Homework 2

PSTAT 115, Spring 2023

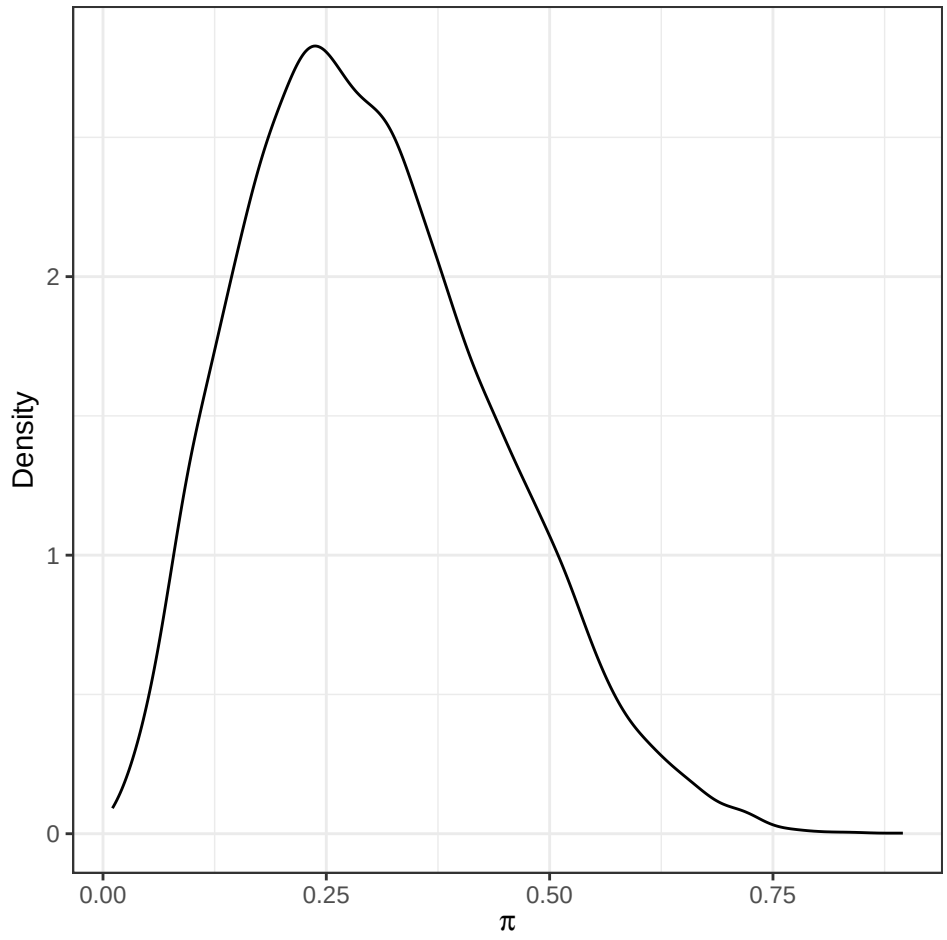
Due on May 5, 2023 at 11:59 pm

1. Knowing someone who is transgender

A September 2016 Pew Research survey found that 30% of U.S. adults are aware that they know someone who is transgender. It is now the 2020s, and Sylvia believes that the current percent of people who know someone who is transgender, π, has increased to somewhere between 35% and 60%.

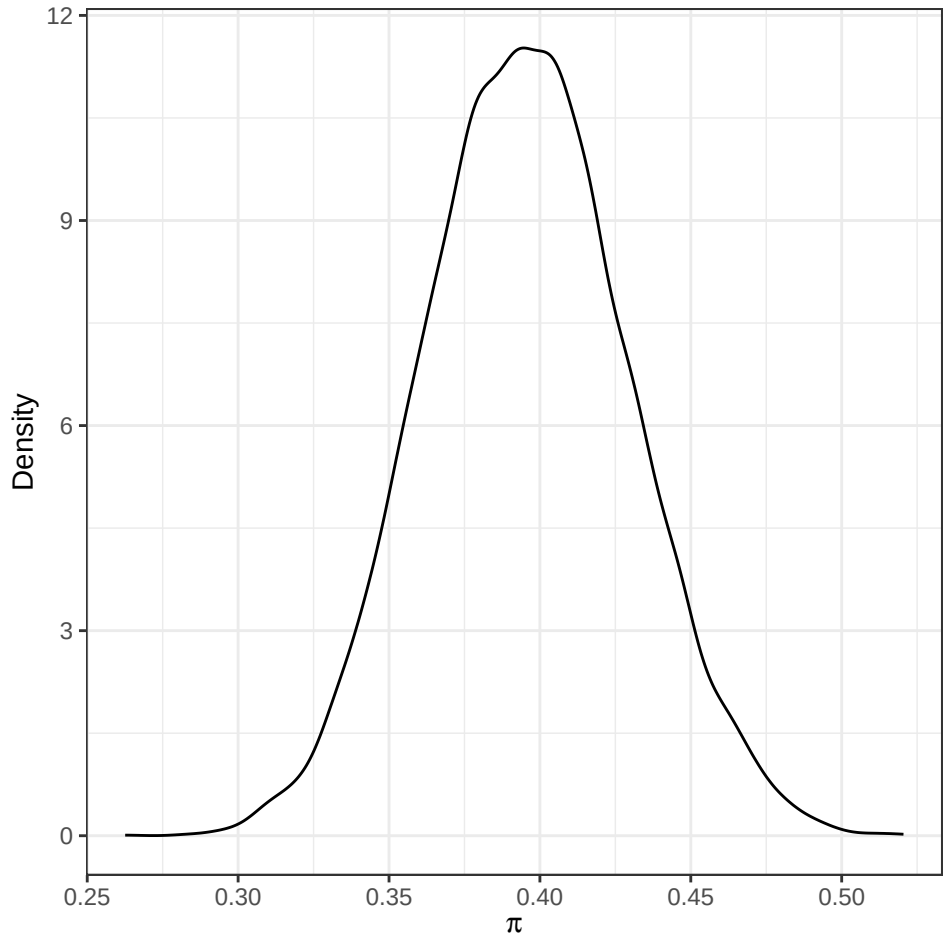
1a. (4pts) Identify and plot a Beta model that reflects Sylvia's prior ideas about π.

```
# sample a beta model with parameter 3 and 7
# plot the beta model

model = rbeta(10000, 3, 7)

ggplot(data.frame(x = model), aes(x)) +
  geom_density() +
  labs(x = expression(pi), y = "Density") +
  theme_bw()
```



1b. (4pts) Sylvia wants to update her prior, so she randomly selects 200 US adults and 80 of them are aware that they know someone who is transgender. Specify and plot the posterior model for π?

```
# update the beta model
updated_model = rbeta(10000, 3 + 80, 7 + 200 - 80)

ggplot(data.frame(x = updated_model), aes(x)) +
  geom_density() +
  labs(x = expression(pi), y = "Density") +
  theme_bw()
```



