Université de Alassane Ouattara de Bouaké

Faculté des sciences économiques **(Travail personnel)**

Exercices d'analyse de données

Emmanuel ASSOHOU

Bouaké, le 12 mars 2024

Liste des tableaux

Table des figures

Exercice 1

1) *Réaliser une ACP (analyse en composantes principales) sur l'ensemble des variables de la base, à l'aide du logiciel STATA, et interpréter les résultats.*

Après avoir effectué le test de l'ACP grâce à la commande :

```
pca popul tact superf nbentr nbbrev chom teleph
```

On obtient donc les résultats suivants : 61,85% de l'inertie est expliquée par la variable **population**, tandis que la variable **taux d'activité** explique 20,42% de l'inertie. Par ailleurs, 14,46% de l'inertie est expliquée par la variable **superficie de la région**, 2,61% de l'inertie est expliquée par la variable **nombre d'entreprises dans la région**, et 0,47% de l'inertie est expliquée par la variable **nombre de brevets déposés au cours de l'année**. Enfin, les variables **taux de chômage** et **nombre de lignes téléphoniques en place dans la région** expliquent respectivement 0,15% et 0,03% de l'inertie.

En outre, en appliquant le critère de Kaiser, nous choisirons 3 composantes car seulement celles-ci ont des valeurs propres supérieures à 1 (4,33 ; 1,43 ; 1,02). Dès lors, k sera égal à 3 au lieu de 7 [1].

Concernant les valeurs des différentes composantes, nous aurons autant de composantes que de variables, dont les équations sont données comme suit :

$$\begin{cases} \mathrm{C}^1 = 0,46\mathrm{Popul} + 0,35\mathrm{tact} - 0,01\mathrm{superf} + 0,46\mathrm{nbentr} + 0,47\mathrm{nbbrev} - 0,14\mathrm{chom} + 0,47\mathrm{telph} \\ \vdots \\ \mathrm{C}^7 = 0,38\mathrm{Popul} + 0,01\mathrm{tact} - 0,03\mathrm{superf} + 0,32\mathrm{nbentr} + 0,16\mathrm{nbbrev} - 0,02\mathrm{chom} - 0,85\mathrm{telph} \end{cases}$$

2) *Réaliser une ACP sur l'ensemble des variables, en permutant les positions de la première et de la deuxième variable de la base.*

Après avoir effectué le test de l'ACP grâce à la commande :

```
pca popul tact superf nbentr nbbrev chom teleph
```

On obtient donc les résultats suivants : 61,85% de l'inertie est expliquée par la variable **taux d'activité**, tandis que la variable **population** explique 20,42% de l'inertie. Par ailleurs, 14,46% de l'inertie est expliquée par la variable **superficie de la région**, 2,61% de l'inertie est expliquée par la variable **nombre d'entreprises dans la région**, et 0,47% de l'inertie est expliquée par la variable **nombre de brevets déposés au cours de l'année**. Enfin, les variables **taux de chômage** et **nombre de lignes téléphoniques en place dans la région** expliquent respectivement 0,15% et 0,03% de l'inertie.

1. Cette valeur désigne le nombre de variables

En outre, en appliquant le critère de Kaiser, nous choisirons 3 composantes car seulement celles-ci ont des valeurs propres supérieures à 1 (4,33 ; 1,43 ; 1,02). Dès lors, k sera égal à 3 au lieu de 7 [2].

Concernant les valeurs des différentes composantes, nous aurons autant de composantes que de variables, dont les équations sont données comme suit :

$$\begin{cases} C^1 = 0,35\text{tact} + 0,46\text{Popul} - 0,01\text{superf} + 0,46\text{nbentr} + 0,47\text{nbbrev} - 0,14\text{chom} + 0,47\text{telph} \\ \vdots \\ C^7 = 0,01\text{tact} + 0,38\text{Popul} - 0,03\text{superf} + 0,32\text{nbentr} + 0,16\text{nbbrev} - 0,02\text{chom} - 0,85\text{telph} \end{cases}$$

3) *Analyser les différences dans les résultats des questions 1 et 2. Commenter.*

Il n'existe pas de différence entre les deux estimations. Cela pourrait s'expliquer par le fait de la non existence d'une variable indépendante et dépendante dans l'Analyse des composantes principales.

