Analyse des données

14 janvier 2022

Table des matières

1 Introduction

Les méthodes statistiques sont utilisées dans presque tous les secteurs de l'activité humaine et font parti des connaissances de base de l'ingénieur, du gestionnaire, de l'économiste, du chercheur ...
A l'issue de la phase de recueil de données, la démarche statistique consiste à traiter et interpreter les informations recueillies. Cette démarche comporte souvant deux grands aspects, d'abord l'aspect descriptif ou exploratoire et dans un second temps l'aspect inférentiel ou décisionnel.

> **La statistique exploratoire** a pour objet de synthetiser, résumer et structurer l'information contenue dans les données et utilise pour cela des représentations des données sous forme de tableau, graphiques et ou indicateurs.
> **La phase décisionnelle** arrive après la phase exploratoire, mais nous n'allons pas la traiter

Depuis plus de 50 ans les méthodes d'analyse de données ont largement démontré leur efficacité dans l'étude de gros volumes de donnés grace a l'informatique. Les méthodes dites **multi-dimensionelles** telles que l'analyse en composante principale (ACP) ou l'analyse factorielle des correspondances (AFC) permettent la mise en relation de nombreuses variables entre elles.
Les méthodes multi-dimensionelles permettent d'obtenir des représentations graphiques qui constituent le meilleur résumé possible de l'information dans un gros tableau de données.
Pour cela le statiticien conscent à une perte d'information afin de gagner en lisibilité, il est alors en mesure de faire apparaitre les principaux phénomènes qu'il cherche à analyser.

Il est possible de diviser les principales méthodes d'analyse de données en deux grands groupes :

> **Les méthodes de classification** qui ont pour objet de réduire la taille de l'ensemble des individus en formant des groupes homogènes
> **Les méthodes factorielles** cherchent à réduire le nombre de variables en les résumant par un petit nombre de composants synthétiques. Selon que les variables soient quantitives ou qualitatives, on utilisera soit l'analyse en composante principale, soit l'analyse factorielle des correspondances.

Le choix d'une méthode statistique dépend de l'objectif que l'on se fixe, mais également de la nature des variables, ainsi que le nombre de variables.

1.1 Les différents types de variables

Une variables statistique décrit une caractéristique pour les différents individus. On distingue deux grands types de variables, les variables qualitatives et quantitatives.

— Les variables quantitatives qualifient des quantités, elles peuvent se classer en variables continues ou discrètes.

— Les variables qualitatives définissent une caractérisitque (des qualités), elles peuvent être de deux natures : nominales ou ordinales.

1.2 Description d'une variable quantitative

Une variable quantitative est décrite par l'ensemble des valeurs qu'elle prend pour n individus. Afin de synthétiser l'information, on peut utiliser la moyenne ou la variance.

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad\qquad V(x) = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2$$

Par définition une variable est centrée si sa moyenne est nulle et est réduite et, si sa variance est égale à 1. L'interet de centrer et réduire permet de donner une interprétation géométrique au coefficient de correlation linéaire.

1.3 Description d'une variable qualitative

Pour décrire les variables qualitatives, on utilise des fréquences relatives et les tableaux disjonctifs complets.

Considérions la variable qualitative «couleur des yeux» avec trois modalités : bleu, vert et marron.

Individu	1	2	3	4	5	6
Couleur des Yeux	Bleu	Vert	Marron	Vert	Bleu	Bleu

L'effectif est le nombre d'individus ayant une modalité dans l'échantillon. On peut donc a partir de l'effectif, calculer la fréquence relative qui est l'effectif divisé par le nombre total d'individus.

La présentation d'une variable qualitative sous sa forme disjonctive complète est celle qui se prête le mieux à des calculs statistiques. Cette dernière s'obtient en définissant une variable indicatrice pour chacune des modalités de la variable («B» pour bleu ...).

Tableau disjonctif complet des données ci-dessus :

$$\mathrm{X} = \begin{array}{c} \begin{matrix} \mathrm{B} & \mathrm{V} & \mathrm{M} \end{matrix} \\ \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \end{array}$$

Propriétés du tableau disjonctif :

— La somme des colonnes du tableau complet est égale à un vecteur colonne de dimension n dont tous les éléments sont égaux à 1. Chaque individu possède une seule modalité, cela signifie que sur une ligne de données ne figure que des 0, à l'exception d'un élément unique égal à 1.