1c. (5pts) What is the mean, mode, and standard deviation of the posterior model?

```
mean(updated_model) # mean
```

```
## [1] 0.3954775
```

```
pmf = density(updated_model)
pmf$x[which.max(pmf$y)] # mode
```

```
## [1] 0.3946099
```

```
sd(updated_model) # sd
```

```
## [1] 0.03362514
```

calculated mean: 3+80/(3+7+200) = 83/210 = 0.3952 calculated mode: 3+80-1/(3+7+200-2) = 82/209 = 0.3923 calculated sd: sqrt((3+80)(7+200-80)/((3+7+200)^2(3+7+200-1))) = 0.0335

1d. (7pts) Describe how the prior and posterior Beta models compare. Hint: in class 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 prior and posterior Beta models are similar in shape, but the posterior model is more concentrated around the mean. This is because the posterior model is updated with the new information from the data. The posterior mean is the weighted average of the prior mean and the sample mean, where the weights are the prior and sample sizes, respectively. Since the sample size is much larger than the prior size, the posterior mean is closer to the sample mean than the prior mean. This is why the posterior model is more concentrated around the sample mean.

2. Sample survey

Suppose we are going to sample 100 individuals from a county (of size much larger than 100) and ask each sampled person whether they support policy Z or not. Let $Y_i = 1$ if person i in the sample supports the policy, and $Y_i = 0$ otherwise.

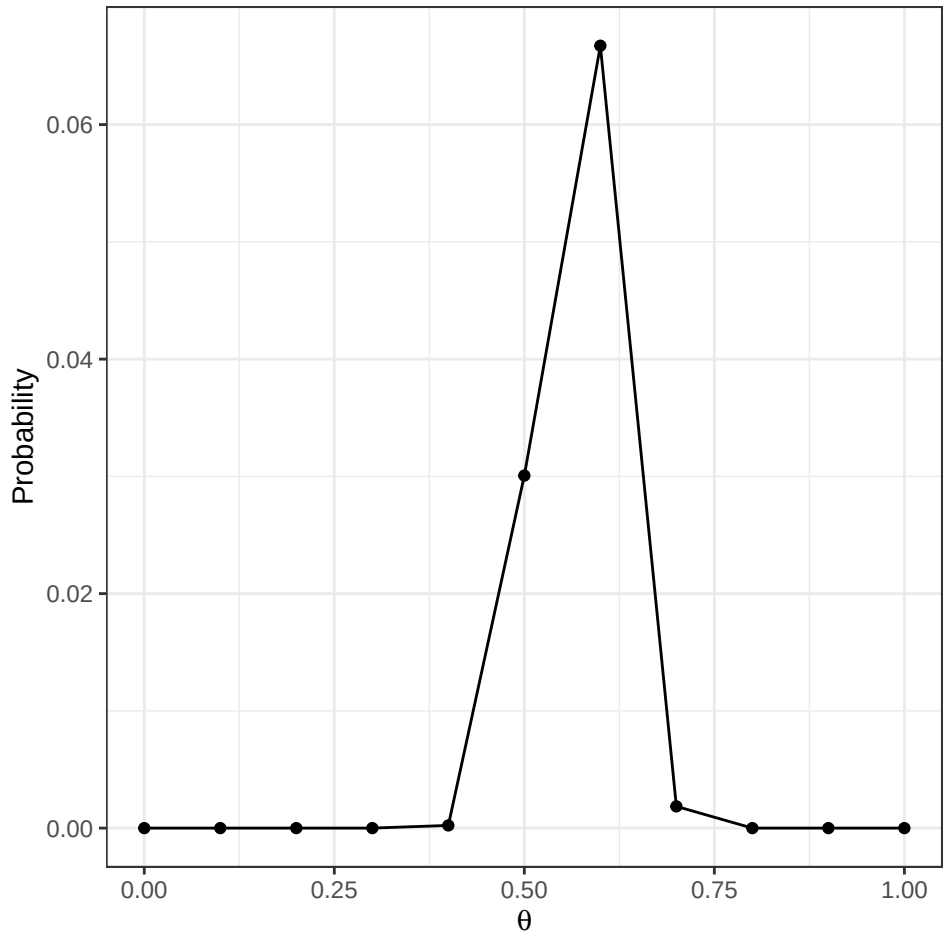
2a. (5pts) Assume $Y_1, \ldots, Y_{100}$ are, conditional on θ, iid binary random variables with expectation θ. Write down the joint distribution of $P(Y_1 = y_1, \ldots, Y_{100} = y_{100}) \mid \theta)$ in a compact form. Also write down the form of $P(\sum_{i=1}^{100} Y_i = y \mid \theta)$.

$$
\begin{aligned}
P(Y_1 = y_1, \ldots, Y_{100} = y_{100}|\theta) &= P(Y_1 = y_1|\theta) \cdots P(Y_{100} = y_{100}|\theta) && (1)\\
&= \prod_{i=1}^{100} P(Y_i = y_i|\theta) && (2)\\
&= \prod_{i=1}^{100} \theta^{y_i}(1-\theta)^{y_i} && (3)\\
&= \theta^{\sum y_i}(1-\theta)^{\sum y_i} && (4)\\
& && (5)
\end{aligned}
$$

$$
P(\sum_{i=1}^{100} Y_i = y|\theta) = \binom{100}{y}\theta^y(1-\theta)^{100-y} \quad (6)
$$

