Faculté
des **sciences économiques** et de **gestion**
Université de Strasbourg

Année Universitaire 2021/2022
Licence Economie et Gestion, L3-S6

UE Techniques quantitatives
Examen : Analyse de données – Session 1 – Mai 2022

Durée de l'épreuve : **2h00.** Enseignant : M. EL OUARDIGHI
Documents autorisés : aucun
Les calculatrices type collège, non programmables et non graphiques, sont *autorisées.*
Barème indicatif : Bonne réponse : **0.5** points ; Absence de réponse : **0** points ; Mauvaise réponse : **-0.25** points.

Notes et consignes importantes

(i) Sept exercices, notés I. à VII. sont proposés pour un total de 40 questions types QCM numérotées de 1 à 40, NB. **Une seule réponse est valide.** (ii) Vos réponses doivent figurer uniquement sur le «formulaire type 9» et **Seul le formulaire type 9 est à remettre.** NB. Pour **éviter la note 0**, il est important de renseigner votre identifiant, i.e. n° d'anonymat, à défaut, reporter vous aux consignes de la dernière page du sujet, i.e. page 7.

Sujet

I. L'objectif des différentes méthodes dans le cadre de l'analyse des données. Considérer les affirmations suivantes et préciser votre choix.

1. La classification hiérarchique (CH) vise à regrouper :
A/ Les observations les plus semblables séquentiellement
B/ Les observations en k groupes simultanément
C/ Les observations les plus proches des centres de gravités des groupes

2. L'analyse factorielle des correspondances (AFC) est une méthode que l'on peut utiliser :
A/ Tout le temps, que les variables soient qualitatives ou quantitatives
B/ Uniquement lorsque les variables sont qualitatives et au nombre de deux
C/ Uniquement lorsque les variables sont qualitatives
D/ Seulement lorsque les variables sont quantitatives et au nombre de deux

3. L'analyse des composantes principales (ACP) est une méthode que l'on peut utiliser :
A/ Uniquement lorsque les variables sont qualitatives
B/ Tout le temps, que les variables soient qualitatives ou quantitatives
C/ Seulement lorsque les variables sont quantitatives

II. Dans le cadre d'une AFC (Analyse factorielle des correspondances), préciser si les affirmations suivantes sont vraies ou fausses.

4. Si l'angle entre deux flèches est aigu, alors il y a une forte association entre les lignes et les colonnes correspondantes.
A/ Vrai
B/ Faux

5. Pour interpréter la distance entre les lignes et les colonnes, on doit projeter perpendiculairement des points lignes sur la flèche de la colonne.
A/ Vrai
B/ Faux

6. On doit d'abord décider si on veut analyser les contributions des lignes ou celles des colonnes.
A/ Vrai
B/ Faux

7. L'hypothèse nulle d'un test de Khi-2 suppose qu'il y a indépendance entre les deux variables. Si cette hypothèse est acceptée, il n'y aura pas d'utilité à réaliser une AFC.
A/ Vrai
B/ Faux

8. Dans une représentation graphique, si les points projetés sont confondus avec le centre de gravité, le test de Khi-2 conclue à une indépendance entre les deux variables étudiées.

