TD10 : Risques et monuments (DS de l'an dernier) (correction)

```
> require(ade4)
> source("fonctions.R")
```

1 Tableaux de contingence « dilués »

On se donne un tableau d'effectifs n_{ij} $(1 \leqslant i \leqslant r, 1 \leqslant j \leqslant s)$ de deux variables, ainsi que leurs marges en ligne $n_{i\cdot}$et en colonne $n_{\cdot j}$. Pour un nombre réel $0 \leq \alpha \leq 1$ donné, on définit une version « diluée » des données par

$$\hat{n}_{ij} = \alpha n_{ij} + (1-\alpha)\frac{n_{i\cdot}n_{\cdot j}}{n},$$

où n est l'effectif total. On voit facilement que $\hat{n}_{ij} = n_{ij}$ quand $\alpha = 1$, et qu'il est égal aux effectifs sous hypothèse d'indépendance $n_{i\cdot}n_{\cdot j}/n$ quand $\alpha = 0$.

Question 1 *Montrer que la marge en ligne $\hat{n}_{i\cdot}$ de $\hat{n}_{ij}$ est égale à la marge en ligne $n_{i\cdot}$ de n_{ij} pour tout α. Même question pour la marge en colonne $\hat{n}_{\cdot j}$.*

On nomme r l'effectif de la première variable et s celui de la seconde. La formule pour la marge en ligne est

$$\begin{aligned}
\hat{n}_{i\cdot} = \sum_{j=1}^{s} \hat{n}_{ij} &= \sum_{j=1}^{s} \left[\alpha n_{ij} + (1-\alpha)\frac{n_{i\cdot}n_{\cdot j}}{n}\right] \\
&= \alpha \sum_{j=1}^{s} n_{ij} + (1-\alpha)\frac{n_{i\cdot}}{n}\sum_{j=1}^{s} n_{\cdot j} \\
&= \alpha n_{i\cdot} + (1-\alpha)\frac{n_{i\cdot}}{n} n = n_{i\cdot}
\end{aligned}$$

On montre de même que la marge en colonne du tableau dilué est $\hat{n}_{\cdot j} = n_{\cdot j}$.

Question 2 *Calculer les profils lignes centrés associés à $\hat{n}_{ij}$ en fonction des profils ligne centrés associés à n_{ij}. On rappelle que le tableau des profils ligne est formé des grandeurs $n_{ij}/n_{i\cdot}$ et que son point moyen est le profil marginal des colonnes (de coordonnées $\hat{n}_{\cdot j}/n$).*

On calcule le profil dilué en utilisant la propriété $\hat{n}_{i\cdot} = n_{i\cdot}$:

$$\begin{aligned}
\frac{\hat{n}_{ij}}{\hat{n}_{i\cdot}} &= \alpha\frac{n_{ij}}{\hat{n}_{i\cdot}} + (1-\alpha)\frac{n_{i\cdot}n_{\cdot j}}{n\hat{n}_{i\cdot}} \\
&= \alpha\frac{n_{ij}}{n_{i\cdot}} + (1-\alpha)\frac{n_{\cdot j}}{n}
\end{aligned}$$

Donc le profil centré est

$$\frac{\hat{n}_{ij}}{\hat{n}_{i\cdot}} - \frac{\hat{n}_{\cdot j}}{n} = \alpha\frac{n_{ij}}{n_{i\cdot}} + (1-\alpha)\frac{n_{\cdot j}}{n} - \frac{n_{\cdot j}}{n} = \alpha\left(\frac{n_{ij}}{n_{i\cdot}} - \frac{n_{\cdot j}}{n}\right).$$

Le profil centré est donc multiplié par α. On peut en fait montrer que les vecteurs propres de l'ACM restent les mêmes, mais que les valeurs propres sont multipliées par α.

Note ce dernier résultat permet (avec un peu de travail) de montrer que le tableau dilué conduit aux mêmes axes propres que le tableau d'origine, mais liés à des valeurs propres multipliées par α^2.