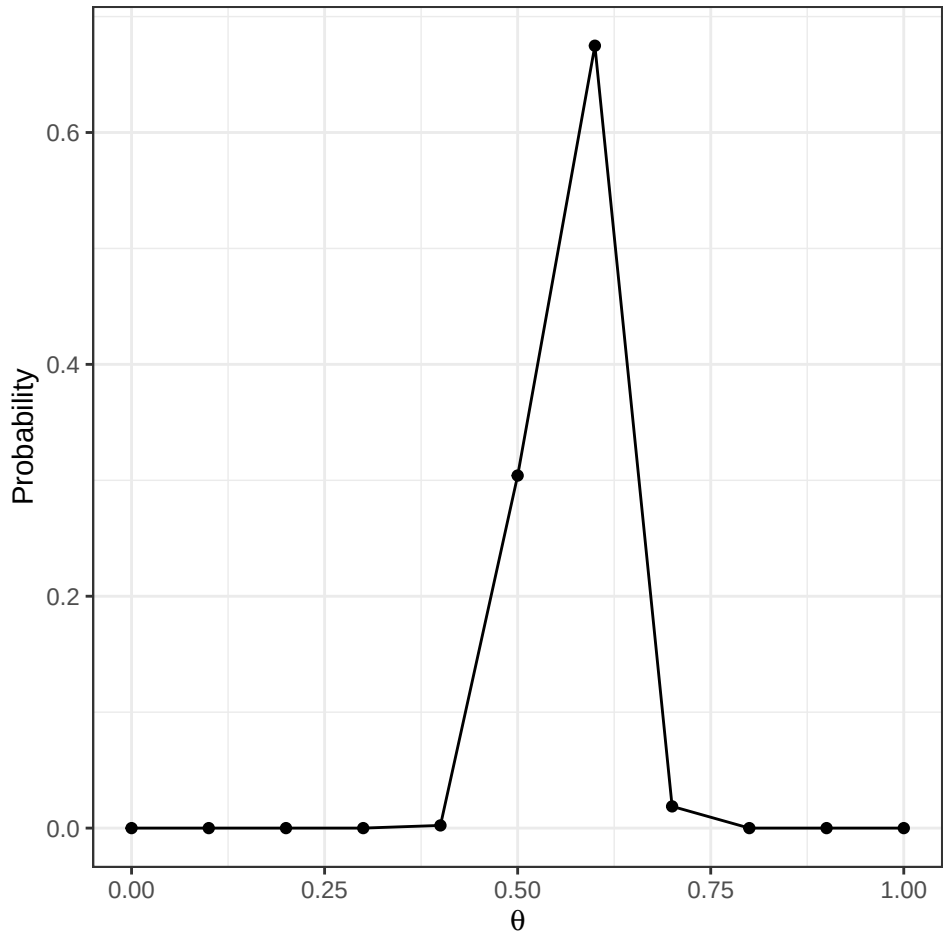
2b. (5pts) For the moment, suppose you believed that $\theta \in \{0.0, 0.1, \ldots, 0.9, 1.0\}$. Given that the results of the survey were $\sum_{i=1}^{100} y_i = 57$, compute $P(\sum_{i=1}^{100} Y_i = 57 \mid \theta)$ for each of the 11 values of θ and plot these probabilities.

```
theta = seq(0, 1, 0.1)
prob = dbinom(57, 100, theta)

ggplot(data.frame(x = theta, y = prob), aes(x, y)) +
  geom_point() +
  labs(x = expression(theta), y = "Probability") +
  geom_line() +
  theme_bw()
```



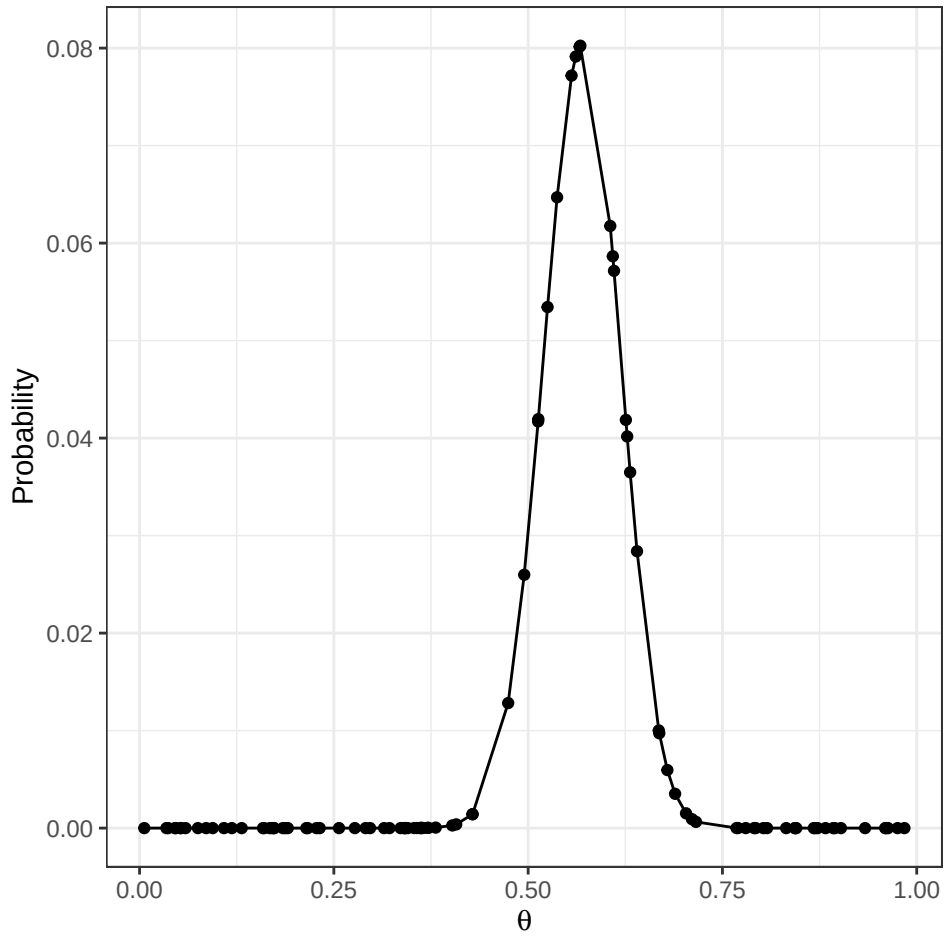
2c. (5pts) Now suppose you originally had no prior information to believe one of these θ-values over another, and so $P(\theta = 0.0) = P(\theta = 0.1) = \ldots = P(\theta = 1.0)$. Use Baye's rule to compute $p(\theta \mid \sum_{i=1}^{100} Y_i = 57)$ for each θ value. Make a plot of this posterior distribution as a function of θ.

```
# compute the posterior distribution
likelihood = prob
prior = 1/11
normalizing_constant = sum(likelihood * prior)
posterior = likelihood * prior / normalizing_constant

ggplot(data.frame(x = theta, y = posterior), aes(x, y)) +
  geom_point() +
  labs(x = expression(theta), y = "Probability") +
  geom_line() +
  theme_bw()
```



2d. (5pts) Now suppose you allow θ to be any value in the interval $[0, 1]$. Using the uniform prior density for θ, so that $p(\theta) = 1$, plot the *kernel* of the posterior density $p(\theta) \times P(\sum_{i=1}^{100} Y_i = 57 \mid \theta)$ as a function of θ.

```
# compute the posterior distribution
n = 100
theta = runif(n, 0, 1)
likelihood = dbinom(57, 100, theta)
prior = 1

kernel = likelihood * prior

ggplot(data.frame(x = theta, y = kernel), aes(x, y)) +
  geom_point() +
  labs(x = expression(theta), y = "Probability") +
  geom_line() +
  theme_bw()
```



