TD8 : les passagers du Titanic (correction)

1 Introduction

Il s'agit de données (sans doute contestables) concernant les 2201 passagers et membres d'équipage du célèbre bateau « le Titanic », qui a coulé le 14 avril 1912. Il faut noter que tout le monde n'est pas d'accord sur le nombre de passagers et sur le nombre de victimes. Les variables sont :

`class`	classe	0=équipage, 1-3=classe
`age`	âge	0=enfant, 1=adulte
`sex`	sexe	0=féminin, 1=masculin
`surv`	survivant	0=non, 1=oui

On donne ci-dessous le tableau de Burt des données ainsi que le poids des catégories (en %).

	class.0	class.1	class.2	class.3	age.0	age.1	sex.0	sex.1	surv.0	surv.1
class.0	885	0	0	0	0	885	23	862	673	212
class.1	0	325	0	0	6	319	145	180	122	203
class.2	0	0	285	0	24	261	106	179	167	118
class.3	0	0	0	706	79	627	196	510	528	178
age.0	0	6	24	79	109	0	45	64	52	57
age.1	885	319	261	627	0	2092	425	1667	1438	654
sex.0	23	145	106	196	45	425	470	0	126	344
sex.1	862	180	179	510	64	1667	0	1731	1364	367
surv.0	673	122	167	528	52	1438	126	1364	1490	0
surv.1	212	203	118	178	57	654	344	367	0	711

	poids(%)
class.0	40.2
class.1	14.8
class.2	12.9
class.3	32.1
age.0	5.0
age.1	95.0
sex.0	21.4
sex.1	78.6
surv.0	67.7
surv.1	32.3

Question 1 *Quelle proportion d'enfants a survécu ? Quelle proportion de femmes a survécu ? Quelle est la proportion de femmes parmi les survivants ?*

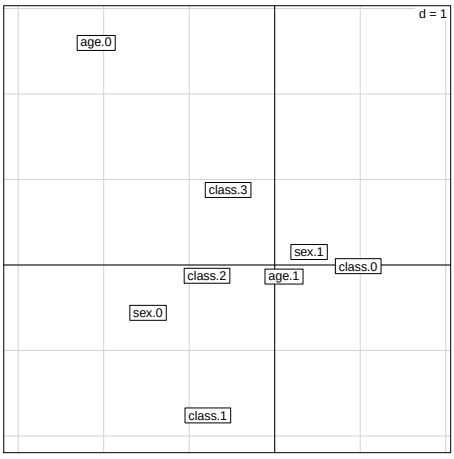
La proportion d'enfants qui ont survécu est $57/109 = 52\%$.

La proportion de femmes qui ont survécu est $344/470 = 73\%$.

La proportion de femmes parmi les survivants est $344/711 = 48\%$.

2 Analyse des correspondances multiples

On fait l'analyse en correspondance multiples des variables `class`, `age` et `sex`. La variable `surv` sera discutée plus loin. On donne ci-dessous les valeurs propres de l'ACM, puis les coordonnées des catégories sur les deux premiers axes (avec la représentation correspondante), ainsi que leur contribution en % à ces axes :



Valeurs propres sur les axes factoriels

	Comp1	Comp2	Comp3	Comp4	Comp5
mu	0.49	0.38	0.33	0.26	0.2

Coordonnées et contributions (en %)

	Comp1	Comp2
class.0	0.98	-0.01
class.1	-0.78	-1.76
class.2	-0.79	-0.13
class.3	-0.54	0.88
age.0	-2.09	2.60
age.1	0.11	-0.14
sex.0	-1.48	-0.56
sex.1	0.40	0.15

	Axis1	Axis2
class.0	26.0	0.0
class.1	6.1	40.0
class.2	5.5	0.2
class.3	6.5	21.6
age.0	14.7	29.2
age.1	0.8	1.5
sex.0	31.8	5.9
sex.1	8.6	1.6

Question 2 *Pourquoi y a-t-il 5 valeurs propres ? Quelle est leur somme ? Combien d'axes est-on conduit à conserver ?*

Le nombre de valeurs propres non nulles est égal à $(4+2+2)-3=5$.

La somme des valeurs propres est $5/3=1,67$.

On décide de retenir les valeurs propres qui sont supérieures à la moyenne, c'est-à-dire $1/3=0,33$. On se contente donc des deux premières valeurs propres.

