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 <- 15
beta <- 75
```

```
. = 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?

```
n <- 90
y <- 30
posterior_alpha <- alpha + y
posterior_beta <- n - y + beta
```

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

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

```
# TO DO: NEED TO CHECK
posterior_mean <- posterior_alpha / (posterior_alpha + posterior_beta)
posterior_sd <- sqrt((posterior_alpha*posterior_beta) / (((posterior_alpha + posterior_beta)^2)*(poster

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 0.032186"
```

```
. = 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.

In a Beta-Binomial Model the Posterior mean can be represented as follows:
$\omega * \hat{\theta}_{MLE} + (1 - \omega) * \hat{\theta}_{priormean}$

The posterior model more closely reflects the data seeing as the posterior mean is 0.25, and the sample mean was 0.3. While the prior mean was 0.15, which clearly suggests the posterior model more closely reflects the 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.

Before seeing any data, I would expect both groups to have the same average of incidence of cancer. This is because the mean for a gamma distribution is a/b, and a/b for both θ_A and θ_B are the same in this scenario. I am more certain about a priori for the Type A group because as stated above, the Typa A group is well studied, while tumor count rates for Type B are unknown.

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 [10.39, 13.41]"

## [1] "Posterior 95% quantile for theta_B is [7.43, 10.56]"
```

