République Algériennne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Mohamed Khider, Biskra

Faculté des Sciences Exactes et des Sciences de la Nature et de la Vie

Départementde Mathématiques

Mémoire présenté pour obtenir le diplôme de

Master en "**Mathematiques Appliquées**"

Option : Statistique

Par

Mouaki Benani Benani Amina

Titre :

Analyse factorielle des correspondances et applications

Devant le Jury :

Mr.	CHERFAOUI Mouloud	Pr.	U. Biskra	Président
Mr.	BENATIA Fateh	Pr.	U. Biskra	Encadreur
Mme.	OUANOUGHI Yasmina	Dr.	U. Biskra	Examinatrice

Soutenu Publiquement le 27/06/2022

Dédicace

Je dédie ce travail :

À mes chers parents pour leur soutient, leur prières et leur encouragement durant mon parcours scolaire.

À mon cher frère et mes chers soeurs ainsi a toute ma famille que dieu les garde pour moi.

À toute mes amies et tous les gens qui m'ont aide dans ma vie.

Remerciements

Tout d'abord, je viens à remercier ***Allah*** *le tout puissant qui m'a donné la santé, la voulanté et la patience pour accomplir ce travail.*

Je tient à exprimé mes profonds remerciments à mon encadrant monsieur le professeur ***Benatia Fateh*** *pour ses précieux conseils et son aide durant la période de travail.*

Je n'oublie pas de remercier sincèrement monsieur le professeur ***Necir Abdelhakim*** *pour son aide et ses conseils et aussi de remercier le Dr madame* ***Benelmir Imen*** *et monsieur le professeur* ***Yahia Djabrane****.*

Je tiens à remercier avec ma plus grande gratitude les membres du jury, monsieur le professeur ***Cherfaoui Mouloud*** *et le Dr madame* ***Ouanoughi Yasmina*** *pour avoir accepter de juger mon travail.*

J'adresse mes remerciments à tous les enseignants de la Faculté des Sciences Exactes et Sciences de la Nature et de la Vie -Département de Mathématiques-.

Merci à tous.

Notations et symbols

Les différentes abréviations et notations utilisées tout au long de ce mémoire sont expliquées ci-dessous :

ACP	:	Analyse en composantes principales.
X	:	Tableau des données.
e_i	:	i^{eme} individu.
x_j	:	j^{eme} variable.
D	:	Matrice des poids.
Z	:	Tableau standard.
Y	:	Matrice centrée.
V	:	Matrice de variance covariance.
R	:	Matrice de corrélation.
D	:	Matrice des poids.
$i.e$	:	C'est à dire.
F_k	:	Sous espace de projection de dimension k.
a_j	:	j^{eme} axe principal.
u_j	:	j^{eme} facteur principal.
c_j	:	j^{eme} composante principale.

AFC	:	Analyse factorielle des correspondances.
p, q	:	Nombre de modalités.
V_1, V_2	:	variables qualitatives.
n_{ij}	:	Effectif observé.
$n_{i:}$	:	Effectif marginal des lignes.
$n_{.j}$	:	Effectif marginal des colonnes.
f_{ij}	:	Fréquence observé.
$f_{i:}$	:	Fréquence marginal des lignes.
$f_{.j}$	:	Fréquence marginal des colonnes.
$f_{i/j}$	:	Fréquence conditionnelle aux profils-lignes.
$f_{j/i}$	:	Fréquence conditionnelle aux profils-colonnes.
pl_i	:	Profils-lignes.
pc_j	:	Profils-colonnes.
X_l	:	Tableau des profils-lignes.
X_c	:	Tableau des profils-colonnes.
g_l	:	Centre de gravité de profils-lignes.
g_c	:	Centre de gravité de profils-colonnes.
vq	:	Variable qualitative.
tr	:	Trace.
$d^2_{\mathcal{X}^2}$	:	Distance de khi-deux.
CTR	:	Contribution.

Table des matières

Table des figures

Liste des tableaux

Introduction

En statistique, on trouve plusieurs méthodes qui permettent d'extraire l'information contenue dans un jeu de données, ces méthodes sont regroupées dans un terme appelé analyse des données. Parmi celles-ci on trouve la méthode de l'analyse factorielle qui traite les tableaux rectangulaires de grande dimensions et qui a pour but de réduire la dimension en gardant le plus d'information possible[2].

En fonction de la nature des variables du tableau de données on distingue deux méthodes : l'analyse en composante principale noté ACP qui permet de résumer et visualiser l'information contenue dans un tableau des individus (en lignes) décrits par des variables quantitatives (en colonne) et l'analyse factorielle des correspondances noté AFC qui permer d'analyser l'information contenue dans un tableau dont toutes ses variables sont qualitatives.

Dans ce travail on s'intéresse à l'analyse factorielle des correspondances appellée aussi analyse des correspondances, cette méthode a été développée en France par J.-P. Benzécri durant la période 1970-1990. L'AFC est une méthode de réduction de dimension, elle sert a analyser les tableaux de contingence appelés tableaux croisées pour but d'étudier la liaison ou la correspondance existant entre les deux variables qualitatives.

Ce mémoire est composé de trois chapitres, deux chapitres théorique et un chapitre

sur la partie pratique.

Le premier chapitre représente les préliminaires avec un passage décrivant le principe de l'ACP.

Le deuxième chapitre est consacré à la méthode de l'AFC et comment la présentée en utilisant l'ACP détaillée en premier chapitre.

On finalise ce travail par une application de l'AFC sur les données réelles nommées "children" en utilisant le language R et ces différents packages.

En fin de ce mémoire on trouve une annexe cantenant la même application présentée auparavant avec le logiciel R mais formalé et calculé grâce au logiciel SPSS.

Chapitre 1

Préliminaire

L'analyse des données est une famille des méhodes statistiques qui permettent de transformer la donnée en information. Dans ce chapitre on s'intéresse à la description de ces données tels que le tableau des données, variables, individus,...etc.

1.1 Données Statistiques

La notion fondamentale en statistique est celle de groupe ou d'ensemble d'objets équivalents appelé population. Ces objets sont appelés des individus. Alors un individu est décrit par un ensemble de caractéristiques appelées variables.[13]

1.1.1 Tableau statistique et matrice des poids

Tableau statistique

On note X le tableau rectangulaire de n lignes et p colonnes contenant les observations de p variables sur n individus.

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ & \ddots & & & \\ x_{i1} & & x_{ij} & & x_{ip} \\ & & & \ddots & \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix} \in \mathcal{M}(n,p).$$

On note x_{ij} la valeur de la variable x_j observée sur l'individu e_i.

Individus et variables

Les individus et les variables sont définies respectivement par[3]

1. Les individus sont les lignes du tableau X, chaque individu est représenté par un vecteur de $\mathbb{R}^p$ noté e_i tel que :

$$e_i = (x_{i1}, ..., x_{ip})^t \in \mathbb{R}^p \quad ; i = 1, ..., n.$$

2. Les variables sont les colonnes du tableau X, chaque variable est représentée par un vecteur de $\mathbb{R}^n$ noté x_j tel que :

$$x_j = (x_{1j}, ..., x_{nj})^t \in \mathbb{R}^n \quad ; j = 1, ..., p.$$

On distingue deux types de variables :

– **Variables quantitatives** : Elles s'expriment par des nombres réels sur lesquels on peut appliquer des opérations arithmétiques (moyenne) qui ont un sens. On a comme exemple l'age, le poids et la taille.

– **Variables qualitatives** : On dit d'une variables aléatoire qu'elle est qualitatives, si ses résultats ne sont pas mesurable, comme la couleur des yeux, le sex

et montion obtenue au bac.

Matrice des poids

Lorsque les réalisations ou éléments du tableau X sont à probabilités égales, alors chaque réalisation x_{ij} a la même importance $1/n$ dans le calcul des caractéristiques de l'échantillon. On peut aussi appliquer un poids p_i différent à chaque réalisation conjointe des variables. Ces poids, qui sont des nombres positifs de somme égale à 1 représentés par une matrice diagonale notée D de taille n définit comme suit :

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

avec $p_i \geq 0$ et $\sum_{i=1}^{n} p_i = 1$ (qu'on va montrer ci-après)[3].
Dans le cas général des poids égaux, la matrice D devient

$$D = \frac{1}{n} I_n,$$

où I_n est la matrice identité, est une matrice carrée symétrique de diagonale 1.

Preuve. Comme les observations ont le même poids $p_1 = p_i = \frac{1}{n}$, alors

$$\begin{aligned} \sum_{i=1}^{n} p_i &= \sum_{i=1}^{n} p_1 \\ &= p_1 \sum_{i=1}^{n} 1 \end{aligned}$$

$$\begin{aligned}\sum_{i=1}^{n} p_i &= p_1 n \\ &= \frac{n}{n} \\ &= 1.\end{aligned}$$

D'où le résulat. ■

Donc la matrice des poids est représentée comme suit :

$$D = \begin{bmatrix} 1/n & & & 0 \\ & 1/n & & \\ & & \ddots & \\ 0 & & & 1/n \end{bmatrix}$$

$$D = \frac{1}{n}\begin{bmatrix} 1 & .. & .. & 0 \\ . & 1 & & . \\ . & & . & . \\ 0 & .. & .. & 1 \end{bmatrix}$$

$$D = \frac{1}{n} I_n.$$

1.1.2 Centre de gravité et tableau standard

Centre de gravité

On note par g le vecteur des moyennes arithmétiques. Il est définit par :

$$g = (\bar{x}_1, \bar{x}_2, ..., \bar{x}_p)^t \in \mathbb{R}^p,$$

où $\bar{x}_j$ désigne la moyenne de la variable x_j, avec

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}.$$

La forme matricielle est

$$g = X^t D 1_n,$$

où 1_n désigne le vecteur de $\mathbb{R}^n$ dont toutes les composantes sont égales à 1[14].

