STAT 535 — Bayesian Data Analysis
Test 2 — Spring 2022

Important: For this exam, you are not allowed to receive help from anyone except me on the exam. For example, you may not talk to other students about the exam problems, and you may not look at other students' exams. Violations of this policy may result in a 0 on the exam, an F for the course, and/or punishment by the USC Office of Academic Integrity. For graduate students, a violation could result in the loss of assistantship.

I will answer queries asking for clarification about the exam questions. Since this is an exam (and not homework), I will probably decline to provide very much help in solving the problems, but I am happy to clarify questions if necessary.

All data are given on my personal course website.

For all questions, show as much work as possible/appropriate! If you use R to solve a problem, please include the R code (e.g., put the code in an appendix). You should upload a Word document or pdf with your answers into Blackboard by Thursday, April 7 at 4:00 p.m.

1. Suppose a gun (if it fires correctly) has an exit velocity (in feet per second) for the bullet that follows a gamma distribution with shape parameter 4 and rate parameter 1/500. However, the gun misfires with probability θ, and if there is a misfire, the recorded exit velocity is 0. Define a variable W that indicates whether the gun fires correctly or no: Let $W = 1$ if the gun fires correctly and let $W = 0$ if the gun misfires. Also let Y be the exit velocity of the bullet; note that Y will be 0 if there is a misfire and Y will be positive if the gun fires correctly. Suppose we have data on 19 attempted firings: We have the indicator values $W_1, W_2, \ldots, W_{19}$ and the exit velocity values $Y_1, Y_2, \ldots, Y_{19}$. The 20th and last attempted firing has not occurred yet. Here are the data:

```
Y.values<-c(975.51,1734.02,2177.37,2965.93,1250.45,2821.97,0.00,2322.25,0.00,
972.20,4352.78,2546.13,2186.04,0.00,869.86,1281.62,3557.40,1234.57,741.99)
W.values <-c(1,1,1,1,1,1,0,1,0,1,1,1,1,0,1,1,1,1,1)
```

The unknown quantities are thus W_{20}, Y_{20}, and the misfire probability θ. Assume that *a priori* we favor all possible values of θ equally. Assume that all gun firing attempts are independent, so that conditional on θ, Y_{20} and W_{20} are independent of the previous data $W_1, W_2, \ldots, W_{19}$ and $Y_1, Y_2, \ldots, Y_{19}$. However, the distribution of θ does depend on $W_1, W_2, \ldots, W_{19}$. Let us consider the full conditional distributions of these quantities.

(a) State the distribution of θ, given the values of W_{20} (and given $W_1, W_2, \ldots, W_{19}$).

(b) Describe the distribution of W_{20}, given θ.

(c) State the distribution of Y_{20} given θ and given that $W_{20} = 1$.

(d) Describe in words the distribution of Y_{20} given θ and given that $W_{20} = 0$. [Hint: What is/are the possible value(s) of Y_{20}, given θ and given that $W_{20} = 0$?]

(e) Based on your answers to parts (a)-(d), write a Gibbs Sampler in R code to draw 20000 values from the full conditional distributions of each of θ, W_{20}, and Y_{20}. Your Markov chain should result in values $\theta^{[1]}$, $W_{20}^{[1]}$, $Y_{20}^{[1]}$, $\theta^{[2]}$, $W_{20}^{[2]}$, $Y_{20}^{[2]}$, $\theta^{[3]}$, $W_{20}^{[3]}$, $Y_{20}^{[3]}$, etc. (So your chain should produce 20000 values of each of Y_{20}, W_{20}, and θ.) Note that for some of these draws, the `ifelse` function in R can be useful. As an example of `ifelse` (which is not at all specific to this problem), the code

```
x1 <- runif(1,min=0,max=1)
x2 <- ifelse(test=(x1>0.5), yes=rnorm(1,mean=0,sd=1), no=rnorm(1,mean=100,sd=1))
```

generates a random Uniform$(0,1)$ value `x1` and a random value `x2` that is a $N(0,1)$ variable if $x1 > 0.5$ and is a $N(100,1)$ variable otherwise. Remember that a test for equality in the `ifelse` statement would have a syntax like `test=(x1==0.5)`.

