Compte-rendu du TP n°1: Statistique descriptive, Analyse en composantes principales

Sarah Richard et Solène Weill

31 mars 2016

1 Introduction

L'objectif de ce premier TP est, d'une part, l'application de la statistique descriptive et, d'autre part, la réalisation de deux ACP en vue de leur exploitation. Il nous permet également de nous familiariser avec l'environnement R et les outils mis à notre disposition pour l'extraction et l'exploitation de données. Dans la première partie, nous étudierons deux jeux de données. Dans la seconde partie, nous nous intéresserons à l'ACP d'un point de vue théorique puis pratique sur un des jeux de données de la première partie.

2 Statistique descriptive

2.1 Le racket du tennis

2.1.1 Question 1

Afin de pouvoir exploiter les données de façon optimale, nous effectuons un prétraitement sur celles-ci pour conserver uniquement les paris dont les matches se sont terminés.

Pour faire une analyse descriptive générale des données, nous avons mis en place différents tableaux : un premier contenant chaque match de façon unique avec les informations qui lui sont associées (perdant, gagnant, année, ID) et un second contenant chaque joueur et le nombre de matches qu'il a joué.

À partir du premier tableau, il est possible de calculer le nombre de matches joués (gagnés et perdus). Pour cela, on compte le nombre d'identifiants de matches différents existant dans le jeu de données. Sur le temps de l'étude, 26532 matches ont été programmés mais seulement 25993 matches ont été joués jusqu'à la victoire d'un des deux joueurs.

Grâce au second tableau, on calcule le nombre de joueurs différents : 1527 noms de joueurs de tennis sont présents dans les données. Mais seulement 1523 d'entre eux ont joué au moins un match qui s'est terminé. Cela signifie que 4 joueurs ont joué des matches qui n'ont jamais aboutis c'est-à-dire que les matches ont été annulés ou reportés.

Les matches regroupés dans le jeu de données se sont déroulés sur une période de 6 ans, entre 2009 et 2015.

2.1.2 Question 2

Pour cette question, nous avons créé un tableau contenant pour chaque joueur son nombre de matches gagnés, perdus et joués. Nous avons ensuite catégorisé les joueurs afin de déterminer leurs niveaux. Pour cela, nous avons calculé le pourcentage de matches gagnés pour chaque joueur

et lui avons attribué un niveau en fonction de ce pourcentage [1]. Afin de visualiser au mieux ces données, nous avons tracé le graphique du nombre de matches joués en fonction du niveau du joueur. On remarque alors que plus un joueur joue de matches, plus sa catégorie est élevée. Afin

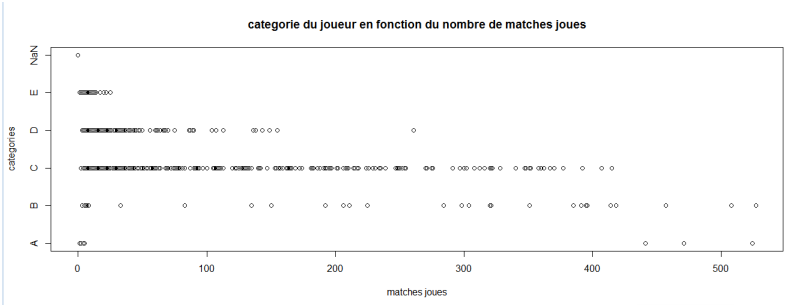

FIGURE 1 – Catégorie du joueur en fonction du nombre de matches joués

que les résultats soient plus représentatifs, il aurait fallu prendre en compte le nombre de matches joués par chaque joueurs. Il est alors pertinent de représenter la propension des joueurs à gagner en fonction du nombre de matches joués. On observe que plus les joueurs ont joué de matches, plus

