

# DS d'analyse de données

mardi 14 mai 2024 — durée : 2 heures — documents **non autorisés**

## 1 AFC : Profession des candidats aux législatives (total 6 points)

On s'intéresse aux données publiées par le journal *Libération* du 13 juin 2017 concernant les candidats au 1<sup>er</sup> tour des élections législatives de 2017. Pour cela, on dispose d'un tableau recensant la profession des 7881 candidats à l'élection, croisée avec leur affiliation politique.

Les affiliations politiques sont, selon la terminologie du ministère de l'intérieur et plus ou moins de la gauche vers la droite : extrême gauche (EXG), parti communiste français (COM), la France insoumise (FI), parti socialiste (SOC), parti radical de gauche (RDG), divers gauche (DVG), écologistes (ECO), divers (DIV), régionaliste (REG), la république en marche (REM), Modem (MDM), union des démocrates et indépendants (UDI), les républicains (LR), divers droite (DVD), debout la France (DLF), front national (FN), extrême droite (EXD).

Les professions retenues sont : permanent politique (perm), avocat (avoc), commerçant (comm), ouvrier (ouvr), secteur agricole (agri), étudiant (etud), ingénieur (inge), secteur médical (medi), enseignement (prof), chef d'entreprise (chef), fonction publique (fonc), sans profession (sans), cadres (cadr), employés du privé (empl), autres (autr), retraités (retr).

### 1.1 Premier regard sur les données (2 points)

On donne ci-dessous le tableau de contingence des candidats, la somme des contributions au  $\chi^2$  par affiliation et le  $\chi^2$  total.

Table de contingence (avec totaux)

	EXG	COM	FI	SOC	RDG	DVG	ECO	DIV	REG	REM	MDM	UDI	LR	DVD	DLF	FN	EXD	TOT
perm	0	15	1	17	1	2	8	5	1	8	0	6	26	9	0	36	1	136
avoc	0	3	5	14	1	6	8	18	2	25	1	6	37	17	7	10	3	163
comm	1	0	2	2	2	2	12	41	3	9	0	0	4	10	9	18	3	118
ouvr	68	8	7	0	0	6	7	22	2	1	0	0	0	0	3	7	4	135
agri	3	1	7	5	1	7	11	14	3	10	1	6	13	9	5	9	4	109
etud	2	11	27	2	2	15	18	120	7	1	0	0	3	20	17	16	6	267
inge	9	5	18	10	1	13	36	91	3	13	4	6	9	23	12	10	7	270
medi	11	5	11	14	5	12	25	45	6	20	9	4	29	20	16	15	7	254
prof	173	68	109	55	7	40	111	110	13	54	6	7	29	42	15	23	7	869
chef	0	1	7	10	2	17	27	72	5	43	6	20	49	58	18	33	8	376
fonc	130	103	83	98	11	65	122	130	22	70	9	20	45	56	43	54	11	1072
sans	3	9	32	19	2	14	45	114	8	13	3	8	26	32	30	41	12	411
cadr	9	21	36	64	10	40	81	131	11	83	15	27	70	82	35	54	12	781
empl	54	58	63	10	2	25	88	155	16	12	2	4	9	29	35	75	28	665
autr	44	29	71	54	11	51	151	257	20	77	11	20	91	95	71	97	33	1183
retr	157	124	77	40	4	60	161	93	27	22	8	14	40	63	73	73	36	1072
TOT	664	461	556	414	62	375	911	1418	149	461	75	148	480	565	389	571	182	7881

Décomposition du  $\chi^2$  par affiliation

chi2
EXG 694.8
COM 204.6
FI 89.7
SOC 113.0
RDG 16.1
DVG 16.1
ECO 44.6
DIV 313.1
REG 10.0
REM 152.3
MDM 38.9
UDI 72.4
LR 250.2
DVD 84.9
DLF 48.3
FN 141.8
EXD 38.7

$\chi^2$  pour les données complètes

chi2
TOT 2329.4

**Question 1** On s'intéresse aux partis *REM*, *MDM* et *UDI* et on cherche à valider à l'aide d'un test du  $\chi^2$  si les profils de leurs candidats sont différents. La table de contingence correspondante a un  $\chi^2$  égal à 43.63. Quelle conclusion peut-on tirer (à 1%, puis à 5%) ? On pourra se référer à la table en fin de sujet.

