Notes de cours et codes sur l'AFC

Zaineb Smida

Table des matières

Packages nécessaires :

```
install.packages(c("FactoMineR", "kableExtra", "RColorBrewer", "vcd"))
```

1 Importation de données

Pour importer un fichier texte sous **R**, on utilise la fonction `read.table()`

```
lessive <- read.table("data/achat_lessive.txt")
```

2 Création de la table de contingence

Pour rappel, ce tableau est un résumé d'un jeu de données avec 391 lignes et deux variable qualitatives, dont les premières lignes ont cette forme :

age	achat
25 à 34 ans	lpdt
60 ans et plus	syst
20 à 24 ans	occas
60 ans et plus	occas
20 à 24 ans	occas
15 à 19 ans	occas

La table de contingence entre deux variables qualitatives s'obtient avec la fonction `table()`. Il s'agit d'un résumé exhaustif du tableau de données initial :

Tableau de contingence :

```
tablessive <- table(lessive)
kableExtra::kbl(addmargins(tablessive))
```

	jamais	lpdt	occas	syst	Sum
15 à 19 ans	9	6	24	6	45
20 à 24 ans	6	25	37	2	70
25 à 34 ans	9	17	25	5	56
35 à 44 ans	3	29	37	12	81
45 à 59 ans	12	45	36	3	96
60 ans et plus	4	19	9	11	43
Sum	43	141	168	39	391

3 Le test du χ^2 d'indépendance

Le test du χ^2 d'indépendance permet d'évaluer si deux variables qualitatives sont indépendantes. Nous présentons d'abord la construction théorique du test, suivie d'une application pratique à l'aide de la fonction `chisq.test()` de `R`.

3.1 Construction théorique du test

Le test du χ^2 consiste à calculer la distance entre la table de contingence observée et la table de contingence théorique. Cette dernière correspond à la table de contingence qu'on devrait avoir, dans l'hypothèse où les deux variables Age et Achat seraient indépendantes et en gardant les mêmes marges que le tableau initial.

Pour obtenir le tableau des effectifs théoriques, on considère le produit matriciel de la marge en colonne par la transposé de la marge en ligne, divisé par l'éffectif total :

```
tab_th <- rowSums(tablessive) %*% t(colSums(tablessive)) / N
row.names(tab_th) <- row.names(tablessive)
kableExtra::kbl(round(tab_th))
```

	jamais	lpdt	occas	syst
15 à 19 ans	5	16	19	4
20 à 24 ans	8	25	30	7
25 à 34 ans	6	20	24	6
35 à 44 ans	9	29	35	8
45 à 59 ans	11	35	41	10
60 ans et plus	5	16	18	4

Remarque : si on calculait les profils lignes et colonnes de cette table, on verrait qu'ils sont tous égaux.

Pour calculer le tableau des contributions, on calcule d'abord l'écart à l'indépendance en considérant la différence entre la table observée et la table théorique, divisée par l'écart-type de la table théorique.

```
ecart <- (tablessive - tab_th) / sqrt(tab_th)
```

Ce tableau nous permet de déterminer quelles sont les couples de modalités les plus sur/sous-représentées.

```
library(kableExtra)
```

```
Warning: le package 'kableExtra' a été compilé avec la version R 4.4.2
```

```
print_table <- function(my_tab, position = "H") {
  my_tab[, 1:ncol(my_tab)] <- lapply(my_tab[, 1:ncol(my_tab)],
                function(x) {
    cell_spec(round(x, 3), bold = T, color = spec_color(abs(x), end = 0.9),
              font_size = spec_font_size(abs(x)))
                  })
    kbl(my_tab, escape = F, align = "c", position = position) %>%
    kable_styling(position = 'center',latex_options = 'HOLD_position') %>%
    kable_classic("striped", full_width = F)
}
print_table(as.data.frame(as(ecart, "matrix")), position = "H")
```

	jamais	lpdt	occas	syst
15 à 19 ans	1.821	-2.539	1.061	0.713
20 à 24 ans	-0.612	-0.048	1.262	-1.885
25 à 34 ans	1.145	-0.711	0.191	-0.248
35 à 44 ans	-1.979	-0.039	0.372	1.379
45 à 59 ans	0.444	1.764	-0.817	-2.125
60 ans et plus	-0.335	0.887	-2.205	3.24

Commentaires : par exemple, la modalité "SYST" est sur-représentée chez les plus de 60 ans, alors qu'elle est sous-représentée chez les 45-59 ans et les 19-24 ans.