2 ACM : risques médicaux et âge

```
> risk=read.csv("risk.csv", colClasses=c("factor", "factor", "factor", "numeric","factor"))
> colnames(risk)=c("CVasc", "Loco", "Neuro", "AgeInt", "Diab", "Id")
> risk1=risk[,c(1:3,5)]
> burt1=acm.burt(risk1, risk1)
> burt2=as.matrix(burt1)
```

```
> burt2[4,2]=NA
> burt2[2,4]=NA
> acm1=dudi.acm(risk1,scannf=F,nf=4)
> acm1=dudi.fixsigns(acm1,sign.co=c(1,-1))
> poids1=as.data.frame(acm1$cw)
> colnames(poids1)="poids"
> inert1=inertia.dudi(acm1, col=T)
> ageint1=risk[,4]
> age1 = as.data.frame(factor(ageint1 %/% 20, label=c("0_19", "20_39", "40_59", "60plus")))
> colnames(age1)="Age"
> suppl1=acm.suppl(acm1,age1)
> corsuppl1=cor(ageint1,acm1$li)
> rownames(corsuppl1)="Age"
```

Une compagnie d'assurance a compilé à propos de ses assurés des données sur leur taux de risque (`0`=normal, `1`=fort) pour le système cardio-vasculaire (`CVas`, cœur), le système locomoteur (`Loco`, risque de paralysie), le système neurologique (`Neuro`, cerveau) et le diabète (`Diab`).

2.1 Les données

On obtient le tableau de Burt suivant, dans lequel une paire de données a été cachée (`NA`) :

```
> burt2

        CVasc.0 CVasc.1 Loco.0 Loco.1 Neuro.0 Neuro.1 Diab.0 Diab.1
CVasc.0   28464       0  27344   1120   26571    1893  22458   6006
CVasc.1       0    8742   7957     NA    7013    1729   6125   2617
Loco.0    27344    7957  35301      0   32186    3115  27312   7989
Loco.1     1120      NA      0   1905    1398     507   1271    634
Neuro.0   26571    7013  32186   1398   33584       0  26303   7281
Neuro.1    1893    1729   3115    507       0    3622   2280   1342
Diab.0    22458    6125  27312   1271   26303    2280  28583      0
Diab.1     6006    2617   7989    634    7281    1342      0   8623
```

Question 3 *Expliquer comment on peut calculer les valeur manquantes du tableau de Burt, indiquées par `NA`, de six manières différentes.*

Les deux valeurs manquantes sont identiques, puisque le tableau de Burt est symétrique. On s'intéresse donc à (`CVasc.1`, `Loco.1`). On peut faire des développements en ligne de la manière suivante :

$$
\begin{aligned}
(\texttt{CVasc.1}, \texttt{Loco.1}) &= (\texttt{CVasc.1}, \texttt{CVasc.1}) - (\texttt{CVasc.1}, \texttt{Loco.0}) = 8742 - 7957 = 785 \\
&= (\texttt{CVasc.1}, \texttt{Neuro.0}) + (\texttt{CVasc.1}, \texttt{Neuro.1}) - (\texttt{CVasc.1}, \texttt{Loco.0}) = 7013 + 1729 - 7957 = 785 \\
&= (\texttt{CVasc.1}, \texttt{Diab.0}) + (\texttt{CVasc.1}, \texttt{Diab.1}) - (\texttt{CVasc.1}, \texttt{Loco.0}) = 6125 + 2617 - 7957 = 785
\end{aligned}
$$

On peut faire la même chose verticalement encore en trois décompositions. Je les donne là, mais on n'est pas obligé de tout développer dans un copie si on explique bien.

