Biostatistics: Exercise 02

Beate Sick, Lisa Herzog

22.09.2020

Exercise 01: R Markdown (voluntary)

R markdown is a notebook interface, which enables to combine text with code to generate a nice output and to perform reproducible research. You can use multiple languages including R. You can do your exercises in R Markdown and save them as .Rmd files or you can stick with the R scripts introduced in the previous exercise. It is not compulsory to do the exercises in R Markdown, but it is helpful to know about it, which is why we introduce it here. To create your own Rmd-file in RStudio, you can do the following:

- Go to `File -> New file -> R Markdown...`. Specify the title of your document, the author and let the default output format in HTML. Then click `Ok`. You should now see so called R chunks and text that is written in Markdown. This file already provides you with some basics.
- Save the file in a folder you want. Then you can translate your file via Knitr into a HTML. Therefore, click on `Knit` in the upper row. Knit your file every time you change something in the text/chunk options to see the differences in the output file.
- In the R chunks you can do all calculations/analyses in R. They are defined with ```` ```{r, ...} ``` ```` Click on the green arrow on the very right of a R chunk. What happens?
- The R chunks have many options. One of the two most important options are `include` and `echo` to control the output of a chunk.
  - What is `include = FALSE` doing in the first chunk?
  - What is `echo = FALSE` doing in the last chunk?
- Markdown allows you to structure your document.
  - What is `##` doing? What happens if you add another `#`.
  - What is `**` doing?
  - Replace `**` with `_` for the word Knit. What happens?

Up to now, you should already be able to create your own .Rmd file and to work with it. If you are interest in working with R Markdown and you aim to learn more about it, you can look into the following tutorials:

- R Markdown: https://rmarkdown.rstudio.com/lesson-1.html
- Markdown: https://www.markdowntutorial.com/

Exercise 02: Univariate & bivariate data visualization

In this exercise we consider a slightly modified version of the data set from last week. It contains a survey of school children and it is stored in CSV format (*survey.csv*). The data set can be downloaded from the webpage.

- Read in the data. You can use `read.table(..., sep=";", header=TRUE)`. Make sure to specify the complete path to your file. In addition, you could use `getwd()`, which shows you the current working directory of R and change that with `setwd()` to the directory of your file.

To gain an overview over the data calculate some characteristic measures of the distribution:

- Determine the mean and the median of `Arm.span` (Hint: `mean(), median()`).
- Calculate the range, variance, standard deviation and interquartile range of `Arm.span` (Hint: `range(), var(), sd(), IQR()`).

Univariate data visualization:

- Visualize the distribution of the variable `Arm.span` using a histogram (Hint: `hist(, breaks=)`). Try out different breaks.
- Visualize the variable `Arm.span` using a boxplot and add notches (Hint: `boxplot(...,notch=TRUE)`). Does a boxplot make sense if you only have one variable? Why - Why not?
- Visualize the four variables `Arm.span, Height, Age, Hand.span` within one figure using a boxplot for each variable. Does this visualization make sense? (Hint: `boxplot(dat[,c("Arm.span",...)])`)

Bivariate data visualization:

- Determine the contingency table between `Eye.color` and `Hair.color` (Hint: `table()`).
- Display the frequencies of the contingency table as mosaic plot (Hint: `mosaicplot()`). What do you observe?
- How does `Arm.span` depend on `Height`? Plot the two variables against each other using a scatterplot (Hint: `plot()`).

Exercise 03: Descriptive analysis

The data set for this exercise is from a study on guinea pigs. The study investigates the effects of Vitamin C consumption on teeth growth. Therefore, the guinea pigs were fed by orange juice (`OJ`) or ascorbic acid (`VC`) using different doses of Vitamin C (0.5, 1.0, 2.0). The data contains the following variables:

R name	Meaning
`len`	mean of teeth length
`supp`	supplement type (OJ or VC)
`dose`	vitamin C dose in mg

In order to access the data, you can use the following code:

```
# The data is contained in the R package data sets. With data(),
# the data is loaded into the workspace.
data("ToothGrowth")

# Then we can assign the data set to a new R object dat
# (easier for coding purposes than working with ToothGrowth directly)
dat <- ToothGrowth

# Consider the first few lines of dat
head(dat)
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
## 4  5.8   VC  0.5
## 5  6.4   VC  0.5
## 6 10.0   VC  0.5
```

- How many guinea pigs have been included into this study?
- Investigate the three variables of the data set graphically using appropriate plots.
- Does the distribution of the tooth length depend on the supplement type? Illustrate your answer with an appropriate plot.
- Does the distribution of the tooth length depend on the Vitamin C dose? Illustrate your answer with an appropriate plot.
  What percentage of guinea pigs in group 3 has longer teeth than 75% of the guinea pigs in group 2?
- Is the Vitamin C dose different for the two supplement types? Take subsets of the data using e.g. `dat$oj<-subset(dat, supp=="OJ")` and `dat$vc<-subset(dat, supp=="VC")` and visualize them.