Université PANTHEON – ASSAS (PARIS II)

Droit-Economie-Sciences Sociales

Janvier 2018

Master 1 Economie-Gestion mention Ingénierie économique et statistique
Analyse des données IES (4008)
(Unité d'Enseignements Complémentaires 1)

Les étudiants sont autorisés à utiliser une machine à calculer et à consulter les notes de cours, sous forme de transparents distribués par l'enseignant, sans ajout de notes personnelles

I Questions de cours (4 points):

1. Expliquer l'utilité du cercle des corrélations dans une analyse en composantes principales. Présenter en détail le phénomène dit d'effet taille (**2 points**).
2. Brièvement, quels sont les objectifs des méthodes factorielles d'analyse de données ? (**1 point**)
3. Définir le point moyen d'un tableau de données noté X, représentation de l'ensemble des valeurs prises pour p variables par n individus (**1 point**).

Exercice 1 (11 points):

On a posé deux questions à un échantillon de plusieurs centaines de personnes :

a. Quelles sont les raisons qui, selon vous, peuvent faire hésiter une femme ou un couple à avoir un enfant ?
b. Quel est votre niveau d'études ?

Pour la deuxième question, les réponses possibles étaient : sans diplôme (SANS), certificat d'études primaires (CEP), brevet d'études du premier cycle (BEPC), baccalauréat ou équivalent (BAC), université, grandes écoles ou équivalent (UNIV). Pour la première question, les réponses ont été analysées.

Pour la première question, on a retenu 15 des mots utilisés :

peur, sante, avenir, argent, emploi, guerre, chômage, travail, égoïsme, finances, logement, difficile, économique, financières, conjoncture. Chaque personne peut avoir utilisé plusieurs de ces mots.

1. Considérant le tableau croisé permettant de mesurer le nombre d'occurrences des mots choisis en fonction du niveau d'études donné ci-dessous, déterminez s'il est possible de calculer le nombre de personnes interrogées (**1 point**)
2. Est-ce que le mot « économique » apparait plus fréquemment dans la bouche des universitaires que dans celle de l'ensemble de la population interrogée ? Quelle est la probabilité qu'un individu ait le Bepc s'il a répondu Emploi ? (**1 point**)

	Sans	Cep	Bepc	Bac	Universitaire	TOTAL
Peur	25	45	38	38	13	159
Santé	18	27	20	19	9	93
Avenir	53	90	78	75	22	318
Argent	51	64	32	29	17	193
Emploi	12	35	19	6	7	79
Guerre	4	7	7	6	2	26
Chômage	71	111	50	40	11	283
Travail	35	61	29	14	12	151
Egoïsme	21	37	14	26	9	107
Finances	10	7	7	3	1	28
Logement	8	22	7	10	5	52
Difficile	7	11	4	3	2	27
Economique	7	13	12	11	11	54
Financières	21	32	42	47	30	172
Conjoncture	1	7	5	5	4	22
TOTAL	344	569	364	332	155	1764

3. Une Analyse Factorielle des Correspondances est réalisée sur ces données. Les résultats mettent à disposition les quatre valeurs propres des axes factoriels suivantes :

	1er axe	2ème axe	3ème axe	4ème axe
Valeur propre	0.0528	0.0119	0.0072	0.0060

Expliquez pourquoi il y a 4 valeurs propres (**1 point**). Définir et donner la valeur de l'inertie totale du nuage (**1 point**). Enfin, trouvez-vous logique de n'interpréter que les deux premiers axes factoriels? Justifier votre réponse (**1 point**)

4. Les coordonnées, centrées et pondérées, des variables de niveaux d'étude et de mots cités dans le premier plan factoriel sont reportées dans les deux tableaux ci-dessous.

Coordonnées des mots

	Axe 1	Axe 2
Peur	0,1399	-0,0939
Santé	0,0673	-0,0155
Avenir	0,1089	-0,1285
Argent	-0,1443	0,0378
Emploi	-0,1754	0,2044
Guerre	0,1522	-0,1129
Chômage	-0,2683	-0,0456
Travail	-0,2430	0,1259
Egoïsme	0,0013	-0,0377
Finances	-0,2664	-0,1538
Logement	-0,0603	0,1098
Difficile	-0,2826	0,0962
Economique	0,3533	0,2582
Financières	0,4543	0,0659
Conjoncture	0,3689	0,2448

Coordonnées des niveaux d'études

	Axe 1	Axe 2
Sans	-0,2358	-0,0449
Cep	-0,1953	0,0544
Bepc	0,1144	-0,0401
Bac	0,2774	-0,1331
Universitaire	0,3774	0,2792

	Axe 3	Axe 4
Sans	0.1402	0,0420
Cep	-0,0721	-0,0527
Bepc	-0.0715	0.1307
Bac	0,0103	-0,0954
Universitaire	0,0995	0,0012

Quelles sont les catégories (mots, niveau d'étude) qui déterminent chacun des deux axes factoriels ? Vous expliquerez la méthode utilisée et caractériserez les axes en termes d'opposition entre mots ou niveaux d'études (**2 points**).

