STAT 491: Final Exam
Name:

1. **Format**: Submit the exam to D2L and include the R Markdown code and a PDF file with output. Submitting the RMD file will increase your chances at partial credit. Please verify that all of the code has compiled and the graphics look appropriate.
2. **Advice**: Be sure to adequately justify your answers and appropriately reference any sources used. Even if you are not able to answer a question completely, do your best to provide an answer and discuss solutions that you tried. Include your code and graphics for each question and create neat output by using options like `kable()` for tables and writing results in line with R commands.
3. **Computer Code / Reproducibility:** Please turn in all relevant computer code to reproduce your results; a reproducible document is a requirement. Include all relevant code and output needed to answer each question and write an answer to each question. *Even if the answer seems obvious from the output, make sure to state it in your narrative as well.*
4. **Resources and Citations:** While the exam is open book and you can use any resources from class or freely available on the internet, this is strictly an individual endeavor and **you should not discuss the problems with anyone outside the course instructor including class members.** All resources, including websites, should be acknowledged.
5. **Exam Questions:** If clarification on questions is required, please email the course instructor: andrew.hoegh@montana.edu.
6. **A note on sharing / reusing code:** This is a huge volume of code is available on the web to solve any number of problems. For this exam you are allowed to make use of any online resources (e.g., StackOverflow) but you must explicitly cite where you obtained any code you directly use (or use as inspiration). Any recycled code that is discovered and is not explicitly cited will be treated as plagiarism. All communication with classmates is explicitly forbidden.
7. **Due Date:** The exam is due Wednesday May 10th at 5PM. Exams submitted after the due date, but before Thursday May 11th at 5PM will have a 10% late charge deducted.

Academic Honesty Statement

Include the following statement at the beginning of your submission.

> I, ___ (your full name here) ___, hereby state that I have not communicated with or gained information in any way from my classmates or anyone other than the course instructor during this exam, and that all work is my own.

In the event that you have inadvertently violated the above statement, you should not sign above and instead discuss the situation with the course instructor.

1. Yelp (22 points)

Suppose you find yourself in Las Vegas and, after mastering Monte Carlo techniques applied to Black Jack, you are looking for a restaurant to celebrate with a 5-star meal. A quick search reveals 7 restaurants in the vicinity. You want to identify restaurants with highest probability of a 5-star review (based on historical yelp reviews).

```r
yelp <- read_csv("https://raw.githubusercontent.com/stat456/Exams/main/yelp_final.csv") %>%
  dplyr::select(-prop, -ID) %>%
  rename(num_reviews = n)
yelp %>% dplyr::select(-categories) %>% kable()
```

name	num_reviews	fivestars
Baja Fresh	6	6
Smooth Eats	57	49
Munch Box	3	2
Hikari	615	290
StripSteak	1091	498
Burger Bar	2440	776
Jenni Pho	343	76

There are several ways to answer this research question, but use a logistic regression model that includes each restaurant in a single model framework.

A. Model Specification (4 points) Write out the statistical model and include clear and complete notation. State and justify appropriate prior distributions for the parameters in the model.

B. MCMC Q (4 points) Why is MCMC necessary/helpful to find posterior distributions in this problem?

C. JAGS Code (4 points) Write JAGS Code for this model.

D. Parameter Summary (4 points) Create a graphic to visualize posterior means and uncertainty intervals for all of your parameters. Provide some intuition about the interpretation of those variables, consider transforming them to be on the probability scale. Include relevant labels, axes, titles, and captions.

E. Bayesian Contrasts (4 points) Construct a posterior predictive distribution and answer the following questions:

1. The probability of a better review at Baja Fresh than Smooth Eats
2. The probability of a worse review at Baja Fresh than Smooth Eats
3. The probability of the same review at Baja Fresh as Smooth Eats
4. The probability of a better review at Jenni Pho than Baja Fresh

F. Summary (2 points) Which of the 7 restaurants do you choose, why?

2. Prior Sensitivity (16 points)

For this question we will re-use the yelp dataset in a beta-binomial framework ($y \sim Bernoulli(\theta)$ and $\theta \sim Beta(a, b)$) and fit separate models for `Smooth Eats` and `Munch Box`. For each of the prior distributions specified below create a plot that overlays the prior distribution and the posterior distribution. Include relevant labels, axes, titles, and captions.

A. Beta(.1, .1) (2 points)

B. Beta(1, 1) (2 points)

C. Beta(5, 1) (2 points)

D. Beta(50, 10) (2 points)

E. Summary (4 points) Write a couple of paragraphs summarizing how the prior distributions change inferences across the 4 settings. Does your restaurant recommendation change based on the prior distribution.

F. Prior Choice (4 points) Which of the four priors would you select, why?

3. Bikes (24 points)

Consider a dataset that contains bike rentals from Capital Bike Share in Washington, D.C. In particular, we are interested in whether there are differences in bike rental times from bikes originating at the Lincoln Memorial and the Washington Memorial (Jefferson Dr & 14th St SW).

```r
bikes <- read_csv('https://raw.githubusercontent.com/stat456/Exams/main/bikes.csv')
bikes %>% head() %>% kable()
```

TripTime	StartStation	EndStation
26.29	Lincoln Memorial	Constitution Ave & 2nd St NW/DOL
26.24	Jefferson Dr & 14th St SW	Jefferson Memorial
27.28	Jefferson Dr & 14th St SW	Jefferson Memorial
27.26	Jefferson Dr & 14th St SW	Jefferson Memorial
20.75	Lincoln Memorial	Jefferson Memorial
21.34	Lincoln Memorial	Jefferson Memorial

A. (4 points) Create a figure that explores whether trip times (in minutes) differ between the two stations. Include relevant labels, axes, titles, and captions.

B. (4 points) Write out a statistical model to explore this question, including appropriate priors.

c. (4 points) Fit the model in JAGS and write a paragraph summary of your results.

d. (4 points) Calculate the probability that the average trip time is shorter for trips originating from the Lincoln Memorial than Jefferson Dr & 14th St SW.

e. (4 points) Construct a posterior predictive distribution and calculate the probability that a single trip time is shorter for trips originating from the Lincoln Memorial than Jefferson Dr & 14th St SW.

f. (4 points) Note the dataset seems to contain bike trips from multiple individuals that rode together (for example see rows 2 - 4 with three bike rentals departing from Jefferson Dr & 14th St SW with almost the same trip time). Which model assumption does this violate? How concerned are you about this violation?