Titanic project

NIANG

2022-12-07

```
## [1] "C:/Users/abdou/Documents/cours analyse de données"
```

1. Tableau Titanic:

```
##   Classe Age Sexe Survie
## 1      1   1    1      1
## 2      1   1    1      1
## 3      1   1    1      1
## 4      1   1    1      1
## 5      1   1    1      1
## 6      1   1    1      1
```

Nous disposons d'un nombre n de personnes présentes sur le Titanic lors de son naufrage en pleine mer, de leur age (adulte ou enfant), du sexe (homme ou femme), de la classe (première, deuxième, troisième ou equipage) et de leur statut (survivant ou décédé).

2. Nous allons renommer les modalités :

Classe	Age	Sexe	Survie
1st class	Adulte	Homme	Survivant
1st class	Adulte	Homme	Survivant
1st class	Adulte	Homme	Survivant
1st class	Adulte	Homme	Survivant
1st class	Adulte	Homme	Survivant
1st class	Adulte	Homme	Survivant

3. Tableau des effectifs croisés *Classe* et *Survie* *étape 1 : Tableau croisé*

```
##
##             Décédé Survivant
##   1st class    122       203
##   2nd class    167       118
##   3rd class    528       178
##   Crew         673       212
```

étape 2 : Somme par colonne

```
##    Décédé Survivant
##      1490       711
```

étape 3 : Ajout de la ligne Total colonne

```
##           Décédé Survivant
## 1st class    122       203
## 2nd class    167       118
## 3rd class    528       178
## Crew         673       212
## Total_c     1490       711
```

étape 4 : Somme par lignes

```
## 1st class 2nd class 3rd class      Crew   Total_c
##       325       285       706       885      2201
```

étape 5 : Ajout de la colonne Total ligne qui nous donne au final le tableau croisé *Classe Survie* suivant :

```
##           Décédé Survivant Total_r
## 1st class    122       203     325
## 2nd class    167       118     285
## 3rd class    528       178     706
## Crew         673       212     885
## Total_c     1490       711    2201
```

4. Tableau des effectifs attendus sous hypothèse H0 d'indépendance entre *Classe* et *Survie*

On sait que sous hypothèse d'indépendance, $f_{ij} = f_{i.}f_{.j}$ ce qui équivaut encore à $n_{ij} = n_{i.}n_{.j}/N$ avec N effectif total : $N = \sum_{i=1}^{p}\sum_{i=1}^{q} n_{ij}$ et n_{ij} le nombre de personnes avec la modalité i de la variable X et la modalité j de Y.

Remarque : Qu'on soit dans le cadre du tableau des effectifs observés ou le tableau des effectifs attendus sous hypothèse d'indépendance, les marges-lignes et marges-colonnes ne changent pas.

En effet, on sait que $n_{.j} = \sum_{i=1}^{p} n_{ij}$

or sous hypothèse H0 $n_{ij} = n_{i.}n_{.j}/N$ donc $n_{.j} = \sum_{i=1}^{p} n_{i.}n_{.j}/N = (n_{.j}/N)\sum_{i=1}^{p} n_{i.} = n_{.j}$.

On peut désormais construire notre tableau des effectifs observés sous hypothèse H0:

```
for(i in 1:4)
{
for(j in 1:2)
{

  Cross_Titanic[i,j] = round((Total_r [i]*Total_c[j])/N,2)
}
}
```

Nous obtenons le tableau suivant alors :

```
##            Décédé Survivant Total_r
## 1st class  220.01    104.99     325
## 2nd class  192.94     92.06     285
## 3rd class  477.94    228.06     706
## Crew       599.11    285.89     885
## Total_c   1490.00    711.00    2201
```

5. Donner le tableau des effectifs croisés *Survie* et *Age* grace a table *étape 1 : Tableau croisé*

```
##
##           Décédé Survivant
##   Adulte    1438       654
##   Enfants     52        57
```

étape 2 : Somme par colonne

```
##    Décédé Survivant
##      1490       711
```

étape 3 : Ajout de la ligne Total colonne

```
##          Décédé Survivant
## Adulte     1438       654
## Enfants      52        57
## Total_c1   1490       711
```

étape 4 : Somme par lignes

```
##   Adulte  Enfants Total_c1
##     2092      109     2201
```

étape 5 : Ajout de la colonne Total ligne qui nous donne au final le tableau croisé *Age Survie* suivant :

```
##          Décédé Survivant Total_r1
## Adulte     1438       654     2092
## Enfants      52        57      109
## Total_c1   1490       711     2201
```

```
##          Décédé Survivant Total_r1
## Adulte     1438       654     2092
## Enfants      52        57      109
## Total_c1   1490       711     2201
```

4. Tableau des effectifs attendus sous hypothèse H0 d'indépendance entre *Age* et *Survie* Par analogie à la question 4, nous avons:

```
for(i in 1:2)
{
for(j in 1:2)
{

  Cross_Titanic_2[i,j] = round((Total_r1 [i]*Total_c1[j])/N,2)
}
}
```

Nous obtenons le tableau suivant sous hypothèse d'indépendance

```
##           Décédé Survivant Total_r1
## Adulte   1416.21    675.79     2092
## Enfants    73.79     35.21      109
## Total_c1 1490.00    711.00     2201
```

Peut-on affirmer, au risque de 5%, que les variables *Classe* et *Survie* sont dépendantes ? : Pour celà, faisons un test du Khi2, du tableau correspondant que nous avons appelé ici Titanic_CS :

Si le ***p-value*** < 0.05 alors nous rejeterons l'hypoyhèse H0 d'indépendance

```
##
##  Pearson's Chi-squared test
##
## data:  Titanic_CS
## X-squared = 190.4, df = 3, p-value < 2.2e-16
```

Le ***p-value*** est pratiquement nul. Donc nous rejetons l'hypothèse d'indépendance. Donc les variables *Classe* et *Survie* sont dépendantes.

