Lecture des résultats d'ACM fournis par FactoMineR

Exemple d'illustration : Mini.ACM

On travaille sur un mini-exemple comportant 10 observations et 3 questions comportant respectivement 2, 2 et 3 modalités. Le tableau de données observées est le suivant :

	Sexe	Revenu	Preference
s1	F	M	A
s2	F	M	A
s3	F	E	B
s4	F	E	C
s5	F	E	C
s6	H	E	C
s7	H	E	B
s8	H	M	B
s9	H	M	B
s10	H	M	A

On réalise une ACM sur ces données, avec toutes les variables et toutes les observations actives, avec le plugin FactoMineR de R Commander. On spécifie 4 axes (le maximum) et on coche toutes les options donnant des résultats numériques. Les lignes de commande correspondantes sont donc :

```
Mini.ACM.MCA<-Mini.ACM[, c("Sexe", "Revenu", "Preference")]
res<-MCA(Mini.ACM.MCA, ncp=4, graph = FALSE)
plot.MCA(res, axes=c(1, 2), col.ind="black", col.ind.sup="blue",
col.var="darkred", col.quali.sup="darkgreen", label=c("ind",
"ind.sup", "quali.sup", "var", "quanti.sup"), invisible=c(""))
plot.MCA(res, axes=c(1, 2), choix="var", col.var="darkred",
col.quali.sup="darkgreen", label=c("var", "quali.sup"),
invisible=c(""))
plot.MCA(res, axes=c(1, 2), choix="quanti.sup", col.quanti.sup="blue",
label=c("quanti.sup"))
res$eig
res$var
res$ind
dimdesc(res, axes=c(1, 2))
remove(Mini.ACM.MCA)
```

Résultats fournis par l'ACM proprement dite

Valeurs propres, inertie relative des valeurs propres et inertie cumulée

```
> res$eig
       eigenvalue percentage of variance cumulative percentage of variance
dim 1 0.60283671              45.212753                           45.21275
dim 2 0.46370802              34.778101                           79.99085
dim 3 0.20295865              15.221899                           95.21275
dim 4 0.06382996               4.787247                          100.00000
```

Résultats relatifs aux modalités (individus colonnes)

Coordonnées des modalités des questions (individus colonnes) selon les 4 axes

```
> res$var
$coord
       Dim 1      Dim 2       Dim 3       Dim 4
F  0.3113818  0.7880231 -0.52133928 -0.10132245
H -0.3113818 -0.7880231  0.52133928  0.10132245
E  0.9378348 -0.1378969 -0.09122967  0.30516785
M -0.9378348  0.1378969  0.09122967 -0.30516785
A -1.0315775  1.0241541  0.21105858  0.41921800
B -0.1925652 -1.0073826 -0.66646281 -0.06265998
C  1.2883311  0.3190227  0.67755850 -0.33567137
```

N.B. Dans les formules littérales ci-dessous, la coordonnée de la j-ième modalité selon l'axe k sera notée : ξ_{jk} . Sur notre exemple, $1 \leq j \leq 7$ et $1 \leq k \leq 4$.

Contributions des modalités des questions à la formation des axes (inertie relative de chaque modalité selon l'axe considéré)

```
$contrib
       Dim 1      Dim 2      Dim 3      Dim 4
F  2.6806221 22.3193779 22.3193779  2.6806221
H  2.6806221 22.3193779 22.3193779  2.6806221
E 24.3165395  0.6834605  0.6834605 24.3165395
M 24.3165395  0.6834605  0.6834605 24.3165395
A 17.6524119 22.6196580  2.1948178 27.5331123
B  0.8201526 29.1798474 29.1798474  0.8201526
C 27.5331123  2.1948178 22.6196580 17.6524119
```

Il s'agit évidemment de pourcentages : la somme de chaque colonne est égale à 100.

