Assignment 2

Aki Vehtari et al.

1 General information

This assignment is related to Lecture 2 and BDA3 Chapters 1 and 2. You may find an additional discussion about choosing priors in a **blog post by Andrew Gelman**.

The maximum amount of points from this assignment is 3.

We prepared a **quarto template specific to this assignment (html, qmd, pdf)** to help you get started.

Tip

Reading instructions:

- **The reading instructions for BDA3 Chapter 1**.
- **The reading instructions for BDA3 Chapter 2**.

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 2. 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/assignments/tests/assignment2.yml", sep="")
set_assignment(assignment_path)
```

The following installs the `aaltobda` package:

```
#| cache: true
# Caching should be fine here
install.packages("aaltobda", repos = c("https://avehtari.github.io/BDA_course_Aalto/", getOption("re
```

2 Inference for binomial proportion

Algae status is monitored in 274 sites at Finnish lakes and rivers. The observations for the 2008 algae status at each site are presented in the dataset `algae` in the `aaltobda` package ('0': no algae, '1': algae present).

Loading the library and the data.

```
library(aaltobda)
data("algae")
# The data are now stored in the variable `algae`.
# These are the values for the prior required in the assignment
prior_alpha = 2
prior_beta = 10
```

The below data is **only for the tests**, you need to change to the full data `algae` when reporting your results.

```
algae_test <- c(0, 1, 1, 0, 0, 0)
```

Let π be the probability of a monitoring site having detectable blue-green algae levels and y the observations in `algae`. Use a binomial model for the observations y and a $Beta(2, 10)$ prior for binomial model parameter π to formulate a Bayesian model. Here it is not necessary to derive the posterior distribution for π as it has already been done in the book and it suffices to refer to that derivation. Also, it is not necessary to write out the distributions; it is sufficient to use label-parameter format, e.g. $Beta(\alpha, \beta)$.

Your task is to perform Bayesian inference for a binomial model and answer questions based on it:

Subtask 2.a)

Formulate

1. the likelihood $p(y|\pi)$ as a function of π,
2. the prior $p(\pi)$, and
3. the resulting posterior $p(\pi|y)$.

Report the posterior in the format $Beta(\alpha, \beta)$, where you replace α and β with the correct numerical values.

Tip

With a conjugate prior, a closed-form posterior has Beta form (see equations in BDA3 and in the slides).

Write the likelihood, the prior and the posterior here!

```
# These are not the actual values for the posterior!
# You will have to compute those from the data!
posterior_alpha = 2
posterior_beta = 10
```

You can do string interpolation using R inline code execution in quarto as such:

α_{prior} is **2** and β_{prior} is **10**. Or string interpolation within math: $Beta(2, 10)$

Subtask 2.b)

What can you say about the value of the unknown π according to the observations and your prior knowledge? Summarize your results with a point estimate (i.e. $E(\pi|y)$) and a 90% posterior interval.

Tip

Posterior intervals are also called credible intervals and are different from confidence intervals.

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

```
# Useful function: qbeta()

beta_point_est <- function(prior_alpha, prior_beta, data) {
    # Do computation here, and return as below.
    # This is the correct return value for the test data provided above,
    # combined with the prior provided above.
    0.2222222
}
beta_interval <- function(prior_alpha, prior_beta, data, prob=0.9) {
    # Do computation here, and return as below.
    # This is the correct return value for the test data provided above,
    # combined with the prior provided above.
    c(0.0846451, 0.3956414)
}
```

Subtask 2.c)

What is the probability that the proportion of monitoring sites with detectable algae levels π is smaller than $\pi_0 = 0.2$ that is known from historical records?

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

```
# Useful function: pbeta()

beta_low <- function(prior_alpha, prior_beta, data, pi_0=0.2) {
    # Do computation here, and return as below.
    # This is the correct return value for the test data provided above,
    # combined with the correct prior.
    0.4511238
}
```

Subtask 2.d)

What assumptions are required in order to use this kind of a model with this type of data?

Tip

No need to discuss exchangeability yet, as it is discussed in more detail in BDA3 Chapter 5 and Lecture 7.

Write your answer here!

Subtask 2.e)

Make prior sensitivity analysis by testing a couple of different reasonable priors and plot the different posteriors. Summarize the results by one or two sentences.

Plot the PDFs here. Explain shortly what you do.

```
# Useful function: dbeta()
```

Rubric

- Is source code included?
- Are the prior, likelihood and posterior forms in a) reported (derivation of posterior not necessary)?
  - No
  - Some missing
  - Yes
- Is the reported resulting posterior correct ?
  - It is not reported, that the posterior distribution is a distribution.
  - It is reported, that the posterior distribution is , but the numerical values for the parameters are incorrect
  - It is reported, that the posterior distribution is , and the numerical values for the parameters are correct.

- In part b), is there at least one point estimate reported. Sample based estimates are also OK. Points should be given if the method is right, even if the result is wrong due to a wrong posterior distribution being used. With the right posterior, mean, median, and mode are all approximately .
- For the b) part, is the 90% posterior interval reported? Sample based estimate is also OK. Points should be given if the method is right, even if the result is wrong because the posterior was wrong in the first place. If the posterior was right, the 90% posterior interval is roughly .
- For the c) part, is the posterior probability Pr(<0.2|y) reported? Points should be given if the method is right, even if the result is wrong because the posterior was wrong. If the posterior was right, the result should be approximately .
- For the d) part, does the report discuss
  - No
  - No, but other reasonable assumptions are discussed
  - Yes, but not quite right or some missing
  - Yes
- For the e) part, is there some comparison and discussion of results obtained with alternative prior parameters?
  - No
  - Yes, but the results and conclusions are clearly wrong
  - Yes

markmyassignment

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

```
mark_my_assignment()
```

3 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 visi-

ble. 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.