Remarque : Cependant, ce test ne nous permet pas d'analyser les relations spécifiques entre deux variables. Il résume simplement si oui ou non il y'a une association.

8. Tester l'indépendance des variables *Sexe Survie* puis *Age Survie* Conclure.

- Pour *Sexe Survie*, créons d'abord la table :

```
##
##         Décédé Survivant
##   Femme    126       344
##   Homme   1364       367
```

```
## X-squared
##  454.4998
```

Le nombre de degrés de liberté est $df = (p-1)(q-1) = 1 = \nu$

Si on se fixe un niveau de $5\% = 0.05$ alors $1 - \alpha = 0.950$ et donc nous cherchons χ_1 tel que $\chi_1 > \chi_{\nu,1-\alpha}$ avec

$\chi_{\nu,1-\alpha} = \chi_{1,0.950} = 0.00$ par lecture sur le tableau du Khi2

Or $\chi_1 = 454.4998$. Donc on bien $\chi_1 > \chi_{\nu,1-\alpha}$.

Conclusion : Les variables *Sexe* et *Survie* sont fortement liées. Nous verrons dans l'AFC en quoi elles le sont.

- Pour *Age Survie* : *Tableau*

```
##
##           Décédé Survivant
##   Adulte    1438       654
##   Enfants     52        57
```

```
## X-squared
##   20.0048
```

Pour les memes raisons que plus haut, les variables *Age Survie* sont fortement liées.

9. Réaliser l'AFC des variables classe et survie. Combien d'axes factoriels peut-on choisir ? Interpreter

D'abord puisque nous avons fait un test du Khi2 (pour evaluer le lien entre les variables), nous allons utiliser les attributs de la fonction khi2 pour analyser de plus près ce lien en particulier le *Résidu* qui est la différence entre les effectifs observés et attendus.

```
##
##                Décédé Survivant
##   1st class -6.607873  9.565772
##   2nd class -1.867159  2.702959
##   3rd class  2.289965 -3.315027
##   Crew       3.018611 -4.369840
```

Nous pouvons dès à présent affirmer qu'il existe une forte attractivité entre *1ère classe* et *Survivant* et une forte répulsion entre *1ère classe* et *Décédé*

Alors que dans le meme temps il une attractivité entre *les membres de l'équipage* et *Décédé* et une répulsion entre *Equipage* et *Survivant*.

```
## **Results of the Correspondence Analysis (CA)**
## The row variable has  4  categories; the column variable has 2 categories
## The chi square of independence between the two variables is equal to 190.4011 (p-value =  4.999928e-
## *The results are available in the following objects:
##
##    name              description
## 1  "$eig"            "eigenvalues"
## 2  "$col"            "results for the columns"
## 3  "$col$coord"      "coord. for the columns"
## 4  "$col$cos2"       "cos2 for the columns"
## 5  "$col$contrib"    "contributions of the columns"
## 6  "$row"            "results for the rows"
## 7  "$row$coord"      "coord. for the rows"
## 8  "$row$cos2"       "cos2 for the rows"
## 9  "$row$contrib"    "contributions of the rows"
## 10 "$call"           "summary called parameters"
## 11 "$call$marge.col" "weights of the columns"
## 12 "$call$marge.row" "weights of the rows"
```

```
##       eigenvalue variance.percent cumulative.variance.percent
## Dim.1 0.08650663              100                          100
```

On peut voir que le tableau *Classe* et *Survie* n'a qu'une seule valeur propre qui a elle seule explique 100% de l'information.

```
## 1st class 2nd class 3rd class      Crew
## 70.991171  5.668177  8.525867 14.814784
```

On peut remarquer que la contribution de la *1ère classe* à la création de l'axe est de 70% environ, c'est la plus forte. Celle de *Equipage* est la deuxième plus forte.

Donc l'axe oppose les *1ère classe* et *Equipage* du point de vue de la survie.

```
##              [,1]
## Décédé    32.3035
## Survivant 67.6965
```

De l'autre coté les *Survivants* participent beaucoup à la création de l'axe.

Donc l'axe oppose *Décédé* et *Survivant* du point de vue de la classe dans le bateau.

Conclusion : L'axe oppose les membres de la *1ère classe* qui pour beaucoup ont survécus aux membres de *Equipage* qui sont décédés pour beaucoup.

```
##           [,1]
## Décédé       1
## Survivant    1
```

```
## 1st class 2nd class 3rd class      Crew
##         1         1         1         1
```

Ici tous les points sont bien représentés.

La représentation se faisant sur un seul axe, nous arreterons notre analyse à la qualité de représentation et à la contribution à l'unique axe factoriel.