Practical Exercises for **Exercise Collection**

Sonja Hartnack, Terence Odoch & Muriel Buri

July 2019

Exercise 1: Terminology

Group the following terminology items into the three categories:

(1) sample & variables

(2) hypothesis testing & statistical modelling

(3) descriptive statistics

- alternative hypothesis
- anova
- barplot
- binary
- binomial
- Bonferroni
- boxplot
- categorical
- Chisquare test
- confounding
- contingency table
- continuous
- correlation coefficient
- count
- data format
- data point
- data type
- degree of freedom
- dependent variable
- effect size
- error
- explanatory variable
- factor
- Fisher's exact test
- histogram
- hypothesis testing
- hypothesis tests
- independent variable
- integer
- interaction
- intercept
- IQR
- linear model
- linear regression
- logistic regression
- mean
- median
- multiple comparison
- nominal
- normal
- null hypothesis
- numeric
- observation
- odds ratio
- ordinal
- outcome
- paired samples
- poisson
- population
- predictor
- proportion
- p-value
- QQ-plot
- quantile
- range
- regression coefficient
- residuals
- response
- sample
- sampling variation
- scatter plot
- significance
- single-sided test
- skewed data
- slope
- standard deviation
- standard error
- student t-distribution
- treatment effect
- t-test
- two-sided test
- unpaired samples
- variable
- variance
- vector

Exercise 2: Getting to know R and `chickwts`

(a) Open R Studio.

(b) Open a new R-Script.

(c) Load data set `chickwts`.

```
# ?chickwts
data(chickwts)
head(chickwts)
```

Exercise 3: Summary Statistics with `chickwts`

(a) Do summary statistics (numerically and graphically).

(b) For advanced R users: Try an anova (are the assumptions fulfilled?) and a Tukey-Anscombe plot. Try a histogram with a density line on top. ...

Exercise 4: Data import to R: `perulung_ems.csv`

(a) 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

(b) What *delimiter* do you need to choose?

(c) Do summary statistics (numerically and graphically).

(d) Plot a boxplot.

Exercise 5: Defining a new data frame

(a) Create a data frame with 3 columns.

Exercise 6: Get to know `bactera` data set

(a) Install package `MASS`.

(b) Load data set `bacteria`.

(c) Describe in your own words what the data set `bacteria` contains.

(d) Do summary statistic (numerically and graphically).

(e) Select only observations collected during the second week.

Exercise 7: Bracket types

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

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

Exercise 8: Factor variables

(a) How many levels has the factor variable `trt` from `bacteria`?

(b) 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`.

(c) Do summary statistics for `placebo` and `treated` group.

Exercise 9: Get to know `ToothGrowth` data set

(a) Load data set `ToothGrowth`.

(b) Do summary statistic (numerically and graphically).

(c) Define additional column `dose.factor` by converting the numeric variable `dose` into a factor variable.

(d) 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)?

Exercise 10: Summary Statistics

Apply the summary statistics to the `perulung_ems` and `ToothGrowth` data set.

Exercise 11: Data Plausibility Checks

(a) What can go wrong?

(b) Identify different strategies for spotting these potential errors.

- Logical errors
- Spelling mistakes

(c) Import the data set `bacteria_plausibility_check.csv` to R.

(d) Detect the **six** errors in the imported data set `bacteria_plausibility_check.csv` in R.

(e) Find possible solutions in R how to handle these challenges.

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

Exercise 12: Missing Values

(a) Check out the difference between the different missing values.

```
y1 <- c(2, 4, 3, NA, 6, 1)
y2 <- c("diseased", "healthy", NA, "NA")
y3 <- c(1, "NA", 0, 1, NaN)
#
is.na(y1)
which(is.na(y1))
is.na(y2)
which(is.na(y2))
is.na(y3)
which(is.na(y3))
is.nan(y3)
```

(b) Create a vector with missing values and determine the mean and median.

(c) If `x = c(22,3,7,NA,NA,67)` what will be the output for the R statement `length(x)`?

(d) If `x = c(NA, 3, 14, NA, 33, 17, NA, 41)` which line of R code removes all occurrences of NA in `x`.

(e) If `y = c(1, 3, 12, NA, 33, 7, NA, 21)` what R statement will replace all occurrences of NA with 11?

(f) If `x = c(34, 33, 65, 37, 89, NA, 43, NA, 11, NA, 23, NA)` then what will count the number of occurrences of NA in `x`?