**Question 2** Donnez un exemple de métier pour lequel les candidats *EXG* sont très différents des autres. Quelle est, en pourcentage, la contribution de *EXG* au  $\chi^2$  ? Expliquez quelle peut être la conséquence de cette valeur sur l'analyse par analogie avec ce que vous avez vu à propos de l'ACP et l'ACM et en expliquant le lien entre le  $\chi^2$  et l'AFC.

On décide par la suite de retirer les candidats *EXG* de l'analyse.

## 1.2 Analyse fonctionnelle des correspondances (4 points)

On réalise une AFC des données *hors candidats EXG* et on obtient les données suivantes : les 6 premières valeurs propres, et — sur les 2 premiers axes — les coordonnées des métiers et des affiliations, ainsi que les qualités de représentation par les axes.

Valeurs propres

	Axe1	Axe2	Axe3	Axe4	Axe5	Axe6
lambda	0.0865	0.0636	0.0282	0.0134	0.0059	0.0049

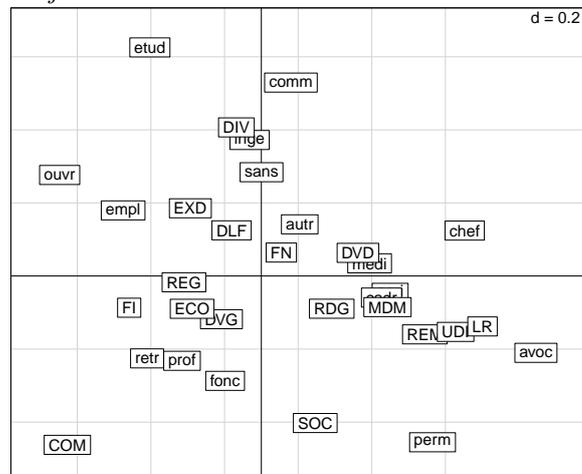
Coordonnées

	Axe1	Axe2
perm	0.47	-0.46
avoc	0.75	-0.21
comm	0.08	0.53
ouvr	-0.55	0.28
agri	0.35	-0.05
etud	-0.30	0.63
inge	-0.03	0.37
medi	0.29	0.03
prof	-0.22	-0.23
chef	0.55	0.12
fonc	-0.10	-0.29
sans	0.00	0.28
cadr	0.33	-0.06
empl	-0.38	0.18
autr	0.10	0.14
retr	-0.31	-0.23

Qualité de représentation

	Axe1	Axe2	Axe1	Axe2
perm	18.8	18.0	COM	52.4 40.3
avoc	78.1	6.2	FI	60.3 3.5
comm	1.5	65.6	SOC	7.8 59.0
ouvr	58.0	14.8	RDG	15.5 3.4
agri	56.9	1.0	DVG	18.5 22.1
etud	16.3	69.6	ECO	53.1 12.1
inge	0.5	66.3	DIV	2.5 89.7
medi	45.8	0.6	REG	54.7 0.4
prof	22.9	26.6	REM	68.9 8.9
chef	80.3	4.1	MDM	25.4 1.6
fonc	7.9	67.2	UDI	67.8 5.7
sans	0.0	77.0	LR	83.0 4.4
cadr	82.8	2.7	DVD	66.2 3.8
empl	64.7	14.8	DLF	6.0 14.3
autr	27.2	50.1	FN	1.4 2.0
retr	44.9	23.7	EXD	16.6 15.3

Projection sur les axes



**Question 3** Expliquez pourquoi on peut se limiter au premier plan principal. Combien d'axes propres y avait-il en tout ? Le  $\chi^2$  de la table sur laquelle l'AFC est faite vaut 1547, quelle qualité globale de représentation obtient-on avec ces deux axes ?

**Question 4** Quelles sont les modalités de ligne et de colonne qui déterminent le premier plan principal ? On précisera les critères utilisés. Faire un commentaire rapide des axes.

