STAT 532: Midterm Exam
Due: October 15 at 1:10 PM
Name:

Please turn in the exam to D2L and include the R Markdown code *and either* a Word or PDF file with output. While the exam is open book, meaning you are free to use any resources from class, this is strictly an individual endeavor and **you should not discuss the problems with anyone outside the course instructor including class members.** The instructor will answer questions related to the data, expectations, and understanding of the exam, but will not fix or troubleshoot broken code.

1. (30 points Avocado Price Analysis)

This question will focus on the analysis of avocado prices using a sample of prices recorded in the Western region of the United States. The price represents the average weekly price for an avocado for a given week. For this question, use the sample of avocado prices for 20 weeks shown below.

```
avo.price <- c(0.99, 1.13, 1.17, 0.93, 1.53, 0.92, 0.79, 0.66, 1.25, 0.85, 1.18, 1.01, 1.04, 0.89, 0.98
```

a. (5 points)

Select and defend a sampling model for the average avocado price.

b. (5 points)

Identify parameters in the sampling model that require prior distributions. Then summarize the distributions you have selected and defend your choice of distribution and the parameters in the distribution.

c. (5 points)

Write a detailed description of the procedure you will use to draw samples from the posterior distribution. Then implement this procedure (include code in document).

d. (5 points)

Plot the posterior distribution and construct a confidence interval for the mean price of avocados in the western region. How does your interval differ, practically and philosophically, from the one obtained by `t.test(avo.price)`?

e. (5 points)

Suppose a posterior predictive distribution puts probability on negative prices for avocados. What does this mean in terms of your prior - sampling model combination and propose a solution.

f. (5 points)

There is another dataset that contains avocado prices for conventional and organic avocadoes for different regions across the country. Write out a sampling model to incorporate this information (account for different prices by avocado type and region) and state what parameters would need prior distributions.

2. (35 points - Speed Camera Violations)

The city of Chicago has established areas designated as Children's Safety Zones https://www.cityofchicago.org/city/en/depts/cdot/supp_info/children_s_safetyzoneporgramautomaticspeedenforcement.html that are near schools and parks. In an effort to reduce accidents in these areas, cameras have been set up to track vehicle speed and red light violations.

For this analysis we will look at a small subset of this dataset available at http://math.montana.edu/ahoegh/teaching/stat532/data/speed_camera.csv. More information and the complete data is available at https://www.kaggle.com/chicago/chicago-red-light-and-speed-camera-data

a. (5 points)

The dataset contains a sample of violations from two different speed cameras. Implement a non-Bayesian procedure to sumarize the behavior of traffic violations at each camera location and assess whether you believe there to be differences across the two locations. Clearly summarize your findings and detail all assumptions made in this model, including a discussion about the implications of the assumed sampling model.

b. (5 points)

Create a plot of the number of traffic violations at each camera. Then choose and defend a sampling model for the observed data. Note: you will have a chance to re-evaluate your sampling model later too.

c. (5 points)

Identify the parameters in your sampling model that need prior distributions. Then select prior distributions and justify the family of the distribution and provide some intuition into the parameters that you selected.

d. (5 points)

Implement and fully describe a procedure to find the posterior distribution for your parameters of interest, either analytically or computationally. This description should include a mathematical justification for what you are doing and be written in a way that your first-year statistics graduate student colleagues could understand. Finally include a plot of the posterior distibution for all parameters of interest.

e. (5 points)

Conduct a posterior predictive check to assess how well your data matches with your prior-sampling model combination. If you are not satisfied, please discuss some options for improving the model.

f. (5 points)

Assume it turns out the city is interested in identifying locations with more than 20 speeding violations, in order to allocate resources. Rather than focusing on the mean of the distribution, use your posterior predictive distribution to compute the following probabilities, where Y_{005} and Y_{068} are future observations from each camera.

- $Pr[Y_{068} > 20]$
- $Pr[Y_{005} > 20]$
- $Pr[Y_{005} > Y_{068}]$

g. (5 points)

Using the Bayesian model you have developed, revisit the analysis from part a. Summarize your findings from the Bayesian model and then the differences between the two approaches.