Homework 2

Carly Greutert

February 5, 2023

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 <- 10.2 # previous percent
beta <- 57.8 # want to equal .15, beta = 17/3*alpha
```

```
. = 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 <- 40.2 # 30 success + prior alpha
posterior_beta <- 117.8 # (90-30) + prior beta
```

```
. = 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*
                                                       (posterior_alpha+posterior_beta+1)))

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

```
## [1] "The posterior mean is 0.254430"
```

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

```
## [1] "The posterior sd is 0.034541"
```

```
. = 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.

```
data_weight <-90/(90+alpha+beta)
prior_weight <-(alpha+beta)/(90+alpha+beta)
data_weight
```

```
## [1] 0.5696203
```

```
prior_weight
```

```
## [1] 0.4303797
```

The posterior model reflects the data more closely since the data is giving more weight to it (0.5696 versus 0.4303). By employing the special way, we are able to see how much each is being considered when calculating the posterior mean.

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 expect group A to have a higher average incidence of cancer since it has a larger shape hyperparameter (120>12) and a larger rate parameter (10>1).

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 8.79"
```

```
## [1] "Posterior variance of theta_B 0.63"
```

```
## [1] "Posterior 95% quantile for theta_A is [10.39, 13.41]"
```

```
## [1] "Posterior 95% quantile for theta_B is [7.3, 10.4]"
```

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

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

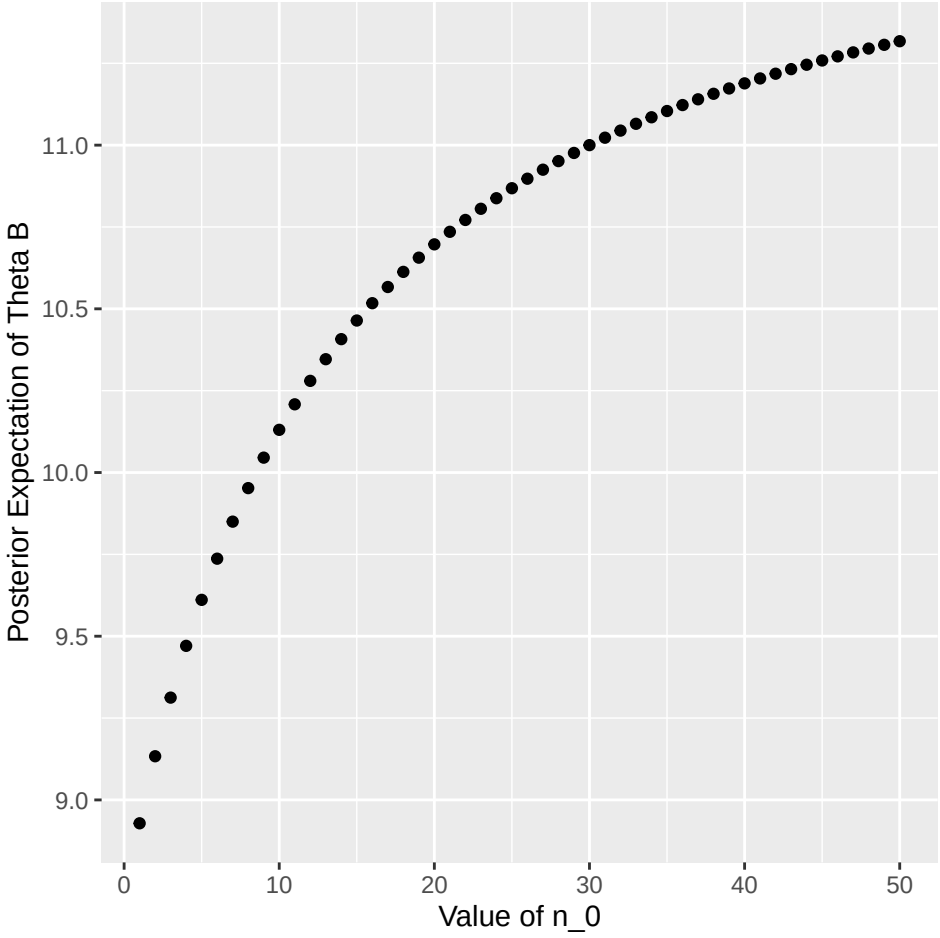
2c. Compute and plot the posterior expectation of θ_B given y_B under the prior distribution gamma$(12 \times n_0, n_0)$ for each value of $n_0 \in \{1, 2, ..., 50\}$. As a reminder, n_0 can be thought of as the number of prior observations (or pseudo-counts).

```
n0 <- 50

posterior_means <- rep(NA, n0)

for (i in 1:n0) {
  alphasim_prior <- 12 * i
  betasim_prior <- i
  alphasim_post <- sum(yB) + alphasim_prior
  betasim_post <- length(yB) + betasim_prior
  sim_mean <- alphasim_post/betasim_post
  posterior_means[i] = sim_mean
}

ggplot(tibble(posterior_means), aes(x= 1:n0, y = posterior_means)) +
  xlab("Value of n_0") + ylab("Posterior Expectation of Theta B") +
  geom_point()
```




