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?

```
# Clicking on the green arrow enables to run the chunk and evaluates
# the code in R
```

- 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?

```
# It does neither include the R code nor the output.
```

- What is `echo = FALSE` doing in the last chunk?

```
# It does not include the R code.
```

- Markdown allows you to structure your document.
  - What is `##` doing? What happens if you add another `#`.

```
# It defines the headings. Subheadings are specified with multiple
# repetitions of #
```

- What is `**` doing?

```
# Makes a part of the text bold.
```

- Replace `**` with `_` for the word Knit. What happens?

```
# Makes a part of the text italic.
```

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.

```
# read in the data. Paste0 combines the string stored in "dir"
# with "data/survey" to get the  complete path to the file
dat <- read.table(file = paste0(dir,"data/survey.csv"),
                  sep = ";", header = TRUE)
```

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()`).

```
# mean
mean(dat$Arm.span)
## [1] 179.5

#median
median(dat$Arm.span)
## [1] 181
```

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

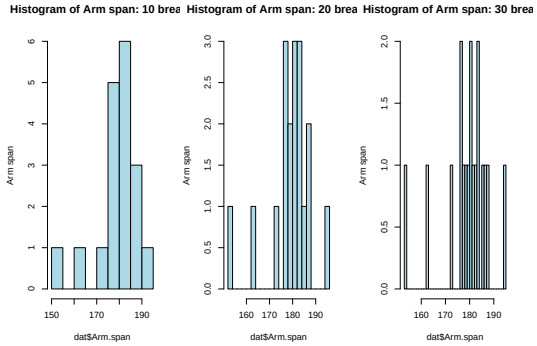
```
# range
range(dat$Arm.span)
## [1] 153 195

# Variance
var(dat$Arm.span)
## [1] 88.61765

# Standard deviation
sqrt(var(dat$Arm.span))
## [1] 9.413695
# or sd(dat$Arm.span)

# IQR
IQR(dat$Arm.span)
## [1] 6.75
```

Univariate data visualization:

- Visualize the distribution of the variable `Arm.span` using a histogram (Hint: `hist(, breaks=)`). Try out different breaks.

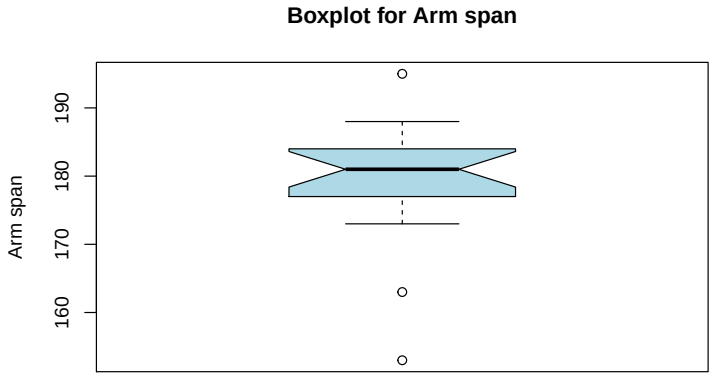
```
# The best plot to visualize the continuous variable is a boxplot
par(mfrow = c(1,3)) # we want the two plots next to each other
```

```r
hist(dat$Arm.span,
     main = "Histogram of Arm span: 10 breaks",
     ylab = "Arm span",
     col = "lightblue",
     breaks = 10)

hist(dat$Arm.span,
     main = "Histogram of Arm span: 20 breaks",
     ylab = "Arm span",
     col = "lightblue",
     breaks = 20)

hist(dat$Arm.span,
     main = "Histogram of Arm span: 30 breaks",
     ylab = "Arm span",
     col = "lightblue",
     breaks = 30)
```



- 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?