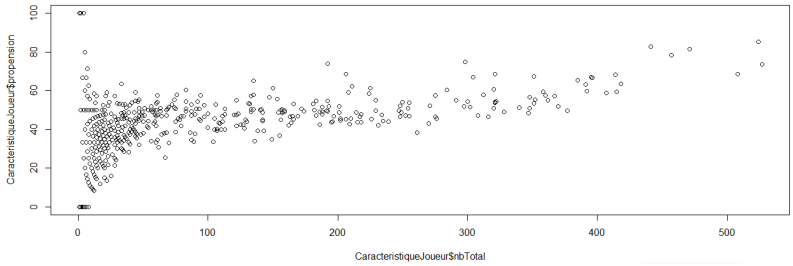

FIGURE 2 – Propension du joueur à gagner ses matches en fonction du nombre total de matches qu'il a joués

leur propension est élevée. Cela concorde avec la réalité car plus un joueur est bien classé plus il jouera de matches.

2.1.3 Question 3

a

Un match est considéré suspect lorsqu'au moins un des paris présente une évolution de probabilité supérieure à 0.1 en valeur absolue. Pour calculer cette évolution, nous avons soustrait la probabilité de gain du match à la fin avec celle de début pour chacun des perdants. Puis nous avons gardé toutes les valeurs dont la valeur absolue est supérieure à 0.1. Nous avons ainsi trouvé 2798 matches suspects.

1. Fonction disponible en annexe.

b

Tous les bookmakers sont impliqués dans les matches suspects mais dans des proportions différentes.

```
> table(tabParis$book)

   A    B    C    D    E    F    G
1007 1363 1117  243  229  218  121
```

FIGURE 3 – Bookmakers impliqués dans les matches suspects

Certains bookmarkers ont effectué plus de paris que d'autres. Il est alors pertinent de représenter par un pourcentage le nombre de paris suspects pris par chaque bookmaker en fonction de leur nombre de pari total. On se rend alors compte que les bookmakers A, B et C sont les plus susceptibles d'être impliqués dans les matchs suspects avec respectivement 4,44%, 6,11% et 5,20% de paris suspects. Les autres bookmakers, D, E, F et G n'ont respectivement que 1,46%, 1,37%, 1,54% et 0,97% de matches suspects.

c

Il est possible qu'un match dépasse une évolution de probabilité supérieure à 0.1 sans pour autant être une malversation. Nous avons donc décidé qu'un joueur associé à une malversation serait un joueur pour lequel un des paris sur son match présente une évolution de probabilité strictement supérieure à 0.15 en valeur absolue. Nous avons ainsi répertorié 540 matches suspects et 286 joueurs suspects sur 1527. Nous avons trouvé 6 joueurs dont le nombre de défaites est supérieur à 10. Ces 6 joueurs sont donc les plus susceptibles d'être associés à des malversations.

```
> DefaiteSup
    players_uid nbdefaite
5    0ffe23c8b8        13
81   63d5267c12        13
125  c9d4889bac        11
141  aa41a60f61        16
154  614c204988        11
270  d967764912        14
```

FIGURE 4 – Joueurs ayant plus de 10 défaites

2.2 Les données crabs

Nous étudions ici une population de crabes, mâles et femelles, de deux espèces différentes selon 5 variables quantitatives représentant la longueur de certaines parties du corps.

2.2.1 Question 1

Afin d'effectuer l'analyse descriptive des données crabs, nous traçons plusieurs histogrammes permettant de voir la répartition du nombre de crabe sur chaque variable quantitative en fonction du sexe et en fonction de l'espèce.

Mais les résultat ne nous permettent pas de tirer de conclusion ni de dégager une réelle tendance.

2.2.2 Question 2

La corrélation entre CW et CL est particulièrement élevée sur les deux graphiques de corrélation. Cela s'explique car les mesures concernent la même partie du corps, ainsi plus l'une est grande, plus l'autre sera grande aussi. Les deux autres valeurs très corrélées sont CW et BD.