By the way, it shouldn't matter what your initial values of Y_{20}, W_{20}, and θ in the chain are. The initial values just need to be possible values for those quantities.

(f) Using your draws, what is your estimate of θ, the overall probability of a misfire? Using your draws, what is the expected exit velocity (including the possibility of both misfires and correct fires) for the 20th gun firing? Briefly explain how you got these.

2. Suppose the amount of liquid squirted from a dispenser (in mg) follows an exponential distribution with mean β:

$$f(y) = \frac{1}{\beta} e^{-y/\beta}$$

if $y > 0$ (and 0 elsewhere). The mean β is unknown, but suppose we are **certain** *a priori* that $\beta > 1$. For our observed data, we have only a single observation $y = 5.2$ mg.

(a) Consider the *rate* parameter $\theta = 1/\beta$. **Based on our prior knowledge**, what values could θ take?

(b) What would be a reasonable prior distribution to choose for θ? If we believe *a priori* that the mean β is around 10 (and quite likely to be between 5 and 15), then suggest reasonable hyperparameter value(s) for the prior on θ. (There is not just one right answer for this, but briefly justify your choice.)

(c) Set up a Metropolis-Hastings algorithm to sample values from the posterior distribution of θ. As the proposal distribution $q(\theta_{propose}|\theta_{curr})$, use a beta distribution with first parameter

$$a_{curr} = \left(\frac{1-\theta_{curr}}{V} - \frac{1}{\theta_{curr}} \right) (\theta_{curr})^2$$

and second parameter

$$b_{curr} = \left(\frac{1-\theta_{curr}}{V} - \frac{1}{\theta_{curr}} \right) \theta_{curr}(1-\theta_{curr})$$

where θ_{curr} is the current value of the chain and where $V = k\theta_{curr}(1-\theta_{curr})$ (where k may be chosen to be any number between 0 and 1) is the variance of the proposal distribution. Write the steps of the algorithm, including the correct form of the acceptance ratio. [Note: To simplify the writing of the acceptance ratio, just use the notation a_{curr}, b_{curr} to represent the expressions above, and (where appropriate in the acceptance ratio) the analogous notation $a_{propose}$ and $b_{propose}$.]

(d) Explain how you could alter the algorithm if the acceptance rate was too high or too low.

EXTRA CREDIT (8 points): Actually code the algorithm in R to sample from the posterior of θ. Provide a point estimate of θ, and a 90% credible interval for θ, as well as diagnostic plots to check the quality of the MCMC algorithm. [This is a bit messy/complicated, but manageable for students who have an excellent understanding of the M-H algorithm. Even if you can't get the algorithm to work fully, if you provide some clearly presented code that is a start to it, I will give you a few extra credit points.]

3. A regression analysis was undertaken to study the relationship between the age (X, in years) of teenage mothers and the baby's birth weight (Y, in kg). The analyst decided to fit (based on data for 10 babies) a linear regression having the model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2), \quad i = 1, \ldots, 10.$$

```
y <- c(2289,3393,3271,2648,2897,3327,2970,2535,3138,3573)/1000
x1 <- c(15,17,18,15,16,19,17,16,18,19)
```

(a) Before looking at the data, the analyst asks an expert to give a guess for the expected birth weights for two hypothetical babies: one whose mother was 15 and another whose mother was 18. The expert guessed expected weights of 2.70 kg and 3.00 kg, respectively. Explain why merely these two "hypothetical prior observations" are sufficient to obtain the necessary prior information on $\boldsymbol{\beta}$ needed for the conjugate analysis we studied in class.

(b) Based on the information given, suggest a prior mean vector for the prior on $\boldsymbol{\beta}$.

