



FACULTÉ DE DROIT
ÉCONOMIE & GESTION

Sujet d'examen - 1^{ère} session
Semestres impairs
Licence : 1-3-5 ; Master : 7-9
Année universitaire 2021-2022

Intitulé de l'épreuve :	Analyse de données qualitatives
Nom de l'enseignant :	Nils Berglund
Mention / Spécialité / Parcours :	M1 ESA
Année d'études :	2021-2022
Durée de l'épreuve :	2 heures
Documents autorisés :	Résumé manuscrit de 4 pages A4
Matériels autorisés :	Calculatrice non programmable
	16/12/2021

P1/ 6

SUJET

Les téléphones portables, ordinateurs et tablettes doivent être éteints durant l'examen.
Il sera tenu compte de la qualité de la rédaction.
Les points sont donnés à titre indicatif.

Problème 1 [4 points]

On considère la matrice

$$A = \begin{bmatrix} 20 & -4 \\ -4 & 5 \end{bmatrix}.$$

1. Déterminer les valeurs propres de A .
2. Déterminer, pour chaque valeur propre de A , un vecteur propre de norme 1.
3. Soit \mathcal{E}_c l'ellipse d'équation

$$\begin{bmatrix} x & y \end{bmatrix} A^{-1} \begin{bmatrix} x \\ y \end{bmatrix} = c^2 \quad (1)$$

Que vaut son demi-grand axe ? Son demi-petit axe ? Esquisser l'ellipse \mathcal{E}_c dans le plan (x, y) pour $c = 1$ et $c = \frac{1}{2}$.

4. Si A est la matrice de covariance d'un jeu de données, que représentent les ellipses \mathcal{E}_c ?

Problème 2 [6 points]

Le tableau de contingence suivant montre les effectifs pour un jeu de données croisant deux variables qualitatives. Chaque variable a 4 modalités. Celles de la première sont notées a, b, c, d, alors que les celles de la seconde sont notées A, B, C, D.

	a	b	c	d
A	20	40	40	100
B	100	0	200	100
C	40	80	80	0
D	40	40	80	40

1. Expliquer la propriété d'équivalence distributionnelle. Utiliser cette propriété afin de simplifier le tableau de contingence ci-dessus.
2. Effectuer un test d'indépendance du chi-deux *sur le tableau simplifié*, en expliquant les étapes principales du calcul. Peut-on rejeter l'hypothèse d'indépendance à un ou plusieurs seuils indiqués ci-dessous ?

Tableau de valeurs de la loi du χ^2 :

Degrés de liberté	90%	95%	97.5%	99%	99.9%
5	9.236	11.070	12.833	15.086	20.515
6	10.645	12.592	14.449	16.812	22.458
7	12.017	14.067	16.013	18.475	24.322
8	13.362	15.507	17.535	20.090	26.125
9	14.684	16.919	19.023	21.666	27.877
10	15.987	18.307	20.483	23.209	29.588
11	17.275	19.675	21.920	24.725	31.264
12	18.549	21.026	23.337	26.217	32.910
13	19.812	22.362	24.736	27.688	34.528
14	21.064	23.685	26.119	29.141	36.123
15	22.307	24.996	27.488	30.578	37.697
16	23.542	26.296	28.845	32.000	39.252

3. Que peut-on dire sur les nuages de points (dans quels espaces se trouvent-ils) ? Expliquer, sans faire de calcul, comment on détermine leur forme.

Problème 3 [7 points]

Les données ci-dessous représentent le symptôme principal ressenti par un échantillon de personnes atteintes d'une infection respiratoire, en fonction de leur classe d'âge. Les symptômes sont : fièvre, toux, pneumonie, malaise général, mal de gorge, nez bouché, diarrhée, mal de tête, douleurs musculaires, et nausée.

range	Fever (n=287)	Cough (n=287)	Pneumonia (n=194)	General malaise (n=195)	Sore throat (n=211)	Nasal discharge and/or congestion (n=150)	Diarrhea (n=154)	Headache (n=128)	Joint and/or Muscle pain (n=142)	Nausea and/or Vomiting (n=145)
under 10	1 (33%)	2 (67%)	0 (0%)	0 (0%)	0 (0%)	2 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
10-19	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (100%)				
20-29	13 (76%)	9 (53%)	2 (22%)	2 (14%)	2 (13%)	2 (29%)	1 (13%)	2 (25%)	1 (13%)	1 (13%)
30-39	14 (78%)	13 (72%)	3 (27%)	3 (25%)	6 (50%)	1 (11%)	2 (22%)	2 (25%)	0 (0%)	0 (0%)
40-49	19 (73%)	18 (69%)	8 (47%)	6 (30%)	3 (14%)	1 (7%)	4 (27%)	3 (20%)	1 (7%)	0 (0%)
50-59	32 (67%)	33 (69%)	24 (60%)	18 (44%)	16 (39%)	6 (19%)	6 (20%)	6 (20%)	1 (3%)	0 (0%)
60-69	37 (69%)	30 (56%)	28 (68%)	20 (48%)	9 (23%)	3 (9%)	7 (21%)	4 (11%)	2 (6%)	2 (6%)
70-79	56 (62%)	53 (58%)	45 (78%)	22 (33%)	14 (22%)	11 (24%)	8 (17%)	6 (13%)	5 (11%)	5 (12%)
80-89	15 (56%)	22 (81%)	10 (50%)	8 (35%)	8 (33%)	3 (15%)	1 (5%)	2 (10%)	0 (0%)	1 (5%)
over 90	1 (100%)	0 (0%)	1 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)

Ci-dessous on a représenté les résultats d'une AFC faite sur ces données.

