DS d'analyse de données (Correction)

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

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

```
> require(ade4)
> source("fonctions.R")
> met=read.csv("metiers-candidats.csv", skip=2)
> met1=met[,-1]
> udirem=t(met[,c("REM", "MDM", "UDI")])
> metc=stats.conting(met)
> met1c=stats.conting(met1)
> nf=2
> coa1=dudi.coa(met1, scannf=F, nf=nf)
> Ig=sum(coa1$eig)
> #coa1=dudi.fixsigns(coa1, sign.co=c(1, -1, 1, 1))
> inert1=inertia.dudi(coa1, col=T, row=T)
```

On s'intéresse aux données publiées par le journal *Libération* du 13 juin 2017 concernant les candidats au 1er 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 χ^2 par affiliation et le χ^2 total.

Table de contingence (avec totaux)

```
> #round(metc$pc*100)
> metc$conting
```

	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 χ^2 par affiliation

```
> chi2.aff=matrix(colSums(metc$marge.col*met1c$tab^2)-1, d
> dimnames(chi2.aff)=list(colnames(met1),"chi2")
> round(chi2.aff,1)
```

	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

χ^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 χ^2 si les profils de leurs candidats sont différents. La table de contingence correspondante a un χ^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.*

La table de contingence à laquelle on s'intéresse a 3 colonnes et 16 lines. Son nombre de degrés de liberté est donc $(16-1)\times(3-1)=30$. Dans la table, sous l'hypothèse d'indépendance, on voit que

$$P(\chi^2_{30} > 50,89) = 0.01.$$

On ne peut donc pas dire que la valeur de 43.63 pour le χ^2 soit trop grande à 1%. Si on regarde maintenant à 5%, on voit que

$$P(\chi^2_{30} > 43.77) = 0.05.$$

On est à peu près à la limite ici. Les variables sont dépendantes avec une marge d'erreur un peu supérieure à 5%. Au total on peu déduire que les métiers exercés par les candidats `REM`, `MDM` et `UDI` sont assez peu différents.

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 χ^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 χ^2 et l'AFC.*

Il y a beaucoup d'ouvriers (`ouvr`) parmi les candidats `EXG`, que ce soit en profil-ligne (50% des candidats ouvriers sont d'extrême gauche) ou en profil colonne ($50/664 = 9\%$ des candidats ouvriers sont d'extrême gauche, alors que la moyenne est $135/7881 = 1,7\%$).

La contribution de `EXG` au χ^2 est $694.77 = 29.8\%$ de la valeur totale 2329.41. C'est une très grosse proportion et on peut s'attendre à ce que les premiers axes principaux fassent la part belle à `EXG` et à `ouvr`. En effet, l'inertie totale (la somme des valeurs propres) est égale à $\varphi^2 = \chi^2/n$. De même que pour les individus sur-représentés de l'ACP, une forte contribution à l'inertie totale n'est pas bonne (en général à partir de 25%).

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.

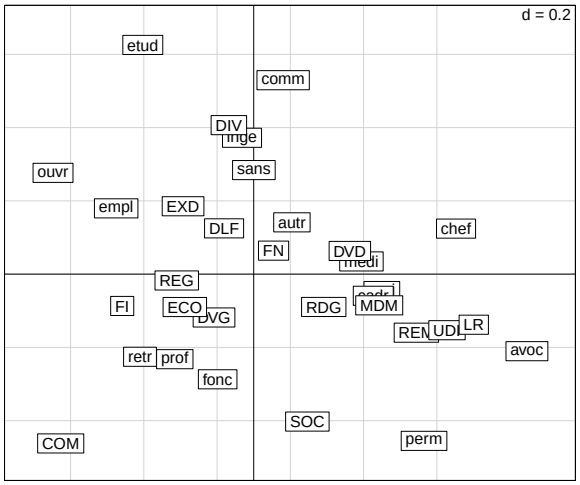
Valeurs propres

```
> neig=6
> eig=matrix(coa1$eig[1:neig],ncol=1)
> colnames(eig)="lambda"
> rownames(eig)=paste0("Axe",1:neig)
> round(t(eig),digits=4)
```

