UNIVERSITY OF EDINBURGH
SCHOOL OF MATHEMATICS
BAYESIAN DATA ANALYSIS

Assignment 1

- To be uploaded to Learn by 17:00, February 28, 2020.
- This assignment is worth 50% of your final grade for the course.
- Assignments should be typed (LaTeX, Word, etc.) and should be no more than 10 pages (including figures but excluding the appended code).
- Answers to questions should be in full sentences.
- Any output (e.g., graphs, tables) from R/JAGS that you use to answer questions must be included with the assignment. Also, please append your R/JAGS code at the end of the assignment.
- The assignment is out of 100 marks.
- You are expected to work independently and not discuss the assignment with others (a plagiarism detection software will compare the submissions of all of the students).

1. (**30 marks**) A study is conducted on a six-sided die being fair (i.e. all of the sides having equal probability) or not. After 120 throws, we obtained the following throw counts for the 6 sides:

Side	1	2	3	4	5	6
Count	17	24	22	17	24	16

   (a) (**8 marks**) We model the dice throws as independent categorical random variables, with the probability of the six sides collected in the vector $\theta = (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6)$. Let $c = (c_1, c_2, c_3, c_4, c_5, c_6)$ denote the counts of the 6 different sides out of N throws.

   Given θ, what is the distribution of c? What is the distribution of the likelihood of θ given c? What would be a suitable conjugate prior for this likelihood? Which parameter choices in this family of conjugate prior correspond to Jeffrey's prior?

   (b) (**7 marks**) Based on the observed throw counts, and using Jeffrey's prior, what is the posterior distribution in θ? What are the expected values of $\theta_1, \ldots, \theta_6$ according to posterior distribution? Compute symmetric 95% credible intervals for them. What are the marginal distributions for these 6 parameters?

   (c) (**5 marks**) With the Null hypothesis being that the die is fair (i.e. $\theta = (1/6, 1/6, 1/6, 1/6, 1/6, 1/6)$), what is the p-value of these observations? Please discuss the implications of this result. [Hint: you might have to simulate from the posterior distribution in R. It is contained in the package extraDistr.]

(d) (**5 marks**) What is the posterior predictive distribution for the next $m = 60$ trials? Please write down the probability mass function. What is the posterior predictive probability that there will be more 5s than 6s among these m trials? (Hint: use R for this calculation).

(e) (**5 marks**) We are given another opportunity to test the die, and obtain the following throw counts out of 120 throws:

Side	1	2	3	4	5	6
Count	22	20	30	16	16	16

What is the posterior distribution for θ after taking into account these samples as well? How do the posterior means and 95% credible intervals for $\theta_1, \ldots, \theta_6$ change?

2. (**35 marks**) The exponential distribution is often used for modelling waiting times. The following list (from Ghitany et al., 2008) shows the waiting times (in minutes) for 100 customers in bank.

0.8	0.8	1.3	1.5	1.8	1.9	1.9	2.1	2.6	2.7	2.9	3.1	3.2
3.3	3.5	3.6	4.0	4.1	4.2	4.2	4.3	4.3	4.4	4.4	4.6	4.7
4.7	4.8	4.9	4.9	5	5.3	5.5	5.7	5.7	6.1	6.2	6.2	6.2
6.3	6.7	6.9	7.1	7.1	7.1	7.1	7.4	7.6	7.7	8	8.2	8.6
8.6	8.6	8.8	8.8	8.9	8.9	9.5	9.6	9.7	9.8	10.7	10.9	11
11	11.1	11.2	11.2	11.5	11.9	12.4	12.5	12.9	13	13.1	13.3	13.6
13.7	13.9	14.1	15.4	15.4	17.3	17.3	18.1	18.2	18.4	18.9	19	19.9
20.6	21.3	21.4	21.9	23.0	27	31.6	33.1	38.5				

We start by modelling these waiting times as being independent, and identically distributed according to an exponential distribution with rate parameter λ.

(a) (**5 marks**) Before collecting this data, two experts were asked about the waiting times. The first said that the typical waiting times are between 5-10 minutes, while the second said that they are between 0-25 minutes. We consider both experts equally trustworthy. What is the conjugate prior for the exponential distribution? Construct a mixture prior based on the opinions of the two experts.

(b) (**8 marks**) Using the prior constructed in part (a), what is the posterior distribution for rate λ of the exponential distribution? Plot the posterior distribution of the expected waiting time $1/\lambda$, compute its posterior mean, and construct a symmetric 95% credible interval. Compute the posterior probability of waiting for longer than 20 minutes.

