FRA-UAS
Fachbereich 2, HIS
Introductory Data Analysis
Prof. Dr. Christina Andersson

Exercise Sheet 1: Descriptive Statistics 1

Objective:
The aim of this exercise sheet is that you shall get some basic knowledge of descriptive statistics.

Theoretical Problems:

1. In a passport, you can find among others the following data about the passport holder:
   Name, residence, height, colour of eyes, date of birth and nationality.

   Describe the scales of these variables!

2. Are the following variables quantitative or qualitative?

   (a) Hair colour.

   (b) Number of children in a family.

   (c) Outdoor temperature.

   (d) Age.

   (e) Weight of a newborn child.

3. Are the following quantitative variables discrete or continuous?

   (a) Number of rooms in a flat.

   (b) Number of children in a family.

   (c) Outdoor temperature.

   (d) Age (in years).

   (e) Weight of a newborn child.

4. We have seven observations:
   1,1,1,4,3,5,4

   (a) Calculate the mean!

   (b) Calculate median!

   (c) Determine the mode!

5. The following observations are given:

   4, 1, 3, 5, 1.

   Calculate

   (a) the arithmetic mean
   (b) the harmonic mean
   (c) the geometric mean

6. The following observations are given:

   2, 8, 5, 3, 8 Calculate

   (a) the 50%-quantile
   (b) the 20%-quantile
   (c) the second quartile

7. In a study concerning petrol consumption, the statistician first becomes the data for 8 test cars. The mean and the standard deviation for these 8 cars were: 6,3 l/100 km and 0,04 l/100 km, respectively. But before these results were published, the results for two additional cars were obtained: 6,1 l/100 km und 6,6 l/100 km.
   Calculate the mean and standard deviation for all 10 cars!

8. We have ten observations with $\bar{x} = 5$ and $\sum_{i=1}^{n} x_i^2 = 350$.

   (a) Calculate the variance!
   (b) Calculate the standard deviation!

R Problems:

1. The following observations are given:
   1,2,3,3,2,45,45,56,67,55,67,56,67,68,55,54,43,32,22,33

   (a) Calculate the arithmethic mean!
   (b) Calculate median!
   (c) Determine the mode!

2. The following observations are given:

   2, 9, 15, 13, 82, 65

   Calculate

   (a) the 50%-quantile,

(b) the 22%-quantile,

(c) the third quartile.

3. The following data show the number of citizens in some German villages and cities in a certain region:

   citizens=c(264, 9338, 445, 475, 5993, 21752, 10728, 537, 7724, 25121, 24923, 19954, 6725, 9363, 17273, 317, 26848, 2213, 5015, 64120, 14127, 2909, 2316, 22774, 25216, 20681, 418, 15786, 25109, 57797, 37194, 450, 8713, 1278, 3327, 2187, 10547, 5960, 5580, 7650, 4024, 31029, 7165, 1409, 8311, 16886, 21132, 19568, 12145, 22476, 1932, 6833, 1002, 3894, 4229, 22084, 6741, 22503, 40480, 6245, 1066, 614, 4185, 13516, 10017, 3033, 2967, 7096, 2727, 11208, 26253, 10666, 23908, 13270, 5817, 2475, 5260, 2996, 12065, 371, 9439, 10425, 5685, 21869, 11580, 7726, 4808, 9482, 8365, 3116, 14974, 6420, 4869, 55583, 2995, 3617, 37414, 25146, 7173, 9817)

   (a) How many cities participated in the study?

   (b) Create a histogram of the number of citizens. What can you say about the skewness of the data? (symmetric, left-skewed, right-skewed)

   (c) Create a histogram with **about** 20 bins of the number of citizens (use *breaks=* in the *hist*-command). What can you now say about the skewness of the data? (symmetric, left-skewed, right-skewed)

4. To find a flat in Frankfurt is sometimes rather difficult - most of you know that from your own experience. Let us assume that the following data show the number of rooms in free flats announced one day in a Frankfurt newspaper:

   number_of_rooms = c(1,2,2,2,2,1,2,3,6,3,1,2,1,3,5,4,1,4,5,2,1,1,2,1,2,5, 1,2,1,2,1,2,1, 3,1,4,2,4,5,4,6,4,2,5,5,4,3,2,3,4,2,3,2,3,2,3,2,4,3,2,3,3,2,8, 2, 2,1,3,4,1,2,3,2,3,2,2,3,4,3,3,3,3,1,1)

   (a) Use the function *table* to summarize the data. Which it the most common number of rooms? How many flats have this number of rooms?

   (b) Use the function *barplot* to illustrate the data.

   (c) Calculate the relative frequencies (= frequencies expressed as percent) of the number of rooms.

   (d) Use a pie chart to illustrate the relative frequencies of the number of rooms. Function: *pie*

5. The following hospital data contains the following variables:

```
age=c(18,19,21,28,23,29,33,31,31,30,39,44,42,22,35,21,23,45,8,45,13,32,31)
gender=c(1,1,1,2,2,1,2,2,1,2,1,1,1,1,1,2,1,2,2,2,1,1,2)
degree=c(1,2,2,4,4,1,1,3,2,5,5,1,2,3,4,4,1,2,3,4,4,2,1)
stay=c(0,2,3,9,11,1,2,3,2,14,12,11,8,8,6,6,5,5,5,5,6,2,3)
diagnosis=c(2,3,1,1,2,2,2,1,2,2,2,4,3,1,2,2,2,3,4,2,3,1,1)
```

which can be summarized to:

```
hospital_data= data.frame(age, gender, degree, stay, diagnosis)
```

The variable *age* describes the age of the patients in years, the variable *gender* is self-explanatory (male=2, female=1), the variable *degree* shows the severity of the illness, the variable *stay* describes the number of days the patient had to stay in the hospital and the variable *diagnosis* contains the code of the illness.

(a) Create the data set *hospital_data* as described above.

(b) What are the dimensions of the data set *hospital_data*?

(c) Explain the scale of each of the variables!

(d) Create for each of the following variables an appropriate diagram and motivate the choice of diagram:

  i. *age*

  ii. *gender*

  iii. *degree*

  iv. *stay*

  v. *diagnosis*

(e) Create and compare the boxplots for the variable *stay* for females and males!

(f) Use the function *quantile* to calculate the three quantiles used in 5e! Compare these quantiles for females and males!