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

Projection sur les axes

```
> s.label(coa1$li)
> s.label(coa1$co, add.p=T)
```



Coordonnées

```
> colnames(coa1$li)=paste0("Axe",1:ncol(coa1$co))
> round(coa1$li,2)
```

	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

```
> round(coa1$co,2)
```

	Axe1	Axe2
COM	-0.53	-0.46
FI	-0.36	-0.09
SOC	0.15	-0.40
RDG	0.19	-0.09
DVG	-0.11	-0.12
ECO	-0.19	-0.09
DIV	-0.07	0.41
REG	-0.21	-0.02
REM	0.44	-0.16
MDM	0.34	-0.09
UDI	0.53	-0.15
LR	0.60	-0.14
DVD	0.26	0.06
DLF	-0.08	0.12
FN	0.05	0.06
EXD	-0.19	0.19

Qualité de représentation

```
> colnames(coa1$co)=paste0("Axe",1:ncol(coa1$co))
> round(abs(inertia.dudi(coa1,row.inertia=T)$row.abs),1)
```

	Axe1	Axe2
perm	18.8	18.0
avoc	78.1	6.2
comm	1.5	65.6
ouvr	58.0	14.8
agri	56.9	1.0
etud	16.3	69.6
inge	0.5	66.3
medi	45.8	0.6
prof	22.9	26.6
chef	80.3	4.1
fonc	7.9	67.2
sans	0.0	77.0
cadr	82.8	2.7
empl	64.7	14.8
autr	27.2	50.1
retr	44.9	23.7

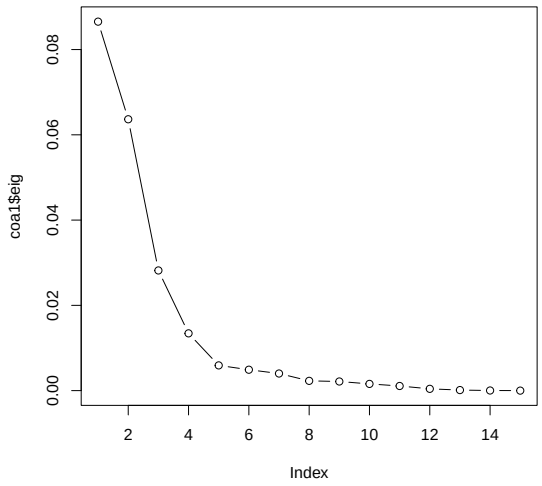
```
> round(abs(inertia.dudi(coa1,col.inertia=T)$col.abs),1)
```

	Axe1	Axe2
COM	52.4	40.3
FI	60.3	3.5
SOC	7.8	59.0
RDG	15.5	3.4
DVG	18.5	22.1
ECO	53.1	12.1
DIV	2.5	89.7
REG	54.7	0.4
REM	68.9	8.9
MDM	25.4	1.6
UDI	67.8	5.7
LR	83.0	4.4
DVD	66.2	3.8
DLF	6.0	14.3
FN	1.4	2.0
EXD	16.6	15.3

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

On représente ci-dessous les valeurs propres :

```
> plot(coa1$eig,t='b')
```



On voit bien une chute des valeurs propres entre la deuxième et la troisième, ainsi qu'un (petit) coude. Il est donc raisonnable de se limiter à deux axes. Il y avait au total $\min(16-1, 16-1) = 15$ axes propres associés à des valeurs propres non nulles.

```
> Ig2=met1c$chi2/(metc$conting["TOT","TOT"]-metc$conting["TOT","EXG"])
> Ig2

[1] 0.2143131
```

On sait que d'inertie totale est χ^2/n, où n est ici l'effectif de la table (tout le monde sauf les `EXG`), c'est-à-dire $7881 - 664 = 7217$. L'inertie expliquée est $0.0865 + 0.0636 = 0.1501$, soit 70% de l'inertie totale $I_{\mathbf{g}} = 1547/7217 = 0.2143$. C'est une qualité plutôt bonne.