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

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

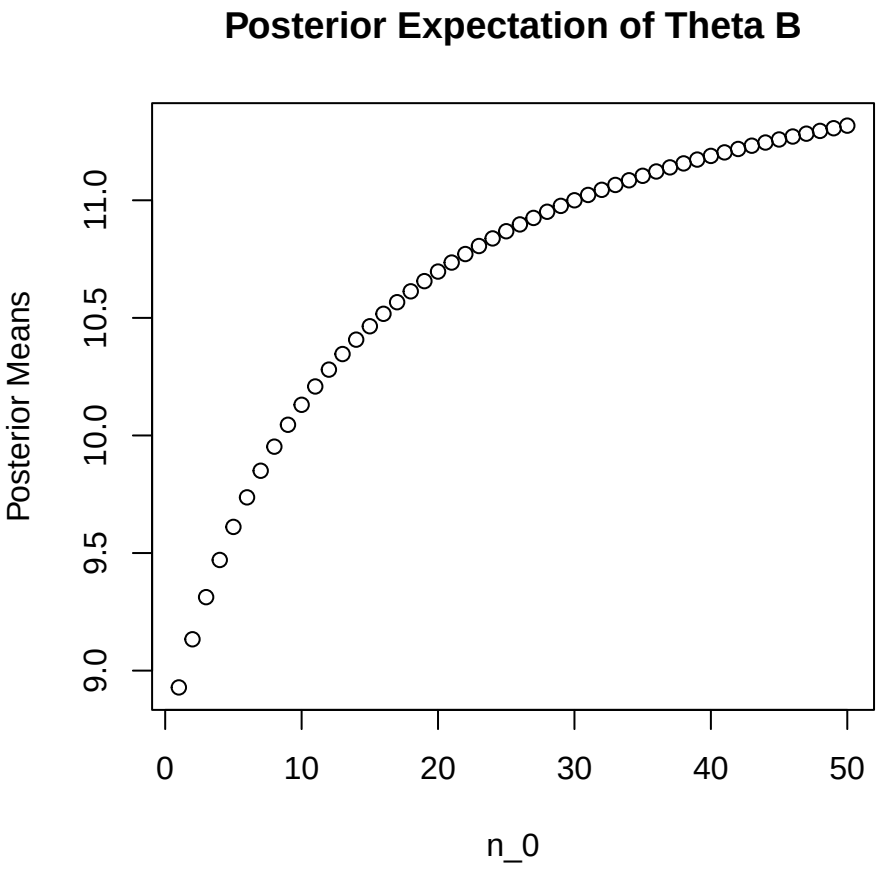
2c. Compute and plot the posterior expectation of θ_B given y_B under the prior distribution gamma($12 \times n_0, n_0$) for each value of $n_0 \in \{1, 2, ..., 50\}$. As a reminder, n_0 can be thought of as the number of prior observations (or pseudo-counts).

```r
posterior_means <- rep(NA, 50)
for(i in 1:50){
  # Prior parameters here
  alpha_B = 12*i
  beta_B = i

  # Posterior parameters here
  alpha_B_posterior = sum(yB) + alpha_B
  beta_B_posterior = length(yB) + beta_B

  ## Posterior mean
  posterior_means[i] <- alpha_B_posterior / beta_B_posterior
}

x <- seq(1, 50, by = 1)
plot(x, posterior_means, xlab = 'n_0', ylab = 'Posterior Means', main = 'Posterior Expectation of Theta
```




```r
. = 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 $p(\theta_A, \theta_B) = p(\theta_A) \times p(\theta_B)$.

Considering both population A and B are mice and are related (as stated in the question), knowledge about population A should give us some value about population B. For example, if we have information about population A, it wouldn't be totally unreasonable to use that information as a prior for population B. However based on my credible intervals there is very little overlap, so there is reason to think that they might be independent. Therefore using the equation $p(\theta_A, \theta_B) = p(\theta_A) \times p(\theta_B)$ would not be ruled out. In addition, if we look at the graph from 2c, as n_0 increases, the posterior mean of $theta_B$ approaches the posterior mean of $theta_A$ only when the prior alpha and beta are 360 and 30, respectively, which is extremely narrow (low variance) distribution, such that it is not affected much by the data.

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:

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

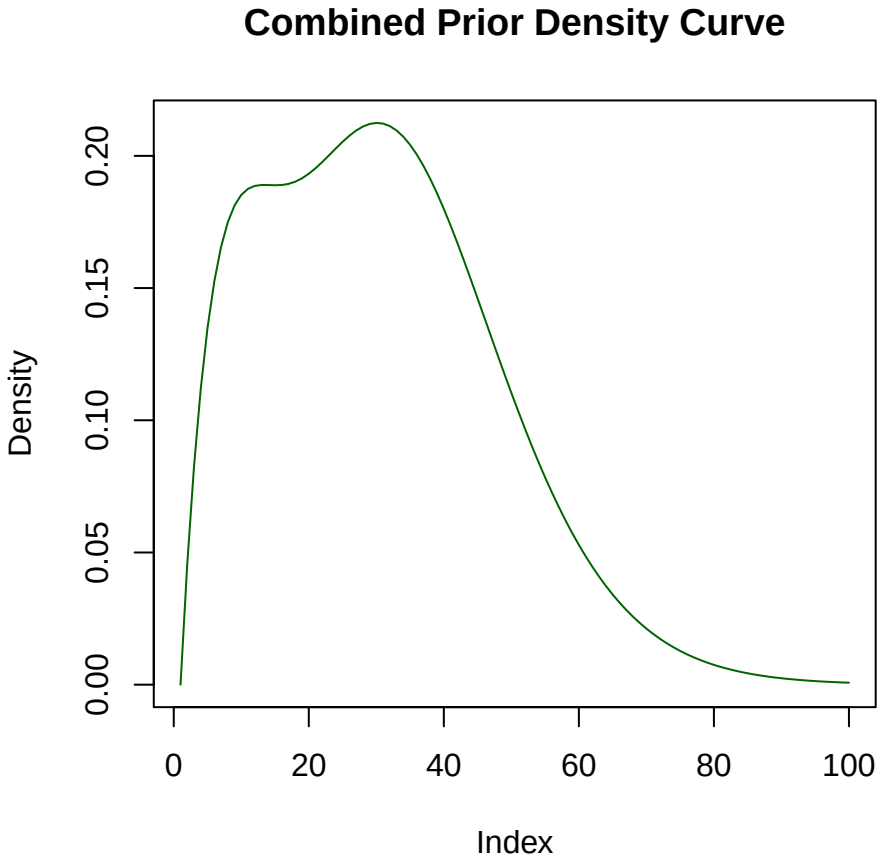
You and your friend argue about a more reasonable prior for λ. You think that $p_1(\lambda)$ with a gamma$(8, 2)$ density is a reasonable prior. Your friend thinks that $p_2(\lambda)$ with a gamma$(2, 1)$ density is a reasonable prior distribution. You decide that each of you are equally credible in 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?

Among my friend and I, I think more goals will be scored on average because I am predicted a gamma$(8, 2)$ which has a mean of 4 goals, while my friend predicts gamma$(2, 1)$ which only has a mean of 2 goals. As stated above, we both decide that we are equally credible and thus implying equal confidence in our assessment of a priori.

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