Tableau standard

Il est pratique de standariser les données avant de les analyser. Pour cela, on va transformer le tableau initial X en un tableau standard Z qui contient des données centrées réduites c'est à dire (une moyenne égale à zéro et un écart type égale à un) définit comme suit[1] :

$$Z = Y D_{1/s},$$

avec Y la matrice centrée définit par :

$$Y = X - 1_n g^t.$$

et $D_{1/s}$ est la matrice diagonale des inverses des écart types tel que :

$$D_{1/s} = \begin{bmatrix} 1/s_1 & .. & .. & 0 \\ . & & & . \\ . & & & . \\ 0 & .. & .. & 1/s_p \end{bmatrix} \in \mathcal{M}(p,p),$$

où s_j^2 est la variance empirique de la variable x_j définit comme suit :

$$s_j^2 = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2 ; \quad j = 1, ..., p,$$

1.1.3 Matrice de variance-covariance et corrélation

Matrice de variance-covariance

Définition 1.1.1 *[14]La matrice de variance-covariance est une matrice carrée symétrique de dimension p notée V, elle est définie par :*

$$V = \begin{bmatrix} s_1^2 & .. & .. & s_{1p} \\ . & & & . \\ . & & & . \\ s_{p1} & .. & .. & s_p^2 \end{bmatrix}.$$

La forme matricielle :$V = Y^t DY = X^t DX - gg^t$.

Preuve. On a $Y = X - 1_n g^t$ alors,

$$\begin{aligned} V &= Y^t DY \\ V &= \left(X - 1_n g^t\right)^t D \left(X - 1_n g^t\right) \\ V &= X^t DX - \left(X^t D 1_n\right) g^t - g\left(1_n DX\right) + g\left(1_n^t D 1_n\right) g^t \end{aligned}$$

comme

$g = X^t D 1_n$ et $1_n^t D 1_n = \sum_{i=1}^{n} p_i = 1$ alors,

$$V = X^t DX - gg^t - gg^t + gg^t$$
$$= X^t DX - gg^t.$$

■

Matrice de corrélation

Définition 1.1.2 *[14]La matrice de corrélation est une matrice carrée symétrique contenant les coeficients de corrélation entre chaque variable et les autres, on la note par R et elle est définit comme suit :*

$$R = \begin{bmatrix} s_1^2 & .. & .. & s_{1p} \\ . & & & . \\ . & & & . \\ s_{p1} & .. & .. & s_p^2 \end{bmatrix}.$$

La forme matricielle : $R = D_{1/s} R D_{1/s}$.

1.2 Nuage des individus

Un individu e_i peut être représenté comme un point de l'espace vectoriel à p dimensions noté $\mathbb{R}^p$. Ses coordonnées sont un vecteur de $\mathbb{R}^p$ appelé l'espace des individus.

L'ensemble de n individus forment un nuage de points dans l'espace $\mathbb{R}^p$ défini par les variables appllé nuage des individus tel que g est son centre de gravité[8].

Ressemblance entre individus

Deux individus se ressemblent d'autant plus qu'ils possèdent des valeurs proches pour l'ensemble des variables[8].

Comment mesurer la distance entre deux individus ?

Pour mesurer cette distance on choisit la distance euclidienne usuelle qui est égale à :

$$d^2\left(e_i, e_{i'}\right) = \sum_{j=1}^{p} \left(x_{ij} - x_{i'j}\right)^2 \;\; ; \quad i,\; i' = \overline{1,n}.$$

Le choix de la distance euclidienne permet de donner le même poids à chacune des variables[3].

1.2.1 Métrique

La métrique notée M est une matrice de taille p carrée symetrique et définie positive, on définit la distance entre deux individues e_i et $e_{i'}$ dans l'espace $\mathbb{R}^p$ par[3] :

$$d^2_M\left(e_i, e_{i'}\right) = \left(e_i - e_{i'}\right)^t M \left(e_i - e_{i'}\right).$$

Généralement en utilise ces deux métriques diagonale I_p et D_{1/s^2} tel que :

I_p : représente la matrice identité d'ordre p et

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

Remarque 1.2.1 [3]*On utilise la métrique $M = I_p$ pour les données centrées réduites du tableau Z et la métrique $M = D_{1/s^2}$ pour les données du tableau initial*

X où les variables ne s'expriment pas avec les mêmes unités.

1.2.2 Inertie

L'inertie mesure la dispersion des points du nuage par rapport à son centre de gravité[4].

On appelle inertie totale du nuage de point la moyenne pondérée des carrées des distances des points au centre de gravité g[14].

Elle est définie par :

$$I_g = \sum_{i=1}^{n} p_i d_M^2 (e_i, g) .$$
$$I_g = \sum_{i=1}^{n} p_i \parallel e_i, g \parallel_M^2 = \sum_{i=1}^{n} p_i \langle e_i - g, e_i - g \rangle_M = \sum_{i=1}^{n} p_i (e_i - g)^t M (e_i - g) .$$

L'inertie en un point "a" quelconque est définie par :

$$I_a = \sum_{i=1}^{n} p_i (e_i - a)^t M (e_i - a) .$$

L'inertie totale du nuage de points égale aussi à la trace de la matrice MV ou VM :

$$I_g = tr (MV) = tr (VM) .$$

Remarque 1.2.2 *on a :*

– *Si $M = I_p$:*

$$\begin{aligned} I_g &= tr (I_p V) \\ &= tr (V) \\ &= \sum_{j=1}^{p} s_j^2. \end{aligned}$$

– *Si* $M = D_{1/s^2}$ *:*

$$\begin{aligned} I_g &= tr\left(D_{1/s^2}V\right) \\ &= tr\left(D_{1/s}VD_{1/s}\right) \\ &= tr\left(R\right) = p. \end{aligned}$$

L'inertie dans ce cas est égale au nombre de variables.

1.3 Nuage des variables

Chaque variable x_j peut être représentée comme un vecteur de l'espace vectoriel à n dimensions noté $\mathbb{R}^n$, dont chaque dimension représente un individu. L'ensemble de p variables constitue un nuage de points sur $\mathbb{R}^n$ appelé nuage des variables[8].

1.3.1 Liaison entre deux variables

On mesure la liaison entre deux variables x_j et $x_{j'}$ par le coefficient de corrélation noté $r_{jj'}$, il est défini comme suit[8] :

$$r\left(x_j, x_{j'}\right) = \frac{cov\left(x_j, x_{j'}\right)}{\sqrt{var\left(x_j\right) var\left(x_{j'}\right)}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_{ij} - \bar{x}_j}{s_j}\right)\left(\frac{x_{ij'} - \bar{x}_{j'}}{s_{j'}}\right), jj' = \overline{1, p}.$$

Avec :

$cov\left(x_j, x_{j'}\right) = \left(x_{ij} - \bar{x}_j\right)\left(x_{ij'} - \bar{x}_{j'}\right)$.

1.3.2 Métrique des variables

Pour étudier la proximité des variables entre elles, il faut munir cet espace d'une métrique, c'est à dire $(i.e)$ trouver une matrice d'ordre n symetrique et définie positive, le choix de cette matrice se porte sur la matrice diagonale des poids D

pour les raisons suivantes[14] :

– Le produit scalaire de deux variables x_j et $x_{j'}$ vaut :

$$\langle x_j, x_{j'} \rangle_D = x_j^t D x_{j'} = \sum_{i=1}^{n} p_i x_{jj} x_{jj'},$$

et lorsque ces variables sont centrées, le produit scalaire sera correspond à la covariance $i.e$ $\langle x_j, x_{j'} \rangle_D = s_{jj'}$ et on a $\|x_j\|_D = s_j$.

– L'angle entre deux variables est donné par :

$$\cos(x_j, x_{j'}) = \cos\theta_{jj'} = \frac{\langle x_j, x_{j'} \rangle_D}{\|x_j\|_D \|x_{j'}\|_D} = \frac{s_{jj'}}{s_j s_{j'}},$$

où $\langle x_j, x_{j'} \rangle_D = \|x_j\|_D \|x_{j'}\|_D \cos(x_j, x_{j'})$.

Donc on remarque que le coefficient de corrélation linéaire n'est autre que le cosinus de l'angle entre les variables x_j et $x_{j'}$ et on écrit :

$$r(x_j, x_{j'}) = \frac{\langle x_j, x_{j'} \rangle_D}{\|x_j\|_D \|x_{j'}\|_D} = \cos(x_j, x_{j'}) = \cos\theta_{jj'}.$$

1.4 Analyse en composantes principales

L'analyse en composantes principales est l'une des méthode statistique qui permet d'analyser les données du tableau (individus/variables) où ces variables sont de nature quantitatives.

1.4.1 Principe de l'ACP

L'ACP est une méthode qui permet de projeter les observations depuis un espace à p dimensions de p variables vers un espace à k dimensions ($k < p$) tel qu'un

maximum d'information soit conservé sur les premières dimensions $i.e$ réduit les dimensions d'une donnée multivariée à deux ou trois composantes principales, qui peuvent être visualiser graphiquement, en perdant le moins possible de l'information.

1.4.2 Construction de sous-espace F_k

Le choix de l'espace de projection s'effectue tel que la moyenne des carrées des distances entre projections soit la plus grande possible. En d'autre terme il faut que l'inertie du nuage projeté sur le sous-espace F_k soit maximale[14].

On définit P la matrice de projection (opérateur) $M-ortogonale$ sur le sous-espace F_k tel que P vérifie :

1. $P^2 = P\,(P - idempotente)$.
2. $P^tM = MP\,(P \text{ est } M - symétrique)$.

On note f_i la projection de e_i sur F_k tel que $f_i = Pe_i$ d'où $f_i^t = e_i^tP^t$, donc[14] :

– le nuage projeté sera associé au tableau noté X_{proj} et on écrit : $X_{proj} = XP^t$.
– Le centre de gravité projeté est donné par : $g_{proj} = Pg$.
– La matrice de variance du tableau projeté est donnée par : $V_{proj} = PVP^t$.
– L'inertie du nuage projeté égale à : $I_{proj} = tr\,(VMP)$.

Pour déterminer F_k if faut trouver P le projecteur $M-ortogonale$ de rang k maximisant l'inertie.

