Homework 2

PSTAT 115, Fall 2020

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 <- 12
beta <-68
```

```
. = 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? XZ

```
posterior_alpha <- 42 # 12 + 30
posterior_beta <- 128 # 68 + 90 - 30
```

```
. = 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_variance <- (posterior_alpha * posterior_beta)/(((posterior_alpha + posterior_beta)^2)*(poste
posterior_sd <- sqrt(posterior_variance)

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

```
## [1] "The posterior mean is 0.247059"
```

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

```
## [1] "The posterior sd is 0.032982"
```

```
. = 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.

The posterior model reflects both the prior information and the observed data. When finding E[theta | y] we find that it gives us w * $\hat{theta}_{MLE}$ + (1-w) * $\hat{theta}_{prior+mean}$. We found the observed data which is expected MLE and prior knowledge, expected prior, or prior guess about successes, both have an influence in the posterior model.

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.

Before seeing any data, I would expect both groups, A and B, to have the same/similar averages. This is because the expected value for a gamma distribution is a/b or 120/10 = 12 for theta_A and 12/1 = 12 for theta_B.

Given that the variance for a gamma distribution is a/bˆ2, I am more certain about a priori for group A because it has a lower variance. i.e. 120/(10)ˆ2 = 1.2 for theta_A and 12/1ˆ2 = 12 for theta_B.

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.

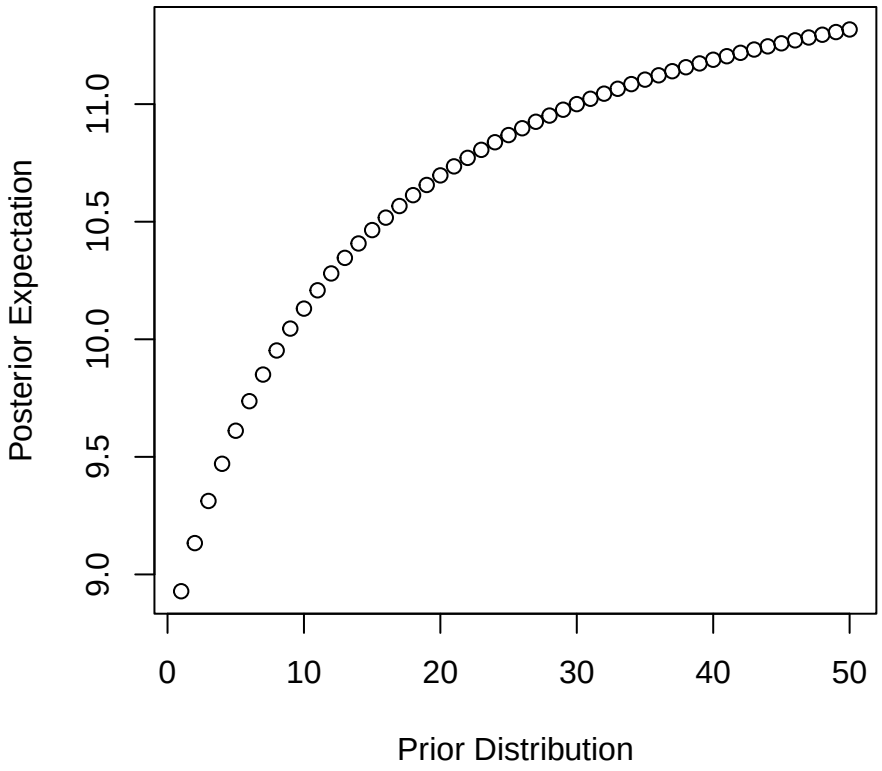
```
## [1] "Posterior mean of theta_A is 11.85"
## [1] "Posterior variance of theta_A is 0.59"
## [1] "Posterior mean of theta_B is 8.93"
## [1] "Posterior variance of theta_B is 0.64"
## [1] "Posterior 95% quantile for theta_A is [0.91, 0.95]"
## [1] "Posterior 95% quantile for theta_B is [0.88, 0.94]"
```

```
. = 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).