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.*

Comme on ne dispose pas des contributions aux axes, on va raisonner ici en fonction des coordonnées. L'analyse étant de qualité moyenne, on se restreindra aux catégories dont la contribution dépasse 2 fois le poids. Comme expliqué dans le cours, on s'intéresse aux axes dont la coordonnée vérifie (pour les lignes, mais les colonnes, c'est pareil) :

$$|a_{ik}| \geq \sqrt{2 \times \lambda_k}$$

Les limites sur les axes seront respectivement 0.4160 et 0.3568. On classe les éléments par coordonnée décroissante (les lignes d'abord).

Axe 1

-	+
`ouvr` (-0.55)	`avoc` (0.75)
`COM` (-0.53)	`chef` (0.55)
[`empl` $(-0,38)$]	`perm` (0.47)
	`LR` (0.60)
	`UDI` (0.53)
	`REM` (0.44)

Axe 2

-	+
`perm` (-0.46)	`etud` (0.63)
`COM` (-0.46)	`comm` (0.53)
`SOC` (-0.40)	`inge` (0.37)
	`DIV` (0.41)

L'axe 1 oppose d'une part les ouvriers et éventuellement les employés, et les candidats communistes (on se souviendra que c'était encore plus vrai de l'extrême gauche, non représentée ici) aux partis de droite et centre droit (UDI, LR, REM), qui sur-représente les avocats, chefs d'entreprise et permanents politiques. C'est une opposition gauche-droite, ou de classe sociale.

L'axe 2, lui, oppose les communistes et les socialistes, qui présentent beaucoup de permanents politiques, avec les candidats divers, chez qui des métiers comme étudiant, commerçant et ingénieur. C'est une opposition professionnel/amateur de la politique (on peut imaginer que les candidats de la partie droite sont moins politisés).

En ce qui concerne les permanents politiques, il y a le souci qu'ils sont opposés au parti communiste sur l'axe 1, et du même côté sur l'axe 2. En fait, on suppose qu'ils sont permanents du parti pour lequel ils se présentent et qu'il n'y a pas vraiment de paradoxe à les voir deux fois (ce ne sont pas les mêmes personnes).

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

On cherche les modalités dont la qualité cumulée (colonne 1 plus colonne 2) est inférieure à 50%, qui est la limite proposée en cours pour une mauvaise représentation. On trouve :

— d'une part `perm` (36.72), `medi` (46.47) et `prof` (49.46) ;

— d'autre part `FN` (3.37), `RDG` (18.94), `DLF` (20.32), `MDM` (26.96), `EXD` (31.91) et enfin `DVG` (40.66).

Parmi ces modalités, perm est notable car il détermine les deux premiers axes. Il y a donc d'autres particularités des permanents politiques que nous ne voyons pas ici. En regardant l'axe 3, on constate en fait qu'il est déterminé par le FN, l'extrême droite et les permanents. Comme remarqué plus haut, chaque parti a ses permanents et ces permanents sont importants là où le parti est important.

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

```
> require(ade4)
> source("fonctions.R")
> options(width=1000)
> hdv2003=read.table("hdv2003.txt", na.string="xx", stringsAsFactors=T)
> hdv0=droplevels(hdv2003[hdv2003$etud != "NA",])
> hdv0=droplevels(hdv0[hdv0$occup != "etud",])
> hdv0=droplevels(hdv0[hdv0$heures.tv != "NA",])
> hdv0$heures.tv=as.numeric(as.character(hdv0$heures.tv))
> hdv1=hdv0[,c(3,4,6,7,12)]
> hdv.suppl=hdv0[,c(13:19)]
> burt1=acm.burt(hdv1, hdv1)
> burt2=burt1[c(15:26),c(15:26)]
> burt2.orig=burt2
> burt2[8,8:9]=NA
> burt2[9,8:10]=NA
> burt2[10,9:10]=NA
> acm1=dudi.acm(hdv1,nf=3,scannf=F)
> acm1=dudi.fixsigns(acm1, sign.co=c(-1, 1, 1))
> inert1 = inertia.dudi(acm1,c=T,r=T)
> suppl1=dudi.suppl(acm1,hdv.suppl)
```