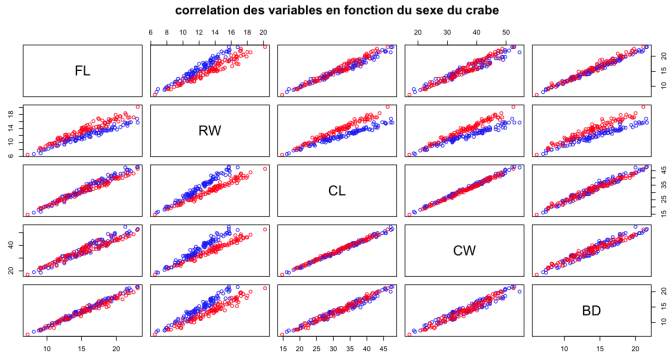

FIGURE 5 – Corrélation des différentes variables en fonction du sexe

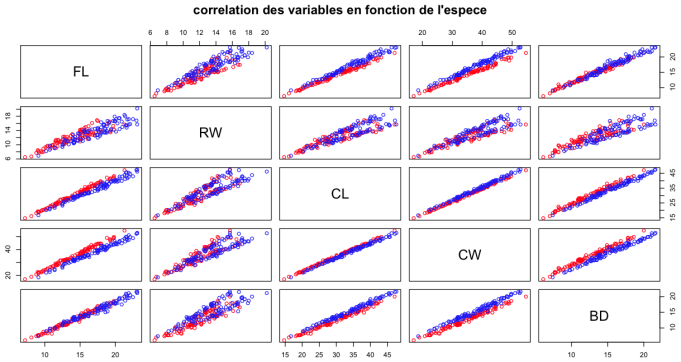

FIGURE 6 – Corrélation des différentes variables en fonction de l'espèce

Il y a donc proportionnalité des différentes parties du corps d'un crabe quelque soit son espèce et son sexe. Plus une de ses parties du corps sera grande plus les autres le seront aussi et inversement.

Afin de s'affranchir de ce phénomère, il serait possible de diviser chacune des valeurs d'un caractère d'un individu par la somme de ses caractères afin d'obtenir un ratio.

3 ACP

3.1 Exercice théorique

3.1.1 Question 1

Afin de pouvoir réaliser notre ACP, la première étape consiste à centrer notre matrice d'entrée. Nous réutilisons pour cela la fonction *centre* codée lors du TP0. Nous calculons ensuite les valeurs

propres et les vecteurs propres de notre matrice centrée.
Voici les pourcentages d'inertie expliquée pour chaque axe :

$$\begin{pmatrix} Axe1 & Axe2 & Axe3 \\ 63.97784 & 29.36972 & 6.652434 \end{pmatrix}$$

Nos valeurs sont donc bien représentée par l'axe 1. Mais aussi par l'axe 2 qui contient une partie élevée du pourcentage d'inertie expliquée. Cela nous conduit à un taux d'inertie expliquée de 93.34756%

3.1.2 Question 2

La matrice des composantes principales est :

$$\begin{pmatrix} 0.09396341 & -0.7993436 & 0.92315798 \\ -0.95412132 & -1.4765829 & -0.63980885 \\ -1.96850984 & 1.6200297 & -0.02174241 \\ 2.82866775 & 0.6558968 & -0.26160672 \end{pmatrix}$$

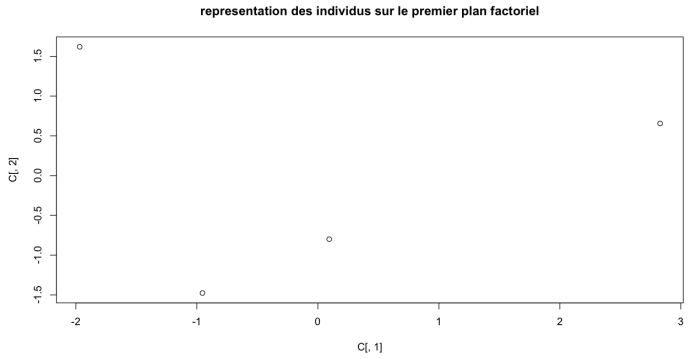

FIGURE 7 – Répartition dans le premier plan factoriel

3.1.3 Question 3

$$\begin{pmatrix} 0.8383226 & 0.3670220 & 0.40312530 \\ -0.7438325 & -0.6043354 & 0.28546782 \\ -0.8139017 & 0.5782244 & 0.05675048 \end{pmatrix}$$

