

Université de Carthage  
Ecole Supérieure de la Statistique et de l'Analyse de l'Information

**Examen d'Analyse des Données**

**1 ère année du cycle de formation d'ingénieurs**

Durée de l'épreuve : 1 heure 30 - Documents non autorisés  
Nombre de pages : 6 - Date de l'épreuve : 17 mai 2023

On a effectué une enquête sur la relation des consommateurs vis-à-vis des magasins *Champion*. Un questionnaire a ainsi été administré à un échantillon représentatif de 60 clients. Ce questionnaire est présenté à l'Annexe I.

Soit  $df$  la base de données sur le logiciel R obtenue à l'issue de ce questionnaire.

Dans la suite, à la question numéro  $i$  de ce questionnaire on associe la variable statistique notée  $Q_i$ .

**Partie I**

On a effectué une Analyse en Composantes Principales (ACP) normée sur les 8 items de la première question. Les résultats de cette ACP sont présentés à l'Annexe II.

1. Quel est l'intérêt de cette ACP ?
2. Compléter la commande PCA par les arguments adéquats pour effectuer l'ACP normée sur les 8 premières variables du dataframe  $df$ .
2. Déterminer le nombre d'axes à retenir.
3. Donner une interprétation des axes retenus.

**Partie II**

On a effectué une classification automatique des individus. La hiérarchie obtenue est donnée à l'Annexe II.

4. Déterminer le meilleur nombre de classes à retenir.
5. On suppose que nous avons retenu la partition en 3 classes issue de cette hiérarchie. Compléter la commande `catdes` par le paramètre adéquat.
6. En vous basant sur les sorties de la commande `catdes` décrire les 3 classes obtenues.

**Partie III**

On a effectué une Analyse Factorielle des Correspondances (AFC) sur les variables  $Q_2$  (Le nombre de fois, par semaine, où vous fréquentez *Champion*) et  $Q_7$  (Catégorie Socio-Professionnelle). Les résultats de cette AFC sont présentés à l'Annexe II.

7. Cette AFC est elle pertinente ?

8. Vérifier que la troisième valeur propre de cette AFC est égal à 0.005 et en déduire le nombre d'axes à retenir.

9. Calculer les tableaux des profils lignes et des profils colonnes et en déduire une interprétation de la carte de représentation simultanée.

### Annexe I : Extrait du questionnaire

1) Veuillez cocher la case qui correspond le plus à votre jugement :

	1	2	3	4	5
<input checked="" type="checkbox"/> 1.a La modernité de l'équipement et le mobilier du magasin					
<input checked="" type="checkbox"/> 1.b L'attractivité et le design du magasin					
1.c La propreté des différents services offerts dans le magasin					
<input checked="" type="checkbox"/> 1.d La disponibilité des marchandises à temps pour la clientèle					
<input checked="" type="checkbox"/> 1.e La disponibilité du personnel à répondre aux questions					
<input checked="" type="checkbox"/> 1.f La qualification et l'expérience du personnel					
<input checked="" type="checkbox"/> 1.g Votre degré de confiance à l'égard du personnel					
<input checked="" type="checkbox"/> 1.h La variété des marchandises					

où (1) = mauvais(e), (2) = moyen(ne) , (3) = normal(e), (4) = acceptable et (5) = excellent(e)

2) Le nombre de fois, par semaine, où vous fréquentez Champion :

1 fois ... 2 fois ... 3 fois ... 4 fois et + ...

3) Le nombre de produits achetés auprès de Champion par semaine ...

4) Quel est votre sexe ?

5) Quel est votre âge ?

6) Quel est votre revenu ?

7) Veuillez indiquer votre Catégorie Socio-Professionnelle

Etudiant ... Retraité ... Cadre ... Ouvrier ... Profession libérale ...

## Annexe II

```
> summary(df)
```

Q1a		Q1b		Q1c		Q1d		Q1e	
Min.	:1.000	Min.	:1.0	Min.	:1.000	Min.	:1.00	Min.	:1.000
1st Qu.	:3.000	1st Qu.	:2.0	1st Qu.	:2.000	1st Qu.	:2.00	1st Qu.	:2.000
Median	:3.000	Median	:3.0	Median	:3.000	Median	:3.00	Median	:2.000
Mean	:2.817	Mean	:2.7	Mean	:2.783	Mean	:2.65	Mean	:2.517
3rd Qu.	:3.000	3rd Qu.	:3.0	3rd Qu.	:3.000	3rd Qu.	:3.00	3rd Qu.	:3.000
Max.	:4.000	Max.	:4.0	Max.	:4.000	Max.	:4.00	Max.	:4.000