$$
\begin{aligned}
(\texttt{CVasc.1}, \texttt{Loco.1}) &= (\texttt{Loco.1}, \texttt{Loco.1}) - (\texttt{CVasc.0}, \texttt{Loco.1}) = 1905 - 1120 = 785 \\
&= (\texttt{Neuro.0}, \texttt{Loco.1}) + (\texttt{Neuro.1}, \texttt{Loco.1}) - (\texttt{CVasc.0}, \texttt{Loco.1}) = 1398 + 507 - 1120 = 785 \\
&= (\texttt{Diab.0}, \texttt{Loco.1}) + (\texttt{Diab.1}, \texttt{Loco.1}) - (\texttt{CVasc.0}, \texttt{Loco.1}) = 1271 + 634 - 1120 = 785
\end{aligned}
$$

Finalement, on obtient le tableau de Burt suivant

```
> burt1

        CVasc.0 CVasc.1 Loco.0 Loco.1 Neuro.0 Neuro.1 Diab.0 Diab.1
CVasc.0   28464       0  27344   1120   26571    1893  22458   6006
CVasc.1       0    8742   7957    785    7013    1729   6125   2617
Loco.0    27344    7957  35301      0   32186    3115  27312   7989
Loco.1     1120     785      0   1905    1398     507   1271    634
Neuro.0   26571    7013  32186   1398   33584       0  26303   7281
Neuro.1    1893    1729   3115    507       0    3622   2280   1342
Diab.0    22458    6125  27312   1271   26303    2280  28583      0
Diab.1     6006    2617   7989    634    7281    1342      0   8623
```

Question 4 *Les personnes ayant un risque locomoteur élevé ont-elles un risque de diabète plus grand ou plus petit que la moyenne ?*

La proportion des personnes ayant un risque locomoteur élevé qui ont aussi un risque de diabète élevé est $634/1905 = 0,33$. Par contre, la proportion de la population générale ayant un risque de diabète élevé est $8623/(28585 + 8623) = 0,23$. On peut donc dire qu'un risque locomoteur élevé augmente le risque de diabète.

2.2 Analyse des correspondances multiples

On réalise une ACM sur les données ci-dessus. On fournit ci-dessous les variances des coordonnées des catégories, leurs poids et leurs contributions (en % pour ces deux derniers) pour tous les axes.

```
>   s.label(acm1$co)
```


```
> eig1=as.data.frame(acm1$eig)
> colnames(eig1)="Variances"
> round(eig1, 2)

  Variances
1      0.34
2      0.24
3      0.23
4      0.20
```

```
> round(poids1*100,1)
        poids
CVasc.0  19.1
CVasc.1   5.9
Loco.0   23.7
Loco.1    1.3
Neuro.0  22.6
Neuro.1   2.4
Diab.0   19.2
Diab.1    5.8
```

```
> round(inert1$col.abs,1)
        Axis1 Axis2 Axis3 Axis4
CVasc.0   7.0   0.0   7.9   8.5
CVasc.1  22.9   0.1  25.8  27.7
Loco.0    1.0   1.7   2.2   0.2
Loco.1   18.6  31.1  41.6   3.6
Neuro.0   3.4   0.1   0.5   5.8
Neuro.1  31.4   0.5   5.0  53.3
Diab.0    3.6  15.4   3.9   0.2
Diab.1   12.0  51.1  13.0   0.7
```

Question 5 *Expliquez pourquoi il n'y a que 4 axes. Combien faut-il en conserver pour l'analyse ? Que peut on dire alors de la qualité globale de la représentation ?*

Il y a 4 valeurs propres et on conserve celles qui sont supérieures à $1/4 = 0,25$. On pourrait se contenter de la première variable, mais comme d'habitude, on ajoutera la seconde pour les besoins de la représentation.

La somme des valeurs propres est $(8-4)/4 = 1$. Les deux premières valeurs propres, dont la somme est $0,58$, représentent 58% de l'inertie totale (alors qu'on est sûr d'avoir au moins 50% avec la moitié des variables !). L'analyse sera donc plutôt mauvaise.