4) *Donner la syntaxe de STATA pour présenter dans un tableau, après réalisation de l'ACP, les statistiques descriptives (moyenne, écart-type, minimum, maximum) des variables utilisées. Le nombre d'observations ne doit pas apparaitre dans le tableau.*

```
1    summarize popul tact superf nbentr nbbrev chom teleph
```

5) *Exécuter la commande « screeplot » après réalisation de l'ACP, et interpréter le résultat obtenu.*

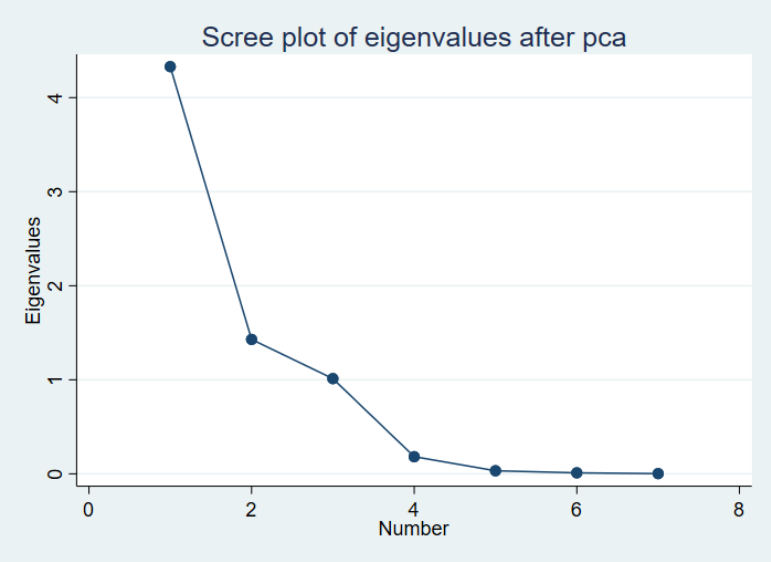

FIGURE 1 – Diagramme de coude des valeurs propres après le test de l'ACP

2. Cette valeur désigne le nombre de variables

INTERPRÉTATION : L'on constate que la variance expliquée par chaque composante principale (COMP$_i$ pour $i \in \{1; 2; ...; 7\}$) est décroissante. Par ailleurs, contrairement aux composantes 4, 5, 6 et 7, dont ($\lambda_i < 1$, pour $i \in \{4; 5; ...; 7\}$) les composantes principales 1, 2 et 3 apportent plus d'information que les variables principales.

Exercice 2

1) Calculer les profils-lignes, présenter les résultats dans un tableau, et les interpréter.

Pour cela, il faudrait trouver les valeurs de x et y du tableau.

Sachant que :

$$\begin{cases} 16 + 12 + y + 17 + x = 50 \\ 23 + 16 + x + 6 + y = 50 \end{cases}$$

On obtiendra alors,

$$\begin{cases} 45 + x + y = 50 \\ 45 + x + y = 5 \end{cases}$$

Ainsi, on conclura que $x + y = 5 \Rightarrow x = 5 - y$ Par conséquent, sachant encore que $4xy - y^2 = 0$ nous aurons :

$$\begin{aligned} & 4(5-y)y - y^2 = 0 \\ & \Rightarrow 20y - 5y^2 - y^2 = 0 \\ & \Rightarrow 20y - 5y^2 = 0 \\ & \Rightarrow y = 0 \text{ ou } y = \frac{20}{5} = 4 \end{aligned}$$

Ainsi, nous obtiendrons $4x \times 4 - 4^2 = 0 \Rightarrow x = 1$

Dès lors $x = 1$ et $y = 4$

	1	2	3	Somme
A	8	3	12	23
B	7	9	1	17
C	1	4	5	10
Somme	16	16	18	50

TABLE 1 – Niveau de diplôme le plus élevé du répondant

	1	2	3	Somme
A	0,35	0,13	0,52	1,00
B	0,41	0,53	0,06	1,00
C	0,10	0,40	0,50	1,00
Somme	0,86	1,06	1,08	3,00

TABLE 2 – Tableau des profils-ligne

Interprétation :

— Parmi les personnes dont la tranche est de 15 à 24 ans, 35% ont un diplôme inférieur au BAC, 13% ont un BAC-BAC +3 et 52% ont un master et plus.

