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.

The results, particularly the third equation, are essential for calculating the posterior distribution is this setting. The third equation results in the marginal probability of the data. After doing this calculation, we can identify the posterior distribution:

$$\begin{aligned}
p(\theta|z,N) &= \frac{p(z,N|\theta)p(\theta)}{p(z,N)} \\
&= \frac{p(z,N|\theta)p(\theta)}{\int p(z,N|\theta)p(\theta)d\theta} \\
&= \frac{\theta^{z}(1-\theta)^{(N-z)} \times (\Gamma(a)\Gamma(b)/\Gamma(a+b))\theta^{(a-1)}(1-\theta)^{(b-1)}}{\int \theta^{z}(1-\theta)^{(N-z)} \times (\Gamma(a)\Gamma(b)/\Gamma(a+b))\theta^{(a-1)}(1-\theta)^{(b-1)}d\theta} \\
&= \frac{(\Gamma(a)\Gamma(b)/\Gamma(a+b))\theta^{z+a-1}(1-\theta)^{b+N-z-1}}{(\Gamma(a)\Gamma(b)/\Gamma(a+b))\int \theta^{z+a-1}(1-\theta)^{b+N-z-1}d\theta} \\
&= \frac{\theta^{z+a-1}(1-\theta)^{b+N-z-1}}{\int \theta^{z+a-1}(1-\theta)^{b+N-z-1}d\theta} \\
&= \frac{\theta^{z+a-1}(1-\theta)^{b+N-z-1}}{\Gamma(a+b+N)/(\Gamma(a+z)\Gamma(N-z+b))} \\
&= \frac{(\Gamma(a+z)\Gamma(N-z+b))\theta^{z+a-1}(1-\theta)^{b+N-z-1}}{\Gamma(a+b+N)} \\
&\sim Beta(a+z, b+N-z)
\end{aligned}$$

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.

This information, along with thinking about a and b, as prior "pseudo" successes is quite helpful in selecting these parameters. In particular, we want to be cognizant of selecting value of a and b that would be "too large" and highly influential with respect to the posterior distribution.

