

Activity 5

Name here

Q1. “Integration”

Solve the following integrals, and write a short summary of why your results are valid.

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

Q2. Posterior Mean - weighted combination

Assume you are interested in estimating θ , a probability of an event occurring. You have placed a beta prior distribution on θ with parameters a and b . You’ve also collected N independent trials and observed z successes. This results in a posterior distribution $(\theta|N, z)$ that is $\text{beta}(a+z, b+N-z)$.

Consider the mean of this posterior distribution $\frac{z+a}{N+a+b}$, which can be rewritten as below.

$$\frac{z+a}{N+a+b} = \frac{z}{N+a+b} + \frac{a}{N+a+b} \quad (1)$$

$$= \frac{N}{N} \frac{z}{N+a+b} + \frac{a+b}{a+b} \frac{a}{N+a+b} \quad (2)$$

$$= \frac{z}{N} \frac{N}{N+a+b} + \frac{a}{a+b} \frac{a+b}{N+a+b} \quad (3)$$

$$(4)$$

Discuss how the posterior mean is a weighted average of the data mean and the prior mean.

What are the weights for each component?

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.

Q3. DBDA Exercise 6.4.

a. [Purpose: To explore an unusual prior and learn about the beta distribution in the process.] Suppose we have a coin that we know comes from a magic-trick store, and therefore we believe that the coin is strongly biased either usually to come up heads or usually to come up tails, but we don’t know which. Express this belief as a beta prior. (Hint: See Figure 6.1, upper-left panel.)

A beta(.1,.1) prior would have this shape. You could also use a mixture of beta distributions.

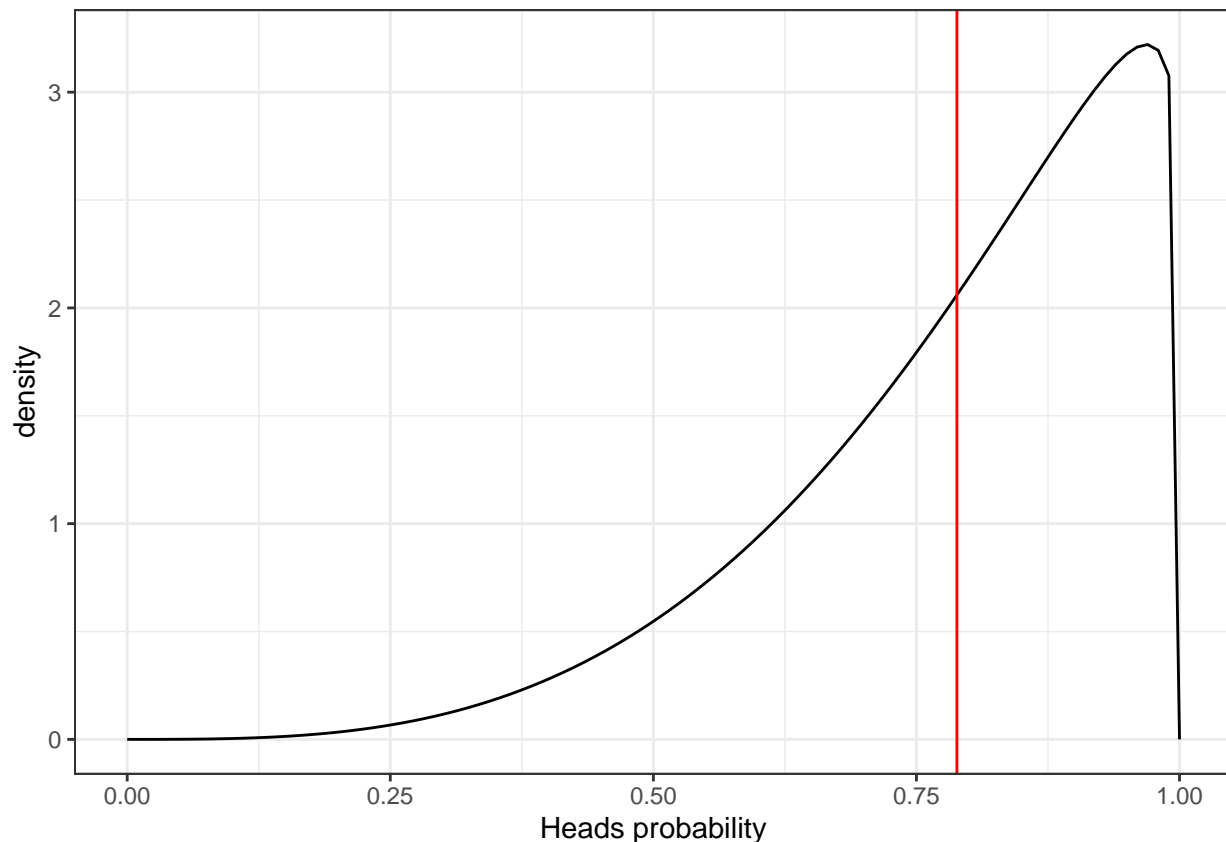
b. Now we flip the coin 5 times and it comes up heads in 4 of the 5 flips. What is the posterior distribution?