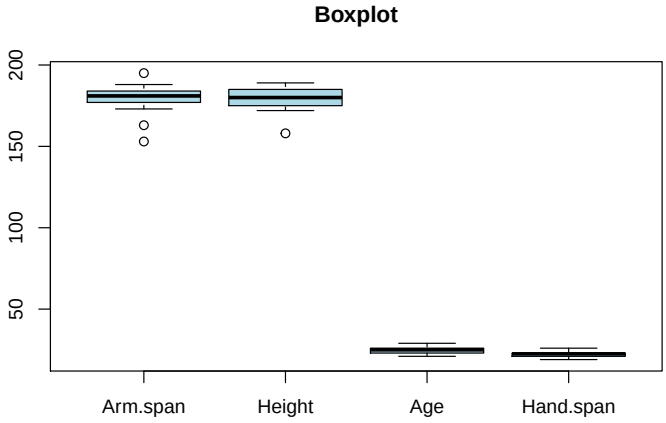
```r
boxplot(dat$Arm.span, notch=TRUE,
        main="Boxplot for Arm span",
        ylab="Arm span",
        col="lightblue")
```




```
# If we only have one variable a histogram is to be preferred
# because it contains much more information than the boxplot.
# However, boxplots are nice to compare multiple continuous
# variables to each other if the distribution of the continuous
# variables are unimodal.
```

- 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",...)])`)

```
boxplot(dat[,c("Arm.span","Height","Age","Hand.span")],
        main="Boxplot",
        col="lightblue")
```




```
# This plot does not make sense since the variables are
# on different scales (age in years, height in cm, etc.)!
```

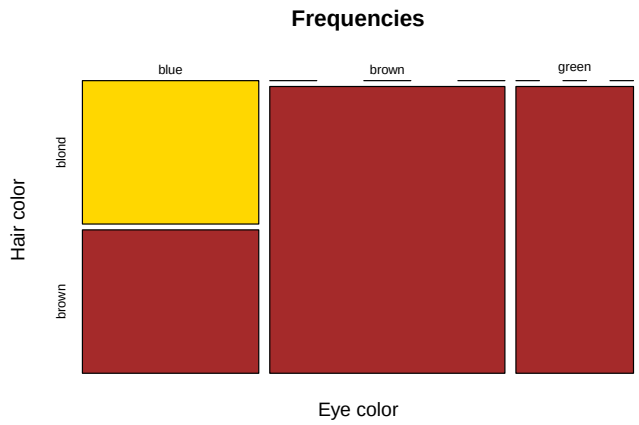
Bivariate data visualization:

- Determine the contingency table between `Eye.color` and `Hair.color` (Hint: `table()`).

```
table(dat$Eye.color, dat$Hair.color)
##
##          blond brown
##   blue       3     3
##   brown      0     8
##   green      0     4
```

- Display the frequencies of the contingency table as mosaic plot (Hint: `mosaicplot()`). What do you observe?

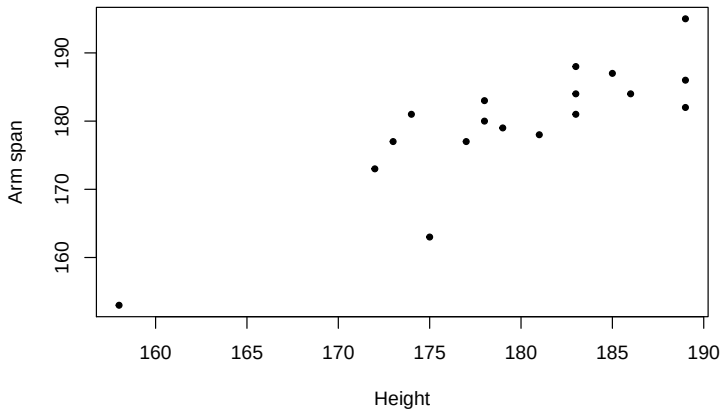
```
mosaicplot(table(dat$Eye.color, dat$Hair.color),
           col = c("gold", "brown"),
           xlab = "Eye color",
           ylab = "Hair color",
           main = "Frequencies")
```




```
# All blondes in the data set have blue eyes while brown haired
# people have blue, brown and green eyes.
```