Pour obtenir F_k on doit chercher E_1 le sous espace de dimension 1 d'inertie maximale, puit E_2 le sous espace de dimension 1 $M-ortogonale$ à E_1 et d'inertie maximale,... etc. La somme directe de ces espaces est F_k et on écrit :

$$F_k = E_1 \oplus ... \oplus E_k.$$

Axes principaux

On cherche une droite passant par g et qui maximise l'inertie de nuage projeté sur elle. Soit $a_1 \in \mathbb{R}^p$ un vecteur directeur unitaire, le projecteur $M-ortogonale$ sur la droite est donc[14] :

$$P_1 = a_1 \left(a_1^t M a_1\right)^{-1} a_1^t M = \frac{a_1 a_1^t M}{a_1^t M a_1}.$$

L'inertie du nuage projeté sur cette droite est égale à :

$$I_1 = tr\left(VMP_1\right) = \frac{a_1^t MVMa_1}{a_1^t M a_1}.$$

Preuve. On montre que $I_1 = \frac{a_1^t MVMa_1}{a_1^t Ma_1}$.

$$\begin{aligned} I_1 &= tr\left(VMP_1\right) \\ &= tr\left(VM\frac{a_1 a_1^t M}{a_1^t M a_1}\right) \\ &= \left(1/a_1^t M a_1\right) tr\left(a_1^t MVMa_1\right) \\ &= \left(a_1^t MVMa_1/a_1^t M a_1\right) \text{ car } a_1^t MVMa_1 \text{ est un scalaire.} \end{aligned}$$

■

Pour obtenir le maximum de l'inertie il suffit d'annuler la dérivé de l'expression $\frac{a_1 a_1^t M}{a_1^t M a_1}$ par rapport à a_1. On applique la règle de dérivation d'une forme quadratique par rapport à un vecteur on obtient : $VMa_1 = \frac{a_1^t MVM}{a_1^t M a_1} a_1$.

On pose $\frac{a_1^t MVM}{a_1^t M a_1} = \lambda \in \mathbb{R}$, alors $VMa_1 = \lambda a_1$.

Donc a_1 est un vecteur propre de VM associée à la plus grande valeur propre λ.

Facteurs principaux

On associe à l'axe principal a_j le facteur principal noté u_j définit par[10] :

$$u_j = Ma_j \in \mathbb{R}^p,$$

vérifiant :

– u_j est $M^{-1} - normé$ de norme 1 et $M^{-1} - orthogonale$.

– Comme a_j est un vecteur propre de VM alors

$$VMa_j = \lambda_j a_j \Rightarrow MVMa_j = \lambda_j Ma_j$$

on obtient :

$$MVu_j = \lambda_j u_j.$$

On remarque donc que les facteurs principaux u_j sont aussi les vecteurs propres de la matrice MV.

Composantes principales

Les composantes principales ce sont les nouvelles variables de $\mathbb{R}^n$ définies par les facteurs principaux comme suit[12] :

$$c_j = Xu_j,$$

où c_j est le vecteur des coordonnées de la projection $M-orthogonale$ des individus du tableau X sur l'axe a_j, vérifiant :

$E(c_j) = 0$, $Var(c_j) = \lambda_j$ et $Cov(c_j c_k) = 0, \quad j \neq k$.

Chapitre 2

Analyse factorielle des correspondances

L'analyse factorielle des correspondances est une méthode statistique qui permet d'étudier la liaison entre deux variables qualitatives, elle peut être vue comme ACP basée sur une métrique spéciale c'est la métrique du khi-deux[2]. Le but de cette méthode est la réduction de dimension.

2.1 Effectifs et fréquences

Soient V_1 et V_2 deux variables qualitatives à p et q modalités respectivement d'un ensemble de n individus, le but d'une AFC est d'étudier la liaison entre ces deux variables qui ne sont pas numériques pour cela on s'intéresse aux effectifs et fréquences.

2.1.1 Effectifs

On appel effectif d'une modalité (ou effectif observé), le nombre de fois que cette modalité apparaît. Autrement dit, c'est le nombre d'individus possédant à la fois la modalité i de la première variable qualitative $(v.q)$ et la modalité j de la seconde $v.q$ et on le note par n_{ij} [16].

Effectif total

L'effectif total noté n c'est la somme de tous les effectifs observés, il est donné par[2] :

$$n = \sum_{i=1}^{p} \sum_{j=1}^{q} n_{ij}.$$

2.1.2 Tableau de contingence

Un tableau de contingence appelé aussi tableau d'effectifs est un tableau à p lignes et q colonnes obtenus en croisant les modalitées des deux variables V_1 et V_2, on le note par N^* [2].

V_1/V_2	y_1	$\cdots$	y_j	$\cdots$	y_q	marge i
x_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1q}	$n_{1.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{iq}	$n_{i.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_p	n_{p1}	$\cdots$	$\vdots$	$\cdots$	n_{pq}	$n_{p.}$
marge j	$n_{.1}$	$\cdots$	$n_{.j}$	$\cdots$	$n_{.q}$	n

TAB. 2.1 – Tableau de contingence de V1 et V2.

Marges

On définit la marge en lignes qui est l'effectif total de la modalité x_i de V_1 et on la note $n_{i.}$ par[15] :

$$n_{i.} = \sum_{j=1}^{q} n_{ij} \ ; \quad i = 1, ..., p,$$

et la marge en colonnes qui est l'effectif total de la modalité y_i de V_2 noté $n_{.j}$ par :

$$n_{.j} = \sum_{i=1}^{p} n_{ij} \ ; \quad j = 1, ..., q.$$

On a aussi

$$\sum_{i=1}^{p} n_{i.} = \sum_{j=1}^{q} n_{.j} = \sum_{i=1}^{p} \sum_{j=1}^{q} n_{ij} = n.$$

2.1.3 Fréquences

La fréquence notée f_{ij} est le rapport (division) entre un effectif (n_{ij}) et l'effectif total (n), elle est donnée par[16] :

$$f_{ij} = \frac{n_{ij}}{n}.$$

Tableau des fréquence

Le tableau des fréquences noté N est un tableau à p lignes et q colonnes, obtenu en divison chaque effectif n_{ij} par l'effectif total n, il est définit comme suit :

V_1/V_2	y_1	$\cdots$	y_j	$\cdots$	y_q	marge i
x_1	f_{11}	$\cdots$	f_{1j}	$\cdots$	f_{1q}	$f_{1.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_i	f_{i1}	$\cdots$	f_{ij}	$\cdots$	f_{iq}	$f_{i.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_p	f_{p1}	$\cdots$	$\vdots$	$\cdots$	f_{pq}	$f_{p.}$
marge j	$f_{.1}$	$\cdots$	$f_{.j}$	$\cdots$	$f_{.q}$	1

TAB. 2.2 – Tableau de fréquence de V1 et V2.

Fréquences marginales

On note $f_{i.}$ la fréquence marginale en lignes et on la définit par[2] :

$$f_{i.} = \sum_{j=1}^{q} f_{ij} = \frac{n_{i.}}{n}; \quad i = 1, ..., p,$$

et on définit la fréquence marginale en colonnes notée $f_{.j}$ par :

$$f_{.j} = \sum_{i=1}^{p} f_{ij} = \frac{n_{.j}}{n}; \quad j = 1, ..., q.$$

On a ainsi

$$\sum_{i=1}^{p} f_{i.} = \sum_{j=1}^{q} f_{.j} = \sum_{i=1}^{p} \sum_{j=1}^{q} f_{ij} = 1.$$

fréquences conditionnelles

Les fréquences conditionnelles sont les nombres définis par[8] :

$$f_{i/j} = \frac{f_{ij}}{f_{.j}} \quad \text{et} \quad f_{j/i} = \frac{f_{ij}}{f_{i.}}.$$

On a aussi

$$\sum_{i=1}^{p} f_{i/j} = 1; \quad j = 1, ..., q,$$

et

$$\sum_{j=1}^{q} f_{j/i} = 1 \ ; \quad i = 1, ..., p.$$

2.2 Indépendance

L'AFC a pour but d'étudier la liaison entre deux variables qualitatives, on exprime cette liaison dans un tableau de contingence ou de fréquence. Si les deux variables ne sont pas liées alors, l'AFC n'a aucun sens.

Définition 2.2.1 [2] *On dit qu'il y a une indépendance entre deux variables qualitatives V_1 et V_2, si pour tout i et j, on a l'égalité :*

$$n_{ij} = \frac{n_{i.} \times n_{.j}}{n} \ ou \ f_{ij} = f_{i.} \times f_{.j}.$$

Remarque 2.2.1 [8] *Sous l'hypothèse d'indépendance on a :*

$$f_{i/j} = f_{.j}; \quad i = 1, ..., p,$$

et

$$f_{j/i} = f_{i.}; \quad j = 1, ..., q.$$

2.3 Profils

En AFC, on doit transformer le tableau des données en deux tableaux appelés profils tel que : pour les lignes on trouve le tableau des profils-ligne et pour les

colonnes le tableau des profils-colonnes. Cette transformation découle de l'objectif qui vise à étudier la liaison entre les deux variables.

2.3.1 Profil-ligne

On appelle profil-ligne le vecteur de $\mathbb{R}^q$ obtenu en divisant les termes f_{ij} de la ligne i par la marge $f_{i.}$ et on le note pl_i. Autrement dit c'est la fréquence conditionnelle $f_{j/i}$, cette valeur représente la probabilité d'avoir la modalité j de la variable V_2 sachant que la modalité de la variable V_1 est i. On définit le i^{eme} profil-ligne par[1] :

$$pl_i = \left(\frac{f_{i1}}{f_{i.}}, ..., \frac{f_{iq}}{f_{i.}} \right)^t ; \quad i = 1, ..., p.$$

Tableau des profils-lignes

On note par X_l le tableau des profils-lignes définit par[6] :

$$X_l = D_l^{-1} N,$$

tel que N la matrice des fréquences et D_l la martice diagonale à p lignes et q colonnes des fréquences marginalles de la $v.q$ V_1. On définit D_l et N comme suit :

$$D_l = \begin{bmatrix} f_{1.} & 0 & .. & 0 \\ 0 & f_{2.} & & . \\ . & & \ddots & . \\ 0 & .. & .. & f_{p.} \end{bmatrix} \text{ et } N = \begin{bmatrix} f_{11} & .. & .. & f_{1q} \\ . & \ddots & & . \\ . & & \ddots & . \\ f_{p1} & .. & .. & f_{pq} \end{bmatrix}.$$

Nuage des profils-lignes

Les profils-lignes forment un nuage de p points de $\mathbb{R}^q$, chaque points de nuage est affecté d'un poids égale à sa fréquence marginale, donc on obtient la matrice des poids égale à D_l[11].

