

Practical Exercises for Exercise Collection

Sonja Hartnack, Terence Odoch & Muriel Buri

July 2019

Exercise 1: Statistical terminologies

Group the following terminology items into the three categories:

- (1) sample & variables
- (2) hypothesis testing & statistical modelling
- (3) descriptive statistics

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

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 for the chickwts data set

- (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 and summary statistics 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).

```
summary(lung)  
lung$sex <- factor(lung$sex, levels = c("0", "1"))  
levels(lung$sex) <- c("female", "male")  
lung$respsymptoms <- factor(lung$respsymptoms, levels = c("0", "1"))  
# Continuous and factor  
tapply(lung$height, lung$sex, mean)  
tapply(lung$height, lung$respsymptoms, mean)  
# Factor and factor  
table(lung$respsymptoms, lung$sex)
```

```
prop.table(table(lung$respsymptoms, lung$sex), 2)
# Continuous and factor
tapply(lung$age, lung$sex, mean)
tapply(lung$age, lung$respsymptoms, mean)
# Continuous and factor
tapply(lung$fev1, lung$sex, mean)
tapply(lung$fev1, lung$respsymptoms, mean)
```

(d) Plot a boxplot.

Exercise 5: Defining a new data frame

(a) Create a data frame with 3 columns.

Exercise 6: Different bracket types within R

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

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

Exercise 7: Data type of perulung_ems data set

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

Exercise 8: Get to know bacteria data set

- (a) Install package MASS. Load data set bacteria.
- (b) Describe in your own words what the data set bacteria contains.
- (c) Do summary statistic (numerically and graphically).
- (d) Select only observations collected during the second week.
- (e) How many levels has the factor variable trt from bacteria?
- (f) 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.
- (g) 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: 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 11: 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 12: 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 13: 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 14: 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 15: 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 `perlung_ems`. Use `fev1` as your response variable y .
- (b) Check the model assumptions.
- (c) Which model is best?

Exercise 16: 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 17: 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 18: Combining everything 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.