A/ Vrai

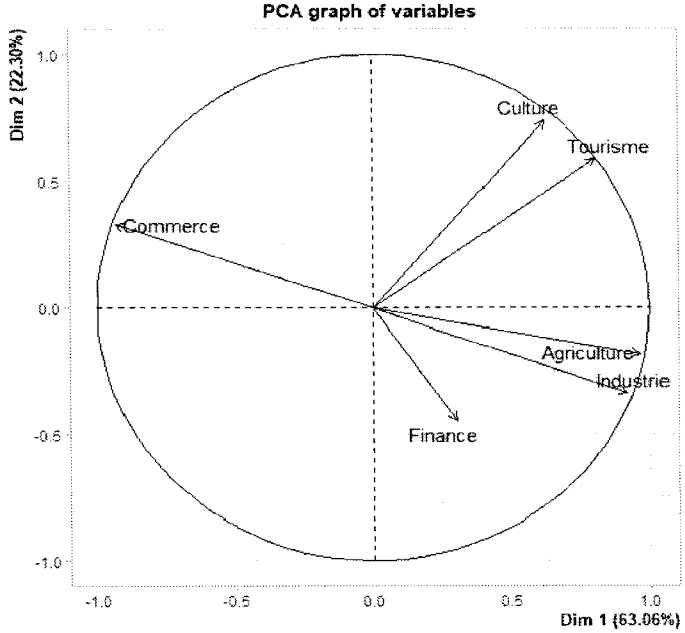
B/ Faux

9. Si le résultat du test indique "*Pearson's Chi-squared test :* $\chi^2_{df=29} = 113.2$, *p-value=0.0046*", on peut dire que les deux variables sont indépendantes.

A/ Vrai

B/ Faux

III. Considérons le « PCA graph » ci-dessous relatif aux caractéristiques prédominantes de certaines régions européennes.



10. Si l'on s'intéresse à la corrélation entre les différentes caractéristiques régionales, nous pouvons dire que la corrélation entre « Commerce » et « Industrie » est proche de :

A/ -1.0 ; B/ -0.5 ; C / 0.0 ; D/ +0.5 ; E/ +1.0 ;

11. Si l'on s'intéresse à la corrélation entre les différentes caractéristiques régionale, nous pouvons dire que la corrélation entre « Tourisme » et « Finance » est proche de :

A/ -1.0 ; B/ -0.5 ; C / 0.0 ; D/ +0.5 ; E/ +1.0 ;

12. Si l'on s'intéresse à la corrélation entre les axes et les caractéristiques régionales, nous pouvons dire que la corrélation entre l'axe 1 (Dim. 1) et « Commerce » est proche de :

A/ -1.0 ; B/ -0.5 ; C / 0.0 ; D/ +0.5 ; E/ +1.0 ;

13. Si l'on s'intéresse à la corrélation entre les axes et les caractéristiques régionales, nous pouvons dire que la corrélation entre l'axe 2 (Dim. 2) et « Tourisme » est proche de :

A/ -1.0 ; B/ -0.5 ; C / 0.0 ; D/ +0.5 ; E/ +1.0 ;

14. « Industrie » et « « Commerce » contribuent quasiment avec les mêmes proportions à l'axe 1.

A/ Vrai

B/ Faux

IV. Soit l'extrait des résultats d'une ACP (Analyse en composantes principales) suivant :

```
> Eig_corr = eigen(cor(Regio))
> Eig$values # valeurs propres
[1] 2.928993277 1.054479100 0.013924984 0.002602638
> Eig$vectors # vecteur propres
[,1]        [,2]       [,3]       [,4]
[1,] -0.4890926  0.5421561  0.2985589  0.6145875
[2,] -0.5178484  0.4314678 -0.4851478 -0.5570460
[3,] -0.4667101 -0.6032110 -0.5038817  0.4054891
[4,] -0.5242168 -0.3950174  0.6493053 -0.3841363
```

15. Les valeurs propres mesurent :

A/ la quantité de variance expliquée par chaque composante principale

B/ le degré de corrélation entre les composantes principales

C/ Aucune des deux réponses n'est valide

16. D'après les résultats, nous pouvons dire que :

A/ Une seule composante principale peut être considérée

B/ Deux composantes principales peuvent être considérées

C/ Trois composantes principales peuvent être considérées

D/ Quatre composantes principales peuvent être considérées

17. 99% de la variance totale sont expliquées par les trois premiers axes

A/ Vrai

B/ Faux

18. En pratique, une ACP normée utilise les données centrées (par rapport à la moyenne) et réduites (rapportées à l'écart-type).

A/ Vrai

B/ Faux

19. Dans le cadre d'une ACP normée, on peut vérifier que la matrice variance-covariance n'est rien d'autre que la matrice de corrélation entre les variables.

