de variables

4. Les
échelles de
mesure

La statistique est une méthode scientifique qui consiste à réunir des données chiffrées sur des ensembles nombreux, puis à analyser, à commenter et à critiquer ces données. Classiquement les méthodes statistiques sont employées soit pour explorer les données (nommée statistique exploratoire) soit pour prédire un comportement (nommée statistique prédictive ou décisionnelle). La statistique exploratoire s'appuie sur des techniques descriptives et graphiques. Elle est généralement décrite par la statistique descriptive qui regroupe des méthodes exploratoires simples, uni- ou bidimensionnelle (moyenne, moments, quantiles, variance, corrélation, ...) et la statistique exploratoire multidimensionnelle.

1- Introduction : (suite)

I-Introduction
à l'analyse
des données

1-
Introduction :

2. Définitions

3. Les types
de variables

4. Les
échelles de
mesure

La statistique classique étudie les variables les unes après les autres, et elle construit autant de graphes (histogrammes) que de variables, mais les techniques dites d'analyse des données permettent de donner une vision globale de l'ensemble des variables.

L'analyse des données peut se définir comme l'ensemble des méthodes permettant une étude approfondie d'informations et de données de nature qualitative ou quantitative.

Dans l'analyse des données, on distingue :

- **l'analyse univariée**, qui porte sur l'étude d'une seule variable ;
- **l'analyse bivariée**, qui a pour objectif d'examiner les relations entre deux variables en même temps ;
- **l'analyse multivariée**, qui vise l'étude de plusieurs variables en même temps.

1- Introduction : (suite)

L'analyse des données recouvre principalement deux ensembles de techniques :

- 1 **analyses factorielles**, qui relèvent de la géométrie euclidienne et conduisent à l'extraction de valeurs et de vecteurs propres. Les méthodes les plus employées de cette technique sont :
 - i) la méthode de l'**analyse en composantes principales (ACP)**
 - ii) la méthode de l'**analyse factorielle des correspondances (AFC)**.
- 2 **classification automatique** sont caractérisées par le choix d'un indice de proximité et d'un algorithme d'agrégation ou de désagrégation qui permettent d'obtenir une partition ou arbre de classification".

2. Définitions

I-Introduction
à l'analyse
des données

1-
Introduction :

2. Définitions

3. Les types
de variables

4. Les
échelles de
mesure

Définition

On appelle **variable** toute application X définie sur P , avec P un ensemble fini appelé **population** ou **univers** ; tout élément ω de P s'appelle un **individu**.

Remarque

X est aussi appelée **caractère statistique**

2. Définitions (suite)

I-Introduction
à l'analyse
des données

1-
Introduction :

2. Définitions

3. Les types
de variables

4. Les
échelles de
mesure

Le caractère désigne une grandeur ou un attribut, observable sur un individu et susceptible de varier prenant ainsi différents états appelés modalités.

Définition

*On appelle **modalité** toute valeur : $x_i \in X(P)$ telle que : $X(P) = \{x_1, x_2, x_3, \dots, x_p\}$ avec p nombre de modalités différentes de X*

A chaque modalité du caractère X , peut correspondre un ou plusieurs individus dans l'échantillon de taille N .

3. Les types de variables

Il existe deux types de variables :

- 1 **quantitatif** : c'est un caractère auquel on peut associer un nombre c'est-à-dire, pour simplifier, que l'on peut "mesurer".

On distingue alors deux types de caractère quantitatif :

- un caractère discret : c'est un caractère quantitatif qui ne prend qu'un nombre fini de valeurs. Par exemple le nombre d'enfants d'un couple.
- un caractère continu : c'est un caractère quantitatif qui, théoriquement, peut prendre toutes les valeurs d'un intervalle de l'ensemble des nombres réels. Ses valeurs sont alors regroupées en classes $[x_{i-1}, x_i[$. Par exemple le poids ou la taille d'un individu.

- 2 **qualitatif** : comme la profession, la couleur des yeux, la nationalité, les groupes sanguins.

4. Les échelles de mesure

Il y a deux échelles de mesures pour les variables qualitatives : L'échelle nominale, L'échelle ordinale.

- L'échelle nominale permet de classer les individus dans des modalités qui sont exprimables par des noms et qui ne sont pas hiérarchisées.

Par exemple :

- a) Le genre des personnes : 1. Femme ; 2. Homme ;
- b) Statut marital : célibataire ; marié ; veuf, ...

- L'échelle ordinale permet de classer les individus dans des modalités et, en plus, d'établir un ordre hiérarchique entre ces modalités. Il y a une gradation dans les modalités utilisées (Elles sont alors hiérarchisées).

Par exemple :

- a) Le niveau de scolarité :
1. Primaire ; 2. Secondaire ; 3. Collégial ; 4. Universitaire.
- b) Niveau d'appréciation d'un produit :
Très bonne qualité, bonne qualité, qualité moyenne,

4. Les échelles de mesure (suite)

Il y a deux échelles de mesures pour les variables quantitatives : L'échelle par intervalles, L'échelle de rapport.

- L'échelle par intervalles : permet non seulement d'identifier la modalité à laquelle appartient l'unité statistique et d'établir un ordre entre les modalités observables mais aussi elle nous informe de l'écart (la distance) séparant deux modalités. Sur cette échelle le zéro est situé de manière arbitraire (une valeur de référence arbitraire, mais ne signifie pas une absence d'un caractère), comme pour la mesure des températures par exemple (échelles Celsius et Fahrenheit).

4. Les échelles de mesure (suite)

I-Introduction
à l'analyse
des données

1-
Introduction :

2. Définitions

3. Les types
de variables

4. Les
échelles de
mesure

- L'échelle de rapport : possède les propriétés d'échelle d'intervalle et le zéro constitue un zéro absolu c'est-à-dire la valeur 0 indique l'absence complète du caractère considéré.

Par exemple : âge, salaire, taille, vitesse, etc...