Questions :

1. Expliquer les lignes suivantes fournies par le logiciel R:

```
Number of cases in table: 729
```

```
Number of factors: 2
```

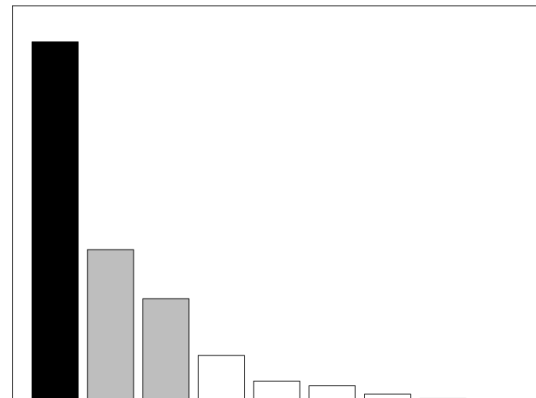
```
Test for independence of all factors:
```

```
Chisq = 88.11, df = 81, p-value = 0.2758
```

Commenter en particulier la signification de $p\text{-value} = 0.2758$.

2. Discuter les valeurs des inerties. Une étude restreinte aux deux premiers axes factoriels vous semble-t-elle pertinente ?
3. Quelles sont les modalités mal représentées sur le plan factoriel ?
4. Quelles sont les modalités contribuant fortement au premier axe factoriel ? Donner une description résumant la signification de cet axe.
5. Quelles sont les modalités contribuant fortement au second axe factoriel ? Que représente cet axe ?
6. Citer des exemples de symptômes mal représentés sur le *premier axe* factoriel. Qu'est-ce que cela signifie ?
7. Que pourrait-on faire afin d'améliorer la qualité de l'analyse, étant donné le nombre de modalités mal représentées ? (Il y a plusieurs réponses possibles).

```
> afcin$TOT
      inertia      cum      cum(%)
Ax1 6.109312e-02 0.06109312 50.54544
Ax2 2.587092e-02 0.08696404 71.94977
Ax3 1.755024e-02 0.10451429 86.46998
Ax4 7.950004e-03 0.11246429 93.04742
Ax5 3.586374e-03 0.11605066 96.01461
Ax6 2.809186e-03 0.11885985 98.33879
Ax7 1.385562e-03 0.12024541 99.48514
Ax8 5.910421e-04 0.12083645 99.97414
Ax9 3.125692e-05 0.12086771 100.00000
```

