Series 1

1. The goal of this exercise is to get acquainted with different abilities of the R statistical software. It is recommended to use the distributed R tutorial as a guide.

R contains more than 50 datasets and more can be loaded using optional packages. The package `VR` is depending on the package `MASS` which contains the dataset `survey`. This dataset comprises of measurements and answers taken from 237 students of statistics at the University of Adelaide. The following variables are available

Sex	gender of student
Wr.Hnd	span width in cm (from thumb to pinky) of the writing hand
NW.Hnd	span width in cm (from thumb to pinky) of the non-writing hand
W.Hnd	writing hand
Fold	When folding your arms - which one is on top?
Pulse	beats per minute
Clap	When clapping your hands - which one is on top?
Exer	How often do you exercise?
Smoke	How often do you smoke?
Height	body length in cm
M.I	Preference of either metric (cm/m) or imperial (feet/inches) units?
Age	age in years

```
> library(MASS)   # makes the datasets of the MASS package available,
                  PC: Install first the package VR
> data()          # shows a list of all available datasets
> help(survey)    # gives a description of the dataset survey
> data(survey)    # makes the dataset survey available
```

Useful functions to get a first overview of the dataset:
str(survey), **summary(survey)**.
The notation `survey$Smoke` accesses the variable `Smoke` in the dataset `survey`.

Some univariate descriptive techniques:

```
> table(survey$Smoke)      # table
> hist(survey$Height)      # histogram
> boxplot(survey$Height)   # boxplot
```

Dealing with missing values (`NA`):

```
> mean(survey$Pulse)                # result is NA
> mean(survey$Pulse, na.rm=T)       # the missing values are removed before calculating the mean
```

Also check the functions **na.omit()**, **is.na()**.

Some useful fuctions for bivariate graphics:

```
> boxplot(split(survey$Height, survey$Sex))   # boxplots of two variables
> plot(survey$Wr.Hnd,survey$NW.Hnd)           # scatter plot
> mosaicplot(~Sex+Smoke,data=survey)          # mosaic plot
> plot(survey$Sex,survey$Height)              # ?
```

Selecting observations:

```
> plot(survey[1:50,2],survey[1:50,3]) # scatter plot based on only the first 50 observation
> boxplot(survey$Height[survey$Sex=="Female"],survey$Height[survey$Sex=="Male"])
# boxplots for males and females separately
```

Do not forget about the online help:

```
> help(survey)
> help(plot)
```

...

Now analyse the dataset `survey` using descriptive methods. Answer the following questions:

a) Do the two oldest students smoke?

b) Which factors might have an influence on the student's pulse? **Hint:** Scatter plots.

c) Is the span width of the writing hand in general larger than the span width of the non-writing hand? **Hint:** Boxplot.

d) It is generally believed that the pulse of an individual decreases with an increase in age. We investigate this using the following code:

```
> Agejung <- survey$Age[survey$Age<30 & !is.na(survey$Pulse)]
> Pulsejung <- survey$Pulse[survey$Age<30 & !is.na(survey$Pulse)]
> cor(Agejung,Pulsejung)
> scatter.smooth(Agejung,Pulsejung,col='red',cex=0.5)
> lmobj <- lm(Pulsejung ~ Agejung); plot(Agejung,Pulsejung); abline(lmobj)
```

Comment on the output. What does the above code do?

Preliminary discussion: Monday, September 29.

Deadline: —.