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), 1er cycle (coll), 2e 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).

```
> burt2
```

	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.*

La première chose à remarquer est que le tableau de Burt est symétrique. On peut donc se contenter de regarder les cases en dessous de la diagonale (diagonale comprise)

- $(\texttt{satis.NA}, \texttt{satis.equ}) = 0$, puisqu'il s'agit d'un terme hors-diagonal de la matrice d'effectifs
- on peut obtenir $(\texttt{satis.NA}, \texttt{satis.NA})$ en sommant toutes les valeurs croisées de cette variable, c'est à-dire $20 + 88 + 254 + 53 + 162 + 129 + 123 + 22 = 851$
- La première vraie difficulté est que, comme il n'y a que 2 variables et que ces variables sont toutes les deux incomplètes, il est n'est pas possible de déduire directement la valeur de $(\texttt{qual.tech}, \texttt{qual.tech})$. Pour y arriver, il faut utiliser le fait que l'effectif total est 1880. On calcule alors l'effectif comme $1880 - 51 - 256 - 583 - 160 - 257 - 288 - 199 = 86$
- on déduite facilement que $(\texttt{satis.equ}, \texttt{qual.tech}) = 86 - 26 - 10 - 22 = 28$
- finalement $(\texttt{satis.equ}, \texttt{satis.equ}) = 1880 - 851 - 115 - 475 = 439$.

Finalement, le tableau de Burt est :

```
> burt2.orig
```

	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	86	28	22	10	26
satis.equ	10	58	151	43	41	71	37	28	439	0	0	0
satis.NA	20	88	254	53	162	129	123	22	0	851	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 7 *à partir de la description des données, quelle hypothèse pouvez vous faire sur les personnes pour lesquelles* `satis=NA` *?*

Seules les personnes qui travaillent peuvent être satisfaites (ou pas) de leur travail. Les autres ne peuvent pas répondre à cette question de satisfaction. C'est justement à cela que servent les réponses `NA` (non applicable). Une grande partie de ces personnes sont donc les personnes pour qui `occup` $\neq$ `prof`.

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

```
> eig=as.matrix(acm1$eig, col=1)
> colnames(eig)=c("")
> rownames(eig)=seq("1", times=nrow(eig))
> round(eig,2)
```

0.49	
0.32	
0.30	
0.23	
0.23	
0.22	
0.21	
0.21	
0.20	
0.20	
0.20	
0.19	
0.19	
0.18	
0.18	
0.17	
0.17	
0.14	
0.10	
0.07	
0.00	

Coordonnées

```
> round(acm1$co,2)
```

	Comp1	Comp2	Comp3
sex.f	-0.21	0.98	0.25
sex.h	0.26	-0.84	-0.31
etud.coll	-0.43	0.08	0.54
etud.lycee	0.23	0.58	0.12
etud.non	-1.28	-0.15	-0.27
etud.prim	-1.04	-0.29	-0.16
etud.sup	0.75	0.61	-1.09
etud.techc	0.29	-0.54	0.79
etud.techl	0.59	-0.07	0.46
occup.chom	-0.76	0.12	0.49
occup.foyer	-1.18	1.27	0.36
occup.inac	-1.24	-0.40	0.40
occup.prof	0.83	0.00	0.19
occup.retr	-0.97	-0.43	-0.75
qual.autr	0.22	0.63	0.09
qual.cadr	0.73	0.44	-1.67
qual.empl	-0.05	0.73	0.71
qual.inter	0.57	-0.03	-0.45
qual.NA	-0.65	0.01	-0.40
qual.ouvq	-0.03	-1.48	0.39
qual.ouvs	-0.75	-0.36	0.39
qual.tech	0.80	-0.84	0.00
satis.equ	0.78	-0.13	0.54
satis.NA	-1.00	0.00	-0.23
satis.non	0.75	-0.47	0.65
satis.oui	0.90	0.23	-0.25

