Bayesian Data Analysis - Assignment 1

The exercises of this assignment are not related to any book chapter, but rather work to test whether or not you have sufficient knowledge to participate the course. Reading the first chapter might anyway help with the first question.

There are four pen and paper exercises and one computer task. The maximum amount of points from this assignment is 3. 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. By anonymity it is meant that the report should not contain your name or student number.

General instructions for reporting:

- Think about the reviewer: Don't write too long answers: Answer to the question asked, no need to repeat the question, make your answer clear and easy to find, especially when your answer is a number.
- Always refer to the attached figures, tables and source code in the main text.
- Always report probabilites as a number between 0 and 1, not as percentages.
- When you are asked to include source code, NEVER take a screenshot of your code editor and attach the code as an image.
- Don't use fancy, editor specific pdf features in the report. Make sure that your pdf opens as it should with all editors and all operating systems.
  - Don't use the attachment feature of adobe.
  - No metadata or comments embedded in the pdf.

---

1. **(Basic probability theory notation and terms)**. This can be trivial or you may need to refresh your memory on these concepts. Note that some terms may be different names for the same concept.

   a) Explain the following terms with one sentence:
   - probability
   - probability mass
   - probability density
   - probability mass function (pmf)
   - probability density function (pdf)
   - probability distribution
   - discrete probability distribution
   - continuous probability distribution
   - cumulative distribution function (cdf)
   - likelihood

b) Answer the following questions in one or two sentence:
   - What is observation model?
   - What is statistical model?
   - What is the difference between mass and density?

2. **(Basic computer skills)** This task deals with elementary plotting and computing skills needed during the rest of the course. You can use either R or Python. For more about Python, see the docs (https://docs.python.org/3/), for R, you can just type ?{function name here}

   a) Plot the density function of Beta-distribution, with mean $\mu = 0.2$ and variance $\sigma^2 = 0.01$. The parameters α and β of the Beta-distribution are related to the mean and variance according to the following equations

$$\alpha = \mu\left(\frac{\mu(1-\mu)}{\sigma^2} - 1\right), \quad \beta = \frac{\alpha(1-\mu)}{\mu}.$$

   Limit x-axis values between 0 and 1 since that is the interval where Beta-distribution is defined. Also, for plotting historgram, use bins that are small enough, but not too small. (Useful Python functions: `numpy.arange` and `scipy.stats.beta.pdf`
   Useful R functions: `seq` and `dbeta`)

   b) Take a sample of 1000 random numbers from the above distribution and plot a histogram of the results. Compare visually to the density function.
   (Useful Python functions: `scipy.stats.beta.rvs` and `matplotlib.pyplot.hist`
   Useful R functions: `rbeta` and `hist`)

   c) Compute and report the sample mean and variance from the drawn sample. Verify that they match (roughly) to the true mean and variance of the distribution.
   (Useful Python functions: `numpy.mean` and `numpy.var`
   Useful R functions: `mean` and `var`)

   d) Estimate the central 95%-interval of the distribution from the drawn samples.
   (Useful Python functions: `numpy.percentile`
   Useful R functions: `quantile`)

3. **(Bayes' theorem)** A group of researchers have designed a new inexpensive and painless test for detecting lung cancer. The test is intended to be an initial screening test for the population in general. A positive result (presence of a lung cancer) from the test would be followed up immediately with medication, surgery or more extensive and expensive test. The researchers know from their studies the following facts:

   - Test gives a positive result in 98% of the time when the test subject has lung cancer.
   - Test gives a negative result in 96 % of the time when the test subject does not have lung cancer.
   - In general population approximately one person in 1000 has lung cancer.

The researchers are happy with these preliminary results (about 97% success rate), and wish to get the test to market as soon as possible. How would you advice them? Base your answer on elementary probability calculus.

4. **(Bayes' theorem)** We have three boxes, A, B and C. There are

   - 2 red balls and 5 white balls in the box A,
   - 4 red balls and 1 white ball in the box B, and
   - 1 red ball and 3 white balls in the box C.

   Consider a random expreriment in which one of the boxes is randomly selected and from that box one ball is randomly picked up. After observing the color of the ball it is put back in the box it came from. Suppose also that on average box A is selected 40% of the time and box B 10% of the time.

   What is the probability of picking a red ball? If a red ball was picked, from which box it most probably came from? (For latter question, it is enough to infer the answer from unnormalized posterior probabilities)

5. **(Bayes' theorem)** On average fraternal twins (two fertilized eggs) occur once in 125 births and identical twins (single egg divides into two separate embryos) once in 300 births. American male singer actor Elvis Presley (1935 – 1977) had a twin brother who died in birth. What is the probability that Elvis was an identical twin? Assume that equal number of boys and girls are born on average.