Homework 2

Julissa Dueñas

Due on October 25, 2020 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.

I would expect both A and B to have around the same average. This is because the expected value of a gamma is a/b being 120/10=12 for A and 12/1=12 aswell for B.

I am more certain about a priori for group A because it has a lower variance. The variance for a gamma distribution is a/bˆ2, 1.2 for A and 12 for B

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 21.55"

## [1] "Posterior variance of theta_A 1.96"

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

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

## [1] "Posterior 95% quantile for theta_A is [0.95, 0.98]"

## [1] "Posterior 95% quantile for theta_B is [0.88, 0.94]"
```

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

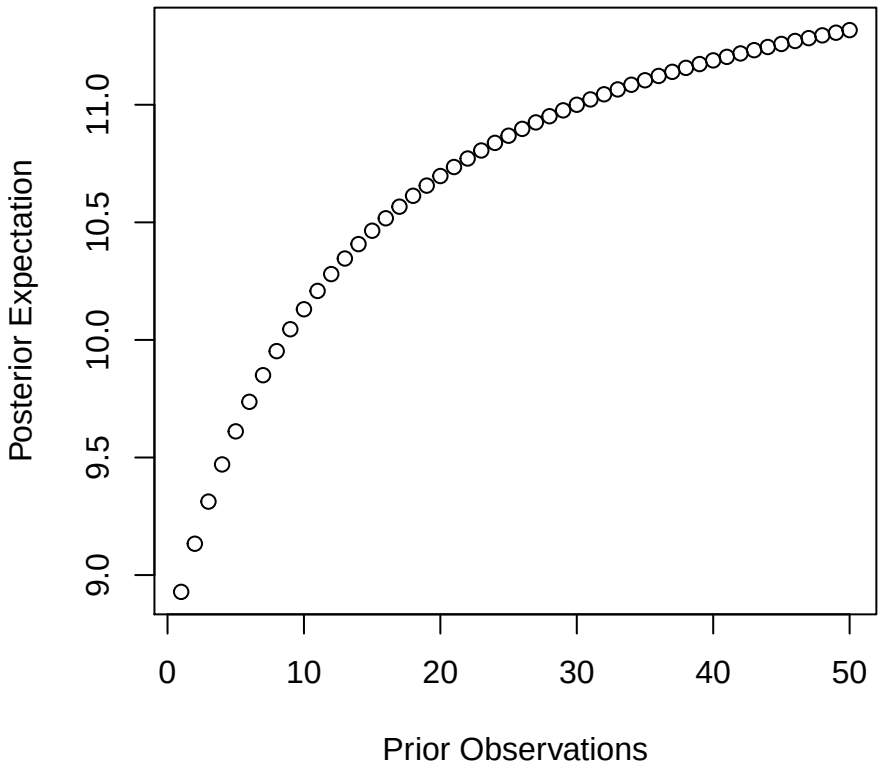
```
## All tests passed!
```

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_b_pos2 <- sum(yB)+alpha_b2
beta_b_pos2 <- 13+beta_b2

posterior_means = alpha_b_pos2/beta_b_pos2

plot(n0,posterior_means,xlab='Prior Observations',ylab='Posterior Expectation')
```




```
. = ottr::check("tests/q1c.R")
```

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

The mice in population B are said to be related to those in population A so knowledge about population A could help us make a guess about population B but the two are independent statistically. The tumore count in

population A does not affect that of population B so yes, it does make sense to have $p(\theta_A, \theta_B) = p(\theta_A) \times p(\theta_B)$

2. A Mixture Prior for Heart Transplant Surgeries

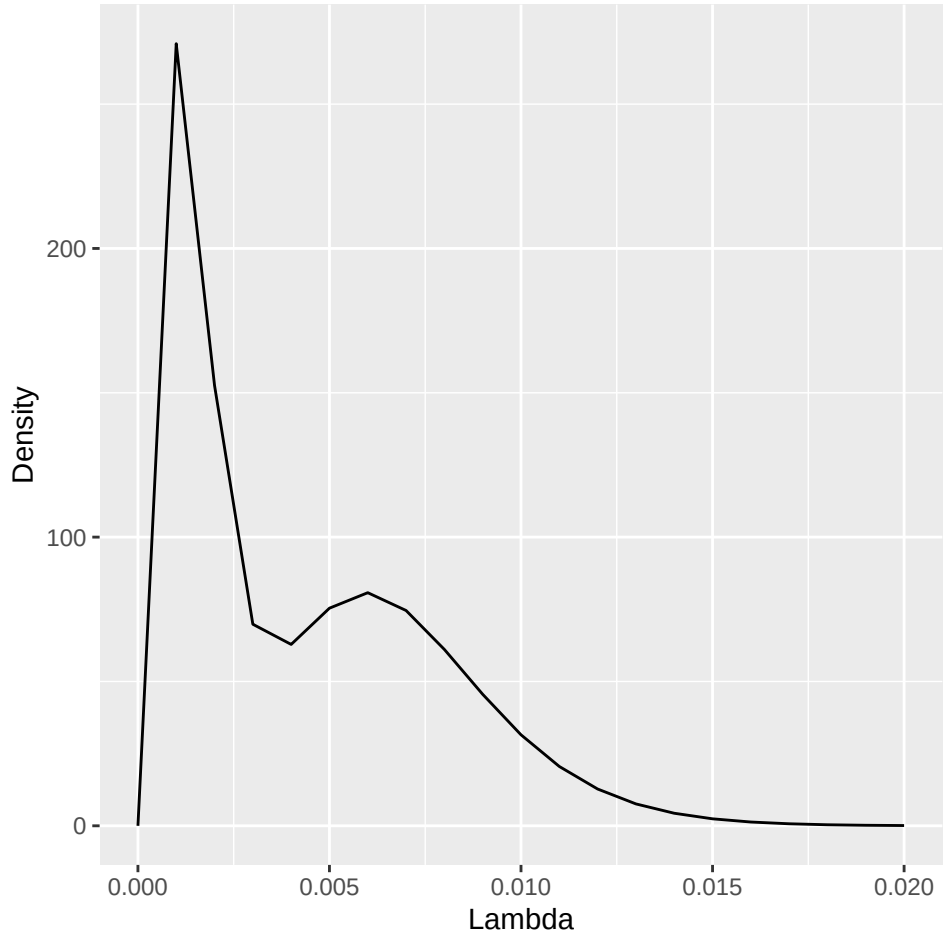
A hospital in the United States wants to evaluate their success rate of heart transplant surgeries. We observe the number of deaths, y, in a number of heart transplant surgeries. Let $y \sim \text{Pois}(\nu\lambda)$ where λ is the rate of deaths/patient and ν is the exposure (total number of heart transplant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

