Assignment 1

Aki Vehtari et al.

1 General information

The exercises of this assignment are meant to test whether or not you have sufficient knowledge to participate in the course. The first question checks that you remember basic terms of probability calculus. The second exercise checks your basic computer skills and guides you to learn some R functions. In the last three ones you will first write the math for solving the problems (you can, for example, write the equations in markdown or include a photo of hand written answers), and then implement the final equations in R (and then you can use `markmyassignment` to check your results). The last question checks that you have found the course book.

The maximum amount of points from this assignment is 3.

We prepared a quarto template specific to this assignment to help you get started. You can inspect this and future templates

- as a qmd file,
- as a rendered html file
- or as a rendered pdf file

or you can download all template `qmd` files and some additional files at templates.zip (also available on Aalto JupyterHub under `/coursedata`).

Tip

Reading instructions:

- **The reading instructions for BDA3 Chapter 1**.

Grading instructions:
The grading will be done in peergrade. All grading questions and evaluations for this assignment are contained within this document in the collapsible **Rubric** blocks.

Further information

- The recommended tool in this course is R (with the IDE RStudio).
- Instead of installing R and RStudio on you own computer, see **how to use R and RStudio remotely**.
- If you want to install R and RStudio locally, download R and RStudio.
- There are tons of tutorials, videos and introductions to R and RStudio online. You can find some initial hints from **RStudio Education pages**.

- When working with R, we recommend writing the report using `quarto` and the provided template. The template includes the formatting instructions and how to include code and figures.
- Instead of `quarto`, you can use other software to make the PDF report, but the the same instructions for formatting should be used.
- Report all results in a single, **anonymous** *.pdf -file and submit it in **peergrade.io**.
- The course has its own R package `aaltobda` with data and functionality to simplify coding. The package is pre-installed in JupyterHub. To install the package on your own system, run the following code (upgrade="never" skips question about updating other packages):

```
install.packages("aaltobda", repos = c("https://avehtari.github.io/BDA_course_Aalto/", getOption("re
```

- Many of the exercises can be checked automatically using the R package `markmyassignment` (pre-installed in JupyterHub). Information on how to install and use the package can be found in the `markmyassignment` documentation. There is no need to include `markmyassignment` results in the report.
- Recommended additional self study exercises for each chapter in BDA3 are listed in the course web page. These will help to gain deeper understanding of the topic.
- Common questions and answers regarding installation and technical problems can be found in Frequently Asked Questions (FAQ).
- Deadlines for all assignments can be found on the course web page and in Peergrade. You can set email alerts for the deadlines in Peergrade settings.
- You are allowed to discuss assignments with your friends, but it is not allowed to copy solutions directly from other students or from internet.
- You can copy, e.g., plotting code from the course demos, but really try to solve the actual assignment problems with your own code and explanations.
- Do not share your answers publicly.
- Do not copy answers from the internet or from previous years. We compare the answers to the answers from previous years and to the answers from other students this year.
- Use of AI is allowed on the course, but the most of the work needs to by the student, and you need to report whether you used AI and in which way you used them (See points 5 and 6 in Aalto guidelines for use of AI in teaching).
- All suspected plagiarism will be reported and investigated. See more about the **Aalto University Code of Academic Integrity and Handling Violations Thereof**.
- Do not submit empty PDFs, almost empty PDFs, copy of the questions, nonsense generated by yourself or AI, as these are just harming the other students as they can't do peergrading for the empty or nonsense submissions. Violations of this rule will be reported and investigated in the same way was plagiarism.
- If you have any suggestions or improvements to the course material, please post in the course chat feedback channel, create an issue, or submit a pull request to the public repository!

Rubric

- Can you open the PDF and it's not blank nor nonsense? If the pdf is blank, nonsense, or something like only a copy of the questions, 1) report it as problematic in Peergrade-interface to get another report to review, and 2) send a message to TAs.
- Is the report anonymous?