A/ Vrai

B/ Faux

20. L'ACP consiste à transformer k variables quantitatives en p composantes principales, avec $p<k$.

A/ Vrai

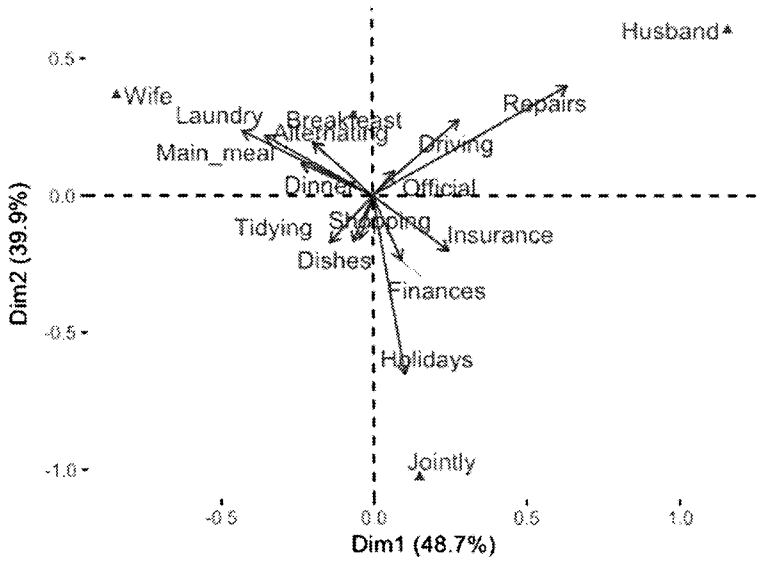
B/ Faux

21. Dans le cadre d'une ACP normée, l'inertie totale est égale au nombre de variables quantitatives (k) moins une, i.e. $k-1$.

A/ Vrai

B/ Faux

V. Considérons le Biplot ci-dessous relatif à une AFC (Analyse factorielle des correspondances). Préciser si les résultats suivants sont corrects, i.e. 'Vrai ' ou erronés, i.e. 'Faux':



22. Les distances entre les points lignes et l'origine du graphique sont liées à leurs contributions aux axes principaux.

A/ Vrai

B/ Faux

23. Plus une flèche est proche (en termes de distance angulaire) d'un axe, plus la contribution de la ligne sur cet axe est importante.

A/ Vrai
B/ Faux

24. Si la flèche est à mi-chemin entre les deux axes, la ligne contribue aux deux axes de manière identique.
A/ Vrai
B/ Faux

25. On peut observer que 'Repairs' a une contribution importante à la première dimension.
A/ Vrai
B/ Faux

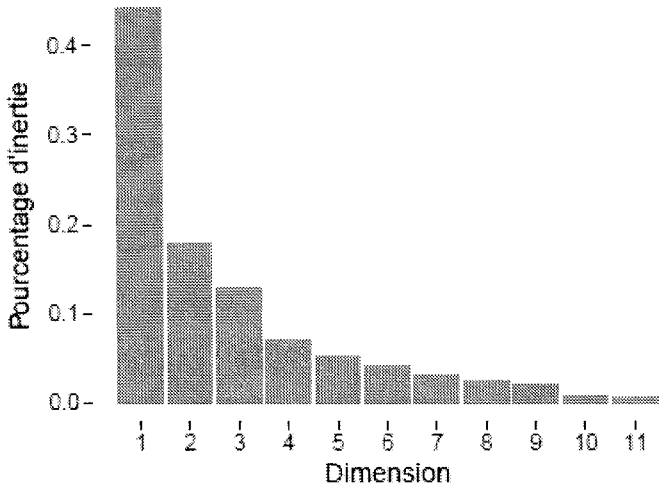
26. La deuxième dimension est principalement définie par la ligne 'Holidays'.
A/ Vrai
B/ Faux

27. Les contributions de 'Shopping' et 'Diner' apparaissent non significatives.
A/ Vrai
B/ Faux

28. les tâches de réparation (Repairs) et de conduite (Driving) sont plutôt du ressort du mari (Husdand) alors que la blanchisserie est réservée à la femme (Wife).
A/ Vrai
B/ Faux