Expert 1 believed the mean rate is $3/2000 = 0.0015$ and Expert 2 thinks the mean rate is $7/1000 = 0.007$. Expert 1 is more confident about the value of λ a priori because there is a smaller variance in his believed density. Expert 1 has a variance of 3/4,000,000 and Expert 2 has a variance of 7/1,000,000.

2b. Plot the mixture prior distribution.

```
lambda <- seq(0,.02,0.001)
mixture_prior <- (0.5*dgamma(lambda,3,2000)+0.5*dgamma(lambda,7,1000))
qplot(lambda,mixture_prior,geom='line',xlab='Lambda',ylab='Density')
```



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.

$P(\lambda|y = 8) \propto L(\lambda)P(\lambda)$

$\propto P(y = 8|\nu\lambda)(0.5p_1(\lambda) + 0.5p_2(\lambda))$

gamma: $f(\lambda, \alpha, \beta) = \frac{\beta^\alpha \lambda^{\alpha-1} e^{-\beta\lambda}}{\gamma(\alpha)}$

poisson: $P(Y = y) = \frac{e^{-\lambda}\lambda^y}{y!} = \frac{e^{-\nu\lambda}\nu\lambda^8}{8!}$

$P(\lambda|y = 8) \propto \frac{e^{-\nu\lambda}(\nu\lambda)^8}{8!} * 0.5(\frac{2000^3\lambda^{3-1}e^{-2000\lambda}}{\Gamma(3)} + \frac{1000^7\lambda^{7-1}e^{-1000\lambda}}{\Gamma(7)})$

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

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

Factor out constants: $e^{-1767\lambda}1767^8\lambda^8(\frac{2000^3\lambda^2e^{-2000\lambda}}{2!} + \frac{1000^7\lambda^6e^{-1000\lambda}}{6!})$

Remove first constant and continue to simplify: $e^{-1767\lambda}\lambda^8(\frac{2^{12}*5^9\lambda^2e^{-2000\lambda}}{2} + \frac{2^{21}*5^{21}\lambda^6e^{-1000\lambda}}{2^4*3^2*5})$

$e^{-1767\lambda}\lambda^8(2^{11} * 5^9\lambda^2e^{-2000\lambda} + \frac{2^{17}*5^{20}\lambda^6e^{-1000\lambda}}{3^2})$

Factor out constants: $2^{11} * 5^9e^{-1767\lambda}\lambda^8(\lambda^2e^{-2000\lambda} + \frac{2^6*5^{11}\lambda^6e^{-1000\lambda}}{3^2})$

Drop constants: $e^{-1767\lambda}\lambda^8(\lambda^2 e^{-2000\lambda} + \frac{2^6 * 5^{11}\lambda^6 e^{-1000\lambda}}{3^2})$

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.

$K = \int L(\lambda; y)p(\lambda)d\lambda$

$= \int e^{-1767\lambda}\lambda^8(\lambda^2 e^{-2000\lambda} + \frac{2^6 * 5^{11}\lambda^6 e^{-1000\lambda}}{3^2})d\lambda$

Proper posterior density: $p(\lambda \mid y) = \frac{L(\lambda;y)p(\lambda)}{K} = \frac{e^{-1767\lambda}\lambda^8(\lambda^2 e^{-2000\lambda} + \frac{2^6 * 5^{11}\lambda^6 e^{-1000\lambda}}{3^2})}{\int e^{-1767\lambda}\lambda^8(\lambda^2 e^{-2000\lambda} + \frac{2^6 * 5^{11}\lambda^6 e^{-1000\lambda}}{3^2})d\lambda}$

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
```