UNIVERSITE POLYTECHNIQUE Année : 2020 – 2021

DE BINGERVILLE (UPB) Licence 2

Filière : ASSRI

ANALYSE DE DONNEES

Sommaire

Chapitre 5 : Analyse factorielle des correspondances (AFC)

1. Présentation de la méthode
2. Technique de la méthode
3. Interprétation des résultats
4. Notion de contribution

Exercices

Introduction

Dans de nombreuses situations, les données sont assez nombreuses pour pouvoir être visualisées (nombre de caractéristiques trop élevées). Il est alors nécessaire d'extraire l'information pertinente qu'elles contiennent. C'est dans ce cadre qu'interviennent les techniques d'Analyse De Données.

L'Analyse De Données, désigne l'ensemble de méthodes descriptives visant à ***résumer*** et ***visualiser*** l'information ***pertinente*** contenue dans une table de données.

Le domaine de l'analyse des données vaste, cependant il repose principalement sur trois catégories de méthodes :

	Méthode d'analyse utilisée	
Objectif de l'analyse	**Variables quantitatives**	**Variables qualitatives ou mixtes**
Repérer et visualiser les corrélations multiples entre variables et /ou les ressemblances entre individus	Analyse en composantes principales (ACP)	Analyse factorielle des correspondances (AFC, AFCM)
Réaliser une typologie des individus	Methodes de classification (CAH,..)	AFC ou AFCM et Méthodes de classification
Caractériser de groupes d'individus à l'aide de variables	Analyse discriminante (AFD,..)	Analyse discriminante (AFD,..)

Dans le présent cours, ce sont les deux premières méthodes (ACP pour les variables quantitatives et sa variante AFC pour les variables qualitatives ou mixtes) qui sont abordées.

Chapitre 1 : Rappel sur les notions de matrice et d'échantillonnage

1. Calcul matriciel

1.1. Définition d'une matrice

On appelle une matrice à n lignes et p colonnes, tout tableau formé par des coefficients réels a_{ij} avec $1 \leq i \leq n$ et $1 \leq j \leq p$. Cette matrice sera notée par $A_{n,p}$

$$A_{n,p} = \begin{pmatrix} a_{11} & \cdots & a_{1p} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{np} \end{pmatrix}$$

1.2. Déterminant d'une matrice

Soit A une matrice d'ordre n (n= p)

1.2.1. Cas d'une matrice d'ordre 2

Soit $A=\begin{pmatrix} a & c \\ b & d \end{pmatrix}$. Le déterminant de A, noté par det(A)= $\begin{vmatrix} a & c \\ b & d \end{vmatrix}$= ad-bc.

1.2.2. Cas d'une matrice d'ordre n (n ≥ 3)

Soit A= $\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix}$. Le déterminant de est défini par :

$$Dét(A)=\begin{vmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{vmatrix} = a_{11}\Delta_{11} - a_{12}\Delta_{12} + a_{13}\Delta_{13} - \ldots\ldots\ldots + (-1)^{1+n} a_{1n}\Delta_{1n}$$

$$= -a_{21}\Delta_{11} + a_{22}\Delta_{12} - a_{23}\Delta_{13} + \ldots\ldots\ldots + (-1)^{2+n} a_{2n}\Delta_{2n}$$

$$\vdots$$

$$\vdots$$

$$= (-1)^{n+1} a_{n1}\Delta_{n1} + (-1)^{n+2} a_{n2}\Delta_{n2} + (-1)^{n+3} a_{n3}\Delta_{n3} + \ldots\ldots\ldots + a_{nn}\Delta_{nn}$$

où Δ_{ij} est le déterminant mineur obtenu en supprimant la $i^{ème}$ ligne et la $j^{ème}$ colonne.

Exemple

$A=\begin{pmatrix} -1 & 2 & 3 \\ 0 & 4 & 5 \\ -2 & 0 & 1 \end{pmatrix}$, dét(A)=0

Remarque : Dans la pratique, on choisit la ligne ou la colonne qui contient le maximum de zéros

1.2.3. Règle de SARRUS

Cette règle est utilisée pour les déterminants d'ordre 3

1.3. Valeurs propres et vecteurs propres d'une matrice

1.3.1. Définition

Soit A une matrice d'ordre n et soit V un vecteur non nul de R^n. On dit que le réel α est une valeur propre associée au vecteur propre V si et seulement si AV= α V <=> AV- α V= 0 <=> (A- αI)V= 0.

Pour déterminer les valeurs propres, il suffit de résoudre l'équation det(A- αI)= 0.

Le det(A- αI) est appelé polynôme caractéristique et sera noté par P(α).

Exemples :
Exemple 1 :

Soit A= $\begin{pmatrix} -2 & 0 \\ 0 & 3 \end{pmatrix}$

Exemple 2 :

Soit A= $\begin{pmatrix} 15 & -2 & 6 \\ 21 & -2 & 9 \\ 28 & 4 & -11 \end{pmatrix}$

1.3.2. Diagonalisation d'une matrice
Soit A une matrice d'ordre n, D la matrice diagonale formée par les valeurs propres de A et P la matrice de passage formée par les vecteurs propres associés. Dans ce cas, $A=PDP^{-1}$.

2. Echantillonnage

2.1. Collecte des informations
Dans la pratique, il existe plusieurs méthodes permettant la sélection d'un échantillon d'individus dans une population dont l'objectif est d'étudier le comportement ou les opinions de certains sous-ensembles d'individus.
L'échantillonnage d'une population n'est pas le seul moyen d'acquérir une information. Il existe d'autres moyens qui feront l'objet de ce paragraphe.
2.1.1. Les principaux modes de collecte
On peut classer les méthodes de collecte en trois catégories : l'observation, l'enquête et l'expérimentation.
2.1.1.1. L'observation : il s'agit de la collecte des informations observées par l'enquêteur. Celui-ci enregistre les activités (exemple : on peut observer le comportement des consommateurs au niveau de la caisse d'un magasin ou les individus au niveau d'une municipalité).
2.1.1.2. L'enquête par sondage : l'objectif de l'enquête par sondage est e décrire une partie de la population, ses comportements, ses opinions et ses attitudes. L'objectif du chercheur est la représentativité de l'échantillon. En effet, on doit arriver à extrapoler les résultats obtenus à toute la population.
2.1.1.3. L'expérimentation : elle consiste à étudier les relations de cause à effet entre une ou plusieurs variables indépendantes qui sont manipulées par le chercheur (exemple : les différents niveaux des prix, les types de promotions).
2.1.2. Les informations utilisées dans les entreprises :
On distingue deux types d'informations utilisées dans les entreprises : les informations primaires et les informations secondaires.
2.1.2.1. Les informations primaires : L'entreprise peut collecter directement certaines données pour un besoin d'information (les études de marché, les tests de produits).
2.1.2.1. Les informations secondaires : Il s'agit d'informations qui ont été collectées et que l'entreprise peut utiliser. Ils sont de deux types : les informations internes (données comptables et financières, rapport des vendeurs) et les informations externes (information économique, banque de données).
2.1.3. Le recours à un échantillon : Il est généralement impossible de collecter les informations auprès de toute la population qui fait l'objet d'une certaine étude. On sera donc

amené à collecter l'information d'un échantillon. C'est à dire un sous-ensemble de la population.
La constitution d'un échantillon est sensiblement différente selon le type d'étude réalisée. On distingue les études qualitatives et quantitatives.
2.1.3.1. Les études qualitatives : Leurs objectif est essentiellement exploratoire. Elles permettent de comprendre le comportement et les motivations d'explorer un secteur d'activité inconnu et d'identifier les grandes dimensions d'un problème. En général, dans ce type d'études, l'échantillon est de taille faible.
2.1.3.1. Les études quantitatives : Leurs objectif est de donner une description quantifiée de certaines caractéristiques de la population étudiée avec une précision qui sera jugée suffisante par les utilisateurs. Pour atteindre cet objectif, on constitue un échantillon représentatif de la population afin de pouvoir extrapoler les résultats de la population entière.

2.2. Les étapes d'une enquête par sondage : Les deux étapes essentielles sont la rédaction des questions et le traitement et l'analyse des données.
2.2.1. La rédaction des questions : Lorsque le recueil des données s'effectue au moyen d'un questionnaire, la rédaction consiste généralement en trois étapes :
- Une première rédaction
- Un pré-test du questionnaire
- La rédaction définitive du questionnaire

2.2.2. Le traitement et l'analyse des données : Cette étape comprend les opérations de vérification, le codage et le traitement informatique. Les analyses statistiques effectuées sont le plus souvent des tris à plat (résultat des questions prises une à une). On peut aussi utiliser des tris croisés (croisement entre deux questions). D'une façon générale, lorsqu'on a un très grand nombre de données, on peut utiliser l'analyse en composantes principales (ACP) ou l'analyse factorielle des correspondances (AFC).

2.3. Le choix de l'échantillon : Faire un sondage c'est observer un sous-ensemble de la population avec comme objectif l'extrapolation des résultats de la population entière. Pour cela, trois étapes sont nécessaires :
- Le choix de la population et des individus
- Le choix de la méthode de sondage (méthode aléatoire ou méthode empirique)

2.3.1. Le choix de la population et des individus : Avant de réaliser la sélection des échantillons au moyen d'une méthode de sondage, il est indispensable de définir les individus qui sont l'objet de l'observation ainsi que la population représentant toutes les catégories d'individus. Le terme individu peut être une personne physique, un ménage ou une pharmacie.
Les unités d'échantillonnage ne sont pas toujours les individus interrogés (exemple : dans le cas d'une consommation alimentaire des enfants moins de trois ans, les unités d'échantillonnage sont les enfants mais les personnes interrogées sont leurs mères).
Dans la pratique, il existe plusieurs méthodes de sondage. Les plus connues sont les méthodes aléatoires et les méthodes empiriques.
2.3.2. La méthode aléatoire de sondage : Les méthodes aléatoires ou probabilistes sont des méthodes dans lesquels chaque individu de la population a une probabilité connue. On peut distinguer les sondages aléatoires simple, stratifié et en grappe.
2.3.2.1. Le sondage aléatoire simple : Un sondage est dit aléatoire simple lorsque tout sous-ensemble de n individus a une même probabilité d'être sélectionné. De plus le tirage de l'échantillon est réalisé sur la base de sondage sans regroupement. Le tirage peut être avec remise (non exhaustif) ou bien un tirage sans remise (exhaustif). Dans la pratique, on utilise

souvent le second tirage car il est absurde de réinterroger le même individu dans une même enquête.
Si on note par N la taille de la population et par n le cardinal du sous-ensemble, il existe C_N^n échantillons possibles. Le rapport $\frac{n}{N}$ est appelé taux de sondage.

Exemple :
Considérons un pays dans lequel ont été recensés 20000 exploitations agricoles. Lors d'une enquête portant sur 400 fermes tirées aléatoirement et constituant un échantillon aléatoire simple sans remise, on interroge les exploitants sur leur consommation d'un certains types d'engrais. Les résultats obtenus dans l'échantillon sont les suivants :
$\bar{X}$= 510 kg, S= 280 kg, N= 20000 et n= 400.
Si on note par μ la consommation moyenne par exploitation dans la population, quel est l'intervalle de confiance de la moyenne μ avec un degré de confiance égale à 95%.