Q1f		Q1g		Q1h		Q2		Q3	
Min.	:1.0	Min.	:1.000	Min.	:1.000	1 fois	:22	Min.	: 0.00
1st Qu.	:2.0	1st Qu.	:2.000	1st Qu.	:2.000	2 fois	:16	1st Qu.	: 3.00
Median	:2.0	Median	:3.000	Median	:2.000	3 fois	: 6	Median	:10.00
Mean	:2.6	Mean	:2.517	Mean	:2.183	4 fois et +:	16	Mean	:17.00
3rd Qu.	:3.0	3rd Qu.	:3.000	3rd Qu.	:3.000			3rd Qu.	:25.25
Max.	:4.0	Max.	:4.000	Max.	:4.000			Max.	:75.00

Q4		Q5		Q6		Q7	
Feminin	:36	Min.	:18.00	Min.	: 50.0	Cadre	:11
Masculin	:24	1st Qu.	:22.00	1st Qu.	: 242.5	Etudiant	:11
		Median	:29.00	Median	: 590.0	Ouvrier	:10
		Mean	:31.22	Mean	: 615.3	Profession liberale	: 9
		3rd Qu.	:37.00	3rd Qu.	: 900.0	Retraite	:19
		Max.	:60.00	Max.	:1750.0		

```
## Analyse en Composantes Principales
> library(FactoMineR)
> res.pca <- PCA(df, quanti.sup=?. , quali.sup=?. , scale =?.)
> round(res.pca$eig[1:6,1],3)
comp 1 comp 2 comp 3 comp 4 comp 5 comp 6
1.827 1.668 1.232 0.860 0.834 0.659
> round(res.pca$var$coord[,1:4],2)
Dim.1 Dim.2 Dim.3 Dim.4
Q1a 0.01 0.80 -0.04 -0.29
Q1b 0.06 0.75 -0.29 -0.19
Q1c 0.04 0.59 0.13 0.77
Q1d 0.02 0.30 0.70 -0.09
Q1e 0.69 -0.06 -0.13 0.14
Q1f 0.72 -0.10 -0.32 0.19
Q1g 0.83 0.09 0.05 -0.29
Q1h 0.38 -0.12 0.72 -0.01
```

```
## Classification des individus
> library(cluster)
> library(FactoMineR)
> classif<-agnes(scale(df[,1:8]), method="ward")
> plot(classif,xlab="individu",main="")
> classes<-cutree(classif,k=3)
```



```
> df.comp<-cbind.data.frame(df, as.factor(classes))
> res.cat=catdes(df.comp, num.var=?)
> res.cat
```

Link between the cluster variable and the categorical variables (chi-square test)  
=====

```
      p.value df
Q7 0.02234873  8
```

#### ✓ Description of each cluster by the categories =====

```
$'1'
      Cla/Mod  Mod/Cla  Global  p.value  v.test
Q7=Retraite 57.89474 61.11111 31.66667 0.002373397 3.039032
Q2=2 fois   6.25000  5.55556 26.66667 0.014097627 -2.454766
```

```
$'2'
      Cla/Mod  Mod/Cla  Global  p.value  v.test
Q2=2 fois   87.50000 37.83784 26.66667 0.01337891 2.47352
Q7=Retraite 42.10526 21.62162 31.66667 0.04212738 -2.03226
```

```
$'3'
      Cla/Mod  Mod/Cla  Global  p.value  v.test
Q7=Ouvrier   30       60 16.66667 0.03085299 2.158964
```

Link between the cluster variable and the quantitative variables  
=====

```
      Eta2      P-value
Q1g 0.5223793 7.142817e-10
Q1a 0.3212950 1.595551e-05
Q1f 0.2880678 6.230180e-05
Q1d 0.2405124 3.934043e-04
Q1c 0.1733503 4.402131e-03
Q1b 0.1578047 7.486275e-03
```

#### ✓ Description of each cluster by quantitative variables =====

```
$'1'
      v.test Mean in category Overall mean sd in category Overall sd      p.value
Q1g  4.851857      3.333333      2.516667      0.5773503  0.8463976 1.223106e-06
Q1f  3.718650      3.277778      2.600000      0.7307192  0.9165151 2.002904e-04
Q1e  2.322074      2.888889      2.516667      0.5665577  0.8060535 2.022895e-02
Q1d -2.126293      2.333333      2.650000      0.4714045  0.7488881 3.347889e-02
Q1c -2.579634      2.444444      2.783333      0.5983516  0.6605974 9.890506e-03
```

```
$'2'
      v.test Mean in category Overall mean sd in category Overall sd      p.value
Q1d  3.498383      2.918919      2.650000      0.7119967  0.7488881 4.680878e-04
Q1c  3.195348      3.000000      2.783333      0.6151247  0.6605974 1.396622e-03
```

