STAT 532: Midterm Exam
Due: Friday October 11 at 11:59 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.

1. (35 points Mount Rainier)

This question will focus on the analysis of groups attempting to climb Mount Rainier. In particular, we are interested in differences in the success probability for two routes: Ingraham Direct and Gibralter Ledges. Each row in the dataset represents a climbing party, where `Route` is the route the party took, `Attempted` is the number of climbers in the party attempting to reach the summit, and `Succeeded` is the number of climbers that reached the summit.

```
rainier <- read_csv('http://math.montana.edu/ahoegh/teaching/stat446/climbing_statistics.csv')
rainier <- rainier %>% filter(Route %in% c('Ingraham Direct', 'Gibralter Ledges')) %>% select( Route, At
```

The goal will be to model the success probability (of a single climber, not climbing party) on each route.

a. (6 points)

Select and defend a sampling model for estimating the success probability of climbers. Include clear notation on how you will handle climbing groups and climbers within a group (note there are a few ways to accomplish this).

b. (6 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. You need to specify prior distributions for the success probability of each route.

c. (6 points)

Implement and fully describe a procedure to find the posterior distribution for your parameters of interest, either analytically or computationally. (include code in document).

d. (6 points)

Plot the posterior distribution and construct a credible interval for the success probability of each route.

e. (6 points)

Now construct a posterior distribution for the difference in the probabilities of success for the two routes. Plot this distribution and construct a credible interval.

f. (5 points)

Summarize your findings from parts c, d, and e in a way that the Superintendent of Mount Rainier National Park could understand. This should be no more than 3 or 4 sentences.

2. (35 points - Snowfall Data)

Given that it is October and ski season is approaching, your dilemma is whether you should buy a season pass at Bridger Bowl or Big Sky. To help you make an informed decision, we have collected snowfall data from nearby Snotel weather stations. The dataset contains snow water equivalent (swe) on January 1 at two Snotel sites: Brackett Creek in the Bridger Range near Bridger Bowl and Lone Mountain near Big Sky Resort. As the amount of water in an inch of snow can vary, swe standardizes the depth in water (not snow).

```
swe_data <- tibble(swe = c(11.0, 4.7, 9.3, 13.3, 11.0, 7.9, 5.2, 11.3, 10.5,
                 10.8, 6.2, 8.7, 8.6, 6.5, 7.7, 5.6, 11.2, 7.6),
        station = c(rep('Brackett Creek',9), rep('Lone Mountain', 9)),
        year = as.character(rep(c(2011:2019), 2)))
```

a. (5 points)

A paired t-test is a common method for assessing differences between two populations. The code below runs a paired t-test on the swe data. Describe the results and your inferential takeaways from the t-test.

```
x = swe_data %>% filter(station == "Brackett Creek") %>% select(swe) %>% pull()

y = swe_data %>% filter(station == "Lone Mountain") %>% select(swe) %>% pull()

t.test(x,y, paired = T)
```

```
##
##  Paired t-test
##
## data:  x and y
## t = 1.6968, df = 8, p-value = 0.1282
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.4507798  2.9618909
## sample estimates:
## mean of the differences
##                1.255556
```

b. (5 points)

Now we will analyze this data from a Bayesian perspective. Choose and defend a sampling model for the observed data. Note you'll need to model the mean snowfall at both stations. (I'm not looking for anything elaborate, so restrict your choices to what we have talked about in class.)

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 hyper-parameters that you selected.

d. (10 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)

Using the Bayesian model you have developed, revisit the analysis from part a. Find the posterior distribution that Brackett Creek (Bridger Bowl) has more swe than Lone Mountain (Big Sky). Summarize your findings from the Bayesian model and then the differences between the two approaches.

e. (5 points)

Assume you are also interested in the following probabilistic statements:

- $Pr[Y_{BC} > 9]$
- $Pr[Y_{LM} > 9]$
- $Pr[Y_{BC} > Y_{LM}]$

where Y_{BC} and Y_{LM} are future observations of swe depth on January 1. Construct posterior predictive distributions and compute the probabilites.