Pour calculer les contributions, il suffit de prendre le tableau des écarts au carré.

```
contrib <- ecart ^ 2
contrib
```

```
                achat
age                    jamais        lpdt        occas        syst
  15 à 19 ans     3.316290965  6.446061200  1.125514554  0.509003869
  20 à 24 ans     0.374621679  0.002338586  1.593648111  3.554990960
  25 à 34 ans     1.310976413  0.505290321  0.036614918  0.061410350
  35 à 44 ans     3.918264306  0.001505729  0.138680459  1.902645711
  45 à 59 ans     0.197079641  3.112941789  0.667724699  4.515351416
  60 ans et plus  0.112350770  0.787112978  4.859839537 10.500690095
```

En prenant la somme du tableau des contributions, on obtient la statistique du test du χ^2 :

```
sum(contrib)
```

```
[1] 49.55095
```

Cette valeur se retrouve directement avec la fonction `chisq.test()`.

3.2 Application pratique avec la fonction chisq.test de R

Tableau : Test du χ^2 d'indépendance

```
res_chi2 <- chisq.test(tablessive)
res_chi2
```

```

    Pearson's Chi-squared test

data:  tablessive
X-squared = 49.551, df = 15, p-value = 1.425e-05
```

On trouve également directement le tableau d'effectif théorique sous l'hypothèse d'indépendance, ainsi que le tableau des écarts.

Tableau : Effectifs théoriques

```
tab_th <- res_chi2$expected
round(tab_th)
```

```
                achat
age              jamais lpdt occas syst
  15 à 19 ans         5   16    19    4
  20 à 24 ans         8   25    30    7
  25 à 34 ans         6   20    24    6
  35 à 44 ans         9   29    35    8
  45 à 59 ans        11   35    41   10
  60 ans et plus      5   16    18    4
```

Tableau : Résultats des écarts

```
ecart <- res_chi2$residuals
ecart
```

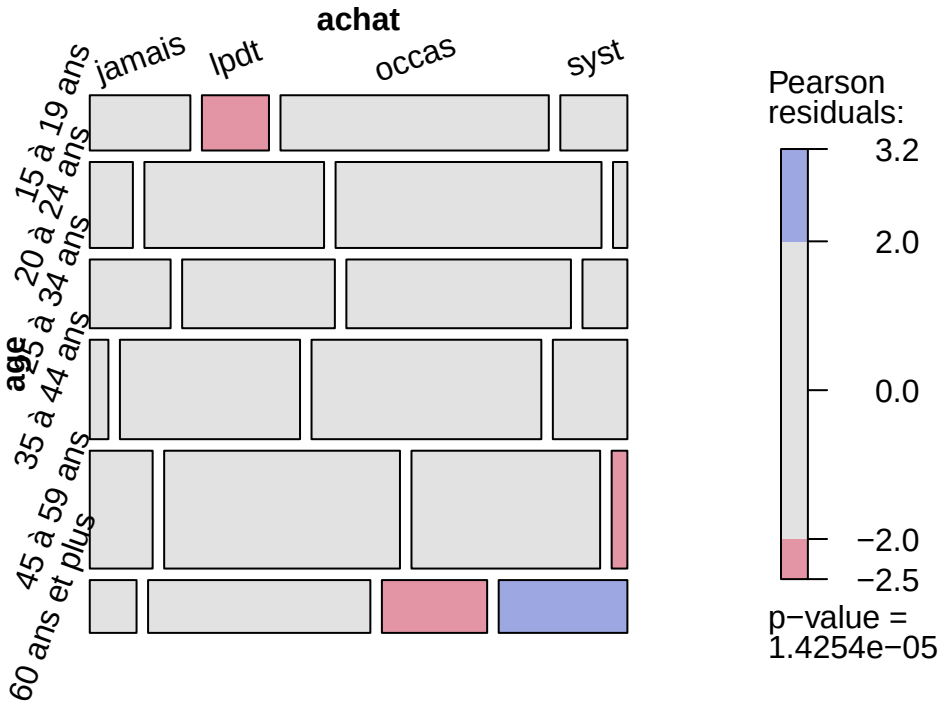
```
               achat
age                  jamais        lpdt       occas        syst
  15 à 19 ans     1.82106863 -2.53890945  1.06090271  0.71344507
  20 à 24 ans    -0.61206346 -0.04835893  1.26239776 -1.88546837
  25 à 34 ans     1.14497878 -0.71083776  0.19135025 -0.24781112
  35 à 44 ans    -1.97946061 -0.03880372  0.37239825  1.37936424
  45 à 59 ans     0.44393653  1.76435308 -0.81714423 -2.12493563
  60 ans et plus -0.33518766  0.88719388 -2.20450437  3.24047683
```

