PSYC 529 R Homework 4

Jacqueline Heitmann
25562334

```r
library(tidyverse)
```

```
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0
──
✔ dplyr     1.1.4     ✔ readr     2.1.5
✔ forcats   1.0.0     ✔ stringr   1.5.1
✔ ggplot2   3.5.1     ✔ tibble    3.2.1
✔ lubridate 1.9.4     ✔ tidyr     1.3.1
✔ purrr     1.0.2
── Conflicts ────────────────────────────────────────── tidyverse_conflicts()
──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all
conflicts to become errors
```

```r
words <- read.csv("../data/word_memory_data_summarized.csv")
words
```

```
       id             condition trials yes_responses prop_yes_resp
1    2002             Memorized     30            27          0.90
2    2002   Related_Unmemorized      6             6          1.00
3    2002 Unrelated_Unmemorized     36             0          0.00
4     LMB             Memorized     30            22          0.73
5     LMB   Related_Unmemorized      6             4          0.67
6     LMB Unrelated_Unmemorized     36             1          0.03
7   egh44             Memorized     30            21          0.70
8   egh44   Related_Unmemorized      6             1          0.17
9   egh44 Unrelated_Unmemorized     36             0          0.00
10 jh3487             Memorized     30            27          0.90
11 jh3487   Related_Unmemorized      6             6          1.00
12 jh3487 Unrelated_Unmemorized     36             0          0.00
13  jhm77             Memorized     30            17          0.57
14  jhm77   Related_Unmemorized      6             2          0.33
15  jhm77 Unrelated_Unmemorized     36             0          0.00
16 mam475             Memorized     30            24          0.80
17 mam475   Related_Unmemorized      6             1          0.17
```

```
18 mam475 Unrelated_Unmemorized     36            0           0.00
19 seb235             Memorized     30           30           1.00
20 seb235   Related_Unmemorized      6            6           1.00
21 seb235 Unrelated_Unmemorized     36            0           0.00
22  spp36             Memorized     30           22           0.73
23  spp36   Related_Unmemorized      6            4           0.67
24  spp36 Unrelated_Unmemorized     36            1           0.03
25 ss3998             Memorized     30           23           0.77
26 ss3998   Related_Unmemorized      6            5           0.83
27 ss3998 Unrelated_Unmemorized     36            0           0.00
28  ssg47             Memorized     30           25           0.83
29  ssg47   Related_Unmemorized      6            4           0.67
30  ssg47 Unrelated_Unmemorized     36            0           0.00
31  wz343             Memorized     30           26           0.87
32  wz343   Related_Unmemorized      6            1           0.17
33  wz343 Unrelated_Unmemorized     36            0           0.00
34 yl2456             Memorized     30           25           0.83
35 yl2456   Related_Unmemorized      6            3           0.50
36 yl2456 Unrelated_Unmemorized     36            0           0.00
37  yq222             Memorized     30           26           0.87
38  yq222   Related_Unmemorized      6            2           0.33
39  yq222 Unrelated_Unmemorized     36            0           0.00
40  zl543             Memorized     30           29           0.97
41  zl543   Related_Unmemorized      6            2           0.33
42  zl543 Unrelated_Unmemorized     36            0           0.00
43  zl627             Memorized     30           25           0.83
44  zl627   Related_Unmemorized      6            4           0.67
45  zl627 Unrelated_Unmemorized     36            5           0.14
```

1. Looking at the data table (you don't need to import it as an R data frame). Define the following as variables in R:

(a) (21/2 points) N: the number of relevant data points (trials)

```
memorized_ds <- words %>%
  filter(condition == "Memorized")

N <- sum(memorized_ds$trials)

N
```

```
[1] 450
```

(b) (21/2 points) z: the number of "yes" responses on relevant trials

```
z <- sum(memorized_ds$yes_responses)
z
```

```
[1] 369
```

(c) (21/2 points) a: the first prior beta hyperparameter
(d) (21/2 points) b: the second prior beta hyperparameter

```
# beta(4,4)
a <- 4
b <- 4
```

2. (10 points) Use these variables to compute the posterior beta distribution hyperparameters (call them a post and b post).

```
a_post <- a + z # prior + successes
b_post <- b + N - z # prior + failures

a_post
```

```
[1] 373
```

```
b_post
```

```
[1] 85
```

3. We now make a combined plot of the prior and posterior distributions. Hint: Use the file "beta conjugate prior demo" on Canvas as a template for doing all of this. That will make this part of the homework much easier and faster

data frame for plotting prior distribution:

```
prior_plot_df = data.frame(theta = seq(from = 0, to = 1, length.out = 200),
                           distribution = "prior") %>%
  mutate(p_theta = dbeta(theta, shape1 = a, shape2 = b))
```

data frame for plotting the posterior distribution:

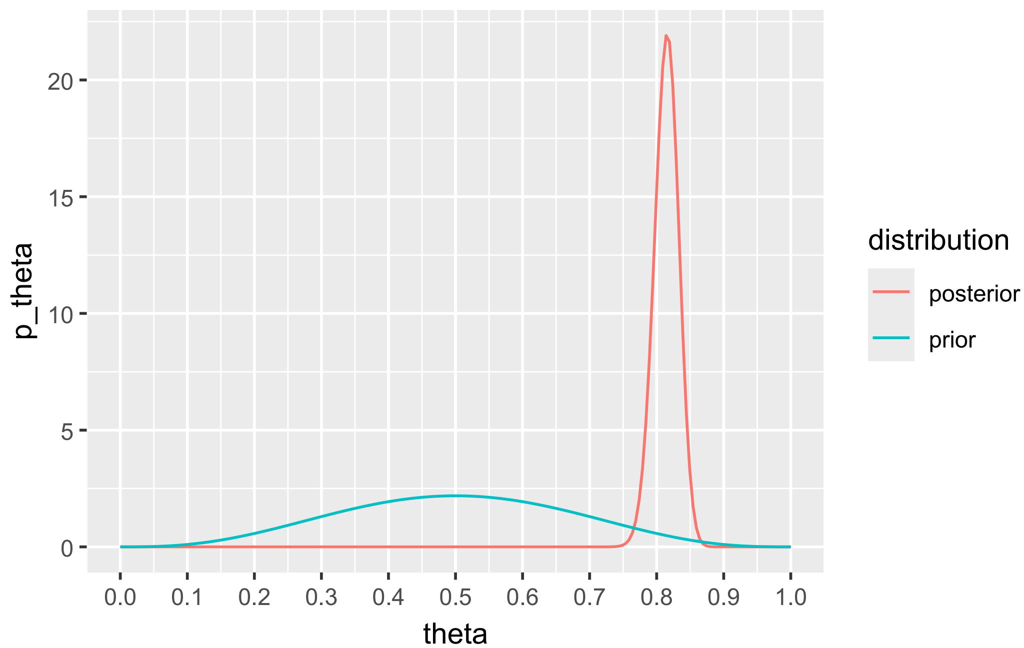
```
post_plot_df = data.frame(theta = seq(from = 0, to = 1, length.out = 200),
                          distribution = "posterior") %>%
  mutate(p_theta = dbeta(theta, shape1 = a_post, shape2 = b_post))
```

combine prior and posterior plotting data frames:

```
combo_plot_df = bind_rows(prior_plot_df,
                          post_plot_df)
```

combined plot:

```
combo_plot_df %>% ggplot(aes(x = theta, y = p_theta, color = distribution)) +
  geom_line() +
  scale_x_continuous(breaks = seq(from = 0, to = 1, by = 0.1)) # more marks on
x axis
```



4. The expected value (mean) and mode (most probable value) are two ways to summarize a probability distribution. Given the hyperparameters a and b of a beta distribution, you can compute the distribution's mode (ω) and expected value (i.e. mean, μ) using formulas given in the textbook (page 129) and summarized in Cliff's notes for chapter 6 (these are on Canvas). Use these formulas and your R variables to compute and display the following:

(a) (10 points) $E[\theta]$ (the prior mean of θ)

```
e_prior <- a / a + b
e_prior
```

```
[1] 5
```

(b) (10 points) the prior mode of θ.

```r
mode_prior <- (a - 1) / (a + b - 2)
mode_prior
```

```
[1] 0.5
```

(c) (10 points) $E[\theta|N, z]$ (the posterior mean of θ)

```r
e_post <- a_post / (a_post + b_post)
e_post
```

```
[1] 0.8144105
```

(d) (10 points) the posterior mode of θ

```r
mode_post <- (a_post - 1) / (a_post + b_post - 2)
mode_post
```

```
[1] 0.8157895
```

5. You can use the R function qbeta to find quantiles of a beta distribution.

For example, let's suppose we want to find the median (0.5 quantile, i.e. 50th percentile) of a beta(3, 2) distribution. Then we use the following code: qbeta(p = 0.5, shape1 = 3, shape2 = 2)

This tells use that, assuming a = 3 and b = 2, we have $p(\theta \leq 0.6142724) = 0.5$. In other words the median of a beta(3, 2) distribution is about 0.61.

Use qbeta to compute the following: (a) (5 points) The 5th percentile (i.e. 0.05 quantile) of the prior distribution

```r
qbeta(p = 0.05, a, b)
```

```
[1] 0.2253216
```

the probability of theta being less than or equal to 0.2253216 is roughly 5% in the prior distribution.

(b) (5 points) The 5th percentile of the posterior distribution

```r
qbeta(p = 0.05, a_post, b_post)
```

```
[1] 0.7837966
```

The probability of theta being less than or equal to 0.78 is roughly 0.05.

(c) (5 points) The 95th percentile (i.e. 0.95 quantile) of the prior distribution

```
qbeta(p = 0.95, a, b)
```

```
[1] 0.7746784
```

(d) (5 points) The 95th percentile of the posterior distribution

```
qbeta(p = 0.95, a_post, b_post)
```

```
[1] 0.8434619
```

Note: We can use these percentiles to construct 90% credible intervals for θ: the probability that θ is between the 5th and 95th percentiles of the distribution is 90%. For the sake of time we won't go into this in detail, but ask me if you have any questions about it (it's very similar to how you derive frequentist confidence intervals, except using the prior or posterior distribution over θ instead of the sampling distribution of some estimator).