(c) (**3 marks**) Check for prior/data conflict by making the prior/normalised likelihood/posterior plot in the rate λ of the exponential distribution. Does the prior seem to be compatible with the posterior?

(d) (**5 marks**) Plot the density function of the posterior predictive distribution describing the waiting times of future customers. Compute its mean, and construct a symmetric 95% credible interval.

(e) (**7 marks**) When modelling waiting times, a popular alternative to the exponential distribution is the Lindley distribution. This is a mixture of exponential(λ) and gamma($2, \lambda$) distributions

with mixing proportions $\left(\frac{\lambda}{\lambda+1}, \frac{1}{\lambda+1}\right)$. Its density and cumulative distribution functions (CDF) are defined for $\lambda > 0$, $x \geq 0$ as

$$f_L(x) = \frac{\lambda^2}{\lambda+1}(1+x)\exp(-\lambda x),$$
$$F_L(x) = 1 - \frac{\lambda+1+\lambda x}{\lambda+1}\exp(-\lambda x).$$

Suppose that we use this distribution for modelling the waiting times, along with the same prior as we used in part (a). Write down the posterior density of the parameter λ, and compute the expected waiting times according to the posterior.

(f) (**7 marks**) Do a Q-Q plot for the data for both models [Hint: this is a plot that compares the quanties of the data with the quantiles of the posterior predictive distribution.]. Which one seems to fit the data better? If $F_1(x)$ and $F_2(x)$ are two CDFs, their Kolmogorov-Smirnov distance is defined as $\sup_{x\in\mathbb{R}} |F_1(x) - F_2(x)|$. Compute the Kolmogorov-Smirnov statistics for this data for both models. Discuss the results. [Hint: use the ks.test function in R].

3. (**35 marks**) In this exercise, we will implement linear regression and Bayesian linear regression on an dataset about abalone (a type of sea-snail). The dataset is included in the file abalone.data, and some description is given in the file abalone.names. There are in total 9 different attributes measured:

Name	Data type	Measurement	Description
Sex	nominal		M (male), F (female), and I (infant)
Length	continuous	mm	Longest shell measurement
Diameter	continuous	mm	perpendicular to length
Height	continuous	mm	with meat in shell
Whole weight	continuous	grams	whole abalone
Shucked weight	continuous	grams	weight of meat
Viscera weight	continuous	grams	gut weight (after bleeding)
Shell weight	continuous	grams	after being dried
Rings	integer		Rings+1.5 gives the age in year

The number of rings can be used to determine the age of the abalone. However this requires cutting the shell through the cone, staining it, and counting the number of rings through a microscope - this is an expensive and time-consuming task. Hence we are interested in predicting the age from the other 8 parameters instead.

(a) (**5 marks**) Load the dataset from abalone.data in R, and perform a linear regression on the age based on the covariates excluding the number of rings [the age is the number of rings + 1.5]. Discuss the summary statistics and the quality of the linear fit. [Hint: you can use the read.table function for this.]

(b) (**5 marks**) Decide on a Bayesian linear regression model for analysing this dataset (with age as the response). Describe the likelihood as well as the priors for all parameters. [Hint: the categorical variable sex (M/I/F) can be included as two dummy variables IsMale and IsFemale. We recommend using Gamma priors for the inverse covariance, and normal priors for the regression coefficients.]

(c) (**5 marks**) Center the data by subtracting the means from each column in the data frame (this improves the mixing of the Markov chain). Fit the model in JAGS and report posterior summaries for all quantities of interest.

(d) (**4 marks**) Check the Gelman-Rubin diagnostics using functions gelman.plot and gelman.diag (Hint: you need to run multiple chains for this. Obtaining a sufficiently low value that is smaller than 1.05 might require increasing the sample size and the length of the burn-in period).

(e) (**3 marks**) Check the sensitivity to the prior distribution (this might include changing the hyper-parameter values and/or the distribution used).

(f) (**5 marks**) Perform model checks (QQ plots based on the residuals).

(g) (**3 marks**) Compute 95% credible intervals for the model parameters (the regression coefficients and the variance parameter). Which one of these intervals contain 0? Discuss the results.

(h) (**5 marks**) We have an additional sample, for which all of the measurements were done except for counting the number of rings. The recorded attributes are as follows.

Sex	M
Length	0.515
Diameter	0.400
Height	0.133
Whole weight	0.531
Shucked weight	0.231
Viscera weight	0.122
Shell weight	0.168

Plot the densitity of the posterior predictive distribution for the age based on the Bayesian linear regression model [Hint: you can use the density function in R for this.] Compute the posterior mean and the symmetric 95% credible interval.