TD9 : les passagers du Titanic

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 %).

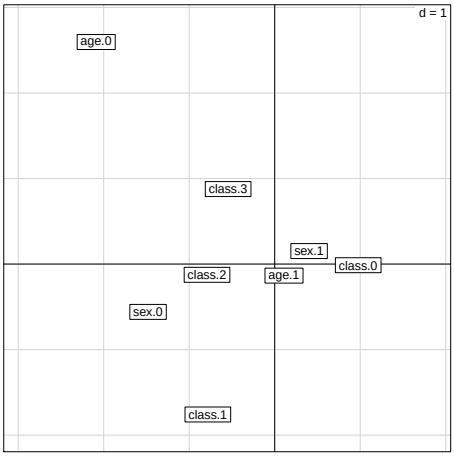
	class.0	class.1	class.2	class.3	age.0	age.1	sex.0	sex.1	surv.0	surv.1
class.0	885	0	0	0	0	885	23	862	673	212
class.1	0	325	0	0	6	319	145	180	122	203
class.2	0	0	285	0	24	261	106	179	167	118
class.3	0	0	0	706	79	627	196	510	528	178
age.0	0	6	24	79	109	0	45	64	52	57
age.1	885	319	261	627	0	2092	425	1667	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 ?*

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 ?*

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 ?*

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 *Que nous indiquent les valeurs-test ?*

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

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 ?*

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.*

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*

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

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.

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