Question 3 *Quelles sont les catégories qui déterminent les deux premiers axes ? Pourquoi pouvait-on prévoir l'importance de* `age.0` *dès la partie 1 ?*

Pour repérer les catégories intéressantes, on compare la contribution des catégories par rapport à leur poids, avec un facteur multiplicatif de 2. Les poids tels qu'ils ont été donnés sous le tableau de Burt ne sont pas utilisables tels quels : il faut les diviser par le nombre de variables, ici 3. En prenant en compte le facteur multiplicatif 2, on obtient le tableau suivant :

```
class.0 class.1 class.2 class.3   age.0   age.1   sex.0   sex.1
   26.8     9.8     8.6    21.4     3.3    63.4    14.2    52.4
```

En cherchant les catégories dont la contribution est supérieure à ces valeurs, on trouve : `age.0` $(14,7\%/3,3\%)$, et `sex.0` $(31,8\%/14,2\%)$ en négatif ; à la limite `class.0` $(26\%/26,8\%)$ en positif.

Les contributions absolues n'apportent rien de plus.

De même pour le second axe on obtient : `class.1` $(40\%/9,8\%)$ en négatif ; en positif `age.0` $(29,2\%/3,3\%)$ et `class.3` $(21,6\%/21,4\%)$.

On voit que `age.0` est important sur les axes. Ceci n'est pas surprenant car cette catégorie a un effectif assez faible par rapport aux autres (109). On sait que la contribution d'une catégorie à l'inertie totale dépend est décroissante avec l'effectif : elle vaut $\frac{1}{p}\left(1-\frac{n_j}{n}\right)=\frac{1}{3}\left(1-\frac{109}{2201}\right)$, soit quasiment la valeur maximale de 1/3.

3 Une variable supplémentaire

On cherche à savoir comment la variable `surv` est représentée sur les axes. On calcule à partir des données d'origine les coordonnées des deux catégories `surv.0` et `surv.1` sur les deux premiers axes principaux et les valeurs tests de ces catégories.

```
        Axis1 Axis2
surv.0   0.25  0.17
surv.1  -0.53 -0.36
```

```
        Axis1 Axis2
surv.0   17.07  11.8
surv.1 -17.07 -11.8
```

Question 4 *Comment les valeurs-test ont-elles été calculées ? Que nous indiquent-elles ?*

On utilise la formule

$$c_j\sqrt{n_j}\sqrt{\frac{n-1}{n-n_j}}$$

qui donne par exemple pour `surv.0` sur le premier axe

$$0,2514\times\sqrt{1490}\sqrt{\frac{2201-1}{2201-1490}}=17,07.$$

Les valeurs-test nous indiquent que la variable `surv` est liée aux deux axes principaux (surtout le premier), puisqu'elles vérifient les conditions suivantes

— calculées sur des variables non actives dans l'analyse
— avec des effectifs >30
— et des valeurs supérieures à 3 en valeur absolue

Les deux modalités sont bien sûr opposées (pour cause de centrage) et on peut dire que les survivants sont placés négativement sur les deux axes.

Question 5 *Comment peut-on interpréter les deux premiers axes principaux ? Expliquez en particulier en quoi la variable* `age.0` *pose un problème.*

Comme la variable supplémentaire est liée aux axes, nous pouvons interpréter les axes en l'utilisant.

— Le premier axe nous apprend qu'il y a beaucoup de survivants parmi les femmes et les enfants, et très peu dans l'équipage.
— Le second axe montre que les survivants sont du coté de la première classe plutôt que de celui des enfants et de la troisième classe.

Donc, alors que les enfants survivent correctement sur le premier axe, le deuxième axe les désavantage, ce qui est paradoxal.

4 Analyse par classe

On s'intéresse uniquement aux passagers de 3e classe. Le tableau de Burt de ce sous-ensemble des passagers est comme suit :

	age.0	age.1	sex.0	sex.1	surv.0	surv.1
age.0	79	0	31	48	52	27
age.1	0	627	165	462	476	151
sex.0	31	165	196	0	106	90
sex.1	48	462	0	510	422	88
surv.0	52	476	106	422	528	0
surv.1	27	151	90	88	0	178

Question 6 *Expliquez ce que ces données apportent de plus que le tableau de Burt de départ. Quel est le genre de liaison entre les variables qui est inaccessible au tableau de Burt, et donc à l'ACM ?*

Le tableau de Burt ne concerne que des couples de variables, comme par exemple `age` et `surv`. Ici la solution est plus complexe puisqu'on lie 3 variables : classe plus deux autres. L'ACM (comme l'ACP d'ailleurs) est incapable de rendre compte des liens entre ces variables.

Question 7 *Quelle proportion d'enfants voyageant en 3e classe a survécu ? Quelle proportion d'enfants voyageant en 1re ou 2e classe a survécu ? En déduire une explication du problème relevé à la question 5.*

La proportion d'enfants de 3e classe ayant survécu est $27/79 = 34\%$.