**Question 5** Quels sont les affiliations et les métiers qui sont mal représentés par l'analyse ? Commentez le cas de *perm*.

## 2 ACM : histoires de vies 2003 (total 9 points)

De février à avril 2003, l'Insee a réalisé une enquête sur la construction des identités, appelée « histoire de vie », pour laquelle des personnes de 18 ans et plus ont été interrogées. On considère ici 1880 personnes et 5 variables.

Les variables utilisées sont

- sexe (*sex*) : femme (*f*), homme (*h*) ;
- niveau d'études (*etud*) : aucun (*non*), primaire (*prim*), 1<sup>er</sup> cycle (*coll*), 2<sup>e</sup> cycle (*lycee*), technique ou professionnel court (*techc*) ou long (*techl*), supérieur (*sup*) ;
- occupation (*occup*) : exerçant une profession (*prof*), chômeur (*chom*), retraité (*retr*), au foyer (*foyer*), autre inactif (*inac*) ;
- qualification (*qual*) : ouvrier qualifié (*ouvq*) ou spécialisé (*ouvs*), technicien (*tech*), employé (*empl*), profession intermédiaire (*inter*), cadre (*cadr*), autre (*autr*), sans objet (NA) ;
- satisfaction au travail (*satis*) : oui (*oui*), non (*non*), équilibré (*equ*), sans objet (NA).

### 2.1 Étude rapide des données (1,5 points)

On donne ci-dessous un tableau de Burt partiel (variables *qual* et *satis*) avec des valeurs manquantes (marquées NA).

	qual.autr	qual.cadr	qual.empl	qual.inter	qual.NA	qual.ouvq	qual.ouvs	qual.tech	satis.equ	satis.NA	satis.non	satis.oui
qual.autr	51	0	0	0	0	0	0	0	10	20	1	20
qual.cadr	0	256	0	0	0	0	0	0	58	88	12	98
qual.empl	0	0	583	0	0	0	0	0	151	254	35	143
qual.inter	0	0	0	160	0	0	0	0	43	53	11	53
qual.NA	0	0	0	0	257	0	0	0	41	162	7	47
qual.ouvq	0	0	0	0	0	288	0	0	71	129	24	64
qual.ouvs	0	0	0	0	0	0	199	0	37	123	15	24
qual.tech	0	0	0	0	0	0	0	NA	NA	22	10	26
satis.equ	10	58	151	43	41	71	37	NA	NA	NA	0	0
satis.NA	20	88	254	53	162	129	123	22	NA	NA	0	0
satis.non	1	12	35	11	7	24	15	10	0	0	115	0
satis.oui	20	98	143	53	47	64	24	26	0	0	0	475

**Question 6** Calculez les valeurs manquantes du tableau de Burt. On détaillera au moins certains calculs pour montrer comment on fait.

**Question 7** à partir de la description des données, quelle hypothèse pouvez vous faire sur les personnes pour lesquelles *satis=NA* ?

## 2.2 Analyse des correspondances multiples (6 points)

On réalise une ACM sur les données complètes. On fournit ci-dessous les valeurs propres ainsi que, pour 3 axes, les coordonnées sur les axes, les contributions aux axes et la qualité de la représentation par les sous-espaces factoriels (en % pour ces deux derniers) pour les catégories.

