Homework 2

PSTAT 115, 2023

Due on Sunday February 5, 2023 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.

```
alpha <- 11
beta <- 63
```

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

```
##
## All tests passed!
```

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?

```
posterior_alpha <- 41
posterior_beta <- 123
```

```
. = ottr::check("tests/q1b.R")
```

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

```
posterior_mean <- posterior_alpha / (posterior_alpha + posterior_beta)
posterior_sd <- sqrt((posterior_alpha * posterior_beta) / (((posterior_alpha + posterior_beta) ** 2) *

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

```
## [1] "The posterior mean is 0.250000"
```

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

```
## [1] "The posterior sd is 0.033710"
```

```
. = ottr::check("tests/q1c.R")
```

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.

The posterior model more closely reflects the data. We added 90 new data points and then updated our prior distribution. We can also see that our posterior mean of 0.25 with a standard deviation of 0.03 more closely reflects the $30/90 = 33.33\%$ of adults in our new data that were same-sex marriage. Therefore our prior is being updated to more accurately model the new data.

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.

I would expect Group A to have a higher average incidence of cancer. This is because based off the given distributions for each group, they have the same mean, but Group A has a much higher shape and rate parameter than Group B, so we can expect there to be more incidences of cancer from Group A.

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.

```
## [1] "Posterior mean of theta_A 11.85"
## [1] "Posterior variance of theta_A 0.59"
## [1] "Posterior mean of theta_B 8.93"
## [1] "Posterior variance of theta_B 0.64"
## [1] "Posterior 95% quantile for theta_A is [10.61, 13.14]"
## [1] "Posterior 95% quantile for theta_B is [7.66, 10.28]"
```

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

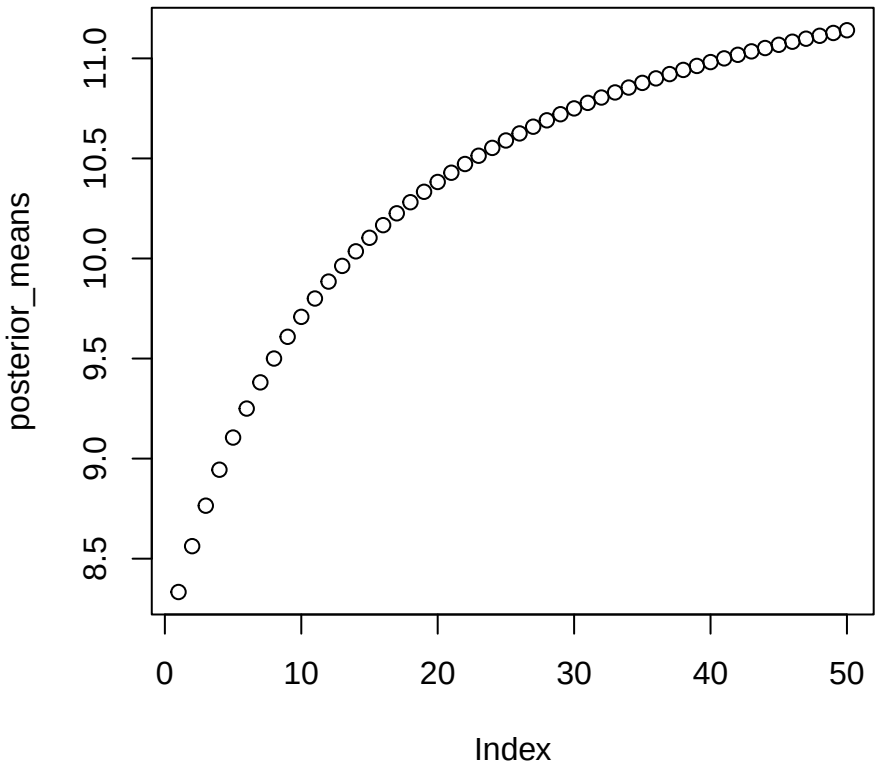
```
##
## All tests passed!
```

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).