La fonction `mosaic()` du package **vcd** offre une manière alternative de visualiser les écarts significatifs :

```
library(vcd)
```

```
Le chargement a nécessité le package : grid
```

```
mosaic(tablessive, shade = T, gdp = shading_Friendly,
       rot_labels = c(20, 90, 0, 70))
```



Cela met en avant l'aspect d'une représentation différente des écarts tout en restant clair et précis.

On peut ainsi définir le tableau des contributions au χ^2 comme suit :

Tableau : Résultats des contributions au χ^2

```
contrib <- ecart^2
print_table(as.data.frame(as(contrib, "matrix")))
```

	jamais	lpdt	occas	syst
15 à 19 ans	3.316	6.446	1.126	0.509
20 à 24 ans	0.375	0.002	1.594	3.555
25 à 34 ans	1.311	0.505	0.037	0.061
35 à 44 ans	3.918	0.002	0.139	1.903
45 à 59 ans	0.197	3.113	0.668	4.515
60 ans et plus	0.112	0.787	4.86	10.501

4 Analyse descriptive : profils lignes et profils colonnes

Il existe plusieurs façons de décrire un tableau de contingence et plus précisément déterminer s'il existe un lien entre les deux variables qualitatives.

4.1 Profils lignes

On peut par exemple calculer les profils lignes et les comparer entre eux :

Tableau : Profils lignes

```
library(kableExtra)
tabPL <- prop.table(tablessive, 1)
kbl(round(addmargins(tabPL, 2), digits = 2))
```

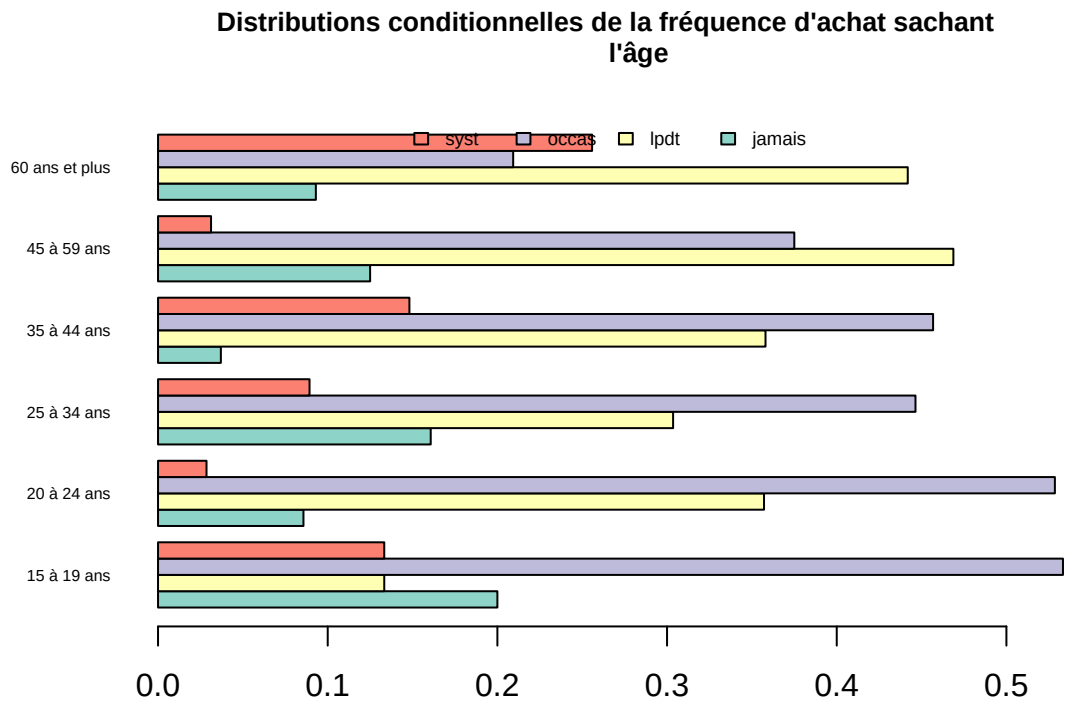
	jamais	lpdt	occas	syst	Sum
15 à 19 ans	0.20	0.13	0.53	0.13	1
20 à 24 ans	0.09	0.36	0.53	0.03	1
25 à 34 ans	0.16	0.30	0.45	0.09	1
35 à 44 ans	0.04	0.36	0.46	0.15	1
45 à 59 ans	0.12	0.47	0.38	0.03	1
60 ans et plus	0.09	0.44	0.21	0.26	1