Question 6 *Quelles sont les catégories qui déterminent les deux premiers axes principaux ? (on détaillera les critères et on cherchera à être précis dans la réponse).*

On choisit de conserver les catégories dont la contribution est supérieure à 2 fois le poids. Cela donne :

Axe 1

-	+
`Loco.1` ($18.6 > 2 \times 1.3$)	
`Neuro.1` ($31.4 > 2 \times 2.4$)	
`CVasc.1` ($22.9 > 2 \times 5.9$)	
`Diab.1` ($12.0 \geq 5.8 \times 2$)	

Axe 2

-	+
`Diab.1` ($51.1 > 2 \times 5.8$)	`Loco.1` ($31.1 > 2 \times 1.3$)

On note que sur l'axe 1, `Diab.1` est limite par rapport aux autres. On ne l'aurait pas eu avec des contributions supérieures à 3 fois le poids.

On ne se pose pas ici la question des catégories sur-représentées, puisque l'on savait dès le départ que les petits effectifs des catégories de risque élevé allaient les mettre en avant.

Question 7 *Calculez à partie des poids la contribution à l'inertie totale de chacune des catégories de risque fort (`Xxx.1`). Calculez la part d'inertie totale qu'elles représentent ensemble et montrez en quoi cela explique en partie la forme du nuage sur le premier plan principal.*

Soit $p = 4$ le nombre de variables. L'inertie totale vaut ici $(8-4)/4 = 1$.

On sait que la contribution à l'inertie d'une catégorie d'effectif n_i (et donc de poids $\frac{n_i}{np}$) est égale à

$$\frac{1}{p}\left(1 - \frac{n_i}{n}\right) = \frac{1}{p} - p_i$$

et donc, si les p_i sont petits, chaque contribution à l'inertie sera proche de $\frac{1}{4}$. Dans le cas des variables Xxx.1, on trouve

```
> diag1=diag(as.matrix(burt1))
> n=diag1[1]+diag1[2]
> risques=c("CVasc.1","Loco.1","Neuro.1","Diab.1")
> contribs=1/4*(1-diag1/n)[risques]
> round(contribs,3)

CVasc.1  Loco.1 Neuro.1  Diab.1
  0.191   0.237   0.226   0.192
```

Leur contribution totale est donc 0.846, soit 84.6% de l'inertie totale. On peut en déduire que ces variables auront une grande importance dans l'analyse. C'est un biais bien connu de l'ACM quand les effectifs sont faibles.

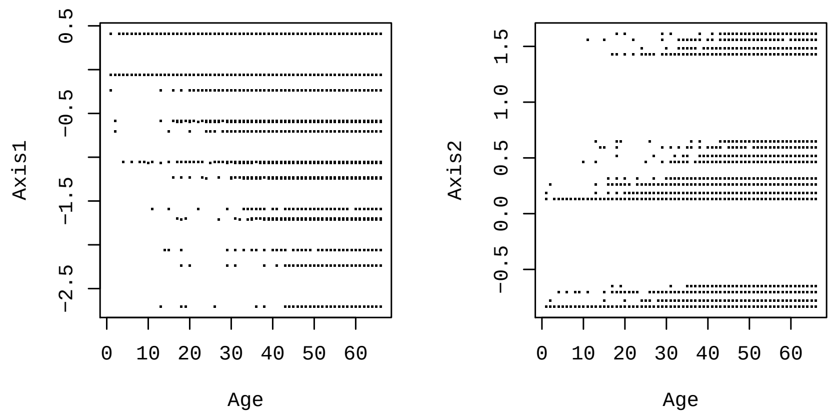
2.3 Une variable supplémentaire : l'âge des assurés

On ajoute à l'analyse une nouvelle variable quantitative : l'âge des assurés. On calcule donc la corrélation de cette variable avec les deux premiers axes, que l'on donne ci-dessous accompagnée d'une représentation des couples (âge, coordonnées factorielles) pour les deux premiers axes.

```
> op=par(mfcol=c(1,2),cex=0.8,family="mono")
> plot(ageint1,acm1$li[,1], xlab="Age", ylab="Axis1", pch='.', cex=2)
> plot(ageint1,acm1$li[,2], xlab="Age", ylab="Axis2", pch='.', cex=2)
> par(op)
```

```
> round(corsuppl1, 2)

    Axis1 Axis2 Axis3 Axis4
Age -0.14  0.12  0.15 -0.09
```



Question 8 *Expliquez pourquoi les points dans les graphiques ci-dessus sont regroupés par lignes.*

Sur les graphes, on distingue respectivement 12 et 16 lignes différentes. En fait, chaque individu est décrit par 4 variables à deux modalités et la formule barycentrique indique que ses coordonnées sont une moyenne des coordonnées de ses catégories. Il y a donc $2^4 = 16$ sortes de clients possibles, ce qui limite le nombre de coordonnées sur les axes. Le fait qu'on ne distingue que 12 lignes sur l'axe 1 est un hasard, en regardant attentivement on peut constater que certaines lignes sont confondues. Par contre, on distingue bien 16 lignes dans le nuage avec l'axe 2.

Question 9 *Que peut-on dire du lien entre la variable* `Age` *et les deux premiers axes ? La forme des nuages de points semble-t-elle donner des informations ?*

La corrélation de la variable `Age` avec les axes sont faibles (inférieures à 0.14) mais non nulles. On peut voir sur les graphes que le nuage ne couvre pas le coin supérieur gauche du carré pour l'axe 1 ; de même, le coin inférieur gauche n'est pas couvert par le nuage sur l'axe 2. La dépendance semble plus simple par rapport à l'axe 1 qu'à l'axe 2. La relation est moins linéaire sur ce dernier axe.

En faisant l'hypothèse que l'âge n'a en fait pas une relation linéaire avec les axes, on regroupe les individus par tranches d'âge de façon à traiter l'âge comme une variable qualitative. On regroupe les individus en 4 groupes : moins de 19 ans (`Age.0_19`), de 20 à 39 ans (`Age.20_39`), de 40 à 59 ans (`Age.40_59`) et plus de 60 ans (`Age.60plus`). On donne ci-dessous les coordonnées des nouvelles catégories sur les axes, leur effectif et enfin la valeur test correspondante.

```
> round(suppl1$li,2)

           Axis1 Axis2 Axis3 Axis4
Age.0_19    0.06 -0.88 -0.63  0.14
Age.20_39   0.29 -0.11 -0.15  0.15
Age.40_59   0.00  0.04  0.00  0.00
Age.60plus -0.21  0.05  0.17 -0.12
```

```
> data.frame(suppl1$eff)

           suppl1.eff
Age.0_19          933
Age.20_39        5116
Age.40_59       23538
Age.60plus       7619
```

```
> round(suppl1$test,2)

            Axis1  Axis2  Axis3  Axis4
Age.0_19     1.93 -27.23 -19.49   4.24
Age.20_39   22.27  -8.22 -11.62  11.44
Age.40_59    0.81  10.50   0.32   0.32
Age.60plus -20.72   5.03  17.08 -11.79
```

Question 10 *Quelles sont les catégories d'âge qui sont liées aux deux premiers axes ? On expliquera ce que sont les valeurs test et pourquoi on peut les utiliser. Quelles interprétation des axes peut-on en déduire ?*

Les valeurs-test données ici permettent de savoir quels sont les catégories liées aux axes. Une catégorie est liée à un axe si
- c'est une catégorie supplémentaire, c'est-à-dire non utilisée dans l'analyse
- son effectif est assez grand (mettons 30)
- sa valeur-test sur l'axe est supérieure à 2 ou 3 en valeur absolue

Les deux premiers points sont valables pour toutes les catégories. Le troisième point nous donne :