- How does `Arm.span` depend on `Height`? Plot the two variables against each other using a scatterplot (Hint: `plot()`).

```
plot(dat$Height, dat$Arm.span,
     xlab = "Height",
     ylab = "Arm span",
     pch = 20)
```




```
# There seems to be a relationship between height and arm span.
# The larger the height, the larger the arm span.
```

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:

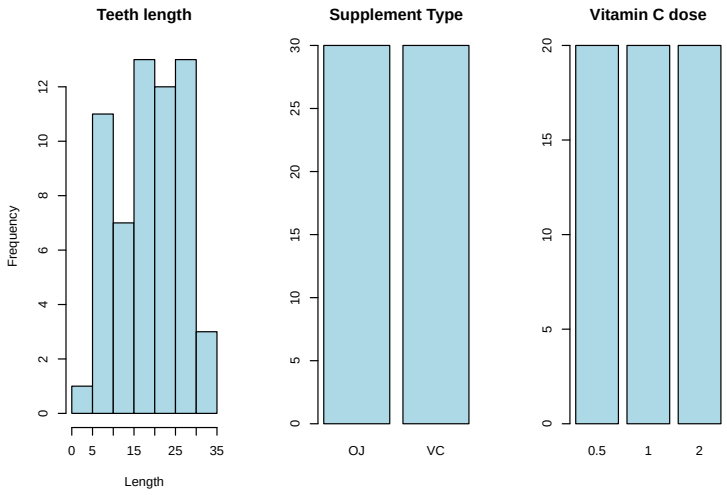
```
# The data is contained in the R package data sets. With data(),
# the data is loaded into the workspace.
data("ToothGrowth")

# 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?

```r
# Each row contains the information of one guinea pig since there are
# 60 rows, 60 guinea pigs have been included
dim(dat)
## [1] 60  3
```

- Investigate the three variables of the data set graphically using appropriate plots.

```r
par(mfrow = c(1,3))

# Since the teeth length is a continuous measure we use a histogram
hist(dat$len,
     col = "lightblue",
     xlab = "Length",
     main = "Teeth length")

# The variables supp and dose are categorical, i.e. we can  visualize them
# using barplots (We first have to calculate the frequencies with table()).
barplot(table(dat$supp),
        main = "Supplement Type",
        col = "lightblue")

barplot(table(dat$dose),
        main = "Vitamin C dose",
        col = "lightblue")
```

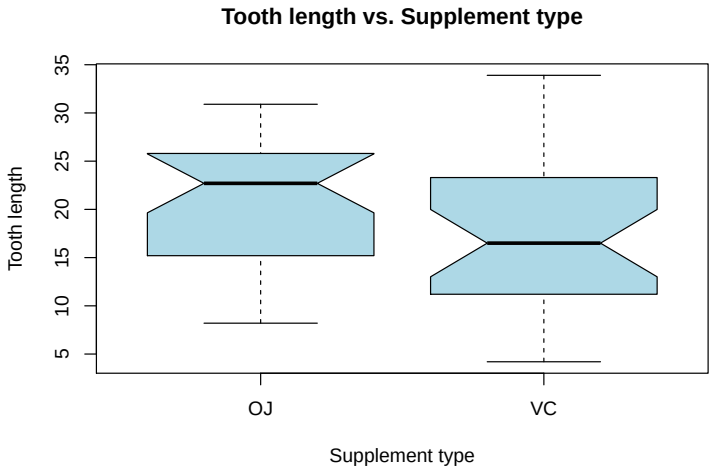



```r
# We see that 30 guines pigs were fed with OJ, 30 with VC
# and each dose was given to 20 guinea pigs.
```

- Does the distribution of the tooth length depend on the supplement type? Illustrate your answer with an appropriate plot.