Une façon de les représenter graphiquement est d'utiliser la fonction barplot(). Un diagramme des profils lignes, représenté de manière juxtaposée, est donné par :

Figure : Diagramme 1 des profils lignes

```
library(RColorBrewer)
 barplot(t(tabPL), beside= T, horiz = T, las = 1, cex.names = 0.5,
         main = "Distributions conditionnelles de la fréquence d'achat sachant
         l'âge", cex.main=0.8, legend.text = T,
```

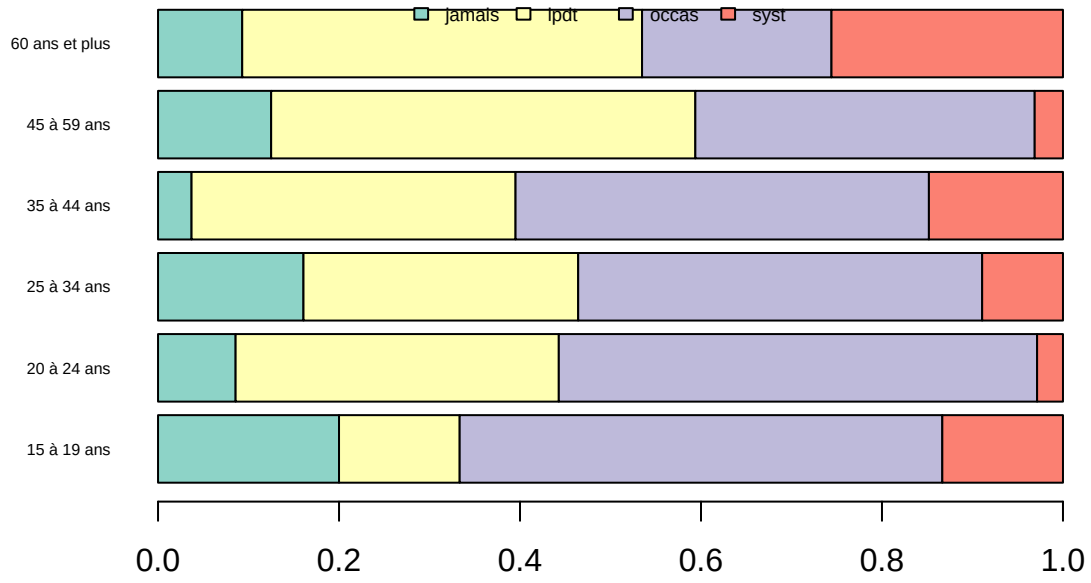
```
        col= brewer.pal(4, name= "Set3"),
        args.legend = list(x = "top", ncol = 4, bty = "n", cex = 0.6))
```



Ou bien, on peut les répresenter comme suit :

Figure : Diagramme 2 des profils lignes

```
library(RColorBrewer)
 barplot(t(tabPL), horiz= T, las = 1, cex.names = 0.5,
        legend.text = T, col= brewer.pal(4, name= "Set3"),
        args.legend = list(x = "top", ncol = 4, bty = "n", cex = 0.6))
```



Commentaires : on voit que les pourcentages sont différents d'une ligne à une autre. Par exemple, la modalité "JAMAIS" est observé dans 20% des cas chez les plus jeunes, alors qu'elle est beaucoup moins représentée dans les autres classes d'âge (par exemple, 4% environ chez les 35-44 ans).

