Frankfurt University of Applied Sciences
Prof. Dr. Christina Andersson

R Exercises

1 Introduction

Make yourself comfortable with the handling of R!

2 Manipulation of Vectors and Numbers

1. (a) Construct a vector x with the elements 8, 9, 7 and 5.
   (b) Construct a vector y with the elements 1, 2, 3 and 4.
   (c) Add the two vectors. Put the result of the addition in the vector s.
   (d) Display the second element of s.
2. (a) Construct a 4×2 matrix with the elements of the first row 1, 2, of the second row 3, 3, of the third row 5, 1 and of the fourth row 4, 2.
   (b) Display the matrix.
   (c) Display the second column.
   (d) Display the second element in the third row.
3. (a) Construct a vector with the entries 5, 6, 7.
   (b) Return the largest element of the vector.
   (c) Reverse the elements iof the vector.
4. (a) Construct a sequence containing all integers from 6 to 76.
   (b) Draw five numbers randomly from the generated sequence. Do this without replacement.
   (c) Draw two numbers randomly from the generated sequence. Do this with replacement.
5. (a) Construct a vector d, containing the integers from 0 to 99.
   (b) Construct a vector e, containing 100 numbers randomly drawn from a normal distribution with expectation 3 and standard deviation 4. Use the function *rnorm*.
   (c) Add the vectors d and e and call the resulting vector f.
6. Create a *data.frame* called *test*, consistng of the two variables

```
x = c(2,3,4)
y = c(5,6,7)
```

7. We are going to use the dataset *airquality* in the package *datasets*.
   (a) Use R commands to determine the number of rows and the number of columns that we have in the dataset.
   (b) Use the function *head* to get an impression from the first six lines of the dataset.

3 Tables and Graphs

1. We investigate in which trial the students in a class passed the statistics exam and these are the results:

```
exam = c(0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,1,0,0,2,0,
0,0,0,0,0,0,1,0,0,0,1,0,0,0,0,0,1,0,0,1,0,0,3,1,1,2,0,0,1,0,0,0,0,2,0,1,0,1,
0,1,0,0,0,0,1,0,0,1,0,0,1,1,0,1,0,0,1,0,0,0,0,0,1,0,1,1,1,2,0,0,1,0,2,1,1,0,
0,0,0,0,1,0,1,1,0,0,0,2,1,1,1,0,2,0,1,1,2,1,1,0,1,0)
```

(a) Construct a table with absolute frequencies.

(b) Construct a table with relative frequencies.

(c) Use a bar plot to illustrate the data graphically.

(d) Illustrate the cumulative distribution function graphically.

2. Peter sells umbrellas in his shop. The following dataset shows the number of umbrellas that he sold, together with the number of days he sold a specific number of umbrellas:

```
umbrella = data.frame(no_of_umbrellas=c(0,3,4,5,6,8),days=c(4,6,8,3,1,2))
```

(a) Construct a table with absolute frequencies. Compute the cumulative frequencies.

(b) Construct a table with relative frequencies. Compute the cumulative frequencies.

(c) Illustrate the cumulative distribution function graphically.

(d) Determine and interpret the value of the empirical distribution function at the position 5.

3. From the country CCClland, we have got data about the number of boot accidents during certain years.

```
boots = data.frame(Year=1985:2008,No_of_Accidents=c(472787,495871,532220,523387,
499666,513587,487654,478721,521972,476544,430432,452165,432589,456436,457422,466064,
519482,343091,328221,522169,415077,387212,415254,423731))
```

(a) Construct a bar plot!

(b) Why is this not an appropriate diagram for this dataset?

(c) Construct a histogram with 9 classes and constant class width. To which class do most of the years belong?

4. We have the following pairwise observations:

```
x=c(1,2,2,3,3,4,4,5,6,6,7,7,8,9,8,9,9)
y =c(1.426865, 2.495512, 2.751945, 3.794935, 3.682121, 3.692246,
4.451148, 5.200307, 5.638318, 7.672076, 6.819001, 7.208195 9.076866,
9.441328, 7.752522, 9.545205, 10.097847)
```

Create a scatter plot!

5. A teacher wants to compare his two classes with regard to the scores in the last maths test. These are the results:

```
group1 = c(12, 23, 34, 33, 35, 33, 32, 31, 30, 29, 28, 28, 27, 27, 6, 26)
group2 = c(13, 13, 24, 30, 31, 31, 30, 30, 31, 28, 28, 29, 26, 25, 26, 26)
```

Create boxplots and use them to compare the groups!

6. (a) Read the following data into R and store the observations in the three variables *Gr1, Gr2* und *Gr3*:

Gr1	Gr2	Gr3
22	21	48
22	23	45
28	21	51
23	24	29
45	22	38
21	20	40
22	21	38

(b) Construct a boxplot for each of the three variables. Discuss measures of central tendency and spread. What about symmetry?

7. The package *datasets* contains a lot of different data sets. We are going to use the datasets *beaver1* and *beaver2*, which describes the body tempeature of two beavers over time.

