# Bayesian data analysis – exercise 3

This exercise is related to Chapters 2 and 3.

The maximum amount of points from this assignment is 9. In addition to the correctness of the answers, the overall quality and clearness of the report is evaluated.

Report all results to a single, **anonymous** *.pdf -file and return it to peergrade.io. Include also source code to the report (either as an attachment or as a part of the answer). By anonymity it is meant that the report should not contain your name or student number.

You may find Frank Harrell's recommendations how to state results in two group comparisons helpful http://www.fharrell.com/2017/10/bayesian-vs-frequentist-statements.html.

## 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`.

We may assume that the observations follow a normal distribution with an unknown standard deviation $\sigma$. We wish to obtain information about the unknown average hardness $\mu$. 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.

a) What can you say about the unknown $\mu$? Summarize your results using Bayesian point and interval estimates and plot the density.

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 and plot the density.

Hint: With a conjugate prior a closed form posterior is Student's $t$ form (see equations in the book). Python users can use `scipy.stats.t` module. It has both `pdf` and `cdf` functions, where you can also set the mean with the loc argument and variance with scale argument. Note that these functions are for standardized Student-$t$ with mean 0 and scale 1, so you need to do the scaling and translation yourself. R users can use the `dt` function after doing input normalisation. We have added a R file having `dtnew` function which does that in the attachments. When you generate samples, use the standard `dt` function and transform them back.

## 2. Inference for 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 were 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 and plot the histogram.

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 Beta form for each group separately (see equations in the book). You can use `scipy.stats.beta.rvs` in Python to sample from the posterior distrbutions of $p_0$ and $p_1$, and use these samples and odds ratio equation to get samples from the distribution of the odds ratio. Python users can make use of commands like `numpy.percentile` to get percentiles. For plotting, matplotlib package is recommended.

## 3. Inference for 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`.

We assume that the samples have unknown standard deviations $\sigma_1$ and $\sigma_2$. 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 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 `scipy.stats.t.rvs` function to sample from the posterior distrbutions of $\mu_1$ and $\mu_2$ in Python, 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.

Python users can use `matplotlib.pyplot.hist` and `numpy.percentile` for visualisation of histogram and computing intervals respectively.