Qualités de représentation (cos^2)

```
$cos2
       Dim 1      Dim 2       Dim 3       Dim 4
F 0.09695864 0.62098047 0.271794644 0.010266239
H 0.09695864 0.62098047 0.271794644 0.010266239
E 0.87953416 0.01901557 0.008322853 0.093127418
M 0.87953416 0.01901557 0.008322853 0.093127418
A 0.45606523 0.44952501 0.019091025 0.075318743
B 0.02472090 0.67654646 0.296115117 0.002617515
C 0.71134161 0.04361805 0.196750936 0.048289400
```

Les cos^2 sont définis comme d'habitude comme les rapports $\frac{OH^2}{OM^2}$ où M désigne le point image de la modalité dans l'espace multidimensionnel et H sa projection sur l'axe considéré.

Valeurs Test

Cette notion est décrite par Escoffier et Pagès en ce qui concerne l'ACP. On obtient ici :

```
$v.test
       Dim 1      Dim 2      Dim 3      Dim 4
F  0.9341455  2.3640694 -1.5640178 -0.3039674
H -0.9341455 -2.3640694  1.5640178  0.3039674
E  2.8135045 -0.4136908 -0.2736890  0.9155036
M -2.8135045  0.4136908  0.2736890 -0.9155036
A -2.0259780  2.0113988  0.4145108  0.8233278
B -0.4716865 -2.4675733 -1.6324938 -0.1534850
C  2.5302321  0.6265481  1.3306985 -0.6592455
```

Pour la modalité j et la dimension k, la valeur test est calculée à partir de la coordonnée de la manière suivante :

$$\text{v.test}(j,k) = \sqrt{\frac{N_j(N-1)}{N-N_j}}\,\xi_{jk}$$

On vérifie par exemple que : $\text{v.test}(1,1) = \sqrt{\frac{5\times 9}{10-5}} \times 0.3113818 = 0.9341454$

Escoffier et Pagès consacrent un paragraphe aux valeurs-tests dans l'ouvrage cité en bibliographie. Cependant, dans leur exposé, la valeur test est calculée pour une classe et fait intervenir l'écart type de la série des coordonnées des éléments de la classe selon la dimension considérée. Dans ces conditions, et sous l'hypothèse H0 : "cette dimension n'a pas de lien avec la partition considérée", la valeur test suit une loi normale centrée réduite, de sorte que l'on peut lui associer une p-value. De telles valeurs tests pourraient être obtenues en divisant les résultats fournis par les racines carrées des valeurs propres considérées. En effet, la valeur propre relative à un axe est la variance (variance d'échantillon, non corrigée) de la série des coordonnées des modalités sur cet axe, pondérées par leur effectif.

Valeurs eta2

Ces valeurs sont calculées pour chaque dimension et chaque question. Elles sont également décrites dans l'ouvrage d'Escoffier et Pagès déjà cité, dans le chapitre relatif à l'ACM.

```
$eta2
                Dim 1      Dim 2       Dim 3      Dim 4
Sexe       0.09695864 0.62098047 0.271794644 0.01026624
Revenu     0.87953416 0.01901557 0.008322853 0.09312742
Preference 0.83201733 0.75112802 0.328758443 0.08809621
```

Calcul des valeurs eta2 : on considère les coordonnées des individus lignes selon la dimension considérée et on forme une partition des individus lignes selon les modalités de la question considérée. On calcule ensuite le rapport de corrélation : $\eta^2 = \frac{\text{Somme des carrés intergroupes}}{\text{Somme des carrés totale}}$.

Ce rapport peut aussi être vu comme le coefficient η^2 de l'ANOVA à un facteur dont la variable dépendante est la série des coordonnées des observations sur la dimension considérée et la variable indépendante est la question elle-même.

En pratique, on peut le calculer comme carré du coefficient de corrélation entre la série des moyennes par modalité de la question et la série des coordonnées des observations sur la dimension spécifiée.
On montre également que ce coefficient peut être calculé de la manière suivante :

$$\eta^2 = \text{Contrib. relative de la question à l'inertie de l'axe} \times \text{Valeur propre de l'axe} \times \text{Nb de questions}$$

La contribution relative d'une question à l'inertie d'un axe est la somme des inerties relatives des différentes modalités de cette question. Ainsi, pour la question Sexe et l'axe 1, on obtient :

$$\eta^2 = \frac{2.6806221 + 2.6806221}{100} \times 0.60283671 \times 3 = 0.09696$$

