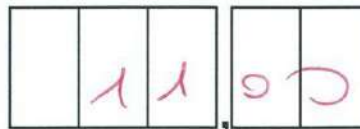


Groupe C

Dnaphr

GAUTIER Arthur
L2 - 2014

DEVOIR ECRIT ANALYSE DE DONNEES - juin 2015.

*Sans calculatrice, ni document. Vous composerez sur le sujet.***Exercice 1 : Démonstration de cours (4 points)**

On rappelle les notations utilisées en AFC :

Considérons deux variables X et Y étudiées sur un échantillon de taille N . On note X_i avec $1 \leq i \leq n$, et Y_j avec $1 \leq j \leq p$ les modalités de chacune de ces variables.L'effectif de la modalité « $X=X_i$ » est noté $k_{i\cdot}$, celui de la modalité « $Y=Y_j$ » $k_{\cdot j}$, celui de la modalité « $X=X_i$ et $Y=Y_j$ » k_{ij} . La fréquence de la modalité « $X=X_i$ » est notée f_i ; la fréquence de la modalité « $Y=Y_j$ » est notée f_j ; la fréquence de la modalité « $X=X_i$ et $Y=Y_j$ » est notée f_{ij} .

1. Rappeler la définition de la matrice notée X^* .
2. Dans le cas de l'analyse des PFL, on diagonalise une matrice notée T^* et dans le cas de l'analyse des PFC, on diagonalise une matrice notée W^* . Donner l'expression de ces deux matrices en fonction de la matrice X^* .
3. Soit λ une valeur propre non triviale de T^* et u un vecteur propre unitaire associé. Démontrer que λ est également valeur propre de la matrice W^* et déterminer un vecteur propre unitaire associé.

Exercice n°2 : (2 points)

Le tableau figurant ci-dessous donne les résultats concernant des variables de différents types pour certains pays du monde en 2007.

1. Expliquer pourquoi on ne peut pas faire une ACP de ces données. Quel est le nom de la méthode d'analyse des données adéquate dans ce cas ?

Ici, on travaille sur des valeurs qualitatives et non quantitatives. On ne peut donc pas faire d'ACP. On fera donc une ACM

2. Construire un tableau de Burt pertinent pour rendre compte de ces données. Vous utiliserez le tableau fourni. Attention, sa dimension n'est peut-être pas correcte !!

Pays	continent	producteur de pétrole	type de pays	
Japon	Asie	Non	x Développé	Asie
Finlande	Europe	Non	x Développé	EU
Russie	Europe	Oui	/ NPI	EU
Venezuela	Amérique Latine	Oui	/ PED	AL
Tunisie	Afrique	Non	x PED	AF
Mexique	Amérique Latine	Oui	/ NPI	AL
Chine	Asie	Oui	/ NPI	AS
Haïti	Amérique Latine	Non	x PED	AL
Bangladesh	Asie	Non	x PED	AS
Angola	Afrique	Oui	/ PED	AF

Tableau de BURT

	Asie	Europe	Amérique	Afrique	Pét.O.	Pét.Nm	Dev	NPI	PED		
Asie	3	0	0	0	1	2	1	1	1		
Europe		2	0	0	1	1	1	1	0		
Amérique			3	0	2	1	0	1	2		
Afrique				2	1	1	0	0	2		
Pét.O.					5	0	0	3	2		
Pét.Nm						5	2	0	3		
Dev							2	0	0		
NPI								3	0		
PED									5		

Exercice n°3 : (4 points)

On effectue une ACM sur une population de 1000 individus ; sur 10 variables et 30 modalités.