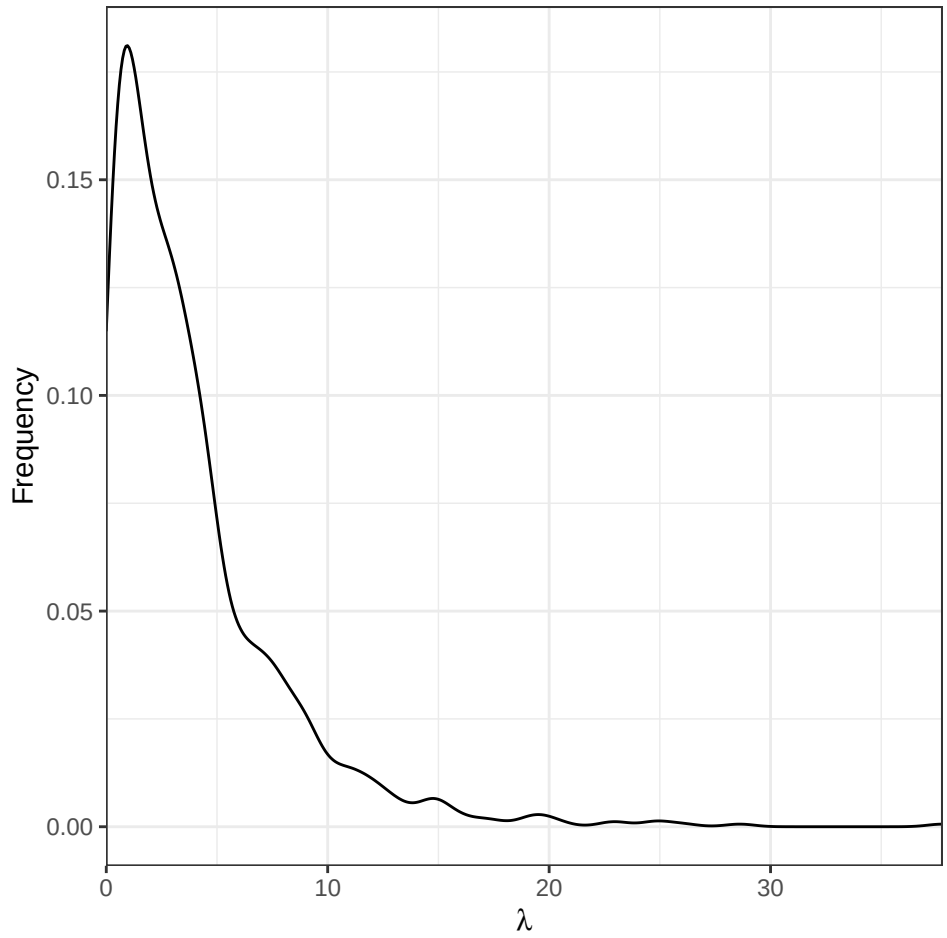
3. Soccer World cup

Let λ be the average number of goals scored in a Women's World Cup game. We'll analyze λ by the following a Y_i is the observed number of goals scored in a sample of World Cup games:

$$Y_i|\lambda \overset{ind}{\sim} \text{Pois}(\lambda)$$
$$\lambda \sim \text{Gamma}(1, 0.25)$$

3a. (5pts) Plot and summarize our prior understanding of λ.

```
lambda = rgamma(1000, 1, 0.25)

ggplot(data.frame(x = lambda), aes(x)) +
  geom_density() +
  labs(x = expression(lambda), y = "Frequency") +
  expand_limits(x = 0) +
  scale_x_continuous(expand = c(0, 0)) +
  theme_bw()
```




```
paste("mean:", 1/0.25)
```

```
## [1] "mean: 4"
```

```
paste("sd:", 1/sqrt(0.25))
```

```
## [1] "sd: 2"
```

```
paste("mode:", (1-1)/0.25)
```

```
## [1] "mode: 0"
```

3b. (5pts) Why is the Poisson model a reasonable choice for our data Y_i?

because the number of goal is a count variable and the poisson distribution is a discrete distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant mean rate and independently of the time since the last event.

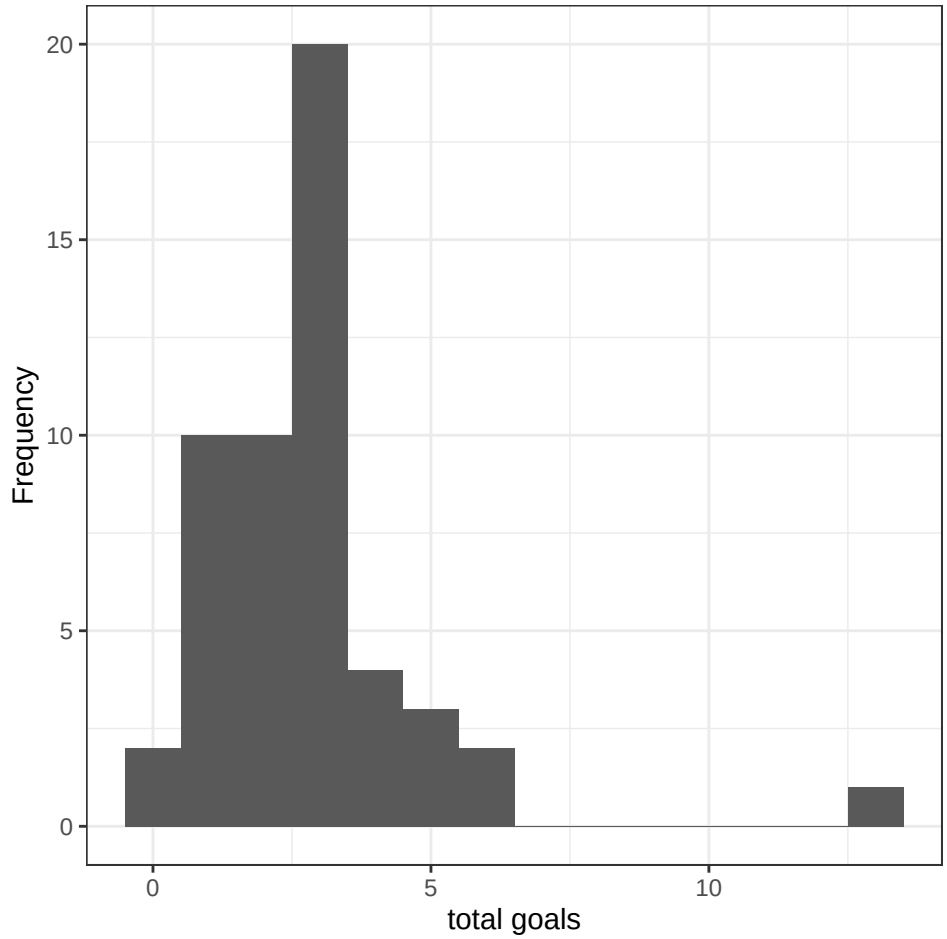
and the conjugate prior for the poisson distribution is the gamma distribution.

3c. (5pts) The `wwc_2019_matches` data in the *fivethirtyeight* package includes the number of goals scored by the two teams in each 2019 Women's World Cup match. Define, plot, and discuss the total number of goals scored per game:

```
library(fivethirtyeight)
```

```
## Some larger datasets need to be installed separately, like senators and
## house_district_forecast. To install these, we recommend you install the
## fivethirtyeightdata package by running:
## install.packages('fivethirtyeightdata', repos =
## 'https://fivethirtyeightdata.github.io/drat/', type = 'source')
```

```
data("wwc_2019_matches")
wwc_2019_matches <- wwc_2019_matches %>%
  mutate(total_goals = score1 + score2)

ggplot(data = wwc_2019_matches, aes(x = total_goals)) +
  geom_histogram(binwidth = 1) +
  labs(x = "total goals", y = "Frequency") +
  theme_bw()
```




```
paste("mean:", mean(wwc_2019_matches$total_goals))
```

```
## [1] "mean: 2.80769230769231"
```

```
paste("var:", var(wwc_2019_matches$total_goals))
```

```
## [1] "var: 3.92307692307692"
```

```
paste("sd:", sd(wwc_2019_matches$total_goals))
```

```
## [1] "sd: 1.98067587532057"
```

The total number of goals plot is a right skewed distribution with a mean of 2.8 and a standard deviation of 1.9.

3d. (5pts) Identify the posterior model, ie, what is the posterior distribution (including its parameters).

The posterior distribution is a gamma distribution with parameters $\alpha = 1 + \sum_{i=1}^{n} Y_i$ and $\beta = 0.25 + n$.

```
alpha = 1 + sum(wwc_2019_matches$total_goals)
beta = 0.25 + nrow(wwc_2019_matches)
paste("alpha:", alpha)
```

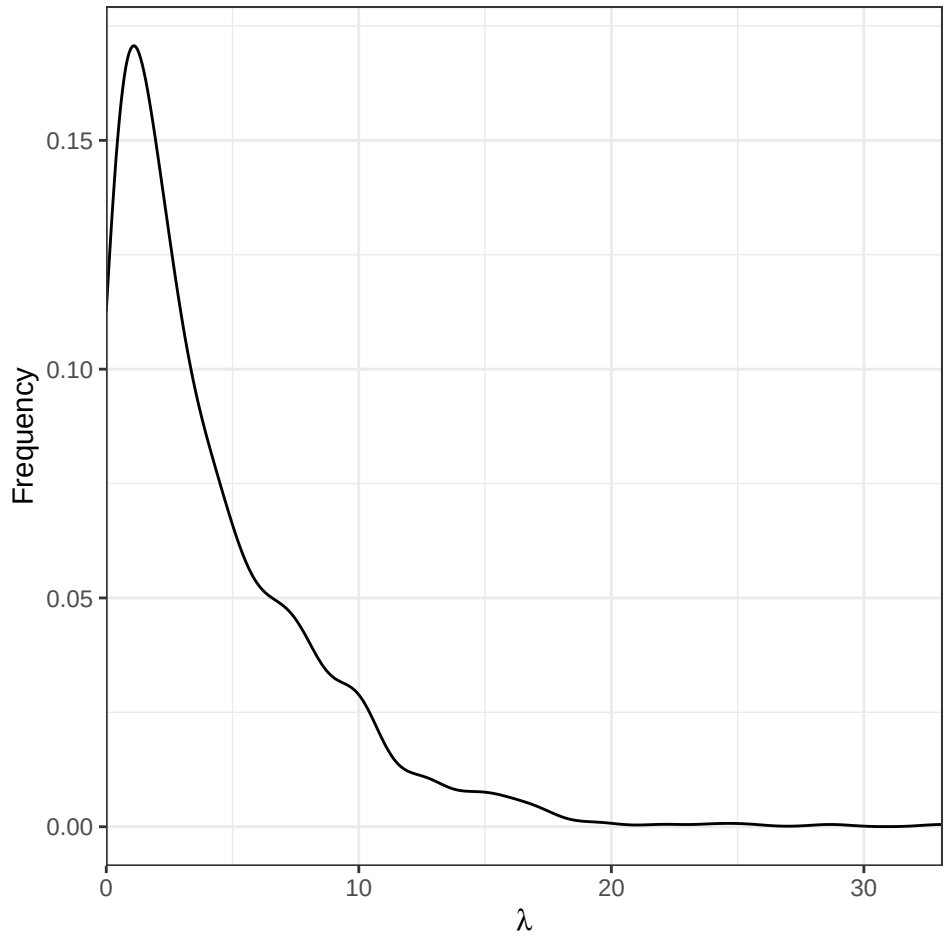
```
## [1] "alpha: 147"
```

```
paste("beta:", beta)
```