Valeurs propres	Coordonnées			Contributions aux axes			Qualités de représentation					
	Comp1	Comp2	Comp3	Axis1	Axis2	Axis3	Axis1	Axis1:2	Axis1:3			
0.49	sex.f	-0.21	0.68	0.25	sex.f	1.0	15.8	2.3	sex.f	5.3	61.8	69.6
0.32	sex.h	0.26	-0.84	-0.31	sex.h	1.2	19.6	2.9	sex.h	5.3	61.8	69.6
0.30	etud.coll	-0.43	0.08	0.54	etud.coll	0.8	0.0	2.1	etud.coll	2.3	2.4	5.9
0.23	etud.lycee	0.23	0.58	0.12	etud.lycee	0.2	2.1	0.1	etud.lycee	0.6	4.2	4.4
0.23	etud.non	-1.28	-0.15	-0.27	etud.non	1.4	0.0	0.1	etud.non	3.5	3.5	3.7
0.22	etud.prim	-1.04	-0.29	-0.16	etud.prim	9.9	1.2	0.4	etud.prim	31.6	34.1	34.8
0.21	etud.sup	0.75	0.61	-1.09	etud.sup	5.3	5.4	18.5	etud.sup	16.9	28.0	64.3
0.21	etud.techc	0.29	-0.54	0.79	etud.techc	0.9	4.5	10.2	etud.techc	2.8	12.3	32.6
0.20	etud.techl	0.59	-0.07	0.46	etud.techl	1.0	0.0	1.0	etud.techl	2.6	2.7	4.2
0.20	occup.chom	-0.76	0.12	0.49	occup.chom	1.6	0.1	1.1	occup.chom	4.3	4.4	6.2
0.20	occup.foyer	-1.18	1.27	0.36	occup.foyer	5.1	9.2	0.8	occup.foyer	13.9	30.1	31.4
0.19	occup.inac	-1.24	-0.40	0.40	occup.inac	2.7	0.4	0.5	occup.inac	7.1	7.8	8.5
0.19	occup.prof	0.83	0.00	0.19	occup.prof	15.3	0.0	1.3	occup.prof	83.5	83.5	87.6
0.18	occup.retr	-0.97	-0.43	-0.75	occup.retr	9.5	2.9	9.3	occup.retr	30.9	37.0	55.6
0.18	qual.autr	0.22	0.63	0.09	qual.autr	0.1	0.7	0.0	qual.autr	0.1	1.2	1.3
0.17	qual.cadr	0.73	0.44	-1.67	qual.cadr	2.9	1.7	25.3	qual.cadr	8.4	11.5	55.6
0.17	qual.empl	-0.05	0.73	0.71	qual.empl	0.0	10.4	10.3	qual.empl	0.1	24.2	46.6
0.14	qual.inter	0.57	-0.03	-0.45	qual.inter	1.1	0.0	1.2	qual.inter	3.1	3.1	5.0
0.10	qual.NA	-0.65	0.01	-0.40	qual.NA	2.4	0.0	1.5	qual.NA	6.8	6.8	9.4
0.07	qual.ouvq	-0.03	-1.48	0.39	qual.ouvq	0.0	21.1	1.5	qual.ouvq	0.0	39.7	42.4
0.00	qual.ouvs	-0.75	-0.36	0.39	qual.ouvs	2.4	0.9	1.1	qual.ouvs	6.7	8.3	10.1
	qual.tech	0.80	-0.84	0.00	qual.tech	1.2	2.0	0.0	qual.tech	3.1	6.5	6.5
	satis.equ	0.78	-0.13	0.54	satis.equ	5.7	0.2	4.5	satis.equ	18.4	18.8	27.6
	satis.NA	-1.00	0.00	-0.23	satis.NA	18.5	0.0	1.5	satis.NA	83.3	83.3	87.6
	satis.non	0.75	-0.47	0.65	satis.non	1.4	0.8	1.7	satis.non	3.7	5.1	7.8
	satis.oui	0.90	0.23	-0.25	satis.oui	8.3	0.8	1.0	satis.oui	27.3	29.0	31.1

**Question 8** Parmi les données ci-dessus, quelles sont celles pour lesquelles vous pouvez donner les sommes en colonne sans effectuer l'addition ?

**Question 9** Combien d'axes propres conseillez-vous de conserver ? Quelle proportion de l'inertie est expliquée par le sous-espace propre correspondant ?

**Question 10** D'une manière générale, quelle est la différence d'usage entre la contribution d'un individu à un axe et la qualité de représentation d'un individu pas un axe ? On expliquera aussi ce qui relie ces deux notions

**Question 11** Quelles sont les catégories qui déterminent les 3 premiers axes principaux ? (on détaillera les critères et on cherchera à être précis dans la réponse).

**Question 12** Quelles sont les catégories qui sont bien représentées sur le sous-espace principal (1, 2, 3) ? Expliquez pourquoi il est logique que ces variables soient bien représentées ensemble.

**Question 13** Comment peut-on décrire et interpréter les 3 premiers axes ?

## 2.3 Catégories supplémentaires (1,5 points)

On cherche à préciser les caractéristiques des axes en termes d'activités ou loisirs. La question posée est : « en dehors du cadre scolaire ou professionnel, au cours des 12 derniers mois, avez-vous pratiqué alors que vous n'y étiez pas obligé-e, l'activité... ». Les activités retenues (parmi une liste bien plus grande) sont arbitrairement : écouter du hard-rock (**hardrock**), lire des bandes dessinées (**lecture.bd**), aller à la pêche ou à la chasse (**peche.chasse**), faire la cuisine pour le plaisir (**cuisine**), bricoler (**brico**), aller au cinéma (**cinema**) et faire du sport (**sport**). Par ailleurs on mesure le nombre d'heures passées par jour à regarder la télévision dans la variable quantitative **heures.tv**.

Les effectifs des catégories supplémentaires et les valeurs test correspondantes sont données ci-dessous, ainsi que les corrélations de **heures.tv** avec les composantes.

Effectifs		Valeurs test			Corrélation avec les composantes				
	Eff		Axis1	Axis2	Axis3	Axis1	Axis2	Axis3	
hardrock.non	1868	hardrock.non	-1.05	-0.28	-1.36	heures.tv	-0.33	-0.09	0.09
hardrock.oui	12	hardrock.oui	1.05	0.28	1.36				
lecture.bd.non	1837	lecture.bd.non	-4.29	-4.04	4.46				
lecture.bd.oui	43	lecture.bd.oui	4.29	4.04	-4.46				
peche.chasse.non	1666	peche.chasse.non	-1.91	9.18	0.22				
peche.chasse.oui	214	peche.chasse.oui	1.91	-9.18	-0.22				
cuisine.non	1054	cuisine.non	-0.75	-8.97	-5.73				
cuisine.oui	826	cuisine.oui	0.75	8.97	5.73				
bricol.non	1062	bricol.non	-11.18	8.14	2.36				
bricol.oui	818	bricol.oui	11.18	-8.14	-2.36				
cinema.non	1150	cinema.non	-17.76	-8.49	2.21				
cinema.oui	730	cinema.oui	17.76	8.49	-2.21				
sport.non	1246	sport.non	-13.91	-3.34	4.69				
sport.oui	634	sport.oui	13.91	3.34	-4.69				

**Question 14** Justifiez l'utilisation des valeurs-test. Quelles catégories supplémentaires sont significatives sur les 3 premiers axes ? Comment les interpréter ?

**Question 15** Que peut-on dire de la variable *heures.tv* ? Quel conseil donneriez-vous pour essayer d'améliorer l'analyse de cette variable ?

### 3 Qualité de représentation et valeurs-tests pour les variables à deux modalités (5 points)

Les questions 17 et 18 sont indépendantes.

On s'intéresse dans le cadre d'une analyse des correspondances multiples à une variable à deux modalités (ou catégories), que l'on notera 1 et 2. On note  $n_1$  et  $n_2$  les effectifs de ces catégories, et  $a_{1k}$  et  $a_{2k}$  leurs coordonnées sur l'axe  $k$ . Que les variables soient actives ou supplémentaires, ces coordonnées de catégories peuvent s'écrire, pour  $j = 1, 2$

$$a_{jk} = \frac{1}{\sqrt{\mu_k}} \frac{1}{n_j} \sum_{i \text{ dans cat. } j} c_{ik},$$

où la somme s'effectue sur tous les individus de catégorie  $j$ . Comme d'habitude,  $\mu_k$  est la valeur propre associée à l'axe  $k$  et  $c_{ik}$  la coordonnée de l'individu  $i$  sur l'axe  $k$ . On rappelle que chaque vecteur  $\mathbf{c}_k$  est centré.