FIGURE 8 – Cor, Matrice de représentation des variables

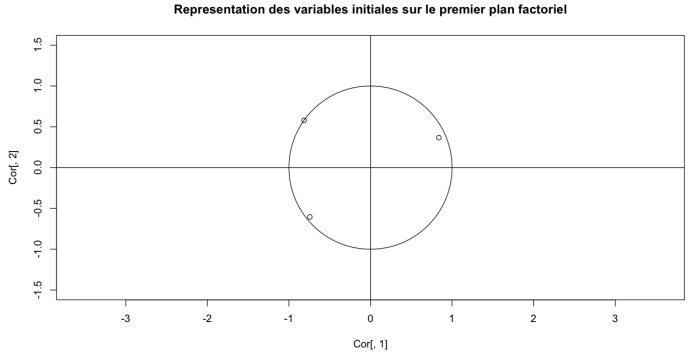

FIGURE 9 – Représentation des trois variables dans le premier plan factoriel

3.1.4 Question 4

Nous appliquons la formule de reconstitution pour k=1,2,3.

$$\begin{pmatrix} 0.5492408 & 0.9521392 & -0.5641390 \\ -2.0000000 & 1.4859869 & 0.1512789 \\ -1.5315826 & 1.0026713 & 3.8436959 \\ 2.9823418 & -3.4407975 & -3.4308358 \end{pmatrix}$$

FIGURE 10 – K=1

$$\begin{pmatrix} -0.2214326 & 0.4403667 & -0.6362579 \\ -1.0000000 & 1.3878624 & -0.4055644 \\ -0.4830087 & 0.0131806 & 2.5032092 \\ 1.7044413 & -1.8414098 & -1.4613869 \end{pmatrix}$$

FIGURE 11 – K=2

$$\begin{pmatrix} 0.5 & 1 & -0.5 \\ -1.5 & 1 & -0.5 \\ -0.5 & 0 & 2.5 \\ 1.5 & -2 & -1.5 \end{pmatrix}$$

FIGURE 12 – K=3

Lorsque K=3 on retrouve notre tableau individu-variable centré initial.

3.2 Utilisation des outils R

La fonction princomp permet de réaliser l'ACP sur un jeu de données quelconque.
$sdev permet d'avoir la racine carré des valeurs propres. Les valeurs propres permettent de calculer le pourcentage d'inertie expliquée pour chaque axe car la valeur propre est égale à l'inertie de chaque axe.
$loadings permet de visualiser la matrice contenant les vecteurs propres.
$scores permet d'obtenir la matrice des composantes principales. Les axes principaux d'inertie correspondent aux colonnes de la matrice des composantes principales et chaque valeur permet d'avoir les coordonnées de représentation de la variable sur cet axe.
plot() permet d'avoir un histogramme donnant la composante principale la plus représentative des données.
biplot() permet de visualiser la projection des données sur le plan factoriel représenté par les composantes, par défaut le premier plan factoriel. Il est possible de choisir les composantes principales sur lesquelles projeter les données en les entrant en paramètres.

3.3 ACP sur les données crabs

3.3.1 Question 1

On constate que l'on peut expliquer l'ACP en grande majorité en fonction de la première composante principale. Le pourcentage d'inertie expliquée pour le premier axe factoriel est de 80%. Les trois premières composantes principales permettent ainsi de regrouper 95% des données. Cela est dû au fait que les variables sont très corrélées entre elles comme nous l'avions déjà souligné à l'exercice 1.2.