Réponse :

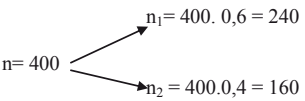
$\bar{X} - 1{,}96\frac{S}{\sqrt{n}}\sqrt{1-\frac{n}{N}} < \mu < \bar{X} + 1{,}96\frac{S}{\sqrt{n}}\sqrt{1-\frac{n}{N}}$ **[1]** => $482{,}84 < \mu < 537{,}16$

Remarque : Si $\frac{n}{N} <$ 10%, on peut négliger le terme $\sqrt{1-\frac{n}{N}}$ qui devient très proche de 1. Dans ce cas la relation [1] devient : $\bar{X} - 1{,}96\frac{S}{\sqrt{n}} < \mu < \bar{X} + 1{,}96\frac{S}{\sqrt{n}}$.

2.3.2.2. Le sondage aléatoire stratifié : La stratification consiste à découper la population en sous populations appelées strates et à réaliser un sondage dans chaque sous population.

Exemple : Reprenons l'exemple précédent et faisons un découpage de la population en deux strates comme le montre le tableau suivant :

Population	Effectif	Superficie moyenne	Ecart-type de la superficie
< 30 hectares	12000 (60%)	12 ha	6 ha
> 30 hectares	8000 (40%)	46 ha	12 ha
Total	20000		



En général, on doit choisir un échantillon stratifié proportionnel. Autrement dit, on doit prélever dans chaque strate un échantillon de taille proportionnelle à l'effectif défini par la répartition de la population.

2.3.2.3. Le sondage aléatoire en grappe et le sondage à plusieurs degrés : Ces deux méthodes sont utilisées lorsqu'on ne dispose pas d'une base de sondage des individus mais d'une liste de sous ensembles appelés grappes.

Exemple : Considérons une étude relative au lancement d'un nouveau produit financier et supposons que la banque commanditaire de sondage comporte 50 agences. Pour des raisons de coûts, on décide que l'enquête se déroule en 10 agences.
Si l'on interroge tous les clients des 10 agences tirés aléatoirement parmi les 50, on dit qu'on a réalisé un sondage aléatoire en grappe.
Si dans chacune des 10 agences, on interroge un échantillon aléatoire de clients, on dit qu'on a réalisé un sondage aléatoire de deux degrés.
L'avantage de ces méthodes est qu'elles permettent de réduire le coût de l'enquête.

2.3.3. Les méthodes empiriques de sondage : Dans ces méthodes appelées aussi méthodes non aléatoires, l'échantillon est sélectionné sans tirage aléatoire des individus. La méthode la

plus utilisée dans la pratique est la méthode des quotas. Le principe de cette méthode consiste à construire un échantillon qui soit un modèle réduit de la population étudiée.
Dans une première étape, on choisit quelques caractéristiques dont on connait la distribution dans la population étudiée (l'âge, le sexe, la région,…).
Dans une seconde étape, on donne à chaque enquêteur un plan de travail qui lui impose le respect de ces mêmes distributions au sein de l'ensemble des interviews qu'il va réaliser.
Exemple : On considère un enquêteur qui veut réaliser 10 interviews auprès des individus de 15 ans et plus. La répartition des individus dans la population est définie par le tableau suivant :

Sexe		Age		Nombre d'individus dans l'échantillon
Hommes	48%	15-24 ans	16,1%	1
		25-34 ans	17,9%	2
Femmes	52%	35-49 ans	26,7%	3
		50-64 ans	19,2%	2
		+ 64 ans	20,1%	2
Total	100%		100%	10

Analyse uni-variée et tests non paramétriques

1. Analyse univariée

1.1. Description d'une série statistique

Lorsqu'on mène une étude statistique, on s'intéresse souvent à des objets appelés individus dont l'ensemble constitue une population qu'on étudie suivant un certain nombre de critères appelés caractères. A chacun de ces critères, on associe une variable pouvant être qualitative ou quantitative. Les réalisations possibles d'une variable qualitative correspondent à un ensemble de modalités. Celles d'une variable quantitative correspondent à un ensemble de valeurs.

1.2. Organisation et représentation des données

1.2.1 Variable qualitative

Dans le cas des variables qualitatives, l'information est résumée dans un tableau de distribution qui comporte, dans une première colonne les modalités et dans une seconde les effectifs. La représentation graphique se fait par le biais des diagrammes en secteurs, des diagrammes en barres ou des diagrammes figuratifs.

1.2.2. Variable quantitative

On peut distinguer deux types de variables quantitatives : des variables discrètes et des variables continues. Les dernières sont susceptibles de prendre n'importe quelle valeur dans un intervalle donné. La représentation se fait par un tableau de distribution et graphiquement, on utilise les diagrammes en bâtons pour les variables discrètes et les histogrammes pour les variables continues. Parfois, on utilise les courbes des effectifs cumulés ou des fréquences cumulées.

1.3. Définition des indicateurs

Les indicateurs les plus utilisés sont les indicateurs de position (mode, médiane et moyenne arithmétique), les indicateurs de dispersion (variance et écart-type) et les indicateurs de forme (coefficient d'asymétrie ou d'aplatissement).

1.3.1. Les indicateurs de position :

1.3.1.1. Le mode : C'est la modalité ou la valeur qui correspond au plus grand effectif. Dans le cas des variables continues, on parle de classe modale (on utilise les effectifs ou les fréquences corrigées).

1.3.1.2. La moyenne arithmétique : C'est l'indicateur le plus utilisé pour décrire la distribution d'une variable quantitative.

Si on note par : x_i : la $i^{ème}$ observation, la moyenne arithmétique est définie par :

$$\bar{X} = \frac{1}{n}\sum_i n_i x_i = \sum_i f_i x_i \text{ avec } n = \sum_i n_i.$$

1.3.1.3. La médiane : C'est la valeur du caractère qui partage une distribution statistique en deux sous ensembles de même effectif. Graphiquement, on peut définir la médiane comme étant l'abscisse du point d'intersection de la courbe des fréquences cumulées croissantes et celle des fréquences cumulées décroissantes.

1.3.2. Les indicateurs de dispersion :

Dans la théorie, il existe plusieurs indicateurs de dispersion : l'écart moyen absolu, l'étendue et la variance. Cependant, ce dernier est le plus utilisé dans la pratique. La variance est définie par :

$$\sigma^2 = \frac{1}{n}\sum_i n_i (x_i - \bar{x})^2.$$

σ est appelé écart-type. Il mesure l'écart par rapport à la moyenne.

1.3.3. Les indicateurs de forme :

1.3.3.1. L'asymétrie : Lorsqu'une distribution est symétrique, ses valeurs sont réparties de part et d'autre d'une valeur centrale qui n'est autre que la moyenne $\bar{X}$. Dans ce cas, le mode, la médiane et la moyenne arithmétique sont confondus.

Pour mesurer l'asymétrie d'une distribution, il existe plusieurs indicateurs. Le plus utilisé dans la pratique est celui de Fisher :

$$F = \frac{\frac{1}{n}\sum_i n_i(x_i - \bar{x})^3}{\sigma^3}$$

- Si F = 0, la distribution est symétrique
- Si F > 0, la distribution est orientée à gauche (étalée à gauche)
- Si F < 0, la distribution est orientée à droite (étalée à droite).

1.3.3.2. L'aplatissement : Pour évaluer l'aplatissement d'une distribution, on compare une distribution quelconque avec celle d'une loi normale. Dans la pratique, on utilise le coefficient de Pearson :

$$P = \frac{\frac{1}{n}\sum_i n_i(x_i - \bar{x})^4}{\sigma^4}$$

- Si P = 0, la distribution est normale
- Si F > 0, la distribution est aigue
- Si F < 0, la distribution est aplatie.

2. Tests non paramétriques

2.1. Test de Khi-deux ($\varkappa^2$)

Le test de Khi-deux est utilisé afin de vérifier la conformité d'une distribution observée à une distribution théorique connue. Pour cela deux hypothèses sont possibles : l'hypothèse nulle notée par H_0 et l'hypothèse alternative notée par H_1.

H_0 : la variable observée est très proche de la loi de khi-deux.

H_1 : la variable observée ne suit pas la loi de khi-deux.

La validation passe par l'analyse des écarts entre la répartition observée et la répartition théorique. Si les écarts sont faibles, on peut considérer que la répartition observée est assimilable à la répartition théorique.

La valeur calculée de khi-deux est définie par :

$$\varkappa^2{}_{cal} = \sum_i \frac{(o_i - t_i)^2}{t_i}$$

où o_i : représente l'effectif observé

t_i : représente l'effectif théorique

Dans la pratique, on doit choisir un échantillon de taille supérieur strictement à 30. Dans ce cas, la variable suit la loi de khi-deux à (n-1) degrés de liberté. La règle de décision consiste à calculer la valeur de khi-deux dans une première étape et dans une seconde étape, de la comparer avec la valeur de la table statistique (valeur tabulée) avec un degré de risque égal à α. Si le $\varkappa^2{}_{cal} \geq \varkappa^2{}_{tab}$, on rejette H_0. Dans le cas contraire, on accepte H_0.

Exemple :

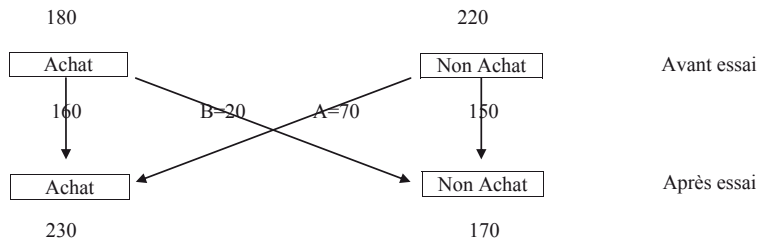
Considérons un échantillon de 1000 personnes interrogées à l'occasion d'une enquête sur la sensibilité à la protection de l'environnement. La répartition est définie par rapport aux différentes professions et catégories sociales comme le montre le tableau ci-dessous.

NB : $t_i = N.f_i$ et $o_i = n.f_i$.

Dans notre cas, le tableau des données comporte 9 modalités. Ce qui signifie que le nombre de degrés de liberté est égal à 8 (9-1).
Pour un risque égal à 5%, le $\varkappa^2{}_{tab}$= 15,3. Or, le tableau monte que $\varkappa^2{}_{cal}$ = 66,2. Ce qui montre que $\varkappa^2{}_{cal} \geq \varkappa^2{}_{tab}$. Autrement dit, on doit rejeter l'hypothèse nulle.
Ce résultat montre que l'écart, entre ce qui est observé et ce qui est théorique, n'est pas dû au hasard mais plutôt au choix de l'échantillon. Autrement dit, l'échantillon choisi n'est pas représentatif.

