Homework 1

PSTAT 115, Spring 2021 by Lina Peroomian and Simranjit Kaur

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.

Before seeing any data, I expect group A to have a higher average incidence of cancer since they are well studied, but group A and B might have very similar average incidences of cancer since the mice are related. Furthermore, we see that both the populations follow gamma distributions. In terms of the parameters, we see that group A has larger parameters in comparison to group B. Given this information, I would say that I am more certain about a priori of group A since the variance is much lower than group B, therefore we are confident of the mean.

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.

```
yA <- c(12, 9, 12, 14, 13, 13, 15, 8, 15, 6)
yB <- c(11, 11, 10, 9, 9, 8, 7, 10, 6, 8, 8, 9, 7)

### Prior parameters here
alpha_A = 120
```

```r
beta_A = 10

alpha_B =12
beta_B = 1

### Posterior parameters here
alpha_A_posterior = sum(yA) + alpha_A
beta_A_posterior =length(yA) + beta_A

alpha_B_posterior = sum(yB) + alpha_B
beta_B_posterior = length(yB) + beta_B

### Posterior mean and variance for each group
A_post_mean <- alpha_A_posterior/beta_A_posterior
A_post_var <- alpha_A_posterior/(beta_A_posterior^2)

### Posterior quantiles for each group
B_post_mean <- alpha_B_posterior/beta_B_posterior
B_post_var <- alpha_B_posterior/(beta_B_posterior^2)

print(paste0("Posterior mean of theta_A ", round(A_post_mean, 2)))
```

```
## [1] "Posterior mean of theta_A 11.85"
```

```r
print(paste0("Posterior variance of theta_A ", round(A_post_var, 2)))
```

```
## [1] "Posterior variance of theta_A 0.59"
```

```r
print(paste0("Posterior mean of theta_B ", round(B_post_mean, 2)))
```

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

```r
print(paste0("Posterior variance of theta_B ", round(B_post_var, 2)))
```

```
## [1] "Posterior variance of theta_B 0.64"
```

```r
# Posterior quantiles
alpha_A_quantile <- c(qgamma(0.05/2, shape =alpha_A_posterior,scale= 1/beta_A_posterior),
                      qgamma(1-(0.05/2),alpha_A_posterior,beta_A_posterior))

alpha_B_quantile <- c(qgamma(0.05/2, shape=alpha_B_posterior,scale=1/beta_B_posterior),
                      qgamma(1-(0.05/2),alpha_B_posterior,beta_B_posterior))



print(paste0("Posterior 95% quantile for theta_A is [", round(alpha_A_quantile[1],2), ", ", round(alpha
```

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

```
print(paste0("Posterior 95% quantile for theta_B is [", round(alpha_B_quantile[1],2), ", ", round(alpha_
```

