Homework 1

Amy Kuang, Shravan Shenoy - PSTAT 115, Spring 2021

Due on April 25, 2021 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. 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:

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

1a. 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.

According to the prior distributions above, θ_A has a mean (average) of 12 and variance of 1.20 while θ_B has a mean (average) of 12 and a variance of 12. Knowing that smaller variance means more accurate mean, we expect that strain B's mice may have a higher average incidence of cancer and we are more certain about strain A's mice priori.

1b. 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"
```

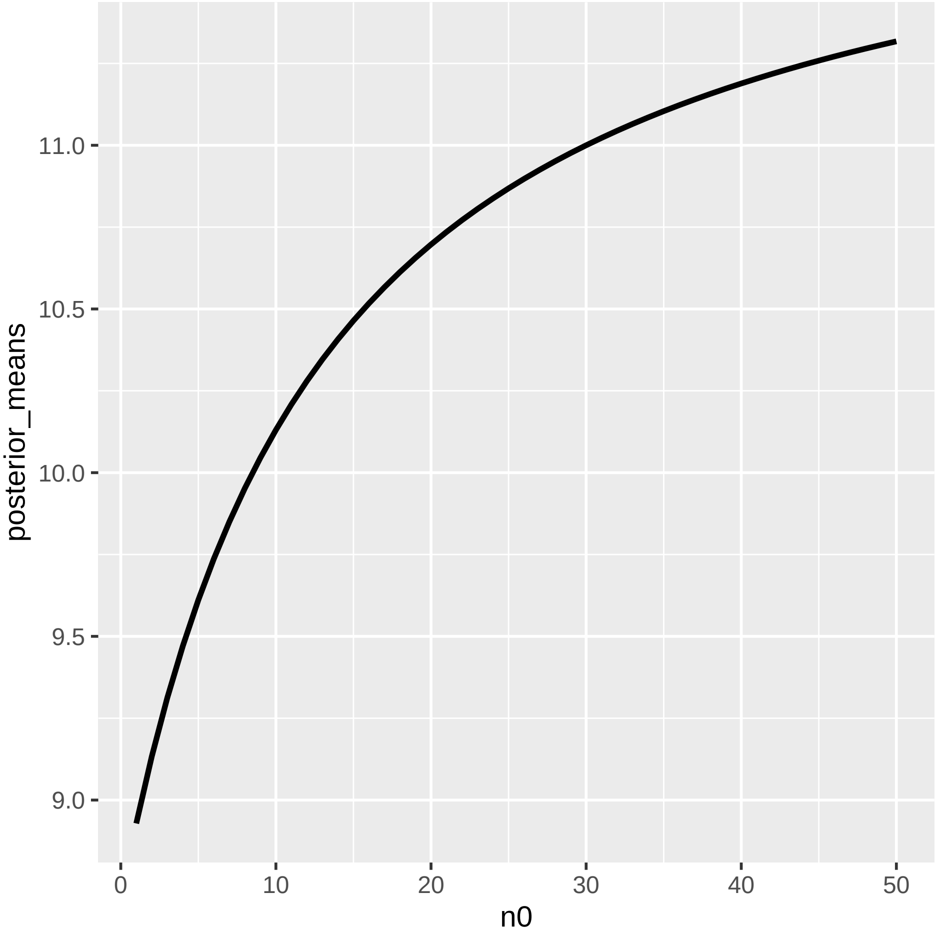
```
## [1] "Posterior mean of theta_B 8.93"
```

```
## [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]"
```

1c. 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
n0 <- c(1:50)
alpha <- 12*n0
beta <- n0
alpha_posterior <- sum(yB) + alpha
beta_posterior <- 13 + beta

posterior_means = alpha_posterior/beta_posterior # YOUR CODE HERE

# YOUR CODE HERE
data_ <- data.frame(n0, posterior_means)
ggplot(data_, aes(x=n0, y=posterior_means)) +
    geom_line(size=1)
```



1d. 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, because population B is said to be related to population A. Thus, they are not independent from each other.

