TP 3: statistiques descriptives

anna.melnykova@univ-avignon.fr

Le but de cette partie c'est de savoir mener un analyse statistique très basique sur une base de données. On va décrire les variables quantitatives, qualitatives, et explorer les rélations entre différentes variables (correlation, covariance).

Quelques commandes utiles:

- `boxplot(x)` — construire une boite à moustaches
- `hist(x)` — construire une histogramme pour visualiser le vecteur `x`.
- `table(x)` — créer une table de contingence pour un vecteur catégorique (ou deux vecteurs catégoriques, avec `table(x,y)`).

Exercice 1: Statistiques descriptives

On va commencer par charger la base de données. Les données décrivent 887 passagers de "Titanic" (naufragé à 1912), comme, par exemple, leur nom, age, prix du ticket et le statut (survecu ou pas le naufrage). Pour que le code marche, vous devez avoir préalablement sauvegardé cette base sur votre machine en la téléchargeant depuis l'espace du cours, et en changeant votre répertoire du travail pour celui qui contient le fichier "titanic.csv":

```
# Télécharger une base des données d'un site internet
titanic <- read.csv("data/titanic.csv")
dim(titanic)
head(titanic)
summary(titanic)
table(titanic$Sex)
```

- Combien il y a des variables dans la base?

- Quelles sont les variables qualitatives? Quantitatives?

- Quel est l'âge moyen de passagers du Titanic? L'âge médian?

- Combien des femmes ont fait le voyage sur Titanic? Hommes?

- Quel pourcentage de passagers a survécu le naufrage?

Maintenant, on va essayer de répondre aux questions comme, par exemple, quel était le prix du billet dans le 1er classe? 3ème classe? Quel était le taux de survie pour les femmes? Hommes? Enfants?

Pour ça, on doit manipuler les vecteurs et accéder aux éléments qui répondent au certain critère. On va commencer par la première question: quel était le prix du billet au premier classe? Dans le code suivant, on extrait d'abord le vecteur qui contient le prix des billets dans une variable séparée (on l'appellera `Billets`). Puis, on calculera les descriptives de cette variable en utilisant l'information de la variable `Pclass` de la base de données:

```
Billets <- titanic$Fare
summary(Billets)
Billets1 <- Billets[titanic$Pclass == 1]
summary(Billets1)
```

- Quel est le prix moyen pour le 1er classe?

- Adaptez le code pour calculer les mêmes caractéristiques pour les classes 2 et 3. Quel est le prix moyen dans ces 2 classes?

Si on veut visualiser la différence entre les classes, on peut utiliser les graphiques de type "boite à moustaches":

```
boxplot(Billets~titanic$Pclass, xlab = "Class", ylab = "Prix")
```

- Comment interprete-on le graphique? Est-ce qu'il y a la différence des prix significative selon la classe?

Maintenant, on va étudier le taux de survie chez les hommes et les femmes. Pour ça, on peut créer une table de contingence avec la commande `table`. Exécutez les commandes suivantes (en completant éventuellement par votre code) et répondez aux questions:

```
TableSurvie <- table(titanic$Sex,titanic$Survived)
TableSurvie
TableSurvie/sum(TableSurvie)
TableSurvie[1,]/sum(TableSurvie[1,])
TableSurvie[,2]/sum(TableSurvie[,2])
plot(TableSurvie)
```

- Combien de femmes ont survécu le naufrage? Hommes?

- Est-ce que le taux de survie est plus élevé chez les femmes ou chez les hommes?

- Quel est le taux de survie observé chez les femmes? Hommes?

- Quel pourcentage de survivants sont de femmes? Des hommes? Quel pourcentage de naufragés?

Maintenant, on va répondre aux même questions, mais pour les enfants et les adultes. Pour ça, on créera un nouveau vecteur qui divise les passagers en 2 categories:

```
IsChild <- (titanic$Age<=18)
TableSurvie2 <- table(IsChild,titanic$Survived)
rownames(TableSurvie2) <- c("Adulte","Enfant")
plot(TableSurvie2)
```

- Adaptez le code écrit pour la première table de survie et répondez aux questions correspondantes (mais pour enfants/adultes cette fois).

Exercice 2: comparer les distributions bivariées

On va étudier 2 bases des données synthétiques que vous devez télécharger depuis l'espace du cours. Chaque base contient des observations d'un paire des variables (X_1, Y_1) et (X_2, Y_2) respectivement.

```
set1 <- read.csv("data/set1.csv")
set2 <- read.csv("data/set2.csv")
# On analyse d'abord la premiere base des données
summary(set1)
sd(set1[,1])
sd(set1[,2])
cor(set1[,1], set1[,2])
# ... et puis la deuxième:
summary(set2)
sd(set2[,1])
sd(set2[,2])
cor(set2[,1], set2[,2])
```

- Est-ce que le paire (X_1, Y_1) et corrélée? Et (X_2, Y_2)?

- Est-ce qu'on peut plutôt dire que les 2 paires suivent la même distribution? Pourquoi?

- Pour mieux visualiser les données, on fera des boxplots de chaque base de données:

```
boxplot(set1)
boxplot(set2)
```

- Est-ce qu'il y a une différence entre les 2 paires des variables, à votre avis?

- Finalement, on va faire le graphique "nuage des points" pour tracer la dépendance

```
plot(set1[,1], set1[,2], lwd = "2", col = "salmon")
plot(set2[,1], set2[,2], lwd = "2", col = "darkblue")
```

- Commentez les boxplots obtenus.