Le centre de gravité de ce nuage noté G_l, est la moyenne pondérée de tous les points sur tous les axes j, il s'interprète comme un profil moyen. Il est donné par :

$$G_l = N^t \mathbf{1}_p = \begin{pmatrix} f_{.1} & f_{.2} & \dots & f_{.q} \end{pmatrix}^t \in \mathbb{R}^q.$$

où $\mathbf{1}_p$ est le vecteur de $\mathbb{R}^p$ dont toutes les composantes vallent 1.

Preuve. On montre que $G_l = N^t \mathbf{1}_p$.

On a

$$\begin{aligned} N^t \mathbf{1}_p &= \begin{bmatrix} f_{11} & .. & .. & f_{p1} \\ . & \ddots & & . \\ . & & \ddots & . \\ f_{1q} & .. & .. & f_{pq} \end{bmatrix} \times \begin{bmatrix} 1 \\ \vdots \\ \vdots \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} \sum_{i=1}^{p} f_{i1} \\ \vdots \\ \vdots \\ \sum_{i=1}^{p} f_{iq} \end{bmatrix} = \begin{bmatrix} f_{.1} \\ \vdots \\ \vdots \\ f_{.q} \end{bmatrix} \\ &= G_l. \end{aligned}$$

■

2.3.2 Profils-colonnes

On appelle profil-colonne le vecteur de $\mathbb{R}^p$ obtenu en divisant les termes f_{ij} de la colonne j par la marge $f_{.j}$ et on le note pc_j. D'autre façon c'est la fréquence conditionnelle $f_{i/j}$, cette valeur représente la probabilité d'avoir la modalité i de la variable V_1 sachant que la modalité de la variable V_2 est j[1]. On définit le j^{eme} profil-colonne par

$$pc_j = \left(\frac{f_{1j}}{f_{.j}}, ..., \frac{f_{pj}}{f_{.j}}\right)^t ; \quad j = 1, ..., q.$$

Tableau des profils-colonnes

On appel tableau des profils-colonnes et on le note par X_c, la matrice à p lignes et q colonnes des fréquence conditionnelle $f_{i/j}$ définit par[6] :

$$X_c = D_c^{-1} N^t,$$

avec $D_c = \begin{bmatrix} f_{.1} & 0 & .. & 0 \\ 0 & f_{.2} & & . \\ . & & \ddots & . \\ 0 & .. & .. & f_{.q} \end{bmatrix}$.

Nuage des profils-colonnes

De même façons que le nuage des profile-ligne. Les profils-colonnes forment un nuage de q points de $\mathbb{R}^p$, chaque points de nuage est affecté d'un poids égale à sa fréquence marginale, donc on obtient la matrice des poids égale à D_c[11].

Le centre de gravité de ce nuage noté G_c, est la moyenne pondérée de tous les

points sur tous les axes i, il s'interprète comme un profil moyen. Il est définit par :

$$G_c = N\mathbf{1}_q = \begin{pmatrix} f_{1.} & f_{2.} & \dots & f_{p.} \end{pmatrix}^t \in \mathbb{R}^p.$$

où $\mathbf{1}_q$ est le vecteur de $\mathbb{R}^q$ dont toutes les composantes sont égale à 1.

Preuve. On a

$$\begin{aligned} N\mathbf{1}_q &= \begin{bmatrix} f_{11} & .. & .. & f_{1q} \\ . & \ddots & & . \\ & & \ddots & \\ . & & . & . \\ f_{p1} & .. & .. & f_{pq} \end{bmatrix} \times \begin{bmatrix} 1 \\ \vdots \\ \vdots \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} \sum_{i=1}^{p} f_{1j} \\ \vdots \\ \vdots \\ \sum_{j=1}^{q} f_{pj} \end{bmatrix} = \begin{bmatrix} f_{1.} \\ \vdots \\ \vdots \\ f_{p.} \end{bmatrix} \\ &= G_c. \end{aligned}$$

■

2.3.3 Ressemblance entre profils

La ressemblance entre les lignes ou les colonnes est mesurée par une distance entre profils, cette distance n'est pas la distance usuelle mais c'est une distance spécifique à l' AFC appellée distance ou métrique du khi-deux notée $\mathcal{X}^2$, elle est définie de façon symétrique pour les lignes et les colonnes[5].

Les distances sont définit comme suit[12] :

– **Distance entre deux profils-lignes i et i' :**

$$d^2_{\mathcal{X}^2}(i, i') = \sum_{j=1}^{q} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2 = \| \, i - i' \, \|^2_{M_l} \, .$$

La métrique correspondante est une matrice diagonale :

$$M_l = D_c^{-1}.$$

– **Distance entre profil-ligne i et son centre de gravité G_l :**

$$d^2_{\mathcal{X}^2}(i, G_l) = \sum_{j=1}^{q} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - f_{.j} \right)^2 = \| \, i - G_l \, \|^2_{M_l} \, .$$

– **Distance entre deux profils-colonnes j et j' :**

$$d^2_{\mathcal{X}^2}(j, j') = \sum_{i=1}^{p} \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}} \right)^2 = \| \, j - j' \, \|^2_{M_l} \, .$$

La métrique correspondante est une matrice diagonale :

$$M_c = D_l^{-1}.$$

– **Distance entre profil-colonne j et son centre de gravité G_c :**

$$d^2_{\mathcal{X}^2}(j, G_c) = \sum_{i=1}^{p} \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - f_{i.} \right)^2 = \| \, j - G_c \, \|^2_{M_l} \, .$$

Pourquoi on utilise la métrique du $\mathcal{X}^2$ plutôt que la métrique euclidienne[10] ?

– Avec la métrique du $\mathcal{X}^2$, la distance entre deux lignes ou deux colonnes ne dépend pas des poids.

– La métrique du $\mathcal{X}^2$ possède la propriété d'équivalence distributionnelle : si on regroupe deux modalités lignes, les distances entre les profils-colonne, ou entre les autres profils-lignes restent inchangées.

2.3.4 Statistique du $\mathcal{X}^2$

Pour mesurer le caractère significatif du liaison entre les deux variables qualitatives on utilise habituellement la statistique du $\mathcal{X}^2$ appliquée au tableau d'effectifs, cette statistique mesure l'écart entre les effectifs observés et les effectifs théoriques[2].

La démarche du test est la suivante :

Soient V_1 et V_2 deux variables qualitatives :

$$\begin{cases} H_0 : V_1 \text{ et } V_2 \text{ sont indépendantes} \\ H_1 : V_1 \text{ et } V_2 \text{ ne sont pas indépendantes} \end{cases}$$

Sous H_0 la statistique du test est définie par :

$$\mathcal{X}^2_{calcul} = \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{(effectif\ obs - effectif\ théo)^2}{effectif\ théo} = \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{(nf_{ij} - nf_{i.}f_{.j})^2}{nf_{i.}f_{.j}}.$$

La région critique de ce test au seuil $\alpha = 0.05$:

$$D = \{|\ \mathcal{X}^2_{calcul}\ | > \mathcal{X}^2_{(p-1)(q-1)}\}.$$

Remarque 2.3.1 *Si les deux variables V_1 et V_2 sont indépendantes alors $\mathcal{X}^2 = 0$.*

En effet, comme V_1 et V_2 sont indépendantes alors on a $f_{ij} = f_{i.}f_{.j}$ on aura donc

$$\mathcal{X}^2 = n\sum_{i=1}^{p}\sum_{j=1}^{q} \frac{(f_{ij} - f_{i.}f_{.j})^2}{f_{i.}f_{.j}} = n\sum_{i=1}^{p}\sum_{j=1}^{q} \frac{(f_{i.}f_{.j} - f_{i.}f_{.j})^2}{f_{i.}f_{.j}} = 0.$$

2.3.5 Inertie totale

L'inertie totale du nuage de points est une quantité qui mesure l'écart à l'indépendace, elle est définie par[12] :

$$\varphi^2 = \frac{\mathcal{X}^2}{n}.$$

On définit l'inertie du nuage des profils-lignes et profils colonnes par rapport aux centres de gravités correspondants par :

$$Inertie\,(X_l, G_l) = \sum_{i=1}^{p} f_{i.} \times d^2_{\mathcal{X}^2}\,(i, G_l)\,,$$

$$Inertie\,(X_c, G_c) = \sum_{j=1}^{q} f_{.j} \times d^2_{\mathcal{X}^2}\,(j, G_c)\,.$$

Proposition 2.3.1 *L'inertie du nuage des profils-lignes est identique à celle des profils colonnes et égale à* φ^2.

Preuve. [12]On a

$$\begin{aligned}
Inertie\,(X_l, G_l) &= \sum_{i=1}^{p} f_{i.} \times d^2_{\mathcal{X}^2}\,(i, G_l) \\
&= \sum_{i=1}^{p} f_{i.} \times \sum_{j=1}^{q} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - f_{.j}\right)^2 \\
&= \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{f_{i.}}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - f_{.j}\right)^2 \\
&= \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{f_{i.}}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i.}f_{.j}}{f_{i.}}\right)^2 \\
&= \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{f_{i.}}{f_{.j}} \left(\frac{f_{ij} - f_{i.}f_{.j}}{f_{i.}}\right)^2
\end{aligned}$$

$$\begin{aligned} Inertie\,(X_l, G_l) &= \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{f_{i.}}{f_{.j}} \frac{(f_{ij} - f_{i.}f_{.j})^2}{(f_{i.})^2} \\ &= \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{1}{f_{.j}} \frac{(f_{ij} - f_{i.}f_{.j})^2}{f_{i.}} \\ &= \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{(f_{ij} - f_{i.}f_{.j})^2}{f_{i.}f_{.j}} \\ &= \frac{\mathcal{X}^2}{n} = \varphi^2. \end{aligned}$$

De la même manière on peut montrer que $Inertie\,(X_c, G_c) = \varphi^2$. ■

2.4 ACP des deux nuages de profils

L'analyse factorielle des correspondances est définie comme étant le résultat d'une double ACP des deux profils qui sont en dualité.