Le coefficient peut également être calculé à partir des coordonnées des individus lignes :

	Dim 1	Sexe	Moy. Dim 1		
1	-0,7118	F	0,24176	R=	0,311375622
2	-0,7118	F	0,24176	R^2=	0,096954778
3	0,4536	F	0,24176		
4	1,0894	F	0,24176		
5	1,0894	F	0,24176		
6	0,8221	H	-0,24176		
7	0,1863	H	-0,24176		
8	-0,6190	H	-0,24176		
9	-0,6190	H	-0,24176		
10	-0,9792	H	-0,24176		

Interprétation : lorsque le coefficient η^2 est proche de 1, les individus correspondant à une même modalité sont très regroupés, et les modalités sont nettement séparées les unes des autres. C'est une situation de liaison forte entre la question et la variable numérique correspondant aux coordonnées sur l'axe considéré. Au contraire, lorsque η^2 est proche de 0, les moyennes des groupes définis par les différentes modalités sont proches les unes des autres, les individus d'un même groupe sont dispersés, et il y a peu de lien entre la question et la série des coordonnées sur l'axe.

Résultats relatifs aux observations

Coordonnées des observations (individus lignes) sur les axes factoriels

```
$coord
         Dim 1      Dim 2       Dim 3       Dim 4
1  -0.7118220  0.9545696 -0.16207658  0.01679254
2  -0.7118220  0.9545696 -0.16207658  0.01679254
3   0.4536393 -0.1748785 -0.94635982  0.18627576
4   1.0894144  0.4744025  0.04808598 -0.17392717
5   1.0894144  0.4744025  0.04808598 -0.17392717
6   0.8220509 -0.2970788  0.81956729  0.09343635
7   0.1862758 -0.9463598 -0.17487851  0.45363928
8  -0.6189826 -0.8113576 -0.03987624 -0.35161909
9  -0.6189826 -0.8113576 -0.03987624 -0.35161909
10 -0.9791856  0.1830883  0.60940473  0.28415606
```

Contributions relatives des observations à la formation des axes (inertie relative de chaque observation selon l'axe considéré)

```
$contrib
        Dim 1      Dim 2       Dim 3       Dim 4
1   8.4051051 19.6503620  1.29429408  0.04417821
2   8.4051051 19.6503620  1.29429408  0.04417821
3   3.4136707  0.6595205 44.12706380  5.43610868
4  19.6873184  4.8534355  0.11392772  4.73925783
5  19.6873184  4.8534355  0.11392772  4.73925783
6  11.2097968  1.9032632 33.09494593  1.36775158
7   0.5755897 19.3138112  1.50683378 32.24012900
8   6.3556097 14.1964566  0.07834674 19.36958692
9   6.3556097 14.1964566  0.07834674 19.36958692
10 15.9048765  0.7228969 18.29801941 12.64996483
```

Qualités de représentation (cos^2) des observations sur les différents axes

```
$cos2
        Dim 1      Dim 2       Dim 3        Dim 4
1  0.35078580 0.63083288 0.018186104 0.0001952234
2  0.35078580 0.63083288 0.018186104 0.0001952234
3  0.17639023 0.02621357 0.767654499 0.0297417064
4  0.82164726 0.15580917 0.001600796 0.0209427657
5  0.82164726 0.15580917 0.001600796 0.0209427657
6  0.46783918 0.06110020 0.465016530 0.0060440900
7  0.02974171 0.76765450 0.026213567 0.1763902283
8  0.32840527 0.56425807 0.001362956 0.1059737031
9  0.32840527 0.56425807 0.001362956 0.1059737031
10 0.66378762 0.02320706 0.257105164 0.0559001555
```

Les cos^2 sont définis comme d'habitude comme les rapports $\frac{OH^2}{OM^2}$ où M désigne le point image de l'observation dans l'espace multidimensionnel et H sa projection sur l'axe considéré.

