TD5 : qu'est-ce qu'être sympathique ?

```
> require(ade4)
> # on lit les données, on ajoute les noms de colonnes
> sympa<-read.table("sympa.txt")
> # on remet les noms en minuscules
> names(sympa)<-c("seri", "gene", "gai", "honn", "intl", "serv", "cour", "comp", "disc", "total")
> # on enlève les totaux
> sympa1<-sympa[1:8,1:9]
> # calcul de l'AFC, seulement 3 axes
> coa1<-dudi.coa(sympa1, scannf=F,nf=3)
> # Ces deux lignes ne sont nécessaires que pour mes TD pour s'assurer de la reproducibilité
> source("fonctions.R")
> coa1<-dudi.fixsigns(coa1, sign.li=c(-1,-1), sign.co=c(-1,-1))
> # calcul des inerties, contributions et qualités
> inert1<-inertia.dudi(coa1, c=T, r=T)
> # poids des lignes et colonnes
> poil=round(as.matrix(coa1$lw*100),1)
> colnames(poil)="Poids"
> poic=round(as.matrix(coa1$cw*100),1)
> colnames(poic)="Poids"
```

1 Les données

Il s'agit d'une recherche sur la représentation sociale. Les personnes interrogées appartenaient à 8 catégories professionnelles différentes : paysan (`PAYS`), ouvrier (`OUVR`), vendeur (`VEND`), commerçant (`COMM`), employé (`EMPL`), technicien (`TECH`), universitaire (`UNIV`), profession libérale (`LIBE`).

Elles avaient à choisir les 3 qualités les plus appropriées à une personne sympathique, parmi une liste de 9 : sérieuse (`seri`), généreuse (`gene`), gaie (`gai`), honnête (`honn`), intelligente (`intl`), serviable (`serv`), courageuse (`cour`), compréhensive (`comp`), discrète (`disc`).

Le tableau suivant indique, pour chaque groupe professionnel, le nombre de fois où chaque qualité a été associée à la représentation d'une personne sympathique.

```
>   sympa
```

	seri	gene	gai	honn	intl	serv	cour	comp	disc	total
PAYS	20	9	9	27	10	16	20	4	8	123
OUVR	42	10	22	51	18	28	38	12	22	243
VEND	11	2	5	14	8	7	5	8	6	66
COMM	8	9	12	23	14	16	14	12	12	120
EMPL	19	10	16	52	32	25	22	25	30	231
TECH	10	5	12	23	20	13	11	13	10	117
UNIV	2	8	7	6	15	6	6	9	4	63
LIBE	8	42	23	24	46	22	22	34	16	237
TOTAL	120	95	106	220	163	133	138	117	108	1200

Question 1 *Combien de personnes ont été interrogées pour cette enquête ? Quelle est la proportion des employés pour qui être honnête rend sympathique ? Quelle est la proportion d'employés parmi les gens qui pensent qu'être honnête rend sympathique ?*

Il y a 1200 réponses à cette enquête. Comme chaque personne a donné 3 réponses, le nombre de personnes est 400. La proportion d'employés pour qui être honnête rend sympathique est égale à $52 \div 231 \times 3 = 67,5\%$. Par contre la proportion d'employés parmi les gens qui pensent qu'être honnête rend sympathique est simplement $52 \div 220 = 24\%$

Question 2 *Commentez l'assertion « il est beaucoup plus courant pour un ouvrier que pour un paysan de penser qu'une personne sérieuse est sympathique ». On se restreindra à une seule interprétation.*

La proportion d'ouvriers qui pensent qu'une personne sérieuse est sympathique est $42/243 \times 3 = 52\%$. La proportion de paysans qui pensent qu'une personne sérieuse est sympathique est $20/123 \times 3 = 49\%$. L'assertion serait donc plus juste en la réécrivant « il est *un peu* plus courant pour un ouvrier que pour un paysan de penser qu'une personne sérieuse est sympathique ».

2 Analyse de correspondances

L'analyse des correspondances du tableau de contingence produit les valeurs propres ci-dessous :

```
>   round(coa1$eig,3)

[1] 0.098 0.022 0.005 0.003 0.001 0.001 0.000
```