(g) Create the vector `x1`. Then, find again the number of missing values and their position.

```
x1 <- c(rnorm(10,5,2), NA, 5:12, NA, 6, 7.5, NA)
```

(h) Now, create the vector `x2` and assess the difference to `x1`.

```
x2 <- c(rnorm(10,5,2), NA, 5:12, NA, 6, 7.5, NA, log(-2))
```

(i) What is the meaning of "NA" versus "NaN"?

(j) Replace the missing values in `x1` with a 0. Check then that the NAs are no longer present. Try two different commands to coerce the NAs into 0.

Exercise 13: Data import to R: `water_errors.csv`

(a) Import the data set `water_errors.csv` to R: A data frame with 61 observations on the following 6 variables.

- **location**: a factor with levels `North` and `South` indicating whether the town is as north as Derby.
- **town**: the name of the town.

- **mortality**: averaged annual mortality per 100.000 male inhabitants.
- **hardness**: calcium concentration (in parts per million).
- **smoker**: If there are any smokers living in town.
- **num.of.cig**: In case, smokers live in town, what number of cigarettes do they smoke per day.

(b) Detect the errors in the imported data set `water_errors.csv` in R.

(c) Find possible solutions in R how to handle these challenges.

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

Exercise 14: t-test in R

(a) Load the data set `ToothGrowth` within R and apply the two-sided two sample t-test to suitable variables of the data set.

```
data(ToothGrowth)
```

(b) Interpret the results.

(c) Read in the data set `perulung_ems` and apply the two-sided t-test to suitable variables of the `perulung_ems` data set and interpret the results.

Exercise 15: Chi-square test in R

(a) Apply the Chi-square test and the fisher exact test to the whole `bacteria` data set.

(b) Apply the Chi-square test and the fisher exact test to the subset of `bacteria` containing only the observations taken in week 2 (cf. Exercise 3). Are there any issues?

(c) Repeat this exercise by using the (previously defined) combined `trt.new` variable (cf. Exercise 5) with the two levels `treated` and `drug`.

(d) Could you also obtain the odds ratios?

(e) Try also a logistic regression in R. Ask Google for help!

Exercise 16: Outside plot frame

(a) Type `demo(graphics)` in your console and press enter. This command shows you a nice demonstration of possible R graphics.

```
# After the demonstration us the following commands:
dev.off()
par(mfrow=c(1,1))
```

(b) Change the x-axis and y-axis labelling of a boxplot plotting the `len` variable of the `ToothGrowth` data set.

```
data("ToothGrowth")
boxplot(ToothGrowth$len)
```

(c) How do you set a main title for your above plot?

(d) What does the following command do?

```
par(mfrow=c(2,2))
```

(e) We have six different feed types in `chickwts`. Try to plot two separate boxplots for `casein` and `horsebean` and set the same minimum and maximum for the y-axis. Use the function `subset` for doing so.

(f) How do you enlarge the font size of the axis as well as the axis labels of the following plot with the `perulung` data set?

```
lung <- read.csv("perulung_ems.csv", sep=";")
par(mfrow=c(1,1))
plot(lung$fev1, lung$height)
```

(g) Label the x-axis of the following plot with ”Vitamin C in μg”. Use the greek letter for μ.

```
plot(ToothGrowth$dose, ToothGrowth$len)
```

(h) Read `http://www.statmethods.net/advgraphs/parameters.html`.

Exercise 17: Inside the square of the plot

(a) Type `demo(graphics)` in your console and press enter. This command shows you a nice demonstration of possible R graphics.

```
# After the demonstration us the following commands:
dev.off()
par(mfrow=c(1,1))
```

(b) Add a legend to the following barplot. Are there several different solutions for this?

```
library("MASS")
data("bacteria")
barplot(prop.table(table(bacteria$y, bacteria$trt),margin=1),
        beside=FALSE, ylim = c(0,0.8))
```

(c) Add a density line to this histogram.

```
hist(ToothGrowth$len, prob = TRUE, col = "grey", ylim = c(0, 0.05))
```

(d) Add a **dotted red** linear regression line to the following plot.

```
plot(lung$height, lung$fev1)
```

(e) Color the points in the following plot according to the `sex` variable.

```
plot(lung$height, lung$fev1)
```

(f) Add two linear regression lines separately for `female` and `male`to the following plot.

```
plot(lung$height, lung$fev1)
```

(g) Color the points in the following plot according to the `supp` variable. Use different point characters (`pch`) based on the `supp` variable.

```
plot(ToothGrowth$len, ToothGrowth$dose)
```

(h) Read `http://www.statmethods.net/advgraphs/parameters.html`.