2.4.1 ACP du nuages de profils-lignes et des profils-colonnes

On définit les éléments dont on a besoin pour faire une ACP sur les deux profils [11] :

1. **Les éléments de l'ACP du nuages des profils-lignes**
   - Tableau des données : $X = X_l = D_l^{-1}N$.
   - Métrique : $M = M_l = D_c^{-1}$.
   - Matrice de poids : $D = D_l$.
   - Centre de gravité : $G = G_l = X_l^t D_l \mathbf{1}_p$.
2. **Les éléments de l'ACP du nuages des profils-colonnes**
   - Tableau des données : $X = X_c = D_c N^t$.
   - Métrique : $M = M_c = D_l^{-1}$.
   - Matrice de poids : $D = D_c$.

– Centre de gravité : $G = G_c = X_c^t D_c \mathbf{1}_q$.

On a besion aussi de la matrice de variance covariance : $V = X^t DX - GG^t$.

Pour faire une ACP sur les profils, on cherche les valeurs propres de la matrice VM, appliquant la méthode expliquée au premier chapitre.

2.4.2 ACP non centrées et facteur trivial

Comme G est $M-orthogonal$ au nuage de points signifie que c'est un vecteur propre de VM (facteur principal) associé à la valeur propre $\lambda = 0$. On prend le cas du nuage profils-ligne et on démontre que $V_l M_l G_l = 0 G_l = 0_{\mathbb{R}^p}$ [3].

Preuve.

$$\begin{aligned}
V_l M_l G_l &= \left(X_l^t DX_l - G_l G_l^t\right) M_l G_l \\
&= \left(X_l^t D_l X_l - G_l G_l^t\right) D_c^{-1} G_l \\
&= \left(X_l^t D_l X_l\right) \mathbf{1}_q - G_l G_l^t \left(D_c^{-1} G_l\right) \text{ , car } D_c^{-1} G_l = \mathbf{1}_q \\
&= \left(X_l^t D_l X_l\right) \mathbf{1}_q - G_l \left(G_l^t D_c^{-1} G_l\right) \\
&= \left(X_l^t D_l X_l\right) \mathbf{1}_q - G_l \parallel G_l \parallel^2_{D_c^{-1}} \\
&= X_l^t D_l X_l \mathbf{1}_q - G_l, \text{ car } G_l \text{ est } D_c^{-1} - orthogonal \text{ au nuage de } pl \\
&= X_l^t D_l \left(D_l^{-1} N\right) \mathbf{1}_q - G_l \text{ ,car } X_l = D_l^{-1} N \mathbf{1}_q \\
&= X_l^t D_l D_l^{-1} N \mathbf{1}_q - G_l \\
&= X_l^t G_c - G_l \text{ , car } N\mathbf{1}_q = G_c \\
&= G_l - G_l \text{ , car } X_l^t G_c = G_l \\
&= 0.
\end{aligned}$$

De même pour le cas du nuage profils-colonnes. Alors, G un vecteur propre de VM. ■

Proposition 2.4.1 [12] *Les deux matrices VM et $X^t DXM$ ont les même vecteurs propres associés aux mêmes valeurs propres, sauf G qui est alors associé à la valeur*

propre 1.

Soit v un vecteur propre de VM, $M-orthogonal$ à G et de valeur propre λ, on montre que $VMv = X^tDXMv$.

Preuve.

$$\begin{aligned} VMv = \lambda v \Rightarrow X^tDXMv &= VMv - GG^tMv \\ &= \lambda v - G\left\langle G, v\right\rangle_M \\ &= \lambda v \text{ car } v \perp G \ (v \text{ et } G \text{ sont deux vecteur propre de } VM). \end{aligned}$$

D'où le résultat. ■

On montre que $VMG = X^tDXMG$

Preuve.

$$\begin{aligned} X^tDXMG &= VMG - GG^tMG \\ &= 0 + G \parallel G \parallel_M^2 \\ &= G. \end{aligned}$$

Ce qui imlique que $\lambda = 1$. ■

Alors, on peut faire l'ACP sans centrer le nuage des profils, on utilise directement dans le cas de nuage profils-lignes la matrice $A_l = X_l^tDX_c^t = N^tD_l^{-1}ND_c^{-1}$ et dans le cas de nuage profils-colonnes la matrice $A_c = X_l^tDX_c^t = ND_c^{-1}N^tD_l^{-1}$[12].

Remarque 2.4.1 *Les deux matrice A_l et A_c ont les même valeurs propres non nulles et on écrit :*

$$rangA_l = rangA_c = \tau + 1.$$

En effet, on a $rangV_lM_l \leq q-1$ car V_lM_l est une matrice carrée de $\mathcal{M}(q \times q)$ admet G_l comme vecteur propre associer à $\lambda = 0$. Avec le même raisonnement, on en conclus que $rangV_cM_c \leq p-1$, ce qui implique

$$0 < \tau \leq \min(p-1, q-1).$$

En pratique on fait l'ACP sur la matrice de plus petite taille[3].

2.4.3 Résumé des deux ACP

On résume les résultats des deux ACP dans le tableau suivant[15] :

	Facteurs principaux	Composantes principales
ACP des pl	Vecteur propres de $D_c^{-1}N^tD_l^{-1}N$	Vecteur propres de $D_l^{-1}ND_c^{-1}N^t$
ACP des pc	Vecteur propres de $D_l^{-1}ND_c^{-1}N^t$	Vecteur propres de $D_c^{-1}N^tD_l^{-1}N$

TAB. 2.3 – Résumé des deux ACP.

On remarque que les deux ACP conduisent aux mêmes valeurs propres et que les facteurs principaux de l'une sont les composantes principales de l'autre.

Formules de transition

Les facteurs de même rang sur les lignes et les colonnes sont liés par des relations appellées relations de transition.

Proposition 2.4.2 [11] *Soit $a = (a_1, ..., a_p)^t$ un vecteur propre associé à la valeur propre $\lambda \neq 0$, M_l-norme 1 de A_r et soit $b = (b_1, ..., b_p)^t$ un vecteur propre associé à la valeur propre $\lambda \neq 0$, M_c-norme 1, de A_c. Les deux formules de transition s'écrivent comme suit :*

$$a = \frac{1}{\sqrt{\lambda}} D_l^{-1} N b \quad et \quad b = \frac{1}{\sqrt{\lambda}} D_c^{-1} N^t a.$$

Preuve. *a est un vectaur propre de $A_l = N^tD_l^{-1}ND_c^{-1}$ alors,*

$$A_l \times a = \lambda a$$

$$N^tD_l^{-1}ND_c^{-1} \times a = \lambda a$$

$$\left(ND_c^{-1}\right) N^t D_l^{-1} N D_c^{-1} a = \lambda \left(ND_c^{-1}\right) a$$
$$X_c^t X_l^t \left(ND_c^{-1} a\right) = \lambda \left(ND_c^{-1} a\right)$$
$$A_c \left(ND_c^{-1} a\right) = \lambda \left(ND_c^{-1} a\right).$$

■

On remarque que $b = kND_c^{-1}a$ est un vecteur propre de la matrice $A_c = ND_c^{-1}N^tD_l^{-1}$, pour que ce vecteur soit de M_c-norme 1 il faut trouver une constante k tel que $\| kND_c^{-1}a \|_{M_c} = 1$.

Preuve.

$$\left(kND_c^{-1}a\right)^t M_c \left(kND_c^{-1}a\right) = 1$$
$$k^2 \left(a^t D_c^{-1} N^t\right) D_l^{-1} \left(ND_c^{-1}a\right) = 1$$
$$k^2 a^t \left(D_c^{-1} N^t D_l^{-1} N D_c^{-1}\right) a = 1$$
$$k^2 a^t D_c^{-1} \left(A_l a\right) = 1$$
$$k^2 a^t D_c^{-1} \left(\lambda a\right) = 1$$
$$\lambda k^2 a^t D_c^{-1} a = 1$$

comme on a $\| a \|_{M_l} = 1$ alors,

$$\lambda k^2 = 1 \Rightarrow k = \frac{1}{\sqrt{\lambda}}.$$

D'où le résultat. ■

Corollaire 2.4.1 *[4]Pour tout $k = 1, 2..., \tau$ on a :*

$$0 \leq \lambda_k \leq 1.$$

Preuve. (voir[4]). ■

2.4.4 La décomposition de l'inertie

On sait que l'inertie totale (et donc la somme des valeurs propres) est égale à φ^2

Comme il y a au plus $\min(p-1, q-1)$ valeurs propres[15], on obtient si $p < q$

$$\varphi^2 = \sum_{k=1}^{p} \lambda_k.$$

La contribution des modalités aux inerties des axes

Les contributions sont définit comme suit[6] :

La contribution de la i^{eme} profil-ligne(pl_i) :

$$CTR(pl_i) = \frac{f_{i.} \times a_{i.}^2}{\lambda}; \quad i = 1, ..., p.$$

Avec $a_{i.}$ la i^{eme} coordonnée de a.

La contribution de la j^{eme} profil-colonne(pc_j) :

$$CTR(pc_j) = \frac{f_{.j} \times b_{.j}^2}{\lambda}; \quad j = 1, ..., q.$$

Avec $b_{.j}$ la j^{eme} coordonnée de b.

En pratique, on considère les modalité ayant les plus fortes contributions, lorsqu'elle dépassent son poids

$$CTR(pl_i) > f_{i.} \text{ et } CTR(pc_j) > f_{.j}.$$

Chapitre 3

Exemple d'application avec R

Dans ce chapitre qui représente la partie pratique de notre travail nous allons appliquer la méthode de l'AFC sur des données réelles contenant dans les packages de R.

