MATH5835M Statistical Computing Practical

http://www1.maths.leeds.ac.uk/~voss/2018/MATH5835M/
Jochen Voss, J.Voss@leeds.ac.uk
2018/19, semester 1

(A) The MATH5835M module is assessed by an exam and a practical. This is the practical for the module. This practical is worth 20% of the final module marks (the other 80% coming from the exam).

(B) You must hand in your solutions by **Monday, 26th November 2018, 10am** at the latest.

(C) Your answer should be in the the form of a short report. Reports must be word-processed (not hand-written) and must be clearly marked with your name and your student ID. You should include your R code and any explanations necessary to understand your conclusions. I would expect the report to be about 5 to 6 pages long.

(D) There are 20 marks per task for correctness and completeness of your answers. In addition there are 20 marks for clarity and tidiness of presentation. The total number of marks for this practical is $4 \times 20 + 20 = 100$.

(E) You can discuss the questions and can share ideas with other students, but the work you hand in must be your own.

In this practical we will use the data set provided at

`https://www.seehuhn.de/maths/letters/`

This data set contains scanned images of handwritten letters. We will study how the variability of pixel intensities within one image depends on the letter shown in the image.

Task 1. The web page provides two versions of the data set, but we will only use the "test data" here. Load the "test data" into R, to get a matrix or an array with $32 \times 54 = 1728$ pixel intensities for each of the 5825 images. Give a short description of the data.

Task 2. For each image, compute the standard deviation of the pixel intensities within the image. Each standard deviation is computed using 1728 pixel values, and there will be 5825 standard deviations, say $s_1, \dots, s_{5825}$, in total. (Hints: you can either use a loop or the `apply()` function here. A loop might be simpler, while the `apply()` function will be more efficient. The standard deviation for the first image is 19.16695.)

Produce three histograms: of all standard deviations, of the standard deviations belonging to images showing the letter "B", and of standard deviations showing the letter "C". Discuss your results.

Task 3. We will use a Bayesian model for the values s_i corresponding to images of a letter $L \in \{\mathrm{A}, \dots, \mathrm{Z}\}$: we assume that these s_i are independent samples from the normal distribution $\mathcal{N}(\mu_L, 13^2)$, where the unknown mean μ_L has prior distribution $\mu_L \sim \mathcal{U}[38, 55]$.

For each letter L, determine the posterior density $p(\mu_L \,|\, s)$ for the mean corresponding to the letter L. Explain how rejection sampling can be used to sample from the posterior $p(\mu_L \,|\, s)$.

Task 4. Use the sampling method from task 3, to compute Monte-Carlo estimates of $\mathbb{E}(\mu_L \,|\, s)$ and $\mathrm{Var}(\mu_L \,|\, s)$ for different letters L. Carefully explain your choice of Monte Carlo sample size and comment on your results. (You can choose yourself which letters to include here; try find a combination which shows some differences between letters.)

Task 5 (for PhD students only). Choose a problem from the area of your PhD, where an expectation or a probability plays a rôle. Using the methods we learned in the module, compute a meaningful Monte Carlo estimate for this expectation or a probability. On one or two pages, give a short description of the context and explain your Monte Carlo estimate. (If you need help in choosing a problem for this question, please come and talk to me.)