TD 2 – Analyse Factorielle des Correspondances et Analyse des Correspondances Multiples

1 Exercice 1. AFC et SVD généralisée

Il s'agit d'un tableau de contingence croisant 5 catégories de salariés (en lignes) et 4 catégories de fumeurs (en colonnes) au sein d'une entreprise :

- Lignes (Salariés) : `SM` : Senior Managers; `JM` : Junior Managers; `SE` : Senior Employees; `JE` : Junior Employees; `SC` : Secretaries.

```
> library(ca)
> data(smoke)
> print_table(smoke)
```

	none	light	medium	heavy
SM	4	2	3	2
JM	4	3	7	4
SE	25	10	12	4
JE	18	24	33	13
SC	10	6	7	2

```
> sum(smoke)
```

```
[1] 193
```

0. On effectue un test préalable d'indépendance du χ^2

```
> chisq.test(smoke, simulate.p.value=TRUE)
```

```
    Pearson's Chi-squared test with simulated p-value (based on 2000
    replicates)

data:  smoke
X-squared = 16.442, df = NA, p-value = 0.1709
```

Que conclut-on quant à la significativité statistique de l'association ? Quel garde-fou méthodologique faut-il garder en tête lors des questions suivantes ?

1. Préparation des données

   a. Donnez le code R ou Python pour construire la matrice F des fréquences, les vecteurs r et c des distributions marginales et la matrice Z des écarts à l'indépendance.

   b. Complétez les entrées manquantes `??` dans les sorties ci-dessous :

Le vecteur r des masses des lignes

```
[1] 0.06  ??  0.26 0.46 0.13
```

Le vecteur c des masses des colonnes

```
[1] 0.32 0.23  ??  0.13
```

La matrice Z des écarts à l'indépendance

	none	light	medium	heavy
SM	0.15	-0.22	-0.15	0.4
JM	-0.3	-0.29	0.21	0.72
SE	??	-0.16	-0.27	-0.39
JE	-0.35	0.17	0.17	0.14
SC	0.27	0.03	-0.13	-0.38

---

2. Représentation des données sur les axes factoriels
   a. La fonction `gsvd` ci-dessous réalise la décomposition en valeurs singulières généralisée d'une matrice réelle Z de dimension $n \times p$ avec les métriques diagonales $N = \mathrm{Diag}(r)$ sur $\mathbb{R}^n$ et $M = \mathrm{Diag}(c)$ sur $\mathbb{R}^p$.

▶ Code

   b. Donnez le code qui calcule, avec la fonction `gsvd`, le vecteur Λ des valeurs singulières ainsi que les matrices X et Y des coordonnées factorielles.

```
> print_table(Lambda)
```

data
0.273
0.100
0.020

et les matrices X et Y des coordonnées factorielles des profils-lignes et profils-colonnes de l'AFC

```
> print_table(X)
```

	dim1	dim2	dim3
SM	-0.066	-0.194	0.071
JM	0.259	-0.243	-0.034
SE	-0.381	-0.011	-0.005
JE	0.233	0.058	0.003
SC	-0.201	0.079	-0.008

```
> print_table(Y)
```

	dim1	dim2	dim3
none	-0.393	-0.030	-0.001
light	0.099	0.141	0.022
medium	0.196	0.007	-0.026
heavy	0.294	-0.198	0.026

   c. Calculez les inerties associées à chaque axe ainsi que le pourcentage d'inertie expliquée par le premier plan factoriel.
   d. Représentez à la main (ou sous R) les profils-lignes et les profils-colonnes sur le premier plan factoriel de l'AFC.
3. Exploitation des sorties `FactoMineR` suivantes :

```
> library(FactoMineR)
>
> res <- CA(smoke, graph = FALSE)
>
> stopifnot(isTRUE(all.equal(abs(unname(res$row$coord)), abs(unname(X)), tolerance = 1e-
      8)))
> stopifnot(isTRUE(all.equal(abs(unname(res$col$coord)), abs(unname(Y)), tolerance = 1e-
      8)))
```

a. Expliquez le code R ci-dessus.

b. On donne :

```
> print_table(res$row$contrib)
```

	Dim 1	Dim 2	Dim 3
SM	0.330	21.356	69.433
JM	8.366	55.115	25.619
SE	51.201	0.300	1.698
JE	33.097	15.177	1.205
SC	7.006	8.052	2.045

```
> print_table(res$col$contrib)
```

	Dim 1	Dim 2	Dim 3
none	65.400	2.934	0.061
light	3.085	46.317	27.282
medium	16.562	0.174	51.140
heavy	14.954	50.575	21.517

Pourquoi ne prend-on pas la peine d'afficher `res$row$cos2` et `res$col$cos2` ?

c. Interprétez les axes factoriels et le graphique réalisé à la question 2.d.