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

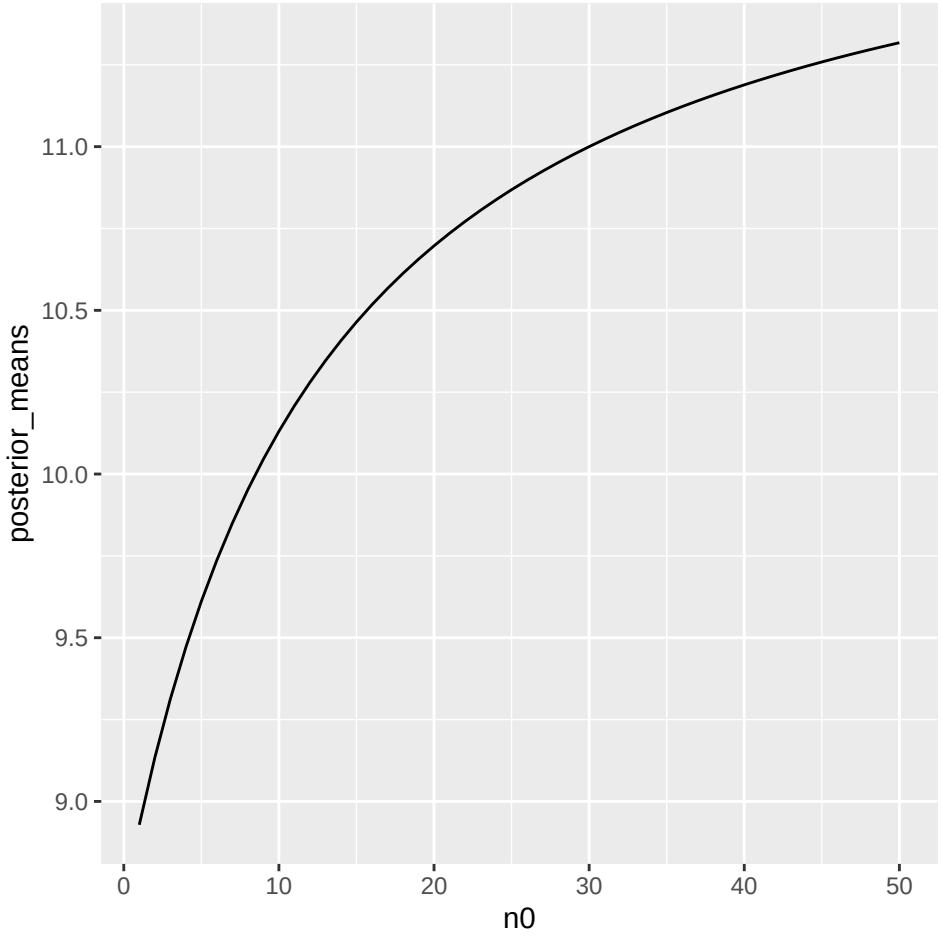
From these outcomes, we can see that posterior mean of θ_A is 11.85. The posterior variance of θ_A is 0.59. The posterior mean of θ_B is 8.93. The posterior variance of θ_B is 0.64. The posterior 95% quantile for θ_A is [10.39, 13.41]. Lastly, the posterior 95% quantile for θ_B is [7.43, 10.56].

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).

```
n0 <- c(1:50)
alpha_B2 <- 12*n0
beta_B2 <- n0
alpha_B2_post <- sum(yB) + alpha_B2
beta_B2_post <- length(yB) + beta_B2

posterior_means = alpha_B2_post/beta_B2_post

new.plot <- data.frame(n0,posterior_means)
ggplot(new.plot, aes(n0,posterior_means))+geom_line()
```



1d. Should knowledge about population A tell us anything about population B? Discuss whether or not it makes sense to have $p(\theta_A, \theta_B) = p(\theta_A) \times p(\theta_B)$.

We were given that the two groups were related, but we can end up seeing that they are close but not the same. I think its fair to say that both population A and B are pretty similar. We can see that they both have the same gamma prior. I think the fact that they have different posteriors could be used to argue that they are acutally differeny, so I am not so sure. If I had to make a claim, I think I would say that I think it would make sense to have $p(\theta_A, \theta_B) = p(\theta_A) \times p(\theta_B)$ since both populations are pretty similar.

