M234-HW1

Zixi Zhang

2023-01-28

```r
# Set working directory
setwd("/Users/bruce/Documents/23winter/M234/234HW/HW1")
getwd()
```

```
## [1] "/Users/bruce/Documents/23winter/M234/234HW/HW1"
```

```r
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(tidyr)
```

Problem 1: Normal data with a normal prior

1. Explain what your measurements will be.

I used a OMRON blood pressure monitor to measure my pulse.

2. Before you collect the data, decide on your prior.

My pulse is a bit higher than normal person, so I guess the mean would be 85, and the standard deviation would be 4. i.e $\mu_0 = 85, \tau = 4$.The prior distribution is $N(85, 16)$.

3. Report the data and the sample mean and variance $(n - 1)$ denominator.

```r
df <- data.frame(
  id = c(1,2,3,4,5,6),
  pulse = c(95,86,88,83,84,86)
  )
df
```

```
##   id pulse
## 1  1    95
## 2  2    86
## 3  3    88
## 4  4    83
## 5  5    84
## 6  6    86
```

```r
pulse_mean <- mean(df$pulse)
pulse_var <- var(df$pulse)
print(c(pulse_mean,pulse_var))
```

```
## [1] 87.0 18.4
```

The data contains 6 measurements, so the sample size = 6. The sample mean is 87.0 and the sample variance $(n-1)$ denominator is 18.4.

4. Now specify the sampling standard deviation σ.

Since we are doing a one parameter model, and since σ is usually not known, we need to do something because we are working with such a simple model. You may either (a) Pick a value for σ yourself, or (b) Set σ to the sample sd of your data set. (c) Specify the exact value for *sigma* that you use in all your calculations (i.e. sqrt(2), 1.41, 1.414, or 1.4)

I set σ to the sample sd of my data($\sqrt{18.4}$) here.

5. Calculate the posterior mean $\bar{\mu}$, posterior variance V, and posterior sd. Show the formulas for the posterior mean and variance with your data values in place of the symbols. Remember that in the likelihood, $\bar{y} \sim N\left(\mu, \sigma^2/n\right)$

$$\bar{\mu} = \left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1} \left(\frac{n\bar{y}}{\sigma^2} + \frac{\mu_0}{\tau^2}\right) \quad = \frac{\frac{n}{\sigma^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}\bar{y} + \frac{\frac{1}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}\mu_0$$

$$V = \frac{\frac{\tau^2\sigma^2}{n}}{\tau^2 + \frac{\sigma^2}{n}}$$

$$sd = \sqrt{V}$$

```r
pluse_sd <- sqrt(pulse_var)
n <- nrow(df)
mu_0 <- 85
tau <- 4
mu_bar <- (n / pulse_var) / ( n / pulse_var + 1 / (tau**2) ) * pulse_mean + (1 / (tau**2) ) / ( n / puls
V <- (((tau**2) * pulse_var) / n)/ ((tau**2)+(pulse_var/n))
```

```r
posterior_sd <- sqrt(V)
list_post <- c(mu_bar,V,posterior_sd)
lapply(list_post,round,3)
```

```
## [[1]]
## [1] 86.678
##
## [[2]]
## [1] 2.573
##
## [[3]]
## [1] 1.604
```