29. L'organisation des vacances est décidée principalement par le mari.
A/ Vrai
B/ Faux

VI. Soit le graphique ci-dessous. Préciser si les affirmations suivantes sont vraies ou fausses :



30. L'inertie mesure la dispersion des points du nuage par rapport à son centre de gravité.
A/ Vrai
B/ Faux

31. Plus l'inertie est grande, plus le nuage est dispersé.
A/ Vrai
B/ Faux

32. Plus l'inertie est petite, moins le nuage est concentré autour de son centre de gravité.
A/ Vrai
B/ Faux

33. L'inertie totale du nuage de points est donc égale à la somme des variances des variables.
A/ Vrai
B/ Faux

34. Le graphique suggère de garder trois axes.

A/ Vrai

B/ Faux

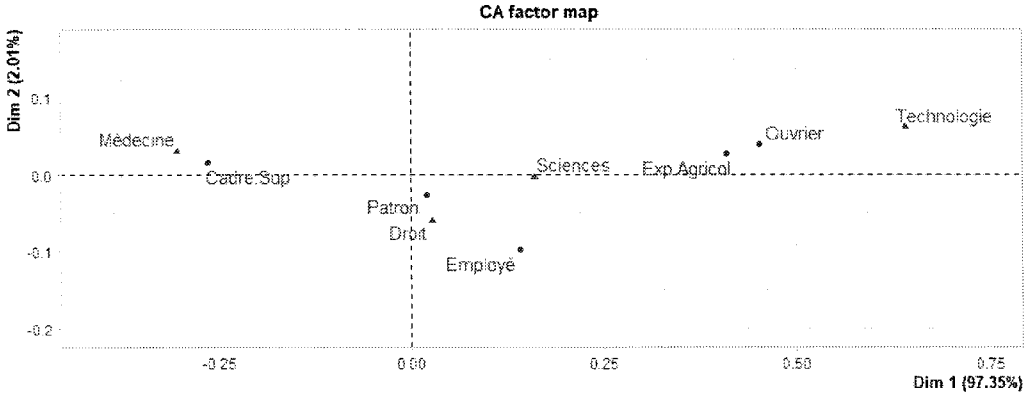
35. Nous pouvons estimer l'inertie captée par les deux premiers axes à environ 74%.

A/ Vrai

B/ Faux

VII. Considérons un jeu de données relatif au choix d'une filière universitaire des enfants selon la CSP (Catégorie socio-professionnelle) des parents. Une analyse a été appliquée à ces données dont certains résultats sont présentés ci-dessous.

```
> Choix
> chisq.test(Choix)
   Pearson's Chi-squared test
data:  Choix
X-squared = 320.27, df = 12, p-value < 2.2e-16
> qchisq(0.95, 12)
[1] 21.02607
> Choix.AFC$eig
       eigenvalue percentage of variance cumulative percentage of variance
dim 1 0.082393603             97.3495522                           97.34955
dim 2 0.001703449              2.0126558                           99.36221
dim 3 0.000539807              0.6377919                          100.00000
> Choix.AFC$col$contrib
                Dim 1      Dim 2    Dim 3
Droit        0.2585182 58.7585597 13.78948
Sciences     7.9446214  0.1115716 66.52097
Médecine    41.5839983 19.2593782  1.86804
Technologie 50.2128622 21.8704905 17.82151
> Choix.AFC$row$contrib
                 Dim 1     Dim 2     Dim 3
Exp.Agricol 16.29200620  3.229165 21.669399
Patron       0.07488716  6.304816 61.867745
Cadre.Sup   40.40124242  6.886489  3.433167
Employé      3.02413561 68.626434 10.316292
Ouvrier     40.20772861 14.953096  2.713396
> Choix.AFC$col$coord
                 Dim 1        Dim 2        Dim 3
Droit        0.02798724 -0.060669159  0.016544781
Sciences     0.16046170 -0.002734195 -0.037582581
Médecine    -0.30312512  0.029661814  0.005200252
Technologie  0.64017413  0.060748795  0.030869913
> Choix.AFC$row$coord
                 Dim 1        Dim 2        Dim 3
Exp.Agricol  0.41011544  0.02625317 -0.038283778
Patron       0.02015079 -0.02658535  0.046880589
Cadre.Sup   -0.26271704  0.01559580 -0.006198846
Employé      0.14209032 -0.09732566 -0.021242138
Ouvrier      0.45148105  0.03958841  0.009493228
```