2. A Mixture Prior for Heart Transplant Surgeries

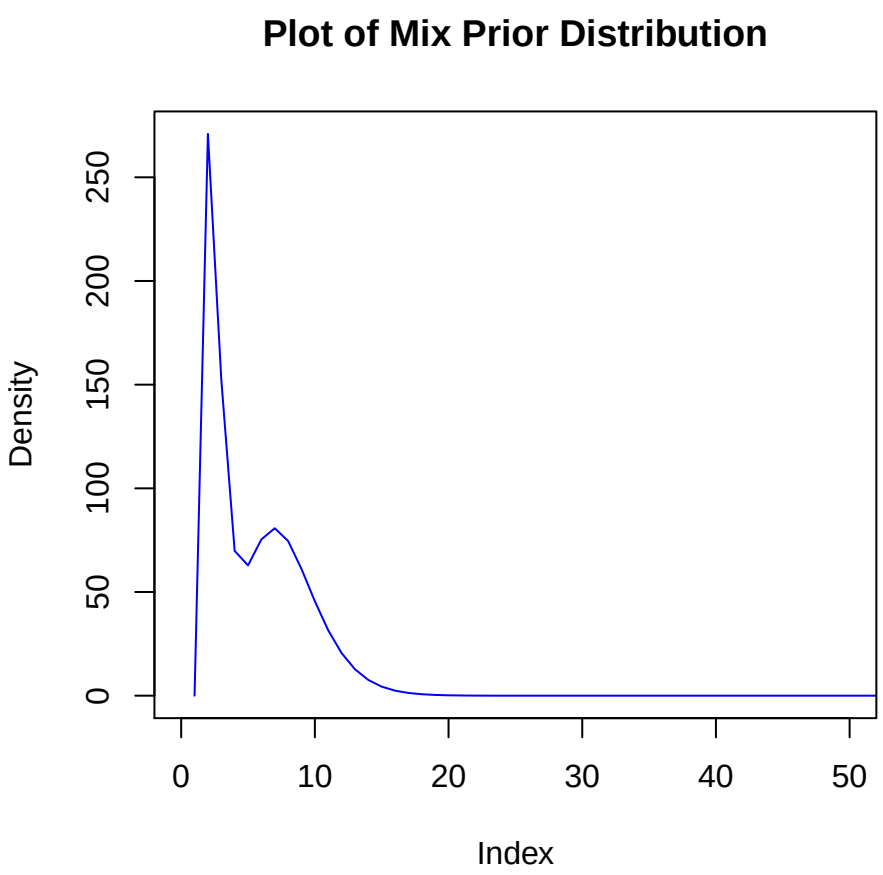
A hospital in the United States wants to evaluate their success rate of heart transplant surgeries. We observe the number of deaths, y, in a number of heart transplant surgeries. Let $y \sim \text{Pois}(\nu\lambda)$ where λ is the rate of deaths/patient and ν is the exposure (total number of heart transplant 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 the mean rate is 3/2000 and the second expert thinks the mean rate is 7/1000. We can see that the first expert is more confident because the variance of their distribution is lower than the variance of the second expert and the first expert's distribtuion is more concentrated around the mean since the beta value is large, which make us more confident.

2b. Plot the mixture prior distribution.

```
sequence1 <- seq(0, 0.06, by = 0.001)
plot(0.5*dgamma(sequence1, 3, 2000) + 0.5 *dgamma(sequence1, 7, 1000),
     main = "Plot of Mix Prior Distribution",
     ylab = "Density",
     xlab = ,
     type = "l",
     col = "blue",
     xlim = c(0, 50))
```



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.

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.

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.

```
new_lambda <- seq(0,.015,0.0001)

dist1_new <- dgamma(new_lambda, shape = 11, rate = 3767)
dist2_new <- dgamma(new_lambda, shape = 15, rate = 2767)

dist1 <- (2000^3/gamma(3))*(gamma(11)/3767^11)
dist2 <- (1000^7/gamma(7))*(gamma(15)/2767^15)

div1 <- dist1/(dist1 + dist2)
div2 <- dist2/(dist1 + dist2)

combined <- div1*(dist1_new) +  div2*(dist2_new)
```

```
updated_dat <- data.frame(new_lambda, combined)
mle <- div1*(11/3767) + div2*(15/2767)
new_plot <- ggplot(updated_dat, aes(new_lambda, combined))+
    xlim(0,.015)+
    geom_line() +
    ggtitle("Plot of Posterior Distribution")
new_plot+
    geom_vline(xintercept =c(0.0015,0.007,mle),linetype='longdash', color=c('red','red','green'))+
    xlab("lambda") + ylab("p(lambda | y)")
```