Description des dimensions

```
> dimdesc(res, axes=c(1, 2))
$`Dim 1`
$`Dim 1`$quali
                  R2      p.value
Revenu     0.8795342 6.058914e-05
Preference 0.8320173 1.942789e-03

$`Dim 1`$category
    Estimate      p.value
E  0.7281590 6.058914e-05
C  0.9836808 7.127695e-04
A -0.8175557 2.320456e-03
M -0.7281590 6.058914e-05


$`Dim 2`
$`Dim 2`$quali
                  R2     p.value
```

```
Sexe       0.6209805 0.006778941
Preference 0.7511280 0.007689817

$`Dim 2`$category
    Estimate     p.value
F  0.5366131 0.006778941
A  0.6211882 0.012849418
H -0.5366131 0.006778941
B -0.7622093 0.003305139
```

Les résultats du type "dim k`$quali"

On utilise ici une variante du modèle linéaire généralisé.
Une question à deux modalités est transformée en une variable indépendante numérique prenant les valeurs 1 et -1. On calcule ensuite le coefficient de détermination (carré du coefficient de corrélation linéaire) entre cette variable et la série des coordonnées des individus, et on teste la significativité de ce coefficient. La valeur de R2 est alors identique au coefficient η^2 vu précédemment. Sous l'hypothèse H0 d'absence de lien entre les modalités de la question et les coordonnées selon la dimension considérée, le rapport $F = (n-2)\frac{R^2}{1-R^2}$ suit une loi de Fisher à 1 et (n-2) degrés de liberté, d'où la valeur de la p.value.

Par exemple, pour Revenu et la dimension 1 : $R^2 = 0.8795342$;
$F = (10-2)\frac{0.8795342}{1-0.8795342} = 58.41$ et on vérifie à l'aide de R que la p-value correspondante est bien $6.06\ 10^{-6}$:

```
> 1-pf(58.41,1,8)
[1] 6.058496e-05
```

Une question à 3 modalités est recodée à l'aide de deux variables numériques indépendantes. Par exemple, pour la question "Préférence", on recode de la manière suivante :

Préférence	v1	v2
A	1	0
B	-1	-1
C	0	1

On remarque que le couple de valeurs (-1, -1) est attribué à la modalité correspondant aux valeurs moyennes les plus proches de 0.

N.B. La manière de construire les contrastes fait partie des options de R. FactoMineR modifie cette option en spécifiant des contrastes de type "somme" dans la procédure condes(). On peut retrouver ces contrastes à l'aide de la fonction contr.sum. Par exemple :

```
> contr.sum(3)
  [,1] [,2]
1    1    0
2    0    1
3   -1   -1
```

On réalise ensuite une régression linéaire multiple de la série des coordonnées sur les deux variables indépendantes. R2 est le coefficient de détermination de cette régression. La p-value

correspondante est celle du rapport $F = (n-3)\frac{R^2}{1-R^2}$ pour une loi de Fisher à 2 et (n-3) degrés de liberté. Par exemple, pour la dimension 1 et la question "Préférence" :

Recodage numérique :

	Sexe	Revenu	Preference	Dim 1	Moy Dim1 par pref	glm-pref1	glm-pref2
s1	F	M	A	0,711822023	0,800943199	1	0
s2	F	M	A	0,711822023	0,800943199	1	0
s3	F	E	B	-0,453639284	0,14951255	-1	-1
s4	F	E	C	-1,08941444	-1,00029327	0	1
s5	F	E	C	-1,08941444	-1,00029327	0	1
s6	H	E	C	-0,822050915	-1,00029327	0	1
s7	H	E	B	-0,186275756	0,14951255	-1	-1
s8	H	M	B	0,618982621	0,14951255	-1	-1
s9	H	M	B	0,618982621	0,14951255	-1	-1
s10	H	M	A	0,97918555	0,800943199	1	0

N.B. Dans ce tableau produit par Statistica, les signes sont inversés par rapport à ceux produits par FactoMineR.

Régression linéaire multiple :

	Synthèse de la Régression; Variable Dép. : Dim 1 R= ,91214984 R²= ,83201733 R² Ajusté = ,78402228 F(2,7)=17,335 p<,00194 Err-Type de l'Estim.: ,38035					
	b*	Err-Type de b*	b	Err-Type de b	t(7)	valeur p
OrdOrig.			-0,0166	0,1214	-0,1369	0,894996
glm-pref1	0,8747	0,1878	0,8176	0,1755	4,6578	0,0023204
glm-pref2	-1,0524	0,1878	-0,9837	0,1755	-5,6043	0,0008123

On retrouve bien ainsi R^2=0.83201 et p.value=0.00194.