Q1a	2.707253	2.972973	2.816667	0.4924559	0.5624846	6.784254e-03
Q1e	-1.998077	2.351351	2.516667	0.8766094	0.8060535	4.570833e-02
Q1g	-2.525018	2.297297	2.516667	0.6091583	0.8463976	1.156923e-02
Q1f	-4.079528	2.216216	2.600000	0.8099094	0.9165151	4.512722e-05

\$'3'

	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
Q1d	-2.628643	1.8	2.650000	0.7483315	0.7488881	8.572619e-03
Q1b	-3.019012	1.8	2.700000	0.7483315	0.6904105	2.536008e-03
Q1g	-3.602725	1.2	2.516667	0.4000000	0.8463976	3.148989e-04
Q1a	-4.185985	1.8	2.816667	0.4000000	0.5624846	2.839325e-05

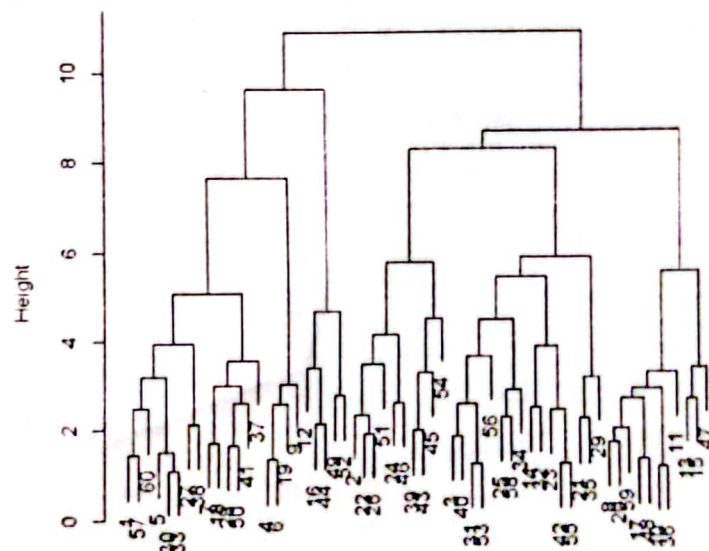


FIGURE 1 - Classification hiérarchique des individus

## Analyse Factorielle des Correspondances  
> table(Q2,Q7)

Q2	Q7				
	Cadre	Etudiant	Ouvrier	Profession liberale	Retraite
1 fois	4	4	3	4	7
2 fois	7	1	4	3	1
3 fois	0	0	1	2	3
4 fois et +	0	6	2	0	8

Somme.

11

11

10

9

19

Somme P2

22

16

6

16

60

RC  
↑

5

profil  
colonne

profil  
ligne  
↓

```
> tab<-table(Q2, Q7)
> res.ca<-CA(tab,graph=T)
> summary(res.ca)
Call:
CA(X = tab, graph = T)
The chi square of independence between the two variables is equal to 25.84882
(p-value = 0.01127501 ).
```

#### Eigenvalues

	Dim.1	Dim.2	Dim.3
Variance	0.336	0.090	?

#### Rows

	Iner*1000	Dim.1	ctr	cos2	Dim.2	ctr	cos2
1 fois	4.507	0.007	0.005	0.004	-0.063	1.596	0.319
2 fois	183.214	0.810	52.101	0.955	0.169	8.430	0.041
3 fois	69.688	-0.124	0.456	0.022	-0.821	74.929	0.968
4 fois et +	173.405	-0.773	47.437	0.919	0.225	15.045	0.078

#### Columns

	Iner*1000	Dim.1	ctr	cos2	Dim.2	ctr	cos2
Cadre	161.191	0.894	43.601	0.909	0.282	16.191	0.090
Etudiant	93.010	-0.596	19.401	0.701	0.385	30.188	0.292
Ouvrier	15.909	0.275	3.739	0.789	0.039	0.279	0.016
Profession liberale	67.382	0.424	8.016	0.400	-0.514	43.968	0.587
Retraite	93.321	-0.517	25.243	0.909	-0.163	9.374	0.090