— Parmi celles dont l'âge est compris entre 25 et 44 ans, 41% ont un diplôme inférieur au BAC, 53% ont un diplôme équivalent au BAC-BAC +3 et seulement 6% ont un master et plus.

— Enfin, parmi les personnes dont l'âge varie entre 45 ans et plus, 10% ont un diplôme inférieur au BAC, 40% ont un BAC-BAC +3 et 50% ont un master et plus.

2) Calculer les profils-colonnes, présenter les résultats dans un tableau, et les interpréter.

	1	2	3
A	0,50	0,19	0,67
B	0,44	0,56	0,06
C	0,06	0,25	0,28
Somme	1,00	1,00	1,00

TABLE 3 – Tableau des profils-colonne

Interpretation :

— Parmi les individus qui ont un diplôme inférieur au BAC, 50% ont un âge variant entre 15 et 24 ans, 44% ont âge compris entre 25 et 44 ans et 6% ont 45 ans et plus.

— Parmi les personnes diplômées d'un BAC-BAC +3, 19% ont un âge compris entre 15 et 24, 56% ont un âge compris entre 25 et 44 ans tandis que 25% ont 45 ans et plus.

— Parmi les personnes ayant un master et plus, 67% ont entre 15 et 24 ans, 6% ont entre 25 et 44 ans et 28% ont 45 ans et plus.

3) Calculer les contributions au χ^2

Pour effectuer cette tâche, nous calculerons les effectifs conjoints donnés par la formule suivante : $n_{ij} = \frac{n_{i.} \times n_{.j}}{n}$ Ce qui permet donc d'avoir le tableau ci-dessous :

	1	2	3	Somme
A	7,36	7,36	8,28	23,00
B	5,44	5,44	6,12	17,00
C	3,20	3,20	3,60	10,00
Somme	16,00	16,00	18,00	50,00

TABLE 4 – Chi-Square Statistic Expected Values

Ensuite, nous faisons la différence entre les valeurs observées et les valeurs attendues en utilisant la formule suivante : $n_{ij} - \frac{n_{i.} \times n_{.j}}{n}$

Enfin, afin d'avoir la contribution au χ^2, nous ferons donc : $\frac{\left(n_{ij} - \frac{n_{i.} \times n_{.j}}{n}\right)^2}{\frac{n_{i.} \times n_{.j}}{n}}$

4) Effectuer le test d'indépendance de V_1 et V_2. On donne $\chi^2_{5\%}(4) = 9,48$; $\chi^2_{5\%}(6) = 12,59$; $\chi^2_{5\%}(8) = 15,51$.

	1	2	3	Somme
A	0,64	-4,36	3,72	0,00
B	1,56	3,56	-5,12	0,00
C	-2,20	0,80	1,40	0,00
Somme	0,00	0,00	0,00	0,00

TABLE 5 – Observed Minus Expected Values

	1	2	3	Somme
A	0,06	2,58	1,67	4,31
B	0,45	2,33	4,28	7,06
C	1,51	0,20	0,54	2,26
Somme	2,02	5,11	6,50	13,63

TABLE 6 – Contributions to the Total Chi-Square Statistic

La formulation du test est la suivante :

$$\begin{cases} \mathrm{H}_0 : V_1 \text{ et } V_2 \text{ sont indépendantes} \\ \mathrm{H}_1 : V_1 \text{ et } V_2 \text{ sont dépendantes} \end{cases}$$

Sachant que $\chi_{5\%}(\text{calculé}) = 13,63$ et que $\chi_{5\%}(L-1)(C-1) = \chi_{5\%}(4) = 9.48$ alors on en déduit que les variables V_1 et V_2 sont dépendantes car $\chi_{5\%}(\text{calculé}) > \chi_{5\%}(\text{théorique})$.