Cette méthode peut être étendue dans le cas d'un k nombre de modalités supérieur à 3. Il y a alors (k-1) variables indépendantes et le test porte sur le rapport $F = (n-k-1)\frac{R^2}{1-R^2}$ qui, sous H0, suit une loi de Fisher à k-1 et (n-k) degrés de liberté.

Les résultats du type $`Dim 1`$category

Les valeurs de la colonne "estimate" sont, au signe près (car les signes des coordonnées sont inverés), les coefficients de l'équation de la régression linéaire multiple. Il semble que la p-value indiquée soit celle correspondant au test de nullité du coefficient concerné (cf. résultats obtenus pour la modalité A ; pour la modalité C, les résultats fournis par FactoMineR et ceux fournis par la régression multiple sous Statistica divergent légèrement).

On notera que, pour l'ensemble des résultats fournis par dimdesc(), seuls les résultats relatifs aux axes choisis pour la représentation graphiques et conduisant à des p-values inférieures à 0.05 sont mentionnés.

N.B. Ces résultats utilisent les fonctions dimdesc() et condes() du package FactoMineR.

Résultats relatifs aux éléments supplémentaires

FactoMineR permet d'indiquer des observations supplémentaires, des variables qualitatives (questions) supplémentaires et des variables quantitatives supplémentaires.

Les résultats concernant ces objets sont généralement les mêmes que pour les objets actifs, à l'exception des contributions, les objets supplémentaires ne contribuant pas à l'inertie des axes. On peut mentionner le traitement réservé aux variables quantitatives supplémentaires : on évalue le coefficient de corrélation de cette variable avec la série des coordonnées de chaque observation, sur chaque axe.

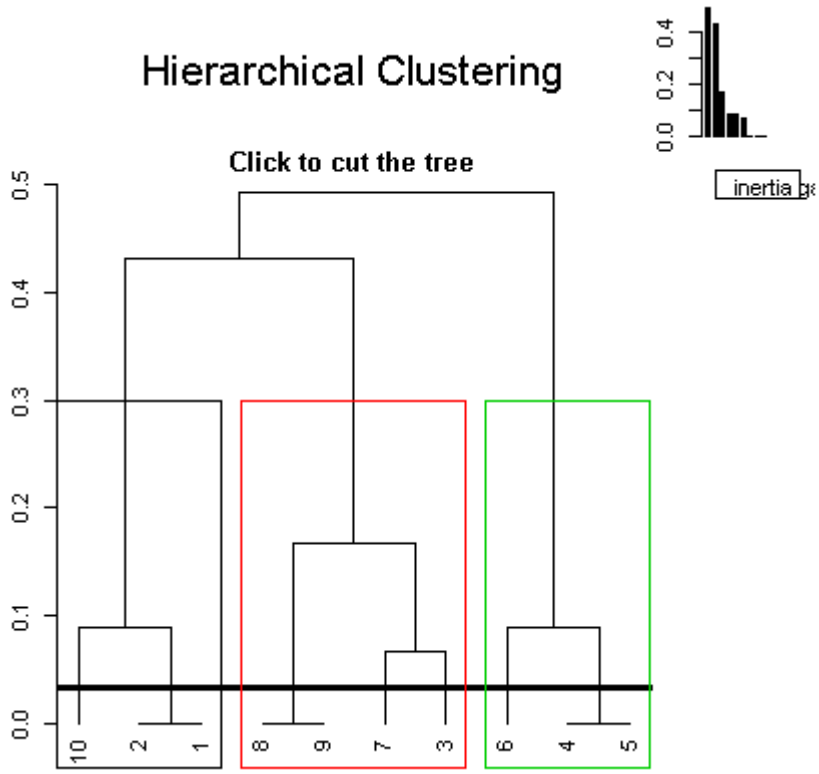
Classification ascendante hiérarchique sur les résultats de l'analyse

On travaille toujours sur les mêmes données (mini-exemple à trois questions et 10 observations) et on spécifie les options comme précédemment. Dans la fenêtre de dialogue relative à l'ACM, on indique "Perform clustering after MCA". On indique un choix interactif du nombre de classes. On coche "consolidate clusters" et on demande l'impression des résultats pour les classes.