PCS	Effectif dans la population	Fréquence dans la population	Effec. obs. dans l'échantillon (o_i)	Effectif théorique (t_i)	(o_i - t_i)	$\frac{(o_i - t_i)^2}{t_i}$
Agriculteurs	516000	0,0732	80	73	7	0,6712
Artisans	350000	0,0497	50	50	0	0,0000
Prof.Sup	22000	0,0031	10	3	7	16,3333
Prof.Sec	130000	0,0184	25	18	7	2,7222
pop.active	4000000	0,5676	500	568	-68	8,1408
Médecins	14000	0,0020	10	2	8	32,0000
Retraités	1100000	0,1561	180	156	24	3,6923
Etudiants	315000	0,0447	45	45	0	0,0000
Sans emploi	600000	0,0851	100	85	15	2,6471
Total	7047000	1	1000	1000	0	66,2070

2.2. Test de Mac-Nemar
Ce test traite des oppositions en termes d'appréciation ou de décision. Les variables traitées sont dichotomiques (oui ou non, acheter ou ne pas acheter). Dans ce type de cas, les échantillons avant et après l'appréciation ne sont pas indépendants. On dit qu'ils sont appariés.
Exemple :
Considérons un échantillon de 400 individus testés sur leur intention d'achat concernant un certain produit. Le test se fait en deux étapes :
- Les sujets sont interrogés à priori sur leur intention d'achat. On recueille les résultats puis on fait un test de ce produit.
- Les sujets sont interrogés à posteriori sur leur dernière intention. A l'issue de ce test, on observe les modifications décisionnelles. L'objectif est de vérifier si l'essai a joué de façon favorable ou défavorable sur les résultats.
Les résultats sont représentés schématiquement comme suit :



Nous observons que sur les 180 individus acceptant le principe de l'achat, dans une première étape, seulement 160 adhèrent réellement à ce principe après essai. Autrement dit, les 20 autres individus, déçus par l'essai, rejettent le produit. Ceci nous amène à définir la matrice suivante :

Après essai

	Achat	Non Achat
Non Achat	A=70	150
Achat	160	B=20

La diagonale formée par les éléments A et B représente le changement décisionnel des individus. La question qui se pose est la suivante : Est-ce que le changement est significatif ou non.

Mac-Nemar, défini l'hypothèse nulle pour laquelle on n'observe pas des différences de comportements.

Le test consiste à calculer le rapport suivant : R= $\frac{(|A-B|-1)^2}{A+B}$.

Ce rapport suit approximativement la loi de khi-deux à (n-1)(p-1) degrés de liberté où n : le nombre de lignes et p : le nombre de colonnes.

Dans notre cas R= $\frac{(50-1)^2}{90}$= 26,67 et n=p=2.

Pour un risque α = 5%, la valeur tabulée de khi-deux à 1 degré de liberté est égale à 3,841. Ceci montre que $\varkappa^2_{cal} \geq \varkappa^2_{tab}$ et donc on doit rejeter l'hypothèse nulle. Autrement dit l'essai a été favorable dans le comportement des individus.

2.3. Test de Kolmogorov-Smirnov

Ce test permet de connaitre l'uniformité des différentes réponses à une question donnée.

Exemple : On cherche à savoir si l'utilité procurée pour un certain produit est jugée de façon uniforme pour un échantillon de la population. Pour cela, on va utiliser une échelle. Les résultats obtenus sot donnés par le tableau suivant :

Le nombre des individus	**Réponse des individus**
18	Très peu utile
21	Peu utile
31	Moyennement utile
27	Plutôt utile
21	Très utile

D'après le tableau, on peut remarquer une différence importante entre la proposition très peu utile (18 individus) et la proposition moyennement (31 individus). La question qui se pose est la suivante : est-ce que cet écart est significatif ?

Le test de Kolmogorov-Smirnov présente une réponse. En effet,

Ce test est basé sur deux hypothèses :

H_0 : l'écart entre les modalités est nul.

H_1 : il existe une différence entre les modalités.

Le principe de ce test consiste à calculer les fréquences observées et théoriques croissantes puis de définir l'écart entre ces deux fréquences.

Notons par D_i : le sup de la plus grande valeur des valeurs absolues des écarts entre les fréquences observées f_o et les fréquences théoriques f_t.

$D_i = Sup_i|f_{oi}^c - f_{ti}^c|$. La table de Kolmogorov-Smirnov est utilisée lorsque le nombre d'observations est inférieur ou égal à 100. Si le nombre est strictement supérieur à 100, on utilise la valeur tabulée comme suit : $D_{tab} = \frac{1,36}{\sqrt{n}}$ (pour un risque d'erreur égal à 5%). Si le risque d'erreur est égal à 1% , $D_{tab} = \frac{1,63}{\sqrt{n}}$.

Si $D_{cal} < D_{tab}$, on accepte l'hypothèse nulle sinon, on rejette H_0.

Notons que dans le cas où on a n modalités, la fréquence théorique est égale à $\frac{1}{n}$. En effet, on suppose que tous les événements sont équiprobables.

Dans notre cas, l'échantillon comporte 118 individus et par conséquent $D_{tab} = \frac{1,36}{\sqrt{118}} = 0,1252$.

Or, $D = Sup_i|f_{oi}^c - f_{ti}^c| = 0,0696$, ce qui montre que $D_{cal} < D_{tab}$ et donc il faut accepter l'hypothèse nulle. Autrement dit, il y a uniformité au niveau des réponses.

Modalités	Effectif (ni)	f_{oi}	f_{ti}	f_{oi}^c	f_{ti}^c	$\lvert f_{oi}^c - f_{ti}^c \rvert$
M_1	18	0,1525	0,2	0,1525	0,2	0,0475
M_2	21	0,1779	0,2	0,3304	0,4	0,0696
M_3	31	0,2627	0,2	0,5931	0,6	0,0069
M_4	27	0,2288	0,2	0,8219	0,8	0,0219
M_5	21	0,1779	0,2	1	1	0
Total	118					

2.4. Test de Spearman (test de rang)

Ce test est basé sur le principe de corrélation linéaire. Dans ce type de test, on utilise les variables ordinales au lieu des variables cardinales utilisées dans la détermination du coefficient de corrélation linéaire (ρ) défini par : $\rho = \frac{cov(x,y)}{\sqrt{v(x).v(y)}}$.

Exemple : On veut étudier la relation entre deux caractéristiques d'un produit : la fiabilité et l'image de marque. L'échantillon étudié comporte 591 personnes. Les notes observées et les effectifs par variable sont définis comme suit :

Notes	1	2	3	4	5	6	7	8	9	10	Total
Fiabilité (x)	20	72	85	100	90	87	85	33	12	7	591
Rx	3	5	6	10	9	8	6	4	2	1	
Image de marque (y)	35	50	71	89	110	120	23	70	23	0	591
Ry	4	5	7	8	9	10	2	6	2	1	
$D=\lvert Rx - Ry \rvert$	1	0	1	2	0	2	4	2	0	0	

La question qui se pose est la suivante : existe-t-il une relation entre la fiabilité et l'image de marque.

- Si on considère que les variables sont cardinales alors on peut utiliser le coefficient de corrélation linéaire $\rho = \frac{cov(x,y)}{\sqrt{v(x).v(y)}} = 0,7018$.
- On peut répondre à cette même question en utilisant un classement des différentes catégories. On donne ainsi, un rang à chaque note en fonction du nombre de personnes représentant chaque catégorie et ce, dans un ordre croissant.

Dans ce cas, le coefficient de corrélation linéaire, appelé coefficient de Spearman, est défini comme suit : $R_s = \frac{cov(Rx,Ry)}{\sqrt{v(Rx).v(Ry)}} = 0,7018$.

Spearman montre que ce coefficient peut être défini de la façon suivante :

$Rs = 1 - \frac{6\sum {d_i}^2}{k^3 - k}$ avec k = le nombre de modalités.

Dans notre cas le coefficient de corrélation de Spearman est égal à 0,7018 montrant ainsi qu'il ya une corrélation entre la fiabilité et l'image de marque.

2.5. Test de Wilcoxon

Le test de wilcoxon est un test de rang qui permet de comparer des notations sur une échelle avant et après essai d'un certain produit.

Exemple : Un enquêteur veut observer la perception d'un produit nouveau avant et après essai. Pour cela, il interroge un groupe d'individus en utilisant une échelle à 7 positions allant d'une perception très défavorable (notée par 1) jusqu'à une perception très favorable (notée par 7). Les résultats sont représentés dans le tableau suivant :

Consommateur	Avant essai	Après essai	d	$\|d\|$	rang	Rang moyen	Rang de d<0	Rang de d>0
A	2	3	1	1	1	1,5		1,5
B	6	7	1	1	2	1,5		1,5
C	4	6	2	2	3	5,5		5,5
D	5	1	-4	4	9	9	-9	
E	3	5	2	2	4	5,5		5,5
F	5	3	-2	2	5	5,5	-5,5	
G	3	5	2	2	6	5,5		5,5
H	2	4	2	2	7	5,5		5,5
I	1	3	2	2	8	5,5		5,5
J	4	4	0	0				

Peut-on dire que l'essai du produit a un effet favorable ou non sur les intentions d'achat.

Ce test passe par les étapes suivantes :

- on calcule les différences entre les observations après et avant essai (d=après-avant)
- on présente les différences sans tenir compte de leur signe ($|d|$)
- on affecte les rangs aux valeurs de $|d|$ obtenues
- on affecte à ces rangs le signe de la différence correspondante (Rang de d<0) et (Rang de d>0).

L'objectif de ce test est de définir la valeur de Γ qui représente la somme des rangs la plus petite en valeur absolue. Wilcoxon montre que Γ suit la loi normale de moyenne $\bar{X}$ et d'écart-type σ avec $\bar{X} = \frac{n(n+1)}{4}$ et $\sigma = \sqrt{\frac{n(n+1)(2n+1)}{24}}$ où n représente le nombre de différences non nulles.

Dans notre exemple, Γ= 14,5, n=9, $\bar{X} = \frac{9(9+1)}{4} = 22,5$ et $\sigma = \sqrt{\frac{9(9+1)(18+1)}{24}} = 8,44$.

En posant $T = \frac{\Gamma - \bar{X}}{\sigma}$ alors T suit la loi normale centrée réduite.

Le test de wilcoxon est utilisé pour vérifier les hypothèses suivantes :

H_0 : il n'y a pas de différence entre les deux sommes des rangs. Autrement dit l'essai du produit n'a rien changé.

H_1 : on observe une différence entre les deux sommes

Revenons à notre exemple : $T = \frac{14,5-22,5}{8,44} = -0,9478$. Si on utilise la table statistique de la loi normale centrée réduit au seuil de 5%, la valeur de t = 1,96. Or, -0,9478 € [-1,96, 1,96] et par conséquent on doit accepter l'hypothèse nulle. Autrement dit, il n'y a pas de différence entre les deux sommes.

Exercices

Exercice 1 :

L'étude de 320 familles ayant 5 enfants s'est traduite par la distribution suivante :

Classe	A	B	C	D	E	F
Nombre de filles	5	4	3	2	1	0
Nombre de garçons	0	1	2	3	4	5
Nombre de familles observées	8	40	88	110	56	18
Nombre de familles théoriques	10	50	100	100	50	10

On veut comparer cette distribution à la distribution théorique qui correspond à l'équiprobabilité de la naissance d'un garçon et de la naissance d'une fille.
1/ Calculer le khi-deux. Sachant que, pour $\alpha = 0.05$, le $\chi^2_{tab} = 11.07$, que peut-on en conclure ?
2/ En utilisant le test de Kolmogorov-Smirnov ($\alpha = 0.05$), que peut-on en conclure ?
3/ Comparer les deux résultats. Quelle est votre conclusion.

Exercice 2 :

Considérons un échantillon de 500 individus testés quant à leurs intentions d'achat concernant un produit. Sur 230 personnes acceptant dans un premier temps le principe de l'achat (avant essai), seulement 180 adhèrent réellement au produit après essai. Les 50 personnes restantes, déçues par l'essai rejettent le produit. Par contre sur les 270 personnes refusant à priori l'achat, 100 de ces personnes changent d'avis après essai.
1/ Déterminer le tableau représentant cette relation croisée (avant et après essai) et prenant en compte les modifications de comportement des acteurs, face à la décision d'achat ou de rejet.
2/ Formuler l'hypothèse nulle.
3/ Sachant que la valeur tabulée de khi-deux, pour $\alpha = 0{,}05$, est de 3,841, donner votre conclusion.