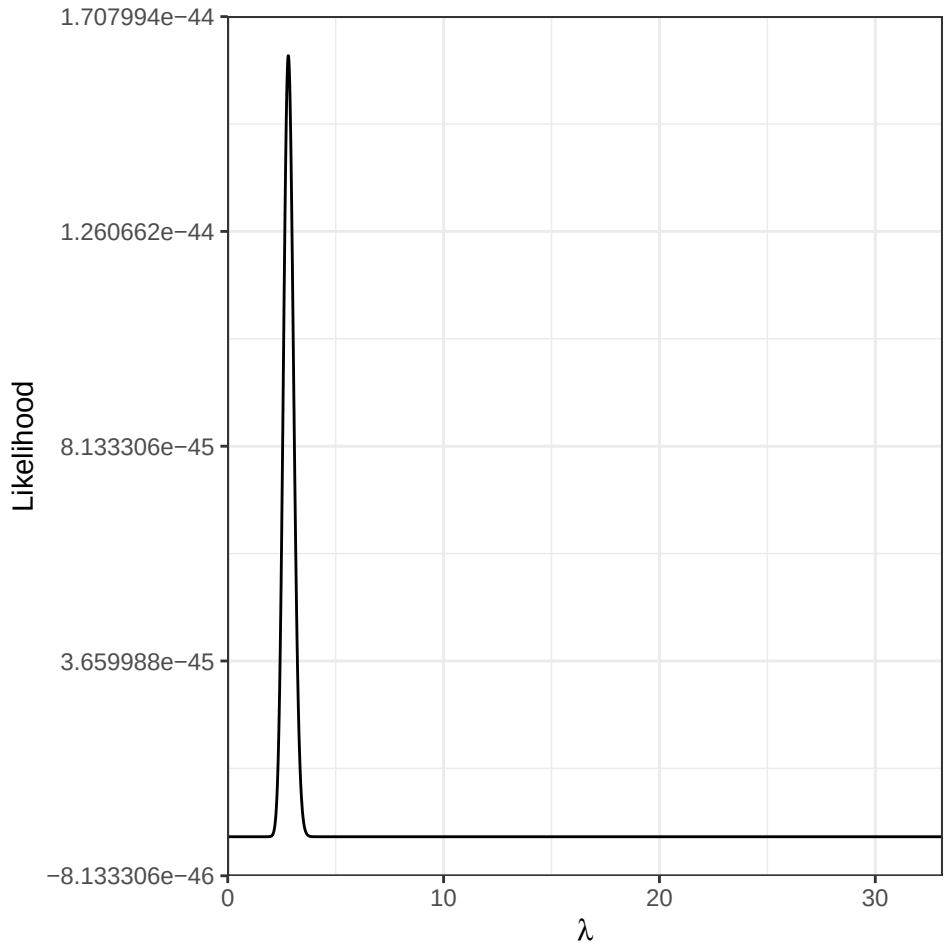
```
## [1] "beta: 52.25"
```

3e. (5pts) Plot the prior, likelihood and posterior for λ. Describe the evolution in your understanding of λ from the prior to the posterior.

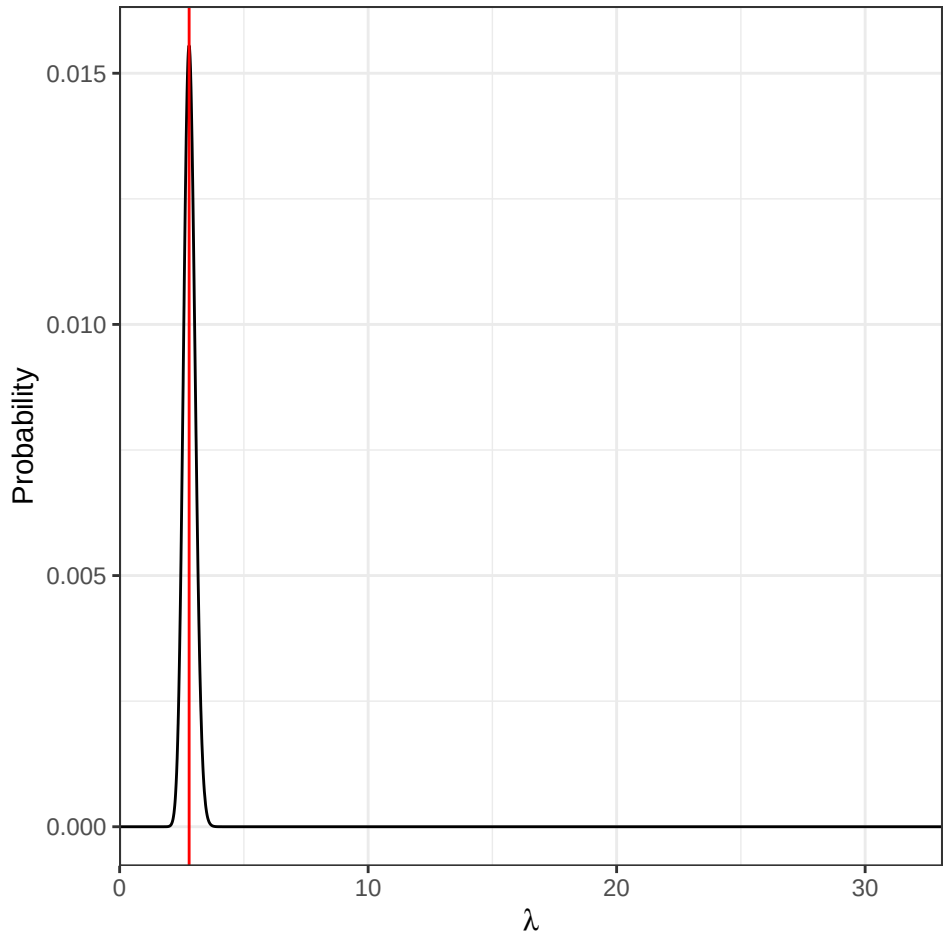
```
# compute the posterior distribution

lambda = rgamma(1000, 1, 0.25)
lambda_prior = dgamma(lambda, 1, 0.25)

likelihood_cal = function (lambda) {
  prod(dpois(wwc_2019_matches$total_goals, lambda))
}

likelihood = sapply(lambda, likelihood_cal)

normalizing_constant = sum(likelihood * lambda_prior)
posterior_lambda = likelihood * lambda_prior / normalizing_constant

# plot prior
ggplot(data.frame(x = lambda), aes(x)) +
  geom_density() +
  labs(x = expression(lambda), y = "Frequency") +
  expand_limits(x = 0) +
  scale_x_continuous(expand = c(0, 0)) +
  theme_bw()
```




```
# plot likelihood
ggplot(data.frame(y = likelihood, x = lambda), aes(x, y)) +
  geom_line() +
  labs(x = expression(lambda), y = "Likelihood") +
  expand_limits(x = 0) +
  scale_x_continuous(expand = c(0, 0)) +
  theme_bw()
```




```
# plot posterior
ggplot(data.frame(x = lambda, y = posterior_lambda), aes(x, y)) +
  labs(x = expression(lambda), y = "Probability") +
  geom_line() +
  expand_limits(x = 0) +
  scale_x_continuous(expand = c(0, 0)) +
  geom_vline(xintercept = lambda[which.max(posterior_lambda)], color = "red") +
  theme_bw()
```



My understanding of prior is that it is a gamma distribution with parameters $\alpha = 1$ and $\beta = 0.25$. By providing evidence for each new number of goals that follows a poisson distribution with rate that connects to the prior, a likelihood function for the parameter λ is constructed. And the posterior distribution is calculated given the likelihood of each λ and the probability in prior distribution for that rate to occur.