Les commandes sont alors :

```
res<-MCA(Mini.ACM.MCA, ncp=4, graph = FALSE)
res.hcpc<-HCPC(res ,nb.clust=0,consol=TRUE,min=3,max=10,graph=TRUE)
res.hcpc$data.clust[,ncol(res.hcpc$data.clust),drop=F]
res.hcpc$desc.var
res.hcpc$desc.axes
res.hcpc$desc.ind
```

Au cours de la procédure, HCPC demande à l'utilisateur de choisir le nombre de classes à retenir, en cliquant dans le dendrogramme. On prend par exemple 3 classes, ce qui paraît raisonnable étant donné le saut de l'indice d'agrégation. A noter : si on laisse le logiciel faire ce choix automatiquement, il choisit 7 classes, ce qui paraît un nombre assez élevé au vu du nombre d'observations. Ce comportement est sans doute dû au fait que certains objets ont des modalités identiques sur l'ensemble des questions, de sorte que la plus forte perte d'inertie est obtenue lors de la première agrégation de deux éléments réellement distincts.



La CAH est faite sur les observations, caractérisées par leurs coordonnées sur l'ensemble des axes retenus pour l'ACM (4 dans notre cas). La métrique utilisée est la distance euclidienne, la méthode d'agrégation est celle de Ward.

Composition des classes

La composition des classes est indiquée de la manière suivante :

```
> res.hcpc$data.clust[,ncol(res.hcpc$data.clust),drop=F]
   clust
10     1
2      1
1      1
8      2
9      2
7      2
3      2
6      3
4      3
5      3
```

Autrement dit, la classe 1 est formée des observations 1, 2 et 10, la classe 2 des observations 3, 7, 8 et 9 et la classe 3 des observations 4, 5 et 6.

Lien entre la partition en classes et l'une ou l'autre des questions.

```
> res.hcpc$desc.var
$test.chi2
               p.value df
```

```
Preference 0.0004993992  4
Revenu     0.0497870684  2
```

On croise la variable "classe" et chacune des questions. On teste alors l'existence d'un lien entre ces deux variables qualitatives à l'aide d'un test du khi-2. Les seuls résultats indiqués sont ceux pour lesquels le résultat du test est significatif à 5%. Par exemple, pour "classe" et Revenu, on obtient comme tableau de contingence :

	Classe 1	Classe 2	Classe 3	Total
M (moyen)	3	2	0	5
E (élevé)	0	2	3	5
Total	3	4	3	10

Le test du khi-2 réalisé sur ce tableau produit comme résultats :

```
> .Table  # Counts
  Classe 1 Classe 2 Classe 3
M        3        2        0
E        0        2        3

> .Test <- chisq.test(.Table, correct=FALSE)
   Pearson's Chi-squared test
data:  .Table
X-squared = 6, df = 2, p-value = 0.04979
```

Ces résultats sont identiques à ceux fournis par FactoMineR. Bien entendu, les effectifs sont ici trop faibles pour tirer de véritables conclusions à partir du résultat du test.

Composition des classes en termes de modalités

```
$category
$category$`1`
             Cla/Mod Mod/Cla Global    p.value  v.test
Preference=A     100     100     30 0.01666667 2.39398

$category$`2`
             Cla/Mod Mod/Cla Global    p.value   v.test
Preference=B     100     100     40 0.00952381 2.592656

$category$`3`
             Cla/Mod Mod/Cla Global    p.value  v.test
Preference=C     100     100     30 0.01666667 2.39398
```

Lecture de ce tableau de résultats :
- La fréquence de la classe 1 dans la modalité "Preference=A" est de 100%.
- La fréquence de la modalité "Preference=A" dans la classe 1 est de 100%.
- La modalité Preference=A représente 30% des observations.
- La p-value est calculée à l'aide d'une loi hypergéométrique (tirages sans remise). Pour Preference=A, le calcul est le suivant :

```
>phyper(2,3,7,3,lower.tail=FALSE)*2
          [1] 0.01666667
```

Plus généralement, considérons une classe C et une modalité M, d'effectifs respectifs N_C et N_M sur un effectif total de N et soit K l'effectif conjoint de C et M. La p-value évalue la

probabilité d'obtenir un effectif conjoint aussi extrême sous l'hypothèse d'un tirage au hasard avec remise de N_M objets dans un ensemble de N objets dont N_C appartiennent à la classe.