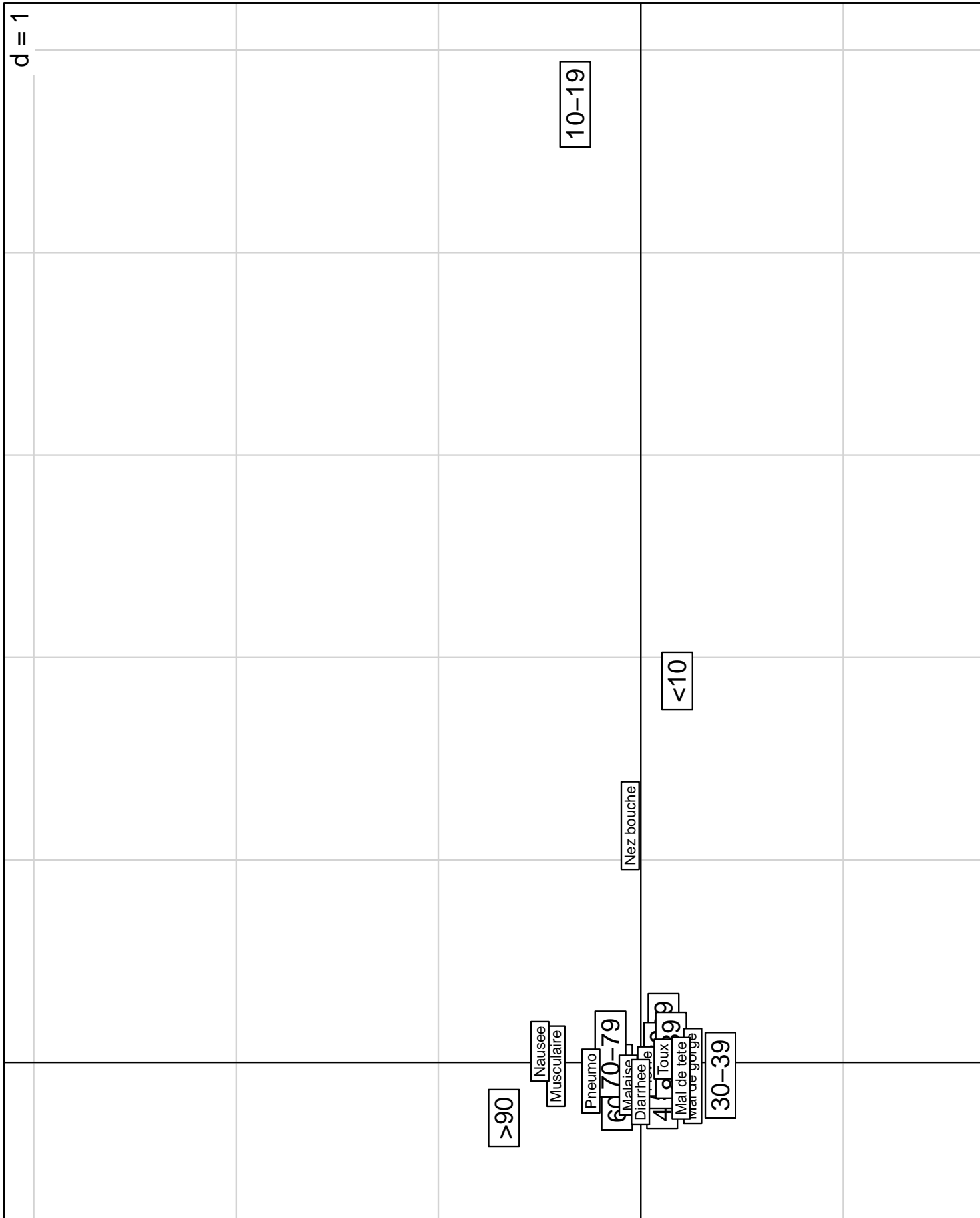


Contributions absolues et relatives des colonnes (multipliées par 10000) :

```
> round(afcin$row.abs*100)      > round(afcin$row.rel*100)
      Axis1(%) Axis2(%)          Axis1 Axis2
<10    3985     86      <10    9617  -87
10-19   5027     56      10-19  9608   45
20-29    128    233      20-29   761 -587
30-39     41   3612      30-39 -235 -8824
40-49    193    371      40-49 -1615 -1314
50-59     7     616      50-59  -54 -2068
60-69    481   1039      60-69 -4033  3687
70-79     84   2676      70-79   566  7630
80-89     20    825      80-89   165 -2932
>90      34    486      >90   -514  3103
```

Contributions absolues et relatives des lignes (multipliées par 10000):

```
> round(afcin$col.abs*100)      > round(afcin$col.rel*100)
      Axis1(%) Axis2(%)          Axis1 Axis2
Fievre      82     82      Fievre    -767  -325
Toux        11   1154      Toux      118 -5075
Pneumo     230   3896      Pneumo   -967  6942
Malaise    221   161      Malaise  -2164  665
Mal de gorge  59   2014      Mal de gorge -339 -4878
Nez bouche  9213     43      Nez bouche  9924   20
Diarrhee    141     0      Diarrhee  -2133   0
Mal de tete  35    524      Mal de tete  -722 -4580
Musculaire   1    936      Musculaire   -9  4519
Nausee       6   1189      Nausee     46  3897
```



Problème 4 [3 points]

Le tableau suivant montre la classe d'âge, le type préféré de vacances (ski, plage, visite de musée), et le budget pour un groupe de 20 personnes.

	Age	Vacances	Budget
Anne	Moins de 20 ans	Ski	Moins de 500 euros
Bertrand	Moins de 20 ans	Ski	Moins de 500 euros
Cédric	Moins de 20 ans	Ski	500 à 1000 euros
Djalil	Moins de 20 ans	Plage	Moins de 500 euros
Emilie	21 à 40 ans	Ski	Moins de 500 euros
Fatima	21 à 40 ans	Ski	500 à 1000 euros
Greg	21 à 40 ans	Plage	Moins de 500 euros
Hubert	21 à 40 ans	Plage	Moins de 500 euros
Inez	21 à 40 ans	Plage	500 à 1000 euros
Jennifer	21 à 40 ans	Musée	500 à 1000 euros
Karine	21 à 40 ans	Musée	Plus de 1000 euros
Luc	41 à 60 ans	Musée	Plus de 1000 euros
Martin	41 à 60 ans	Plage	Plus de 1000 euros
Nadia	41 à 60 ans	Plage	500 à 1000 euros
Oscar	41 à 60 ans	Musée	500 à 1000 euros
Pascal	41 à 60 ans	Musée	Plus de 1000 euros
Quentin	Plus de 60 ans	Plage	500 à 1000 euros
Rita	Plus de 60 ans	Musée	500 à 1000 euros
Stéphane	Plus de 60 ans	Musée	Plus de 1000 euros
Tatiana	Plus de 60 ans	Musée	Plus de 1000 euros

1. Déterminer le tableau de Burt associé à ce jeu de données.
2. Quelle est l'utilité du tableau de Burt pour l'analyse des correspondances multiples ?