5. Déterminer si les catégories de niveau d'études Bepc et Universitaire sont bien ou mal représentées par le premier plan de l'analyse. Expliquer soigneusement les critères que vous utilisez et la signification géométrique de la qualité de représentation. (**2 points**).

6. Les mots peur et guerre d'une part et argent et CEP, d'autre part, sont proches sur le plan. Dans chacun des cas, quelles sont les données qui permettent d'expliquer cette proximité ? (**2 points**)

Exercice 2 (6 points):

On étudie les résultats des élections présidentielles de 2017 à Marseille. Les variables représentant les candidats sont : **Arthaud** (Artha, Extrême gauche), **Poutou** (Pouto, Extrême gauche), **Mélenchon** (Mélen, Gauche), **Hamon** (Hamon, Socialiste), **Macron** (Macro, Centre), **Lassalle** (Lassalle, Centre droit), **Fillon** (Fillo, Droite), **Dupont-Aignan** (Dupon, Droite souverainiste), **LePen** (LePen, Extrême droite).

Les individus sont les 16 arrondissements de Marseille (I à XVI).

Dans cette analyse, les données sont les pourcentages obtenus par les 11 candidats dans les 16 lieux étudiés. Le tableau de données est une matrice dont les colonnes sont les variables correspondant aux 11 candidats et les individus sont les groupes de votants des 16 arrondissements étudiés. Nous donnons ci-dessous la matrice des corrélations entre les variables.

Matrice des corrélations entre variables

	Artha	Pouto	Mélen	Hamon	Macro	Lassa	Fillo	Dupon	LePen
Artha	1.00	0,55	0,46	0,67	-0,12	-0,84	-0,79	-0,44	0,51
Pouto	0,55	1.00	0,53	0,36	0,40	-0,63	-0,52	-0,42	0,40
Mélen	0,46	0,53	1.00	0,06	0,02	-0,58	-0,84	-0,25	0,84
Hamon	0,67	0,70	0,45	1.00	0,83	-0,78	-0,34	-0,78	-0,16
Macro	-0,12	0,40	0,06	0,83	1.00	-0,70	-0,70	-0,47	0,05
Lassa	-0,12	-0,63	-0,58	-0,78	-0,70	1.00	0,69	0,74	-0,63
Fillo	-0,79	-0,52	-0,84	-0,34	-0,70	0,69	1.00	0,22	-0,82
Dupon	-0,45	-0,42	-0,25	-0,78	-0,47	0,74	0,22	1.00	0,12
LePen	-0,32	0,40	0,84	-0,16	0,05	-0,63	-0,82	0,12	1.00

1. Comment se groupent les variables du point de vue des signes de corrélation ? Quelle est la particularité du vote pour le candidat d'extrême droite LePen (**1 point**)?

On procède à une analyse en composantes principales sur variables centrées-réduites mesurant les scores en pourcentage des différents candidats. On donne ci-après les valeurs propres associées aux composantes principales, sauf la valeur de la quatrième (en gras ci-dessous).

Axes	1	2	3	4	5	6	7	8	9
Valeurs propres	5,14	2,27	0,82	**Xxx**	0,14	0,07	0,03	0,02	0

2. Sachant que les variables étudiées sont centrées et réduites, en déduire l'inertie totale ou somme des valeurs propres. Quelle est alors la valeur manquante de la valeur propre associée au quatrième axe ? (**1 point**)

3. Combien d'axes doit-on garder pour l'analyse ? Quelle sera alors la qualité globale de la représentation du nuage sur ce sous ensemble d'axes ? (**1 point**)

4. Le tableau ci-dessous donne les coefficients de corrélation entre les neuf variables centrées réduites et les trois premières composantes principales. A l'aide de ce tableau, expliquer les variables qui déterminent les deux premières composantes principales (précisez clairement les critères utilisés). (**1 point**)

Corrélations variables / composantes principales			
	1ère comp	2ème comp	3ème comp
Macro	0.84	0.46	-0.17
Hamon	-0.71	0.66	0.01
Fillo	0.95	-0.14	0.13
LePen	-0.54	-0.81	0.13
Artha	-0.92	0.05	0.34
Pouto	-0.85	0.19	0.44
Mélen	-0.66	-0.66	-0.29
Dupon	0.63	-0.52	0.52
Lassalle	-0.71	-0.48	-0.31

5. Comment interpréter ces deux premiers axes principaux en fonction des variables ? (**2 points**)