Setup

This is the template for assignment 1. You can download the qmd-file or copy the code from this rendered document after clicking on `</> Code` in the top right corner.
Please replace the instructions in this template by your own text, explaining what you are doing in each exercise.
The following will set-up `markmyassignment` to check your functions at the end of the notebook:

```
library(markmyassignment)
assignment_path = paste("https://github.com/avehtari/BDA_course_Aalto/",
"blob/master/tests/assignment1.yml", sep="")
set_assignment(assignment_path)
```

```
Assignment set:
assignment1: Bayesian Data Analysis: Assignment 1
The assignment contain the following (3) tasks:
- p_red
- p_box
- p_identical_twin
```

2 Basic probability theory notation and terms

This can be trivial or you may need to refresh your memory on these concepts (see, e.g. Aalto course *First Course in Probability and Statistics*). Explain each of the following terms with one sentence:

- probability
- probability mass (function)
- probability density (function)
- probability distribution
- discrete probability distribution
- continuous probability distribution
- cumulative distribution function (cdf)
- likelihood

Rubric

- How is the answer?
  - Totally wrong/has not tried
  - Something sensible written
  - All/almost all are correct (>=70% correct)

3 Basic computer skills

This task deals with elementary plotting and computing skills needed during the rest of the course. You can use either R or Python, although R is the recommended language in this course and we will only guarantee support in R. For documentation in R, just type `?{function name here}`.

> Subtask 3.a)
>
> Plot the density function of the Beta-distribution, with mean $\mu = 0.2$ and variance $\sigma^2 = 0.01$. The parameters α and β of the Beta-distribution are related to the mean and variance according to the following equations
>
> $$\alpha = \mu\left(\frac{\mu(1-\mu)}{\sigma^2} - 1\right), \quad \beta = \frac{\alpha(1-\mu)}{\mu}.$$

Plot the PDF here. Explain in text what you do.

```
# Useful functions: seq(), plot() and dbeta()
```

> Subtask 3.b)
>
> Take a sample of 1000 random numbers from the above distribution and plot a histogram of the results. Compare visually to the density function.

Sample and plot the histogram here. Explain in text what you do.

```
# Useful functions: rbeta() and hist()
```

> Subtask 3.c)
>
> Compute the sample mean and variance from the drawn sample. Verify that they match (roughly) to the true mean and variance of the distribution.

Compute the sample mean and variance here. Explain in text what you do.

```
# Useful functions: mean() and var()
```

> Subtask 3.d)
>
> Estimate the central 95% probability interval of the distribution from the drawn sample.

Compute the central interval here. Explain in text what you do.

```
# Useful functions: quantile()
```

> Rubric
>
> - Is the source code for the solutions included?
> - Does the plot in a) look something like this:
> - Does the plot in b) look something like this:
> - Is the computed mean in c) close to ?

- Is the variance in c) close to ?
- Is the probability interval in d) roughly ? Remember that since the interval is computed from random sample, there can be small variation, but the answers should be roughly the same!

Further formatting recommendations

- Please try to include as much code and output as needed, but as little as possible.
- Please make sure that the plots are properly labeled and are easily legible and understandable. This means
  - they should have x- and y-labels,
  - the text within should be of a size comparable to the size of the surrounding text and
  - each plot should have a concise but descriptive caption or title.
- Please make sure to report a sensible number of digits when reporting numbers. You will get more precise instructions later on, but for now think independently about how many digits of your results are important for the assignment.

4 Bayes' theorem 1

A group of researchers has designed a new inexpensive and painless test for detecting lung cancer. The test is intended to be an initial screening test for the population in general. A positive result (presence of lung cancer) from the test would be followed up immediately with medication, surgery or more extensive and expensive test. The researchers know from their studies the following facts:

- Test gives a positive result in 98% of the time when the test subject has lung cancer.
- Test gives a negative result in 96% of the time when the test subject does not have lung cancer.
- In general population approximately one person in 1000 has lung cancer.

Subtask 4.a)

The researchers are happy with these preliminary results (about 97% success rate), and wish to get the test to market as soon as possible. How would you advise them? Base your answer on Bayes' rule computations.

Tip

Relatively high false negative (cancer doesn't get detected) or high false positive (unnecessarily administer medication) rates are typically bad and undesirable in tests.

Compute the quantities needed to justify your recommendation here. Explain in text what you do. You can do the computation with pen and paper or in R. Either way, you have to explain why you compute what you compute.

If you use pen and paper, you can include scans or pictures as follows (see also assignment_instructions#fig-workflow):

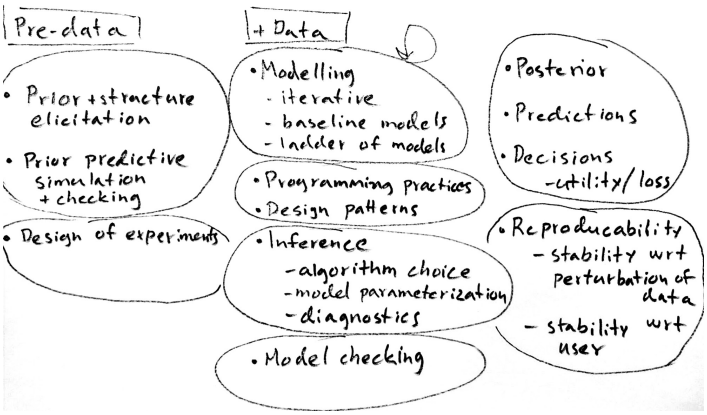

Figure 1: Parts of Bayesian workflow

See Figure 1 for illustration of parts of Bayesian workflow.