4.2 Profils colonnes

On peut aussi calculer les profils colonnes et les comparer entre eux :

Tableau : Profils colonnes

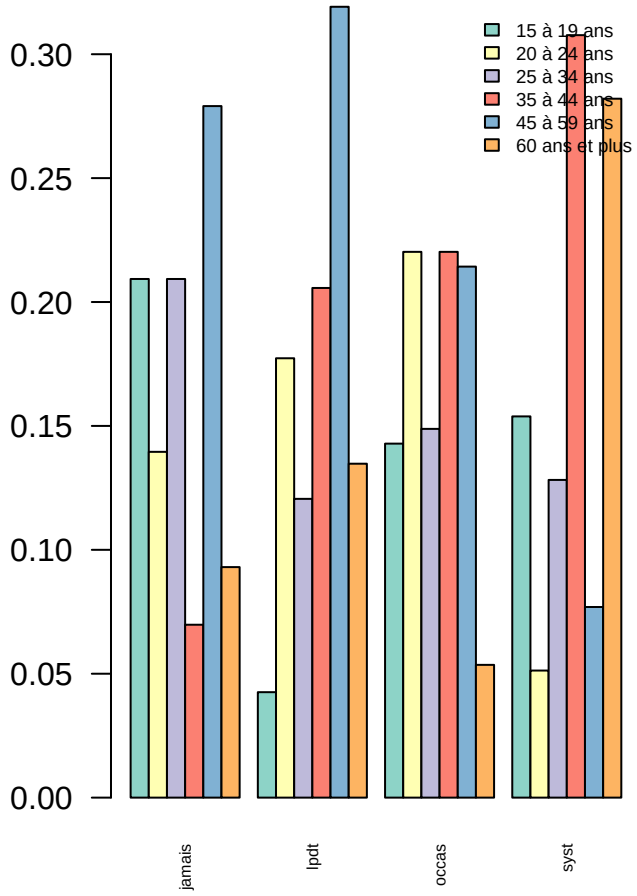
```
tabPC <- prop.table(tablessive, 2)
kbl(round(addmargins(tabPC, 1), digits = 2))
```

	jamais	lpdt	occas	syst
15 à 19 ans	0.21	0.04	0.14	0.15
20 à 24 ans	0.14	0.18	0.22	0.05
25 à 34 ans	0.21	0.12	0.15	0.13
35 à 44 ans	0.07	0.21	0.22	0.31
45 à 59 ans	0.28	0.32	0.21	0.08
60 ans et plus	0.09	0.13	0.05	0.28
Sum	1.00	1.00	1.00	1.00

On peut également les représenter graphiquement de manière juxtaposée comme suit :

Figure : Diagramme 1 des profils colonnes

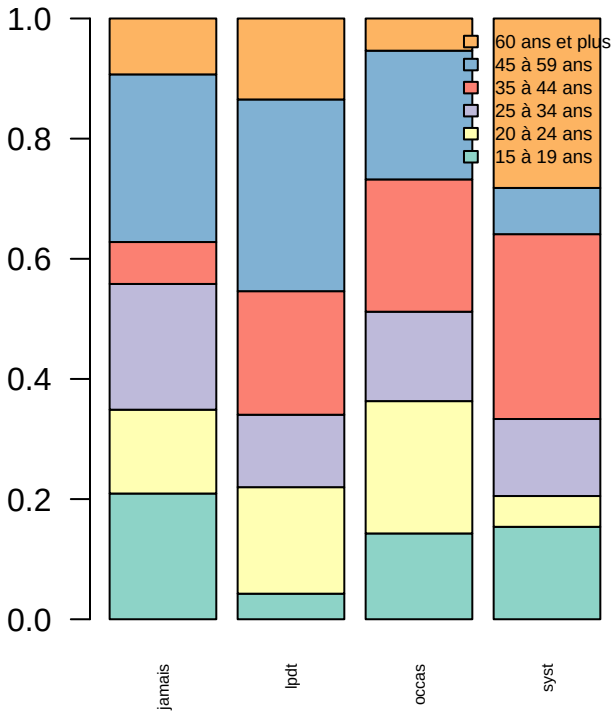
```
library(RColorBrewer)
 barplot(tabPC, beside= T, horiz = F, las = 2, cex.names = 0.5,
         legend.text = T, col= brewer.pal(6, name= "Set3"),
         args.legend = list(x = "topright", ncol = 1, bty = "n", cex = 0.6))
```