2. A Mixture Prior for Heart Transplant Surgeries

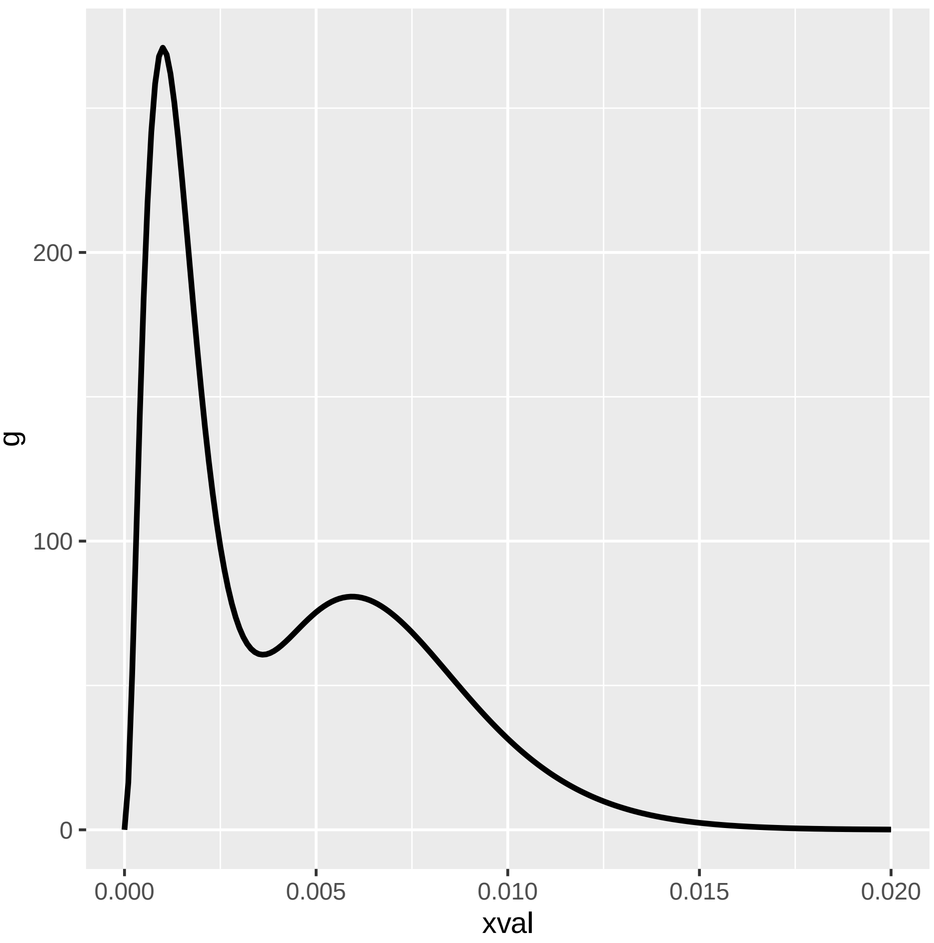
A hospital in the United States wants to evaluate their success rate of heart transplant surgeries. We observe the number of deaths, y, in a number of heart transplant surgeries. Let $y \sim \text{Pois}(\nu\lambda)$ where λ is the rate of deaths/patient and ν is the exposure (total number of heart transplant 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: $p(\lambda) = 0.5 * p_1(\lambda) + 0.5 * p_2(\lambda)$

2a. 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 0.0015 (or 3/2000) while the second expert thinks that the mean rate is 0.007 (or 7/1000). The first expert is more confident about the value of lambda since the variance for the first expert's distribution is 7.5e-07 while the variance for second expert's distribution is 7e-06 – since higher variance means less confidence.$

2b. Plot the mixture prior distribution.

```
# YOUR CODE HERE
xval <- seq(0, 0.02, 0.0001)

a1 <- 3
b1 <- 2000
g1 <- dgamma(xval, a1, rate= b1)

a2 <- 7
b2 <- 1000
g2 <- dgamma(xval, a2, rate= b2)

g <- 0.5*g1 + 0.5*g2

df <- data.frame(xval, g)
ggplot(df, aes(x=xval, y=g)) +
    geom_line(size=1)
```



2c. 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. *Warning:* be very careful about what constitutes a proportionality constant in this example.