3.1 Logiciel R

Le logiciel R est un language de programmation et un environnement mathématique riche par ses fonctions qui permettent de manipuler des données, obtenir des représentations graphiques et enfin réalisation des analyses statistiques telles que : les tests d'hypothèses, l'analyse de la variance, les méthodes de régression linéaire (simple et multiple)...etc, et mieu d'autre applications .

Ce logiciel a été créé en 1996 par Ross Ihaka et Robert Gentleman à l'Université d'Auckland, en Nouvelle-Zélande, et est actuellement développé par AT&T Bell Laboratories en 1988. Il est disponible gratuitement (on peut le télécharger par internet) et il fonctionne sous les systèmes d'exploitation Linux, windows et MacOS[7].

3.1.1 Les packages

Dans le logiciel R on trouve plusieurs packages qui permettent de réaliser l'AFC, parmi ceux-là on peut sité[4] :

- Le package **FactoMineR** (Factor analysis and Data Mining with R).
- le package **ade4** (Analysis of Environmental Data : Exploratory and Euclidean method).
- le package **MASS**.
- le package **ca**

Les fonctions

- La fonction **CA()** du package **FactoMineR**.
- La fonction **dudi.coa()** du package **ade4**.
- La fonction **corresp()** du package **MASS**.
- La fonction **ca**() du package **ca.**

Dans notre exemple on va utiliser le package **FactoMineR** pour faire l'AFC et **factoextra** pour représenter graphiquement les données, tout d'abord, il faut installer et charger ces packages exécutant les commandes suivantes[9] :

- Pour installer les package :

  > install.packages("FactoMineR")

  > install.packages("factoextra")

- Pour charger les packages :

  > library(FactoMineR)

  > library(factoextra)

3.2 Les données

Pour réaliser l'AFC on va utiliser le jeu de donnée **"children"** trouver dans le package **FactoMineR**.

Ce jeu de données est un tableau de contingence qui résume les réponses données par différentes catégories de personnes à la question suivante : selon vous, quelles sont les raisons qui peuvent faire hésiter une femme ou un couple à avoir des enfants ? Les lignes du tableau représent les raisons pour ne pas faire d'enfants et les colonnes le niveau d'éducation de la personne répondant.

$> data(children)$

$> children$

	N-qualifié	CEP	BEPC	BAC	Univ	30	50	< 50
Argent	51	64	32	29	17	59	66	70
Avenir	53	90	78	75	22	115	117	86
Chômage	71	111	50	40	11	79	88	177
Circonstances	1	7	5	5	4	9	8	5
Dure	7	11	4	3	2	2	17	18
Économique	7	13	12	11	11	18	19	17
Égoïsme	21	37	14	26	9	14	34	61
Emploi	12	35	19	6	7	21	30	28
Finances	10	7	7	3	1	8	12	8
Gueurre	4	7	7	6	2	7	6	13
Logement	8	22	7	10	5	10	27	17
Craindre	25	45	38	38	13	48	59	52
Santé	18	27	20	19	9	13	29	53
Travail	35	61	29	14	12	30	63	58
Confort	2	4	3	1	4	NA	NA	NA
Désaccord	2	8	2	5	2	NA	NA	NA
Monde	1	5	4	6	3	NA	NA	NA
Vivre	3	3	1	3	4	NA	NA	NA

TAB. 3.1 – Tableau de contingence des raisons et niveau d'éducation.

Où

N-qualifié : non qualifier.

CEP : certificat d'études primères.

BEPC : brevet d'études du premier cycle.

BAC : baccalauréat.

Univ : université.

On considère les deux variables suivants :

$V_1 :=$ les raisons pour ne pas faire d'enfants.

$V_2 :=$ le niveau d'éducation de la personne répondant.

On va étudier la liaison entre ces deux variable.

3.2.1 Code R pour faire l'AFC

On prend les trois dernières colonnes qui expriment la classe d'âge comme colonne supplémentaire. On effectue l'AFC en utilisant la fonction CA qui donne une liste contenant les valeurs propres, les coordonnées des lignes et colonnes, la qualité de représentation et les contributions de deux profils...etc.

$res.afc < -CA(children[, 1:5], graph = FALSE)$

Parmi les résultats données par CA on trouve le résultat du test du khi deux qui est égale à 123.8249 ($p - value = 4.145257e - 05 < \alpha = 0.05$), ce qui assure qu'il y a une liaison forte entre les deux variables.

– Calcule des valeurs propres :

Les valeurs propres non nulles calculer indiquent le nombre des axes principaux à retenir, telle que la plus grande valeur propre est asocciée au premier axe

principale et la deuxième valeur propre est asocciée au deuxième axe et ainsi de suite. Alors, les valeurs propres expriment la quantité d'information extraite par chaque axe.

$> res.afc\$eig$

	Eigenvalues	% of *var*	cumulative % of *var*
dim 1	0.038	51.894	51.894
dim 2	0.018	25.287	77.182
dim 3	0.008	11.900	89.082
dim 4	0.008	10.917	100.000

TAB. 3.2 – Tableau des valeurs propres et pourcentages de variances.

Selon le tableau 3.2, les deux premières dimensions représentent 77.18% de la variance totale, c'est un pourcentage acceptable et 89.08% de la variance totale est exprimé par les trois premières dimensions, l'analyse est donc de bonne qualité.

Résultats pour les lignes

– Les coordonnées (les composantes principales en lignes) :

$> res.afc\$ \text{row} \$coord$

	Dim 1	Dim 2	Dim 3	Dim 4
Argent	-0.117	0.066	0.092	0.074
Avenir	0.140	-0.145	-0.036	0.022
Chômage	-0.233	-0.047	0.018	-0.025
Circonstances	0.426	0.204	-0.118	-0.118
Dure	-0.245	0.107	0.046	-0.022
Économique	0.386	0.268	-0.019	0.097
Égoïsme	0.052	-0.024	0.174	-0.108
Emploi	-0.130	0.148	-0.259	-0.091

	Dim 1	Dim 2	Dim 3	Dim 4
Finances	-0.279	-0.110	0.0129	0.353
Gueurre	0.180	-0.136	-0.087	0.050
Logement	0.003	0.092	0.050	-0.215
Craindre	0.174	-0.110	-0.028	0.010
Santé	0.097	-0.012	0.012	0.053
Travail	-0.20	0.117	-0.070	-0.001
Confort	0.300	0.669	-0.125	0.170
Désaccord	0.158	0.054	0.128	-0.331
Monde	0.525	0.027	0.021	-0.120
Vivre	0.398	0.565	0.358	0.142

– Les contributions :

$> res.afc\$\,\text{row}\,\$contrib$

	Dim 1	Dim 2	Dim 3	Dim 4
Argent	4.142	2.715	11.210	8.018
Avenir	9.791	21.646	2.859	1.232
Chômage	23.965	2.029	0.660	1.383
Circonstances	6.237	2.944	2.108	2.292
Dure	2.529	1.002	0.402	0.099
Économique	12.539	12.431	0.139	3.792
Égoïsme	0.461	0.198	22.175	9.337
Emploi	2.109	5.543	36.173	4.860
Finances	3.413	1.085	0.031	25.930
Gueurre	1.320	1.558	1.364	0.496
Logement	0.001	1.410	0.885	17.814

	Dim 1	Dim 2	Dim 3	Dim 4
Craindre	7.513	6.188	0.891	0.139
Santé	1.374	0.047	0.098	1.984
Travail	10.266	6.647	5.122	0.003
Confort	1.971	20.041	1.504	2.998
Désaccord	0.738	0.178	2.117	15.470
Monde	8.171	0.046	0.059	2.030
Vivre	3.453	14.283	12.194	2.113

– Le cosinus carré (cos2) :

Le cos2 est vu comme un indicateur de qualité de représentation d'un nuage ou d'un point par un axe.

$> res.afc\$\,\mathrm{row}\,\$cos2$

	Dim 1	Dim 2	Dim 3	Dim 4
Argent	0.426	0.136	0.264	1.734e-01
Avenir	0.460	0.496	0.030	1.220e-02
Chômage	0.943	0.038	0.005	1.146e-02
Circonstances	0.722	0.166	0.055	5.584e-02
Dure	0.807	0.155	0.029	6.656e-03
Économique	0.645	0.311	0.001	4.107e-02
Égoïsme	0.060	0.012	0.668	2.582e-01
Emploi	0.149	0.191	0.587	7.237e-02
Finances	0.362	0.056	0.000	5.800e-01
Gueurre	0.528	0.304	0.125	4.180e-02
Logement	0.000	0.148	0.043	8.078e-01
Craindre	0.698	0.280	0.018	2.733e-03
Santé	0.747	0.012	0.012	2.272e-01

	Dim 1	Dim 2	Dim 3	Dim 4
Travail	0.699	0.220	0.080	4.439e-05
Confort	0.155	0.768	0.027	4.961e-02
Désaccord	0.161	0.019	0.106	7.128e-01
Monde	0.946	0.002	0.001	4.947e-02
Vivre	0.252	0.509	0.204	3.256e-02

Résultats pour les colonnes

– Les coordonnées (les composantes principales en colonnes) :

	Dim 1	Dim 2	Dim 3	Dim 4
N-qualifié	-0.226	-0.004	0.097	0.114
CEP	-0.133	0.043	-0.028	-0.103
BEPC	0.078	-0.083	-0.154	0.080
BAC	0.259	-0.159	0.102	-0.046
Univ	0.330	0.383	0.027	0.051

– Les contributions :

$> res.afc\$\,\mathrm{col}\,\$contrib$

	Dim 1	Dim 2	Dim 3	Dim 4
N-qualifié	26.439	0.026	21.356	32.214
CEP	15.522	3.403	2.999	44.480
BEPC	3.203	7.407	53.534	15.829
BAC	31.380	24.393	21.381	4.750
Univ	23.453	64.768	0.728	2.725

– Le cosinus carré :

$> res.afc\$\,\mathrm{col}\,\$cos2$

	Dim 1	Dim 2	Dim 3	Dim 4
N-qualifié	0.693	0.000	0.128	0.177
CEP	0.570	0.060	0.025	0.343
BEPC	0.142	0.161	0.547	0.148
BAC	0.638	0.241	0.099	0.020
Univ	0.420	0.566	0.002	0.010

Représentation graphique des données

Pour visualiser les données on va utiliser le package **factoextra**.