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

Because the problem states that A and B mice are related, knowledge about population A should give us some information about population B. I think it would make sense to treat θ_A and $theta_B$ independently since they are two separate strains of mice, and one of them having a tumor does not influence whether one in another group does too, but incorporating our prior knowledge may lead to a more helpful model where we treat them as dependent populations.

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:

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

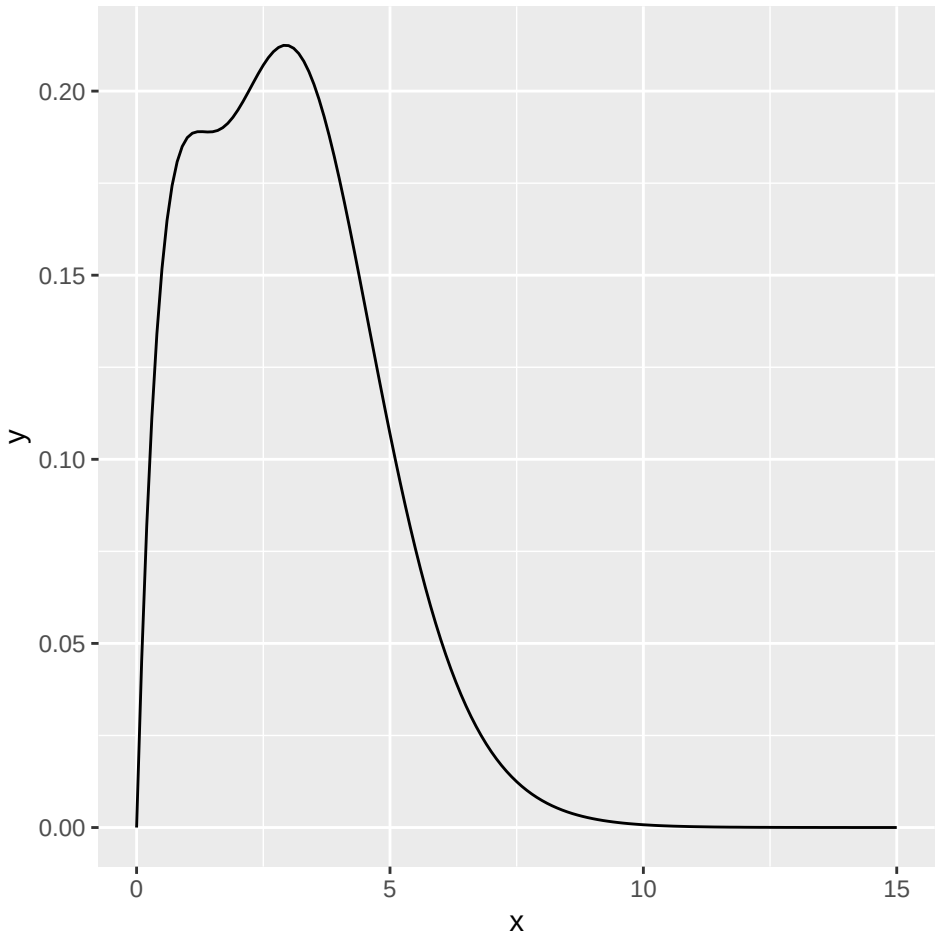
You and your friend argue about a more reasonable prior for λ. You think that $p_1(\lambda)$ with a gamma$(8, 2)$ density is a reasonable prior. Your friend thinks that $p_2(\lambda)$ with a gamma$(2, 1)$ density is a reasonable prior distribution. You decide that each of you are equally credible in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

