EISTI	ING1-GM I : DATA EXPLORATION - EXAMEN 2017-2018
Rédigé par : l'équipe enseignante	Durée : 1h30
A l'intention de : Elèves d'ING1-GM I	Document ou matériel autorisés : Calculatrice de l'EISTI uniquement

Exercice 1 (8,5)

Dans cet exercice, on considère les points de $\mathbb{R}^2$ suivants :

M1(2,0), M2(0,1), M3(0,2) ,M4(2,2), M5(7,0), M6(8,0), M7(0,8) et M8(8,8).

On applique la méthode de k-means pour répartir les points en deux classes, avec l'initialisation G1=M7 et G2=M8, et distance de Manhattan entre les points :

$$d_{Manhattan}(M_i,M_j)=|x_i-x_j|+|y_i-y_j|$$

(2,5) 1) Quels sont les clusters calculés à l'initialisation ?

	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8
$G_1=M_7$	10	9	6	8	15	16	0	8
$G_2=M_8$	14	15	14	12	9	8	8	0

$C_1 = \{M_1, M_2, M_3, M_4, M_7\}$

$C_2 = \{M_5, M_6, M_8\}$

(2) 2) Quels sont les nouveaux centres de gravité ?

$G_1 = \frac{1}{5}(2+0+0+2+0, 0+1+2+2+8) = (0.8 ; 1.625)$

$G_2 = \frac{1}{3}(7+8+8 ; 0+0+8) = (7.7 ; 2.7)$

(2) 3) Quand doit-on stopper l'algorithme ?

Quand les points ne changent plus de cluster

ou

Quand les centres de gravité ne bougent plus

(2,5) 4) Citez au moins un inconvénient à l'algorithme des k-means ?

- Initialisation aléatoire des centres de gravités
- nb de clusters à fixer a priori

Exercice 2 (8)

Une association de consommateurs soumet 20 piles de trois marques différentes à un même usage et mesure la ~~durée~~ durée de vie des piles. Les observations sont les suivantes.

Durée de vie pour la marque A									
~~65.1~~	~~58.4~~	~~64.9~~	~~76~~	~~67.8~~	~~75.1~~	~~76.7~~	~~64.2~~	~~74.9~~	77.6
~~58.1~~	~~68.1~~	~~73.3~~	~~75.4~~	~~76~~	~~59.4~~	~~65.4~~	~~74.7~~	~~76.6~~	81.3
Durée de vie pour la marque B									
64.4	69.1	66.9	67.5	65.8	70.4	67.8	61.8	68.7	65.3
63.7	68.5	72	67.5	71.8	64	69.5	66.8	64.9	63
Durée de vie pour la marque C									
62.8	58.6	63.3	65.3	78.8	63.1	76.3	64.2	61.8	73.9
73.8	76.9	78.4	69.3	63.7	73.7	70.9	63	74.4	64.4

Le tableau suivant donne les résumés numériques classiques pour chaque marque.

	Marque A	Marque B	Marque C
Médiane : M	74	67.2	67.3
Moyenne : $\bar{x}$	70.45	66.97	68.83
(x_{min}, x_{max}) E	(58.1 ;81.3) 23.2	(61.8 ;72) 10.2	(58.6 ;78.8) 20.2
(Q_1, Q_3) IQ	(• ;•) •	(64.65 ;68.9) 4.25	(63.2 ;74.15) 10.95
Variance : s	6.91	2.78	6.28

(2,5)

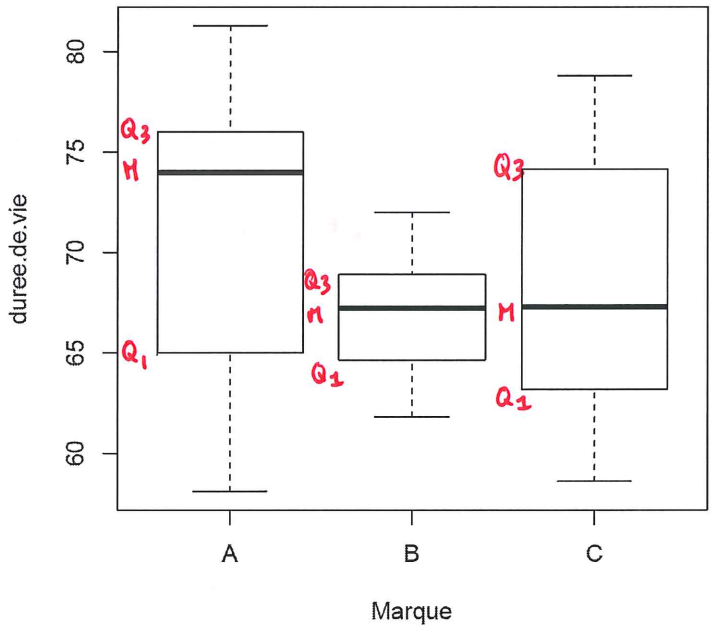
1) Complétez la case manquante de la marque A. Justifiez votre réponse

série ordonnée : 58.1 ; 58.4 ; 59.4 ; 64.2 ; 64.9 ; 65.1 ; 65.4 ; 67.8 ; 68.1 ; 73.3 ; 74.7 ; 74.9 ; 75.1 ; 75.4 ; 76 ; 76 ; 76.6 ; 76.7 ; 77.6 ; 81.3