```r
n0 <- c(1:50)
alpha_b2 <- 12*n0
beta_b2 <- n0
alpha_b2_posterior <- sum(yB)+alpha_b2
beta_b2_posterior <- 13+beta_b2

posterior_means = alpha_b2_posterior/beta_b2_posterior

plot(n0,posterior_means,xlab = "Prior Distribution", ylab = "Posterior Expectation")
```




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

We are told that the mice in population B is related to the mice in population A so knowledge about population A could help us make a guess about population B. The reason why it may not matter about knowledge beforehand is that the two are independent statistically. The tumor count in population A does not affect that of population B, therefore, $p(\theta_A, \theta_B) = p(\theta_A) \times p(\theta_B)$ makes sense.

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)?

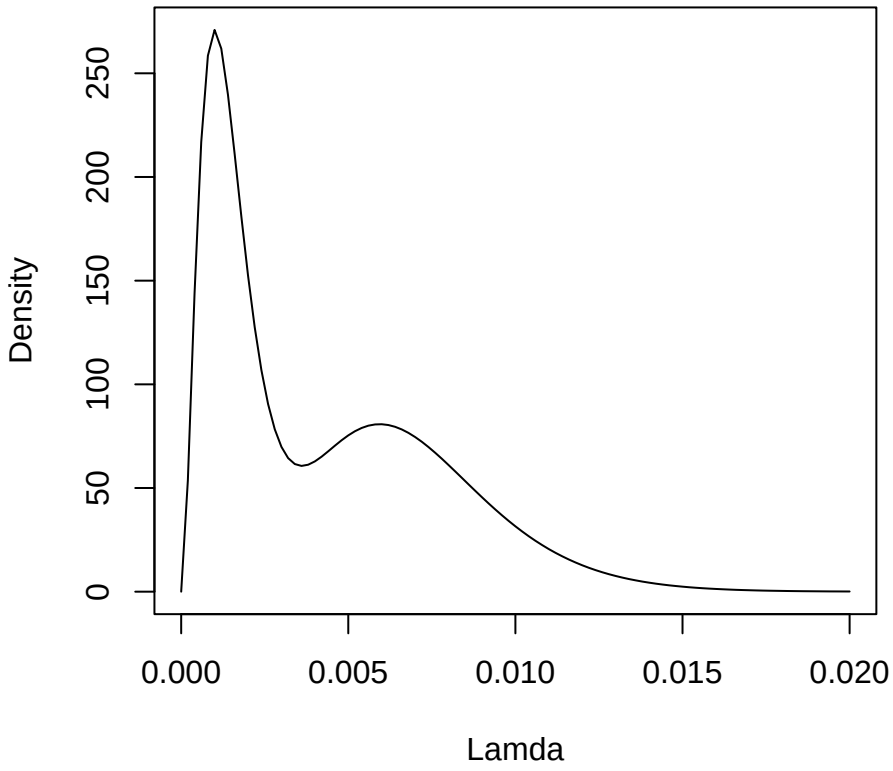
Expert 1 thinks the mean rate is $3/2000 = 0.0015$ and Expert 2 thinks the mean rate is $7/1000 = 0.007$. Lets look at their 95% credible intervals using their priors.

```
upperlim1 <-  qgamma(.975, 3,2000)  #0.003612344
lowerlim1 <- qgamma(.025, 3, 2000)  #0.0003093361



upperlim2 <- qgamma(.975, 7, 1000) #0.01305947
lowerlim2 <- qgamma(.025, 7, 1000) #0.002814363
```

Expert 1 credible interval is (.0003, .003) and expert 2 credible interval is (.0028, .013). Comparing these intervals, we notice that expert 2 is more confident because they has a smaller interval indicating greater confidence.

3b. Plot the mixture prior distribution.