Si la modalité M est sur-représentée dans la classe $\left(\frac{K}{N_C} > \frac{N_M}{N}\right)$, la p-value est le double de la probabilité cumulée P(X>=K) de la queue à droite de la distribution hypergéométrique de paramètres N, K et N_C/N.

Ainsi, dans l'exemple Etudiants.Ville.RData, avec une partition en 4 classes :
- La modalité Type=Cité a un effectif de 41 ;
- La classe 1 a un effectif de 45 ;
- L'effectif conjoint de Type=Cité et de Classe=1 est de 38.
- L'effectif total est de 383 observations.

Dans ce cas, la p.value est 2 * phyper(38-1, 45, 338, 41, lower.tail=FALSE), c'est-à-dire 2.19e-41.

Si la modalité M est sous-représentée dans la classe $\left(\frac{K}{N_C} \leq \frac{N_M}{N}\right)$, la p-value est le double de la probabilité cumulée P(X<=K) de la queue à gauche de la distribution hypergéométrique de paramètres N, K et N_C/N.

Ainsi, dans l'exemple Etudiants.Ville.RData, avec une partition en 4 classes :
- La modalité Type=Appart a un effectif de 116 ;
- La classe 1 a un effectif de 45 ;
- L'effectif conjoint de Type=Appart et de Classe=1 est de 1.
- L'effectif total est de 383 observations.

Dans ce cas, la p.value est 2 * phyper(1, 45, 338, 116, lower.tail=TRUE) c'est à dire 1.25e-6.

Les valeurs tests sont les valeurs d'une variable normale centrée réduite correspondant à la même p-value bilatérale. Sur le mini-exemple :

```
          - qnorm(0.01666667 /2)
         [1] 2.39398
```

Description des axes factoriels

```
> res.hcpc$desc.axes
$quanti
$quanti$`1`
         v.test Mean in category  Overall mean sd in category Overall sd
Dim.2  2.011399        0.6974091  3.884696e-17      0.3636798  0.6809611
Dim.1 -2.025978       -0.8009432 -5.551115e-17      0.1260364  0.7764256
         p.value
Dim.2 0.04428335
Dim.1 0.04276703

$quanti$`2`
         v.test Mean in category Overall mean sd in category Overall sd
Dim.2 -2.467573       -0.6859884 3.884696e-17      0.3001922  0.6809611
         p.value
Dim.2 0.01360324

$quanti$`3`
        v.test Mean in category  Overall mean sd in category Overall sd
Dim.1 2.530232         1.000293 -5.551115e-17      0.1260364  0.7764256
         p.value
Dim.1 0.01139871
```

Pour chaque axe factoriel et chaque classe, on calcule la moyenne des coordonnées des observations constituant la classe $\overline{x_C}$, la moyenne des coordonnées de l'ensemble des observations $\overline{x}$, l'écart type (non corrigé) des coordonnées des observations constituant la classe s_C et l'écart type (non corrigé) *s* de l'ensemble des coordonnées des observations. On peut noter que $\overline{x}$ est égale à 0, tandis que *s* est la racine carrée de la valeur propre associée à l'axe.
On calcule ensuite la valeur test selon la formule donnée par Escoffier et Pagès :

$$\text{v.test} = \sqrt{\frac{N_C (N-1)}{N - N_C}} \frac{\overline{x_C} - \overline{x}}{s}$$

On calcule ensuite la p-value correspondante pour un test bilatéral, en utilisant une loi normale centrée réduite. Pour la dimension 2 et la classe 1, le calcul sous Excel donne :

