Homework 2

Jessica Nguyen and Tristan Chen

Due on January 30, 2022 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*0.6
beta <- 85*0.6
print(sprintf("Alpha is %f", alpha))
```

```
## [1] "Alpha is 9.000000"
```

```
print(sprintf("Beta is %f", beta))
```

```
## [1] "Beta is 51.000000"
```

```
. = 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 <- 30 + alpha
posterior_beta <- 60 + 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) * (post

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

```
## [1] "The posterior mean is 0.260000"
```

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

```
## [1] "The posterior sd is 0.035696"
```

```
. = 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.

```
# YOUR CODE HERE
w1 <- 90 / (90 + alpha + beta)
print(w1)
```

```
## [1] 0.6
```

The weight for the posterior is 0.6, which tells us that 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.

I expect that the groups should have about the same average incidence of cancer because $120/10 = 12$ and $12/1 = 12$.

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 $\text{gamma}(12 \times n_0, n_0)$ 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
# YOUR CODE HERE
n0 = c(1:50)
new_alpha = n0 * 12
new_beta = n0
alpha_B_posterior2 = new_alpha + sum(yB)
beta_B_posterior2 = new_beta + length(yB)
posterior_means = alpha_B_posterior2 / beta_B_posterior2

# PLOTTING
plot(n0, posterior_means, xlab = "Number of Observations", ylab = "Posterior Mean", type = 'l')
```
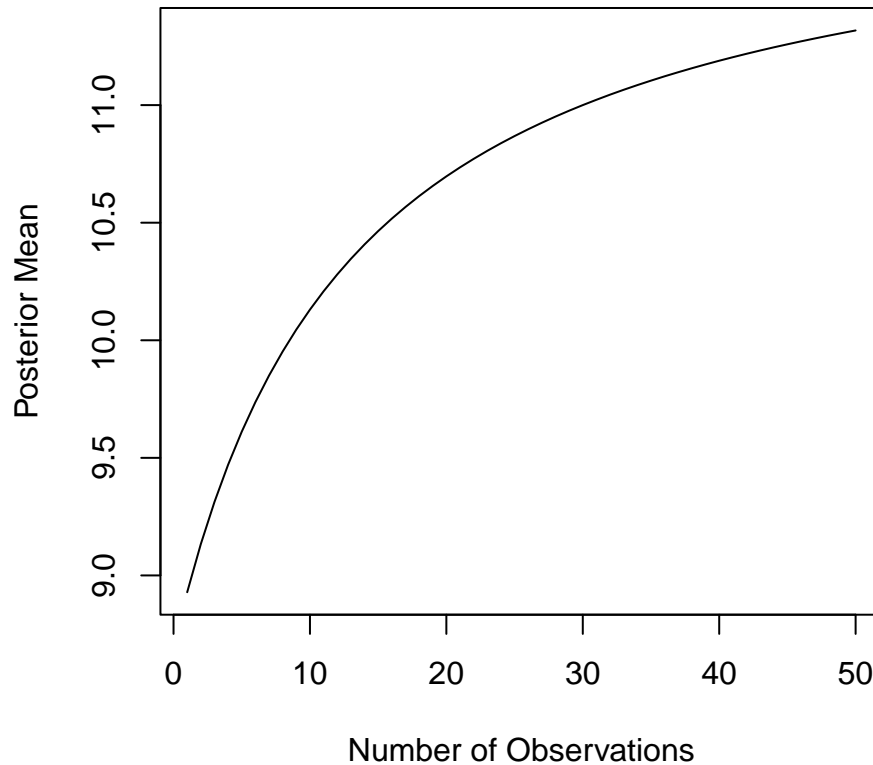


```r
. = ottr::check("tests/q2c.R")
```

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

3

Since population B is related to population A, knowledge about population A should tell us something about population B. There is a correlation between population A and population B, which means that the two populations are not independent of each other, therefore it does not make sense to have $p(\theta_A, \theta_B) = p(\theta_A) \times p(\theta_B)$.

3. A Mixture Prior for Heart Transplant Surgeries

A hospital in the United States wants to evaluate their success rate of heart transplant surgeries. We observe the number of deaths, y, in a number of heart transplant surgeries. Let $y \sim \text{Pois}(\nu\lambda)$ where λ is the rate of deaths/patient and ν is the exposure (total number of heart transplant patients). When measuring rare events with low rates, maximum likelihood estimation can be notoriously bad. We'll tak a Bayesian approach. To construct your prior distribution you talk to two experts. The first expert thinks that $p_1(\lambda)$ with a gamma$(3, 2000)$ density is a reasonable prior. The second expert thinks that $p_2(\lambda)$ with a gamma$(7, 1000)$ density is a reasonable prior distribution. You decide that each expert is equally credible so you combine their prior distributions into a mixture prior with equal weights: $p(\lambda) = 0.5 * p_1(\lambda) + 0.5 * p_2(\lambda)$

3a. What does each expert think the mean rate is, *a priori*? Which expert is more confident about the value of λ a priori (i.e. before seeing any data)?

