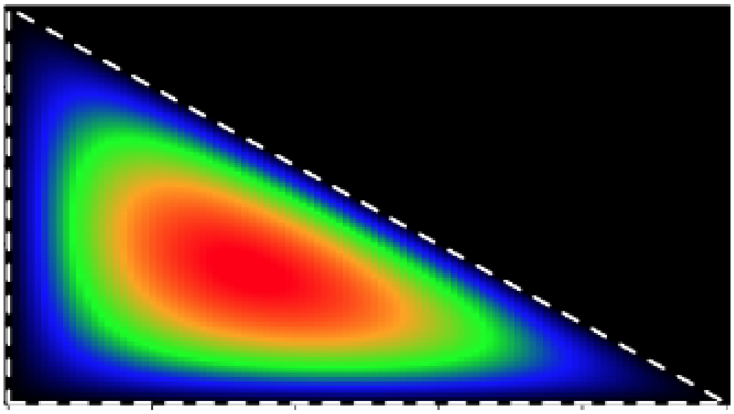
Studio 7 Conjugate priors
18.05, Spring 2024



Overview of the studio

This studio explores numerical integration and conjugate priors through the beta-binomial and Dirichlet-categorical distributions.

R introduced in this studio

The R needed is introduced in studio7-samplecode.r. We will use apply a simple numerical integration scheme and

functions: `dbeta(), ddirichlet()`

Download the zip file

- You should have downloaded the studio7 zip file from our site.
- Unzip it in your 18.05 studio folder.
- You should see the following R files
  studio7.r, studio7-samplecode.r, studio7-test.r

  and the following other files

  studio7-instructions (this file), studio7-test-answers.html

Prepping R Studio

- In R studio, open studio7-samplecode.r and studio7.r
- Using the Session menu, set the working directory to source file location. (This is a good habit to develop!)
- Answer the questions in the detailed instructions just below. Your answers should be put in studio7.r

- Solution code will be posted tomorrow at 10 pm

Detailed instructions for the studio

• Go through **studio7-samplecode.r** as a tutorial. Pay special attention to parts about preforming numerical integration.

Problem 1.

In this problem, we revisit the bent coin, our go-to metaphor for any model with two potential outcomes. The coin has unknown probability θ of heads. Given a $Beta(\texttt{a},\texttt{b})$ prior distribution on θ, we will explore two ways to compute the posterior and study the interplay between the prior and the observed data. After a series of flips, let `n_heads` and `n_tails` be the respective number of observed heads and tails.

Problem 1a. Here you will finish the code for the function

```
studio7_problem_1a(a, b, n_heads, n_tails, theta, m)
```

In this problem we will estimate the posterior density at `theta` using standard Bayesian updating. Thus, you will need to compute the likelihood function and the Bayes numerator BN, and then normalize by the integral

$$\int_0^1 \text{BN}(\theta) d\theta.$$

While this integral has a closed expression in our setting, in many cases it doesn't. So, we will use a simple numerical integration (Riemann sums) to estimate it.

We discretize the interval $[0,1]$ into `m` sub-intervals $[\frac{i-1}{m}, \frac{i}{m}]$ with i going from 1 and `m`, and estimate the posterior predictive probability by

$$\int_0^1 \text{BN}(\theta)\, d\theta \simeq \sum_{i=1}^{m} \text{BN}\left(\frac{i}{m}\right) \cdot \frac{1}{m}.$$

Use the cat statements to print the likelihood function at `theta` and then estimate the posterior pdf at `theta`.

Problem 1b. Here you will finish the code for the function

```
studio7_problem_1b(a, b, n_heads, n_tails, theta)
```

In this problem we will again calculate the posterior density at `theta`, but this time using the fact that the Beta distribution is a conjugate prior for the binomial distribution.

Use conjugacy and the `dbeta()` function to return the value of the posterior density at `theta`.

Compare the outputs of problem 1a and 1b when you make the parameter `m` larger, and appreciate the benefit of conjugate priors.

Problem 1c. Here you will finish the code for the function

```
studio7_problem_1c(a, b, n_heads, n_tails)
```

In this problem, you will plot the posterior pdf. You may want to use the output of the Problem 1b to facilitate plotting.

Problem 1d. Here you compete the function

`studio7_problem_1d()`

Here you will run your code from problem 1c to explore the influence of the prior and outcomes on the posterior.

Begin by setting a *flat prior* (`a` $=$ `b` $= 1$), and increasing the number of tosses, *while keeping a fixed proportion.* Start with `n_heads` $= 1$ and `n_tails` $= 3$ and then increase the number of throws to $(2, 6)$, $(3, 9)$,...,$(10, 30)$ and so on. Use the first cat statement to describe how the shape of the posterior pdf evolves.

Now repeat this starting from a *biased prior* with `a` $= 30$, `b` $= 10$. Use the second cat statement to explain the differences between the two cases.

Problem 2.