Exercise 18:

(a) Load the below data set and for further information check the command `?water`.

```
# install.packages("HSAUR3")
library("HSAUR3")
data("water")
str(water)
head(water)
summary(water)
```

(b) Try to plot the variables `mortality` against `hardness` from the `water` data set.

(c) Add a main title to the above plot (`mortality` against `hardness`).

(d) Change the ...

   (a) font size of the axis annotation

   (b) font size of the x- and y-axis labels

   (c) the point sizes within the plot

   ... of the above plot (`mortality` against `hardness`).

(e) Looking at the above plot: Do you think the two variables `hardness` and `mortality` correlate? What function do you use to find out the correlation coefficient? Do they have a positive or a negative correlation coefficient? How do you interpret the correlation coefficient in your own words?

(f) In the `water` data set, can you graphically find out if there is a difference between the the two variables `hardness` and `mortality` conditional on the `location` (`North`, `South`).

(g) Add a legend to the above plot so that you can easily differentiate the locations (`North` or `South`) of the observations.

(h) Do a barplot of the variable `location` from the `water` data set.

(i) ADDITIONAL: Try if any of these following plotting functions can be applied to the data sets `perulung` or `ToothGrowth`.

Exercise 19: Anova in R

(a) Open the .R file `ANOVA_with_chickwts.R` from your RCourse folder and have another look on how we applied the anova to the `chickwts` data set. Check line for line.

(b) Load the `ToothGrowth` data set into R and encode the numeric variable `dose` as a factor variable. Define the new factor variable as `dose.factor` with the three levels `low`, `med` and `high` and add it to the data frame of `ToothGrowth`.

```
data(ToothGrowth)
```

(c) Visualize the variable `len` per `dose.factor` level in a boxplot.

(d) With the help of the R-commands written in the `ANOVA_with_chickwts.R` file, apply a analysis of variance (ANOVA) to the data set `ToothGrowth`

Exercise 20: Linear Regression Model I

(a) Reuse the commands from the lecture slides to fit a simple as well as a multiple linear regression model to the data set of `perulung_ems`. Use `fev1` as your response variable y.

(b) Check the model assumptions.

(c) Which model is best?

Exercise 21: Linear Regression Model II

(a) Load the `ToothGrowth` data set and run the following four linear regression models.

```
data(ToothGrowth)
ToothGrowth$dose.factor <- factor(ToothGrowth$dose, levels = c(0.5, 1.0, 2.0),
                                  labels = c("low", "med", "high"))
mod1 <- lm(len ~ dose.factor, data = ToothGrowth)
mod2 <- lm(len ~ supp, data = ToothGrowth)
mod3 <- lm(len ~ dose.factor + supp, data = ToothGrowth)
mod4 <- lm(len ~ dose.factor*supp, data = ToothGrowth)
```

(b) Have a look at the summary of these models.

(c) How do you interpret the model coefficients?

(d) Which model is best?

Exercise 22: Linear Regression Model III

(a) Load the `water` data set and fit a multiple linear regression model. Use `mortality` as your response variable and add `hardness` and `location` as an explanatory variable.

(b) Check the underlying model assumptions.

(c) Add an interaction term between `hardness` and `location` to the above estimated multiple linear regression model.

(d) Interpret the interaction coefficient `hardness:locationSouth`.

(e) Check the underlying model assumptions.

(f) Which one is the better model? With or without the interaction term?

(g) How to derive confidence intervals for the regression coefficient of `hardness` and `location`?

Exercise 23: Combining everthing we have learnt

Hypothetical example from Kirkwood and Sterne, Medical Statistics, 2nd ed., p. 177

(a) Read in the data set `lepto`. This study presents a serology survey of leptospira sero-prevalence in rural and urban areas of the west indies.

(b) Encode the numeric variable `antibodies` as a factor with levels 0 and 1.

(c) Make a crosstable with the risk factor `exposure` and `antibodies`.

(d) Run a Chi-squared test, a Fisher's exact test and a logistic regression (`glm`) to assess if the `exposure` (living in rural vs. urban areas) is a risk factor.

(e) Create a subset for `male` and `female` based on the variable `gender`.

(f) Repeat the crosstable (2-by-2 table), Chi-squared test, Fisher's exact test and a logistic regression (`glm`) for the subsets **separately**.

(g) Does the conclusion of your research question change with the analysis of the subsets? (Research question: Is the `exposure` (rural and urban areas) a risk factor?)

(h) Fit a logistic regression model (`glm`) with `exposure` and `gender` as explanatory variables.