```
alpha <- 12 * seq(1,50,1) + sum(yB)
beta <- seq(2,51,1) + length(yB)

posterior_means = alpha / beta

plot(posterior_means)
```




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

```
## Test q2c - 1 passed
##
##
## Test q2c - 2 passed
```

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

Yes, the knowledge about population A should tell us something about population B. Since population B is related, but not directly related, the 2 groups could be considered independent. For example, humans are monkeys are related, but we consider them independent groups. However, the case could also be made that they are not independent, but rther dependent on each other. There could be a lot of shared traits / things we are looking for which could make grouping population A and B into a single group better.

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. You decide that each of you are equally credible in your prior assessments and so you combine your prior distributions into a mixture prior with equal weights: $p(\lambda) = 0.5 * p_1(\lambda) + 0.5 * p_2(\lambda)$

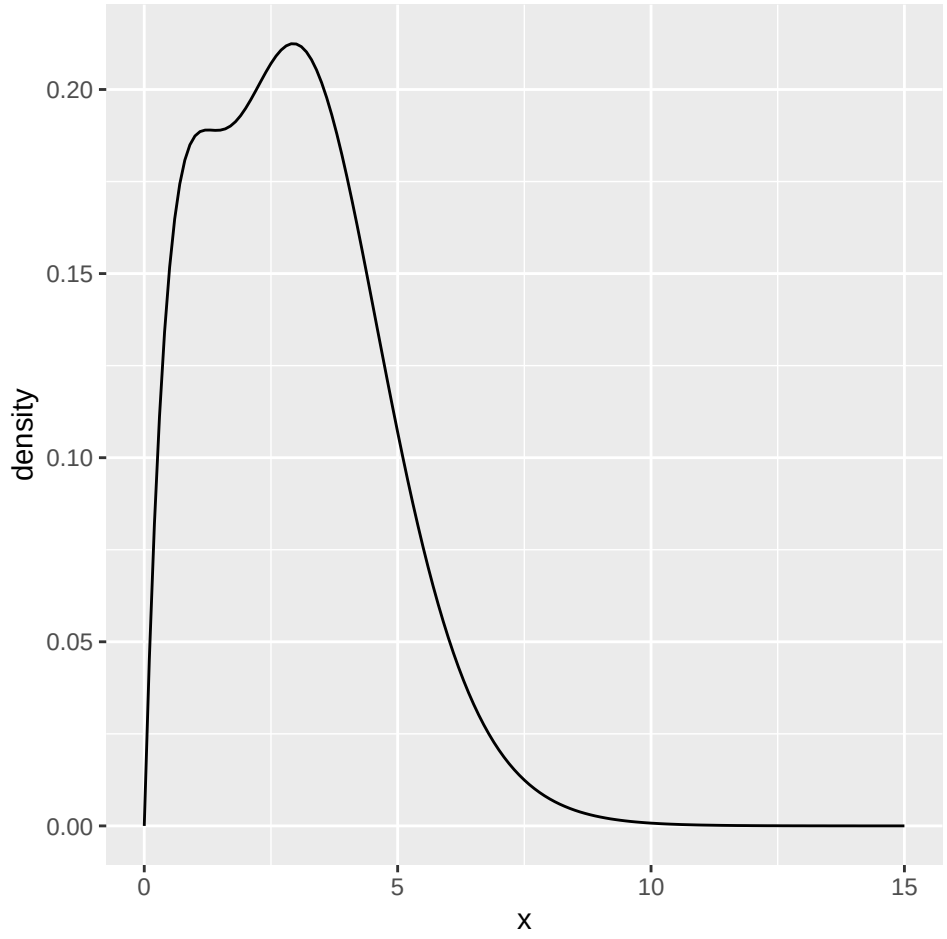
3a. Which of you thinks more goals will be scored on average? Which of you is more confident in that assessment a priori?

I believe that more goals will be scored on average. Both distributions have the same mean, but we have a much higher rate and shape parameters which leads the graph to be less skewed and therefore possible more goals on average. In terms of confident in the assessment a priori, we are more confident.

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

