Activity 5

Week 4 Recap: Conditional Probability and Bayes Rule

- Marginal, Joint, Conditional Probability
- Bayes Rule
- Bayes Rule with data

Week 5 Overview: Binomial Probability Exact Analysis

- Bayes Rule with data
- "Calculus"
- Bayesian Data Analysis with binary data

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

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

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 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.

$$\begin{aligned} \frac{z+a}{N+a+b} &= \frac{z}{N+a+b} + \frac{a}{N+a+b} && (1)\\ &= \frac{N}{N}\frac{z}{N+a+b} + \frac{a+b}{a+b}\frac{a}{N+a+b} && (2)\\ &= \frac{z}{N}\frac{N}{N+a+b} + \frac{a}{a+b}\frac{a+b}{N+a+b} && (3)\\ & && (4) \end{aligned}$$

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

What are the weights for each component?

---

from 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.)

b.

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

c.

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

---

from DBDA 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.

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 uniformative prior (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.

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)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
```

```
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
seattle <- read_csv('http://www.math.montana.edu/ahoegh/teaching/stat408/datasets/SeattleHous
  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)
```

---

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

Beta(1,1)

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

1. Mean = .5, interval = (.025, .975)

b

Beta(10,1)

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

c

Beta(10,90)

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

d

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