— Le produit matriciel de la transposée de la matrice X fois X est une matrice diagonale dont les éléments sont les effectifs de chacune des modalités. Exemple avec les données précédentes :

$$\text{X'X} = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

1.4 Relation entre deux variables quantitatives

Une variable x prenant n valeurs peut être représentée par un vecteur dans un ensemble $\mathbb{R}^n$. Dans cet espace, le produit scalaire entre deux vecteurs $\vec{x}$ et $\vec{y}$, est égal à la somme pour l'ensembles des individus de x et y

$$\vec{x} \cdot \vec{y} = \sum_{i=1}^{n} x_i y_i$$

En statistiques, le produit scalaire utilisé est :

$$\vec{x} \cdot \vec{y} = \frac{1}{n} \sum_{i=1}^{n} x_i y_i$$

Ce produit scalaire permet de donner une interpréation géométrique du coefficient de corrélation linéaire. Le cosinus de l'angle formé entre les deux variables est égal au coefficient de correlation entre ces deux variables tel que : $\cos(\text{X,Y}) = r_{\text{X,Y}}$.

Si le coefficient de corrélation est égal a 1, les deux vecteurs sont colinéaires (les valeurs prises par x_i et y_i sont proportionelles). L'abscence de correlation se traduit par un coefficient de corrélation nul et par un angle droit entre x et y.
Notes 14/01/22

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

$$Y = \begin{bmatrix} 0 & 1 & \\ 0 & 1 & \\ 0 & 1 & \\ 1 & 0 & \\ 1 & 00 & 1 \\ 1 & 0 & \\ 1 & 0 & \\ 1 & 0 & \\ 0 & 1 & \end{bmatrix}$$

On peut obtenier le tableau de contingence : $C = X'Y$ 13 21 21 = 1010100001 0000011010 0101000100 * 01 01 01 10 10 01 10 10 10 01
Tableau de contingence ou des effectifs observés
Tableau des frequences observées

1.4.1 Le tableau de BURT

A partir du tableau disjoncitf complet de trois variabkes qualitatives X, Y, Z on peut construire le tableau de BURT : $B = (X, Y, Z)' \cdot (X, Y, Z)$

$$X$$

X Y Z 010 100 100 010 100 001 001 010 100 100

$$B = \begin{bmatrix} 4 & 0 & 0 & 1 & 3 & 3 & 1 & 0 \\ 0 & 3 & 0 & 2 & 1 & 1 & 0 & 2 \\ 0 & 0 & 3 & 2 & 1 & 1 & 2 & 0 \\ & & & & & & & \\ 1 & 2 & 2 & 5 & 0 & 2 & 2 & 1 \\ 3 & 1 & 1 & 0 & 5 & 3 & 1 & 1 \\ & & & & & & & \\ 3 & 1 & 1 & 2 & 3 & 5 & 0 & 0 \\ 1 & 0 & 2 & 2 & 1 & 0 & 3 & 0 \\ 0 & 2 & 0 & 1 & 1 & 0 & 0 & 2 \end{bmatrix}$$

Le dernier élément concerne la transformation d'une variable quantiative en qualitative ou d'une variable qualitative en quantitative. Pour transformer une variable quantitative en variable qualitative, on fait des classes. Exemple : Avis sur l'enseignement pendnant la prériode covid

- Très satisfait = 2
- Moyennement satisfait = 1
- Pas dutout satisfait = 0

Le saut entre niveau de satisfaction est de 1 L'interet des vraibles qualitatives est d'autoriser un certain nombre de non linéarité.

1.5 Relation entre une variable quantitative expliquée et un ensemble de variables quantitatives explicatives

Si on considère un tableau a n lignes et p colones

$$\begin{bmatrix} X_{11} & X_{12} & \dots & X_{1p} \\ \cdots X_{i1} & X_{ij} & \dots & X_{ip} \\ X_{n1} & X_{nj} & \dots & X_{np} \end{bmatrix}$$

X La distance entre deux individus i et i' revient à calculer la distance : $d^2(i,i') = \sum_{j=1}^{p}(X_{ij} - X_{i'j})^2$ Ce qui différencie les méthodes, c'est la distance que l'on prend.