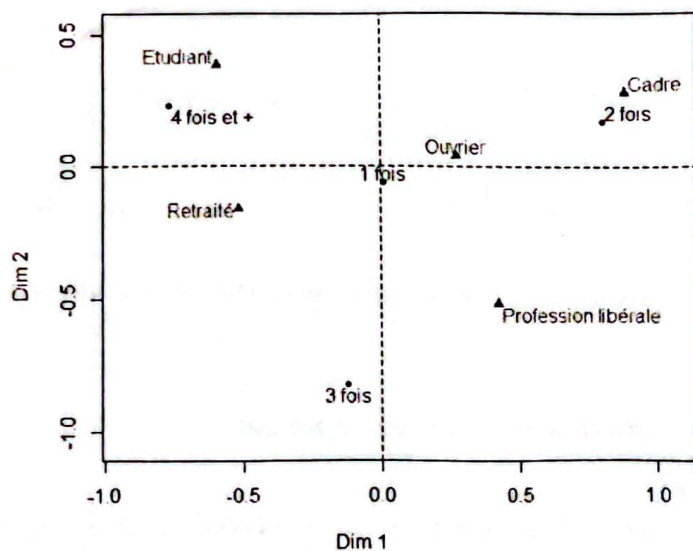


FIGURE 2 – Carte de représentation simultanée



① l'intérêt de cette ACP.

→ Déterminer un petit groupes de nouvelles variables ( - - - ) décrivant les - - - .

→ Regrouper les - - - ayant les mêmes caractéristiques puis décrire ces groupes.

④

Demande de l'ACP.

→ pertinence.

→ choix des axes

→ justification.

→ Interp des axes } qualité

→ carte d'indiv et interp. } correction.

② quant. sup = 10, 12, 13.

qual. sup = 9, 11, 14.

Scale = TRUE.

Variables

dans l'étude de l'ACP (1-8)

... max - min : كان تلقى ←

→ quant. sup

كان تلقى حادثة أخرى ←

→ qual. sup.

① choix des axes en se basant sur 3 critères

→ Critère de Kaiser

→ l'axe d'inertie cumulé

→ le boude.

③ choix des axes :

→ critère de Kaiser :  $\lambda_1, \lambda_2, \dots, \lambda_p$  : valeurs propres :

il ya 3 up supérieur à 1 Alors on retient 3 axes.

→ critère de l'axe cumulé :

$$\sum_{i=1}^p \lambda_i$$

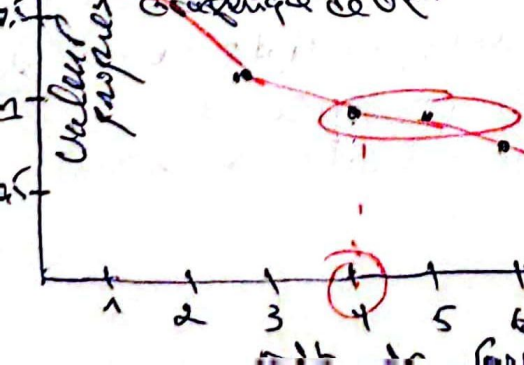
$$\frac{\sum_{i=1}^p \lambda_i}{\sum_{i=1}^p \lambda_i} = 0,7 \dots \rightarrow 70\%$$

les 4 premiers axes, ressume approx de 70% de l'inf Générale.

Alors c'est légitime de retenir les 4 premiers axes.

→ Critère de boude :

Graphique de UP.



$$\text{axe 1 : } \frac{1,828}{8} =$$

$$\text{axe 2 : } \frac{1,668}{8} =$$

$$\text{axe 3 : } \frac{1,238}{8} =$$

$$\text{axe 4 : } \frac{0,860}{8} =$$

$$\text{axe 5 : } \frac{0,834}{8} =$$

$$\text{axe 6 : } \frac{0,659}{8} =$$

le boude est au niveau de 4ème axe.

Alors on retient soit 3 ou 4

→ on retient 4.




- ④ interprétation: res. pca \$ var \$ load...
- axe 1:  $\{Q1e, Q1f, Q1g\}$  sont fortement corréliés avec l'axe 1 et entre eux.
- axe 2:  $\{Q1a, Q1b\}$  sont "la qualité personnelle" avec l'axe 2, et entre eux.
- axe 3:  $\{Q1d, Q1h\}$  "le décor du magasin"
- axe 4:  $\{Q1c\}$  "la disponibilité et la variété des marchandises"
- "hygiène des services"

Partie II:

- ① meilleur nb de class:
- on loupe au niveau de 3: car la répartition apparaît intéressante. en 3 classes.