En se reportant au premier tableau de Burt, le nombre d'enfants qui étaient en 1re ou 2e classe est $6+24 = 30$. Le nombre d'enfants de 1re ou 2e classe ayant survécu est $57 - 27 = 30$. Tous les enfants de ces classes ont survécu.

Cette relation complexe entre âge, classe et survie est l'explication du problème : les enfants qui apparaissent sur l'axe 1 sont ceux des classes « de luxe » et ceux qui ont moins de chance sur l'axe 2 sont ceux de la 3e classe.

5 Contribution des individus à l'inertie en ACM

On considère l'ACM de p variables qualitatives mesurées sur n individus. On a calculé dans le cours la contribution des catégories et des variables à l'inertie totale. On cherche ici à calculer la contribution des individus à cette même inertie. Dans le cas de l'ACM, l'inertie totale s'écrit sur les profils lignes

$$I_{\mathbf{g}} = \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{e}_i - \mathbf{g}_\ell\|^2_{\chi_\ell}, \text{ avec } \|\mathbf{e}_i - \mathbf{g}_\ell\|^2_{\chi_\ell} = \sum_{\text{toutes les catég. } j} \frac{np}{n_j}\left(\frac{x_i^j}{p} - \frac{n_j}{np}\right)^2,$$

où x_i^j vaut 1 si l'individu i appartient à la catégorie j et 0 sinon, et n_j est le nombre total d'individus de catégorie j.

Question 8 *Montrer que*

$$\left(\frac{x_i^j}{p} - \frac{n_j}{np}\right)^2 = \frac{x_i^j}{p^2} + \frac{n_j^2}{n^2p^2} - 2\frac{x_i^j n_j}{np^2}.$$

On sait, d'après la formule classique $(a+b)^2 = a^2 + b^2 + 2ab$ que

$$\left(\frac{x_i^j}{p} - \frac{n_j}{np}\right)^2 = \left(\frac{x_i^j}{p^2}\right)^2 + \left(\frac{n_j}{np}\right)^2 - 2\frac{x_i^j n_j}{np^2}.$$

Or, comme $x_i^j = 0$ ou 1, on a $\left(x_i^j\right)^2 = x_i^j$. La formule ci-dessus se simplifie donc comme demandé dans la question.

Question 9 *En déduire que la contribution de l'individu i à l'inertie totale est*

$$\left(\frac{1}{np}\sum_{j \text{ catég. de } i} \frac{n}{n_j}\right) - \frac{1}{n},$$

où la somme est faite sur les catégories auxquelles appartient i.

La contribution de l'individu i à l'inertie totale est

$$\frac{1}{n}\sum_{\text{toutes les catég. } j}\frac{np}{n_j}\left(\frac{x_i^j}{p}-\frac{n_j}{np}\right)^2=\frac{1}{n}\sum_{\text{toutes les catég. } j}\frac{np}{n_j}\left(\frac{x_i^j}{p^2}+\frac{n_j^2}{n^2p^2}-2\frac{x_i^j n_j}{np^2}\right)$$
$$=\sum_{\text{toutes les catég. } j}\left(\frac{x_i^j}{pn_j}+\frac{n_j}{n^2p}-2\frac{x_i^j}{np}\right).$$

La somme se calcule somme suit :
— premier terme : le x_j^i revient à ne garder que les catégories auxquelles i appartient ; il reste donc $\frac{1}{np}\sum_j$ catég. de $_i\,\frac{n}{n_j}$.
— deuxième terme : $\sum_{\text{toutes les catég. } j} n_j = np$ et le terme restant est donc $1/n$.
— troisième terme : $\sum_{\text{toutes les catég. } j} x_i^j = p$ (chaque individu appartient à p catégories) ; il reste donc $-2/n$.

En additionnant les termes, on retrouve la valeur souhaitée.

Question 10 *Expliquez pourquoi cette contribution est toujours positive. Comment peut-on caractériser les individus dont la contribution à l'inertie totale est grande ?*

La raison la plus simple pour laquelle la contribution d'un individu est positive est qu'elle vaut $\frac{1}{n}\|\mathbf{e}_i-\mathbf{g}_\ell\|^2_{\chi_\ell}>0$. On peut aussi voir sur la formule finale qu'on a toujours $n/n_i\geq 1$ et donc que $\frac{1}{p}\sum_j$ catég. de $_i\,\frac{n}{n_j}$, qui est la moyenne de p valeurs supérieures à 1, est aussi supérieure à 1. Ceci implique la positivité de la contribution.

Les individus dont la contribution à l'énergie totale est grande sont ceux qui appartiennent le plus à des catégories a petit effectif. En effet la contribution croît quand l'effectif des catégories diminue.