```r
# Teeth length is continuous while the supplement type is categorical.
# We therefore choose a boxplot to investigate the association.
boxplot(dat$len ~ dat$supp, notch=TRUE,
        xlab = "Supplement type",
        ylab = "Tooth length",
        main = "Tooth length vs. Supplement type",
```

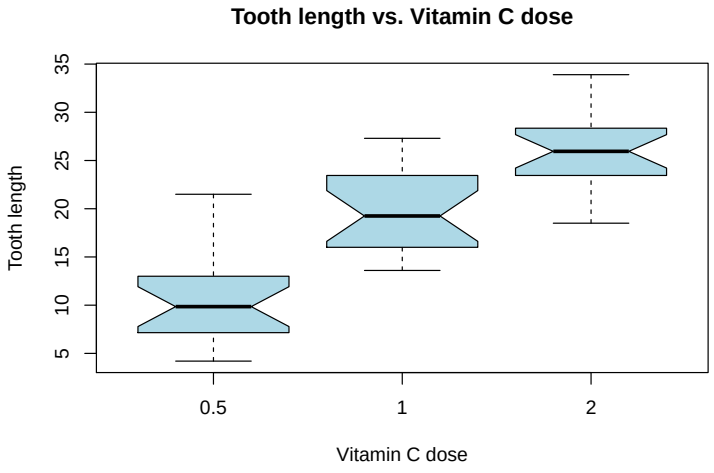
```
        col = "lightblue")
```




```
# OJ seems to have a higher effect on teeth growth than VC.
```

- 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?

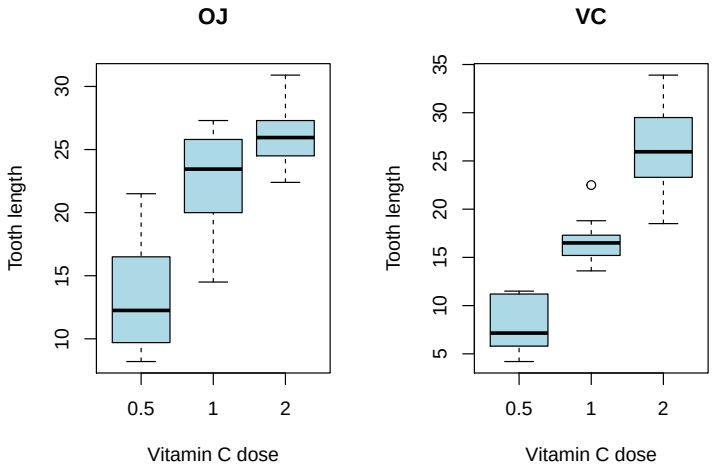
```
boxplot(dat$len~dat$dose, notch=TRUE,
        xlab="Vitamin C dose",
        ylab="Tooth length",
        main="Tooth length vs. Vitamin C dose",
        col="lightblue")
```




```
# Obviously, the higher the Vitamin C dose, the larger the tooth growth.

# From the boxplot we can see that the 75% quantile of dose 2 group is
# on the same value as the 25% of dose 3 group. Therefore, 75% of the
# guinea pigs in group 3 have larger teeth than 75% in group 2.
```

- Describe the effect of Vitamin C dose on tooth length 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.

```
# Take subsets of the data

# consider guinea pigs which were fed with OJ
dat_oj <- subset(dat, supp=="OJ")
dat_oj[1:3,]
##     len supp dose
## 31 15.2   OJ  0.5
## 32 21.5   OJ  0.5
## 33 17.6   OJ  0.5
# consider guinea pigs which were fed with VC
dat_vc <- subset(dat, supp=="VC")
dat_vc[1:3,]
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5

# plot the tooth length in both groups
par(mfrow=c(1,2))

boxplot(len ~ dose,
        xlab = "Vitamin C dose",
        ylab = "Tooth length",
        main = "OJ",
        col = "lightblue",
        data = dat_oj)

boxplot(len ~ dose,
        xlab = "Vitamin C dose",
        ylab = "Tooth length",
        main = "VC",
        col = "lightblue",
        data = dat_vc)
```



```
# In both groups, an increase in Vitamin C dose leads to an increase
# in teeth length. In The VC group, the effect seems to be slightly
# higher (larger maximum). In the OJ group a Vitamin C dose of 1
# results in a higher teeth growth than in group VC.
```