(c) For the diagonal matrix $\mathbf{D}$ that plays a role in the conjugate analysis, choose a diagonal matrix with all 1's along the diagonal. What does such a choice of $\mathbf{D}$ imply about our level of certainty in our prior information about $\boldsymbol{\beta}$?

(d) Choose a gamma prior on the error precision parameter τ that has first parameter $a = 0.5$ and second parameter $b = 0.01663$. What does the value $a = 0.5$ imply about our level of certainty in our prior information about τ? [Alternatively, if using `stan_glm`, you could choose a default exponential prior on the error standard deviation, but still answer the question above about $a = 0.5$.]

(e) Estimate the model: Write your fitted model with point estimates for the model coefficients, and provide a 90% credible interval for the coefficient of mother's age. Use the model to predict the birth weight for a baby whose mother is 17.5 years old.

(f) Check the model fit and predictive accuracy using your favorite techniques. Be clear and complete in your explanations.

4. A multiple regression model is being built to predict the response variable $Y =$ **common (base-10) logarithm** of the survival time in days using a set of 4 candidate predictors (see the `survivaldatabayes.txt` file on the course website for data). The data set consists of 54 patients. Perform a Bayesian regression with noninformative priors for $\boldsymbol{\beta}$ and σ^2.

```
x1 = blood-clotting index
x2 = prognostic measurement
x3 = enzyme measurement
x4 = liver function measurement
y = survival time (had been measured in days, before common (base-10) log transformation)
```

(a) Explain why the analyst may have chosen to define the response variable Y as the (common) logarithm of survival days rather than as survival days itself.

(b) Based on your R analysis (specifically the credible intervals for the β's) which of the predictor variables seem to have higher posterior probabilities of being "unimportant?" Explain your answer.

(c) Give a point estimate for the expected survival time (in days) for a patient with blood-clotting index 6.0, prognostic measurement 65, enzyme measurement 2.00, and liver function measurement 150. Show how you got your answer.

(d) Give a point estimate for the ratio of expected survival time (in days) for patients with enzyme measurement 3.00 to expected survival time (in days) for patients with enzyme measurement 2.00 (holding other predictors constant). Explain briefly how you got your answer.

(e) Consider selecting a set of predictor variables in the model. If we only consider first-order terms as potential predictors (no interactions), then use any model selection techniques to choose among the class of possible models. What model is chosen as best? How does this relate to your answer in part (a)?

5. In class, we fit a Poisson regression model to explain the number of sparrow offspring using age as a predictor, with both linear and quadratic age terms in the model. The course website gives that same data set, but including an additional predictor, a binary variable about whether the sparrow lives in an urban environment or not (1=urban, 0=rural). Fit a Bayesian Poisson regression with linear and quadratic age terms AND the "urban" predictor in the model. Incorporate however much prior information you would like; just clearly specify how you are incorporating it. Do things like model selection, checking model fit, etc., and write your conclusions about the regression model, including explaining the effect of the "urban" variable on expected offspring. In particular, discuss the posterior predicted number of offspring for a 3-year-old rural sparrow. This is somewhat open-ended, so you have some flexibility to do whatever you'd like in the analysis; just explain your modeling choices and conclusions in clear and understandable writing. The data are at:

   `https://people.stat.sc.edu/hitchcock/sparrowdatamore.txt`

6. For parameter μ, suppose you have a prior model that is Normal with mean 10 and variance 10^2. After seeing the data, the posterior model is Normal with mean 5 and variance 5^2. You wish to test $H_0 : \mu \geq 6$ vs. $H_a : \mu < 6$.

   (a) Give the posterior probability that the alternative hypothesis is true. Also calculate and interpret the posterior odds of the alternative hypothesis.

   (b) Calculate and interpret the prior odds of the alternative hypothesis.

   (c) Calculate the Bayes Factor for the alternative hypothesis. Interpret this in plain English for someone with little familiarity with Bayesian statistics.

   (d) If we had wanted the Bayes Factor for the **null** hypothesis, what value would this be?