```
x1 <- seq(0, 10, length = 100)
prior_d <- 0.5*dgamma(x1, 8, 2) +  0.5*dgamma(x1, 2, 1)
plot(prior_d, type = 'l', col = 'darkgreen', main = 'Combined Prior Density Curve', ylab = 'Density') #
```



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?

A poisson model would be a reasonable model for our data Y_i because we are trying to model the number of soccer goals scored in a game, which fits the poisson distribution model. A poisson model would be too simple in this case, because we would be assuming that the rate at which goals are scored is constant. Having a constant rate at which goals are scored may not be the case when teams of varying skill level play against each other.

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.

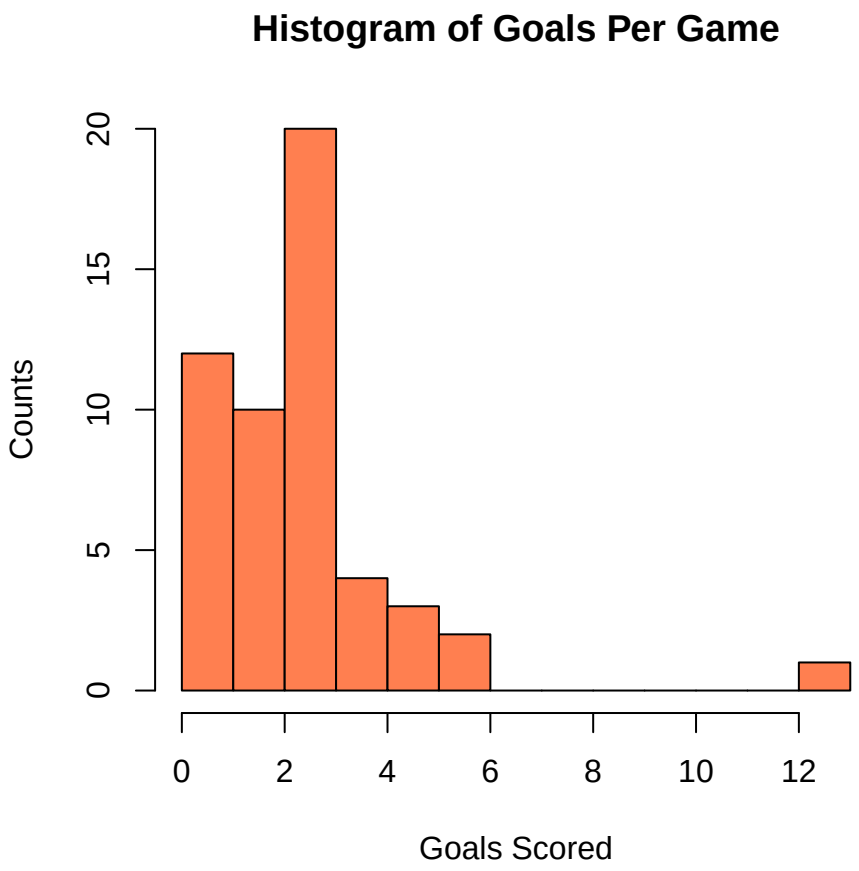
```
library(fivethirtyeight)
```

```
## 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
```

```r
# Histogram
hist(total_goals, xlab = 'Goals Scored', ylab = 'Counts', main = 'Histogram of Goals Per Game', col = '
```




```r
soccer_mle <- sum(total_goals) / length(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.

$$
\begin{aligned}
f(\lambda|y) &\propto f(\lambda)L(\lambda|y) \\
&= \frac{r^s}{\gamma(s)}\lambda^{s-1}e^{-r\lambda} \times \frac{\lambda^{\Sigma y_i}e^{-n\lambda}}{\prod y_i!} \\
&\propto \lambda^{s-1}e^{-r\lambda} \times \lambda^{\Sigma y_i}e^{-n\lambda}
\end{aligned}
$$