Ou bien, on peut les répresenter comme suit :

Figure : Diagramme 2 des profils colonnes

```
library(RColorBrewer)
 barplot(tabPC, horiz = F, las = 2, cex.names = 0.5,
         legend.text = T, col= brewer.pal(6, name= "Set3"),
         args.legend = list(x = "topright", ncol = 1, bty = "n", cex = 0.6))
```



On constate également que les profils colonnes varient entre eux. Ainsi, les personnes qui achètent “SYST” ou “LPDT” font majoritairement partie des classes d’âge supérieures à 35 ans.

5 Analyse Factorielle des Correspondances

La fonction `CA()` du package **FactoMineR** s’utilise directement sur le tableau de contingence.

```
library("FactoMineR")
res_afc <- CA(tablessive, graph = F)
```

5.1 AFC : Valeurs propres

On représente les valeurs propres et le pourcentage d'inertie expliquée :

Tableau : AFC-Valeurs propres

```
res_afc$eig
```

```
      eigenvalue percentage of variance cumulative percentage of variance
dim 1 0.06257418               49.37646                            49.37646
dim 2 0.04405526               34.76343                            84.13989
dim 3 0.02009933               15.86011                           100.00000
```

Commentaires : On garde deux axes selon le critère de la part d'inertie expliquée.

5.2 AFC : Intérprétation des axes

Nous présentons ici les informations relatives aux profils-lignes et aux profils-colonnes.

5.2.1 AFC : Coordonnées

On représente les coordonnées des profils-lignes :

Tableau : AFC-Coordonnées des profils-lignes

```
res_afc$row$coord[,1:2]
```

```
                    Dim 1       Dim 2
15 à 19 ans    -0.07049537  0.49304658
20 à 24 ans    -0.24571980 -0.05282357
25 à 34 ans    -0.08229516  0.13179860
35 à 44 ans     0.16344750 -0.01651419
45 à 59 ans    -0.14598104 -0.22618665
60 ans et plus  0.59897998 -0.06554885
```

On peut également afficher les coordonnées des profils-colonnes :

Tableau : AFC-Coordonnées des profils-colonnes

```
res_afc$col$coord[,1:2]
```

```
             Dim 1      Dim 2
jamais -0.15943505  0.2526960
lpdt    0.04497942 -0.2711430
occas  -0.15842603  0.1086131
syst    0.69562002  0.2338009
```

5.2.2 AFC : Contributions

On représente les contributions des profils-lignes :

```
res_afc$row$coord[,1:2]
```

```
                    Dim 1       Dim 2
15 à 19 ans    -0.07049537  0.49304658
20 à 24 ans    -0.24571980 -0.05282357
25 à 34 ans    -0.08229516  0.13179860
35 à 44 ans     0.16344750 -0.01651419
45 à 59 ans    -0.14598104 -0.22618665
60 ans et plus  0.59897998 -0.06554885
```

De plus, il est possible d'ajouter des couleurs pour mieux visualiser les contributions des profils-lignes :

Tableau : AFC-Contributions des profils-lignes

```
print_table(as.data.frame(res_afc$row$contrib)[, 1:2])
```

	Dim 1	Dim 2
15 à 19 ans	0.914	63.506
20 à 24 ans	17.275	1.134
25 à 34 ans	1.55	5.647
35 à 44 ans	8.844	0.128
45 à 59 ans	8.362	28.512
60 ans et plus	63.055	1.073

