Exercises for introduction to R, day 2

Exercise 1

At <http://www.folk.uib.no/nzlkj/IMRkurs> you will find a spreadsheet file called rats.xls. This dataset contains body mass measures (g) of female and male individuals of *Rattus norvegicus*.

a) Import

Import the dataset to R via text file. Remember that you must first create a text file of your data and then do the R syntax needed to import it. Remember to write all your R syntax in a text editor and copy it over to the commands window of R.

Tip! If you have trouble with importing the data from a text file, it may help to have a look at the text file you have created by opening it in a text editor (not a Word processor!). By doing so, you will see what is used as separator between variables. Depending on the format of the text file it could be comma (`sep=","`), tabulator (`sep="\t"`), semicolon (`sep=";"`) etc. You will also see what is the decimal symbol. Normally that is either period (`dec="."`), or comma (`dec=","`) . The separator between variables and decimal symbol will never be the same. For a Norwegian computer setup a csv-file will probably be formatted with semicolon as separator and comma as decimal symbol (`sep=";", dec=","`), while the English standard for a csv-file is comma as separator and period as decimal symbol (`sep=",", dec="."`)

After importing the data, have a look at them to see that the import went well, e.g. by using the `str` function on the dataset. For example, assuming that you called your dataset for rats.df:

```
str(rats.df)
```

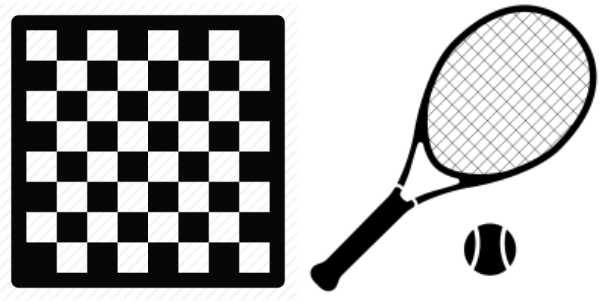
b) Plot

Plot the relationship between response and predictor. Use a boxplot since the predictor is categorical (use the geom_boxplot function in ggplot).

c) Analysis

Test the following hypothesis by using R: H0: The body mass in *Rattus norvegicus* is the same for males and females.

Exercise 2



At <http://www.folk.uib.no/nzlkj/IMRkurs> you will find a spreadsheet file called "solving.time.xls". This file contains data on time (s) spent solving a mathematical problem for professional chess, tennis and squash players.

a) Import

Create a text file of the dataset and import this text file data into R. After importing the data, use the `str` function to check that everything looks OK with the data.

b) Plot

Make a plot of time spent solving the mathematical problem depending on type of player. Use a boxplot since the predictor is categorical (the geom_boxplot function in ggplot).

c) Analysis

Test the following hypothesis by using R: H0: The time spent solving a mathematical problem is the same for professional chess, squash and tennis players. In case your main test showed an overall effect of type of player - which groups (type of player) differ from each other? Use an unplanned multiple comparisons test (library multcomp) for this latter task.

d) Subsetting followed by new plot and analysis

You were just about to submit the results from b) and c) to a psychology journal when the squash players said that they have cheated on the test and used their cellphones. Thus, the results from the squash players is not to be trusted. You therefore decide to do the whole analysis again but just with the chess and tennis players. Make a subset of the data (from within R) and do the plot again. Do also find a suitable test for this reduced dataset for testing the following hypothesis: H0: The time spent solving a mathematical problem is the same for professional chess and tennis players.

Exercise 3 (Yes, this exercise is the same as yesterday except for the statistical analysis)

At <http://www.folk.uib.no/nzlkj/IMRkurs> you'll find a spreadsheet file called “TCB.xls”. It contains sample results for the amount of TCB (thermotolerant coliform bacteria), measured in CFU (coliform forming units) from two locations in Møllendal river; one above and one below a pipeline with water running out of it.

a) Import

Create a text file of the dataset and import it into R. After importing the data, use the `str` function to check that everything looks OK with the data.

b) Plot

Make a plot of the data.

c) Analysis

Test the following hypothesis by using R: H0: The amount of TCB in Møllendal river is the same above and below pipeline A.

Exercise 4

At <http://www.folk.uib.no/nzlkj/IMRkurs> you will find a spreadsheet file called "occurrence.xls". It contains data on occurrence of blue tit (*Cyanistes caeruleus*) nests depending on summer mean temperature.

a) Import

Create a text file of the dataset and import it into R. After importing the data, use the `str` function to check that everything looks OK with the data.

b) Plot

Make a plot of the data. Since the predictor represent continuous data, it is reasonable to think about using the geom_point function in ggplot. However, since this is binary data, it is probably better to replace geom_point with geom_jitter. This function adds some noise to the data so that the placement of points is not exactly correct. In our case, it will be useful to add some noise to the vertical placement of points. This will make it easier to see the density of points along the range of temperatures. Thus, you may use the following to get a good illustration of the data: `geom_jitter(height=0.01)`

To make good illustrations of binary data is not always easy. The figure you have created will look better if you add a line that illustrate the probability of occurrence depending on temperature. This can be done by the following syntax:

```
geom_smooth(method = "glm", method.args = list(family = "binomial"), formula = y~x)
```

c) Analysis

Test the following hypothesis by using R: H0: The occurrence of blue tit nests does not depend on summer mean temperature.

Exercise 5 (difficult for a beginner, but look at the lecture notes before you evt look in the answers or ask the lecturer)

You have found out that you make a lot of figures of the type you created for the rats data in Exercise 1. You therefore decide that you want to make a function for this type of boxplot, i.e. boxplot for a single categorical predictor. Make such a function and try if it works by using it on the rats dataset. Let the function contain the following three arguments; dataset, response, and predictor).