Bayesian 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 **here**.
- You can write the report with your preferred software, but the outline of the report should follow the instruction in the R markdown template that can be found **here**.
- Report all results in a single, **anonymous** *.pdf -file and return it to **peergrade.io**.
- Many of the exercises can be checked using the R package `markmyassignment`. Information on how to install and use the package can be found **here**.
- The course has its own R package with data and functionality to simplify coding. To install the package just run the following:
  1. `install.packages("devtools")`
  2. `devtools::install_github("avehtari/BDA_course_Aalto", subdir = "rpackage")`
- 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**.
- Additional self study exercises and solutions for each chapter in BDA3 can be found **here**.
- If you have any suggestions or improvements to the course material, please feel free to create an issue or submit a pull request to the public repository!!

Information on this assignment

This exercise is related to Chapters 2 and 3. The maximum amount of points from this assignment is 9. You may find Frank Harrell's recommendations on how to state results in two group comparisons helpful, they can be found **here**.

Reading instructions: Chapter 2 and 3 in BDA3, see **here** and **here**.

Grading instructions: The grading will be done in peergrade. All grading questions and evaluations for assignment 3 can be found **here**.

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

```
> library(markmyassignment)
> exercise_path <-
    "https://github.com/avehtari/BDA_course_Aalto/blob/master/exercises/tests/ex3.yml"
> set_assignment(exercise_path)
```

1. Inference for normal mean and deviation (3 points)

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 `aaltobda` R package as follows:

```
> library(aaltobda)
> 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. Formulate the Bayesian model with uninformative or weakly informative prior and answer the following questions below. Here it is not necessary to derive the posterior distribution 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`.

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

a) What can you say about the unknown μ? Summarize your results using Bayesian point and interval estimates (95%) and plot the density. A test example can be found below for an uninformative prior.

```
> 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 and interval estimates (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
```

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 a R function `dtnew()` in the `aaltobda` R package which does that. When you generate samples, use the standard `rt` function and transform them.

2. Inference for the difference between proportions (3 points)

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).

a) Summarize the posterior distribution for the odds ratio, $(p_1/(1-p_1))/(p_0/(1-p_0))$. Compute the point and interval estimates (95%) and plot the histogram. Below is a test case on how the odd ratio should be computed.

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

```
[1] 2.676
```

```
> posterior_odds_ratio_interval(p0, 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.

3. Inference for the difference between normal means (3 points)

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. 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 Bayesian point and interval estimates (95%) and plot the histogram.

b) Are the means the same? Explain your reasoning with a couple of sentences.

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 $\mu_d = \mu_1 - \mu_2$. The equivalent function in R is the `rt` function. Be careful to scale them and shift them according to their mean and variance values in R, as described above.