Exercice 3 :

L'étude de 320 familles ayant 5 enfants s'est traduite par la distribution suivante :

Classe	A	B	C	D	E	F
Nombre de filles	5	4	3	2	1	0
Nombre de garçons	0	1	2	3	4	5
Nombre de familles observées	8	40	88	110	56	18
Nombre de familles théoriques	10	50	100	100	50	10

On veut comparer cette distribution à la distribution théorique qui correspond à l'équiprobabilité de la naissance d'un garçon et de la naissance d'une fille.
1/ Calculer le khi-deux. Sachant que, pour $\alpha = 0.05$, le $\chi^2_{tab} = 11.07$, que peut-on en conclure ?
2/ En utilisant le test de Kolmogorov-Smirnov ($\alpha = 0.05$), que peut-on en conclure ?
3/ Comparer les deux résultats. Quelle est votre conclusion.

Exercice 4 :

On veut tester une campagne de prévention chez la femme enceinte vis à vis du tabac. On mesure le caractère fumeur au 3ème et 8ème mois sur un échantillon de 100 femmes. On obtient les résultats suivants :

3ème mois	**8ème mois**	
négatif	négatif	35
négatif	positif	5
positif	négatif	15
positif	positif	45

L'objectif est de tester l'hypothèse nulle suivante : **il y a autant de femmes qui ont arrêté de fumer que de femmes qui se sont mises à fumer**.
Sachant que la valeur de khi-deux, pour $\alpha = 5\%$, est de 3.841, donner votre conclusion.

Analyse bivariée

Introduction

Dans les chapitres précédents, nous avions utilisé des séries statistiques à une seule variable. Dans le cas où le problème étudié comporte deux variables, on parle d'une série statistique bivariée. Dans ce cas, on va manipuler des couples de la forme (x_i, y_j) avec i = 1, ….., p et j = 1,……., q.

Exemple : Une enquête réalisée auprès de 100 individus travaillant dans une entreprise afin de définir des profils de poste. Parmi les questions posées, on trouve celles dont l'objectif est de mesurer les variables suivantes :

- Le sexe de la personne interrogée
- L'occupation d'un poste de direction (si oui, on note par 1 et si non, on note par 2)
- L'âge de la personne
- Le revenu annuel de la personne

Le département des ressources humaines suppose un certain nombre de questions :

- Y a-t-il une prédominance de poste de direction confié à des hommes ?
- Pour un même âge et un même niveau, les salaires sont-ils réellement égaux entre femmes et hommes ?
- L'âge est-il un facteur fondamental dans le niveau des revenus.

1. Cas de deux variables nominales

Si on s'intéresse au cas des deux premières variables, on peut définir un tableau de deux variables : la première représente le sexe et la seconde représente la prédominance du poste. Dans ce cas les couples (x_i,y_j) ne pouvant prendre que 4 modalités : (F,1), (F,0), (H,1) et (H,0). Or, dans la pratique, ces modalités se présentent plusieurs fois. Par conséquent, on doit trier ces données de façon à rassembler tous les individus qui ont les mêmes réponses. On obtient ainsi un tableau appelé tableau de contingence.

1.1. Tableau de contingence : Soit (x_i,y_j) les modalités de deux variables nominales et soit n_{ij} l'effectif ou le nombre de fois que le couple (x_i,y_j) a été observé. Ces éléments peuvent être rassemblés dans un tableau défini comme suit :

	y_1	y_2	.	y_j	.	y_q
x_1	n_{11}	n_{ij}				n_{1q}
x_2	n_{21}					n_{2q}
.						
x_i				n_{ij}		
.						
x_p	n_{p1}					n_{pq}

Exemple :

	1	0	
F	2	30	32
H	8	60	68
	10	90	100

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

$n_{i.}$ est appelé effectif marginal par rapport à la ligne i.

$n_{.j}$ est appelé effectif marginal par rapport à la colonne j.

Interprétation des résultats : Dans le cas de notre exemple, la valeur 32 de la première ligne peut être interprétée comme étant l'effectif des femmes et la valeur 10 de la première colonne représente l'effectif des individus qui occupent un poste de direction.
De même, on peut introduire la notion de fréquence marginale définie comme suit :
$\frac{n_{i.}}{n}$: est appelé fréquence marginale par rapport à la ligne i.
$\frac{n_{.j}}{n}$: est appelé fréquence marginale par rapport à la ligne i.

1.2. Fréquence conditionnelle et profil : L'exemple ci-dessus montre qu'il est difficile de comparer la répartition des postes de direction parmi les femmes et les hommes et ce, résulte du fait que le nombre d'individus dans chacune des catégories n'est pas le même. C'est la raison pour laquelle nous devons introduire les fréquences conditionnelles. Les distributions associées à ces fréquences sont appelées profils. On distingue les profils lignes et les profils colonnes.

1.2.1. Profil ligne :

	1	0	
F	2/32	30/32	1
H	8/68	60/68	1

La valeur 2/32 représente le pourcentage du nombre des femmes qui occupent un poste de direction, 60/68 représente le pourcentage du nombre des hommes qui n'occupent pas un poste de direction.

1.2.2 Profil colonne :

	1	0
F	2/10	30/90
H	8/10	60/90
	1	1

La valeur 2/10 représente le pourcentage des femmes parmi ceux qui occupent un poste de direction.

2. Cas de deux variables ordinales

Dans le cas des variables ordinales, les modalités (x_i,y_j) sont qualitatives mais peuvent être ordonner. Dans ce cas, il suffit de définir une série ordonnée sur laquelle on élabore des mesures d'association ou de liaison. Le test le plus approprié est celui de Spearman.

3. Cas de deux variables quantitatives

Lorsque la série bivariée est formée par des variables quantitatives, on peut utiliser une représentation graphique des couples (x_i,y_j). Par la suite, on définie la droite de régression linéaire : $Y = \hat{a}X + \hat{b}$ ou bien on utilise le coefficient de corrélation linéaire $\rho = \frac{cov(x,y)}{\sqrt{v(x).v(y)}}$.

4. Test d'indépendance de deux variables

4.1. Cas de deux variables quantitatives : Supposons qu'on nous demande de tester s'il y a ou non une corrélation entre les variables X et Y. La procédure consiste à définir les deux hypothèses suivantes :
H_0 : X et Y sont indépendantes
H_1 : X et Y sont dépendantes.

Le principe consiste à calculer le rapport défini par : $\frac{\rho\sqrt{n-2}}{\sqrt{1-\rho^2}}$ où ρ est le coefficient de corrélation linéaire et n est le nombre d'observations. Ce rapport suit la loi de Student à n-2 degrés de liberté.

4.2. Cas de deux variables nominales : Le problème qui se pose est toujours le même à savoir de tester s'il y a ou non une corrélation entre les variables nominales. Le test utilisé dans ce genre de cas est celui de khi-deux.

Notons que $\varkappa^2{}_{cal} = \sum_{i=1}^{p}\sum_{j=1}^{q}\frac{\left(n_{ij(th)}-n_{ij(obs)}\right)^2}{n_{ij(th)}}$.

Pour un niveau de risque α, on doit comparer la valeur ci-dessus avec la valeur de khi-deux tabulée à (p-1)(q-1) degrés de liberté.

Analyse en composantes principales

Introduction

Le principe de cette méthode consiste à décrire les données contenues dans un tableau d'individus et de caractères (ou variables). Ce tableau est appelé matrice des données. Il se compose de lignes représentant les individus (personnes, entreprises, journaux,..) et de colonnes formées par les variables (âge, taille,..). Pour obtenir une meilleure représentation des données, on prend les premières composantes principales données par les deux meilleures valeurs propres en terme de pourcentage.

L'analyse en composantes principales cherche à représenter dans un espace de dimension faible (deux ou trois axes) un nuage de points représentant les n individus ou objets décrits par les p variables numériques et ce, par l'utilisation des corrélations entre les différentes variables. L'analyse des données étudie les proximités entre les variables qui sont, dans ce chapitre, des variables quantitatives.

1. Nature des données étudiées

L'ACP s'intéresse principalement aux données quantitatives. L'objectif principal est de former des groupes homogènes d'individus ou d'unités statistiques et d'analyser les liaisons entre les différentes variables.

Ces données sont constituées d'observations de plusieurs grandeurs notées par : X_1, X_2, ……X_p sur un ensemble d'unités statistiques ou d'individus de taille n. Dans la pratique, ces données sont représentées dans un tableau où les lignes représentent les individus et les colonnes correspondent aux variables.

	X_1	X_2	X_3	……	X_j	……	X_p
1	$X_1(1)$	$X_2(1)$					$X_p(1)$
2	$X_1(2)$	$X_2(2)$					$X_p(2)$
3							
.							
i					$X_j(i)$		
.							
n	$X_1(n)$	$X_2(n)$					$X_p(n)$

$X_j(i)$ représente l'observation de la variable X_j sur l'individu ou l'unité statistique i.

Exemple : On a recueilli le poids, la taille, l'âge et la note moyenne de 10 élèves. Les données se présentent comme suit :

	Poids	Taille	Age	Note
1	45	1.50	13	14
2	50	1.60	13	16
3	50	1,65	13	15
4	60	1,75	15	9
5	60	1,70	14	10
6	60	1,70	14	7
7	70	1 ,60	14	8
8	65	1,60	13	13
9	60	1.55	15	17
10	65	1,70	14	11

Tableau 1

Les données traitées par l'ACP doivent être quantitatives pour que la notion de la moyenne ait un sens. Les variables quantitatives peuvent être homogènes (même unité de mesure) ou hétérogènes. Dans notre cas, le tableau représente des variables hétérogènes.
L'ACP ne donnera des résultats intéressants que sur des tableaux suffisamment grands (le nombre d'unités statistiques > 20).
Remarque : On peut introduire dans un tableau des données appelées éléments supplémentaires pour faciliter les interprétations des résultats. On peut, à titre d'exemple, introduire le centre de gravité, les différentes moyennes de variables dans un groupe bien déterminé (taille moyenne, poids moyen, note moyenne,…).
On peut aussi introduire d'autres éléments comme le sexe des individus ou au sein d'une même famille.

2. Adéquation des données : Avant de réaliser l'analyse, il faut s'assurer que les données forment un ensemble cohérent pour qu'on puisse déterminer des dimensions communes. La matrice des données doit comporter suffisamment de corrélations pour justifier la réalisation d'une ACP.
Certains indicateurs peuvent être utilisés comme le test de Bartlett, la « Measure of Sampling Adequacy (MSA) » ou Kaiser-Meyer-Olkin (KMO).

2.1. Test de Bartlett: ce test examine la matrice des corrélations dans son intégralité et fournit la probabilité de l'hypothèse nulle selon laquelle toutes les corrélations sont nulles.

2.2. MSA ou KMO: Cette mesure indique dans quelles proportion les variables retenues forment un ensemble cohérent et mesurent de manière adéquate un concept.

Des valeurs de KMO en dessous de 0,5 sont inacceptables, médiocres entre 0,5 et 0,6, moyennes entre 0,6 et 0,7, bonnes entre 0,7 et 0,8, très bonnes entre 0,8 et 0,9 et excellentes au-delà de 0,9.

3. Présentation de la méthode : L'interprétation des variables à travers l'ACP est basée sur le principe de la notion de distance ou de proximité. En effet, une distance très petite signifie une ressemblance entre les individus ou les variables.