Id	Dim2	Classe
1	0,9545696	1
2	0,9545696	1
3	-0,1748785	2
4	0,4744025	3
5	0,4744025	3
6	-0,2970788	3
7	-0,9463598	2
8	-0,8113576	2
9	-0,8113576	2
10	0,1830883	1

Moyenne	2E-08		
Ecart type	0,680961123	v.test	2,011398781
Moy classe 1	0,697409167	p.value	0,044283352
ET classe 1	0,363679773		

Formules utilisées :

`Moyenne`	`=MOYENNE(B2:B11)`		
`Ecart type`	`=ECARTYPEP(B2:B11)`	`v.test`	`=(B15-B13)/ B14*RACINE(3*9/7)`
`Moy classe 1`	`=MOYENNE(B2;B3;B11)`	`p.value`	`=2*(1- LOI.NORMALE.STANDARD(D14))`
`ET classe 1`	`=ECARTYPEP(B2;B3;B11)`		

N.B. Escoffier et Pagès indiquent également la formule :

$$s^2_{\overline{x_c}} = \frac{s^2}{N_C} \frac{N - N_C}{N-1}$$

d'où une autre formule pour calculer les valeurs tests. Cependant, les deux formules n'aboutissent pas aux mêmes résultats numériques, et l'égalité précédente est sans doute une estimation d'un paramètre statistique et non pas une égalité algébrique.

Description des individus

```
> res.hcpc$desc.ind
```

```
$para
cluster: 1
        2         1        10
0.3849002 0.3849002 0.7698004
------------------------------------------------------------
cluster: 2
        8         9         7         3
0.6454972 0.6454972 0.6454972 1.0408330
------------------------------------------------------------
cluster: 3
        5         4         6
0.3849002 0.3849002 0.7698004
```

Pour chaque classe et chaque observation appartenant à cette classe, on affiche la distance (sur notre exemple, distance euclidienne dans R^4) de l'observation au centre de classe. Le calcul équivalent sous Excel donne :

	Dim1	Dim2	Dim3	Dim4	clust	Centre1	Centre2	Centre3	Centre4	Dist au centre
1	-0,7118	0,9546	-0,1621	0,0168	1	-0,8009	0,6974	0,0951	0,1059	0,3849
2	-0,7118	0,9546	-0,1621	0,0168	1	-0,8009	0,6974	0,0951	0,1059	0,3849
3	0,4536	-0,1749	-0,9464	0,1863	2	-0,1495	-0,6860	-0,3002	-0,0158	1,0408
4	1,0894	0,4744	0,0481	-0,1739	3	1,0003	0,2172	0,3052	-0,0848	0,3849
5	1,0894	0,4744	0,0481	-0,1739	3	1,0003	0,2172	0,3052	-0,0848	0,3849
6	0,8221	-0,2971	0,8196	0,0934	3	1,0003	0,2172	0,3052	-0,0848	0,7698
7	0,1863	-0,9464	-0,1749	0,4536	2	-0,1495	-0,6860	-0,3002	-0,0158	0,6455
8	-0,6190	-0,8114	-0,0399	-0,3516	2	-0,1495	-0,6860	-0,3002	-0,0158	0,6455
9	-0,6190	-0,8114	-0,0399	-0,3516	2	-0,1495	-0,6860	-0,3002	-0,0158	0,6455
10	-0,9792	0,1831	0,6094	0,2842	1	-0,8009	0,6974	0,0951	0,1059	0,7698

```
$dist
cluster: 1
       1        2       10
1.740051 1.740051 1.536591
------------------------------------------------------------
cluster: 2
       7        8        9        3
1.592808 1.592808 1.592808 1.446580
------------------------------------------------------------
cluster: 3
       4        5        6
1.740051 1.740051 1.536591
```

Ce tableau fait appel à la fonction distinctivness(), définie comme fonction interne dans HCPC(). Il donne, pour chaque élément de la classe, le minimum de la distance de cet élément à un centre d'une autre classe.

Bibliographie

Escoffier B., Pagès J., Analyses factorielles simples et multiples. Objectifs, méthodes et interprétation, 3è édition, 1998, Dunod, Paris.