1. Si on effectue l'AFC à partir du tableau disjonctif complet, quelle est l'inertie totale ?

Inertie totale : nombre de modalités variables

2. Quel est le nombre de valeurs propres que l'on doit obtenir ?

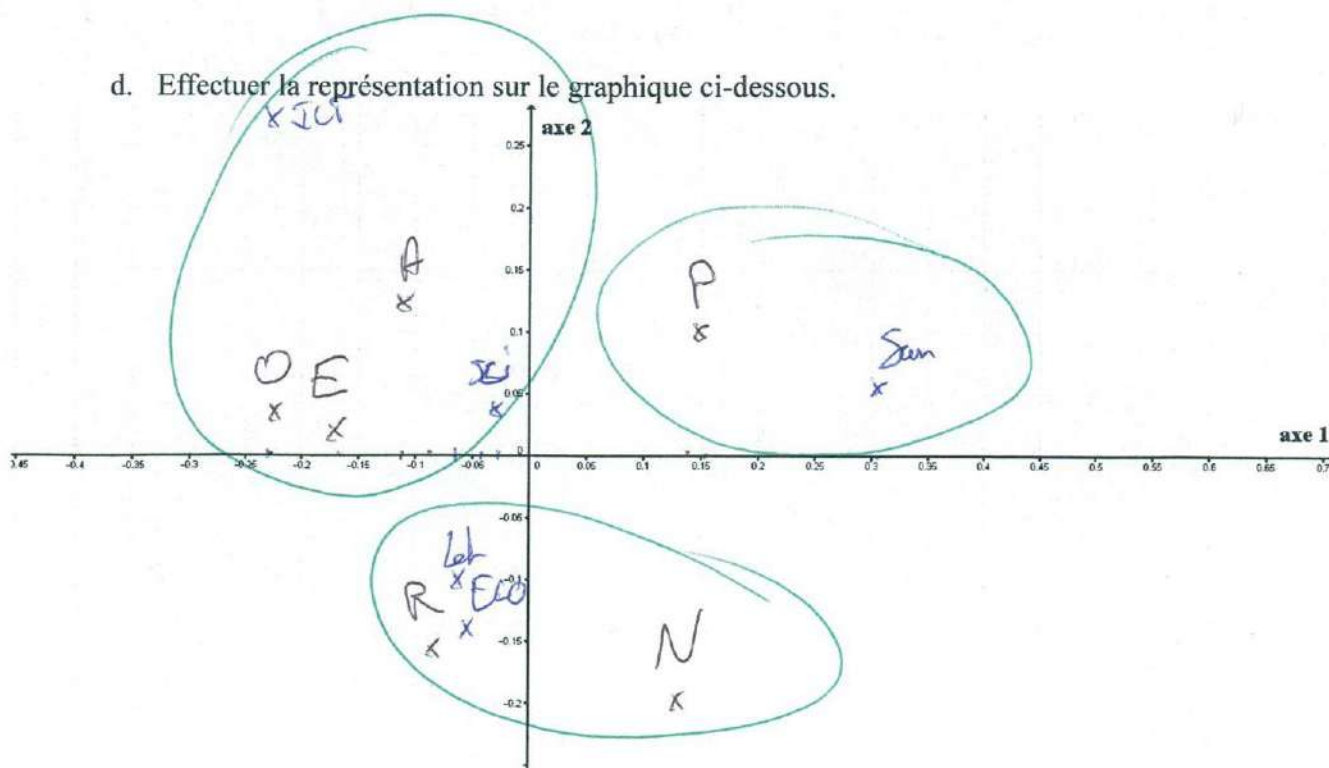
On doit obtenir 10 valeurs propres

	Axe 1			Axe 2			7th cumulé
	coord.	qlt	cr	coord.	qlt	cr	
Droit	0,035	0,046	0,009	0,079	0,233	0,065	0,279
Eco	-0,031	0,033	0,007	-0,140	0,657	0,197	0,690
Let	-0,068	0,260	0,068	-0,105	0,628	0,243	0,888
Sci	-0,027	0,195	0,008	0,044	0,502	0,029	0,697
San	0,305	0,950	0,682	0,052	0,027	0,029	0,977
IUT	-0,230	0,424	0,227	0,261	0,547	0,436	0,971

c. Quels profils doit on sélectionner ?

Il faut choisir les profils avec une qlt importante $c-a-d > 0,65$

d. Effectuer la représentation sur le graphique ci-dessous.



11. Finir l'interprétation de ces données.

On peut distinguer 3 groupes:

- Le groupe P-Sou: Professeurs titulaires - cadres supérieurs et Santé
- Le groupe O-E-A-Si-IUT: Ouvriers, employés, agriculteurs et Sciences, IUT
- Le groupe R-N-Let-Eco: Retraités, Non Renseignés et Lettres - Économie

On remarque en effet un lien entre la catégorie socio-professionnelle du père et le domaine d'étude de l'enfant.

Les pères retraités vont avoir tendance à ce que leurs enfants étudient les lettres et l'économie

Les pères ouvriers, employés ou agriculteurs auront des enfants qui se tourneront plus vers les Sciences ou les IUT qui sont des formations techniques

Les pères cadres supérieurs ou professeurs titulaires auront des enfants qui étudient plus le domaine de la Santé

3. On suppose que les 6 plus grandes valeurs propres sont les suivantes :

0,8	0,5	0,3	0,2	0,1	0,05
-----	-----	-----	-----	-----	------

- a. Calculer la part d'inertie portée par les deux plus grandes valeurs propres ; calculer la part totale d'inertie portée par les valeurs propres manquantes.