   (a) Load the package *datasets*.

   (b) Compare the body temperature of the two beavers using boxplots.

8. We use the dataset *iris* in the package *MASS*.

   (a) Load the package *datasets*.

   (b) What is the maximum of the variable *Sepal.Width* in the dataset *iris*?

   (c) Create a scatter plot of the variables *Sepal.Width* and *Sepal.Length*.

   (d) Create a scatter plot of the variables *Sepal.Width* and *Sepal.Length* again. This time you shall use different colors of the dots for the different values of the variable *Species*.

4 Measures of Central Tendency

1. The data for the number of strike days in a specific year are given for four fictive countries:

```
strike= data.frame(Country=c("Aland", "Bland", "Cland", "Dland"),
Days=c(77, 45, 76, 83))
```

   (a) Calculate the median.

   (b) Calculate the arithmetic mean.

   (c) Interpret the results.

2. We have received a data set with some data about a group of students:

```
students=data.frame(name=c("Anton", "Kim", "Harald", "Inga",
"Mona", "Sigrid"), height=c(170,167,169,172,171,170),
weight=c(70,75,120,87,88,75))
```

   (a) Calculate the arithmetic mean and the median for the variable weight.

   (b) Exclude the student Harald from the data set. Calculate again the arithmetic mean and the median for the variable weight, now for the remaining data.

   (c) Interpretation?

3. These are the grades for some pupils in a school class:

```
grades<-c(2,3,3,3,4,1,5,2,4,2,2,2,3,4,4,3,2,1,3,3,3,2,2)
```

(a) Construct a table with absolute frequencies.

(b) Calculate the arithmetic mean, the median and the mode.

4. Mr. Cucumber has a shop selling fruit and vegetables called *Green and Good*. He is thinking about how the new kind of pumpkin sells in the shop. During 30 days, he observes the sales pattern of the new pumpkin. These are his observations:

```
pumpkins=data.frame(no_of_pumpkins=c(0,1,2,3,4,5,6,7),
days=c(10,2,3,5,4,2,4,5))
```

   Determine and calculate the arithmetic mean of the number of sold pumpkins.

5. The package *datasets* contains a lot of different data sets. The data set presents daily air quality measurements in New York, May to September 1973.

   (a) Load the package *datasets*.

   (b) Calculate the arithmetic mean of the temperature.

   (c) Calcullate the median of the solar radiation.

   (d) In the previous task, you encountered a problem in the calculation of the median. Which problem? Recalculate the median of the solar radiation, taking this problem into account.

6. In a small company, the number of server problems during different days was recorded. During the observation period, it turned out that 90% of the days, there was no server problem, 4% of the days there was one server problem and 6% of the days, there were two server problems. Calculate and interpret the arithmetic mean for these data.

7. The students Klara and Simon have studied the flat advertisments during three monzs in X-City. Here is the resultng table:

No of rooms:	1	2	3	4	5	6	7
No of flats:	56	55	35	22	12	6	2

   How many rooms did a flat have on average?

8. Calculate the geometric mean of the following dataset:

```
mydata=data.frame(obs_values=c(50,45,25,20),frequency=c(2,4,2,2))
```

9. Calculate the geometric mean of the following dataset:
   5,7,8,9,9

10. Calculate the harmonic mean of the following observations:
    5,7,8,9,9

11. Calculate the quartiles of the following observations:
    5,8,9,9,9,4,5,6,6,76,43,56,65,65,3,34,45

12. A lecturer wants to give a scholarship to the best 4% of the class. These are the scores in the exam, settling the possibility to get the scholarship:
    43,12,11,22,23,34,34,33,34,23,33,32,11,9, 45,56,48,23,23,43,23,21,21,45,23,22,32,32,21,43,11,47

    Which scores did those students, who get a scholarship, reach?

5 Measures of Spread

1. A student participates in five tests. These are the scores he reaches in the tests: 56, 87, 88, 91 and 66

   (a) Calculate the standard deviation for the scores.

   (b) Calculate the variance for the scores.

   (c) Calculate the range for the scores.

2. The package *datasets* contains a lot of different data sets. We are going to use the dataset *beavers*, which describes the body tempeature of two beavers over time.

   (a) Load the package *datasets*.

   (b) Compare the body temperature of the two beavers concerning spread.

   (c) Interpret the results.

3. The student Greta reaches the following scores in five tests: 17, 23, 33, 24 and 78. Compare the range and the interquartile range for the data set with and without the outlier. Interpret the results.

6 Correlation

1. These data show the prices for food in two different supermarkets. The first item in the vector for supermarket A corresponds to the first item in the vector for supermarket B and so on.

   A=c(1, 5, 6.6, 4, 10, 12, 13, 21, 1, 3)

   and

   B=c(1.2, 57, 4.6, 4.1, 10.4, 12.9, 11.9, 22, 1.4, 4)

   (a) Calculate the correlation of the prices in the two supermarkets. Interpret the correlation.

   (b) Construct a scatter plot for the data.