Contributions aux axes

```
> round(inert1$col.abs,1)
```

	Axis1	Axis2	Axis3
sex.f	1.0	15.8	2.3
sex.h	1.2	19.6	2.9
etud.coll	0.8	0.0	2.1
etud.lycee	0.2	2.1	0.1
etud.non	1.4	0.0	0.1
etud.prim	9.9	1.2	0.4
etud.sup	5.3	5.4	18.5
etud.techc	0.9	4.5	10.2
etud.techl	1.0	0.0	1.0
occup.chom	1.6	0.1	1.1
occup.foyer	5.1	9.2	0.8
occup.inac	2.7	0.4	0.5
occup.prof	15.3	0.0	1.3
occup.retr	9.5	2.9	9.3
qual.autr	0.1	0.7	0.0
qual.cadr	2.9	1.7	25.3
qual.empl	0.0	10.4	10.3
qual.inter	1.1	0.0	1.2
qual.NA	2.4	0.0	1.5
qual.ouvq	0.0	21.1	1.5
qual.ouvs	2.4	0.9	1.1
qual.tech	1.2	2.0	0.0
satis.equ	5.7	0.2	4.5
satis.NA	18.5	0.0	1.5
satis.non	1.4	0.8	1.7
satis.oui	8.3	0.8	1.0

Qualités de représentation

```
> round(inert1$col.cum[,1:3],1)
```

	Axis1	Axis1:2	Axis1:3
sex.f	5.3	61.8	69.6
sex.h	5.3	61.8	69.6
etud.coll	2.3	2.4	5.9
etud.lycee	0.6	4.2	4.4
etud.non	3.5	3.5	3.7
etud.prim	31.6	34.1	34.8
etud.sup	16.9	28.0	64.3
etud.techc	2.8	12.3	32.6
etud.techl	2.6	2.7	4.2
occup.chom	4.3	4.4	6.2
occup.foyer	13.9	30.1	31.4
occup.inac	7.1	7.8	8.5
occup.prof	83.5	83.5	87.6
occup.retr	30.9	37.0	55.6
qual.autr	0.1	1.2	1.3
qual.cadr	8.4	11.5	55.6
qual.empl	0.1	24.2	46.6
qual.inter	3.1	3.1	5.0
qual.NA	6.8	6.8	9.4
qual.ouvq	0.0	39.7	42.4
qual.ouvs	6.7	8.3	10.1
qual.tech	3.1	6.5	6.5
satis.equ	18.4	18.8	27.6
satis.NA	83.3	83.3	87.6
satis.non	3.7	5.1	7.8
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 ?*

On regarde les 4 familles de données

— les valeurs propres ont pour somme q/p, ou $p = 5$ est le nombre de variables et $q = 26 - 5 = 21$ est le nombre de catégories moins le nombre de variables ; la somme vaut donc $21/5 = 4.2$;

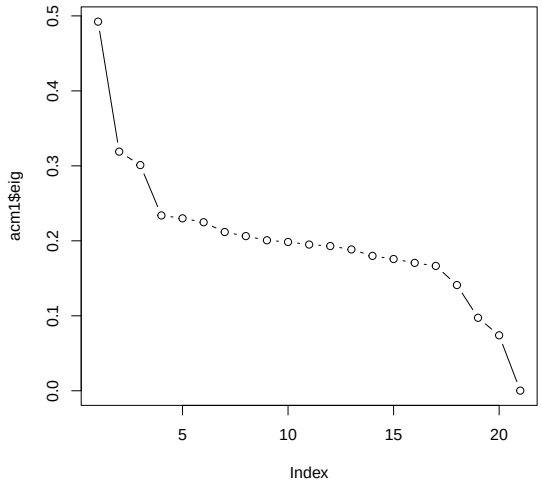
— la somme des coordonnées n'est pas connue ici ; si chaque ligne avait le même poids (comme c'est le cas en général pour les individus en ACP), alors la somme serait 0, puisque les variables sont centrées ;

— la somme des contributions aux axes est égale à 1 (ou ici à 100%) ;

— la somme des qualités de représentation n'est pas connue.