- b. Déterminer les six plus grandes valeurs propres de l'AFC effectuée à partir du tableau de Burt. Calculer alors l'inertie totale et la part d'inertie portée par les deux plus grandes valeurs propres.

Exercice n°4 :

Les tableaux de données figurant ci-dessous indiquent les effectifs et les fréquences d'étudiants dans les principales filières de l'enseignement supérieur en France en 2009-2010 en fonction de la catégorie socioprofessionnel du père. L'effectif total des étudiants est de 1 386 065.

EFFECTIFS	Droit	Economie	Lettres	Sciences	Santé	IUT
Agriculteurs	2881	3299	6088	5778	3003	3130
Artisans, commerçants, chefs d'entreprise	16141	15884	24627	19233	12232	10501
Professions libérales, cadres supérieurs	67395	45903	102119	82953	80692	33026
Professions intermédiaires	19398	17971	51178	37786	23985	19269
Employés	22479	21463	50725	33588	15306	17733
Ouvriers	16181	18927	40679	27815	10392	16766
Retraités, inactifs	28118	28388	66877	35671	18139	10916
Non renseigné	20894	34212	64847	37350	37329	6798

FREQUENCES	Droit	Economie	Lettres	Sciences	Santé	IUT	Total
Agriculteurs	0,21%	0,24%	0,44%	0,42%	0,22%	0,23%	1,74%
Artisans, commerçants, chefs d'entreprise	1,16%	1,15%	1,78%	1,39%	0,88%	0,76%	7,11%
Professions libérales, cadres supérieurs	4,86%	3,31%	7,37%	5,98%	5,82%	2,38%	29,73%
Professions intermédiaires	1,40%	1,30%	3,69%	2,73%	1,73%	1,39%	12,24%
Employés	1,62%	1,55%	3,66%	2,42%	1,10%	1,28%	11,64%
Ouvriers	1,17%	1,37%	2,93%	2,01%	0,75%	1,21%	9,43%
Retraités, inactifs	2,03%	2,05%	4,82%	2,57%	1,31%	<u>0,79%</u>	13,57%
Non renseigné	1,51%	2,47%	4,68%	2,69%	2,69%	0,49%	14,53%
Total	13,96%	13,42%	29,37%	20,21%	14,51%	8,52%	100,00%

Vous répondrez aux questions suivantes en argumentant à partir des documents fournis, des représentations graphiques que vous serez amenés à faire et des connaissances du cours.

1. Quelle est la problématique de l'étude ?

L'influence de la catégorie socio-professionnelle du père sur le domaine d'étude de ses enfants.

2. Les tableaux suivants sont les tableaux des profils : l'un est celui des PFL, l'autre celui des PFC. Après avoir rappelé ce que sont les PFL et les PFC vous identifierez chacun de ces tableaux en expliquant vos raisons et vous indiquerez les PFC ou PFL moyens dans le tableau correspondant.

Prof: PFL: Fréquence de la variable catégorie socio-professionnelle dans le domaine d'étude
Prof: PFC: Tercenaire

PROFIL <i>PFL</i>	Droit	Economie	Lettres	Sciences	Santé	IUT	
Agriculteurs	1,49%	1,77%	1,50%	2,06%	1,49%	2,65%	17,6%
Artisans, commerçants, chefs d'entreprise	8,34%	8,54%	6,05%	6,86%	6,08%	8,89%	7,11%
Professions libérales, cadres supérieurs	34,83%	24,67%	25,08%	29,61%	40,13%	27,96%	29,73%
Professions intermédiaires	10,03%	9,66%	12,57%	13,49%	11,93%	16,31%	12,24%
Employés	11,62%	11,54%	12,46%	11,99%	7,61%	15,01%	11,64%
Ouvriers	<u>8,36%</u>	10,17%	9,99%	9,93%	5,17%	14,19%	9,43%
Retraités, inactifs	14,53%	15,26%	16,43%	12,73%	9,02%	9,24%	13,57%
Non renseigné	10,80%	18,39%	15,93%	13,33%	18,56%	5,75%	14,53%
	100%	100%	100%	100%	100%	100%	100%

PROFIL	Droit	Economie	Lettres	Sciences	Santé	IUT	
Agriculteurs	11,92%	13,64%	<u>25,18%</u>	23,90%	12,42%	12,95%	100%
Artisans, commerçants, chefs d'entreprise	16,37%	16,11%	24,97%	19,50%	12,40%	10,65%	100%
Professions libérales, cadres supérieurs	16,35%	11,14%	24,78%	20,13%	19,58%	8,01%	100%
Professions intermédiaires	11,44%	10,60%	30,18%	22,28%	14,14%	11,36%	100%
Employés	13,94%	13,31%	31,45%	20,82%	9,49%	10,99%	100%
Ouvriers	12,37%	14,47%	31,11%	21,27%	7,95%	12,82%	100%
Retraités, inactifs	14,95%	15,09%	35,55%	18,96%	9,64%	5,80%	100%
Non renseigné	10,37%	16,98%	32,19%	18,54%	18,53%	3,37%	100%
	13,96%	13,62%	29,37%	20,21%	16,51%	8,52%	100%

