SY09 Printemps 2018
TP 2 — Analyse exploratoire

1 Analyse exploratoire des Iris de Fisher

1.1 Résumés numériques et analyses graphiques

Charger le jeu de données `iris` en mémoire, puis observer le résultat des instructions suivantes :

```
> data(iris)
>
> class(iris)
> names(iris)
> iris[, 1]
> iris$Sepal.Length
> class(iris[, 1])
> class(iris$Species)
>
> summary(iris)
> apply(iris[, 1:4], 2, mean)
> cor(iris[, 1:4])
> print(cor(iris[, 1:4]), digits = 3)
> plot(iris)
> boxplot(iris)
```

Il est possible d'afficher plusieurs graphiques sur la même fenêtre :

```
> def.par <- par(no.readonly = T)
> par(mfrow = c(2, 2))
> for (i in 2:5) {
>     hist(iris[, i])
> }
> par(def.par)
```

Que fait la commande suivante ?

```
> barplot(summary(iris$Species))
```

Que font les commandes suivantes ? Comment fonctionnent-elles ?

```
> quartz()  # ou x11()
> plot(iris[, 1:4], col = c("red", "green", "blue")[iris$Species])
> quartz()  # ou x11()
> pairs(iris[, 1:4], main = "Iris de Fisher", pch = 21, bg = c("red", "green3",
  ↪  "blue")[iris$Species])
```

La séquence d'instructions suivante permet de faire un histogramme plus élaboré de manière à représenter la distribution d'un caractère en distinguant les espèces :

```
> attach(iris)
>
> # Histogrammes avec les espèces
> inter <- seq(min(Petal.Length), max(Petal.Length), by = (max(Petal.Length) -
  ↪  min(Petal.Length))/10)
> h1 <- hist(plot = F, Petal.Length[Species == "setosa"], breaks = inter)
> h2 <- hist(plot = F, Petal.Length[Species == "versicolor"], breaks = inter)
> h3 <- hist(plot = F, Petal.Length[Species == "virginica"], breaks = inter)
> barplot(rbind(h1$counts, h2$counts, h3$counts), space = 0, legend =
  ↪  levels(Species), main = "LoPe", col = c("blue", "red", "yellow"))
>
> # Graphique sur un fichier Postscript
> postscript("exemple.eps", horizontal = F, width = 12/2.5, height = 12/2.5)
> pairs(iris[2:5], main = "Les Iris", pch = 21, bg = c("red", "green3",
  ↪  "blue")[Species])
> dev.off()
>
> detach(iris)
```

1.2 Exercice : histogrammes avec variables qualitatives

Utiliser le code précédent pour définir une fonction nommée `hist.factor` qui, à partir d'une variable quantitative et d'une variable qualitative, affiche un histogramme permettant de visualiser dans chaque « bin » les effectifs selon les modalités de la variable qualitative.

2 Analyse des notes du médian de SY02 (Printemps 2014)

Charger le jeu de données contenu dans le fichier `median-sy02-p2014.csv` avec les noms de colonnes « branche » et « note » en spécifiant que les entrées contenant « ABS » (pour absent) doivent être mises à `NA` (argument `na.strings`).

En utilisant la fontion `is.na` éliminer les individus n'ayant pas de notes.

Convertir l'information de branche en chaîne de caractères et extraire l'information de branche. On pourra utiliser l'instruction

```
> substr(notes$branche, 1, 2)
```

pour extraire l'information de branche. Finalement, convertir en variable qualitative.

Visualiser les résultats de notes en fonction de la branche au moyen de la fonction `hist.factor`, puis au moyen de la fonction `boxplot`. Semble-t-il y avoir une influence de la branche sur les notes ?

Confirmer cette analyse en effectuant un test du χ^2 d'indépendance sur un tableau de contingence que l'on constituera à partir de ce jeu de données (on pourra pour cela s'inspirer du code de la fonction hist.factor).

Comparer les différences de notes entre les étudiants des branches GI et GP. Ces différences semblent-elles significatives ?

3 Analyse des données babies

3.1 Données

Charger le jeu de données babies23.data déjà prétraité.

Les variables disponibles sont :

1. bwt : le poids de naissance (birth weight) en onces,
2. gestation, la durée de la gestation en jours,
3. parity : le nombre de grossesses précédentes,
4. age : l'âge de la mère à la fin de la grossesse,
5. height : la taille de la mère en pouces,
6. weight : le poids de la mère en livres,
7. smoke : la mère a-t-elle fumé pendant la grossesse ?,
8. ed : le niveau d'éducation de la mère (0 : less than 8th grade ; 1 : 8th to 12th grade - did not graduate ; 2 : High School graduate, no other schooling ; 3 : High School + trade ; 4 : High School + some college ; 5 : College graduate ; 6 and 7 : Trade school, HS unclear).

3.2 Questions

Effectuer une analyse exploratoire des données. Vous pourrez ainsi isoler des sous-populations dans le jeu de données (par exemple, en fonction du niveau d'études, ou du tabagisme pendant la grossesse), pour ensuite en faire des résumés numériques ou des représentations graphiques. L'utilisation de tests statistiques permettra de tester la significativité des différences observées. Quelques suggestions :

1. dans un premier temps, on décrira le jeu de données : taille, nature des variables, quantité de données manquantes, ... ;
2. dans un second temps, on pourra effectuer des analyses univariées ;
3. on étudiera ensuite les liens pouvant exister entre certaines variables ; par exemple :
   — le lien entre tabagisme et niveau d'étude,
   — le lien entre tabagisme et poids du nouveau-né,
   — le lien entre tabagisme et temps de gestation ;
   — les liens entre poids, taille, ou âge de la mère et poids du nouveau-né ;
   — l'influence du temps de gestation sur le poids du nouveau-né.

3.3 Annexe : extrait de l'édition du New York Times datée du 1er mars 1995

Infant deaths tied to premature births, low weights not solely to blame

A new study of more than 7.5 million births has challenged the assumption that low birth weights per se are the cause of the high infant mortality rate in the United States. Rather, the new findings indicate, prematurity is the principal culprit.

Being born too soon, rather than too small, is the main underlying cause of stillbirth and infant deaths within four weeks of birth.

Each year in the United States about 31,000 fetuses die before delivery and 22,000 newborns die during the first 27 days of life.

The United States has a higher infant mortality rate than those in 19 other countries, and this poor standing has long been attributed mainly to the large number of babies born too small, including a large proportion who are born "small for date", or weighing less than they should for the length of time they were in the womb. The researchers found that American-born babies, on the average, weigh less than babies born in Norway, even when the length of pregnancy is the same. But for a given length of pregnancy, the lighter American babies are no more likely to die than are the slightly heavier Norwegian babies.

The researchers, directed Dr. Allen Wilcox of the National Institute of Environmental Health Sciences in Research Triangle Park, N.C., concluded that improving the nation's infant mortality rate would depend on preventing preterm births, not on increasing the average weight of newborns.

Furthermore, he cited an earlier study in which he compared survival rates among low-birthweight babies of women who smoked during the pregnancy.

Ounce for ounce, he said, "the babies of smoking mother had a higher survival rate". As he explained this paradoxical finding although smoking interferes with weight gain, it does not shorten pregnancy.