We will now generalize the setting of Problem 1 to a 3-sided coin (your imagination is required to do some heavy lifting here) with 3 possible outcomes that we call *heads*, *tails*, and *wings*. The coin is bent (as far as 3-sided coins can bend) and we do not know the probabilities of each outcomes. We will denote by p_1, p_2 and p_3 the possible probabilities assigned respectively to the outcomes heads, tails, and wings. We know these are non-negative and $p_1 + p_2 + p_3 = 1$.

As in PSet 7, the likelihood function for a 3-sided coins is a categorical distribution. The *Dirichlet* distribution is a conjugate prior. Here, for integers α_1, α_2, α_3, the Dirichlet distribution $\text{Dir}(3, (\alpha_1, \alpha_2, \alpha_3))$ has density

$$f(p_1, p_2, p_3) = c p_1^{\alpha_1 - 1} p_2^{\alpha_2 - 1} p_3^{\alpha_3 - 1},$$

where c is a normalizing constant.

Suppose that we set $\text{Dir}(3, (\alpha_1, \alpha_2, \alpha_3))$ as the prior distribution and observe the outcomes `ns =c(n_heads, n_tails, n_wings)`. Since the Dirichlet is the conjugate prior, we have a closed form expression for the posterior: $\text{Dir}(3, \alpha_1 + \texttt{n\_heads}, \alpha_2 + \texttt{n\_tails}, \alpha_3 + \texttt{n\_wings})$.

Problem 2a. Here you will finish the code for the function

`studio7_problem_2a(alphas, ns, probs)`

Use the conjugacy above to compute and return the value of the posterior pdf for the prior $\text{Dir}(3, \alpha_1, \alpha_2, \alpha_3)$ and outcomes `ns` at the input `probs`$=(p_1, p_2, p_3)$. The input `alphas` is also a vector $(\alpha_1, \alpha_2, \alpha_3)$.

We have supplied the `ddirichelt()` function in the file, as its not part of base R (though it is included in many packages). The normalization constant involves something called the gamma function. This function takes as input `x` and `alpha`, both vectors of *equal lengths* and returns the density of $\text{Dir}(\texttt{length}(\texttt{alpha}), \texttt{alpha})$ at `x`.

Problem 2b. Here you will finish the code for the function

`studio7_problem_2a(alphas, ns)`

In this problem, you will use a heat map to visualize the posteror pdf. Recall that Dirichlet pdf is only positive for vectors (p_1, p_2, p_3) which satisfy $p_1, p_2, p_3 \geq 0$ and $p_1 + p_2 + p_3 = 1$.

Thus, instead of plotting a heat map with a 3-dimensional domain (our imagination can only take us so far), we will draw a heat map for the first two values (p_1, p_2) and use the fact that $p_3 = 1 - p_1 - p_2$.

Remark: in the answers the heat map includes a triangle bordering the admissible region $p_1, p_2 \geq 0$ and $p_1 + p_2 \leq 1$. You do not have to include a similar triangle in your submission, but you are welcome to. The lines() or polygon() functions can prove handy.

Problem 2c. Here you will finish the code for the function

```
studio7_problem_2c()
```

As before, we will compare the shape of posterior with respect to different priors and outcomes.

Again, we start with *flat prior* and set $\alpha_1 = \alpha_2 = \alpha_3 = 1$. Draw the heatmap for the posterior when `n_heads` $= 1,$ `n_tails` $= 3$, and `n_wings` $= 7$. Proceed by increasing the number of tosses, *while keeping a fixed proportion*, first to $(2, 6, 14)$, then to $(3, 9, 21)$, and so on. Use the first cat statement to report and describe the shape of the posterior density.

Now repeat the same procedure with the *biased prior* $\alpha_1 = 7, \alpha_2 = 3$, and $\alpha_3 = 1$. Use the second cat statement to explain the differences between the two cases.

Testing your code

For each problem, we ran the problem function with certain parameters. You can see the function call and the output in studio7-test-answers.html. If you call the same function with the same parameters, you should get the same results as in studio7-test-answers.html – if there is randomness involved the answers should be close but not identical.

For your convenience, the file studio7-test.r contains all the function calls used to make studio7-test-answers.html.

Before uploading your code

1. Make sure all your code is in studio7.r. Also make sure it is all inside the functions for the problems.
2. Clean the environment and plots window.
3. Source the file.
4. Call each of the problem functions with the same parameters as the test file studio7-test-answers.html.
5. Make sure it runs without error and outputs just the answers asked for in the questions.
6. Compare the output to the answers given in studio7-test-answers.html.

Upload your code

Upload your code to Gradescope.

Leave the file name as studio7.r.

You can upload more than once. We will grade the last file you upload.

Due date

The goal is to upload your work by the end of class.

If you need extra time, you can upload your work any time before 9 PM ET the day after the studio.

Solutions uploaded: Solution code will be posted on at 10 PM the day after the studio.