TP: Analyse descriptive univariée et bivariée

Analyse des données

Master ISEFAR - M1

1 Packages R utiles

```
rm(list = ls())
library(knitr)
library(tidyverse)
library(WVPlots)
library(corrplot)
library(questionr)
library(scales)
```

2 Durée de vie de piles

Une association de consommateurs soumet 20 piles de trois marques différentes à un même usage et mesure la durée de vie (en min) des piles. Les données sont à télécharger sur cours en ligne (fichier "piles.txt").

1. Importer le jeu de données à l'aide de la fonction suivante:

```
piles = read.table(Nom du jeu de donnees,header = .. ,sep ="")
```

On réorganise le jeu de données afin de n'avoir que deux colonnes: une colonne `duree` et une colonne `Marque` en faisant

```
piles.G <- piles %>% gather(key='Marque',value='duree')
head(piles.G)
```

2. Calculer les moyennes et écart-types par marque
3. Représenter graphiquement la distribution de la durée de vie en fonction des marques (histrogramme, boxplot). Qu'observe-t-on?

3 Les iris

Les données utilisées ici sont célèbres. Ce sont les mesures en centimètres des 4 variables suivantes: longueur du sépale (variable `Sepal.Length`), largeur du sépale (variable `Sepal.Width`), longueur du pétale (variable `Sepal.Length`) et largeur du pétale (variable `Petal.Width`) pour trois espèces d'iris (variable `Species`): setosa, versicolor et virginica. Les données sont issues du package `MASS` et peuvent être récupérées via la commande `data(iris)`.

1. Charger les données, donner le nombre d'observations et la nature des différentes variables.
2. Variable `Species`.
   a. Combien de modalités possède cette variable?
   b. Donner le nombre d'observations par modalité.
   c. Représenter la distribution de cette variable à l'aide d'un diagramme en baton et d'un camenbert.
3. Chaque variable quantitative (en fonction de la variable `Species`).
   a. Calculer quelques statistiques simples sur les variables quantitatives et représenter la distribution des différentes variables.
   b. Faire la même chose en fonction de l'espèce.
4. Lien entre variables quantitatives 2 à 2 (en fonction de la variable `Species`).
   a. Etudier les corrélation entre les variables quantitatives et les représenter.
   b. Faire des graphes des variables 2 à 2 en coloriant selon l'espèce.

```
PairPlot(iris, colnames(iris)[1:4],"Les iris", group_var = "Species")
```

4 Couleur des cheveux et des yeux

Nous étudions le jeu de données `HairEyeColor` (les données peuvent être récupérées via la commande `data(HairEyeColor)`. Il décrit la couleur des cheveux et des yeux suivant le sexe de 592 individus. Les données sont stockées dans un format array de R qui contient 3 tables correspondants à 3 tableaux de contingence:

```
dim(HairEyeColor)
is.array(HairEyeColor)
HairEyeColor[1,,] # tableau de contingence Eyes/Sex pour la couleur de cheveux "Black"
HairEyeColor[,1,] # tableau de contingence Hair/Sex pour la couleur des yeaux "Brown"
HairEyeColor[,,1] # tableau de contingence Hair/Eyes pour le sexe "Male"
```

On cherche à étudier le lien entre entre la couleur des cheveux et des yeux pour les 592 individus, sans prendre en compte le sexe. Pour obtenir la table d'intérêt, qui est la table des effectifs croisés entre les variables `Eyes`et `Hair`, faire

```
HEC = HairEyeColor[,,1] +HairEyeColor[,,2]
knitr::kable(HEC)
```

On peut les représenter avec la commande suivante

```
HEC %>% as.data.frame() %>% ggplot(aes(x=Hair,fill=Eye))+geom_bar(aes(y=Freq),stat="identity")
```

1. Calculer et représenter
   a. les fréquences conjointes
   b. les fréquences marginales
   c. les fréquences conditionnelles (profils lignes et colonnes)

2. Représenter ces fréquences par un graphique en mosaïque:

    a. Donnez les largeurs et les longueurs des bandes dans chacun des deux graphiques.

    b. Quelle est la proportion d'individus ayant les cheveux bruns? ayant les yeux verts?

    c. Parmi les individus ayant les cheveux bruns, quelle est la proportion d'individus ayant les yeux verts?

    d. Parmi les individus ayant les yeux verts, quelle est la proportion d'individus ayant les cheveux bruns?

```
mosaicplot(HEC, main = "Relation between hair and eye color")
mosaicplot(t(HEC), main = "Relation between hair and eye color")
```

3. Au vu des graphiques, est-ce que les deux variables semblent indépendantes?

4. On cherche ici à tester statistiquement l'indépendance entre les deux variables à l'aide d'un test du $\chi 2$.

```
h.chi = chisq.test(HEC)
```