Introduction to R **University of Bremen**
FB03 Mathematics and Computer Science Eike Voß, evoss@uni-bremen.de

Exercise 03

Programming tasks should be written in a `.R`-script. Please include your lastname(s) in the file name (e.g. `Ex03_Voss.R`) and submit it via **DoIT!** in stud.ip, where you also find the submission deadline.
To receive sample solutions for the Presence Tasks throughout the course, you need to submit a solution in **DoIT!**.

Presence task 1:

Roll a die 1000 times and store the results in the vector `x`. Store the results less than or equal to 5 in the vector `y` and all sixes in `z` and calculate how many times a six is thrown. Use `sample(1:6,1000, replace = T)` to simulate the 1000 die throws.

Presence task 2:

Read in the data set `airquality` from the `datasets` package (probably already installed **and** loaded) and read the documentary of this data set.

a. Create a new data set `air1` with all rows where `Wind` is less than 10.

b. Create the data set `air2` with `Temp` greater than 70 (F).

c. Create the data set `air3` that contains only the data for the variables/columns `Month` and `Ozone` where `Wind` is less than 10 and `Temp` is greater than 70 (F).

Presence task 3:

Load the `airquality` dataset. We want to create three different sub-datasets (all from `airquality`):

a. Create the dataset `air4` without missing values (see `?na.omit`).

b. Create the dataset `air5` without missing values in the 1st column (`!is.na()`).

c. Create the dataset `air6` without missing values in the 1st and 2nd columns.

d. Add a new column `TempC` to `air3` which contains the temperature in Celsius. It is $°C = \frac{5}{9}(°F - 32)$.

e. In `air3`, when (month and day) was the wind (or temperature, ozone concentration, solar radiation) the strongest?

Homework task 1 (9 Points):

Let the height (in cm) of a population be normally distributed with $\mu = 170$ and $\sigma = 10$. Set the seed `1528` and use the following command to generate a sample of size $n = 2000$ from this population: `pop1<-rnorm(2000,170,10)`.

a) Round the height to integers by using `round()` (and overwrite the values in `pop1`).

b) Execute `which(pop1 > 190)` and interpret the result.

c) What subgroup is created by running `pop1[pop1 > 190]`?

d) Create two subgroups: 1) `height < 155` and 2) `height > 185`.

e) Interpret the results:
`rev(pop1); unique(pop1); duplicated(pop1); pop1[duplicated(pop1)]; pop1[!duplicated(pop1)]`.

f) Simulate another sample `pop2` of size $n = 1000$ from a population whose height is $\mathrm{N}(\mu = 172, \sigma = 12)$-distributed.

g) Round the height in the second population to integers.

h) **Randomly** assign 1000 individuals in sample 1 to those in sample 2 by defining a matrix `X` with two columns, one containing the individuals from sample 1 and one containing the individuals from sample 2 that the individuals from sample 1 were assigned to. *Hint: Use `sample()` for the randomization and `cbind()` for the assignment.*

i) We consider each row of the matrix `X` as a pair, where individuals from sample 1 and 2 are seen as height values by men and women, respectively. How many pairs are there where the man is shorter than the woman?

Homework task 2 (4+2 (+2) Points)

a) Write a function `getOdd` that returns all the odd numbers in a `numeric`-vector `x`. This function should first check whether the vector is indeed a vector and of type `numeric`. If `x` does not meet these two requirements return the string `"x has to be a numeric vector!"`, otherwise convert the numeric vector to an integer and return a vector of all odd numbers that occur in `x` without duplicates (see `unique()` from task 1). Note: An integer (a whole number) $z \in \mathbb{Z}$ is odd if z `modulo` $2 = 1$ (so if $z/2$ has a remainder $= 1$).
You get 2 extra points if you manage to write this function without a loop.

b) Generate a random sample of 50 numbers from the set $\{-50, -49, \ldots, 0, \ldots, 49, 50\}$ using `sample()` (without replacement). Check whether your sample contains any negative values. Also, apply the function from a) to your sample and from the resulting vector extract all numbers that are divisible by 3 (all z with z `mod` $3 = 0$).

Homework task 3 (2+3P)

The variable `Wind` in `airquality` gives the wind speed in mph (mile per hour). This is a metric-scaled variable. Form a new variable (ordinal), `Wind`, from this variable:

$$\texttt{Wind2} = \begin{cases} \texttt{weak}, & \text{if } \texttt{Wind} \leq 12 \\ \texttt{strong}, & \text{if } \texttt{Wind} > 12 \end{cases}$$

and add this variable to the `airquality` data set.
Now define the ordinal-scaled feature `Wind3` like this:

$$\texttt{Wind3} = \begin{cases} \texttt{weak}, & \text{if } \texttt{Wind} \leq 4 \\ \texttt{mild}, & \text{if } 4 < \texttt{Wind} \leq 18 \\ \texttt{strong}, & \text{else.} \end{cases}$$