3.1. Notion de distance entre deux unités statistiques : Considérons les individus 4, 5 et 6 du tableau 1.
- $d^2(4,5) = 2{,}0025$, $d^2(4,6) = 5$ et $d^2(5,6) = 9$. Ces résultats montrent que l'élève 6 est plus proche, en terme de ressemblance, de l'élève 4 que l'élève 5 ($5 < 9$).
- Reprenons les mêmes distances mais au lieu d'utiliser, comme unité, le mètre pour la taille, on va utiliser le cm. Dans ce cas on obtient :
$d^2(4,5) = 27$, $d^2(4,6) = 30$ et $d^2(5,6) = 9$. Ces résultats montrent que l'élève 6 est plus proche, en terme de ressemblance, de l'élève 5 que l'élève 4 ($9 < 30$).
Cette contradiction résulte du changement d'unité. En effet, en changeant l'unité du mètre au cm, l'analyste n'arrive plus à conclure si l'élève 6 est plus proche de l'élève 4 ou 5. Autrement dit, la notion de distance dépend de l'unité choisie. Pour résoudre ce genre de problème, on fait appel aux variables centrées réduites. En effet, ces dernières sont indépendantes de l'unité choisie. Il suffit donc de poser :
$X'_j(i) = \frac{X_j(i) - \mu_j}{\sigma_j}$ où μ_j représente la moyenne des variables X_j et σ_j est l'écart-type.
Les moyennes et les variances sont définies par le tableau suivant :

Variables	Moyenne	Ecart-type	Variance
Poids	58.5	7.433	55.25
Taille	1.635	0.074	0.005
Age	13.8	0.748	0.56
Note	12	3.316	11

Tableau 2

3.2. Description de la méthode : Dans l'exemple ci-dessus, lorsqu'on s'intéresse aux individus deux à deux, le nombre de distances à calculer est égal à 45 (10.9/2). En effet, lorsqu'on a n individus, on doit calculer $\frac{n(n-1)}{2}$ distances.

L'ACP fournit un système d'axes orthonormés conservant l'ensemble de ses distances. Les axes possédant des propriétés supplémentaires représentent les droites les plus proches des observations suivant le critère des moindres carrées.

Chaque individu ou unité statistique possède des coordonnées sur les axes. Puisque le système est orthonormé, le carré de la distance entre deux individus est donné par la somme des carrées des différences des coordonnées.

3.3. Définition des axes principaux : Les droites les plus proches des unités statistiques sont appelées axes principaux. Leurs vecteurs directeurs sont appelés vecteurs principaux.

3.4. Définition d'une composante principale : On appelle composante principale, la liste des coordonnées des unités statistiques sur l'axe principal engendré par le vecteur principal associé.

Application : Reprenons l'exemple ci-dessus

	Axe 1	Axe 2	Axe 3	Axe 4
.	.	.	.	.
.	.	.	.	.
4	2.083	0.078	1.201	0.192
5	0.987	-0.420	0.296	-0.053
6	1.474	-0.816	0.061	0.555
.	.	.	.	.

Tableau 3

	Axe 1	Axe 1,2	Axe 1,2,3	Axe1,2,3, 4
.	.	.	.	.
.	.	.	.	.
$d^2(4,5)$	1.201	1.449	2.268	2.328
$d^2(4,6)$	0.371	1.170	2.470	2.601
$d^2(5,6)$	0.237	0.394	0.449	0.819
.	.	.	.	.

Tableau 4

Ces résultats montrent que les distances sont mieux reconstruites que le nombre d'axes considérés est suffisamment grand. D'après le tableau 4, on remarque que le dernier axe modifie peu les carrées des distances. Autrement dit, l'analyste peut se contenter des trois premiers axes pour analyser les proximités entre les individus.

L'élément essentiel dans l'analyse des données est la composante principale. En effet, les composantes principales des unités statistiques permettent d'expliquer les relations entre les différentes variables et de justifier la formation des groupes homogènes. Pour cela, on utilise

les coefficients de corrélations linéaires entre les composantes principales et les variables (voir tableau 5).

	Axe 1	Axe 2	Axe 3	Axe 4
Poids	0.785			
Taille				
Age				0.150
Note	-0.832			

Tableau 5

On remarque que le poids est corrélé positivement avec l'axe 1 contrairement à la note qui corrélée négativement. D'autre part, on remarque une absence de corrélation entre l'âge et l'axe 4.

3.5. Définition de la valeur propre : La variance d'une composante principale est appelée valeur propre ou inertie expliquée par l'axe de même rang. Dans la pratique, les valeurs propres sont ordonnées dans l'ordre décroissant. La somme des valeurs propres est égale au nombre des variables.

- L'inertie expliquée par rapport à chaque valeur propre est définie par le rapport $\frac{\alpha_l}{\sum_{l=1}^{n} \alpha_l}$ où α_l est la valeur propre relative à l'axe n.
- A chaque valeur propre α_l est associé un vecteur propre. Dans ce cas, la composante principale d'ordre l par rapport à l'individu i est définie par :

$C_l(i) = \sum_{j=1}^{p} X'_j(i).U_l^j$. Inversement, on peut définir $X'_j(i) = \sum_{l=1}^{k} C_l(i).U_l^j$ où l représente le nombre d'axes.

Application : Revenons à l'exemple ci-dessus.

- Les 4 valeurs propres son définies comme suit :

$\alpha_l = 2{,}391$, $\alpha_2 = 0{,}750$, $\alpha_3 = 0{,}584$ et $\alpha_4 = 0{,}274$.

- Les vecteurs propres :

U_l = (0,508 ; 0,503 ; 0,445 ; -0,538) , U_2 = (0,306 ; -0,464 ; 0,705 ; 0,438)

U_3 = et U_4 =

	Variable $X_j(i)$	Moyenne	Ecart-type	Variable $X'_j(i)$
Poids	$X_1(1) = 45$	$\mu_1 = 58{,}5$	$\sigma_1 = 7{,}43$	$X'_1(1) = -1{,}816$
Taille	$X_2(1) = 1{,}50$	$\mu_2 = 1{,}635$	$\sigma_2 = 0{,}07$	$X'_2(1) = -1{,}816$
Age	$X_3(1) = 13$	$\mu_3 = 13{,}8$	$\sigma_3 = 0{,}74$	$X'_3(1) = -1{,}069$
Note	$X_4(1) = 14$	$\mu_4 = 12$	$\sigma_4 = 3{,}316$	$X'_4(1) = 0{,}603$

Tableau 6

$C_1(1) = \sum_{j=1}^{4} X'_j(1).U_1^j = -2{,}638$

$C_2(1) = \sum_{j=1}^{4} X'_j(1).U_2^j =$

$C_3(1) = \sum_{j=1}^{4} X'_j(1).U_3^j =$

$C_4(1) = \sum_{j=1}^{4} X'_j(1).U_4^j =$

4. Interprétation des résultats :

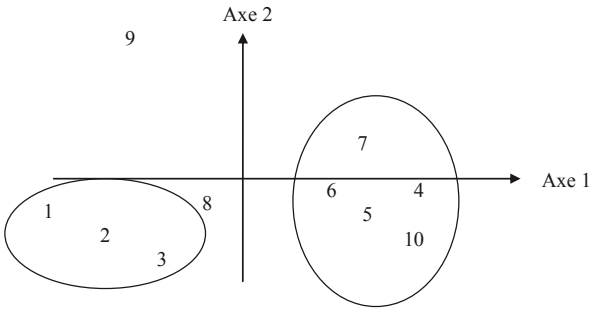
Avant l'interprétation des résultats, il faut s'assurer de la cohérence interne de l'instrument de mesure afin de réduire l'ensemble des termes d'erreur. Pour cela, nous utilisons le concept de fiabilité qui se mesure à travers le coefficient alpha de CRONBACH. Cette valeur mesure la cohérence interne d'une échelle construite à partir d'un ensemble d'items. Plus la valeur de l'alpha est proche de 1, plus la cohérence interne de l'échelle (sa fiabilité) est forte. On

élimine donc les items qui diminuent le score et on conserve ceux qui contribuent à augmenter l'alpha. Pour une étude exploratoire, un coefficient proche de 0,7 est acceptable alors que dans le cadre d'une recherche fondamentale, ce coefficient doit être supérieur à 0,8. Les résultats peuvent être résumés dans le tableau suivant :

$\alpha < 0{,}6$	Insuffisant
$0{,}6 \leq \alpha < 0{,}65$	Faible
$0{,}65 \leq \alpha < 0{,}7$	Acceptable
$0{,}7 \leq \alpha < 0{,}8$	Bon
$0{,}8 \leq \alpha < 0{,}9$	Très bon
$> 0{,}9$	Réduction du nombre d'items

4.1. Représentation graphique : Dans l'ACP, deux représentations graphiques sont possibles :
- <u>On représente les unités statistiques</u> dans un plan formé par les deux axes principaux. Les coordonnées de l'unité statistique « i » définies par les composantes principales $C_l(i)$ seront définies par rapport à l'axe l. L'origine des axes principaux est caractérisée par l'unité statistique dont les valeurs sont les moyennes des variables initiales.
- <u>On représente les variables initiales</u> à l'aide d'un cercle de corrélation. Les coordonnées d'une variable sont les coefficients de corrélations.

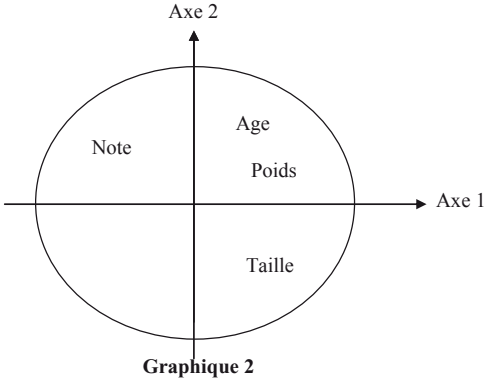
Application 1 : Plan principal 1x2



Graphique 1

Le graphique 1 montre qu'on peut distinguer un premier groupe homogène formé par les individus 1,2 et 3 et un deuxième groupe formé par les individus 7,4,10,5 et 6. On remarque aussi que les individus 9 et 8 sont isolés par rapport aux autres. Le graphique montre que l'individu 6 est plus proche de l'individu 5 que le 4.
Notons que les unités statistiques les plus éloignées sur les axes sont les mieux reconstituées (tableau 4). Inversement, les unités se trouvant proches de l'origine sont les mal reconstruites.

Application 2 : Cercle de corrélation



Graphique 2

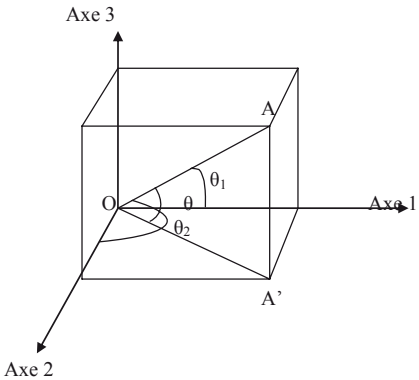
Le cercle de corrélation montre que l'âge, le poids et la taille sont corrélés positivement avec le premier axe principal. Inversement, la note est corrélée négativement avec cet axe. Par rapport à l'axe 2, la taille est corrélée négativement contrairement à la note, l'âge et le poids qui sont corrélés positivement.
L'analyse des deux graphiques 1 et 2 montre que le premier groupe formé par les individus 1,2 et 3 représente les individus ayant des fortes moyennes contrairement au deuxième groupe formé par les individus 7, 4, 10, 5 et 6 mais d'un autre coté, on peut dire que les individus du premier groupe sont peu développés physiquement. L'individu 9 qui est un peu isolé par rapport aux autres, a une bonne note puisque cette variable est corrélée positivement par rapport à la deuxième composante mais il a la caractéristique d'un individu âgé et un peu gros.

