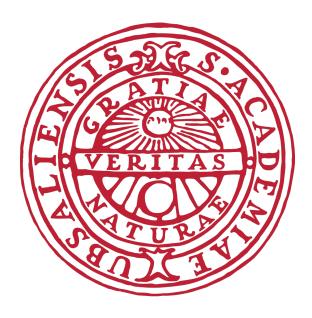BAYESIAN STATISTICS AND DATA ANALYSIS

Assignment 3

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 or solve 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 2 and 3.

   Reading instructions: Chapter 2 and 3 in BDA3, see reading instructions. Use Frank Harrell's recommendations on how to state results in Bayesian two group comparisons (and note that there is no point null hypothesis testing in this assignment).

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

```r
library(markmyassignment)
assignment_path <-
  paste("https://github.com/MansMeg/BSDA/",
        "blob/main/assignments/tests/assignment3.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 normal mean and deviation

A factory has a production line for manufacturing car windshields. A sample of windshields has been taken for testing hardness. The observed hardness values $\mathbf{y}_1$ can be found in the **windshieldy1** dataset. The data can be accessed from the **bsda** R package as follows:

```
library(bsda)
data("windshieldy1")
head(windshieldy1)

## [1] 13.357 14.928 14.896 15.297 14.820 12.067
```

We may assume that the observations follow a normal distribution with an unknown variance σ^2. We wish to obtain information about the unknown average hardness μ. For simplicity we assume standard uninformative prior discussed in the book, that is, $p(\mu, \sigma) \propto (\sigma^2)^{-1}$. It is not necessary to derive the posterior distribution in the report, as it has already been done in the book.

Below are test examples that can be used. The functions below can also be tested with *markmyassignment*. **Note!** This is *only* a test case. You need to change to the full data **windshieldy** above when reporting your results.

```
windshieldy_test <- c(13.357, 14.928, 14.896, 14.820)
```

In the report, formulate (1) model likelihood, (2) the prior, and (3) the resulting posterior.

a) What can you say about the unknown μ? Summarize your results using Bayesian point estimate (i.e. $E(\mu|y)$), a posterior interval (95%), and plot the density. A test example can be found below for an uninformative/vague Normal prior. **Note!** Posterior intervals are also called credible intervals and are different from confidence intervals.

```
mu_point_est(data = windshieldy_test)

## [1] 14.5

mu_interval(data = windshieldy_test, prob = 0.95)

## [1] 13.3 15.7
```

b) What can you say about the hardness of the next windshield coming from the production line before actually measuring the hardness? Summarize your results using Bayesian point estimate, a *predictive* interval (95%), and plot the density. A test example can be found below.

```
mu_pred_point_est(data = windshieldy_test)

## [1] 14.5

mu_pred_interval(data = windshieldy_test, prob = 0.95)

## [1] 11.8 17.2
```

Note! Predictive intervals are different from posterior intervals.

Hint With a conjugate prior a closed form posterior is Student's t form (see equations in the book). R users can use the `dt` function after doing input normalisation. We have added an R function `dtnew()` in the `bsda` R package which does that. For generating samples, you can use the corresponding `rtnew` function.

2. Inference for the difference between proportions

An experiment was performed to estimate the effect of beta-blockers on mortality of cardiac patients. A group of patients was randomly assigned to treatment and control groups: out of 674 patients receiving the control, 43 died, and out of 680 receiving the treatment, 22 died. Assume that the outcomes are independent and binomially distributed, with probabilities of death of p_0 and p_1 under the control and treatment, respectively. Set up a noninformative or weakly informative prior distribution on (p_0, p_1).

In the report, formulate (1) model likelihood, (2) the prior, and (3) the resulting posterior.

a) Summarize the posterior distribution for the odds ratio, $(p_1/(1-p_1))/(p_0/(1-p_0))$. Compute the point estimate, a posterior interval (95%), and plot the histogram. Use Frank Harrell's recommendations how to state results in Bayesian two group comparison. Below is a test case on how the odd ratio should be computed. **Note!** This is *only* a test case. You need to change to the real posteriors when reporting your results.

```
set.seed(4711)
p0 <- rbeta(100000, 5, 95)
p1 <- rbeta(100000, 10, 90)

posterior_odds_ratio_point_est(p0 = p0, p1 = p1)

## [1] 2.676

posterior_odds_ratio_interval(p0 = p0, p1 = p1, prob = 0.9)

## [1] 0.875 6.059
```

b) Try a couple of other reasonable priors. Discuss the sensitivity of your inference to your choice of prior density with a couple of sentences.

Hint With a conjugate prior, a closed-form posterior is the Beta form for each group separately (see equations in the book). You can use `rbeta()` to sample from the posterior distributions of p_0 and p_1, and use these samples and odds ratio equation to get samples from the distribution of the odds ratio.

3. Inference for the difference between normal means

Consider a case where the same factory has two production lines for manufacturing car windshields. Independent samples from the two production lines were tested for hardness. The hardness measurements for the two samples $\mathbf{y}_1$ and $\mathbf{y}_2$ are given in the files `windshieldy1.txt` and `windshieldy2.txt`. These can be accessed directly with

```
data("windshieldy1")
data("windshieldy2")
```

We assume that the samples have unknown standard deviations σ_1 and σ_2. In the report, formulate (1) model likelihood, (2) the prior, and (3) the resulting posterior. Use uninformative or weakly informative priors and answer the following questions:

a) What can you say about $\mu_d = \mu_1 - \mu_2$? Summarize your results using a Bayesian point estimate, a posterior interval (95%), and plot the histogram. Use Frank Harrell's recommendations how to state results in Bayesian two group comparison.

b) Given the model used, what is the probability that the means are exactly the same $(\mu_1 = \mu_2)$? Explain your reasoning.

Hint With a conjugate prior, a closed-form posterior is Student's t form for each group separately (see equations in the book). You can use `rt()` function to sample from the posterior distributions of μ_1 and μ_2, and use these samples to get samples from the distribution of the difference $\mu_d = \mu_1 - \mu_2$. Be careful to scale them and shift them according to their mean and variance values in R, as described above.

Hint Posterior distributions of μ_1 and μ_2 are continuous, and thus the posterior distribution of the difference $\mu_d = \mu_1 - \mu_2$ is also continuous. What is the probability that $\mu_d = 0$?