Bayesian 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 tons of tutorials, videos and introductions to R and R-Studio online. You can find some initial hints from **RStudio Education pages**.

- Instead of installing R and RStudio on you own computer, see **how to use R and RStudio remotely**.

- When working with R, we recommend writing the report using R markdown and the provided **R markdown template**. The remplate 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 the the same instructions for formatting should be used. These instructions are available also in **the PDF produced from the R markdown template**.

- Report all results in a single, **anonymous** *.pdf -file and return it to **peergrade.io**.

- The course has its own R package `aaltobda` 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("avehtari/BDA_course_Aalto",`
     `subdir = "rpackage", upgrade="never")`

- Many of the exercises can be checked automatically using the R package `markmyassignment`. Information on how to install and use the package can be found **here**. 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.

- We collect common questions regarding installation and technical problems in a course Frequently Asked Questions (FAQ). This can be found **here**.

- Deadline for all assignments are **Sunday at 23.59**.

- 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! Useful feedback will be thanked with bonus points.

Information on this assignment

This assignment is related to Chapters 1 and 2. The maximum amount of points from this assignment is 3. You may find an additional discussion about choosing priors by Andrew Gelman useful, they can be found **here**.

Reading instructions: Chapter 1 and 2 in BDA3, see reading instructions **here** and **here**.

Grading instructions: The grading will be done in peergrade. All grading questions and evaluations for assignment 2 can be found **here**

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

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

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 file `algae.txt` ('0': no algae, '1': algae present). The data can also be accessed from the `aaltobda` R package as follows:

```
> library(aaltobda)
> data("algae")
> # 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 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($\cdot, \cdot$).

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 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?

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.

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