Posterior distribution:

$$= p(\lambda|y) \propto p(|\lambda) * p(\lambda) = \frac{(\nu_i\lambda)^{y_i} e^{-\nu_i\lambda}}{y_i!} * [0.5 * p_1(\lambda) + 0.5 * p_2(\lambda)]$$

(after plugging in y and ν)

$$\propto p(\lambda|y = 8) \propto p(y = 8|\lambda) * p(\lambda) = \frac{(1767\lambda)^8 e^{-1767\lambda}}{8!} * [0.5 * p_1(\lambda) + 0.5 * p_2(\lambda)]$$

$$\propto \lambda^8 e^{-1767\lambda} * [p_1(\lambda) + p_2(\lambda)]$$

$$\propto \lambda^8 e^{-1767\lambda} * [\frac{2000^3}{\gamma(3)} * \lambda^2 e^{-2000\lambda} + \frac{1000^7}{\gamma(7)} * \lambda^6 e^{-1000\lambda}]$$

$$\propto \lambda^{10} e^{-3767\lambda} * \frac{2000^3}{\gamma(3)} + \lambda^{14} e^{-2767\lambda} \frac{1000^7}{\gamma(7)}$$

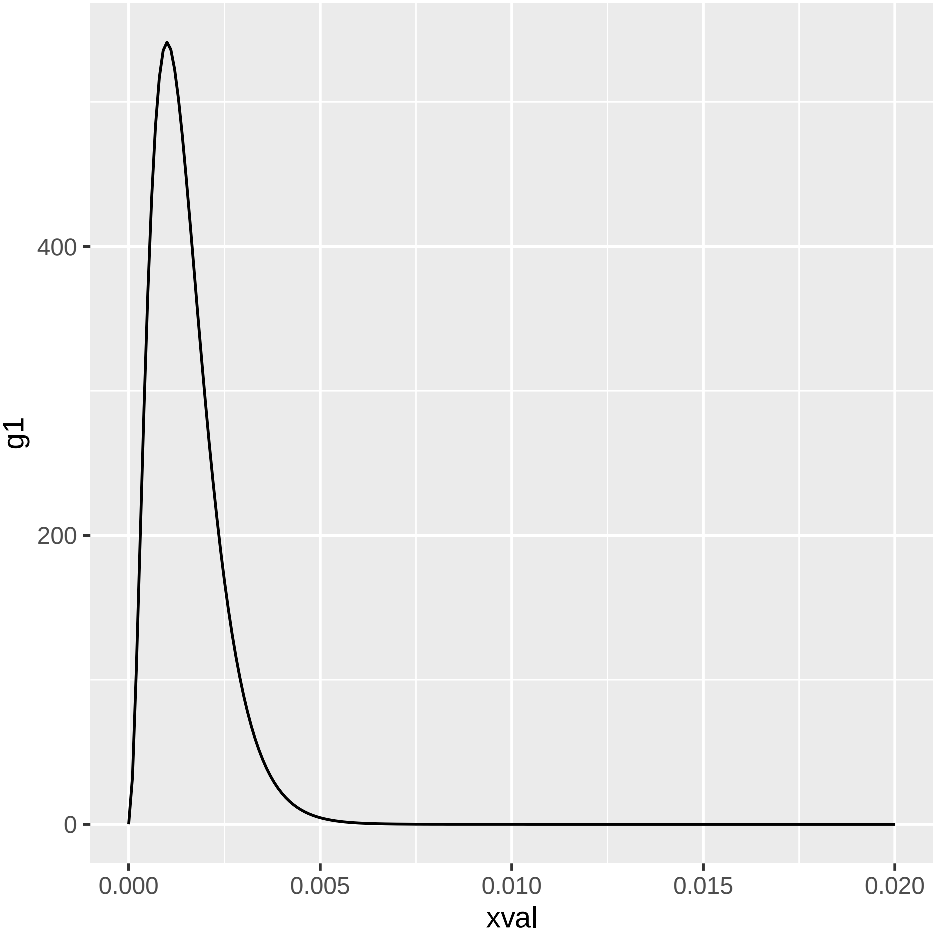
2d. Let $K = \int L(\lambda; y)p(\lambda)d\lambda$ be the integral of the proportional posterior. Then the proper posterior density, i.e. a true density integrates to 1, can be expressed as $p(\lambda \mid y) = \frac{L(\lambda;y)p(\lambda)}{K}$. Compute this posterior density and clearly express the density as a mixture of two gamma distributions.

Posterior density:

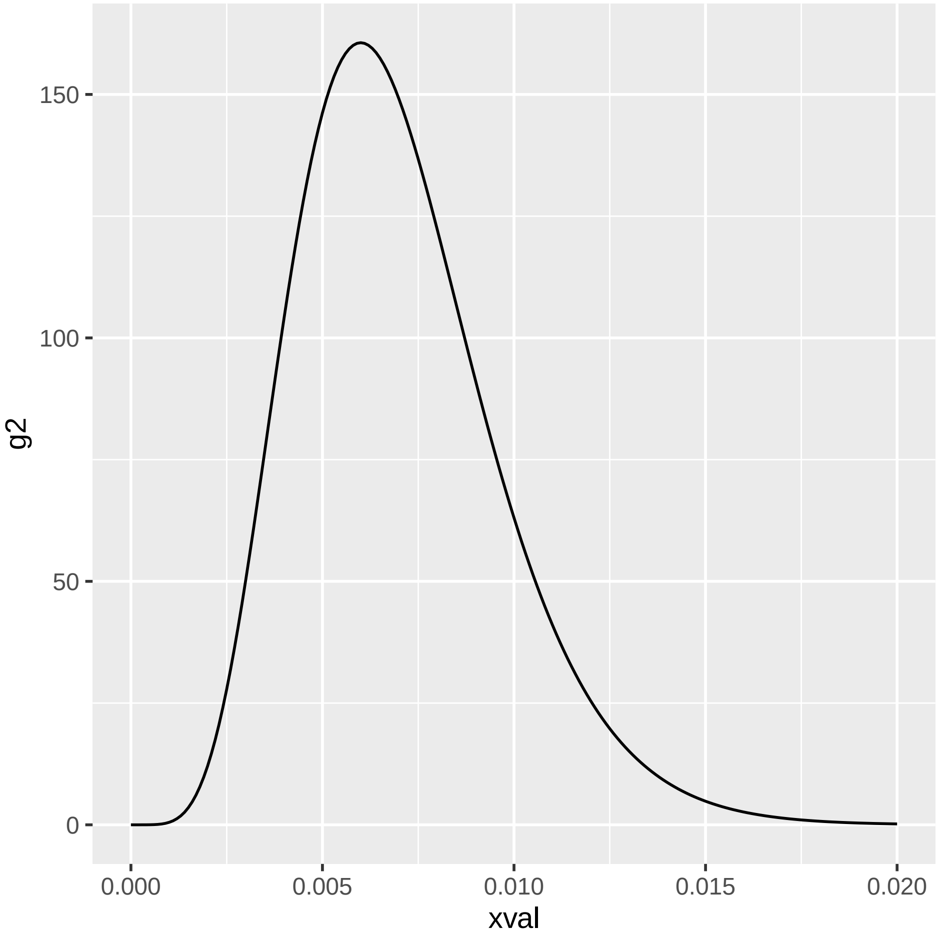
$$\propto \lambda^{10} e^{-3767\lambda} * \frac{2000^3}{\gamma(3)} + \lambda^{14} e^{-2767\lambda} \frac{1000^7}{\gamma(7)}$$

2e. Plot the posterior distribution. Add vertical lines clearly indicating the prior means from each expert. Also add a vertical line for the maximum likelihood estimate.