Axe 1

-	+
`Age.60plus` $(20,72)$	`Age.20_39` $(22,27)$

Axe 2

-	+
`Age.0_19` $(27,23)$	`Age.40_59` $(10,50)$
`Age.20_39` $(8,22)$	`Age.60plus` $(5,03)$

L'interprétation que l'on peut tirer de ces données est que l'axe 1 sépare les personnes âgées de 60 ans et plus, qui cumulent tous les sur-risques, d'une part, et les jeunes adultes (20 à 39 ans), qui ont peu de risques. On notera que les catégories non représentées ici sont les enfants (moins de 19 ans) et les adultes entre 40 et 59 ans. Il y a donc un effet non linéaire en fonction de l'âge.

L'axe 2, lui, montre que le risque de diabète décroît avec l'âge, alors que le risque locomoteur augmente.

3 AFC : monuments historiques

```
> monum=read.table("monuments.dat")
> monum1=monum[1:5,]
> stats1=stats.conting(monum1)
> coa1=dudi.coa(monum1,nf=2,scannf=F)
> inert1=inertia.dudi(coa1,r=T,c=T)
```

3.1 Les données

On considère la répartition de 12387 monuments historiques en fonction de deux variables :

- leur propriétaire : `COMM` (municipalité), `PRIV` (privé), `ETAT` (état), `DEPA` (département), `ETPU` (établissement public) ;
- leur type : `anti` (antiquités), `chat` (châteaux), `mili` (architecture militaire), `reli` (monuments religieux), `civi` (architecture civile), `dive` (divers).

On donne ci-dessous le tableau de contingence, les profils marginaux des lignes et des colonnes (en %). Le χ^2 d'écart à l'indépendance vaut 4550.06.

```
> monum1
```

	anti	chat	mili	reli	civi	dive
COMM	490	289	351	5022	563	967
PRIV	956	964	76	426	956	242
ETAT	161	82	59	160	138	109
DEPA	32	58	7	49	26	40
ETPU	23	40	2	30	59	10

```
> round(stats1$ml/stats1$tot*100,1)
```

COMM	PRIV	ETAT	DEPA	ETPU
62.0	29.2	5.7	1.7	1.3

```
> round(stats1$mc/stats1$tot*100,1)
```

anti	chat	mili	reli	civi	dive
13.4	11.6	4.0	45.9	14.1	11.0

Question 11 *Avec une erreur inférieure à 1%, montrez que les variables type et propriétaire sont liées. On pourra s'aider de la table de* χ^2 *ci-dessous.*

On sait que le χ^2 d'écart à l'indépendance possède $(6-1)\times(5-1) = 20$ degrés de liberté. On cherche d_c tel que

$$P(\chi^2_{20} > d_c) = 0.01$$

dans la table donnée à la fin du sujet. On obtient, à la ligne $n = 20$ la valeur $37,566$. On en déduit que les variables ne sont pas indépendantes en suivant le raisonnement suivant : si elles étaient indépendantes, alors le χ^2 d'écart à l'indépendance aurait une probabilité 1% d'être supérieure à $37,566$. La valeur est ici supérieure à 4500, et les variables ne sont certainement pas indépendantes.

On peut en fait remarquer dans la table des contributions au χ^2 que beaucoup de valeurs sont à elles seules assez grandes pour conclure à la dépendance entre les deux variables.

