

Partie - II : Analyse des données

Chapitre V : Data mining

Prof. Abdelkamel ALJ

Updated: 2020/03/02

FSJES - UMI

Meknès

1. Introduction

1. Introduction

2. Analyse univariée

1. Introduction

2. Analyse univariée

3. Analyse bivariée

3.1 Analyse multivariée

1. Introduction

J-P. Fénelon " L'analyse des données est un ensemble de techniques pour découvrir la structure, éventuellement compliquée, d'un tableau de nombres à plusieurs dimensions et de traduire par une structure plus simple et qui la résume au mieux. Cette structure peut le plus souvent, être représentée graphiquement"

L'analyse univariée :

Analyse la distribution de fréquences, mode, médiane, moyenne et écart-type, test Khi-deux d'ajustement, sans oublier les représentations graphiques.

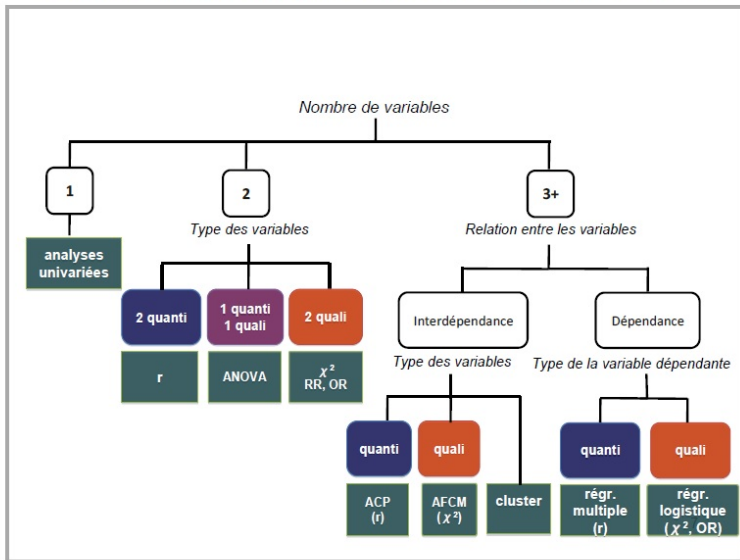
L'analyse bivariée

Analyse les relations entre un couples de variables : test Khi-deux d'indépendance, tests t, coefficient de corrélation et régression simple).

L'analyses multivariées

- ⊙ L'analyses multivariées des interdépendances : analyse en composantes principales (ACP), analyse factorielle des correspondances multiples (AFCM) et analyse de classification.
- ⊙ L'analyses multivariées des dépendances : régression multiple et régression logistique.

Techniques d'analyse appliquées selon le nombre et le type de variables



2. Analyse univariée



2. Analyse univariée

- *Variable qualitative* :

Les outils descriptifs d'une variable qualitative se limitent habituellement à la fréquence (les effectifs) de ses modalités.

Variable : « Sexe »

Code : sexe

Valeurs : 0 : Masculin
1 : Féminin

Type : dichotomique

Statistiques

Sexe

N	Valide	1440
	Manquante	0
Moyenne		,52
Ecart-type		,500

Le tableau " Statistiques " (moyenne, écart-type) se lit comme suit : sur un total de 1440 observations valides, il n'y a aucune valeur manquante pour la variable sexe. La moyenne de 0,52 correspond à la fréquence des femmes, codées 1, dans l'échantillon.

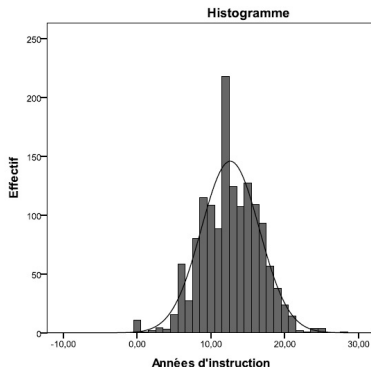
- Variable quantitative :

Les paramètres (moyenne, écart-type) de la distribution de cette variable est décrite dans le tableau intitulé " Statistiques ", et sa distribution est représentée par un " Histogramme ".