   (c) Does the scatter plot confirm the result in a)?

   (d) Remove the outlier and perform the tasks a) - c) again!

2. The following pairwise observations of the variables x and y are given:

x	1	2	3	5	10
y	4	4	7	11	11

   (a) Construct a scatter plot for x and y. Interpret the scatter plot.

   (b) Calculate and interpret the Pearson's correlation coefficient and the rank correclation coefficient!

   (c) Explain why the rank 1.5 exists!

3. The following pairwise observations tell the ranks eight athlets reached in the sports W and V during a certain international competition.

W	1	2	3	3	6	1	3	8
V	4	4	5	3	8	1	5	6

Calculate an appropriate correlation measure for the data. Conclusions?

4. Following observations are given:

```
d=c(3,8,4,4,2,2,4,3,5,6,7,8)
e=c(1,3,4,4,5,6,4,3,2,8,1,9)
```

   (a) Calculate the rang correlation of d and e. Interpret the result.
   (b) Calculate the rang correlation of d and e. Interpret the result.
   (c) Compare the results in a) and b).
   (d) When is usually the correlation calculated as in a) and when as in b)?

7 Regression

1. Following observations are given::

```
d=c(1,1,1,2,2,2,3,3,5,6,7,8)
e=c(2,3,4,4,5,6,6,7,8,8,8,9)
```

   (a) Determine the estimated line in a regression analysis with d as independent variable and e as dependent variable.
   (b) Interpret the estimated regression coefficients!
   (c) Calculate the correlation between d and e!
   (d) Interpret the correlation coefficient!
   (e) Calculate the covariance between d and e!
   (f) Interpret the covariance!

2. In a regression analysis, we use e as dependent variable and j as independent variable. The following data is available:

e	5	2	3	4	2	1	5	4	7	5	8	9	8	8
j	5	6	3	4	1	1	1	6	7	8	8	7	8	9

   (a) Perform the regression analysis and evaluate the model! Motivate your evaluation of the model carefully, i.e. using at least one graph.
   (b) Interpret the estimated regression line, d.h. explain the meaning of the coefficient coefficients!
   (c) How can you use the results of the regression analysis to find the correlation between e and j?

8 Probability Distributions

1. Let X be Bin(100, 0.4)-distributed. Calculate

   (a) $P(X \leq 35)$
   (b) $P(X > 39)$
   (c) $P(36 \leq X \leq 38)$

2. Let X be N(mean=10, sd=4)-distributed. Calculate

   (a) $P(X \leq 11)$

   (b) $P(X > 13)$

   (c) $P(10 \leq X \leq 12)$

3. Which x-value corresponds to the probability 0.50 of a normal distribution with expectation 0 and standard deviation 10?

4. Draw 10 numbers randomly from an exponential distribution with parameter 5!

5. What is the 40th quantile of the Bin(100, 0.3)-distribution?

6. Suppose screws produced at the company ACCD have weights that are normally distributed with mean 35.42 grams and variance 16 grams. What is the probability that a randomly chosen screw weighs more then 36 grams?

9 Hypothesis Tests

1. Suppose the mean weight of a certain pumpkin we grew last year was 15.4 kg. In a sample of 35 pumpkins this year we got the mean weight 14.6 kg. Historical data allows to consider the standard deviation t be known to be 2.5 kg. We want to perform a hypothesis test at significance level 5%.
   Can we reject the null hypothesis that tells that the mean pumpkin weight does not differ from the value 15.4 kg?

2. Instead of computing the critical value in the previous problem, apply the *pnorm* function to compute the two-tailed p-value of the test statistic.

3. We assume that a company claims that a certain electronical component, which they use, has a mean lifetime that is more than 10,000 hours. We collect a sample of 30 components. The sample mean for the life time is only 9,900 hours! The sample standard deviation is 125 hours. At the significance level 5%, can we reject what this company claims?

4. Instead of computing the critical value in the previous problem, apply the *pnorm* function to compute the p-value of the test statistic.

5. (a) Load the package *datasets*.

   (b) What is the sample mean temperature of the beaver in the dataset *beaver1*?

   (c) Use the function *t.test* to test if the population mean temperature of the beaver in the dataset *beaver1* equals 37.

6. We want to study if *gender* and *chocolate behavior* (i.e. if the person eats a lot of chocolate or not) are two variables that are independent of each other. The data show that there were 208 females that eat a lot of chocolate and 230 females that don't eat a lot of chocolate. For the males, the corresponding numbers are 282 and 241, respectively. Test if the two variables are independent. Use significance kevel 5% in the test.

7. We are going to use the dataset *Aids2* in the package *MASS*. Load the package *MASS* and test if the variables *state* and *status* are independent.

10 Confidence Intervals

1. (a) Load the package *datasets*.
   (b) What is the sample mean temperature of the beaver in the dataset *beaver1*?
   (c) Use the function *t.test* to construct a confidence for the population mean temperature of the beaver in the dataset *beaver1*. Use confidence level 99%