Exam: Data Analysis with R

Sonja Hartnack, Terence Odoch & Muriel Buri

October 2018

Please provide the solutions for the following two exercises 1 and 2 as an R-Script (.R).

Exercise 1

(a) Open a new R-Script and read in the `lepto_exam.csv` data file. This study data presents a serology survey of leptospira sero-prevalence in rural and urban areas. Within the data set you find the following variables:

- `gender`: factor variabe with two levels "female" and "male"
- `age`: integer variable (years)
- `exposure`: factor variabe with two levels "urban" and "rural"
- `antibodies`: factor variabe with two levels "absent" and "presence"
- `num.rats`: integer variable (amount of rats seen in the last 3 months)

```
rm(list = ls())
lepto_exam <- read.csv("lepto_exam.csv", sep = ",")
```

(b) Save the `lepto_exam` data frame as `lepto` data frame.

```
lepto <- data.frame(lepto_exam)
```

(c) Have a look at the `str(...)`, `summary(...)`, `head(...)`.

```
str(lepto)
summary(lepto)
head(lepto)
```

(d) How many observations are in the `lepto` data frame?

```
dim(lepto)
```

(e) What data type does each variable have? Is this appropriate?

```r
# typeof(lepto$X)
# typeof(lepto$gender)
# typeof(lepto$age)
# typeof(lepto$exposure)
# typeof(lepto$antibodies)
# typeof(lepto$num.rats)

class(lepto$X)
class(lepto$gender)
class(lepto$age)
class(lepto$exposure)
class(lepto$antibodies)
class(lepto$num.rats)

sapply(lepto, class)

# gender: factor
# age: integer
# exposure: factor
# antibodies: factor
# num.rats: integer
```

(f) Plot in one graph, two boxplots showing the distribution of `age` for both `gender`.

```r
boxplot(lepto$age ~ lepto$gender)
```

(g) Make a 2-by-2 table for `exposure` and `antibodies`.

```r
table(lepto$exposure, lepto$antibodies)
```

(h) Quantify the effect of the two potential risk factors `exposure` and `gender` with an odds ratio (OR). What is the 95%-confidence interval? Interpret the odds ratio as well as the 95%-confidence interval.

```r
table(lepto$exposure, lepto$antibodies)
fisher.test(table(lepto$exposure, lepto$gender))
fisher.test(table(lepto$exposure, lepto$antibodies))
# CAREFUL WITH THE INTERPRETATION!
```

Exercise 2

(a) Simulate a continuous vector with length $n = 100$ representing the body weight of adult goats. Before simulating, set the seed by `set.seed(2018)`.

```
rm(list = ls())
set.seed(2018)
n <- 100
gw <- rnorm(n, mean = 35, sd = 10)
hist(gw, prob = TRUE)
lines(density(gw), lwd=2, col="red")
abline(v=35, lwd=3, col="blue")
abline(v=35+10, lwd=3, lty = 2, col="green")
abline(v=35-10, lwd=3, lty = 2, col="green")
```

(b) Assess if the above simulated variable is normally distributed.

```
qqnorm(gw)
qqline(gw, col = "red", lwd = 4)
shapiro.test(gw)
```

(c) Replace the first 10 observations of your previously simulated vector with the values 200.

```
gw.200 <- gw
gw.200[1:10] <- 200
# gw.200 <- replace(gw, 1:10, 200)
hist(gw.200, prob = TRUE)
lines(density(gw.200), lwd=2, col="red")
abline(v=35, lwd=3, col="blue")
abline(v=35+10, lwd=3, lty = 2, col="green")
abline(v=35-10, lwd=3, lty = 2, col="green")
```

(d) Check again for normality.

```
qqnorm(gw.200)
qqline(gw.200, col = "red", lwd = 4)
shapiro.test(gw.200)
```