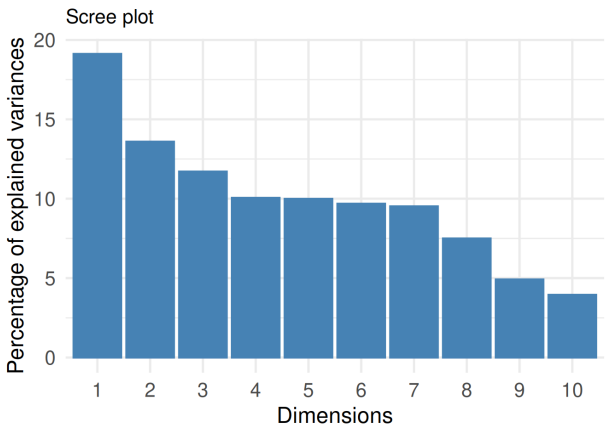
2 Exercice 2.

On considère un jeu de données référençant $1\,044$ passagers du *Titanic*, décrivant les variables suivantes :

- `Survived` : Survie du passager ($0 = \text{Non}$, $1 = \text{Oui}$)
- `Pclass` : Classe du billet ($1^{\text{ère}}$, $2^{\text{ème}}$ ou $3^{\text{ème}}$)
- `Sex` : Sexe du passager
- `Age` : Tranche d'âge ($[0\,;10]$, $[11\,;18]$, $[19\,;29]$, $[30\,;49]$, ≥ 50)
- `Fare` : Prix du ticket (variable quantitative supplémentaire)
- `Embarked` : Port d'embarquement ($\text{C} = \text{Cherbourg}$, $\text{Q} = \text{Queenstown}$, $\text{S} = \text{Southampton}$)

1. On affiche d'abord les résultats suivants.

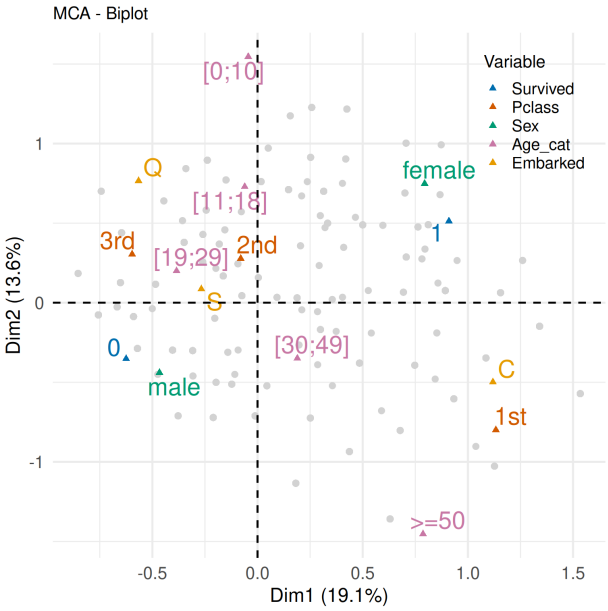
	eigenvalue	percentage of variance	cumulative percentage of variance
dim 1	0.382	19.116	19.116
dim 2	0.272	13.589	32.705
dim 3	0.234	11.707	44.412
dim 4	0.201	10.051	54.463
dim 5	0.200	9.989	64.452
dim 6	0.194	9.682	74.134
dim 7	0.190	9.518	83.651
dim 8	0.150	7.493	91.144
dim 9	0.098	4.913	96.057
dim 10	0.079	3.943	100.000