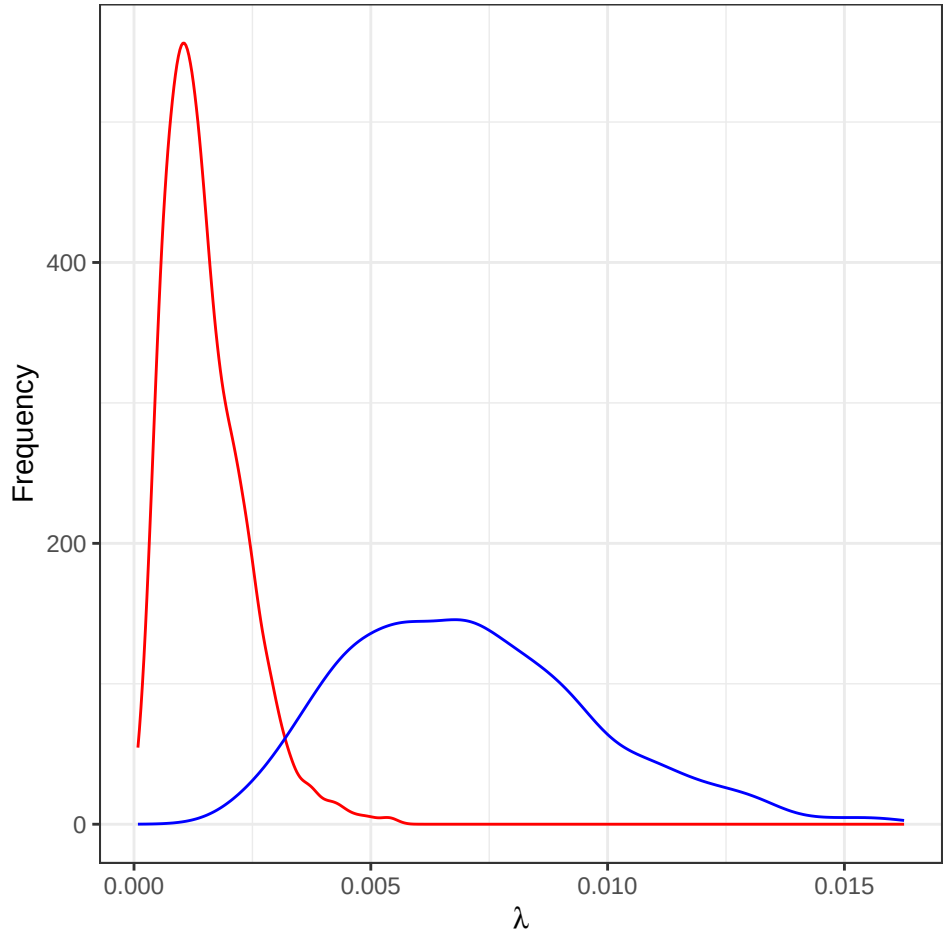
4. A Mixture Prior for Heart Transplant Surgeries

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: $p(\lambda) = 0.5 * p_1(\lambda) + 0.5 * p_2(\lambda)$

4a. (10pts) 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)?

```
expert1 = rgamma(1000, 3, 2000)
expert2 = rgamma(1000, 7, 1000)
```

```
# plot both
ggplot(data.frame(e1 = expert1, e2 = expert2)) +
  geom_density(aes(x = e1), color = "red") +
  geom_density(aes(x = e2), color = "blue") +
  labs(x = expression(lambda), y = "Frequency") +
  theme_bw()
```




```
sd(expert1)
```

```
## [1] 0.0008387552
```

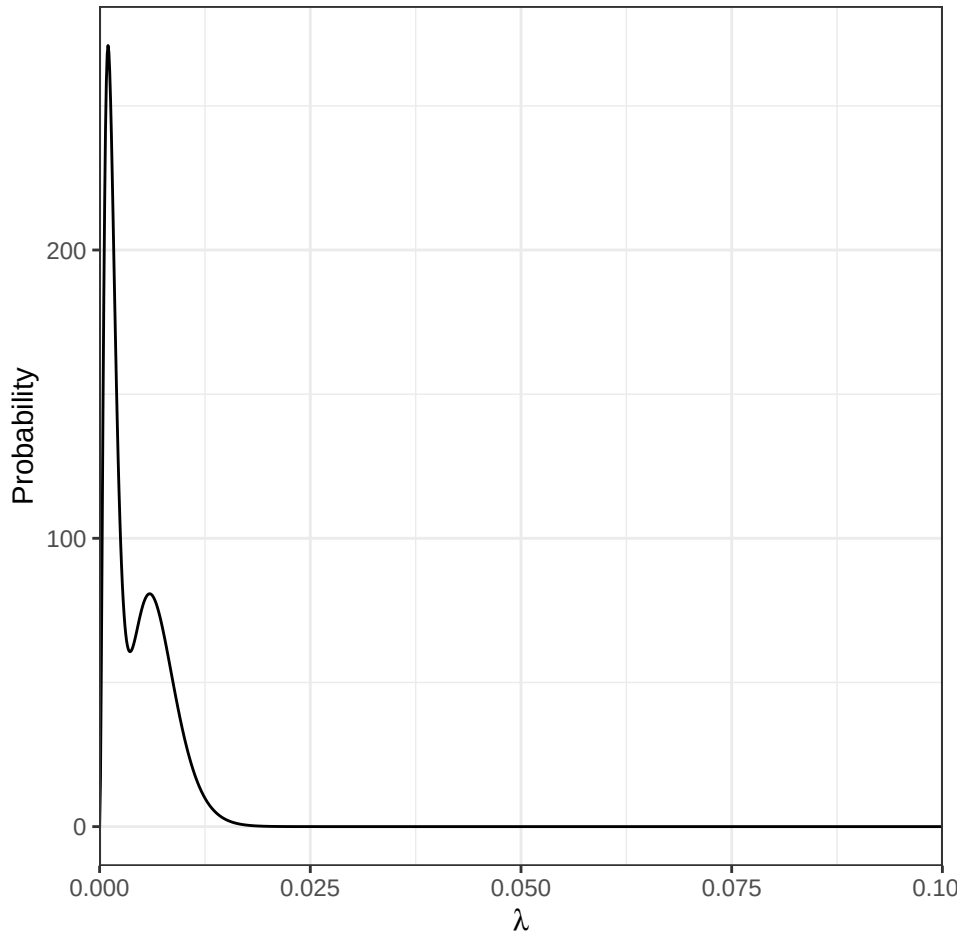
```
sd(expert2)
```

```
## [1] 0.002629382
```

The first expert thinks that the mean rate is 0.0015 and the second expert thinks that the mean rate is 0.007. The first expert is more confident about the value of λ a priori because the variance is smaller.

4b. (5pts) Plot the mixture prior distribution.