3.2 Analyse factorielle des correspondances

On réalise une analyse factorielle des correspondances sur ces données où on se limite aux deux premiers axes factoriels. On fournit ci-dessous, pour les lignes puis pour les colonnes : les coordonnées des modalités, leurs contribution aux axes (en %), la qualité de leur représentation par les deux premiers axes principaux (en % encore).

```
> round(coa1$li, digits=2)
> round(inert1$row.abs,1)
> round(abs(inert1$row.rel[,1:2]),1)
> round(coa1$co, digits=2)
> round(inert1$col.abs,1)
> round(abs(inert1$col.re
```

	Axis1	Axis2
COMM	0.46	0.01
PRIV	-0.84	0.04
ETAT	-0.34	-0.38
DEPA	-0.42	-0.03
ETPU	-0.71	0.21

	Axis1	Axis2
COMM	36.5	1.2
PRIV	58.8	4.6
ETAT	1.9	87.9
DEPA	0.9	0.2
ETPU	1.9	6.1

	Axis1	Axis2
COMM	99.9	0.1
PRIV	99.7	0.2
ETAT	44.0	55.2
DEPA	45.5	0.3
ETPU	73.7	6.4

	Comp1	Comp2
anti	-0.68	-0.09
chat	-0.90	0.12
mili	0.24	-0.30
reli	0.54	0.05
civi	-0.63	0.02
dive	0.22	-0.14

	Axis1	Axis2
anti	17.4	10.5
chat	26.3	16.6
mili	0.7	38.5
reli	38.4	11.7
civi	15.7	0.6
dive	1.5	22.1

	Axis1	Axis2
anti	96.8	1.6
chat	97.0	1.7
mili	38.5	60.8
reli	99.1	0.8
civi	96.5	0.1
dive	59.9	24.8

Question 12 *La valeur propre associée au premier axe principal est 0.352. Montrez que la deuxième vaut à peu près 0.0096 à partir des données fournies.*

Pour calculer la deuxième valeur propre, on utilise la formule des contributions aux axes

$$\text{contribution} = \frac{n_{i\cdot}}{n} \frac{a_{ki}^2}{\lambda_k},$$

qui donne pour l'axe 2 et la modalité `ETAT`

$$87,9 = 5,7 \frac{(-0,38)^2}{\lambda_2},$$

équation qui est facile à résoudre :

$$\lambda_2 = 5,7 \frac{(-0,38)^2}{87,9},$$

On trouve $\lambda_2 = 0,00936$ et la vraie valeur est 0.0096.

Question 13 *Calculez l'inertie totale ; quelle est la part d'inertie expliquée par le premier plan principal ?*

L'inertie totale est χ^2/n =0.3673. La part d'inertie expliquée est donc 0.3616, soit 98.4% du total.

Question 14 *Quelles sont les modalités qui déterminent les deux premiers axes principaux ?*

On regarde les modalités qui influencent le plus les axes (lignes et colonnes à la fois) : on cherche celles dont la contribution est supérieure à deux fois le poids.

Axe 1	
-	+
`chat` (26.3/11.6)	
`PRIV` (58.8/29.2)	

Axe 2	
-	+
`ETAT` (87.9/5.7)	`ETPU` (6.1/1.3)
`mili` (38.5/4.0)	
`dive` (22.1/11.0)	

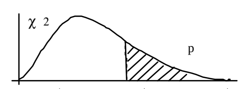
On en déduit que le premier axe principal caractérise les propriétaires privés qui possèdent les châteaux. Le second axe principal est plus difficile à interpréter, mais de toute façon il est peu important. Il oppose l'état et l'architecture militaire d'une part, aux établissements publics.

Question 15 *Quels sont les types de monuments et de propriétaires qui sont mal représentés par le premier plan principal ?*

On dira que l'individu est mal représenté si sa qualité de représentation est inférieure à 50%. Ici, comme on s'intéresse au premier plan principal, il faut additionner les valeurs pour les axes 1 et 2. Cette addition est déjà faite dans le tableau fourni, il suffit donc de lire les données.

La seule modalité mal représentée est `DEPA` (45, 8%), ce qui n'est pas étonnant car la qualité globale de représentation est très élevée. On constate que ce point n'est pas trop proche du centre de la projection sur le premier plan principal (comparé aux autres). Sa mauvaise représentation est donc probablement réelle. Toutes les autres modalités sont bien représentées, puisque qu'on observe des valeurs toutes supérieures à 80%.

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