Statistiques

Années d'instruction

N	Valide	1435
	Manquante	5
Moyenne		12,6185
Ecart-type		3,92869
Minimum		,00
Maximum		28,00



- ⊙ Le nombre d'années d'études comporte 5 données manquantes et que ses valeurs varient de 0 à 28 années, avec une moyenne de 12,6 années.
- ⊙ L'histogramme montre une distribution assez symétrique, qui pourra sans doute être considérée comme approchant une distribution normale.

3. Analyse bivariable

- Le tableau de contingence

Un tableau de contingence est une méthode de représentation de données issues d'un comptage permettant d'estimer la dépendance entre deux caractères.

$X Y$	Y_1	Y_2	\dots	Y_j	\dots	Y_q		
X_1	n_{11}	n_{12}		n_{1j}		n_{1q}	$n_{1.}$	$f_{1.}$
X_2	n_{21}	n_{22}		n_{2j}		n_{2q}	$n_{2.}$	$f_{2.}$
\vdots								
X_i	n_{i1}	n_{i2}		n_{ij}		n_{iq}	$n_{i.}$	$f_{i.}$
\vdots								
X_p	n_{p1}	n_{p2}		n_{pj}		n_{pq}	$n_{p.}$	$f_{p.}$
	$n_{.1}$	$n_{.2}$		$n_{.j}$		$n_{.q}$	$n_{..}$	
	$f_{.1}$	$f_{.2}$		$f_{.j}$		$f_{.q}$		100

3.1 Deux variables qualitatives

Le test du Khi-deux

Le test du Khi-deux mesure le niveau de signification d'une relation bivariée.

3.2. Une variable qualitative et une variable quantitative

- Le Test t de Student

Ce test paramétrique repose sur des comparaisons de moyennes.

- ⊙ Le test-t de Student pour échantillon unique
- ⊙ Le test-t de Student comparant deux groupes d'échantillons indépendants (on parle de test de Student non apparié)
- ⊙ Le test-t de Student comparant deux groupes d'échantillons dépendants (on parle de test de Student apparié).

- Analyse de la variance "ANOVA"

L'analyse de la variance a pour but la comparaison des moyennes de k populations, à partir d'échantillons aléatoires et indépendants prélevés dans chacune d'elles.

- Conditions d'applications de l'ANOVA :

- ⊙ les populations étudiées suivent une distribution normale
- ⊙ les variances des populations sont toutes égales (Homoscédasticité)
- ⊙ les échantillons sont prélevés aléatoirement et indépendamment dans les populations.

3.3. Deux variables quantitatives

- Les coefficients de corrélations

Les coefficients de corrélation permettent de donner une mesure synthétique de l'intensité de la relation entre deux caractères et de son sens lorsque cette relation est linéaires .

Le coefficient de corrélation 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

- La régression linéaire simple

Un modèle de régression linéaire simple est un modèle de régression d'une variable expliquée sur une variable explicative.

4.1. Variables quantitatives

- La régression multiple

L'objectif général de la régression multiple est d'en savoir plus sur la relation entre plusieurs variables indépendantes ou prédictives et une variable dépendante.

- Analyse en composantes principales

Un outil d'analyse des relations entre plusieurs variables quantitatives sans leur attribuer de rôles de dépendante et d'indépendante(s) comme dans les analyses de régression et d'autre part c'est un outils de visualisation (graphique) des données.

4.2. Variables qualitatives et quantitatives

- Régression logistique

La régression logistique est une technique prédictive. Elle vise à construire un modèle permettant de prédire / expliquer les valeurs prises par une variable cible qualitative (le plus souvent binaire, on parle alors de régression logistique binaire ; si elle possède plus de 2 modalités, on parle de régression logistique polytomique) à partir d'un ensemble de variables explicatives quantitatives ou qualitatives (un codage est nécessaire dans ce cas).

4.3. Variables qualitatives

- Analyse factorielle des correspondances multiples

Ils sont des outils d'analyse des relations s'établissant entre deux ou plusieurs variables qualitatives sans leur attribuer les rôles de dépendante et d'indépendante(s) comme dans les analyses de régression.