– Représentation des valeurs propres :

On utilise la fonction barplot() qui permet de tracer l'histogramme des valeurs propre.

$> barplot(res.afc\$eig[,1], main = "Valeurspropres", col = "steelblue")$

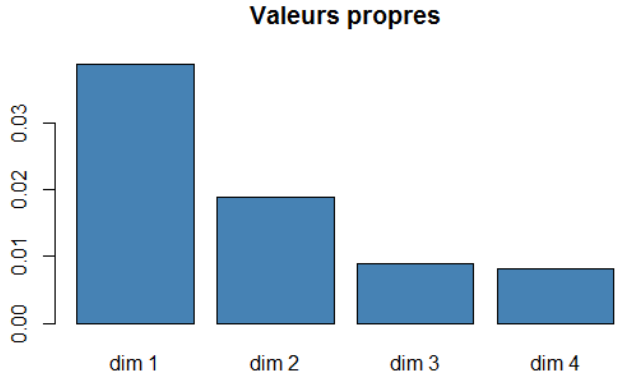

FIG. 3.1 – Les valeurs propres et les pourcentages de variances.

– Représentation des pourcentages d'inerties :

Pour tracer le diagramme en barre des pourcentages d'inerties on utilise la fonction fviz_screeplot() trouvée dans le package factoextra.

$> fviz\_screeplot(res.afc, addlabels = TRUE)$

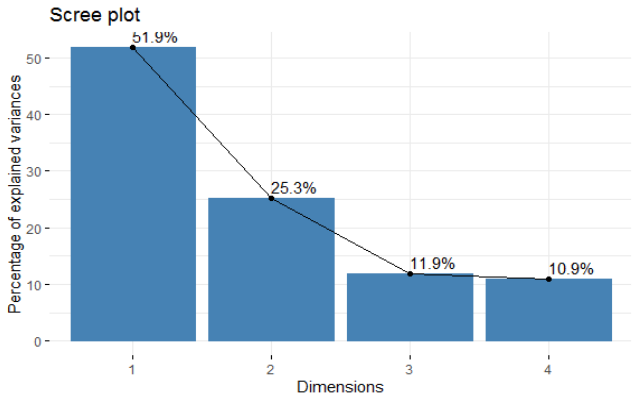

FIG. 3.2 – Pourcentage d'inertie associé à chaque axe.

– Représentation de cos2 des lignes :

On utilise la commande suivante pour tracer le diagramme en barres de cos2 des lignes sur le premier plan.

$> fviz\_cos2(res.afc, choice = "row", axes = 1:2)$.

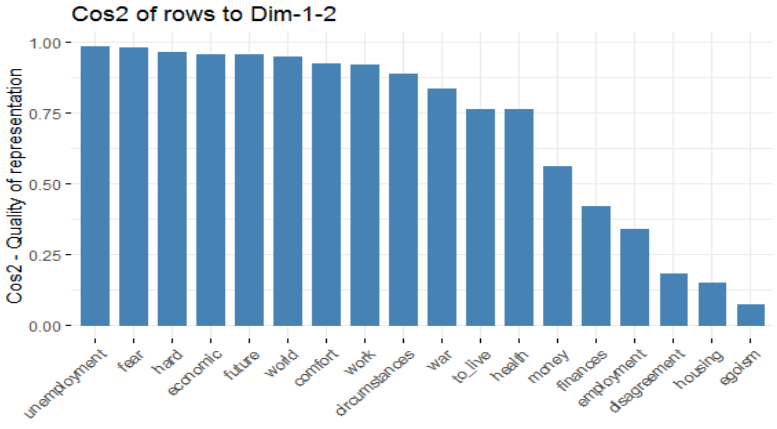

FIG. 3.3 – Qualité de représentation des lignes sur le premier plan

Si la qualité de représentation est proche de 1 on dit que le point est bien représenté sur cet axe.

– Représentation des contribution :

Pour visualiser la contribution des lignes sur le premier axe on tape la commande suivante :

$> fviz\_contrib(res.afc, choice = "row", axes = 1)$

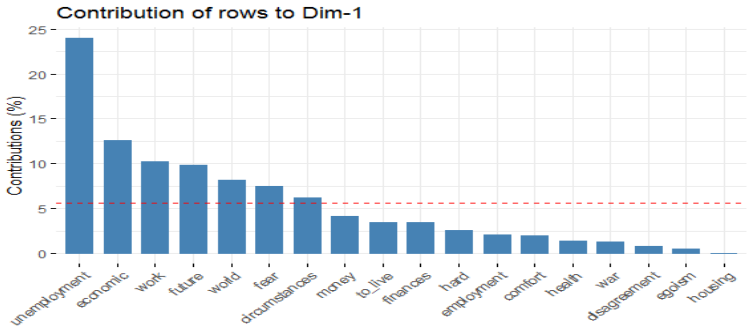

FIG. 3.4 – Contributions des lignes sur le premier axe.

– Représentation graphique :

Pour la visualisation superposée des lignes et des colonnes sur le premier plan, on tape la commande suivante :

$> plot(res.afc)$

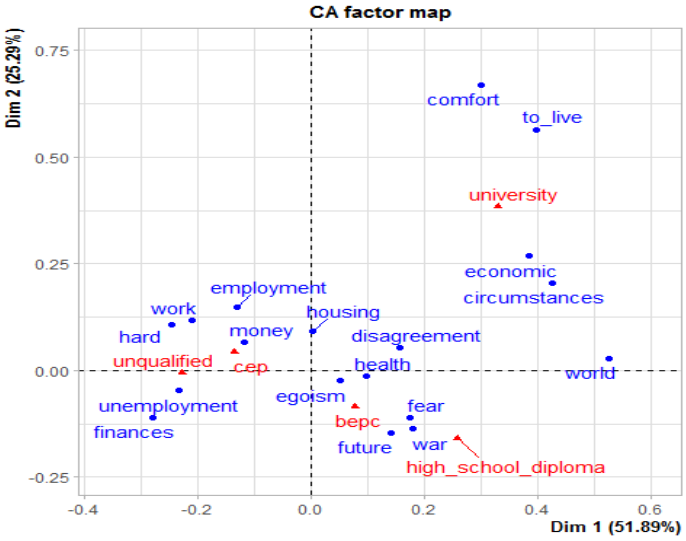

FIG. 3.5 – Représentation superposée de V_1 et V_2 sur le premier plan

Dans le graphique ci-dessus, les points en bleus représentent les lignes et les triangles en rouges représentent les colonnes.

Interprétation du plan factoriel

le graphique 3.5 représente la projection des modalités sur le premier plan factoriel, tel qu'il exprime un pourcentage de 77.18% d'inertie totale. En remarque que :

- Les raisons (chômage, finance) sont associés au niveau d'éducation non-qualifier.
- Les raisons (argent, travail, dure, emploi et Logement) sont associés au niveau CEP.
- Les raisons (égoïsme, santé et désaccord) sont associer au niveau BEPC.
- Les raisons (guerre, craindre, avenir et monde) sont associés au niveau baccalauréat.
- Les raisons (économie, circonstance, vivre et confort) sont associés au niveau universitaire.

Suivant cette représentation des données, on peut conclure que les personnes sans emploi sont plus inquiètes par le chômage en générale par contre les personnes diplomés sont plus inquiètes par le niveau économie globale, les personnes de haut niveau (université) indique que la perte de confort et le mal vivre sont des causes importantes pour ne pas avoir d'enfant.

Conclusion

L'analyse factorielle des correspondances est une méthode utilisée dans différents domaines, tels que l'économie, le marketing, la gestion,...etc et qui a pour but l'étude de la liaison entre les deux variables ainsi que la réduction de dimention à condition de conserver le maximum d'information.

Dans ce mémoire on a essayé de présenter le principe de l'AFC ainsi que la méthode de calcule et on a illustrer ce travail avec une application de cette méthode sur des données réelles en utilisant le language R, ainsi qu'une version en SPSS présentée en annexe.

Lorsque le tableau de données a plus de deux variables qualitatives, il existe une autre méthode appelée analyse factorrielle multiple qui traite ce genre de problème.

Bibliographie

[1] Alain, B. (2010). Statistique Descriptive Multidimensionnelle, L'Institut de Mathématiques de Toulouse.

[2] Arnaud, M. L'analyse de données Polycopié de cours ENSIETA-Réf : 1463. Septembre 2004

[3] Baey, C. (2019). Analyse de données. M2 Ingénierie Statistique et Numérique. Université de Lille.

[4] Boumaza, R. (2007). *Analyse des données* (Vol. 16). Centre de publication universitaire.

[5] Bounkhala, A. (2017). Méthodes ACP et AFC en statistiques et leurs applications, Tlemcen.

[6] Chavent, M. (2014-2015). Notions de base pour l'analyse d'un tableau de contingence, Université de Bordeaux-MASTER MIMSE-2^{eme}année.

[7] De Micheaux, P. L., Drouilhet, R., Liquet, B., & DODGE, Y. (2014). *Le logiciel R : maîtriser le langage, effectuer des analyses (bio) statistiques*. Springer.

[8] Escofier, B., & Pagès, J. (2008). *Analyses factorielles simples et multiples*. Dunod, Paris.

[9] Kassambara, A. (2017). *Practical guide to principal component methods in R : PCA, M (CA), FAMD, MFA, HCPC, factoextra* (Vol. 2). Sthda.

[10] Lasgouttes, J. M. (2019-2020). Cours d'analyse de données.

[11] Lebart, L., Morineau, A., & Piron, M. (1995). *Statistique exploratoire multidimensionnelle* (Vol. 3). Paris : Dunod.

[12] Necir, A. (2021).Cours de master 1 Modèle Linéaire. Université Mohammed Khider Biskra.

[13] NEGGAZ, N. (2021). Cours d'Analyse des Données.

