Practical Exercises for **Day 2**

Sonja Hartnack, Terence Odoch & Muriel Buri

October 2017

Exercise 4

What is conceptionally the difference between these bracket types ([...], (...))?

```
chickwts[, 2]
summary(aov(weight ~ feed, data = chickwts))
```

```
# SOLUTION: [...] the squared brackets we need to select rows and columns of a
# data frame.
# (...) the round brackets we need around function calls,
# e. g.:
subset(...) # to define a subset
c(...) # to define a vector
data.frame(...) # to define a data frame
```

Exercise 5

- How many levels has the factor variable `trt` from `bacteria`?

```
str(bacteria)
head(bacteria$trt)
table(bacteria$trt)
levels(bacteria$trt)
nlevels(bacteria$trt)
```

- Define a new variable `trt.new` in which you combine the levels `drug` and `drug+` into one single level and label it as `treated`. The new variable `trt.new` should in the end have two levels: `placebo` and `treated`.

```
table(bacteria$trt)
# OPTION 1:
# Test how many levels are in the variable "trt"?
levels(bacteria$trt)
```

```
bacteria$trt.new <- bacteria$trt
# Overwrite the levels "placebo", "drug", "drug+" with new
# levels called "placebo", "drug", "drug" --> combine "drug" and "drug+"
levels(bacteria$trt.new) <- c("placebo", "drug", "drug")
# Do table for variable "trt" and "trt.new" to see if you combined correctly
table(bacteria$trt)
table(bacteria$trt.new)
# Rename the levels from "placebo", "drug" to "placebo", "treated"
levels(bacteria$trt.new) <- c("placebo", "treated")
# Do another table to check if you did everything correctly:
table(bacteria$trt.new)
```

- Do summary statistics for `placebo` and `treated` group.

```
summary(bacteria)
table(bacteria$trt.new)
barplot(table(bacteria$trt.new))
table(bacteria$trt.new, bacteria$ap)
table(bacteria$trt.new, bacteria$y)
plot(table(bacteria$trt.new, bacteria$y))
```

Exercise 6

- Load data set `ToothGrowth`.

```
data(ToothGrowth)
str(ToothGrowth)
head(ToothGrowth)
```

- Do summary statistic (numerically and graphically).

```
table(ToothGrowth$supp)
tapply(ToothGrowth$len, ToothGrowth$supp, mean)
tapply(ToothGrowth$len, ToothGrowth$supp, median)
tapply(ToothGrowth$len, ToothGrowth$supp, sd)
```

```
tapply(ToothGrowth$len, ToothGrowth$dose, mean)
tapply(ToothGrowth$len, ToothGrowth$dose, median)
tapply(ToothGrowth$len, ToothGrowth$dose, sd)

barplot(table(ToothGrowth$supp))
hist(ToothGrowth$len)
# install.packages("graphics")
library("graphics")
coplot(len ~ dose | supp, data = ToothGrowth, panel = panel.smooth,
       xlab = "ToothGrowth data: length vs dose, given type of supplement")
```

- Define additional column `dose.fac` by converting the numeric variable `dose` into a factor variable.

```
table(ToothGrowth$dose)
class(ToothGrowth$dose)
ToothGrowth$dose.fac <- factor(ToothGrowth$dose, levels = c("0.5", "1", "2"))
class(ToothGrowth$dose.fac)
table(ToothGrowth$dose.fac)
```

- Are the tooth length measurements normally distributed within the treatment (`supp`: VC or OJ) and within in the different doses (`dose`: 0.5, 1, 2)?

```
# supp: VC, OJ
sub.OJ <- subset(ToothGrowth, supp == "OJ")
sub.VC <- subset(ToothGrowth, supp == "VC")
# graphically
qqnorm(sub.OJ$len)
qqline(sub.OJ$len)
qqnorm(sub.VC$len)
qqline(sub.VC$len)
# with a statistical test
shapiro.test(sub.OJ$len)
shapiro.test(sub.VC$len)
# dose: 0.5, 1, 2
sub.0.5 <- subset(ToothGrowth, dose.fac == "0.5")
sub.1 <- subset(ToothGrowth, dose.fac == "1")
```

```r
sub.2 <- subset(ToothGrowth, dose.fac == "2")
# graphically
qqnorm(sub.0.5$len)
qqline(sub.0.5$len)
qqnorm(sub.1$len)
qqline(sub.1$len)
qqnorm(sub.2$len)
qqline(sub.2$len)
# with a statistical test
shapiro.test(sub.0.5$len)
shapiro.test(sub.1$len)
shapiro.test(sub.2$len)
```

Exercise 7

- Import the data set `perulung_ems.csv` (taken from Kirkwood and Sterne, 2nd edition) into R.
  Data from a study of lung function among children living in a deprived suburb of Lima, Peru.
  Variables:
  - `fev1`: in liter, "Forced Expiratory Volume in 1 second" measured by a spirometer. This is the maximum volume of air which the children could breath out in 1 second
  - `age`: in years
  - `height`: in cm
  - `sex`: 0 = girl, 1 = boy
  - `respsymp`: respiratory symptoms experienced by the child over the previous 12 months
- What *delimiter* do you need to choose?

```r
# OPTION 1:
# install.packages("readr")
library("readr")
lung <- read_delim("~/Dropbox/201710_Makerere/03_Exercises/data/perulung_ems.csv",
                   ";", escape_double = FALSE, trim_ws = TRUE)
lung <- data.frame(lung)
# OPTION 2:
# Import .csv file with the help of the read.csv function
```

```
# Be sure to add sep = ";" so that we separate the columns.
lung <- read.csv("C:\\Users\\Exercises\\data\\perulung_ems.csv", sep = ";")
head(lung)
str(lung)
```

- Do all variables have the correct data type (numeric, integer, factor)? If not, do correct and / or define them.

```
head(lung)
str(lung)
lung$sex <- factor(lung$sex, levels = c("0", "1"))
# levels(lung$sex) <- c("female", "male")
# levels(lung$sex)[levels(lung$sex)=="0"] <- "female"
# levels(lung$sex)[levels(lung$sex)=="1"] <- "male"
# tapply(lung$fev1, lung$sex, mean)
lung$respsymptoms <- factor(lung$respsymptoms, levels = c("0", "1"))
```

```
library(usdm)
# check for multicollinearity by using variance inflation factors
# cerate a dataframe just with the three continuous/numeric variables fevs, age and height
try.vif <- lung[,c("fev1","height","age")];
# perform scatterplots for these three variables
pairs(try.vif)
# get the three VIF, as a rule of thumb they should be < 3
vif(try.vif)
```

```
# Check for heteroscedascity or homogeneity of variances
?bartlett.test
data("chickwts")
bartlett.test(weight ~ feed, data = chickwts)
```