The first expert thinks that the mean rate is 3/2000, or 0.0015, while the second expert thinks that the mean rate is 7/1000, or 0.007. I believe that the first expert is more confident, because they have a higher beta value than the second expert.

3b. Plot the mixture prior distribution.

```
p <- seq(0, 0.025, 0.0001)
gamma1 <- 0.5 * dgamma(p, 3, 2000)
gamma2 <- 0.5 * dgamma(p, 7, 1000)
mixgamma <- gamma1 + gamma2

ggplot() +
  geom_line(aes(p, mixgamma, color = "mixture"))
```
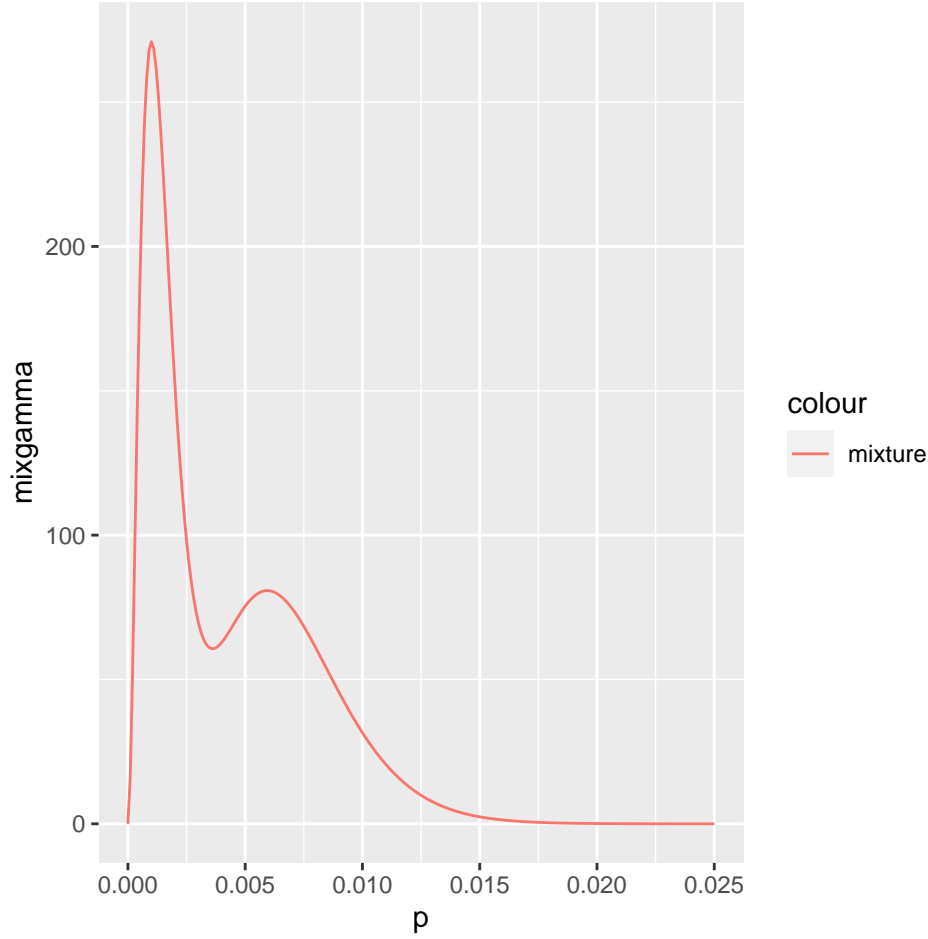
3c Suppose the hospital has $y = 8$ deaths with an exposure of $\nu = 1767$ surgeries performed. Write the posterior distribution up to a proportionality constant by multiplying the likelihood and the prior density. Plot this unnormalized posterior distribution and add a vertical line at the MLE. *Warning:* be very careful about what constitutes a proportionality constant in this example.