$\frac{n}{4} = \frac{20}{4} = 5$, donc $Q_1 \rightarrow Q_1 = 64.9$; $Q_3 \rightarrow Q_3 = 76$

2) Placez les éléments qui ont permis de tracer les box-plots sur les graphiques associés.

(1,5)



3) Commentez les box-plots et le tableau ci-dessus. Quelle marque de pile auriez-vous achetée ?

(1)

la durée médiane est moins importante pour les marques B et C. On aura tendance à privilégier la marque A. Attention cependant car la marque A présente une grande variabilité de ses performances

4) Calculez la durée de vie moyenne toutes marques confondues.

(0,5)

$\bar{x} = \left(n_A \bar{x}_A + n_B \bar{x}_B + n_C \bar{x}_C \right) \times \frac{1}{n}$

$= (70.45 + 66.97 + 68.83) \times \frac{20}{60} = 68.75$

si problème avec question 5 => 2,5 pour questions 1 à 7 même si pas dans le bon ordre

5) Calculez la variance ~~totale~~. intra

$$V_{intra} = \frac{1}{n}\sum_{\ell=1}^{p} n_\ell s_\ell^2 = \frac{1}{60}\left(20\times 6.91 + 20\times 2.78 + 20\times 6.88\right) = 5,32$$

(1)

6) Calculez la variance inter-groupes $\frac{1}{n}\sum_{\ell=1}^{p} n_\ell (\bar{y}_\ell - \bar{y})^2 = \frac{1}{60}\left(20(70.45-68.75)^2 + 20(66.97-68.75)^2 + 20(68.83-68.75)^2\right) = 2,0216 = V_{inter}$

(1)

7) Quelles part de variance de la durée de vie des piles est expliquée par la marque ?

$V_{TOT} = V_{intra} + V_{inter} = 7,34$

(0,5) $\Rightarrow R^2 = \frac{V_{inter}}{V_{TOT}} = 0,275$

27,5 % de la variabilité de la durée de vie des piles est expliqué par la marque

Exercice 3 (5)

1) Une analyse en composantes principales d'un nuage d'individus décrit par des variables quantitatives cherche à :

A) Augmenter l'inertie

B) Maximiser l'inertie du nuage en projection dans des plans

C) Représenter des individus dans un cercle

(1)

2) En ACP, l'inertie correspond à :

A) La variance des variables

B) L'espérance des variables

C) La corrélation des variables

(1)

3) En ACP, un vecteur propre correspond à :

A) Un axe de projection optimale du nuage

B) La corrélation entre deux composantes principales

C) La variance d'une composante principale

(1)

4) Une valeur propre correspond à :

A) Un axe de projection optimale du nuage

B) La corrélation entre deux composantes principales

C) La variance d'une composante principale

(1)

5) Une ACP normée est une ACP pour laquelle :

A) Toutes les variables suivent une loi normale

B) Toutes les variables ont pour variance 1

C) Toutes les variables varient entre 0 et 1

(1)

Exercice 4

Une entreprise de vente par correspondance vend 5 produits (P_1,P_2,P_3,P_4,P_5) dans 4 régions (Ouest, Sud, Est, Nord). Un échantillon aléatoire de 680 commandes a permis d'établir tous les résultats donnés en annexex.

- Les tableaux 1 et 2 présentent les effectifs et les fréquences observés.
- Les tableaux 3 et 4 donnent les profils lignes et les profils colonnes.
- Le tableau 5 représente les écarts du chi-deux
- Le tableau 6 donne les seuils de comparaison en fonction des dégrés de liberté (d.d.l.).
- Le tableau 7 affiche les résultats sur les valeurs propres d'une AFC effectuée avec R
- Le tableau 8 affiche les résultats sur les profils-lignes d'une AFC effectuée avec R
- Le tableau 8 affiche les résultats sur les profils-colonnes d'une AFC effectuée avec R
- La figure 1 représente le plan principal de l'AFC

1) Quelles affirmations sont vraies ?