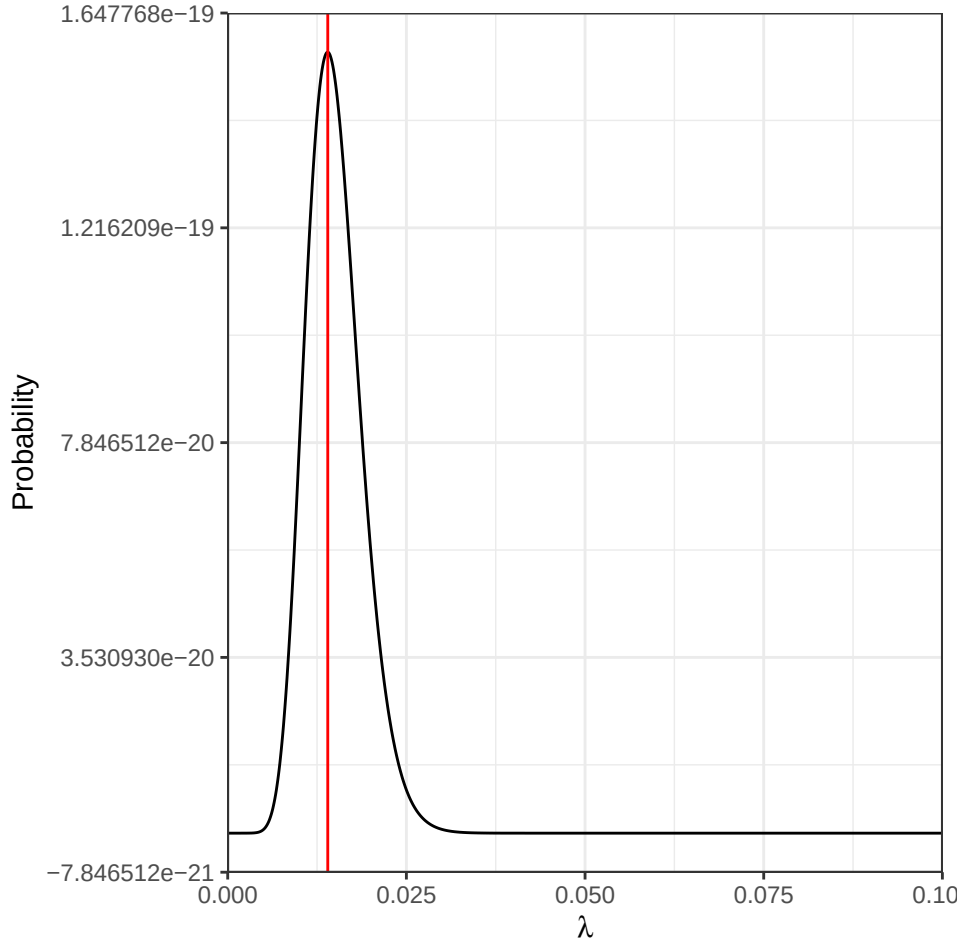
```r
# plot mixture
lambda = seq(0, 0.1, 0.0001)
p1 = 0.5*dgamma(lambda, 3, 2000)
p2 = 0.5*dgamma(lambda, 7, 1000)

p = p1 + p2

ggplot(data.frame(x = lambda, y = p), aes(x, y)) +
  geom_line() +
  labs(x = expression(lambda), y = "Probability") +
  expand_limits(x = 0) +
  scale_x_continuous(expand = c(0, 0)) +
  theme_bw()
```



4c. (10pts) 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.

```r
lambda_prior = p1+p2
likelihood_cal = function (lambda) {
  prod(dpois(8, lambda))
```

```
}
likelihood = sapply(lambda, likelihood_cal)
unnornmalized_posterior = likelihood * lambda_prior

mle = lambda[which.max(unnornmalized_posterior)]

ggplot(data.frame(x = lambda, y = unnornmalized_posterior), aes(x, y)) +
  labs(x = expression(lambda), y = "Probability") +
  geom_line() +
  expand_limits(x = 0) +
  scale_x_continuous(expand = c(0, 0)) +
  geom_vline(xintercept = mle, color = "red") +
  theme_bw()
```



$$f(\lambda|y_i) \propto \frac{1}{\prod y_i!}(\nu\lambda)^{\sum y_i} e^{-n\nu\lambda}\left(\frac{\beta_1^{\alpha_1}}{2\Gamma(\alpha_1)}\lambda^{\alpha_1-1}e^{\beta_1\lambda_1} + \frac{\beta_2^{\alpha_1}}{2\Gamma(\alpha_2)}\lambda^{\alpha_2-1}e^{\beta_2\lambda_1}\right) \tag{7}$$

$$\propto \lambda^{\sum y_i+\alpha_1-1}e^{-(\beta_1+n\nu)\lambda} + \lambda^{\sum y_i+\alpha_2-1}e^{-(\beta_2+n\nu)\lambda} \tag{8}$$

$$\propto \lambda^{10}e^{-(3767)\lambda} + \lambda^{14}e^{-(2767)\lambda} \tag{9}$$

4e. (10pts) 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.

$$p(\lambda|y) = \frac{p(y|\lambda=\lambda)p(\lambda)}{\int p(y|\lambda=\lambda)p(\lambda)d\lambda}$$

$$\int_0^\infty p(y|\lambda=\lambda)p(\lambda)\,d\lambda$$

$$= \frac{\sqrt{\sum y_i}}{\prod y_i}\left(\frac{b_0^{a_0}}{\Gamma(a_0)}\int_0^\infty \lambda^{a_0+\sum y_i-1} e^{-(b_0+n)\lambda}\right.$$

$$\left. + \frac{b_1^{a_1}}{\Gamma(a_1)}\int_0^\infty \lambda^{a_1+\sum y_i-1} e^{-(b_1+n)\lambda}\right)$$

$$p(\lambda|y) = \frac{p(y|\lambda=\lambda)p(\lambda)}{\int p(y|\lambda=\lambda)p(\lambda)d\lambda}$$

$$= \frac{\beta_0^{a_0}}{\Gamma(a_0)}\left(\lambda^{a_0+\Sigma y_i-1}e^{-(\beta_0+nv)\lambda}\right) + \frac{\beta_1^{a_1}}{\Gamma(a_1)}\left(\lambda^{a_1+\Sigma y_i-1}e^{-(\beta_1+nv)\lambda}\right)$$

$$\frac{\beta_0^{a_0}}{\Gamma(a_0)}\int_0^\infty \lambda^{a_0+\Sigma y_i-1}e^{-(\beta_0+nv)\lambda}\Big) + \frac{\beta_1^{a_1}}{\Gamma(a_1)}\int_0^\infty \lambda^{a_1+\Sigma y_i-1}e^{-(\beta_1+nv)\lambda}d\lambda$$

$$= \frac{\beta_0^{a_0}}{\Gamma(a_0)}\left(\lambda^{a_0+\Sigma y_i-1}e^{-(\beta_0+nv)\lambda}\right) + \frac{\beta_1^{a_1}}{\Gamma(a_1)}\left(\lambda^{a_1+\Sigma y_i-1}e^{-(\beta_1+nv)\lambda}\right)$$

$$\frac{\beta_0^{a_0}}{\Gamma(a_0)}\frac{\Gamma(a_0+\Sigma y_i)}{(\beta_0+nv)^{a_0+\Sigma y_i}} + \frac{\beta_1^{a_1}}{\Gamma(a_1)}\frac{\Gamma(a_1+\Sigma y_i)}{(\beta_1+nv)^{a_1+\Sigma y_i}}$$