$$PriorDensity = 0.5 * p_1(\lambda) + 0.5 * p_2(\lambda)$$

$$PriorDensity = 0.5 \frac{2000^3}{\Gamma(3)} \lambda^{3-1} e^{-2000\lambda} + 0.5 \frac{1000^7}{\Gamma(7)} \lambda^{7-1} e^{-1000\lambda}$$

$$L(\lambda) = e^{-\nu\lambda} * \frac{(\nu\lambda)^8}{8!}$$

$\nu = 1767$ so when we plug it in, the likelihood becomes

$$L(\lambda) = e^{-1767\lambda} * \frac{(1767\lambda)^8}{8!}$$

The posterior is the likelihood multiplied by the prior density, therefore

$$posterior \propto e^{-1767\lambda} \frac{(1767\lambda)^8}{8!} \left(0.5 \frac{2000^3}{\Gamma(3)} \lambda^{3-1} e^{-2000\lambda} + 0.5 \frac{1000^7}{\Gamma(7)} \lambda^{7-1} e^{-1000\lambda} \right)$$

$$posterior \propto e^{-1767\lambda} \frac{(1767\lambda)^8}{8!} \left(0.5 \frac{2000^3}{\Gamma(3)} \lambda^{2} e^{-2000\lambda} + 0.5 \frac{1000^7}{\Gamma(7)} \lambda^{6} e^{-1000\lambda} \right)$$

We took out the proportionality constants in order to get

$$posterior \propto e^{-1767\lambda}\lambda^8 \left(\frac{2000^3}{\Gamma(3)}\lambda^2 e^{-2000\lambda} + \frac{1000^7}{\Gamma(7)}\lambda^6 e^{-1000\lambda} \right)$$

```
gamma_fun3 <- function(l) {exp(-1767*l)*l^(8)*(2000^3/gamma(3) * l^(2) * exp(-2000 * l) + 1000^7/gamma(
```

```
plot(gamma_fun3, xlim = c(0, 0.03))
p <- seq(0, 0.03, 0.0001)
abline(v=p[which.max(gamma_fun3(p))],col='red')
```
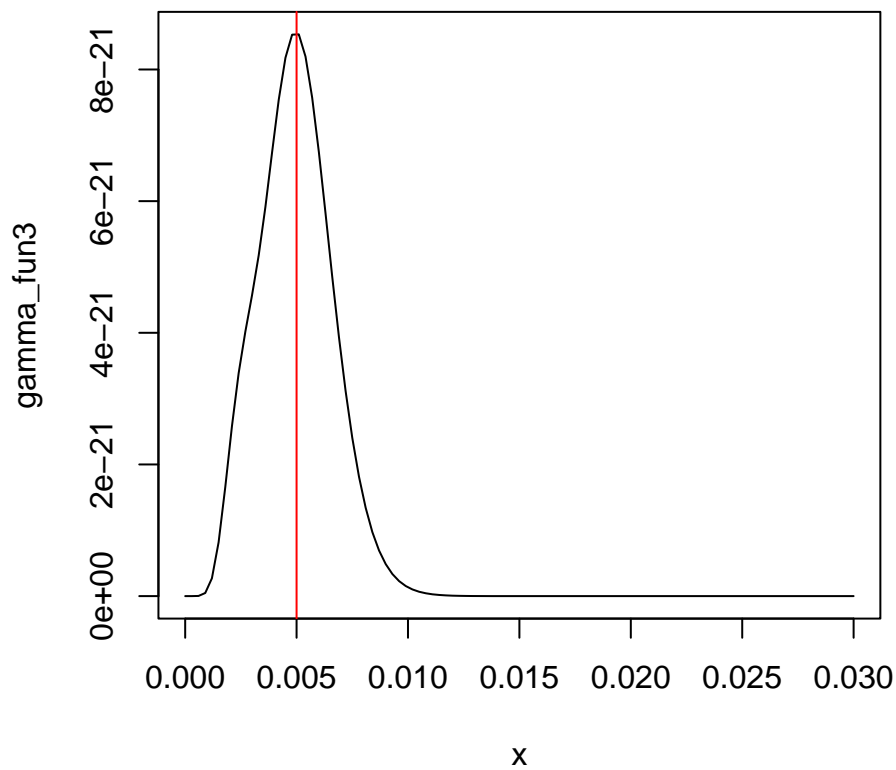

Extra Credit Let $K = \int L(\lambda; y)p(\lambda)d\lambda$ be the integral of the proportional posterior. Then the proper posterior density, i.e. a true density integrates to 1, can be expressed as $p(\lambda \mid y) = \frac{L(\lambda;y)p(\lambda)}{K}$. Compute this posterior density and clearly express the density as a mixture of two gamma distributions.

Type your answer here, replacing this text.