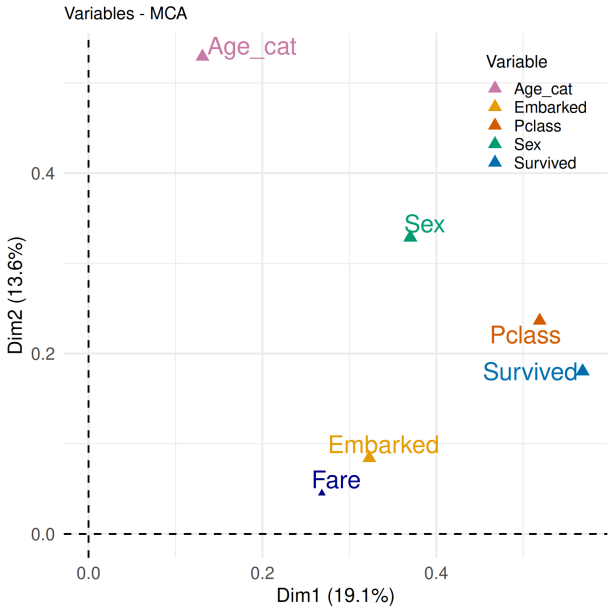


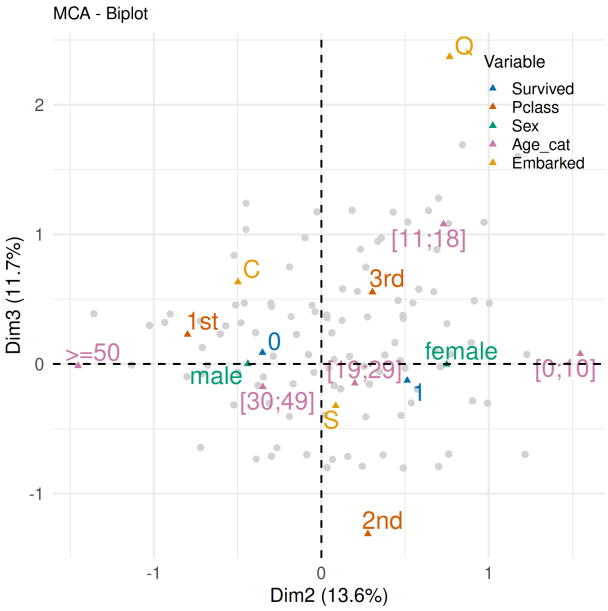
a. Justifiez le fait qu'il y ait 10 lignes dans ce tableau.

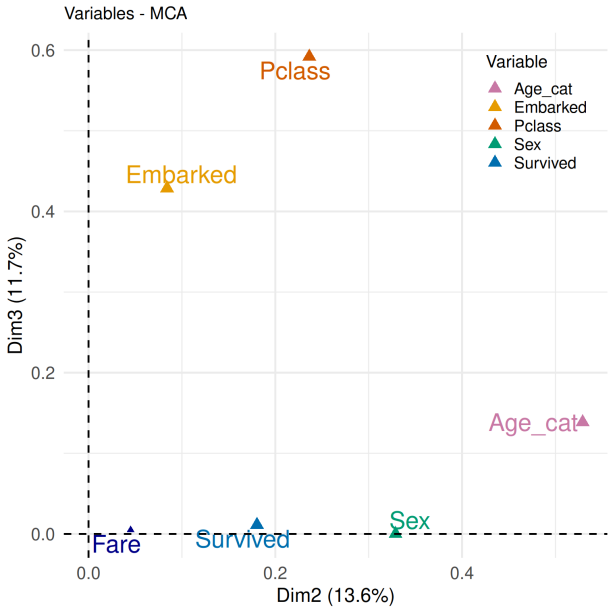
b. À partir de ces affichages, combien de directions principales aurait-on envie de conserver ?

2. On trace les représentations ci-dessous.









a. Commentez la représentation des individus dans ces plans.
b. Rappelez comment sont calculées les positions des modalités sur ces graphiques.
c. Commentez la représentation des variables dans ces deux plans.
d. Interprétez les positions des différentes modalités par rapport à celles de la variable `Survived`.

