Uppsala University

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 tons of 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 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**.
- We supply a Google Colab notebook that can also be used for the assignments. We have included installation of all nessecary R packages and 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.
- 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`. Information on how to install and use the package can be found **here**. There is no need to include `markmyassignment` results in the report.
- Common questions and answers regarding installation and technical problems can be found in Frequently Asked Questions (FAQ).
- Deadlines and information on how to turn in the assignments can be found in Studium.
- You are allowed to discuss assignments with your friends, but it is not allowed to copy solutions directly from other students or from internet. 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 with urkund. All suspected plagiarism will be reported and investigated.
- 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**

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:

```
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.

General information

See formating instructions in the template **here** on what to include.

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 file `windshieldy1.txt`. The data can also 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 standard deviation σ. We wish to obtain information about the unknown average hardness μ. For simplicity we assume standard uninformative prior discussed in the book, that is, $p(\mu, \sigma) \propto \sigma^{-1}$. 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 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.

```r
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 `aaltobda` R package which does that. For generating samples, you can use the corresponding `rtnew` function.

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, 39 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.

```r
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) 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.

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 $\mu_d = 0$?