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

```
> plot(acm1$eig, type="b")
```



La règle la plus classique consiste à conserver les axes associés aux valeurs propres supérieures à $1/p$, où p est le nombre de variables actives (5 ici, d'où une limite de 0.20). On serait amené à conserver jusqu'à 8 axes ! Toutefois, on voit bien ici que la décroissance des valeurs propres est très lente. Comme on peut constater un décrochement visible après la troisième valeur propre, il est raisonnable de se contenter de ces 3 premiers axes.

Comme on l'a vu plus en question 8, l'inertie totale est $I_{\mathbf{g}} = 4.2$. En gardant 3 axes, on décrit $0.49 + 0.32 + 0.30 = 1.11 = 26\% I_{\mathbf{g}}$, ce qui est très faible.

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*

Le contribution d'un individu à un axe mesure l'importance de l'individu dans la variance de l'axe. Elle permet de savoir à quel point l'individu définit l'axe

La qualité de représentation mesure à quel point l'axe (ou un ensemble d'axes) permet de représenter fidèlement l'individu.

Le lien entre ces deux grandeurs est qu'elle évaluent l'importance de la valeur $p_i c_{ik}^2$: pour la contribution à l'axe on divise par la somme des autres valeurs sur l'axe k, alors que pour la qualité de représentation on divise par la somme des autres valeurs pour l'individu i.

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).*

On va raisonner ici en fonction des coordonnées. L'analyse étant de mauvaise qualité, on se restreindra aux catégories dont la contribution dépasse 2 fois le poids. Comme expliqué dans le cours, on s'intéresse aux axes dont la coordonnée vérifie

$$|a_{ik}| > \sqrt{2 \times \mu_k}$$

Les limites sur les axes seront respectivement 0.99, 0.80, et 0.78. On classe les éléments par coordonnée décroissante.

Axe 1

$\ominus$	$\oplus$
`etud.non` (-1.28)	
`occup.inac` (-1.24)	
`occup.foyer` (-1.18)	
`etud.prim` (-1.04)	
`satis.NA` (-1.00)	
[`occup.retr` (-0.97)]	

Axe 2

$\ominus$	$\oplus$
`qual.ouvq` (-1.48)	`occup.foyer` (1.27)
`qual.tech` (-0.84)	
`sexe.h` (-0.84)	

Axe 3

$\ominus$	$\oplus$
`qual.cadre` (-1.67)	`etud.techc` (0.79)
`etud.sup` (-1.09)	
[`occup.retr` (-0.75)]	

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.*

La représentation des catégories par le premier plan principal peut être lue dans la troisième colonne du dernier tableau de données fournie. La représentation est mauvaise, ce qui est cohérent avec ce que l'on a trouvé dans la question 9. Les seules variables bien représentées ($> 80\%$) sont `occup.prof` $(87,6\%)$ et `statis.NA` $(87,6\%)$. Ces qualités sont les mêmes, ce qui fait penser à la question 17.

On a vu dans la question 7 que ces deux catégories sont en quelque sorte opposées, c'est-à-dire que les gens qui n'ont pas rempli `satis` sont probablement ceux qui ne travaillent pas (cf question 7). Dans ce sens, comme décrire les gens qui travaillent est équivalent à décrire les gens qui ne travaillent pas (par défaut), ces deux catégories de personnes auront tendance à être représentées avec la même qualité.