- ⑤ catdes (... , num.var = 15)
- 14 variable de df
- 1: df.comp < (... , as.factor)

- ⑥  $v.test > 0 \rightarrow$  variable de présence important.
- $v.test < 0 \rightarrow$  variable de présence faible. 

cluster 1: on remarque une présence important de  $Q1g, Q1f$  et  $Q1e$ . ( $v.test > 0$ ) et une présence faible de  $Q1d$  et  $Q1c$  ( $v.test < 0$ ).

$\rightarrow$  le 1<sup>er</sup> cluster focalise sur la qualité des personnes consiste principalement des gens retraités.

cluster 2: " de  $Q1d, Q1c$  et  $Q1a$  ( $v.test > 0$ ) et une présence faible de  $Q1e, Q1f$  et  $Q1g$  ( $v.test < 0$ ) de plus: on remarque une faible présence de retraite ( $v.test < 0$ ) et présence forte de 2 fois ( $v.test \neq 0$ ).

$\rightarrow$  2<sup>e</sup> cluster focalise sur l'hygiène, disponibilité et la modernité des services (et la diversité)



cluster 3: " une faible présence de Q15, Q18, Q1a et Q1d  
 (v. test < 0) de plus, une forte présence des variables (v. test > 0)  
 les variables qui sont neutres pour la modernité  
 et disponibilité des services et la qualité de  
 personnel.

Partie III :

7) AFC est pertinente :

Value de test Khi-deux entre les 2 variables est très supérieur  
 à ce minimum.  $\hookrightarrow$  Donc il y a une dépendance entre les  
 variables met en question.

8)  $UP_3 = 0,005$

ou a :  $\sum_{i=1}^2 UP_i = 0,336 + 0,08 = 0,426$ .

$\left[ \text{Inter} \times 1000 \right] \hookrightarrow \sum \frac{\text{Inter} \times 1000}{1000} = X$ .

Alors  $X = 0,426 \approx 0,05 = \underline{\underline{UP_3}}$ .

9) Tableau de profil lignes.

	Cadre	étudiant	ouvrier	prof	Retraite	Somme
1 fois	$\frac{4}{100} \cdot 100 = 4$	$\frac{4}{22} \cdot 100 = 18$	$\frac{3}{22} \cdot 100 = 14$	$\frac{4}{22} \cdot 100 = 18$	$\frac{2}{22} \cdot 100 = 9$	100
2 fois	$\frac{7}{76} \cdot 100 = 44$	6	25	19	6	100
3 fois	0	0	17	33	50	100
4 fois +	0	38	13	0	50	100
profil moyen	15	15	17	18	35	100

Tableau de profil colonnes.

						profil moyen
1 fois	$\frac{4}{11} \cdot 100 = 36$	$\frac{4}{11} \cdot 100 = 36$	$\frac{3}{10} \cdot 100 = 30$	$\frac{4}{9} \cdot 100 = 44$	$\frac{2}{19} \cdot 100 = 11$	37
2 fois	$\frac{7}{71} \cdot 100 = 99$	$\frac{2}{11} \cdot 100 = 18$	$\frac{4}{40} \cdot 100 = 10$	33	5	30
3 fois	0	0	10	22	16	10
4 fois +	0	55	20	0	42	23
profil moyen	1m	1m	1m	1m	1m	1m

Interpretation :

② axe à retenir : les 2 premières axes resument 99 % de l'information  
générale en effet :  $\frac{0,336 + 0,09}{0,432} = 99\%$

• en gauche et en haut : présence des variables

"Etudiant" et "4 fois et +"

↳ les étudiants sont ceux qui visitent  
fréquemment le magasin.

avec le "retraite" entre "4 fois et +"  
et "3 fois".

↳ les étudiants et les retraités ont  
le plus de temps libre et moins  
de responsabilités que les autres

• en droite en haut : présence des variables.

"cadre" et "2 fois" et en bas la présence de  
variable "profession libérale" entre "2 fois"  
et "3 fois".

↳ les cadres et les prof lib ont  
moins ~~temps~~ de temps libre que  
les étudiants et les retraités.

"ouvrier" ou "trouvé".

à cause de manque de temps

"ouvrier" et "1 fois"

↳ est expliqué par le fait que  
les ouvriers ont moins de  
temps libre à cause de  
travail et manque d'argent  
↳ visite seulement 1 fois  
par semaine.

ADAM BACCOUCH  
Yahia  
Chammou