2 L'analyse en composantes principales (ACP)

L'ACP est sans aucun doute la méthode d'AD la plus connue et la plus utilisée, c'est à Pearson et Hotelling que l'on doit les premieres publications à son sujet. Cette technique est connue depuis plus d'un siècle, mais s'est développée au cours des 50 dernières années, grace a l'informatoqie. C'est à partir des années 60 que la technique se développe

But et interêt de la méthode La présentation synthétique d'un grand ensemble de données, résultant de l'étude de plusieurs caractères quantitatifs sur une population n'est pas chose facile. L'ACP a donc pour objet de reveler les inter relations entre ces différentes variables quantitatives et proposer une solution

2.1 A la recherche d'une structure

Un des interets majeurs de la SCP est de fournir une méthode de représentation d'une population décrite pa run enesemble de carctere quantitattif afin de :

1. Reperer des groupes d'individus homogènes vis a vis de l'ensemble des caractères.
2. Cette méthode acte a révéler des différences entre individus ou groupe d'individus, relativement à l'ensemble des caractères,
3. Apte a mettre en évidence des individus au comportement atypique ce comportement étant du a la précesence de données abérantes soit a des causes qu'il conviendra d'expliquer.
4. Réduire l'information qui permet de décrire la position d'un individu dans l'ensemble de la population.

L'ACP permet de construire des variables artificielles qui expliquent l'ensemble des variables statistiques prises en compte dans l'ACP. Ces var permettent une réduction du tableau des données brutes. Puisqu'au prix d'une perte d'information que l'on saura mesurer, il sera possible de remplacer l'ensemble des variables statistiques de départ par un nombre en général beaucoup plus faible de var statistiques artificielles. Finalement les apports principaux de l'ACP sont de deux types :

1. L'élaboration d'une ou plsr représentation des individus analyse du nuage des individus, cette analyse permet de chercher la str
2. La construction de variables artificielles qui expliquent les variables statistiques mesurées sur la population. Analyse du nuage des variables.

En résumé, l'analyse en composante principale est la base de toutes les analyses multi-factorielles, elle s'applique à des tableaux a deux dimensions croisant des individus et des varaiables quantitatives. Elle consiste à regrouper des vabriables quantiatives en combinaisons linairaes, que l'on va appeller composantes principales ou axes factoriels.

2.2 Données des marges retenues dans l'ACP et démarche utilisées

Du point de vue de sa notation mathématique, la position du prob est simple, elle consiste à partir de p variables quantitatives quelconques, constituant un repère à p dimensions, à passer a un repère orthonormé à p dimensions également, mais ces nouvelles dimensions sont appellés les facteurs ou composantes principales.

2.2.1 Les données et les notations

On suppose que l'on dispose de l'observation de p variables quanti pour n individus

$$X_{n,p} = W_{11}X_1jX_{1p}X_{i_1}XijX_{ip}X_{n_1}X_{np}$$

La ligne i décrit la valeur prise par l'individu petit i pour les p variables qui sont en colonne. La colonne j décrit bien la valeur de la variable $X_1 jpourlesnindividusdel'chantillon.Pourdesraiso$

$$Z_{ij} = X_{ij} - \overline{X_j}\sigma\left(X_j\right) \implies Z_{N,p} = \begin{bmatrix} Z_{11} = \frac{X_{11}-\overline{X_1}}{\sigma(X_1)} \\ Z_{ij} = \frac{X_{ij}-\overline{X_j}}{\sigma(X_j)} \end{bmatrix}$$

A partir de cette matrice Z il est possible de construire de construire la matrice des coef de cor lin (r), cette matrice r est :

$$R_{p,p} = \frac{1}{n}Z'_{p*n}Z_{n*p} = \begin{bmatrix} 1 r_{1,2} \\ r_{2,1} \end{bmatrix}$$

$r_{1,2} = r_{2,1}$ est donc symétrique A partir de ce moment là on peut savoir si l'on peut réaliser l'ACP Il est également possible de calculer la matrice de dispersion des individus, appellée matrice V $V_{n,n} = \frac{1}{n}ZZ'$. cette matrice n'est pas une matrice de corrélation (c'est une matrice de dispersion). On travaille plus souvent sur R car l'on a plus d'individus que de variables dim R $<$ dim V