36. Les résultats révèlent que le choix des enfants et la CSP des parents ne sont pas dépendants.

A/ Vrai

B/ Faux

37. Les deux premières dimensions concentrent presque la totalité de l'inertie.

A/ Vrai

B/ Faux

38. Les résultats font ressortir une préférence des enfants du « Cadre.Sup » pour :

A/ Médecine

B/ Technologie

C/ Sciences

D/ Droit

39. Les résultats font ressortir une préférence des enfants des exploitants agricoles et des ouvriers pour la filière

A/ Médecine

B/ Technologie

C/ Sciences

D/ Droit

40. L'axe 1 est principalement déterminé par le choix des filières

A/ Technologie et Médecine

B/ Technologie et Sciences

C/ Technologie et Droit

D/ Médecine et Sciences

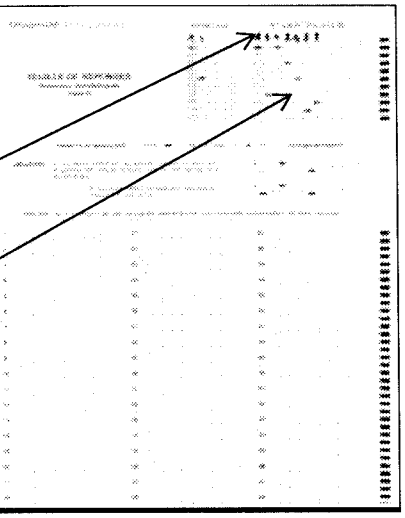
E/ Médecine et Droit

F/ Sciences et Droit

Consignes relatives aux renseignements à reporter sur le formulaire type 9

Identification

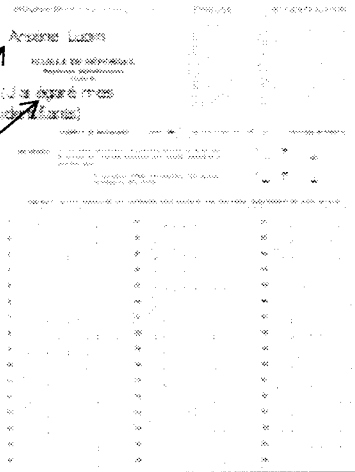
- Inscrivez dans la grille en haut à droite le numéro d'anonymat ou l'identifiant qui vous a été attribué.
- Puis codez chacun de ses caractères dans la colonne qu'il surplombe.
- Ne cochez pas plus d'une case par colonne !
- Le cas échéant, faites de même avec votre code épreuve selon les instructions de vos surveillants.



Identification

- Si vous ne pouvez utiliser le cadre d'identification, i.e. vous n'avez pas ou vous avez oublié vos identifiants, vous pouvez écrire dans la zone située sous le bloc « Université de Strasbourg», i.e. Nom & Prénom

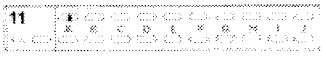
- N'écrivez jamais dans la marge : votre copie pourrait se voir attribuer une note aléatoire.



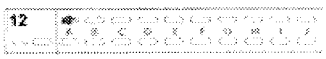
Codage d'une réponse

- Pour chaque réponse, deux lignes de dix cases sont proposées.
- La deuxième ligne sert à remplacer si nécessaire la réponse donnée à la première.
- Veillez à noircir correctement les cases de vos réponses pour atteindre le seuil de détection.

sera interprétée correctement en « A »



pourrait être interprétée en « A » ou en abstention



pourrait être interprétée en « A » ou en « AB »