4.2. Paramètres d'aide à l'interprétation : En général, il n'y a pas de règle précise pour limiter le nombre d'axes considérés. Le raisonnement se fait selon les résultats empiriques obtenus. Le premier paramètre qu'on utilise est relatif au choix des valeurs propres. Après avoir ordonné ces valeurs dans un ordre décroissant, on calcule l'inertie expliquée par rapport à chaque valeur. Par la suite, on détermine l'inertie expliquée cumulée.
Application :

Valeurs propres	% d'inertie expliquée	% d'inertie cumulée
$\alpha_1 = 2{,}391$	60	60
$\alpha_2 = 0{,}750$	19	79
$\alpha_3 = 0{,}584$	14	93
$\alpha_4 = 0{,}274$	7	100

Tableau 7

La première valeur propre montre qu'on peut reconstruire 60% de la somme des carrées des distances. L'utilisation d'un plan principal 1x2 montre qu'on peut reconstruire 79% la somme des carrées des distances. Toutefois plusieurs plans sont envisageables : plan 1x2, plan 1x3, plan 2x3,……
La question qui se pose est lequel choisir ?
Pour résoudre ce genre de problème, on peut étudier les projections des unités statistiques sur les différents plans et de définir l'angle de projection par rapport à chaque axe.



Parfois l'unité statistique n'est pas bien représentée dans un plan principal. Pour faire face à ce problème, on doit tenir compte de l'angle que fait la demi-droite (OA) avec les différents axes (A représente une unité statistique).
Notons par θ_1 l'angle que fait la demi-droite (OA) avec l'axe 1, θ_2 l'angle que fait la demi-droite (OA) avec l'axe 2 et θ l'angle que fait la demi-droite (OA) avec (OA') où A' est la projection de A sur le plan principal 1x2.
On montre que $\cos^2\theta = \cos^2\theta_1 + \cos^2\theta_2$.
Les $\cos^2$ peuvent être utilisés pour analyser la position d'une unité statistique par rapport à un axe principal. Une unité statistique très bien représentée dans un plan principal lorsque le $\cos^2$ est très proche de 1. Autrement dit l'angle θ est très proche de zéro.
Notons que dans le cas des variables initiales, on utilise le carré des coefficients de corrélation et la démarche es analogue à celle utilisée dans le cas des unités statistiques.

Exercices

Exercice 1 :

Une enquête a été réalisée sur les indicateurs économiques en France. Elle propose 7 variables pour caractériser les 17 régions françaises. Pour ces 17 régions, une analyse en composante principale a été réalisée. Le tableau des données et les résultats sont présentés dans **l'annexe**.

1/ Est-il indispensable d'effectuer l'ACP sur les données centrées réduites ? Justifier votre réponse.

2/ En examinant les valeurs propres, déterminer le nombre d'axes pouvant contenir une information pertinente. Justifier votre réponse.

3/ Définir les variables les plus corrélées entre elles.

4/ Représenter le cercle de corrélations des variables avec les composantes principales. En déduire les variables les mieux représentées sur le cercle.

5/ Quelles sont les régions mal représentées sur le plan principal 1x2?

6/ Commenter cette ACP.

Annexe

Tableau des données

Région	Population	PopActiv	Superficie	Entreprises	Brevets	TauxChô	LignTéléph
Alsace	1 624,00	39,14	8 280,00	35 976,00	241,00	5,20	700,00
Aquitaine	2 795,00	36,62	41 308,00	85 351,00	256,00	10,20	1 300,00
Auvergne	1 320,00	37,48	26 013,00	40 494,00	129,00	9,30	600,00
Normandie	1 390,00	38,63	17 589,00	35 888,00	91,00	9,00	600,00
Bourgogne	1 600,00	38,26	31 582,00	40 714,00	223,00	8,10	750,00
Bretagne	2 795,00	36,62	27 208,00	73 763,00	296,00	9,50	1 300,00
Limousin	720,00	38,06	16 942,00	21 721,00	73,00	7,90	350,00
Lorraine	2 300,00	34,34	23 547,00	48 353,00	185,00	8,60	950,00
Pyrénnées	2 430,00	37,14	45 348,00	78 771,00	237,00	9,00	1 100,00
PasDeCalais	3 960,00	32,05	12 414,00	78 504,00	278,00	12,60	1 600,00
Picardie	1 810,00	34,39	19 399,00	36 285,00	139,00	9,80	750,00
CôteD'Azur	4 260,00	34,96	31 400,00	132 552,00	610,00	11,00	2 300,00
RhôneAlpes	5 350,00	39,44	48 698,00	159 634,00	1 474,00	7,40	2 500,00
Ardenne	1 340,00	37,85	25 606,00	24 060,00	155,00	9,30	550,00
Rousillon	2 110,00	32,12	27 376,00	62 202,00	179,00	13,20	1 000,00
Charentes	1 590,00	36,82	25 809,00	44 592,00	133,00	10,10	750,00

Matrice de Corrélation

	Popula	PopActiv	Superfi	NbreEntre	NbreBrev	TauxChôm	LignTélé
Population	1,000	-0,215	0,483	0,952	0,837	0,224	0,984
PopActive	-0,215	1,000	0,199	-0,075	0,226	-0,832	-0,173
Superficie	0,483	0,199	1,000	0,652	0,572	0,041	0,526
NbreEntrepri	0,952	-0,075	0,652	1,000	0,868	0,169	0,979
NbreBrevets	0,837	0,226	0,572	0,868	1,000	-0,173	0,844
TauxChôm	0,224	-0,832	0,041	0,169	-0,173	1,000	0,216
LignesTéléph	0,984	-0,173	0,526	0,979	0,844	0,216	1,000

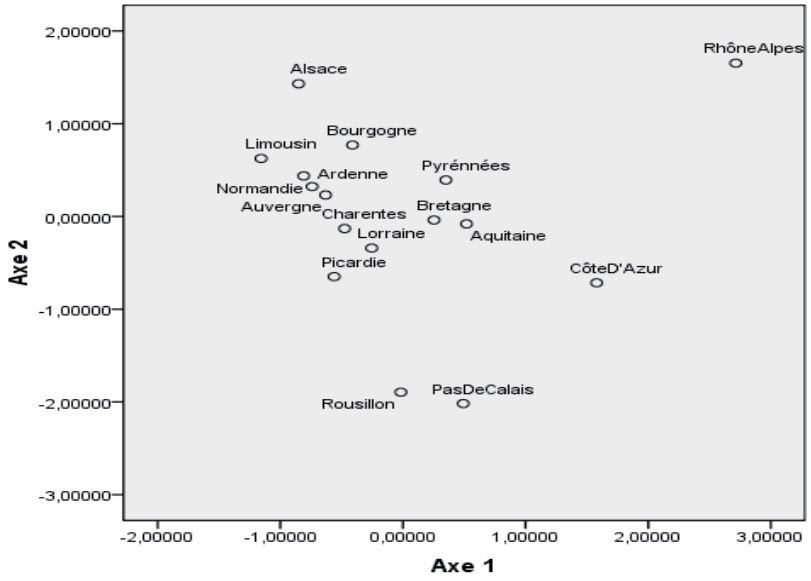
Matrice des composantes

Variables	Axe 1	Axe 2
Population	0,962	-0,127
PopulationActive	-0,081	0,963
Superficie	0,674	0,235
NbreEntreprises	0,991	-0,004
NbreBrevets	0,897	0,330
TauxChômage	0,169	-0,925
LignesTéléphoniques	0,976	-0,096

Variance totale expliquée

Composante	Valeurs propres
1	4,153
2	1,971
3	0,633
4	0,126
5	0,089
6	0,023
7	0,005

Carte des régions



Exercice 2 :

Une enquête a été réalisée sur certains indicateurs économiques dans 15 pays de la CEE. Elle propose les 8 variables suivantes :

- **Mariage** : nombre de mariages pour 1000 habitants
- **Divorce** : nombre de divorces pour 1000 habitants
- **Esp.Vie** : espérance de vie à la naissance pour les femmes
- **Mort.Inf** : mortalité infantile pour 1000 enfants nés vivants
- **Médecin** : nombre de médecins pour 1000 habitants

- **Chômage** : taux de chômage en % (moyenne annuelle)
- **Act.Fem** : taux d'activité professionnelle des femmes en %
- **Temp** : température moyenne estivale (avr-sept).

Une analyse en composante principale a été réalisée. Le tableau des données et les résultats sont présentés dans **l'annexe**.

1/ En examinant les valeurs propres, déterminer le nombre d'axes pouvant contenir une information pertinente. Justifier votre réponse.

2/ Représenter le cercle de corrélations des variables avec les composantes principales 1 et 2. En déduire les variables les mieux représentées sur le cercle.

3/ Quelles sont les pays mal représentées sur le plan principal 1x2?

4/ Commenter cette ACP.

Annexe

Tableau des données

Pays	Mariage	Divorce	Esp.Vie	Mort.Inf	Médecin	Chômage	Act.Fem	Temp
Belgique	5,1	2,2	79,8	7,6	3,5	10,2	10,2	16
Danemark	6,8	2,6	77,8	5,5	2,6	6,7	58,3	13
Allemagne	5,4	1,9	79,6	5,6	3,5	8,3	47,8	17,4
Grèce	5,7	0,7	79,9	8,3	4,3	8,9	3,5	23,2
Espagne	5	0,7	80,9	7,2	3,8	22,7	3,5	23,1
France	4,4	1,9	81,8	6,4	2,6	11,5	47,9	20,2
Irlande	4,6	0	77,9	5,9	1,4	14,4	38,6	13,4
Italie	5	0,4	81,2	6,5	1,1	11,9	33,7	22,8
Luxembourg	5,8	1,7	79,1	5,3	2,1	3,9	38,4	14,9
Hollande	5,4	2,4	80	5,6	2,5	7	47,7	15,2
Autriche	5,4	2,2	79,7	6,3	3,6	8,3	5	17,2
Portugal	6,7	1,2	78,2	8,7	2,9	7,2	49,6	20
Finlande	4,9	2,7	79,5	4,7	2,5	17,2	56,1	11,1
Suède	3,9	2,5	81,3	4,9	2,5	9,2	56,9	12,9
GBR	5,9	3,1	78,9	6,2	2,2	8,8	53,5	14,8

Matrice de Corrélation

	Mariage	Divorce	EspVie	Mort.Inf	Médecin	Chômage	Act.Fem	Temp
Mariage	1,000	0,125	-0,688	0,342	0,170	-0,474	0,070	0,058
Divorce	0,125	1,000	-0,040	-0,466	0,101	-0,347	0,450	-0,612
Esp.Vie	-0,688	-0,040	1,000	-0,031	0,111	0,307	-0,219	0,447
Mort.Inf	0,342	-0,466	-0,031	1,000	0,466	0,041	-0,570	0,732
Médecin	0,170	0,101	0,111	0,466	1,000	0,063	-0,581	0,348
Chômage	-0,474	-0,347	0,307	0,041	0,063	1,000	-0,281	0,212
Act.Fem	0,070	0,450	-0,219	-0,570	-0,581	-0,281	1,000	-0,586
Temp	0,058	-0,612	0,447	0,732	0,348	0,212	-0,586	1,000

