Chapter 1: Descriptive statistics

Learning objectives of this chapter:

- Calculating the mean, mode, median, quartiles and range by hand
- Get started with R: set the working directory, inspect and load data
- Calculating the mean, mode, median and range in R

Assignment 1.1: Descriptive statistics by hand

Calculate by hand (or by the use of a pocket calculator) the

- **mean**
- **mode**
- **median**
- **range**
- **lower and upper quartiles**
- **interquartile range**

For the following data sets:

1.1 a) 2 7 4 5 8 10 10 7 9 2 8 8 9 4 6
1.1 b) 7 7 6 5 2 1 3 7 5 9 9 10

Answer 1.1a:

Mean: ________________ Range: ________________

Mode: ________________ Lower quartile: ________________

Median: ________________ Upper quartile: ________________

Interquartile range: ________________

Answer 1.1b:

Mean: ________________ Range: ________________

Mode: ________________ Lower quartile: ________________

Median: ________________ Upper quartile: ________________

Interquartile range: ________________

1.1 c) Describe which question (1.1a or 1.1b) was more difficult to calculate and why.

Answer 1.1c:

____

____

____

____

1.1 d) Are these data sets **positively skewed** or **negatively skewed** ? Circle the correct answer and explain why you chose this answer using the relation between the **mean** , the **median** , and the **mode** .

Answer 1.1d:

These data sets are **positively** / **negatively** skewed.

Explanation:

____

____

____

Assignment 1.2: Descriptive statistics of small data sets in R

This assignment assumes you opened a new script in the code editor in RStudio. Write and run your own code to answer the following questions.

In R, a one-dimensional row of numbers is represented by a `vector` .

1.2 a) Use the `c()` function to enter the numbers from assignment 1.1a in a new vector called `dataset1` . Next, run the following code in R and explain what you see.

```
View(dataset1)
```

Hint 1.1: Check Part I of the R help for more information on how to make a `vector` *.*

R code 1.2a:

Answer 1.2a:

____

____

1.2 b) Write and run your own code in R to find the **`mean`** , **`mode`** , **`median`** , and **`range`** for the vector `dataset1` . Compare your answers with those of assignment 1.1a.

Hint 1.2: Check part IV of the R help on page 100 for descriptive statistics functions.

Hint 1.3: There is no **`mode`** *function in R, but you can find the mode in a frequency* `table` *.*

R code 1.2b:

Answer 1.2b:

Mean: ____________ Median: ____________

Mode: ____________ Range: ____________

1.2 c) Run the following code in R and explain what you see.

```
quantile(dataset1, type = 6)
```

Answer 1.2c:

____

____

1.2 d) Use the `c()` function to create a vector `dataset2` with the data from assignment 1.1b and find the **mean** , **mode** , **median** , **range** , and **quartiles** for these data. Compare your answers with those of assignment 1.1b.

R code 1.2d:

Answer 1.2d:

Mean: ____________ Range: ____________

Mode: ____________ Lower quartile: ____________

Median: ____________ Upper quartile: ____________

Assignment 1.3: Descriptive statistics of larger data sets in R

For this assignment, you need the `bloodPressure.csv` data set that you can download from the online resources. This data set contains measurements of the age, blood pressure, cholesterol level, gender, and description of a random selection of people. It is normally used to look for relationships between these variables. Note that this is fake data and does not contain actual measurements.

Let's start with importing the data set (which is available in the online resources) into R.

1.3 a) Inspect and run the following code in R to import the blood pressure data and store it in the object `dataset3` . Explain how the code works and describe what the `bloodPressure.csv` file contains.

```
dataset3 <- read.csv(file.choose())
```

Answer 1.3a:

____

____

This method of importing data can be a lot of work if there are many files or if the script will be run many times. Faster methods exist though, for example by providing the full file path.

1.3 b) Describe and test ways this code can be improved to make importing a file easier.

Hint 1.4: Look at Part I of the R help to find out more functions for importing data.

Answer 1.3b:

____

____

R code 1.3b:

The functions you used for descriptive statistics on the small data sets in assignment 1.1 can also be applied to the data set that is currently stored in `dataset3` .

1.3 c) Find the **`mean`** , **`mode`** , **`median`** , **`range`** , and **`quartiles`** for the column `Age` in `dataset3` . Describe this variable in running text using these statistics.

Hint 1.5: First find out how to extract (index) a specific column in a data frame using the `$` *sign.*

R code 1.3c:

Answer 1.3c:

____

____

____

____

For large data sets, it becomes a lot of work finding the **`mode`** in a frequency table each time. It is possible to import a package into the R session that contains a function for calculating the **`mode`** automatically. However, it is also possible to create a function that calculates the **`mode`** ourselves.

Run the following lines of R code together:

```
getMode <- function(x){
   uniqx <- unique(x)
   uniqx[which.max(tabulate(match(x, uniqx)))]
}
```

You have now created your first R function and you will see it displayed separately in the R environment. This function will give you the **`mode`** for any **numeric** vector or column. It works by first extracting all unique values, counting their frequency, and then selecting the value with the highest frequency. Note that you can use this function, but will not be required to understand or make functions like this. However, for the interested reader, part III of the R help contains more information on how to create your own functions.

1.3 d) Use the new `getMode()` function to determine the **mode** for column `Age` in `dataset3` and check if it is consistent with your answer for assignment 1.3c.

R code 1.3d:

Answer 1.3d:

Mode: ________________

1.3 e) Find the **mean** , **mode** , **median** , **range** , and **quartiles** for the column `BloodPressure` in `dataset3` . Also use the new `getMode()` function.

R code 1.3e:

Answer 1.3e:

Mean: ________________ Range: ________________

Mode: ________________ Lower quartile: ________________

Median: ________________ Upper quartile: ________________

1.3 f) Is the distribution of the values in the `BloodPressure` column **positively skewed** or **negatively skewed** ? Explain your answer using the relation between the **mean** , **median** , and **mode** .

Answer 1.1f:

These data sets are **positively** / **negatively** / **not** skewed.

Explanation:

____

____

____

1.3 g) Determine the **variance** and **standard deviation** for the column `Cholestrol` in `dataset3` .

R code 1.3g:

Answer 1.3g:

```
Variance:               ________________

Standard deviation:     ________________
```

1.3 h) Validate the relation between the **variance** and the **standard deviation** by performing a calculation in R.

R code 1.3h: