Lab 4

Name here

Q1. (4 points)

In lecture and the activity this week, we showed the following results:

1. $\int \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \theta^{(a-1)}(1-\theta)^{(b-1)}d\theta = 1$
2. $\int \theta^{(a-1)}(1-\theta)^{(b-1)}d\theta = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$
3. $\int \theta^{(a-1)}(1-\theta)^{(b-1)}\theta^{z}(1-\theta)^{(N-z)}d\theta = \frac{\Gamma(a+b+N)}{\Gamma(a+z)\Gamma(N-z+b)}$

One of your classmates asked, we do we care? [my paraphrasing] Address this question or, in particular, how are these three results useful in finding the posterior distribution when θ is a probability and our data consist of z successes from N trials.

Q2. (4 points)

Recall that if you are interested in estimating θ, a probability of an event occurring, when you've collected N independent trials and observed z successes and placed a beta prior distribution on θ with parameters a and b, then the posterior distribution for $\theta|N, z$ is beta$(a + z, b + N - z)$.

Furthermore, the mean of the posterior distribution can be written as

$$E[\Theta|N, z] = \left(\frac{z}{N}\right)\left(\frac{N}{N+a+b}\right) + \left(\frac{a}{a+b}\right)\left(\frac{a+b}{N+a+b}\right)$$

such that the posterior mean is a weighted average of the data mean ($\frac{z}{N}$) and the prior mean ($\frac{a}{a+b}$), where the weights are ($\frac{N}{N+a+b}$) for the data piece and ($\frac{a+b}{N+a+b}$) for the prior.

Write a short paragraph discussing how the formulation, along with knowledge of N, will impact your choice of a and b.

Q3. (10 points)

Use a dataset containing homes in the Seattle, WA area http://www.math.montana.edu/ahoegh/teaching/stat408/datasets/SeattleHousing.csv for this question.

Estimate the posterior distribution for the probability that houses in Seattle have more than 2 bathrooms.

```
library(tidyverse)
seattle <- read_csv('http://www.math.montana.edu/ahoegh/teaching/stat408/datasets/SeattleHousing.csv')
  mutate(more_than2baths = bathrooms > 2)

z <- sum(seattle$more_than2baths)
N <- nrow(seattle)
```

a. (2 pts) Justify your prior distribution.

b. (2 pts) State the probability model you will using. You can, but don't need, to write out the full functional form of the probability mass/distribution function.

c. (2 pts) What is the form of your posterior distribution?

d. (2 pts) Plot your prior and posterior distributions on the same figure.

e. (2 pts) Pretend your cousin has recently accepted a new job that requires relocating to Seattle. Summarize your findings (with regard to probability of finding a house with more than 2 bathrooms) in a non-technical manner avoiding statistical lingo.

Q4. (14 points)

Recall that, under a classical statistical perspective, an approximate 95% confidence interval for binary data can be calculated as $\hat{p} \pm 1.96\sqrt{\hat{p}(1-\hat{p})/n}$, where $\hat{p}$ is the maximum likelihood estimate $\sum y_i/n$ if y_i are binary data.

If we consider a scenario with 100 trials and 50 successes this can be calculated by hand `.5 - qnorm(.975) * sqrt(.5^2 / 100), .5 + qnorm(.975) * sqrt(.5^2 / 100)` or with software `binconf(x = 50, n= 100, alpha = .05, method = 'asymptotic')` to achieve the same result.

a. (4 points) Select 2 different prior distributions and calculate a 95% interval using a Bayesian perspective. Comment on differences or similarity with your results.

b. (4 points) Now suppose we have observed 5 trials and 0 successes, what is the 95% confidence interval in this situation? Do you have any concerns about this interval?

b. (6 points) Compare the 95 % posterior intervals, based on 5 trials and 0 successes, for the following three prior distributions.

- $\theta \sim beta(.01, .01)$
- $\theta \sim beta(1, 1)$
- $\theta \sim beta(.01, 10)$

Do you have any concerns about these intervals?