Homework 2

PSTAT 115, 2024

Due on February 11, 2024 at 11:59 pm

Note: If you are working with a partner, please submit only one homework per group with both names and whether you are taking the course for graduate credit or not. Submit your Rmarkdown (.Rmd) and the compiled pdf on Gauchospace.

1. Trend in Same-sex Marriage

A 2017 Pew Research survey found that 10.2% of LGBT adults in the U.S. were married to a same-sex spouse. Now it's the 2020s, and Bayard guesses that π, the percent of LGBT adults in the U.S. who are married to a same-sex spouse, has most likely increased to about 15% but could reasonably range from 10% to 25%.

1a. Identify a Beta model that reflects Bayard's prior ideas about π by specifying the parameters of the Beta, α and β.

```r
alpha <- NULL # YOUR CODE HERE
beta <- NULL # YOUR CODE HERE
```

```r
. = ottr::check("tests/q1a.R")
```

1b. Bayard wants to update his prior, so he randomly selects 90 US LGBT adults and 30 of them are married to a same-sex partner. What is the posterior model for π?

```r
posterior_alpha <- NULL # YOUR CODE HERE
posterior_beta <- NULL # YOUR CODE HERE
```

1c. Use R to compute the posterior mean and standard deviation of π.

```r
posterior_mean <- NULL # YOUR CODE HERE
posterior_sd <- NULL # YOUR CODE HERE

print(sprintf("The posterior mean is %f", posterior_mean))
print(sprintf("The posterior sd is %f", posterior_sd))
```

1d. Does the posterior model more closely reflect the prior information or the data? Explain your reasoning. Hint: in the recorded lecture we showed a special way in which we can write the posterior mean in a Beta-Binomial model. How can this help? Check the lectures notes.

Type your answer here, replacing this text.

```r
# YOUR CODE HERE
```

2. Cancer Research in Laboratory Mice

A laboratory is estimating the rate of tumorigenesis (the formation of tumors) in two strains of mice, A and B. They have tumor count data for 10 mice in strain A and 13 mice in strain B. Type A mice have been well studied, and information from other laboratories suggests that type A mice have tumor counts that are approximately Poisson-distributed. Tumor count rates for type B mice are unknown, but type B mice are related to type A mice. Assuming a Poisson sampling distribution for each group with rates θ_A and θ_B. Based on previous research you settle on the following prior distribution:

$$\theta_A \sim \text{gamma}(120, 10),\ \ \theta_B \sim \text{gamma}(12, 1)$$

2a. Before seeing any data, which group do you expect to have a higher average incidence of cancer? Which group are you more certain about a priori? You answers should be based on the priors specified above.

Type your answer here, replacing this text.

2b. After you the complete of the experiment, you observe the following tumor counts for the two populations:

$$y_A = (12, 9, 12, 14, 13, 13, 15, 8, 15, 6)$$
$$y_B = (11, 11, 10, 9, 9, 8, 7, 10, 6, 8, 8, 9, 7)$$

Compute the posterior parameters, posterior means, posterior variances and 95% quantile-based credible intervals for θ_A and θ_B. Same them in the appropriate variables in the code cell below. You do not need to show your work, but you cannot get partial credit unless you do show work.

```
. = ottr::check("tests/q2b.R")
```

2c. Compute and plot the posterior expectation of θ_B given y_B under the prior distribution gamma$(12 \times n_0, n_0)$ for each value of $n_0 \in \{1, 2, ..., 50\}$. As a reminder, n_0 can be thought of as the number of prior observations (or pseudo-counts).

```
# YOUR CODE HERE

posterior_means = NULL # YOUR CODE HERE

# YOUR CODE HERE
```

```
. = ottr::check("tests/q2c.R")
```

2d. Should knowledge about population A tell us anything about population B? Discuss whether or not it makes sense to have $p(\theta_A, \theta_B) = p(\theta_A) \times p(\theta_B)$.

Type your answer here, replacing this text.

3. Soccer World cup

Let λ be the expected number of goals scored in a Women's World Cup game. We'll analyze λ by the following a Y_i is the observed number of goals scored in a sample of World Cup games:

$$Y_i|\lambda \overset{ind}{\sim} \text{Pois}(\lambda)$$

You and your friend argue about a more reasonable prior for λ. You think that $p_1(\lambda)$ with a gamma$(8, 2)$ density is a reasonable prior. Your friend thinks that $p_2(\lambda)$ with a gamma$(2, 1)$ density is a reasonable prior distribution. Consider the case in which each of you are equally credible in your prior assessments and so you combine your prior distributions into a mixture prior with equal weights: $p(\lambda) = 0.5 * p_1(\lambda) + 0.5 * p_2(\lambda)$.

3a. Which of you thinks more goals will be scored on average? Which of you is more confident in that assessment a priori?

Type your answer here, replacing this text.

3b. Plot the combined prior density, $p(\lambda)$, that you and your friend have created.

```
# YOUR CODE HERE
```

3c. Why might the Poisson model be a reasonable model for our data Y_i? In what ways might this model for Y_i be too simple?

3c. The `wwc_2019_matches` data in the *fivethirtyeight* package includes the number of goals scored by the two teams in each 2019 Women's World Cup match. Create a histogram of the number of goals scored per game. What is the maximum likelihood estimate for the expected number of goals scored in a game? You do not need to show your work for computing the MLE.

```
library(fivethirtyeight)
data("wwc_2019_matches")
wwc_2019_matches <- wwc_2019_matches %>%
  mutate(total_goals = score1 + score2)

## This is your y_i
total_goals <- wwc_2019_matches$total_goals
```

```
soccer_mle <- NULL # YOUR CODE HERE
```

3d. Write the posterior distribution up to a proportionality constant by multiplying the likelihood and the combined prior density created by you and your friend. Plot this unnormalized posterior distribution and add a vertical line at the MLE computed in the previous part. *Warning:* be very careful about what constitutes a proportionality constant in this example.

Type your answer here, replacing this text.

```
# YOUR CODE HERE
```

3e. Based on the plot above would you say that the prior had a large impact on conclusions or only a small one? Reference pseudo-counts and the proposed prior to argue why it makes sense that the prior did or did not have a big effect.

Type your answer here, replacing this text.