[14] Saporta, G. (2006). *Probabilités, analyse des données et statistique*. Editions technip.

[15] THIARE, O. (16 avril 2020). Analyse factorrielle des correspondance(AFC).

[16] Tillé, Y. (2010). Résumé du cours de statistique descriptive.

Annexe : Application avec SPSS

Logiciel SPSS

L'SPSS (Satistical Package for Social Science) est un logiciel utiliser pour les analyses statistiques. Ce language a été créé en 1968 par Norman.H.Nie.c.Hadlaï (Tex) et Dale.H.Bent.

Ce logiciel permi d'éffecuer plusieurs analyses sur de grand bases de données. Il fonctionne sous les systèmes d'exploitation Linux, MacOS et Windows.

Les données et leur résultats

On prend les même données **"children"** contenant dans le logiciel R et on esseye d'éffectuer l'AFC avec le logiciel SPSS.

– Tableau des données :

Le tableau ci-dessous croise les deux variables A1 et A2 comme suit :

Table des correspondances

A1	A2 bepc	cep	high_school_ diploma	university	unqualified	Marge active
Argent	25	35	24	15	32	131
Avenir	38	39	37	20	33	167
Chômage	31	40	29	11	36	147
Circonstances	5	7	5	4	1	22
Confort	3	4	1	4	2	14
Craindre	28	30	28	13	21	120
Désaccord	2	8	5	2	2	19
Dure	4	11	3	2	7	27
Économique	12	13	11	11	7	54
Égoïsme	14	27	22	9	19	91
Emploi	17	26	6	7	12	68
Finances	7	7	3	1	10	28
Gueurre	7	7	6	2	4	26
Logement	7	20	10	5	8	50
Monde	4	5	6	3	1	19
Santé	18	23	17	9	16	83
Travail	24	34	14	12	26	110
Vivre	1	3	3	4	3	14
Marge active	247	339	230	134	240	1190

FIG. 3.6 – Tableau de contingence.

– Tableau récaptulatif :

Récapitulatif

Dimension	Valeur singulière	Inertie	Khi-deux	Sig.	Proportion d'inertie Représentation	Proportion d'inertie Cumulé	Valeur singulière de confiance Ecart type	Valeur singulière de confiance Corrélation 2
1	,155	,024			,435	,435	,028	,017
2	,116	,013			,245	,679	,028	
3	,103	,011			,193	,872		
4	,084	,007			,128	1,000		
Total		,055	65,673	,557[a]	1,000	1,000		

a. 68 degrés de liberté

FIG. 3.7 – Tableau des résultats de l'analyse.

Le tableau 3.7 résume les inerties, les inerties expliquées par chaque axe, le résulat du test de khi-deux,...etc, d'aprés le résultat donné par le test de khi-deux on conclue qu'il y a une liaison entre les deux variables A1 et A2.

– Résultats pour les lignes :

Présentation des points de ligne[a]

		Score de la dimension			Contribution				
					Du point vers l'inertie de la dimension		De la dimension vers l'inertie du point		
A1	Masse	1	2	Inertie	1	2	1	2	Total
Argent	,110	,151	-,062	,001	,016	,004	,294	,038	,332
Avenir	,140	-,155	-,276	,002	,022	,092	,234	,558	,792
Chômage	,124	,310	-,255	,003	,077	,069	,657	,334	,991
Circonstances	,018	-,935	,421	,003	,104	,028	,772	,118	,890
Confort	,012	-,694	1,200	,004	,037	,146	,207	,464	,671
Craindre	,101	-,199	-,280	,002	,026	,068	,314	,469	,783
Désaccord	,016	-,414	,340	,003	,018	,016	,142	,071	,213
Dure	,023	,674	,585	,003	,067	,067	,524	,296	,820
Économique	,045	-,736	,215	,005	,159	,018	,777	,050	,827
Égoïsme	,076	-,042	-,192	,002	,001	,024	,010	,150	,160
Emploi	,057	,313	,756	,006	,036	,281	,147	,644	,791
Finances	,024	1,118	-,267	,005	,190	,014	,866	,037	,903
Gueurre	,022	-,097	-,303	,001	,001	,017	,030	,217	,247
Logement	,042	-,038	,419	,003	,000	,064	,003	,259	,262
Monde	,016	-1,074	-,225	,003	,119	,007	,860	,028	,889
Santé	,070	-,046	-,083	,000	,001	,004	,173	,424	,597
Travail	,092	,332	,295	,003	,066	,069	,547	,323	,871
Vivre	,012	-,902	,330	,004	,062	,011	,336	,034	,369
Total actif	1,000			,055	1,000	1,000			

a. Normalisation symétrique

FIG. 3.8 – Résultats pour lignes

– Résultats pour les colonnes :

Présentation des points de colonne[a]

A2	Masse	Score de la dimension 1	Score de la dimension 2	Inertie	Contribution : Du point vers l'inertie de la dimension 1	Contribution : Du point vers l'inertie de la dimension 2	Contribution : De la dimension vers l'inertie du point 1	Contribution : De la dimension vers l'inertie du point 2	Contribution : De la dimension vers l'inertie du point Total
bepc	,208	,047	-,083	,006	,003	,012	,011	,026	,038
cep	,285	,123	,381	,009	,028	,355	,074	,533	,607
high_school_diploma	,193	-,396	-,508	,012	,196	,430	,380	,469	,848
university	,113	-,713	,387	,014	,370	,145	,617	,136	,753
unqualified	,202	,557	-,182	,013	,403	,057	,734	,059	,792
Total actif	1,000			,055	1,000	1,000			

a. Normalisation symétrique

FIG. 3.9 – Résultats pour les colonnes

Les deux tableaux 3.8 et 3.9 nous donnent une idée sur l'importance de chacune des modalités des deux variables étudiées.

– La représentation superposée des lignes et colonnes sur le plan :

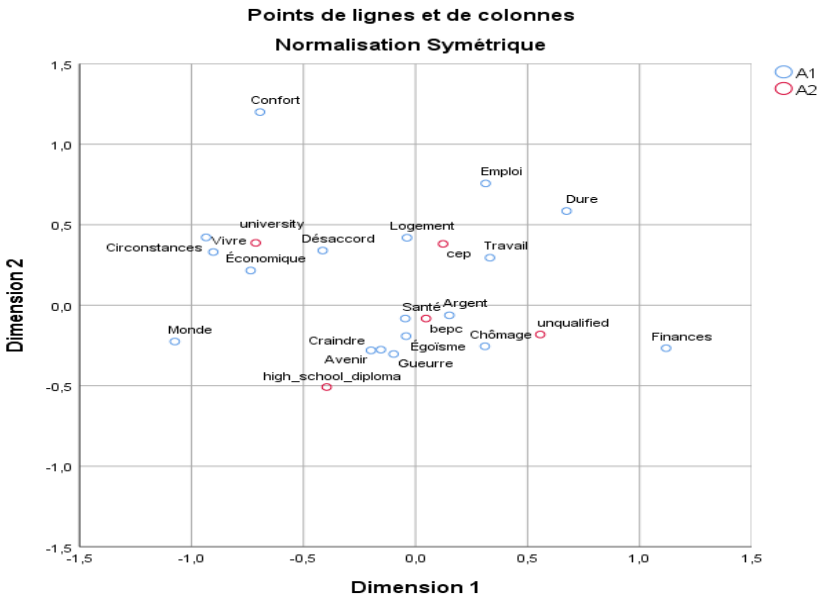

FIG. 3.10 – La représentation des lignes et des colonnes sur le 1^{er} plan

Selon le graphe3.10 on cite les remarques suivantes :

- Les raisons (vivre, circonstance, économie,confort et désaccord) sont proche l'une de l'autre alors on peut les regrouper ensemble et son associer au niveau d'éducation universitaire.
- Les raisons (logement, travail,emploi et dure) sont aussi proche l'une de l'autre et sont associer au niveau CEP.
- Les raisons (santé,égoïsem et argent) sont associer au niveau BEPC.
- Les raisons (chomâge et finance) sont associer au niveau non qualifier.
- Les raisons (craindre, avenir et guerre) sont proche et on peut les associer au niveau baccalauréat.

ملخص

في هذه المذكرة، حاولنا التكلم عن إحدى طرق تحليل العوامل وهي التحليل العاملي للمراسلات. بدأنا بتعريف الطريقة والغرض منها، بعدها قمنا بإعطاء بعض المفاهيم الأساسية التي تسهل تحقيق هذه الطريقة مثل طريقة تحليل المركبات الرئيسية، جدول تقاطع البيانات، وقياس مربع كاي... إلخ. ثم تطرقنا إلى مبدأ التحليل العاملي للمراسلات وكيفية تنفيذه، وأخيرا قمنا بتوضيح هذا العمل بتطبيق الطريقة على البيانات الحقيقية باستخدام البرنامجين R و SPSS.

الكلمات المفتاحية: التحليل العاملي للمراسلات، تحليل المركبات الرئيسية، قياس مربع كاي.

Résumé

Dans ce mémoire on a essayé de représenté une des méthodes factorielle qui est l'analyse factorielle des correspondances. On a commencé par la définition et le but de la méthode, après on a essayé de données quelque notions de base qui facilitent la réalisation de cette méthode tels que l'ACP, le tableau de contingence, les profils et la métrique de Khi-deux…etc. En suite, on a parlé du principe de l'AFC et comment on l'effectuer et à la fin on a illustré ce travail avec une application de la méthode sur des données réelles en utilisant les deux logiciels: R et SPSS.

Les mots clés: Analyse factorielle des correspondances, analyse en composantes principales, métrique de khi-deux.

Abstract

In this memory we tried to represent one of the factorial methods which is the factorial analysis of correspondences. We started with the definition and the purpose of the method, then we tried some basic notions that facilitate the realization of this method such as the PCA, the contingency table, the profiles and the metric of Chi-square …etc. Then, we talked about the principle of CA and how to perform it and finally we illustrated this work with an application of the method on real data using the two software: R and SPSS.

Key words: Correspondence analysis, principal component analysis, chi-square metric.