```
x <- seq(0,15,0.1)
density <- (0.5 * dgamma(x,8,2)) + (0.5 * dgamma(x,2, 1))

df <- data.frame(x = x, vals = density)
plot <- ggplot(data = df, aes(x = x, y = density)) + geom_line()
plot
```



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?

The poisson model might be a reasonable model for our data Y_i because it is a rate of goals being scored. Each goal is independent of each other and can be said to occur at certain rate so a poisson model could fit our data well. It also has one parameter of λ so it is simple, but that is also why it might be too simple. Since there is only one parameter, it doesn't capture a multitude of other factors. For example different teams might have different skill level, the location may have different temperatures, the time of day could even affect it. A game in broad daylight when the players are fresh can be different than at night with floodlights when players have already been up for a while.

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.

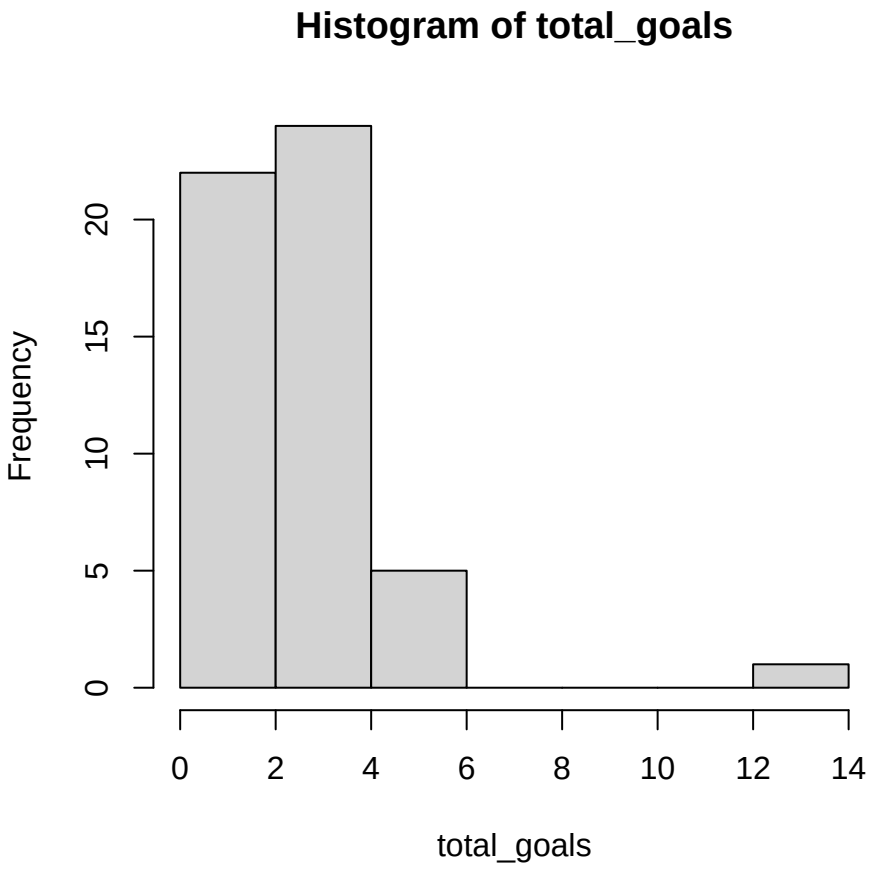
```
library(fivethirtyeight)
```

```
## Some larger datasets need to be installed separately, like senators and
## house_district_forecast. To install these, we recommend you install the
## fivethirtyeightdata package by running:
## install.packages('fivethirtyeightdata', repos =
## 'https://fivethirtyeightdata.github.io/drat/', type = 'source')
```