   A) La probabilité qu'une commande choisie au hasard provienne de la région « Ouest » est 0.21

   B) La probabilité qu'une commande choisie au hasard concerne le produit « P_1 » est 0.24

   C) Il y a 12% de chance qu'un produit P_2 provienne de la région « Sud »

   D) Il y a 21% de chance qu'une commande choisie au hasard concerne un produit P_1 provenant de la région « Sud »

2) On note X la variable « région » et Y la variable « produit ». A quel type de probabilités correspondent les lignes du tableau 3 ?

   A) P(X|Y) B) P(Y|X) C) P(X∩Y) D) P(X∪Y)

3) La ~~première~~ dernière case du tableau 3 vaut

   A) 0.09 B) 0.19 C) 0.14 D) 0.018

4) Dans le tableau 3, pour une colonne donnée, peut-on faire la somme des lignes ?

   A) Oui, la somme vaut 1

   B) Non, cela ne veut rien dire

   C) Oui, la somme vaut la dernière ligne du tableau 2

5) Dans le tableau 3, pour une ligne donnée, peut-on faire la somme des colonnes ?

   A) Oui, la somme vaut 1

   B) Non, cela ne veut rien dire

   C) Oui, la somme vaut la dernière colonne du tableau 2

6) Dans le tableau 5, que valent les cases vides (aux arrondis près):

   A) (*) = 0.45 et (**) = 6.03 B) (*)=5.51 et (**)=1.01 C) (*)=0.19 et (**)=6.33

7) A l'aide du tableau 6, déterminez le seuil pour la distance du chi-deux

   A) 9.49 B) 11.07 C) 7.82 D) 18.31 E) 21.03 F) 26.3 G) 31.41

8) Peut-on affirmer que les deux variables sont dépendantes ?

   A) OUI B) NON

Cette partie concerne les résultats de l'AFC

9) Pourquoi y-a-t-il 3 dimensions ?

nb dimension = $\min\{p-1, q-1\} = \min\{5-1, 4-1\} = 3$
où p et q = nb modalités des variables

10) A quoi correspondent les pourcentages affichés sur les axes 1 et 2 de la figure 1 ?

Cela correspond au pourcentage de la distance du $\frac{\chi^2}{n}$ expliquée par les axes

11) Quels axes faudrait-il afficher pour avoir une bonne repprésenation de la région « Ouest ». Justifiez.

La région ouest est bien représentée sur l'axe 1 (47%) et l'axe 3 (53%)

12) A l'aide des tableaux 3, 4 et 5, expliquez pourquoi :

a. P_4 et P_2 sont les plus éloignés du cetre du graphique
b. « Est » est la région la plus éloignée du centre du graphique
c. « Est » et P_2 sont proches

a) On compare les profils colonnes (tab 4) avec le profil moyen des régions
On remarque que P_2 s'éloigne beaucoup du profil moyen avec une sur-représentation de la région Est (45% contre 21%). Idem pour P_4 avec une sur-représentation de la région Sud (55% contre 25%) et une sous-représentation de la région Est (9% contre 21%)

b) On compare le profil ligne de la région Est (tab 3) avec le profil moyen des produits. La région Est vend plus de produits P_2 que la moyenne (45% contre 21%) mais moins de produits P_4 (8% contre 17%) et moins de produit P_5 (5% contre 12%)

c) Car leur distribution théorique est très éloignée de la distribution observée (tab. 5). Cette case contribue donc très fortement au χ^2.

ANNEXE

	P_1	P_2	P_3	P_4	P_5	**Total**
Ouest	28	14	45	33	12	**132**
Sud	36	21	25	64	23	**169**
Est	21	64	38	11	7	**141**
Nord	79	42	67	9	41	**238**
Total	**164**	**141**	**175**	**117**	**83**	**680**

Tableau 1 : Effectifs observés

	P_1	P_2	P_3	P_4	P_5	**Total**
Ouest	0,041	0,021	0,066	0,049	0,018	**0,194**
Sud	0,053	0,031	0,037	0,094	0,034	**0,249**
Est	0,031	0,094	0,056	0,016	0,010	**0,207**
Nord	0,116	0,062	0,099	0,013	0,060	**0,350**
Total	**0,241**	**0,207**	**0,257**	**0,172**	**0,122**	**1**

Tableau 2 : fréquences observées

	P_1	P_2	P_3	P_4	P_5
Ouest	0,21	0,11	0,34	0,25	•
Sud	0,21	0,12	0,15	0,38	0,14
Est	0,15	0,45	0,27	0,08	0,05
Nord	0,33	0,18	0,28	0,04	0,17