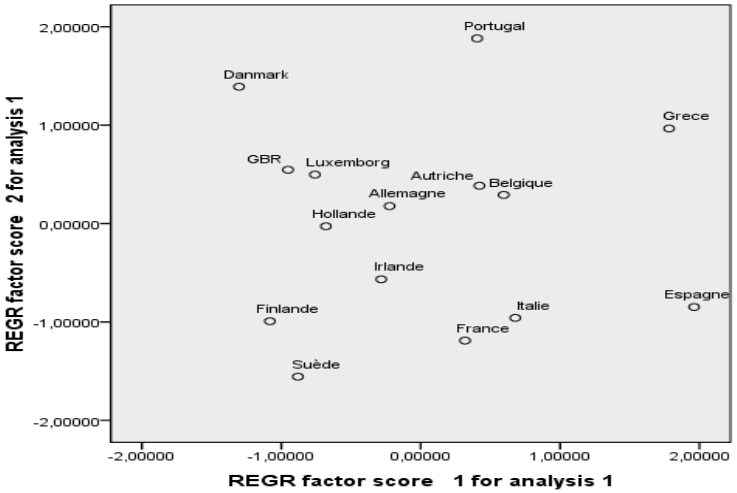
Matrice des composantes

Variables	Axe 1	Axe 2	Axe 3	Axe 4
Mariage	-0,091	0,934	-0,082	0,004
Divorce	-0,655	0,080	0,693	0,017
Esp.Vie	0,390	-0,714	0,352	0,443
Mort.Inf	0,775	0,481	-0,082	0,095
Médecin	0,536	0,316	0,697	-0,242
Chômage	0,412	-0,580	-0,146	-0,577
Act.Fem	-0,833	-0,069	-0,178	0,223
Temp	0,890	0,062	-0,099	0,358

Variance totale expliquée

Composante	Valeurs propres
1	3,134
2	2,065
3	1,167
4	0,775
5	0,437
6	0,222
7	0,186
8	0,015

Carte des Pays



Exercice 3 :

Un concessionnaire de voitures a réalisé une étude permettant de déterminer les liens entre les 6 caractéristiques suivantes d'une voiture (*Cylindrée, Puissance, Vitesse, Poids, Longueur et Largeur*).

L'étude a porté sur les 24 modèles suivants de voitures (Honda Civic ; Renault 19 ; Fiat Tipo ; Peugeot 405; Renault 21; Citroën BX ; BMW 530i ; Rover 827i ; Renault 25 ; Opel Omega ; Peugeot 405 Break ; Ford Sierra ; BMW325iX ; Audi 90Quattro ; Ford Scorpio ; Renault Espace ; Nissan Vanette ; VW Caravelle ; Ford Fiesta ; Fiat Uno ; Peugeot 205 ; Peugeot 205 Rallye ; Seat Ibiza SX I ; Citroën AX Sport).

Les résultats de l'application d'une ACP sur ces données sont présentés dans l'annexe ci-joint.

NB : ***A chaque réponse, Indiquer la source d'information (valeur ou tableau)***

1/ Est-il indispensable d'effectuer l'ACP sur les données centrées réduites ? Justifier votre réponse.
2/ Déterminer les corrélations les plus élevées, les plus faibles et les plus proches de zéro entre les variables. Interpréter les résultats obtenus.
3/ En examinant les valeurs propres, déterminer le nombre d'axes pouvant contenir au moins 90% de l'information totale.
4/ Calculer l'inertie absorbée par le premier axe principal et par le premier plan principal.
5/ Tracer le cercle des corrélations.
6/ Quelles sont les variables les mieux représentées.
7/ Quelle est la variable la plus corrélée avec chacune des 6 composantes principales.
8/ Représenter les 24 modèles sur le premier plan principal.
9/ Quels sont les modèles de voiture les mieux représentés ? Justifier votre réponse.
10/ Commenter cette ACP.

Annexe

Tableau 1 : Statistiques descriptives

	Minimum	**Maximum**	**Moyenne**	**Ecart-type**
Cylindrée	1116	2986	1906,1250	527,9087
Puissance	50	188	113,6667	38,7844
Vitesse	135	226	183,0833	25,2154
Poids	730	1510	1110,8333	230,2912
Longueur	350	473	421,5833	41,3405
Largeur	155	184	168,8333	7,6537

Tableau 2 : Valeurs propres

	1	***2***	***3***	***4***	***5***	***6***
Valeurs propres	*4.656*	*0.9152*	*0.2404*	*0.1027*	*0.0647*	*0.0210*

Tableau 3: Matrice des corrélations

	Cylindrée	**Puissance**	**Vitesse**	**Poids**	**Longueur**	**Largeur**
Cylindrée	1					
Puissance	0,861	1				
Vitesse	0,693	0,894	1			
Poids	0,905	0,746	0,491	1		
Longueur	0,864	0,689	0,532	0,917	1	
Largeur	0,709	0,552	0,363	0,791	0,864	1

Tableau 4 : Corrélations des variables avec les composantes

Variables	Axe 1	Axe 2	Axe 3	Axe 4	Axe 5	Axe 6
Cylindrée	0,96	0,03	-0.20	-0.02	0.20	0,00
Puissance	0,89	0,40	-0.02	-0,16	-0,08	0,08
Vitesse	0,74	0,63	0,18	0,10	0,00	-0,07
Poids	0,93	-0,24	-0,24	-0.04	-0,12	-0,08
Longueur	0,93	-0,28	0.02	0,23	-0,04	0,06
Largeur	0,81	-0,46	0,33	-0,12	0,03	-0,02

Tableau 5 : Coordonnées des unités statistiques

		Axe 1		Axe 2		Axe 3		Axe 4	
		Coord	**Cos²**	**Coord**	**Cos²**	**Coord**	**Cos²**	**Coord**	**Cos²**
1	**Honda Civic**	-2.02	0.88	0.32	0.02	0.53	0.06	-0.41	0.04
2	**Renault 19**	-0.78	0.66	-0.13	0.02	0.44	0.21	0.21	0.05
3	**Fiat Tipo**	-1.29	0.76	-0.43	0.09	0.47	0.10	-0.19	0.02
4	**Peugeot 405**	-0.27	0.11	-0.46	0.31	0.19	0.05	0.61	0.53
5	**Renault 21**	0.18	0.03	-0.64	0.42	-0.06	0.00	0.63	0.42
6	**Citroën BX**	-0.50	0.48	-0.21	0.08	0.15	0.04	0.42	0.33
7	**BMW 530i**	3.95	0.94	0.84	0.04	-0.52	0.02	-0.14	0.00
8	**Rover 827i**	3.19	0.94	0.77	0.06	-0.01	0.00	0.01	0.00
9	**Renault 25**	3.44	0.94	0.61	0.03	0.63	0.03	-0.19	0.00
10	**Opel Omega**	1.50	0.67	-0.78	0.18	0.51	0.08	0.40	0.05
11	**P 405 Break**	0.59	0.62	0.14	0.04	0.35	0.22	0.19	0.07
12	**Ford Sierra**	0.74	0.63	-0.43	0.22	0.11	0.01	0.32	0.12
13	**BMW 325i**	1.71	0.50	1.36	0.32	-0.98	0.17	-0.16	0.00
14	**Audi 90**	1.41	0.57	1.09	0.34	0.15	0.01	0.03	0.00
15	**Ford Scorpi**	2.80	0.92	-0.12	0.00	-0.39	0.02	-0.04	0.00
16	**Ren Espace**	0.92	0.45	-0.89	0.42	0.26	0.04	-0.40	0.08
17	**Nissan Vane**	-0.02	0.00	-1.82	0.64	-1.25	0.30	-0.10	0.00
18	**VW Carave**	1.22	0.19	-2.38	0.73	0.30	0.01	-0.67	0.06
19	**Ford Fiesta**	-3.50	0.93	-0.90	0.06	-0.07	0.00	-0.12	0.00
20	**Fiat Uno**	-3.76	0.98	-0.01	0.00	-0.51	0.02	0.14	0.00
21	**Peugeot 205**	-2.62	0.89	0.42	0.02	-0.80	0.08	0.00	0.00
22	**P 205 Rallye**	-2.29	0.70	1.48	0.29	0.10	0.00	0.11	0.00
23	**Seat Ibiza**	-1.93	0.79	0.90	0.17	-0.05	0.00	-0.35	0.03
24	**Citr AX Spo**	-2.65	0.78	1.30	0.19	0.45	0.02	-0.31	0.01

Analyse factorielle des correspondances

Introduction

L'AFC est une méthode d'analyse factorielle de données multidimensionnelles. Cette méthode est essentiellement descriptive. Elle généralise le test d'indépendance de khi-deux sur les tableaux de contingence. Autrement dit, des tableaux donnant la répartition d'une population statistique suivant des modalités de deux variables qualitatives.

Les données initiales sont constituées de deux variables qualitatives définies sur un ensemble d'unités statistiques. Ces variables qualitatives sont définies par des modalités dont on ne peut calculer la moyenne.

1. Présentation de la méthode

1.1. Exemple introductif

L'exemple que nous allons traités rassemble les réponses de 60 individus à la question suivante : indiquer votre source principale (journal, radio ou TV) pour vous informer sur les sujets suivants :

- La politique internationale
- La politique régionale
- Les faits divers
- Les résultats sportifs

Les résultats obtenus sont résumés dans le tableau suivant :

		Q_2			
		Journal	Radio	Télévision	
Q_1	PI	10	20	30	60
	PR	10	30	20	60
	FD	20	10	30	60
	RS	20	30	10	60
		60	90	90	240

Tableau 1 : tableau de contingence

Notations :
- La somme de la ligne i est appelée effectif marginal et sera notée par $n_{i.}$
- La somme de la colonne j est appelée effectif marginal et sera notée par $n_{.j}$

Par analogie avec l'analyse en composantes principales, on peut introduire, dans le tableau de contingence, des données appelées éléments supplémentaires.

Dans notre cas, on peut ajouter une colonne supplémentaire qui regroupe la radio et la télévision. Cette colonne représente l'audiovisuel. De même, on peut ajouter une ligne représentant la politique qui regroupe la politique internationale et régionale. On peut aussi regrouper les faits divers et les résultats sportifs sous une rubrique appelée événements.

	Journal	Radio	Télévision	
PI	10	20	30	50
PR	10	30	20	50
FD	20	10	30	40
RS	20	30	10	40
PO	20	50	50	
EV	40	40	40	

Tableau 2

Ces éléments supplémentaires ont la caractéristique de faciliter la tâche de l'analyste au niveau de l'interprétation des résultats mais in ne faut pas tenir compte de ces éléments dans la conclusion finale.