On peut également représenter les contributions des profils-colonnes :

Tableau : AFC-Contributions des profils-colonnes

```
print_table(as.data.frame(res_afc$col$contrib)[, 1:2])
```

	Dim 1	Dim 2
jamais	4.467	15.94
lpdt	1.166	60.179
occas	17.234	11.505
syst	77.132	12.376

Commentaires : sur l'Axe 1, on retrouve les modalités 60 ans et plus, 20-24 ans pour les profils lignes et la modalité "SYS" pour les profils colonnes. Sur l'axe 2, on retrouve les modalités 15-19 ans, 45-59 ans pour les profils lignes et la modalité "LPDT" pour les profils colonnes.

5.3 Qualité de représentation

On représente les mesures de la qualité de représentation `cos2` des profils-lignes et des profils-colonnes sur les axes principaux :

Tableau : ACP-Cos2 des profils lignes

```
print_table(as.data.frame(res_afc$row$cos2)[, 1:2])
```

	Dim 1	Dim 2
15 à 19 ans	0.02	0.96
20 à 24 ans	0.765	0.035
25 à 34 ans	0.198	0.508
35 à 44 ans	0.363	0.004
45 à 59 ans	0.241	0.578
60 ans et plus	0.949	0.011

Tableau : ACP-Cos2 des profils colonnes

```
print_table(as.data.frame(res_afc$col$cos2)[, 1:2])
```

	Dim 1	Dim 2
jamais	0.118	0.297
lpdt	0.026	0.955
occas	0.501	0.235
syst	0.897	0.101

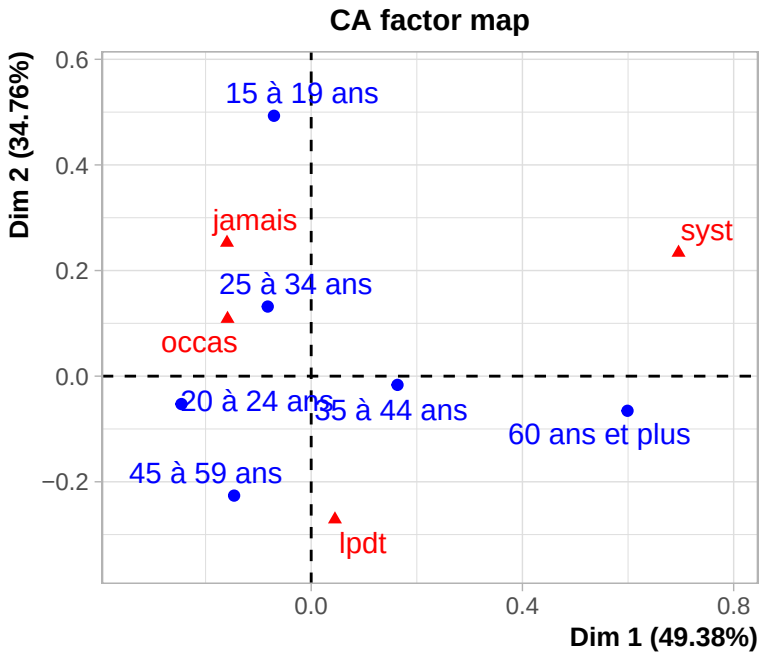
Commentaires : sur l'Axe 1, les modalités les mieux représentées sont : 20-24 ans, 60 ans et plus, "systématiquement", "occasionnellement", et sur l'axe 2 : 15-19 ans, 25-34 ans, 45-49 ans, "lpdt", "jamais".

5.4 Représentation graphique

Les coordonnées des profils-lignes et des profils-colonnes sont représentées simultanément :

AFC-Graphique des modalités (profils)

```
plot(res_afc)
```



Commentaires :

- **Axe 1 :** sur-représentation des acheteurs systématiques de lessive écologique parmi les plus de 60 ans au contraire des 20-24 ans ;
- **Axe 2 :** sous-représentation des 15-19 ans achetant la plupart du temps de la lessive écologique au contraire des 45-59 ans.