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

```
lepto <- data.frame(lepto_exam)</pre>
```

- (c) Have a look at the str(...), summary(...), head(...).
- (d) How many observations are in the lepto data frame?
- (e) What data type does each variable have? Is this appropriate?
- (f) Plot in one graph, two boxplots showing the distribution of age for both gender.
- (g) Make a 2-by-2 table for exposure and 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.

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).
- (b) Assess if the above simulated variable is normally distributed.
- (c) Replace the first 10 observations of your previously simulated vector with the values 200.
- (d) Check again for normality.