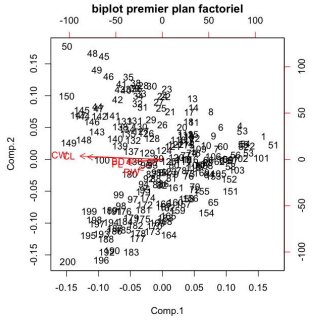

FIGURE 13 – Biplot dans le premier plan factoriel

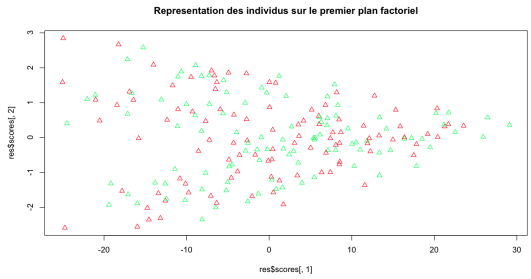

FIGURE 14 – Représentation des individus dans le premier plan factoriel

Sur ces premiers graphes, on observe que les variables sont très corrélées par rapport à la première composante. En effet, les vecteurs des variables sont tous orientés dans le même sens et de même longueur sur le premier graphe. Les individus d'espèces et sexes différents sont donc mélangés.

On peut ainsi voir que les variables initiales sont très corrélées par rapport à la première composante principale. Il nous est donc impossible de déduire une représentation permettant de distinguer visuellement les différents groupes avec nos données courantes.

$$\begin{pmatrix} -0.9807030 & -0.10531583 & 0.14511899 & 0.077273085 & 0.009972848 \\ -0.9093833 & -0.38266517 & -0.16095478 & -0.021199167 & -0.015279391 \\ -0.9987428 & 0.03170926 & 0.02463182 & -0.007419358 & -0.029080277 \\ -0.9970222 & 0.04166015 & -0.06242942 & 0.005870192 & 0.016708970 \\ -0.9827251 & -0.05315073 & 0.15970570 & -0.068132733 & 0.035754547 \end{pmatrix}$$

FIGURE 15 – Corrélation entre les variables initiales et les composantes principales

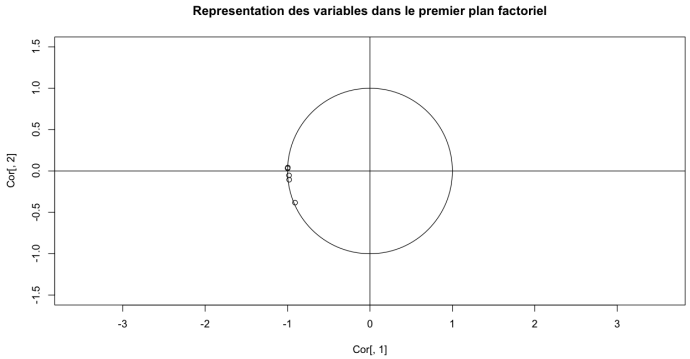

FIGURE 16 – Représentation des variables dans le premier plan factoriel

3.3.2 Question 2

Nous réalisons un pré-traitement des données crabs avant de faire l'ACP. Celui-ci consiste à faire la somme des valeurs de chaque variable par individus. On divise ensuite chaque valeur de chaque individu par la somme correspondante. On obtient ainsi un ratio des valeurs.