3. Quelle est la signification des nombres soulignés et grisés dans les 3 tableaux précédents ?

0,79 : 0,79% des étudiants sont en IUT et ont un père retraité ou inactif

8,36 : 8,36% des étudiants ont un père ouvrier

25,18 : 25,18% des pères agriculteurs ont des enfants en lettres

4. Faire une analyse des PFC et des PFL. Vous pourrez utiliser les tableaux précédents et faire une synthèse.

PFC : - Une 30% des étudiants en études supérieures ont un père cadre supérieur ou professionnel.
 - 7% ont un père artisan, commerçant ou chef d'entreprise.

PFL : - 30% des étudiants en études supérieures ont un père en lettres
 - 8,52% sont en IUT

5. Va-t-on effectuer l'AFC à partir des PFL ou à partir des PFC ? Argumenter.

À partir des PFC car il y a plus de lignes que de colonnes.

6. Quelle est le nombre de valeurs propres que l'on va obtenir ?

On aura 6 valeurs propres dont la valeur propre la plus grande est égale à 1

7. Les 4 plus grandes valeurs propres sont 0,020; 0,013; 0,003 et 0,002. Effectuer un test d'indépendance de ces données. Que peut-on en déduire?

χ^2	$\chi^2_{0,995}$	$\chi^2_{0,99}$	$\chi^2_{0,975}$	$\chi^2_{0,95}$	$\chi^2_{0,90}$	$\chi^2_{0,75}$	$\chi^2_{0,50}$	$\chi^2_{0,25}$	$\chi^2_{0,10}$	$\chi^2_{0,05}$	$\chi^2_{0,025}$	$\chi^2_{0,01}$	$\chi^2_{0,005}$
30	53,7	50,9	47,0	43,8	40,3	34,8	29,3	24,5	20,6	18,5	16,8	15,0	13,8
40	66,8	63,7	59,3	55,8	51,8	45,6	39,3	33,7	29,1	26,5	24,4	22,2	20,7
50	79,5	76,2	71,4	67,5	63,2	56,3	49,3	42,9	37,7	34,8	32,4	29,7	28,0
60	92,0	88,4	83,3	79,1	74,4	67,0	59,3	52,3	46,5	43,2	40,5	37,5	35,5

$$\chi^2 = NR \approx 6 \cdot 10^2 \times 1,6 \cdot 10^6 = 5,6 \cdot 10^4 \gg \chi^2_{0,90}$$

$$R = 0,027 + 0,013 + 0,003 + 0,002$$

$$= 0,038$$

$$N = 1386065$$

Donc il n'y a pas d'indépendance entre le domaine d'études et la

$$V = (p-1)(q-1) = (6-1)(8-1) = 5 \times 7 = 35$$

8. Justifier le choix des deux premiers axes.

$$q_1 = \frac{0,02}{0,038} \times 100 \approx 50\% \quad q_2 = \frac{0,013}{0,038} \times 100 \approx 30\%$$

On a 80% d'informations sur l'axe 1 et 2

9. Quelle est la différence entre « q1 » et « cr » ? A quoi servent ces paramètres ?

La q1 sert à vérifier qu'un point est bien placé par rapport à l'axe

La cr sert à calculer la contribution de chaque point à l'axe donc son importance dans l'axe

10. On donne ci-dessous les composantes principales des profils.

		Axe 1			Axe 2			
		coord.	q1	cr	coord.	q1	cr	
Agriculteurs	A	-0,110	0,291	0,011	0,126	0,381	0,021	0,672
Artisans, commerçants, chefs d'entreprise		-0,058	0,135	0,012	0,060	0,145	0,019	0,280
Professions libérales, cadres supérieurs	P	0,152	0,691	0,345	0,097	0,282	0,209	0,973
Professions intermédiaires		-0,062	0,172	0,024	0,080	0,287	0,059	0,459
Employés	E	-0,160	0,979	0,151	0,018	0,012	0,003	0,991
Ouvriers	O	-0,229	0,945	0,250	0,032	0,019	0,007	0,964
Retraités, inactifs	R	-0,090	0,193	0,055	-0,158	0,600	0,254	0,793
Non renseigné	N	0,144	0,320	0,153	-0,198	0,603	0,428	0,923