3. On affiche à présent les sorties suivantes.

```
> print_table(
+     titanic.mca$var$contrib[,1:3]
+ )
```

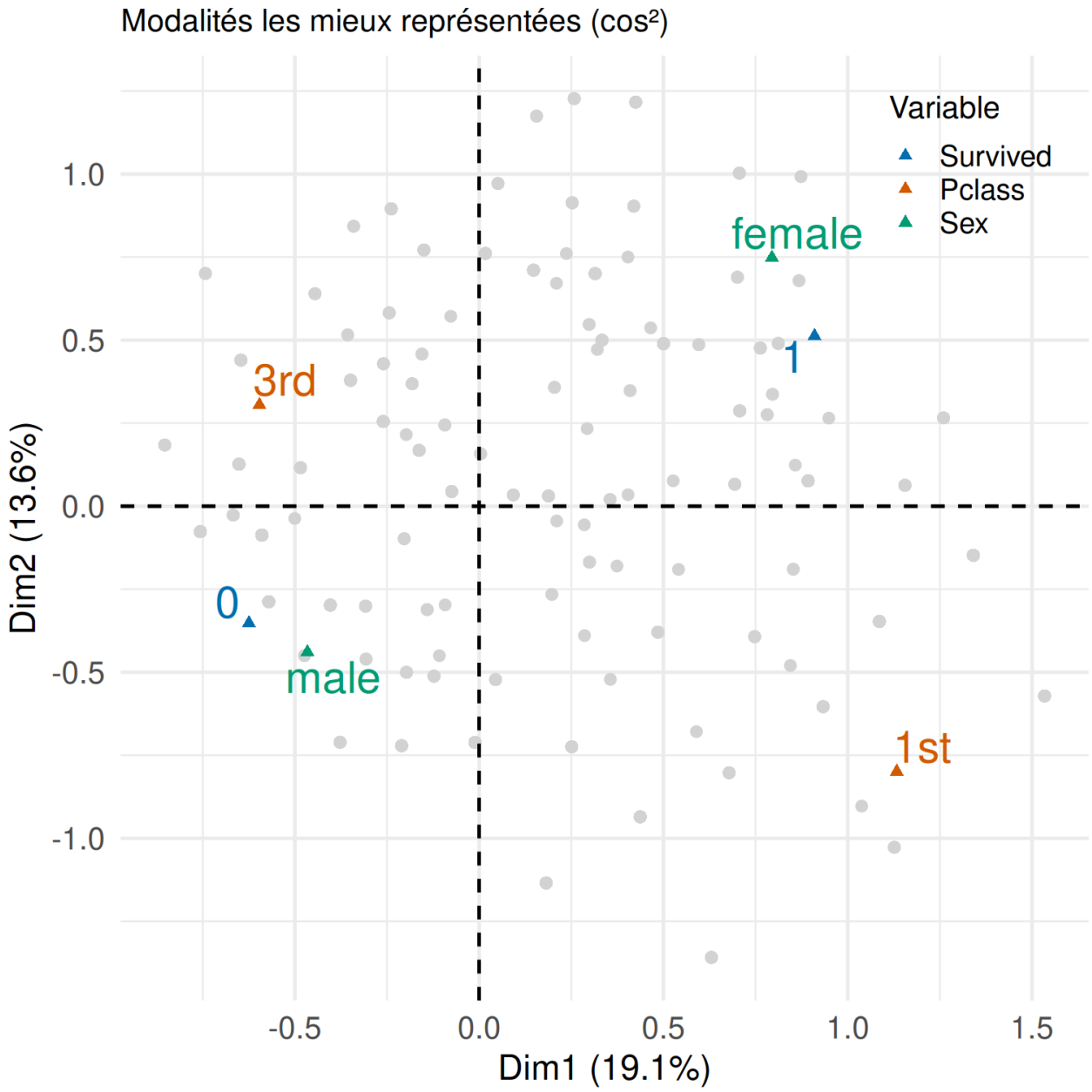
	Dim 1	Dim 2	Dim 3
0	12.104	5.404	0.393
1	17.629	7.871	0.572
1st	18.141	12.693	1.193
2nd	0.083	1.421	36.734
3rd	8.924	3.285	12.638
female	12.205	15.250	0.000
male	7.159	8.946	0.000
[0;10]	0.008	14.495	0.042
[11;18]	0.020	4.011	10.186
[19;29]	2.791	1.062	0.683
[30;49]	0.657	3.151	0.929
>=50	3.381	16.221	0.002
C	13.300	3.712	6.943
Q	0.802	2.068	23.003
S	2.797	0.409	6.681

```
> print_table(
+     titanic.mca$var$cos2[,1:3]
+ )
```

	Dim 1	Dim 2	Dim 3
0	0.568	0.180	0.011
1	0.568	0.180	0.011
1st	0.475	0.236	0.019
2nd	0.002	0.026	0.573
3rd	0.328	0.086	0.284
female	0.370	0.329	0.000
male	0.370	0.329	0.000
[0;10]	0.000	0.215	0.001
[11;18]	0.000	0.061	0.133
[19;29]	0.083	0.023	0.012
[30;49]	0.019	0.066	0.017
>=50	0.072	0.246	0.000
C	0.319	0.063	0.102
Q	0.016	0.030	0.283
S	0.213	0.022	0.312

Quelle tendance semble se dégager ?

4. Synthétisez votre analyse en quelques lignes.


Modalités les mieux représentées (cos²)
Variable
Survived
Pclass
Sex
female
1
3rd
0
male
1st
1.0
0.5
0.0
-0.5
-1.0
Dim2 (13.6%)
-0.5
0.0
0.5
1.0
1.5
Dim1 (19.1%)