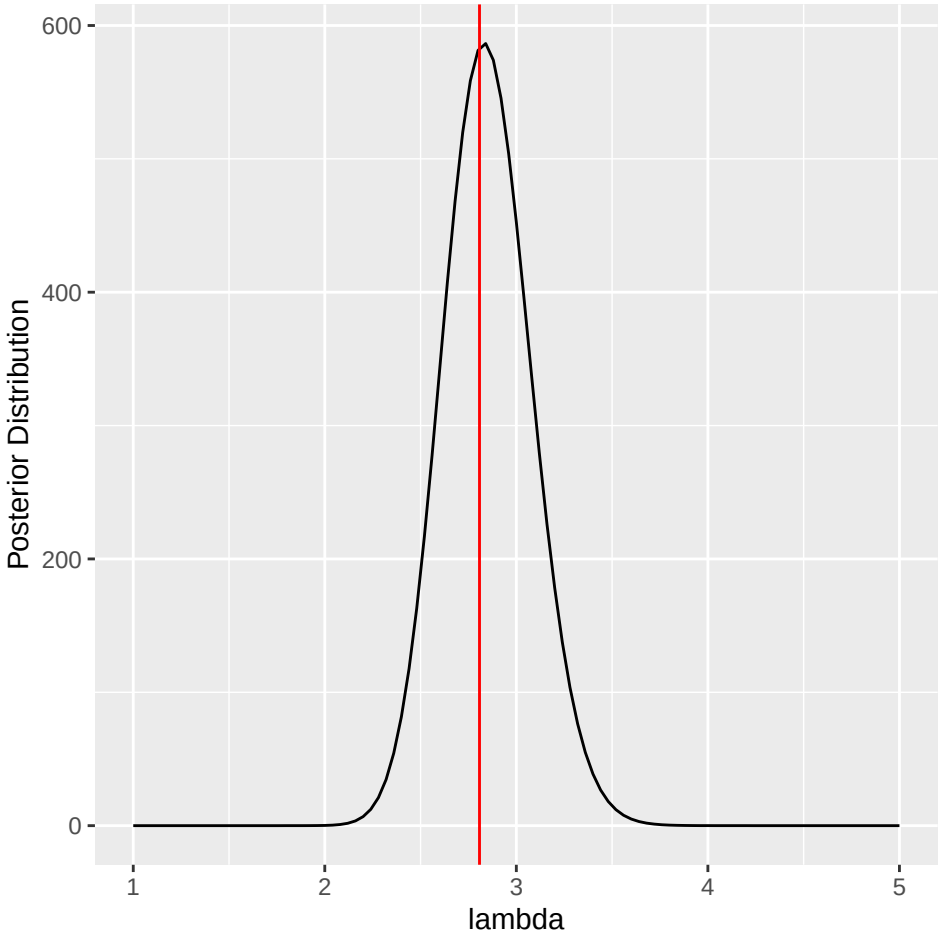
$$
with\ 0.5gamma(8,2) + 0.5gamma(2,1)
$$

$$
\begin{aligned}
&(\lambda^7 e^{-2\lambda} + \lambda e^{-\lambda})(\lambda^{\Sigma y_i}e^{-n\lambda}) \\
&= \lambda^{7+\Sigma y_i}e^{-\lambda(n+2)} + \lambda^{\Sigma y_i+1}e^{-\lambda(n+1)}
\end{aligned}
$$

```r
n <- length(total_goals)
y <- total_goals
post_dist <- function(lambda){
  lambda^(7 + sum(y))*exp(-lambda*(2 + n)) + lambda^(1 + sum(y))*exp(-lambda*(1 + n))
```

```
}

ggplot(data = data.frame(lambda = 0), mapping = aes(x=lambda)) + stat_function(fun = post_dist) +
     xlim(1,5) + scale_y_continuous(name = "Posterior Distribution") + geom_vline(xintercept = soccer_m
```



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.

Based on this graph, the MLE is very close to the actual peak of the posterior distribution. This means that the Likelihood function has a higher weight and is more impactful while the prior has much less impact, and it pulled the posterior distribution distribution just a little bit towards the mean of the two priors. The mean of the two priors is the mean of 2 and 4.