

MATH1712 Probability and Statistics II

Homework 1

<http://www1.maths.leeds.ac.uk/~voss/2017/MATH1712/>

Jochen Voss, J.Voss@leeds.ac.uk

2017/18, semester 2

This exercise sheet will be discussed in the tutorials of the week beginning 29th January.

If you have a laptop which you can run R on, please bring this laptop with you to the tutorial class. If possible, try to install RStudio from on your laptop before the class (there are different versions, you need the one called “RStudio Desktop”).

During the tutorial. For tasks (1) and (2) you should work in small groups, sharing laptops as needed. For task (3) it might be best to discuss together on the board.

(1) Install RStudio on your laptop, if you haven’t done so before, and make sure you know how to start it.

(2) The data from the R questionnaire is available at <https://goo.gl/2WMu5s> (in CSV format). Use the `read.csv()` command (see overleaf) to read the data into R. Determine the number of responses. Inspect the data and discuss what you find. (If you need help with this, have a look at the section “Handling Data in R” in the lecture notes.)

(3) Let $X \geq 0$ be a random variable which only takes positive values. Try to show that $\mathbb{E}(X) = 0$ if and only if $P(X = 0) = 1$.

Homework questions. Your solutions to these questions contribute towards your final mark for the module. Please hand in your solutions **to your tutor** via the silver pigeon holes (down the stairs from the maths reception) by **Monday, 5th February, 4pm**.

- a) Clearly mark your solution with your name, your student ID, and your tutor’s name.
- b) Staple the sheets of your solution together. Do not use plastic sleeves *etc.*
- c) Write clearly and legibly, and leave margins for the marker to write comments in.
- d) Write complete sentences, including correct punctuation.
- e) Explain how you obtained your solution. Just giving the final answer is not enough, all intermediate steps are also required.

Exercise 1. The data from the R questionnaire is available at <https://goo.gl/2WMu5s> (in CSV format). We are interested in how the answers to the questions “How would you describe your programming skills?” and “How much have you enjoyed using R so far?” relate to each other.

- a) Load the data into R.
- b) Use the command `table()` (see the section “Handling Data in R” in the lecture notes) to produce a table which shows how often each combination of answers occurs.
- c) Discuss your results.

Your solution should show the R commands you use, as well as the results you get.

Exercise 2. Let X be a random variable with mean μ and variance σ^2 . For $a \in \mathbb{R}$, consider the expectation $\mathbb{E}((X - a)^2)$.

- a) Write $\mathbb{E}((X - a)^2)$ in terms of a , μ and σ^2 .
- b) For which value a is $\mathbb{E}((X - a)^2)$ minimal?
- c) For the value a from part (b), what is $\mathbb{E}((X - a)^2)$?

Exercise 3. The sample variance of $x_1, \dots, x_n \in \mathbb{R}$ is given by

$$\hat{\sigma}^2 := \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2,$$

where \bar{x} is the sample mean. Show that the sample variance equals 0, if and only if all observations in the sample coincide.

read.csv()

The `read.csv()` command can be used to import `.csv` files into R: if we use the command

```
x <- read.csv("file.csv")
```

then the contents of the file `file.csv` will be stored in the data frame `x`. The most important things to know about the function `read.csv()` are:

- a) The filename can either denote a file on the local computer, or a file available for download from the internet. If you want R to read the file directly from the internet, replace the file name with the web address (starting with `http://` or `https://`). If the file is on the local computer, you need to change R's current directory to the directory where the file is stored before calling `read.csv()`; use the R menu "file ▷ Change dir..." to do this.
- b) By default, R uses the first row of the `.csv` file to set the column names of the data frame and assumes that the actual data starts in row 2 of the `.csv` file. If the file does not contain column names, you can use the `header=FALSE` option with `read.csv()` to tell R that the column names are not included in the data:

```
x <- read.csv("file.csv", header=FALSE)
```

You can see whether this option is needed by opening the file in Excel and looking whether the first row contains headers or not. Alternatively you can inspect the column names and the contents of the first data row in R to see whether everything looks right after importing the data.

- c) Sometimes, the columns in a `.csv` file are separated not by a comma, but using a semicolon instead. In this case you need to use the option `sep=";"` when you import the data:

```
x <- read.csv("file.csv", sep=";")
```

- d) Missing values should be represented by empty cells in the `.csv` file and are represented as special NA values in the data frame. If the `.csv` file uses a different encoding for missing values, the `na.strings` option can be used to tell `read.csv()` which cell values should be interpreted as missing values. For example, `read.csv("file.csv", na.strings=c("", "-"))` can be used for a file where missing values are indicated by either empty cells or cells containing a hyphen.

Further details about the function `read.csv()` can be found using the command

```
help(read.csv)
```

This will show the R online help for `read.csv()`, which has the complete description of how the command can be used.