On fournit ci-dessous, pour les profils lignes et les profils colonnes, les poids des modalités (en %) et, sur les 3 premiers axes, les coordonnées des modalités, leurs contributions aux axes (en %) et la qualité de leur représentation par les axes factoriels (en % aussi).

```
> round(poil,2)
     Poids
PAYS  10.2
OUVR  20.2
VEND   5.5
COMM  10.0
EMPL  19.2
TECH   9.8
UNIV   5.3
LIBE  19.8
```

```
> round(coa1$li,3)
      Axis1  Axis2  Axis3
PAYS -0.303 -0.213 -0.034
OUVR -0.357 -0.111  0.018
VEND -0.191  0.120  0.189
COMM  0.015  0.046 -0.105
EMPL -0.060  0.215 -0.056
TECH  0.015  0.147  0.075
UNIV  0.461  0.001  0.092
LIBE  0.498 -0.114 -0.006
```

```
>   round(inert1$row.abs,1)
      Axis1 Axis2 Axis3
PAYS   9.6  21.4   2.5
OUVR  26.3  11.6   1.3
VEND   2.1   3.6  40.3
COMM   0.0   1.0  22.6
EMPL   0.7  40.8  12.6
TECH   0.0   9.7  11.4
UNIV  11.4   0.0   9.1
LIBE  49.9  11.8   0.2
```

```
>   round(abs(inert1$row.rel[,1:2]),1)
      Axis1 Axis2
PAYS  63.5  31.5
OUVR  89.6   8.7
VEND  35.9  14.1
COMM   1.0   9.9
EMPL   6.6  83.3
TECH   0.6  58.4
UNIV  90.4   0.0
LIBE  94.4   5.0
```

```
> round(poic,2)
     Poids
seri  10.0
gene   7.9
gai    8.8
honn  18.3
intl  13.6
serv  11.1
cour  11.5
comp   9.8
disc   9.0
```

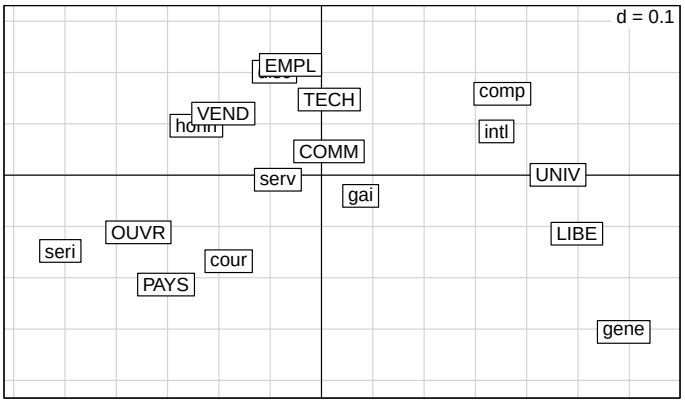
```
>   round(coa1$co,3)
      Comp1  Comp2  Comp3
seri -0.509 -0.148  0.133
gene  0.589 -0.306 -0.063
gai   0.076 -0.040  0.036
honn -0.244  0.096 -0.039
intl  0.341  0.085  0.070
serv -0.085 -0.009 -0.046
cour -0.181 -0.168 -0.056
comp  0.352  0.158  0.062
disc -0.092  0.202 -0.091
```

```
>   round(inert1$col.abs,1)
      Axis1 Axis2 Axis3
seri  26.4  10.0  36.2
gene  28.0  34.0   6.6
gai    0.5   0.7   2.3
honn  11.1   7.8   5.7
intl  16.1   4.5  13.8
serv   0.8   0.0   4.9
cour   3.8  14.9   7.4
comp  12.4  11.3   7.6
disc   0.8  16.8  15.4
```

```
>   round(abs(inert1$col.rel[,1:2]),1)
      Axis1 Axis2
seri  85.7   7.2
gene  77.0  20.7
gai   25.4   7.1
honn  82.6  12.9
intl  87.0   5.4
serv  51.6   0.6
cour  48.1  41.4
comp  78.8  15.9
disc  13.1  63.4
```

Le diagramme ci-dessous est la projection jointe des points-lignes et des points-colonnes sur le premier plan factoriel.

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



Question 3 *Pourquoi y a-t-il 7 valeurs propres ?*

Si les deux variables ont respectivement m_1 et m_2 modalités, le nombre de valeurs propres est $\min(m_1 - 1, m_2 - 1)$. Ici le nombre est donc $\min(7, 8) = 7$.

Question 4 *Quelles sont les modalités qui définissent le premier axe factoriel ? Et le deuxième ? On précisera sur quel(s) critère(s) on se fonde.*

méthode classique on compare les contributions des modalités à 2 fois leur poids. Le facteur 2 est ici arbitraire, on essaie de choisir un facteur (supérieur à 1, bien sûr) qui permettre de garder un nombre de modalités ni trop grandes ni trop petites). On trouve pour le premier axe :

- en négatif : aucune catégorie professionnelle n'est significative (seul `OUVR` $(26,3\%)$ a une contribution supérieure à son poids $(20,2\%)$) ; dans les qualités on a `seri` $(26,4\%$ pour $10\%)$;
- en positif : `LIBE` $(49,9\%$ pour $19,8\%)$ et `UNIV` $(11,4\%$ pour $5,3\%)$, d'une part, et `gene` $(28\%$ pour $7,9\%)$ d'autre part.

En ce qui concerne le second axe, on a

- en négatif : `PAYS` $(21,4\%$ pour $10,2\%)$ et aussi `gene` $(34\%$ pour $7,9\%)$;
- en positif : `EMPL` $(40,8\%$ pour $19,2\%)$, d'une part, et à la limite `disc` $(16,8\%$ pour $9\%)$ d'autre part.

méthode par les coordonnées La contribution de chaque modalité (de coordonnée a_i) à un axe factoriel (associé à la valeur propre λ) s'écrit

$$\frac{n_{i\cdot}}{n}\frac{(a_i)^2}{\lambda}$$

et doit être comparée à son poids $n_{i\cdot}/n$. Si on cherche celles pour lesquelles le rapport est supérieur à 2, par exemple, un calcul simple montre qu'il faut s'intéresser aux modalités dont les coordonnées sur l'axe vérifient

$$|a_i| > \sqrt{2}\sqrt{\lambda}.$$

On compare les coordonnées sur les axes à aux racines carrées des valeurs propres, soit respectivement $0,3130$ et $0,1476$. En multipliant par $\sqrt{2}$, les seuils souhaités sont $0,44$ et $0,20$. On trouve pour le premier axe :

- en négatif : `seri` $(-0,509)$;
- en positif : `LIBE` $(0,498)$ et `UNIV` $(0,461)$, d'une part, et `gene` $(0,589)$ d'autre part.

En ce qui concerne le deuxième axe, on a

- en négatif : `PAYS` $(-0,213)$ et aussi `gene` $(-0,306)$;
- en positif : `EMPL` $(0,215)$, d'une part, `disc` $(0,202)$ d'autre part.

Question 5 *Quelles sont les modalités (lignes et colonnes) qui sont particulièrement mal représentées par le premier plan factoriel ?*

Il s'agit de regarder les $\cos^2$ de l'angle entre le point et le plan factoriel. Toutefois, ce que nous avons ici sont les $\cos^2$ des angles entre les points et les axes. Il faut donc additionner les valeurs des deux colonnes pour avoir ce que nous cherchons. On regarde ici les valeurs inférieures à 50%.

Ceci posé, la catégorie professionnelles formant un grand angle avec le premier plan factoriel est `COMM` $(10,9\%)$. Toutefois, `COMM` est proche du centre de gravité et il faut prendre cette mesure avec prudence.

Pour les qualités, seule `gai` $(32,5\%)$ est mal représentée. Là encore, le points est proche du centre de gravité et il est difficile de conclure.

Question 6 *Que peut-on déduire du fait que `OUVR` et `PAYS` sont proches sur le graphique ? Même question pour `VEND` et `honn`.*

Le fait que `OUVR` et `PAYS` soient proches laisse penser que ces deux modalités d'une même variable ont des caractéristiques proches. Ceci est en fait vrai parce que ces modalités sont très bien représentées par le premier plan factoriel (leur qualité de représentation est respectivement $98,3\%$ et 95%). Si les variables étaient mal représentées, la proximité des variables n'aurait aucune signification.

Par contre, on ne peut rien déduire de la proximité de `VEND` et `honn`, puisque ce sont des modalités de deux variables différentes. La formule du barycentre nous apprend seulement que `honn` est, a un coefficient près, un barycentre pondéré des catégories professionnelles.

3 Quelques démonstrations de propriétés du cours

Question 7 *Montrer que pour toutes variables réelles $x_1,\dots,x_n$ et $y_1,\dots,y_n$ on a la propriété*

$$\textit{Si pour tout } i,\ \frac{x_i}{y_i} = K \textit{ alors } \frac{\sum_{i=1}^n x_i}{\sum_{i=1}^n y_i} = K.$$

On a, pour tout i, $x_i = Ky_i$. Si on somme sur toutes les valeurs de i on a alors

$$\sum_{i=1}^n x_i = \sum_{i=1}^n Ky_i = K\sum_{i=1}^n y_i.$$

On en déduit alors le résultat demandé.

Question 8 *Montrer que l'inertie du nuage sous la métrique du χ^2 vérifie*

$$I_{\mathbf{g}} = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{n_{i.}n_{.j}} = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} \frac{n_{ij}^2}{n_{i.}n_{.j}} - 1 = I_{\mathbf{0}} - 1.$$

On développe simplement l'expression de $I_{\mathbf{g}}$:

$$\begin{aligned}
\sum_{i=1}^{m_1}\sum_{j=1}^{m_2} \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{n_{i.}n_{.j}} &= \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} \left[\frac{n_{ij}^2}{n_{i.}n_{.j}} + \frac{n_{i.}n_{.j}}{n^2} - 2\frac{n_{ij}}{n}\right] \\
&= \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} \frac{n_{ij}^2}{n_{i.}n_{.j}} + \frac{\sum_{i=1}^{m_1} n_{i.} \times \sum_{j=1}^{m_2} n_{.j}}{n^2} - 2\frac{\sum_{i=1}^{m_1}\sum_{j=1}^{m_2} n_{ij}}{n} \\
&= I_{\mathbf{0}} + 1 - 2 = I_{\mathbf{0}} - 1
\end{aligned}$$