Par conséquent, les variables «tranches d'âge » et «niveau de diplôme élevé du répondant » sont liées.

Exercice 3

1. a) Préciser l'espace des individus et l'espace des variables.

Il y a quatre individu donc l'espace des individus est $\mathbb{R}^4$

Il y a trois variables, donc l'espace des variables est $\mathbb{R}^3$

1. b) Calculer la matrice R des corrélations.

Afin de calculer la matrice de corrélation, nous allons d'abord centrer-réduire les différentes valeurs grâce à la formule : $x^j = \frac{X^j - \overline{X}^j}{\sigma_{X^j}}$

Pour cela, il faudra l'écart-type et la moyenne de chaque colonne du tableau grâce aux formules respectives : $\sigma_{X^j} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X^j - \overline{X}^j\right)^2}$ et $\overline{X}^j = \frac{1}{n}\sum_{i=1}^{n} X^j$

Dès lors, nous pouvons calculer la matrice de corrélation R $= \frac{1}{n}x'x$

Mean	1,00	3,00	-2,00
std-Error	2,31	1,15	1,15

TABLE 7 – Moyennes et écart-type de chaque colonne

-0,87	0,87	0,87
-0,87	-0,87	0,87
0,87	0,87	-0,87
0,87	-0,87	-0,87

TABLE 8 – Valeurs centrées réduites

1,00	0,00	-1,00
0,00	1,00	0,00
-1,00	0,00	1,00

TABLE 9 – Matrice de corrélation

2. Donner l'analogie d'une part, et la différence d'autre part, entre une variable discriminante et une composante principale.

Objectif :

- Les variables discriminantes sont utilisées dans l'analyse discriminante pour trouver des combinaisons linéaires de variables qui maximisent la séparation entre les groupes de données connus.
- Les composantes principales sont utilisées dans l'analyse en composantes principales pour réduire la dimensionnalité des données en trouvant des directions qui capturent le plus de variance possible.

Nature :

— Les variables discriminantes sont spécifiques à un problème de classification particulier. Elles sont sélectionnées pour leur capacité à séparer les groupes de données.

— Les composantes principales sont des axes de variation dans l'ensemble de données d'origine. Elles ne sont pas spécifiques à un problème de classification particulier, mais visent plutôt à réduire la complexité des données.

Interprétation :

— Les variables discriminantes sont souvent interprétées en termes de contribution à la séparation entre les groupes de données.

— Les composantes principales sont interprétées en termes de contribution à la variance totale des données.

3) Démontrer que la diagonale principale d'une matrice des corrélations est un vecteur unitaire.

Sachant que lorsqu'on fait la moyenne du produit de la la matrice centrée-réduite et de sa transposée, on obtient donc la matrice de corrélation R dont la diagonale est égale à : $\frac{1}{n}\sum_{i=1}^{n}\left(\frac{X^j-\overline{X}^j}{\sigma_X^j}\right)^2$ Par ailleurs, sachant que $x^j = \frac{X^j-\overline{X}}{\sigma_X^2} \Rightarrow x \rightsquigarrow \mathcal{N}(0,1)$ Or sachant que sur la diagonale nous avons : $d_j = \frac{1}{n}\sum_{i=1}^{n}\left(x^j\right)^2$ et comme $x^j \rightsquigarrow \mathcal{N}(0,1)$ alors $\left(x^j\right)^2 \rightsquigarrow \chi^2(1)$ Ainsi,

$$d_j = \frac{1}{n}\sum_{i=1}^{n}\chi^2(1) = \frac{1}{n}\chi^2(n) = \frac{1}{n}\times n = 1$$

C'est pourquoi la diagonale de la matrice de corrélation est égale au vecteur unitaire.

4) Quelle est l'utilité du test d'indépendance du Chi-deux (χ^2) dans la mise en œuvre de l'analyse factorielle des correspondances ?

Le test d'indépendance du χ^2 permet de vérifier la dépendance entre les deux variables. Il est donc déterminant dans le processus de décision d'effectuer une (AFC) ou non.