The posterior distribution is $Beta(4.1, 1.1)$

c. Create a plot of your posterior distribution and create a line to denote the posterior mean.

```
prob_seq <- seq(0, 1, length.out = 100)

tibble(density = dbeta(prob_seq, 4.1, 1.1),
       prob_seq = prob_seq) %>%
  ggplot(aes(y = density, x = prob_seq)) +
  geom_line() + theme_bw() + xlab('Heads probability') +
  geom_vline(xintercept = 4.1 / 5.2, color = 'red')
```



4. Exercise 6.5.

[Purpose: To get hands on experience with the goal of predicting the next datum, and to see how the prior influences that prediction.]

a. Suppose you have a coin that you know is minted by the government and has not been tampered with. Therefore you have a strong prior belief that the coin is fair. Say, $Beta(10,10)$ You flip the coin 10 times and get 9 heads. What is your predicted probability of heads for the 11th flip (as a 95% interval)? Explain your answer carefully; justify this choice of prior.

This prior corresponds to the equivalent of 20 coin flips, with equal probability of heads or tails. The posterior distribution would be $Beta(19,11)$ and would have an 95 % interval of (0.46, 0.79).

b. Now you have a different coin, this one made of some strange material and marked (in fine print) “Patent Pending, International Magic, Inc.” _You opt for an uninformative prior ($\text{Beta}(1,1)$). You flip the coin 10 times and get 9 heads. What is your predicted probability of heads for the 11th flip (as a 95% interval)? Explain your answer carefully; justify this choice of prior.

This prior corresponds to the equivalent of 2 coin flips, with equal probability of heads or tails. The posterior distribution would be $\text{Beta}(10,2)$ and would have an 95 % interval of (0.59, 0.98).

5. Data Analysis

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 3 bedrooms.

```
library(tidyverse)
seattle <- read_csv('http://www.math.montana.edu/ahoegh/teaching/stat408/datasets/SeattleHousing.csv')
mutate(more_than3beds = bedrooms > 3)

## Rows: 869 Columns: 14
## -- Column specification -----
## Delimiter: ","
## dbl (14): price, bedrooms, bathrooms, sqft_living, sqft_lot, floors, waterfr...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
z <- sum(seattle$more_than3beds)
N <- nrow(seattle)
```

With a uniform prior, this results in a Beta distribution with $a = 275$ and $b = 596$

6. Monte Carlo Techniques

We have previously used simulation, in the form of Monte Carlo techniques, to estimate probabilities of outcomes. Soon we will use Monte Carlo techniques to assist with summarizing posterior distributions.

For the following distributions:

1. calculate an exact mean and 95% interval (a quantile based-approach is fine)
2. simulate 1000 samples from the distribution, then calculate an approximate mean and 95% interval
3. simulate 10000 samples from the distribution, then calculate an approximate mean and 95% interval
4. simulate 100,000 samples from the distribution, then calculate an approximate mean and 95% interval

a

$\text{Beta}(1,1)$

```
sims2a <- rbeta(1000, 1, 1)
sims3a <- rbeta(10000, 1, 1)
sims4a <- rbeta(100000, 1, 1)
```

1. Mean = .5, interval = (.025, .975)
2. Mean = 0.495, interval = (0.021, 0.976)
3. Mean = 0.501, interval = (0.025, 0.973)

4. Mean = 0.501, interval = (0.024, 0.975)

b

Beta(10,1)

```
sims2b <- rbeta(1000, 10, 1)
sims3b <- rbeta(10000, 10, 1)
sims4b <- rbeta(100000, 10, 1)
```

1. Mean = 0.909, interval = (0.692, 0.997)
2. Mean = 0.913, interval = (0.712, 0.998)
3. Mean = 0.91, interval = (0.693, 0.998)
4. Mean = 0.909, interval = (0.692, 0.998)

c

Beta(10,90)

```
sims2c <- rbeta(1000, 10, 90)
sims3c <- rbeta(10000, 10, 90)
sims4c <- rbeta(100000, 10, 90)
```

1. Mean = 0.1, interval = (0.05, 0.166)
2. Mean = 0.099, interval = (0.047, 0.168)
3. Mean = 0.1, interval = (0.049, 0.166)
4. Mean = 0.1, interval = (0.05, 0.165)

d

Comment how effective the Monte Carlo techniques are at estimating these properties of the distributions.

It seems pretty good, especially when the number of samples increases.