The person anticipating the prior should be gamma$(8, 2)$ is likely expecting more goals to be scored on average since the expected value is 0.8, which is a higher rate than 0.667 once put in the poisson distribution. Also person one has a higher shape and rate. I think gamma$(2, 1)$ is a better assessment of the prior, but the person assuming gamma$(8, 2)$ is likely more confident.

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

```
p1 <- dgamma(x = seq(0,15,0.1), 8, 2)
p2 <- dgamma(x = seq(0,15,0.1), 2, 1)
p <- 0.5 * p1 + 0.5 * p2
ggplot(data.frame(x = seq(0,15,0.1), y = p), aes(x, y)) + geom_line()
```



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 is a reasonable model for this case because it is counting the rate of the number of goals scored in a certain period of time. It may too simple, though, because there are lots of factors that may influence how many goals are scored. For instance, we assume the games are independent, but if a team is on a winning or losing streak, this may affect the model. Weather and location (home or away) may also affect the goal rate. These and more are not taken into account with just a poisson model.

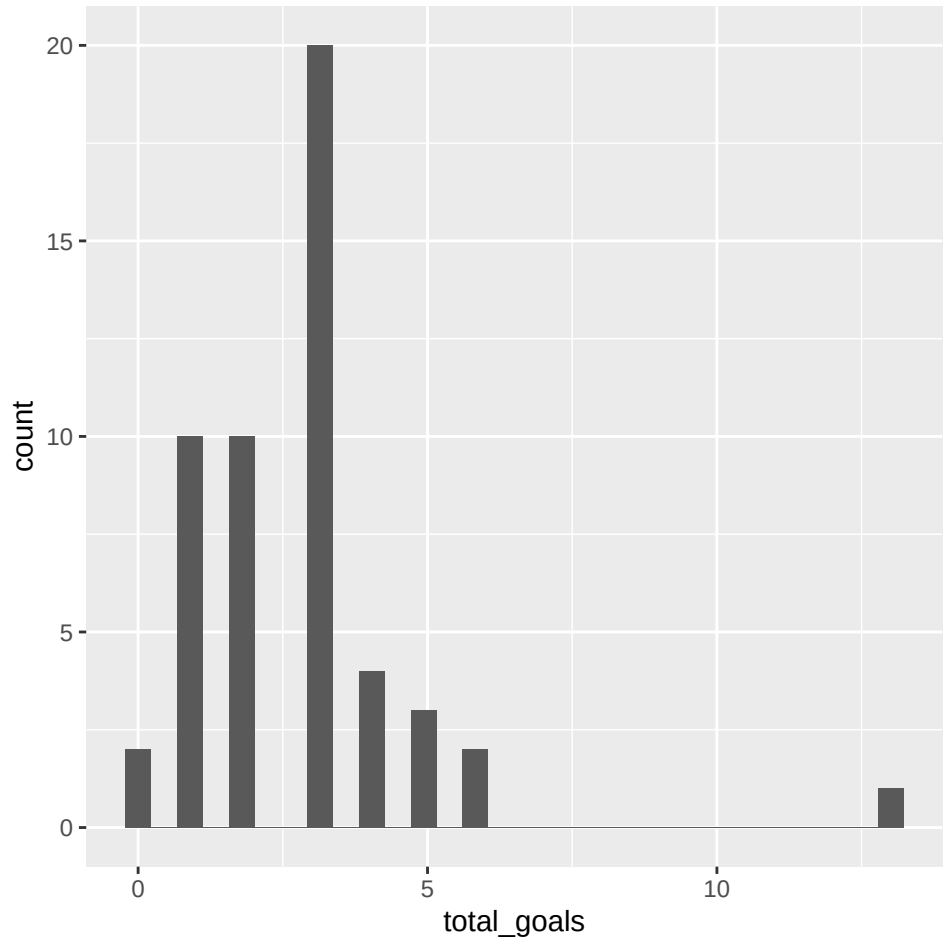
3c. The `wwc_2019_matches` data in the *fivethirtyeight* package includes the number of goals scored by the two teams in each 2019 Women's World Cup match. Create a histogram of the number of goals scored per game. What is the maximum likelihood estimate for the expected number of goals scored in a game? 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

ggplot(data = wwc_2019_matches, aes(x = total_goals)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