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.

```r
library(tidyverse)
library(scales)
library(Hmisc)
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.

I have little knowledge of this parameter, and hence, will opt to use a uniform prior (Beta(1,1)). With a large data set, this will have minimal impact on the posterior distribution, relative to the data.

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.

We assume each house results in a Bernoulli trial with a probability term corresponding to having more than 2 bathrooms. Or alternatively, we could consider the collection of houses using a binomial distribution.

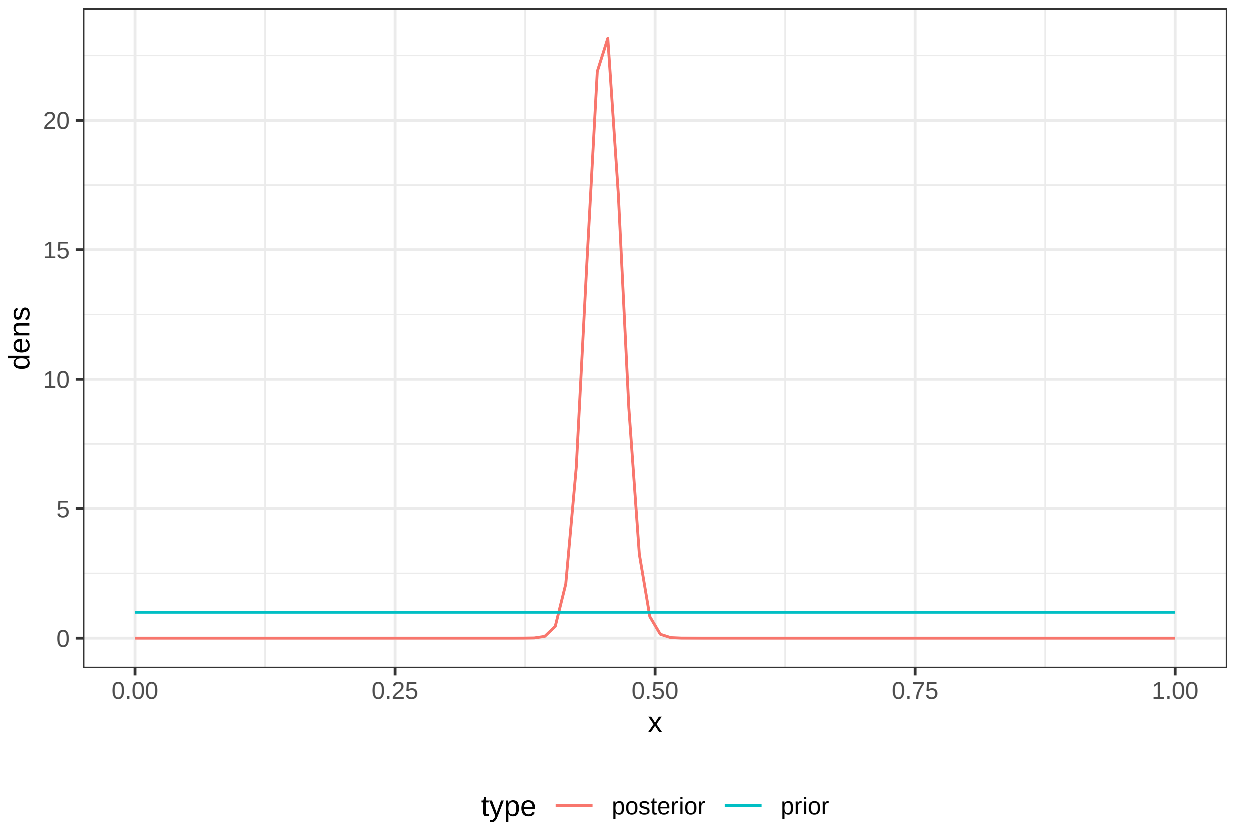
c. (2 pts) What is the form of your posterior distribution?

Given this prior distribution and probability model for our data, the resultant posterior distribution is also a beta distribution, with parameters 393 and 478

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

```r
num_x <- 100
x_seq <- seq(0,1, length.out = num_x)

tibble(dens = c(dbeta(x_seq, 1 + z, 1 + N - z),dbeta(x_seq, 1, 1)),
       x = c(x_seq, x_seq),
       type = c(rep('posterior', num_x), rep('prior', num_x))) %>%
  ggplot(aes(y=dens, x = x, color = type)) +
  geom_line() + theme_bw() + theme(legend.position = 'bottom')
```



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.

Hello Hans,

Congratulations on the new job - and the new baby girl. You are going to need extra bathrooms when those kids become teenagers. I did a quick analysis and found that you should be able to find a place with more than two bedrooms, roughly 42% to 48% of houses have more than 2 bathrooms. On the other hand, I hope the job pays well, because Seattle is very expensive.

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.

- With a uniform prior, a 95% interval would be (0.404, 0.596)
- With a beta(.1,.1) prior, a 95% interval would be (0.403, 0.596)

Both are very similar to what we'd get with the asymptotic classical 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?

```
binconf(x = 0, n= 5, alpha = .05, method = 'asymptotic')
```

```
##  PointEst Lower Upper
##         0     0     0
```

It makes sense that our point estimate would be at zero, but the confidence interval is very troubling. In particular, only having a point mass at zero doesn't make sense.

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

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

A 95% interval would be (0, 0.01)

- $\theta \sim beta(1, 1)$

A 95% interval would be (0.004, 0.459)

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

A 95% interval would be (0, 0.003)

Do you have any concerns about these intervals?

All three are more reasonable than the previous setting. However, the differences can be fairly substantial. So it will be important to give consideration to selecting prior distributions.