```
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

hist(total_goals)
```




```
soccer_mle <- mean(wwc_2019_matches$total_goals)
```

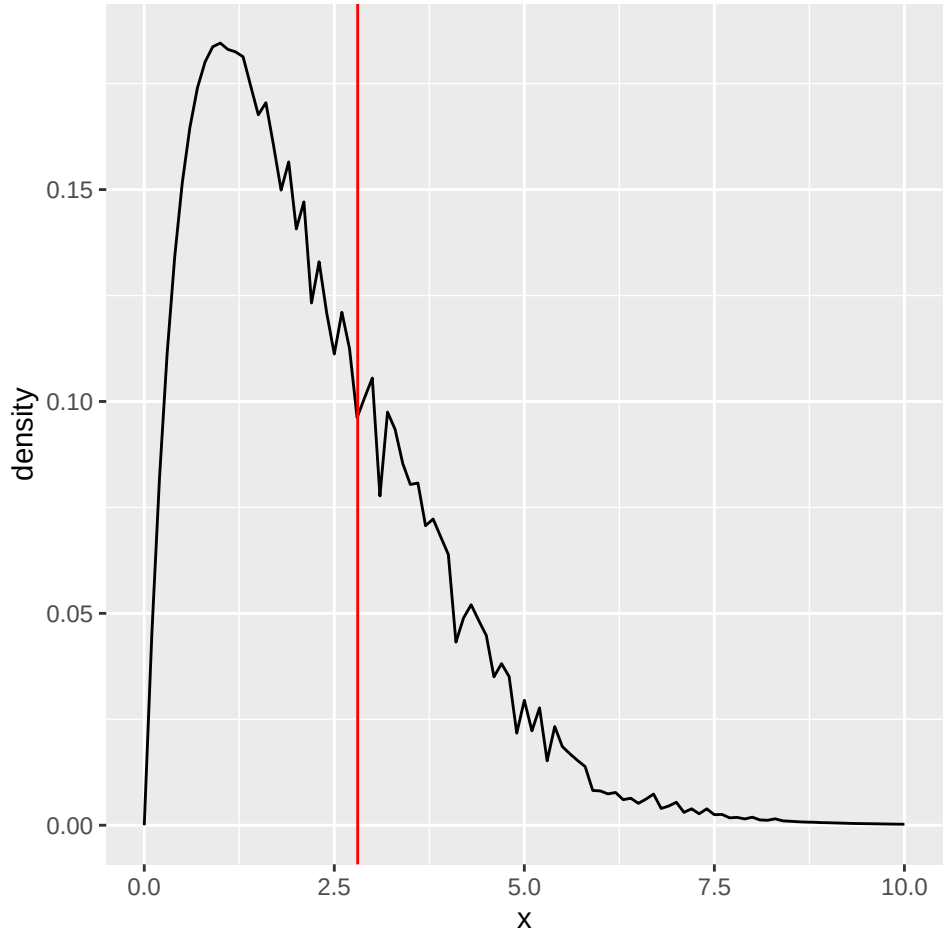
3d. Write the posterior distribution up to a proportionality constant by multiplying the likelihood and the combined prior density created by you and your friend. Plot this unnormalized posterior distribution and add a vertical line at the MLE computed in the previous part. *Warning:* be very careful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

$$\prod_{i=1}^{n} \frac{\lambda^{y_i} e^{-\lambda}}{y_i!} \times \frac{1}{2}\frac{256}{\Gamma(8)} y^7 e^{-2y} + \frac{1}{2}\frac{ye^{-y}}{\Gamma(2)} = \prod_{i=1}^{n} \frac{\lambda^{y_i} e^{-\lambda}}{y_i!} \times \frac{128}{\Gamma(8)} y^7 e^{-2y} + \frac{ye^{-y}}{2} \propto \lambda^{\Sigma y_i} e^{-\lambda n} \times \frac{128}{\Gamma(8)} y^7 e^{-2y} + \frac{ye^{-y}}{2}$$

```
x <- seq(0,10,0.1)
pois <- dpois(total_goals,x)
g1 <- dgamma(x,8,rate =2)
g2 <- dgamma(x,2,rate = 1)

post <- pois * 0.5 * g1 + 0.5 * g2

df <- data.frame(x = x, density = post)
plot <- ggplot(data = df, aes(x = x, y = density)) + geom_line()
plot + geom_vline(xintercept = (soccer_mle), color = 'red')
```



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.

I would say that prior had a small impact on conclusions. Thhis is because the posterior graph above resembles more of the poisson likelihood than our prior combined model. I would say this because the data must be overwhelmingly "poisson-like" and therefore our model only made a small impact. We also had large pseudo counts of 8-2 which looking at the data again is very large. Therefore, since it doesn't match well with the data, the distribution made from it wouldn't have that much of an impact either.