Exercises for the course **"An introduction to R"** Sheet 02

- What is the first, the seventeenth and the last entry of the vector data?
- What is the maximum of the vector data? At which index is the maximum attained?
- Plot the vector data with plot(data) and visually confirm your last result.
- At which indices are the entries of data between 2000 and 2500?
- Define a new vector **p** by **p** <- data/sum(data). Calculate

$$\sum_{i=1}^{100} \mathbf{p}[i], \quad m < - \sum_{i=1}^{100} i \cdot \mathbf{p}[i], \quad s < - \sum_{i=1}^{100} i^2 \cdot \mathbf{p}[i], \quad \text{and} \quad \sqrt{s - m^2}$$

Exercise 8: Create the following matrices:

$$\begin{pmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 2 & 3 & 4 \\ 7 & 8 & 9 \end{pmatrix}$$
$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 4 & 9 & 16 \\ 1 & 4 & 9 & 16 \\ 1 & 4 & 9 & 16 \\ 1 & 4 & 9 & 16 \end{pmatrix}$$

Define a new matrix **m** by

m <- matrix(11:35, nrow=5, byrow=TRUE)</pre>

What is the entry in the third row and forth column? Briefly describe in words what

m[2:4,3:5]

returns. Calculate the matrix product of m with itself.

Exercise 9: The most important distribution is the normal distribution. So let's study it.

- Plot the density of the (standard) normal distribution between -3 and 3.
- In statistics, null hypotheses are rejected if a certain probability is below 5%. Thus it is important to know which centered interval supports 95% of the mass of the normal distribution. Find the value x such that pnorm(x)-pnorm(-x)=0.95. Recall the symmetry of the (standard) normal density and explain why this value x is equal to qnorm(0.975). Round the value x to 3 significant digits and remember this number.
- Sample 1000 random values from the standard normal distribution and denote the vector of these values as \mathbf{x} . Calculate the mean, the variance, the standard deviation and the quartiles of this vector. Then visualize the quartiles of \mathbf{x} with a boxplot. Finally plot the histogram and the empirical distribution function of \mathbf{x} .

Exercise 10: Data frames are the typical R representation of data sets. Here we create a data frame "by hand" to become familiar with data frames.

Use the command data.frame() to create a data frame students with the following entries

name	degree	mat.nr	grade
Leonie	Master	1111	2.3
Luka	Master	1112	3.0
Leon	Bachelor	1114	2.0
Lea	Bachelor	1113	1.3
Luis	Master	1116	2.7
Laura	Master	1115	1.0

- Get an overview of students with the commands names(), str() and summary().
- Which command returns the fifth element of the vector 'mat.nr'?
- Check existence of the variable degree by entering it into the R command line. Now copy students into the search path with the command attach(). Check again whether degree is a known variable.
- Calculate the average grade of all students with a master degree
- Define a new data frame named 'ma.students' which consists of all students with degree Master (without using the command data.frame()). As all students in ma.students have degree Master the variable degree is not needed in ma.students.
- Write the data frame students into the file 'studentsfile.txt'. Then read the data frame from this file into the new variable students2. If you used the right command, then students and students2 are identical. Check this using the command all().
- Lea and Leon just received their Masters degree. Use fix() or edit() to update students accordingly. Note that these two commands do not work on all systems.
- We wish to change 'degree' into 'deg' to save typing work. Use the command names() to accomplish this change. You might need to consult the help page ?names to find out how to do this.

Exercise 11: Reading and writing data

Download the files 'olympic.txt', 'olympic0.dat', 'olympic1.txt', 'olympic2.txt' and 'olympic3.csv' from the course web page.

- Read the file 'olympic.txt' into a data frame.
- Read the file 'olympic0.dat' into a data frame. In that data frame, replace the first column by a column containing the respective years (integers between 1896 and 1992) and denote this column as "year". Then write this modified data frame to the file 'olympic1new.txt'.
- Read the file 'olympic1.txt' into a data frame. Produce a file 'olympicHighJump.txt' consisting only of the columns "Since1900" and "HighJump".
- Read the file 'olympic2.txt' into a data frame. Produce a file 'olympic2new.txt' containing a header with appropriate names (of your choice).
- Read the file 'olympic3.csv' into a data frame. Using this data frame, caculate the mean value of all long jump records between 1896 and 1972.

Exercise 12: Let **p** be the vector of the weights on 0, 1, ..., 100 of a binomial distribution with 100 trials and success probability $\frac{1}{2}$.

- Use the command dbinom() to define p. Plot the vector p.
- Recall from the script the true mean and variance of this binomial distribution. Sample 100 random values from this binomial distribution and denote the resulting vector as y. Calculate the mean and the variance of y and compare these values with the respective true quantities.
- What is the correlation between y and -2*y+4? What is the correlation between y and 3*y?
- Sample another 500 random values and another 10000 random variables from the binomial distribution at hand (100 trials and success probability $\frac{1}{2}$) and denote the resulting vectors as z1 and z2, respectively. One method to compare y, z1 and z2 is to compare the boxplots. This is done with the command boxplot(y,z1,z2). Briefly describe in words the differences and similarities of the three boxplots (you might want to find out whether your observations are generally true by repeating the sampling of y, z1 and z2)