1.2. Tableau des probabilités et des profils

Dans la pratique, on doit transformer le tableau des effectifs en un tableau de probabilités en utilisant la formule suivante : $P_{ij} = \frac{n_{ij}}{n}$. On obtient ainsi le tableau suivant :

	Journal	Radio	Télévision	
PI	1/24	1/12	1/8	1/4
PR	1/24	1/8	1/12	1/4
FD	1/12	1/24	1/8	1/4
RS	1/12	1/8	1/24	1/4
	1/4	3/8	3/8	1

Tableau 3

Notations :
- La somme de la ligne i est appelée probabilité marginale et sera notée par $P_{i.}$
- La somme de la colonne j est appelée probabilité marginale et sera notée par $P_{.j}$

Le tableau des probabilités nous permet de définir les tableaux des profils lignes et colonnes. En effet, l'utilisation des probabilités conditionnelles nous permette de définir deux types de tableaux. Le tableau des profils lignes qui comporte les probabilités définies par le rapport $\frac{P_{ij}}{P_{i.}}$ et le tableau des profils colonnes qui comporte les probabilités définies par le rapport $\frac{P_{ij}}{P_{.j}}$. On obtient ainsi les deux tableaux suivants :

	Journal	Radio	Télévision	
PI	1/6	1/3	1/2	1
PR	1/6	1/2	1/3	1
FD	1/3	1/6	1/2	1
RS	1/3	1/2	1/6	1

Tableau 4 : tableau des profils lignes

	Journal	Radio	Télévision
PI	1/6	2/9	1/3
PR	1/6	1/3	2/9
FD	1/3	1/9	1/3
RS	1/3	1/3	1/9
	1	1	1

Tableau 5 : tableau des profils colonnes

2. Technique de la méthode

La méthodologie consiste à mesurer la distance entre deux profils afin de comparer les termes de même rang. Pour cela on va généraliser le test classique de khi-deux.

Définition : On appelle distance de khi-deux, de centre P_j entre les deux profils lignes i et i', la valeur définie par :

$d^2(i, i') = \sum_j \frac{\left(P_j^i - P_j^{i'}\right)^2}{P_{.j}}$. De même, on peut définir la distance de khi-deux de centre P_i entre les deux profils colonnes j et j' comme suit : $d^2(j, j') = \sum_i \frac{\left(P_i^j - P_i^{j'}\right)^2}{P_{i.}}$

Application : Calculons la distance entre les deux profils colonnes Radio et Télévision. Les tableaux 5 et 3 montrent que :

$d^2(R,T) = (2/9\text{-}1/3)^2/_{1/4} + (1/3\text{-}2/9)^2/_{1/4} + (1/9\text{-}1/3)^2/_{1/4} + (1/3\text{-}1/9)^2/_{1/4} = 0{,}494$.

Les autres distances de khi-deux entre les différents profils colonnes sont définies dans le tableau suivant :

	Journal	Radio	Télévision
Journal	0		
Radio	0,321	0	
Télévision	0,321	0,494	0

Tableau 6

3. Interprétation des résultats

Pour interpréter les différents résultats, on utilise les mêmes techniques que celles utilisées dans le cas de l'ACP. On définit les axes principaux comme étant les droites les plus proches des profils selon le critère des moindres carrées. De même pour les valeurs propres, en les ordonne dans un ordre décroissant et par la suite, on détermine l'inertie expliquée et l'inertie cumulée.

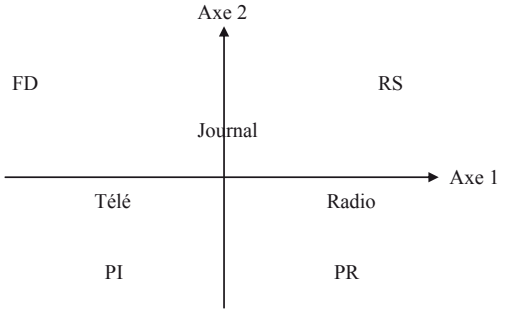
Notons que la somme des valeurs propres est définie par : $\sum_i \alpha_i = \frac{X^2}{n}$ où X^2 est la statistique définie par le test de khi-deux.

$$\varkappa^2{}_{cal} = \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{\left(n_{ij} - nP_{i.}P_{.j}\right)^2}{nP_{i.}P_{.j}}$$

A noter aussi que dans un tableau de contingence à p lignes et q colonnes, le nombre de valeurs propres est inférieur ou égal à inf(p-1,q-1). Dans le cas de l'exemple introductif, ce nombre est inférieur ou égal à 2.

La représentation graphique des différentes modalités peut être définie sur deux graphiques séparés : le premier pour la première variable et le second pour la deuxième variable. Toutefois, il est plus intéressant de les représenter dans un seul plan principal.

Application : En ce qui concerne l'exemple introductif, la représentation se fait comme suit :



Graphique 1

Au niveau de l'interprétation, on se base sur les mêmes constatations que dans le cas de l'ACP. Les modalités les mieux représentées sont celles qui sont éloignées par rapport aux axes principaux. Inversement, les mal reconstruites sont celles qui sont très proches de l'origine.
On peut aussi utiliser les projections des modalités sur les différents axes pour apprécier le degré de représentation et définir la position par rapport à un plan.

Application :
- Dans notre cas, on peut dire qu'il y a une opposition de la télé et la radio par rapport à la deuxième composante principale (axe 2). Ce dernier axe représente en grande partie le journal. Ce même axe décèle une opposition entre les faits divers et la politique internationale d'une part et les résultats sportifs et la politique régionale d'autre part.
- D'un autre coté, on peut remarquer une opposition entre le journal d'une part et la télé et la radio d'autre part et ce, par rapport à la première composante principale (axe 1). Par rapport à ce même axe, on peut noter une opposition entre les faits divers et les résultats sportifs d'un coté et d'un autre coté la télé et la radio.
La représentation graphique montre que les individus interrogés s'intéressent à la télé pour s'informer de la politique internationale. Ceux qui veulent s'informer sur la politique régionale écoutent plutôt la radio.
On peut dire aussi que la moitié des interrogés s'intéressant aux résultats sportifs et les faits divers cherchent l'information dans un journal.

Annexes des tableaux :

		Axe 1			Axe 2		
	Poids	Corrélation	Cos²	Contribution	Corrélation	Cos²	Contribution
PI	1/4	-0,192	0,5	10	-0,192	0,5	25
PR	1/4	0,192	0,5	10	-0,192	0,5	25
FD	1/4	-0,385	0,8	40	0,192	0,2	25
RS	1/4	0,385	0,8	40	0,192	0,2	25

Tableau 7 : profil ligne

		Axe 1			Axe 2		
	Poids	Corrélation	Cos²	Contribution	Corrélation	Cos²	Contribution
J	1/4	0	0	0	0,333	1	75
R	3/8	0,351	0,909	50	-0,111	0,091	12,5
T	3/8	-0,351	0,909	50	-0,111	0,091	12,5

Tableau 8 : profil colonne

4. Notion de contribution :
La contribution est un paramètre qui sert à mesurer l'influence d'une modalité dans la définition d'un axe principal l. Dans l'AFC, les profils sont affectés de poids et interviennent de façon plus ou moins importante dans le calcul des axes.
La somme des carrés des coordonnées des profils pondérés par les poids est égale à la valeur propre relative à l'axe l définie par :

$$\alpha_l = \sum_{i=1}^{p} P_{i.} g_l^2(i)$$

La contribution relative d'un profil à l'inertie expliquée par l'axe principal l est exprimée en pourcentage comme suit:

$$Contr(i) = \frac{P_{i.} g_l^2(i)}{\alpha_l} * 100$$

Exercices

Exercice 1 :

Afin de répondre à l'augmentation de son activité, une entreprise souhaite modifier ses horaires d'ouverture. Quatre solutions sont envisagées :
– Matin : ouvrir une heure plus tôt le matin ;
– Midi-deux : ne pas fermer entre midi et deux ;
– Soir : ouvrir une heure plus tard le soir ;
– Samedi : ouvrir le samedi matin.
Une enquête est effectuée auprès des 150 salariés de l'entreprise pour connaître leurs préférences.
50 salariés ont répondu (anonymement) à l'enquête. Parmi les données collectées, 2 variables ont attiré l'attention du directeur des ressources humaines
– Statut : le statut du salarié (ouvrier, employé, cadre) ;
– Préférence : La solution préférée par le salarié parmi les 4 solutions envisagées ci-dessus.
Les résultats de l'étude sont présentés dans **le tableau 1**:

Tableau 1

Statut	**Préférence**			
	Matin	Midi-deux	Soir	Samedi
Ouvrier	14	2	3	1
Employé	1	13	3	4
Cadre	1	1	6	1

1/ Déterminer le tableau des profils lignes et celui des profils colonnes.
2/ Quel est le nombre des valeurs propres ?
3/ **Le tableau 2** donne les coordonnées des points lignes et colonnes par rapport aux deux axes principaux.
Représenter graphiquement les modalités lignes et les modalités colonnes dans le plan principal 1x2.
4/ Calculer les contributions des différentes modalités (ligne et colonne) dans la définition des deux axes principaux
5/ Interpréter le graphique obtenu.

Tableau 2

Points lignes			**Points colonnes**		
	Axe 1	Axe 2		Axe 1	Axe 2
Ouvrier	-0,961	0,281	Matin	-1,127	0,371
Employé	0,875	0,362	Midi-deux	0,861	0,500
Cadre	0,095	-1,469	Soir	0,037	-1,208
			Samedi	0,634	0,091

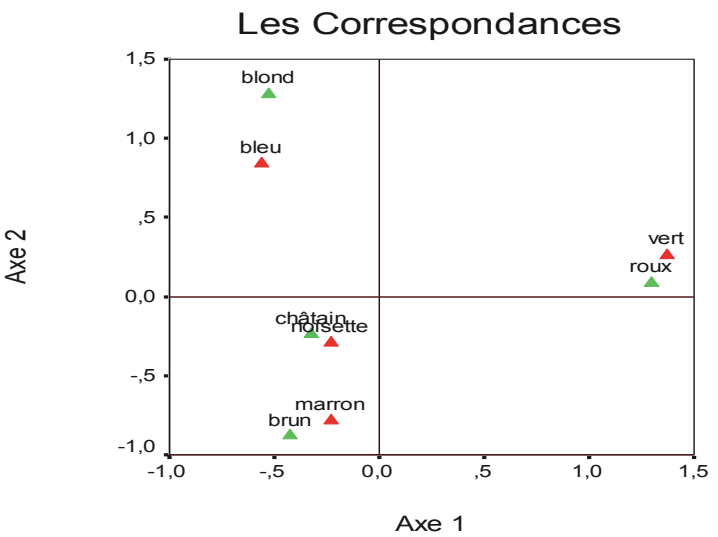
Exercice 2 :

Le tableau 1 de contingence est obtenu en ventilant une population de 691 femmes suivant leurs couleurs des yeux et des cheveux. Les résultats sont donnés dans **l'annexe** ci-joint.

Tableau 1				
cheveux / yeux	blond	brun	châtain	roux
bleu	92	20	84	17
marron	7	78	119	36
noisette	10	15	56	14
vert	16	5	29	93

1/ Compléter le tableau par ces marges.
2/ Donner les tableaux des profils lignes et profils colonnes et leurs matrices des poids (probabilités) associées.
3/ Calculer l'inertie totale.
4/ Commenter les résultats de l'AFC.

Annexe



Caractéristiques des points lignes

		Axe 1		Axe2		
Yeux	Poids	Coord	contrib	coord	contrib	Inertie
bleu	0,308	-0,563	0,190	0,838	0,476	0,149
marron	0,347	-0,227	0,035	-0,783	0,467	0,108
noisette	0,137	-0,233	0,015	-0,295	0,026	0,019
vert	0,207	1,374	0,761	0,261	0,031	0,207

Caractéristiques des points colonnes

		Axe 1		Axe2		
Cheveux	Poids	coord	contrib	coord	contrib	Inertie
blond	0,181	-0,524	0,097	1,280	0,651	0,161
brun	0,171	-0,422	0,059	-0,882	0,292	,0082
chatain	0,417	-0,321	0,084	-0,241	0,053	0,038
roux	0,232	1,299	0,760	0,085	0,004	0,202