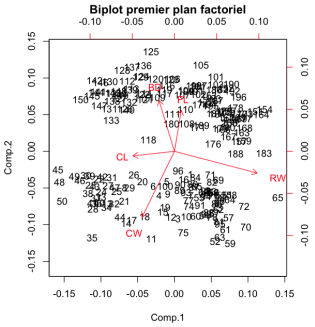

FIGURE 17 – Biplot premier plan factoriel

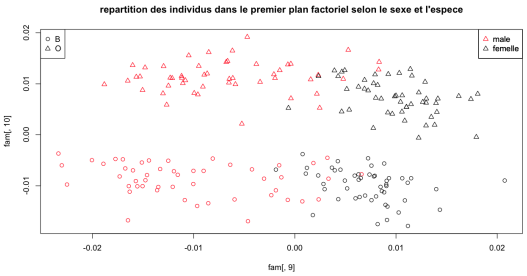

FIGURE 18 – Répartition des individus dans le premier plan factoriel selon le sexe et l'espèce

Le biplot nous permet de différencier 4 groupes de crabes différents. De plus, grâce aux vecteurs, on peut observer plus facilement les différences morphologiques de chaque crabe en fonction de leurs caractéristiques.
On observe, sur le graphe de la répartition, une séparation par espèce et par sexe. Il nous est donc possible de distinguer visuellement les différents groupes de crabes.

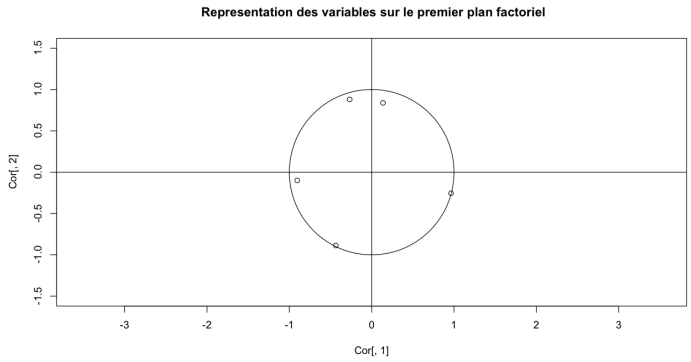

FIGURE 19 – Représentation des variables dans le premier plan factoriel

La représentation des variables initiales dans le premier plan factoriel nous permet d'observer qu'elles ne sont pas corrélées et qu'elles s'expriment différemment en fonction des composantes.

4 Conclusion

Grâce aux différentes manipulations effectuées, nous avons pu prendre en main le logiciel R et les différents outils qu'il propose. De plus, nous avons fait l'analyse descriptive des jeux de données fournis afin de pouvoir observer des tendances. Cela nous a permis d'obtenir des résultats assez fiables.
Nous avons ensuite utilisé l'ACP qui nous permet de confirmer les résultats de l'analyse descriptive. On peut ainsi exploiter un jeu de données de taille restreinte sans perte de qualité.
L'analyse descriptive et l'ACP sont complémentaires. En effet, dans le cas des données Crabs, sans prétraitement l'ACP n'est pas efficace. On a donc dû utiliser l'analyse descriptive afin d'observer qu'un traitement préalable des données était nécessaire pour exploiter les résultats.
Les graphiques créés à l'aide de R nous permettent ainsi de visualiser facilement les résultats.

5 Annexe

5.1 Fonction de catégorisation des joueurs

```
cat<-function(tab, col, num){
        i=1;
        cate<-c()
        while (i<length(col)+1){
                print(tab[,num][i]);
                if(tab[,num][i]=='NaN'){
                        cate<-c(cate, 'NaN');
                        print('NaN');
                }
                else if(tab[,num][i]>=80) {
                        cate<-c(cate, 'A');
                        print('A');
                }
                else if(80>tab[,num][i] & tab[,num][i]>=60) {
                        cate<-c(cate, 'B');
                        print('B');
                }
                else if(60>tab[,num][i] & tab[,num][i]>=40) {
                        cate<-c(cate, 'C');
                        print('C');
                }
                else if(40>tab[,num][i] & tab[,num][i]>=20) {
                        cate<-c(cate, 'D');
                        print('D');
                }
                else if(20>tab[,num][i] & tab[,num][i]>=0) {
                        cate<-c(cate, 'E');
                        print('E');
                }
                i<-i+1;
        }
        cate;
}
```