Tableau 3 : Tableau des profils lignes

	P_1	P_2	P_3	P_4	P_5
Ouest	0,17	0,10	0,26	0,28	0,14
Sud	0,22	0,15	0,14	0,55	0,28
Est	0,13	0,45	0,22	0,09	0,08
Nord	0,48	0,30	0,38	0,08	0,49

Tableau 4 : Tableau des profils colonnes

	P_1	P_2	P_3	P_4	P_5	**Total**
Ouest	(*)	6,485	3,632	4,684	1,042	
Sud	0,566	5,632	7,878	41,768	0,266	
Est	4,923	41,713	0,092	7,209	(**)	
Nord	8,166	1,072	0,556	24,915	4,930	
Total						**171.84**

Tableau 5 : Tableau des écarts du chi-deux

d.d.l.	1	2	3	4	5	6	7	8	9	10
seuil	3.84	5.99	7.82	9.49	11.07	12.59	14.07	15.51	16.92	18.31
d.d.l.	11	12	13	14	15	16	17	18	19	20
seuil	19.68	21.03	22.36	23.68	25	26.3	27.59	28.87	30.14	31.41

Tableau 6. Seuils pour la distance du chi-deux

```
> round(res.ca$eig,3)
eigenvalue percentage of variance cumulative percentage of variance
dim 1       0.155                61.425                           61.425
dim 2       0.080                31.776                           93.201
dim 3       0.017                 6.799                          100.000
```

Tableau 7. Résultats sur les valeurs propres de l'AFC

```
> Résultats sur les lignes
> round(cbind(res.ca$row$coord[, 1:3], res.ca$row$cos2[, 1:3] ), 2)
Dim 1 Dim 2 Dim 3 Dim 1 Dim 2 Dim 3
Ouest  0.24  0.02 -0.25  0.47  0.00  0.53
Sud    0.55 -0.13  0.12  0.91  0.05  0.05
Est   -0.49 -0.43  0.01  0.57  0.43  0.00
Nord  -0.23  0.33  0.05  0.32  0.66  0.01
```

coord cos2

Tableau 8. Résultats sur les profils-lignes de l'AFC :
Les 3 premières colonnes correspondent aux coordonnées et les 3 dernières au cos²

```
> Résultats sur les colonnes
> round(cbind(res.ca$col$coord[, 1:3], res.ca$col$cos2[, 1:3] ), 2)
Dim 1 Dim 2 Dim 3 Dim 1 Dim 2 Dim 3
P1 -0.03  0.28  0.06  0.01  0.94  0.04
P2 -0.47 -0.39  0.09  0.58  0.40  0.02
P3 -0.14  0.08 -0.21  0.29  0.08  0.63
P4  0.77 -0.28  0.00  0.89  0.11  0.00
P5  0.08  0.34  0.17  0.04  0.77  0.19
```

Tableau 9. Résultats sur les profils-colonnes de l'AFC
Les 3 premières colonnes correspondent aux coordonnées et les 3 dernières au cos²

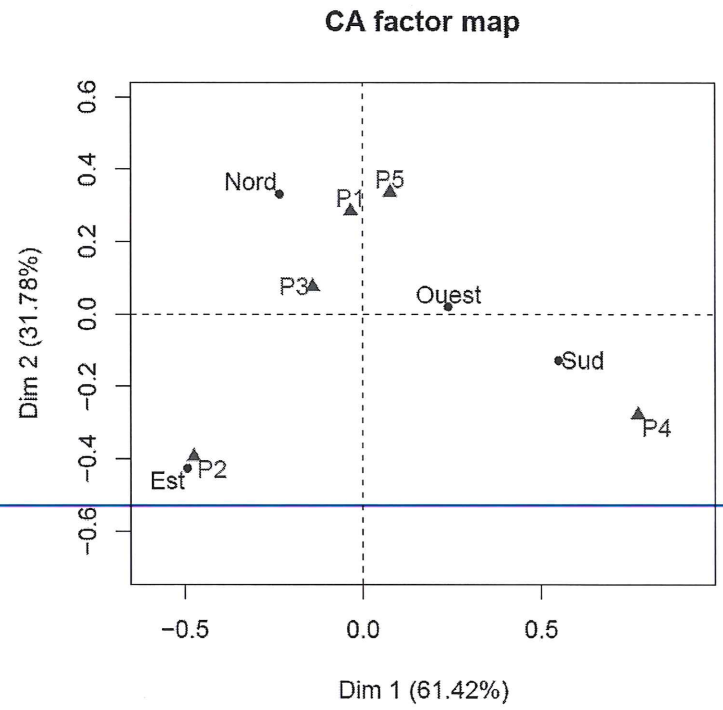


Figure 1. Représentation des résultats de l'AFC sur le plan principal