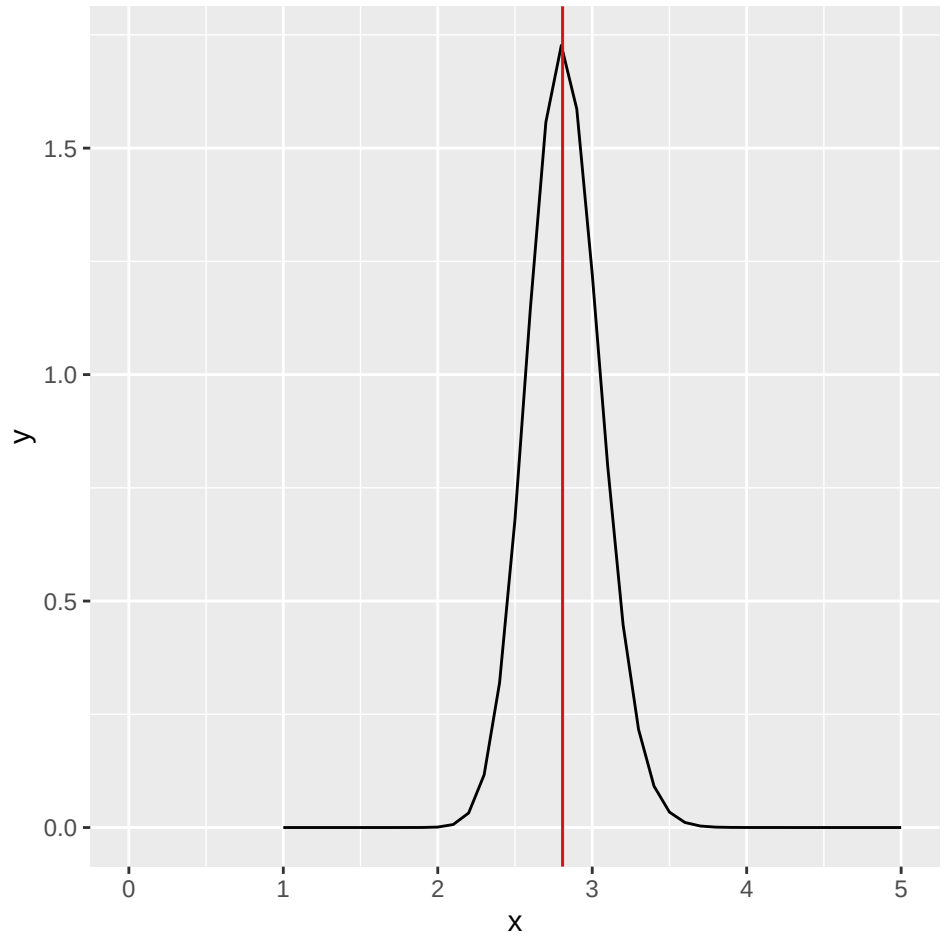



```
soccer_mle <- sum(total_goals)/length(total_goals)
```

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.

```
n <- length(total_goals)
yi <- sum(total_goals)
val <- seq(1, 15, .1)
posterior <- 0.5*dgamma(val, 8+yi, 2+n) + 0.5*dgamma(val, 2+yi, 1+n)
plot <- ggplot(data.frame(x = val, y = posterior), aes(x, y)) + xlim(0, 5) + geom_line()
plot+geom_vline(xintercept = soccer_mle, color = 'red')
```

```
## Warning: Removed 100 rows containing missing values (`geom_line()`).
```



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 the prior did not have much of an impact on conclusions, since they were vastly different and given equal weight. Furthermore, the data is only based on 52 observations. Thus 8 and 2 are relatively large pseudo counts in comparison. The data seemed to influence the results more.