```
curve(.5*dgamma(x,shape = 3, rate=2000) + .5*dgamma(x, shape = 7, rate=1000), from=0, to=.02, xlab="Lam
```



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.

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

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

We know this is a Poisson-gamma model. The Likelihood function is a possion model meant as exposure:

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

where $y_i = 8$ is actually a proportionality constant we can get rid of so:

$$= e^{-\nu\lambda}\nu\lambda^8$$

we can then find a gamma distribution proportional to our prior.

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

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

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

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

Factor out constants:

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

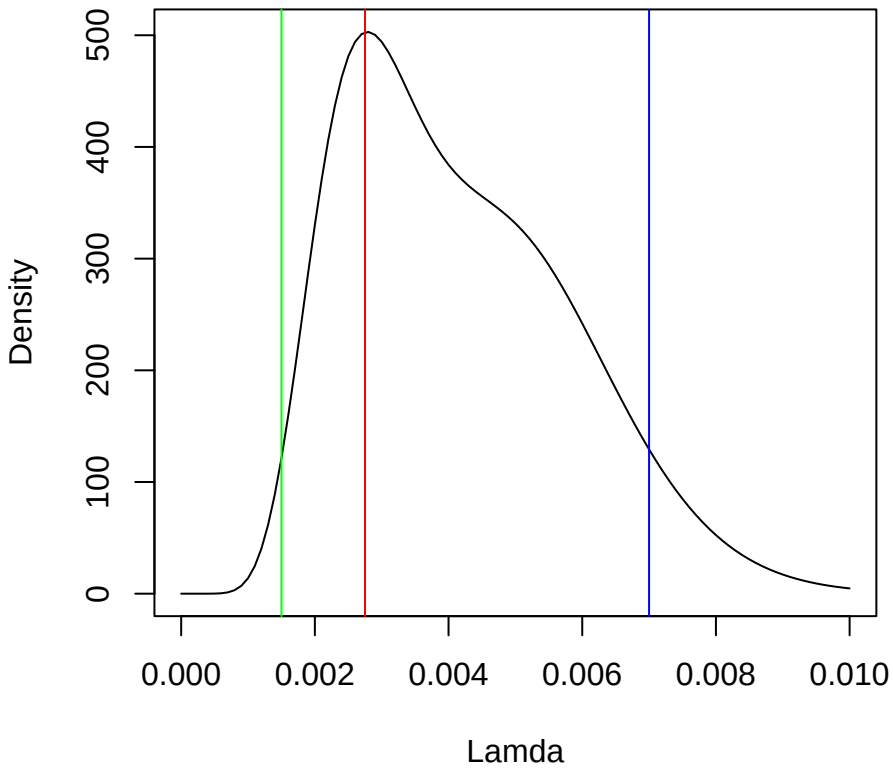
Remove constant (1767)^8 and simplify:

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

Simplify further and distribute:

$$\frac{2000^{3}\lambda^{10}e^{-3767\lambda}}{2!} + \frac{1000^{7}\lambda^{14}e^{-2767\lambda}}{6!}$$

```
curve(dgamma(x,shape = 11, rate=3767) + dgamma(x, shape = 15, rate=2767), from=0, to=.01, xlab="Lamda",
abline(v =(((7/1000)-(3/2000))/2), col = "red")
abline(v =(3/2000), col = "green") #Expert 1 mle from part a, aka mean
abline(v =(7/1000), col = "blue") #Expert 2 mle from part a, aka mean
```



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.

$$p(\lambda \mid y = 8) \propto \lambda^8 e^{-1767\lambda}(\lambda^2 e^{-2000\lambda} + \lambda^6 e^{-1000\lambda})$$

Since we know a true density integral evaluates to 1, then by multiplying both sides by the reciprocal of K we get:

$$\frac{1}{K} \times K = \frac{1}{K} \int_0^\infty L(\lambda|y) \times p(\lambda) d\lambda$$

$$1 = \int_0^\infty \frac{L(\lambda|y) \times p(\lambda)}{K} d\lambda$$

$$1 = \frac{1}{K} \int_0^\infty (\lambda^{10} e^{-3767\lambda} + \lambda^{14} e^{-2767\lambda} d\lambda)$$

$$1 = \frac{1}{K} (\int_0^\infty \lambda^{10} e^{-3767\lambda} d\lambda + \int_0^\infty \lambda^{14} e^{-2767\lambda} d\lambda)$$

This is equal to the mixture model of a gamma posterior distribution:

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

With exposure:

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

If we set the integral with limits 0 to infinity, we can let every constant be proportion to K:

$$f(\lambda\nu) = Gamma(\alpha, \beta) = \frac{1}{K} \int_0^{\infty} \lambda^{\alpha-1} e^{-\lambda\nu}$$

We set α=11, and ν=3767 for integral 1 and α=15 and ν = 2767 for integral 2.

Gamma(11, 3767) + Gamma(15, 2667) from *3c*.