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

— Le premier axe décrit en négatif des personnes qui sont en dehors du marché du travail (inactif, au foyer ou retraité) et qui n'ont pour ainsi dire pas fait d'études ;
— Le second axe oppose des hommes ouvriers qualifiés ou techniciens à (des femmes ?) au foyer ;
— le troisième axe oppose des cadres ayant fait des études supérieures (et étant éventuellement à la retraite) à des personnes ayant fait des études techniques courtes.

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

```
> m1=as.matrix(suppl1$eff)
> colnames(m1)=c("Eff")
> m1
                      Eff
hardrock.non         1868
hardrock.oui           12
lecture.bd.non       1837
lecture.bd.oui         43
peche.chasse.non     1666
peche.chasse.oui      214
cuisine.non          1054
cuisine.oui           826
bricol.non           1062
bricol.oui            818
cinema.non           1150
cinema.oui            730
sport.non            1246
sport.oui             634
```

Valeurs test

```
> round(suppl1$test,2)
                    Axis1 Axis2 Axis3
hardrock.non        -1.05 -0.28 -1.36
hardrock.oui         1.05  0.28  1.36
lecture.bd.non      -4.29 -4.04  4.46
lecture.bd.oui       4.29  4.04 -4.46
peche.chasse.non    -1.91  9.18  0.22
peche.chasse.oui     1.91 -9.18 -0.22
cuisine.non         -0.75 -8.97 -5.73
cuisine.oui          0.75  8.97  5.73
bricol.non         -11.18  8.14  2.36
bricol.oui          11.18 -8.14 -2.36
cinema.non         -17.76 -8.49  2.21
cinema.oui          17.76  8.49 -2.21
sport.non          -13.91 -3.34  4.69
sport.oui           13.91  3.34 -4.69
```

Corrélation avec les composantes

```
> cortv=cor(hdv0$heures.tv,acm1$li)
> rownames(cortv)=c("heures.tv")
> round(cortv,2)
           Axis1 Axis2 Axis3
heures.tv -0.33 -0.09  0.09
```

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

Les valeurs test permettent de savoir si des catégories supplémentaires sont corrélées de manière significatives avec les axes principaux. On peut les utiliser si

— on les utilise sur des variables qui n'ont pas pris part à l'analyse : c'est le cas ici ;
— les effectifs des catégories sont assez importants (> 30) : la seule petite catégorie est `hardrock.oui` (12), mais de toute façon la valeur test associée est trop faible pour être utile.
— on considère une valeur comme significative si elle est supérieure à 2 ou 3 en valeur absolue. Ici, comme les valeurs test sont probantes, on se limitera à celles qui sont supérieures à 3.

On remarque (voir question 18) que les valeurs test relatives à `oui` et `non` sont identiques au signe près. C'est parce que décrire les gens qui ont une activité est équivalent à décrire les gens qui n'ont pas cette activité. On reportera donc uniquement les catégories `oui` dans les tableaux ci-dessous, étant entendu que les `non` leur sont opposées.

Axe 1

$\ominus$	$\oplus$
	`cinema.oui` (17.76)
	`sport.oui` (13.91)
	`bricol.oui` (11.18)
	`lecture.bd.oui` (4.29)

Axe 2

$\ominus$	$\oplus$
`peche.chasse.oui` (-9.18)	`cuisine.oui` (8.97)
`bricol.oui` (-8.14)	`cinema.oui` (8.49)
	`lecture.bd.oui` (4.04)
	`sport.oui` (3.34)

Axe 3

$\ominus$	$\oplus$
`sport.oui` (-4.69)	`cuisine.oui` (5.73)
`lecture.bd.oui` (-4.46)	

L'interprétation peut être faite de la manière suivante :

— axe 1 : les personnes à faible niveau d'études et hors du marché du travail (négatif) opposées à la pratique de loisir (cinéma, le sport, bricolage et lecture de bandes dessinées). Il est notable qu'aucun des loisirs testés n'apparaît à gauche de l'axe, alors qu'on peut supposer ces personnes ont du temps libre. On peut imaginer une question de moyens ou de capital culturel.
— axe 2 : les hommes ouvriers qualifiés et techniciens associés à la pratique de la chasse et la pêche ainsi que le bricolage, en opposition aux personnes au foyer (peut-être des femmes), qui elles sont liées à la cuisine, au cinéma la lecture de BD et le sport. On voit des loisirs assez genrés.
— axe 3 : les cadres ayant fait des études supérieures pratiquent le sport et la lecture de BD bien plus que les personnes qui ont fait des études techniques courtes, qui elles s'adonnent à la cuisine.