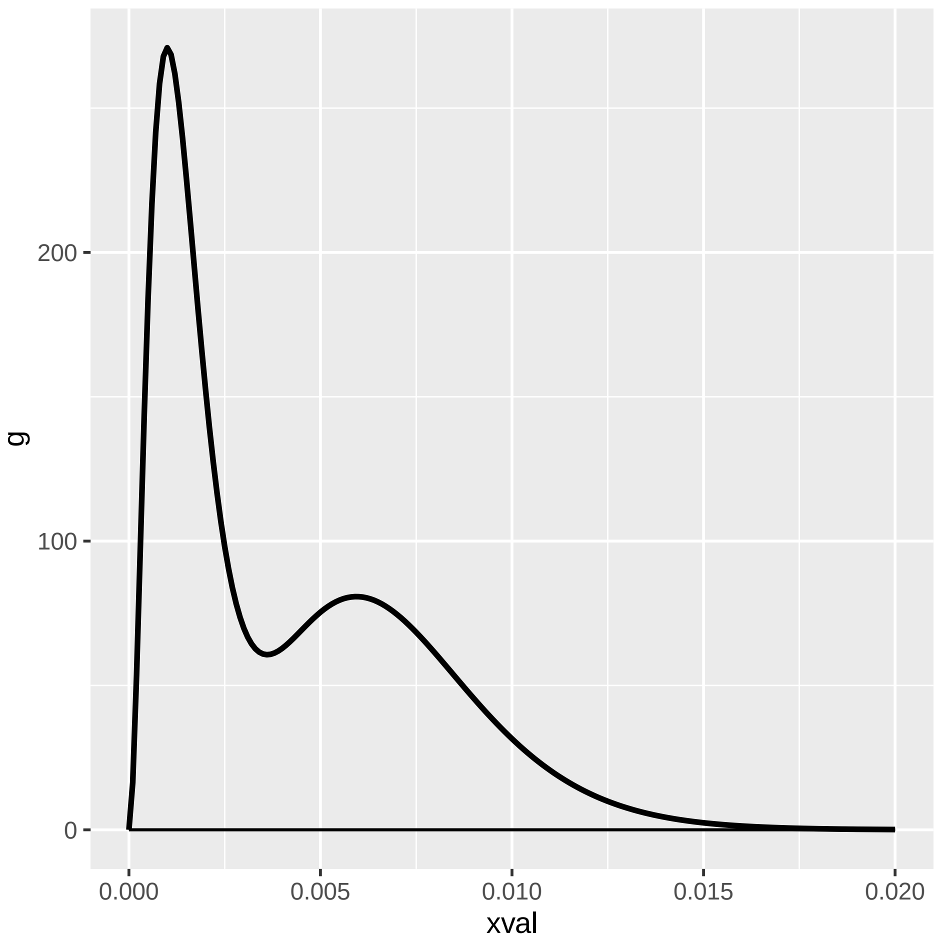
```
# YOUR CODE HERE

prior1_mean <- a1/b1
prior2_mean <- a2/b2

df1 <- data.frame(xval, g1)
df2 <- data.frame(xval, g2)

ggplot(df1, aes(x=xval, y=g1)) + geom_line()
```



```
ggplot(df2, aes(x=xval, y=g2)) + geom_line()
```



```
ggplot(df, aes(x=xval, y=g)) + geom_line(size=1) + geom_line(y=prior1_mean) + geom_line(y=prior2_mean
```


200
100
0
g
0.000
0.005
0.010
0.015
0.020
xval