Uppsala University

Bayesian Statistics and Data Analysis

Assignment 2

General information

- The recommended tool in this course is R (with the IDE R-Studio). You can download R **here** and R-Studio **here**. There are many tutorials, videos and introductions to R and R-Studio online. You can find some initial hints from **RStudio Education pages**.
- When working with R, we recommend writing the report using R markdown and the provided **R markdown template**. The template includes the formatting instructions and how to include code and figures.
- Instead of R markdown, you can use other software to make the PDF report, but you should use the same instructions for formatting. These instructions are also available in **the PDF produced from the R markdown template**.
- We supply a Google Colab notebook that you can also use for the assignments. We have included the installation of all necessary R packages; hence, this can be an alternative to using your own local computer. You can find the notebook **here**. You can also open the notebook in Colab **here**.
- Report all results in a single and *anonymous* pdf. Note that no other formats are allowed.
- The course has its own R package `bsda` with data and functionality to simplify coding. To install the package, just run the following (upgrade="never" skips question about updating other packages):

  1. `install.packages("remotes")`
  2. `remotes::install_github("MansMeg/BSDA", subdir = "rpackage", upgrade="never")`

- Many of the exercises can be checked automatically using the R package `markmyassignment`. you can find information on how to install and use the package **here**. There is no need to include `markmyassignment` results in the report.
- You can find common questions and answers regarding the installation and technical problems in Frequently Asked Questions (FAQ).
- You can find deadlines and information on how to turn in the assignments in Studium.
- You are allowed to discuss assignments with your friends, but it is not permitted to copy solutions directly from other students or the internet. Try to solve the actual assignment problems with your code and explanations. Do not share your answers publicly. We compare the answers with the "urkund" system. We will report all suspected 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 **here**.

- It is *mandatory* to include the following parts in all assignments (these are included already in the template):

  1. Time used for reading: How long time took the reading assignment (in hours)
  2. Time used for the assignment: How long time took the basic assignment (in hours)
  3. Good with assignment: Write one-two sentences of what you liked with the assignment/what we should keep for next year.
  4. Things to improve in the assignment: Write one-two sentences of what you think can be improved in the assignment. Can something be clarified further? Did you get stuck on stuff unrelated to the content of the assignment etc.

- You can find information on how each assignment will be graded and how points are assigned **here**. **Note!** This grading information can change during the course, for example, if we find errors or inconsistencies or do additions to the assignments. Please feel free to comment on these grading instructions, ideally before turning in your assignment, if you think something is missing or is incorrect.

- To pass (G) the assignment, you need 70% of the total points. To pass with distinction (VG), you need 90% of the total points. See the grading information on the point allocations for each assignment.

- On cheating:

  – You are not allowed to show your assignments (text or code) to anyone. Only discuss the assignments with your fellow students. The student that show their assignment to anyone else could also be considered to cheat. Similarly, on zoom labs, only screen share when you are in a separate zoom room with teaching assistants.

  – You are not allowed to use large language models, such as ChatGPT, to write assignments.

- All mathematics need to be done in digital form to simplify grading and commenting. Hence, it is not allowed to write math on paper and add an image in the assignment. If you have difficulties to write math in latex, see https://editor.codecogs.com/.

- Do not copy the text from the PDF in the assignment files. It can contain hidden numbers and texts to make cheating with ChatGPT harder. Only read the assignments from the plain text document.

Information on this assignment

This assignment is related to Chapters 1 and 2.

Reading instructions: Chapter 1 and 2 in BDA3, see reading instructions. You may find an additional discussion about choosing priors by Andrew Gelman useful, they can be found **here**.

To use markmyassignment for this assignment, run the following code in R:

```
library(markmyassignment)
assignment_path <-
  paste("https://github.com/MansMeg/BSDA/",
        "blob/main/assignments/tests/assignment2.yml", sep="")
set_assignment(assignment_path)
# To check your code/functions, just run
mark_my_assignment()
```

Don't include `markmyassignment` results in the report.

---

1 Inference for binomial proportion (Computer)

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 `algae` dataset ('0': no algae, '1': algae present). The data can be accessed from the `bsda` R package as follows:

```
library(bsda)
data("algae")
head(algae)

## [1] 0 1 1 0 0 0

# the data is now stored in the variable 'algae'
```

So that you can test the correctness of your code implementations, we provide some results for the following **test data**. It is also possible to check the functions you need to implement with `markmyassignment`.

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

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

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 $\text{Beta}(2, 10)$ 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. $\text{Beta}(\cdot, \cdot)$.

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

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 $\text{Beta}(\cdot, \cdot)$, where you replace $\cdot$'s with the correct numerical values.

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. **Note!** Posterior intervals are also called credible intervals and are different from confidence intervals. **Note!** In your report, use the values from the data `algae`, not `algae_test`.

```
beta_point_est(prior_alpha = 2, prior_beta = 10, data = algae_test)

## [1] 0.2222222

beta_interval(prior_alpha = 2, prior_beta = 10, data = algae_test, prob = 0.9)

## [1] 0.0846451 0.3956414
```

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?

```
beta_low(prior_alpha = 2, prior_beta = 10, data = algae_test, pi_0 = 0.2)

## [1] 0.4511238
```

d) What assumptions are required in order to use this kind of a model with this type of data? **Hint!** What does the likelihood assume? (No need to discuss exchangeability yet, as it is discussed in more detail in BDA Chapter 5 and later Lectures)

e) Make prior sensitivity analysis by testing a couple of different reasonable Beta priors (i.e. test reasonable hyperparameters) and plot the different posteriors. Summarize the results by one or two sentences.

Hint! With a conjugate prior, a closed-form posterior is Beta form (see equations in the book). Useful functions: `dbeta`, `pbeta`, `qbeta` in R.

2 Inference for count data (Computer)

As a next step we will analyze count data. We will use the `warpbreaks` data from the R `datasets` package that describe the number of breaks when warping (use `?warpbreaks` to get more information. We will analyze the `breaks` variable using a Poisson model.

```
library(datasets)
data("warpbreaks")
head(warpbreaks$breaks)

## [1] 26 30 54 25 70 52

mean(warpbreaks$breaks)

## [1] 28.14815

length(warpbreaks$breaks)

## [1] 54
```

a) Derive posterior (analytically) using a Poisson likelihood with parameter λ and a Gamma prior with α (`shape` in the R `*gamma`-functions) and β (`rate` in the R `*gamma`-functions). Start with the likelihood and prior and show analytically that the prior and posterior is conjugate.

b) Compute the (analytical) posterior for the `warbreaks breaks` variable for any prior hyperparameter α and β.

c) Create a 95% credible interval, using R, for the λ parameter using the 2.5% upper and lower percentile of the posterior distribution for λ. Use a $Gamma(\alpha = 1.5, \beta = 0.1)$ prior. Include your code on how you compute this.

d) What distribution is the posterior predictive distribution, $p(\tilde{y}|y)$ in this case (see Section 2.6 in BDA3)? Describe the distribution and the parameters you have for the predictive distribution conditioned on the `warpbreaks` data. In R the distribution can be found in the stats library, but differently parametrized. Hence, compute the R parametrized distribution, i.e. using the `prob` and `size` parameter, where `rate = prob/(1 - prob)` and `shape = size` (see details in the R function for more information).

e) Simulate data from the predictive distribution and create one histogram for the predictive distribution and one for the `breaks` variable with the same range in both plots on the x axis (see `xlim = c(5,75)`). Is the Poisson model a good model for the breaks data? Why or why not?