Here are some probability values that can help you figure out if you copied the right conditional probabilities from the question.

- P(Test gives positive | Subject does not have lung cancer) = 4%
- P(Test gives positive **and** Subject has lung cancer) = 0.098% this is also referred to as the **joint probability** of *test being positive* and the *subject having lung cancer*.

Rubric

- Is p(has cancer|test result is positive) computed using Bayes' formula (or its complement p(does not have cancer|test result is positive))?
- Is the result p(has cancer|test result is positive)= (or p(does not have cancer|test result is positive)=)
- Is the result motivated with something like

5 Bayes' theorem 2

We have three boxes, A, B, and C. There are

- 2 red balls and 5 white balls in the box A,
- 4 red balls and 1 white ball in the box B, and
- 1 red ball and 3 white balls in the box C.

Consider a random experiment in which one of the boxes is randomly selected and from that box, one ball is randomly picked up. After observing the color of the ball it is replaced in the box it came from. Suppose also that on average box A is selected 40% of the time and box B 10% of the time (i.e. $P(A) = 0.4$).

You will need to change the numbers to the numbers in the exercise.

```
boxes_test <- matrix(c(2,2,1,5,5,1), ncol = 2,
    dimnames = list(c("A", "B", "C"), c("red", "white")))
```

Subtask 5.a)

What is the probability of picking a red ball? Implement an R function to compute that probability.

Keep the below name and format for the function to work with `markmyassignment`:

```
p_red <- function(boxes) {
    # Do computation here, and return as below.
    # This is the correct return value for the test data provided above.
    0.3928571
}
```

Subtask 5.b)

If a red ball was picked, from which box did it most probably come from? Implement an R function to compute the probabilities for each box.

Keep the below name and format for the function to work with `markmyassignment`:

```
p_box <- function(boxes) {
    # Do computation here, and return as below.
    # This is the correct return value for the test data provided above.
    c(0.29090909,0.07272727,0.63636364)
}
```

Rubric

- Is the source code available?
- How is the answer for probability of picking a red ball?
  - No answer
  - Probability rules
  - Probability rules
- How is the answer for what box is most probable?
  - No answer
  - Bayes rule used to compute probabilities for all boxes given that the picked ball is red, but the answers are not
  - Bayes rule used to compute probabilities for all boxes given that the picked ball is red, the answers are p
  - Bayes rule used to compute probabilities for all boxes given that the picked ball is red, the answers are

6 Bayes' theorem 3

Assume that on average fraternal twins (two fertilized eggs and then could be of different sex) occur once in 150 births and identical twins (single egg divides into two separate embryos, so both have the same sex) once in 400 births (**Note!** This is not the true value, see Exercise 1.6, page 28, in BDA3). American male singer-actor Elvis Presley (1935 – 1977) had a twin brother who died in birth. Assume that an equal number of boys and girls are born on average.

Subtask 6.a)

What is the probability that Elvis was an identical twin? Show the steps how you derived the equations to compute that probability and implement a function in R that computes the probability.

You will need to change the numbers to the numbers in the exercise.

```
fraternal_prob = 1/125
identical_prob = 1/300
```

Keep the below name and format for the function to work with `markmyassignment`:

```
p_identical_twin <- function(fraternal_prob, identical_prob) {
    # Do computation here, and return as below.
    # This is the correct return value for the test data provided above.
    0.4545455
}
```

Rubric

- How is the answer for probability of Elvis having had an identical twin brother?
  - No answer
  - Probability that Elvis had an identical twin brother is computed using Bayes rule, but the result is not roughly
  - Probability that Elvis had an identical twin brother is computed using Bayes rule, and the result is roughly

7 The three steps of Bayesian data analysis

Subtask 7.a)

Fill in the three steps of Bayesian data analysis (see BDA3 section 1.1):

1.
2.
3.