Toutes les catégories importantes sur l'axe 3 étaient déjà sur l'axe 2. Par contre, `cuisine.oui` se retrouve opposé aux deux autres sur cet axe. On voit que l'interprétation du troisième axe est un peu fragile ici, on aurait peut être pu se contenter des deux premiers.

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

On peut remarquer que le seul axe avec lequel `heures.tv` est un peu corrélé est le premier. La faiblesse de la corrélation est beaucoup moins grave que pour des variables actives d'ACP, qui sont sensées par construction être corrélées avec les axes. De plus, le fait que l'effectif total soit très élevé rend les corrélations exploitables (même pour les axes 2 et 3, en fait). On peut s'en persuader avec un test statistique, mais nous n'avons pas vu ça en cours.

On peut interpréter ces corrélations en disant que les personnes inactives identifiées sur la gauche de l'axe 1 à la question précédente regardent plus que la moyenne la télévision au lieu de pratiquer d'autres activités.

Toutefois les corrélations sont faibles et on peut se demander si l'effet de la variables `heures.tv` est linéaire. Une manière de vérifier cela serait de partager les heures en paquets (par exemple $[0,4[$, $[4,8[$ et $[8,12[$) et de calculer des valeurs test.

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, μ_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,*

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

On sait que, comme le vecteur $\mathbf{c}_k$ est centré, $\sum_{i=1}^{n} c_{ik} = 0$. Comme les individus appartiennent nécessairement soit à la catégorie 1 soit à la 2, on peut écrire

$$0 = \sum_{i \text{ dans cat. } 1} c_{ik} + \sum_{i \text{ dans cat. } 2} c_{ik} = \sqrt{\mu_k} n_1 a_{1k} + \sqrt{\mu_k} n_2 a_{2k},$$

et on en déduit la propriété demandée en divisant par $\sqrt{\mu_k}$.

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*

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

La qualité de la représentation de la catégorie 1 sur l'axe k s'écrit

$$\frac{a_{1k}^2}{a_{11}^2 + \cdots + a_{1q}^2}$$

et on sait grâce à la question 16 que $a_{1k}^2 = n_2^2 a_{2k}^2 / n_1^2$. On peut donc réécrire la qualité de représentation comme

$$\frac{\frac{n_2^2}{n_1^2} a_{2k}^2}{\frac{n_2^2}{n_1^2} a_{21}^2 + \cdots + \frac{n_2^2}{n_1^2} a_{2q}^2} = \frac{a_{2k}^2}{a_{21}^2 + \cdots + a_{2q}^2}.$$

Les deux catégories ont donc la même qualité de représentation sur les axes.

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*

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

La valeur-test associée à la catégorie 1 est égale à

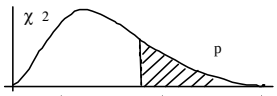
$$a_{1k}\sqrt{n_1}\frac{\sqrt{n-1}}{\sqrt{n-n_1}},$$

où là encore $a_{1k} = -n_2 a_{2k}/n_1$ et $n - n_1 = n_2$. La valeur test ci-dessus devient donc

$$-a_{2k}\frac{n_2}{n_1}\sqrt{n_1}\frac{\sqrt{n-1}}{\sqrt{n_2}} = -a_{2k}\sqrt{n_2}\frac{\sqrt{n-1}}{\sqrt{n_1}}.$$

Pour chaque axe, les deux catégories ont donc la même valeur test au signe près.

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



n p	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 $\sqrt{2\chi^2} - \sqrt{2n\text{-}1} \approx N(0,1)$