**Question 16** Montrer que, pour chaque  $k$ ,

$$n_1 a_{1k} + n_2 a_{2k} = 0.$$

**Question 17** En déduire que, dans le cas d'une variable active, les qualités de représentation des catégories 1 et 2 par un axe sont égales. On rappelle que la qualité de représentation de la catégorie  $j$  par l'axe  $k$  est

$$\frac{a_{jk}^2}{a_{j1}^2 + a_{j2}^2 + \dots + a_{jq}^2}.$$

**Question 18** Montrer que, dans le cas d'une variable supplémentaire, les valeurs tests associées aux catégories 1 et 2 sont elles aussi égales en valeur absolues (mais opposées en signe). On rappelle que la valeur-test associée à la catégorie  $j$  sur l'axe  $k$  est

$$a_{jk} \sqrt{n_j} \frac{\sqrt{n-1}}{\sqrt{n-n_j}},$$

TABLE DU CHI-DEUX :  $\chi^2(n)$



n <sup>p</sup>	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.02	0.01
1	0,0158	0,0642	0,148	0,455	1,074	1,642	2,706	3,841	5,412	6,635
2	0,211	0,446	0,713	1,386	2,408	3,219	4,605	5,991	7,824	9,210
3	0,584	1,005	1,424	2,366	3,665	4,642	6,251	7,815	9,837	11,341
4	1,064	1,649	2,195	3,357	4,878	5,989	7,779	9,488	11,668	13,277
5	1,610	2,343	3,000	4,351	6,064	7,289	9,236	11,070	13,388	15,086
6	2,204	3,070	3,828	5,348	7,231	8,558	10,645	12,592	15,033	16,812
7	2,833	3,822	4,671	6,346	8,383	9,803	12,017	14,067	16,622	18,475
8	3,490	4,594	5,527	7,344	9,524	11,030	13,362	15,507	18,168	20,090
9	4,168	5,380	6,393	8,343	10,656	12,242	14,684	16,919	19,679	21,666
10	4,865	6,179	7,267	9,342	11,781	13,442	15,987	18,307	21,161	23,209
11	5,578	6,989	8,148	10,341	12,899	14,631	17,275	19,675	22,618	24,725
12	6,304	7,807	9,034	11,340	14,011	15,812	18,549	21,026	24,054	26,217
13	7,042	8,634	9,926	12,340	15,119	16,985	19,812	22,362	25,472	27,688
14	7,790	9,467	10,821	13,339	16,222	18,151	21,064	23,685	26,873	29,141
15	8,547	10,307	11,721	14,339	17,322	19,311	22,307	24,996	28,259	30,578
16	9,312	11,152	12,624	15,338	18,418	20,465	23,542	26,296	29,633	32,000
17	10,085	12,002	13,531	16,338	19,511	21,615	24,769	27,587	30,995	33,409
18	10,865	12,857	14,440	17,338	20,601	22,760	25,989	28,869	32,346	34,805
19	11,651	13,716	15,352	18,338	21,689	23,900	27,204	30,144	33,687	36,191
20	12,443	14,578	16,266	19,337	22,775	25,038	28,412	31,410	35,020	37,566
21	13,240	15,445	17,182	20,337	23,858	26,171	29,615	32,671	36,343	38,932
22	14,041	16,314	18,101	21,337	24,939	27,301	30,813	33,924	37,659	40,289
23	14,848	17,187	19,021	22,337	26,018	28,429	32,007	35,172	38,968	41,638
24	15,659	18,062	19,943	23,337	27,096	29,553	33,196	36,415	40,270	42,980
25	16,473	18,940	20,867	24,337	28,172	30,675	34,382	37,652	41,566	44,314
26	17,292	19,820	21,792	25,336	29,246	31,795	35,563	38,885	42,856	45,642
27	18,114	20,703	22,719	26,336	30,319	32,912	36,741	40,113	44,140	46,963
28	18,939	21,588	23,647	27,336	31,391	34,027	37,916	41,337	45,419	48,278
29	19,768	22,475	24,577	28,336	32,461	35,139	39,087	42,557	46,693	49,588
30	20,599	23,364	25,508	29,336	33,530	36,250	40,256	43,773	47,962	50,892

Pour  $n > 30$ , on peut admettre que  $\sqrt{2\chi^2} - \sqrt{2n-1} \approx N(0,1)$