Rubric

- Are the three steps listed as follows:
  1.
  2.
  3.

markmyassignment

The following will check the functions for which `markmyassignment` has been set up:

```
mark_my_assignment()
```

```
v | F W  S  OK | Context

/ |          0 | task-1-subtask-1-tests
/ |          0 | p_red()
x | 1        3 | p_red()
--------------------------------------------------------------------------------
Failure ('test-task-1-subtask-1-tests.R:21:3'): p_red()
p_red(boxes = boxes) not equivalent to 0.5.
1/1 mismatches
[1] 0.393 - 0.5 == -0.107
Error: Incorrect result for matrix(c(1,1,1,1,1,1), ncol = 2)
--------------------------------------------------------------------------------

/ |          0 | task-2-subtask-1-tests
/ |          0 | p_box()
x | 1        3 | p_box()
--------------------------------------------------------------------------------
Failure ('test-task-2-subtask-1-tests.R:19:3'): p_box()
p_box(boxes = boxes) not equivalent to c(0.4, 0.1, 0.5).
3/3 mismatches (average diff: 0.0909)
[1] 0.2909 - 0.4 == -0.1091
[2] 0.0727 - 0.1 == -0.0273
[3] 0.6364 - 0.5 ==  0.1364
Error: Incorrect result for matrix(c(1,1,1,1,1,1), ncol = 2)
--------------------------------------------------------------------------------

/ |          0 | task-3-subtask-1-tests
/ |          0 | p_identical_twin()
x | 2        3 | p_identical_twin()
--------------------------------------------------------------------------------
Failure ('test-task-3-subtask-1-tests.R:16:3'): p_identical_twin()
p_identical_twin(fraternal_prob = 1/100, identical_prob = 1/500) not equivalent to 0.2857143.
1/1 mismatches
[1] 0.455 - 0.286 == 0.169
Error: Incorrect result for fraternal_prob = 1/100 and identical_prob = 1/500

Failure ('test-task-3-subtask-1-tests.R:19:3'): p_identical_twin()
p_identical_twin(fraternal_prob = 1/10, identical_prob = 1/20) not equivalent to 0.5.
1/1 mismatches
[1] 0.455 - 0.5 == -0.0455
Error: Incorrect result for fraternal_prob = 1/10 and identical_prob = 1/20
--------------------------------------------------------------------------------

== Results =====================================================================
-- Failed tests ----------------------------------------------------------------
Failure ('test-task-1-subtask-1-tests.R:21:3'): p_red()
p_red(boxes = boxes) not equivalent to 0.5.
1/1 mismatches
[1] 0.393 - 0.5 == -0.107
Error: Incorrect result for matrix(c(1,1,1,1,1,1), ncol = 2)
```

```
Failure ('test-task-2-subtask-1-tests.R:19:3'): p_box()
p_box(boxes = boxes) not equivalent to c(0.4, 0.1, 0.5).
3/3 mismatches (average diff: 0.0909)
[1] 0.2909 - 0.4 == -0.1091
[2] 0.0727 - 0.1 == -0.0273
[3] 0.6364 - 0.5 ==  0.1364
Error: Incorrect result for matrix(c(1,1,1,1,1,1), ncol = 2)

Failure ('test-task-3-subtask-1-tests.R:16:3'): p_identical_twin()
p_identical_twin(fraternal_prob = 1/100, identical_prob = 1/500) not equivalent to 0.2857143.
1/1 mismatches
[1] 0.455 - 0.286 == 0.169
Error: Incorrect result for fraternal_prob = 1/100 and identical_prob = 1/500

Failure ('test-task-3-subtask-1-tests.R:19:3'): p_identical_twin()
p_identical_twin(fraternal_prob = 1/10, identical_prob = 1/20) not equivalent to 0.5.
1/1 mismatches
[1] 0.455 - 0.5 == -0.0455
Error: Incorrect result for fraternal_prob = 1/10 and identical_prob = 1/20

[ FAIL 4 | WARN 0 | SKIP 0 | PASS 9 ]
```

8 Overall quality of the report

Rubric

- Does the report include comment on whether AI was used, and if AI was used, explanation on how it was used?
  - No
  - Yes
- Does the report follow the formatting instructions?
  - Not at all
  - Little
  - Mostly
  - Yes
- In case the report doesn't fully follow the general and formatting instructions, specify the instructions that have not been followed. If applicable, specify the page of the report, where this difference is visible. This will help the other student to improve their reports so that they are easier to read and review. If applicable, specify the page of the report, where this difference in formatting is visible.
- Please also provide feedback on the presentation (e.g. text, layout, flow of the responses, figures, figure captions). Part of the course is practicing making data analysis reports. By providing feedback on the report presentation, other students can learn what they can improve or what they already did well. You should be able to provide constructive or positive feedback for all non-empty and non-nonsense reports. If you think the report is perfect, and you can't come up with any suggestions how to improve, you can provide feedback on what you liked and why

you think some part of the report is better than yours.