$\bar{\mu} = 86.678$, posterior variance $V = 2.573$
posterior sd $= 1.604$

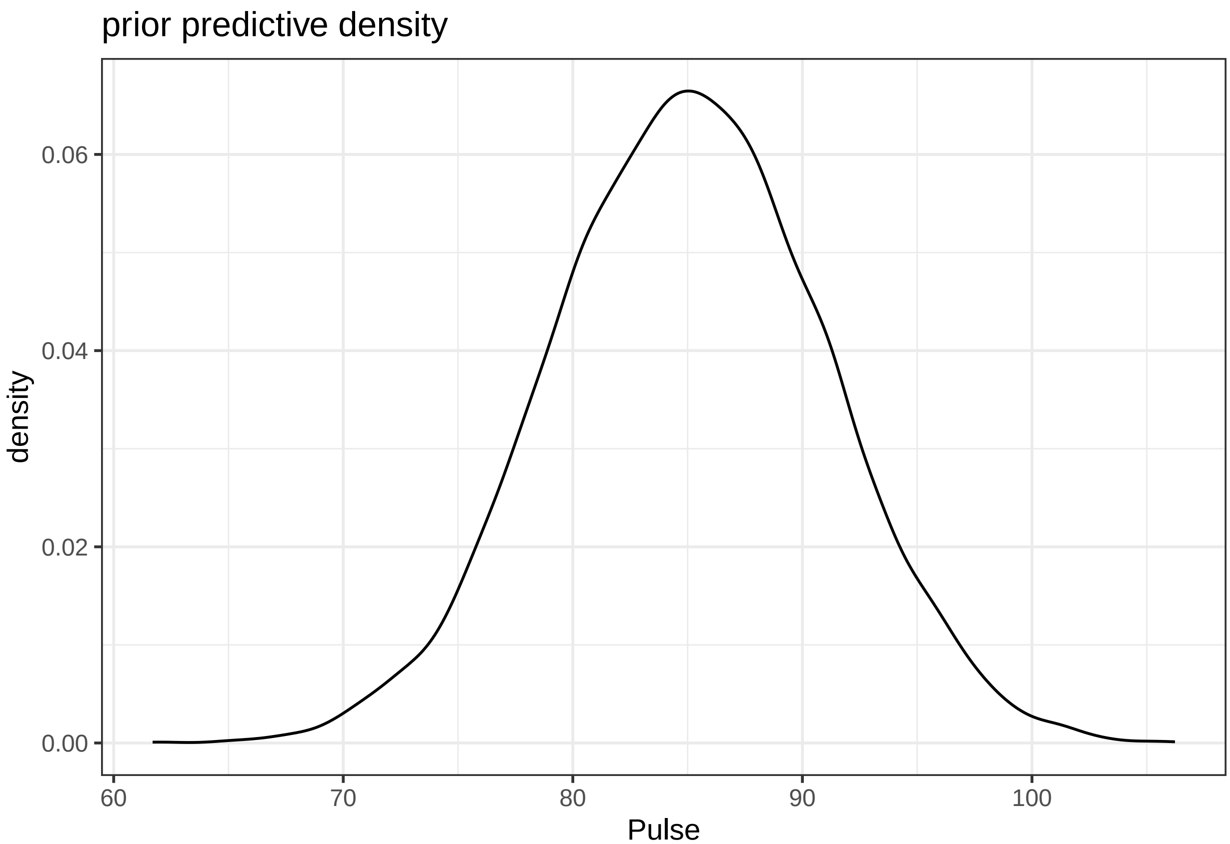
6. The prior predictive density is the density that you predict for a single observation before seeing any data. In this model, the prior predictive for a single observation is $y \sim N\left(\mu_0, \sigma^2 + \tau^2\right)$.

$y \sim N(85, 34.4)$

```r
set.seed(1997)
prior_dist <- data.frame(Pulse =rnorm(10000, mean = mu_0, sd = sqrt(pulse_var+tau**2)))
head(prior_dist)
```

```
##      Pulse
## 1 80.49845
## 2 90.30849
## 3 78.50958
## 4 84.34272
## 5 74.80928
## 6 86.34655
```

```r
ggplot(prior_dist, aes(x=Pulse)) +
  geom_density(alpha = 0.4) + labs(title="prior predictive density") +theme_bw()
```



7. Construct a table with means, sds and vars for the (i) posterior for μ, (ii) the prior for μ, (iii) the prior predictive for y, and (iv) the likelihood of μ.

```
df_posterior <- data.frame(posterior = c(mu_bar,V,posterior_sd))
df_prior <- data.frame(prior = c(mu_0,tau**2,tau))
df_priorpred <- data.frame(priorpredictive = c(mu_0,pulse_var+tau**2,sqrt(pulse_var+tau**2)))
df_likelihood <- data.frame(likelihood = c(pulse_mean,pulse_var,sqrt(pulse_var)))
df_table = cbind(df_posterior,df_prior,df_priorpred,df_likelihood)
rownames(df_table) <- c("Mean","Variance","Std. Deviation")
knitr::kable(df_table,
             caption = "Table of means, sds and vars for posterior, prior, prior predictive and likelih
             digits = 3,
             align = "cccc",
             col.names = c("posterior","prior","prior predictive", "likelihood"))
```

Table 1: Table of means, sds and vars for posterior, prior, prior predictive and likelihood

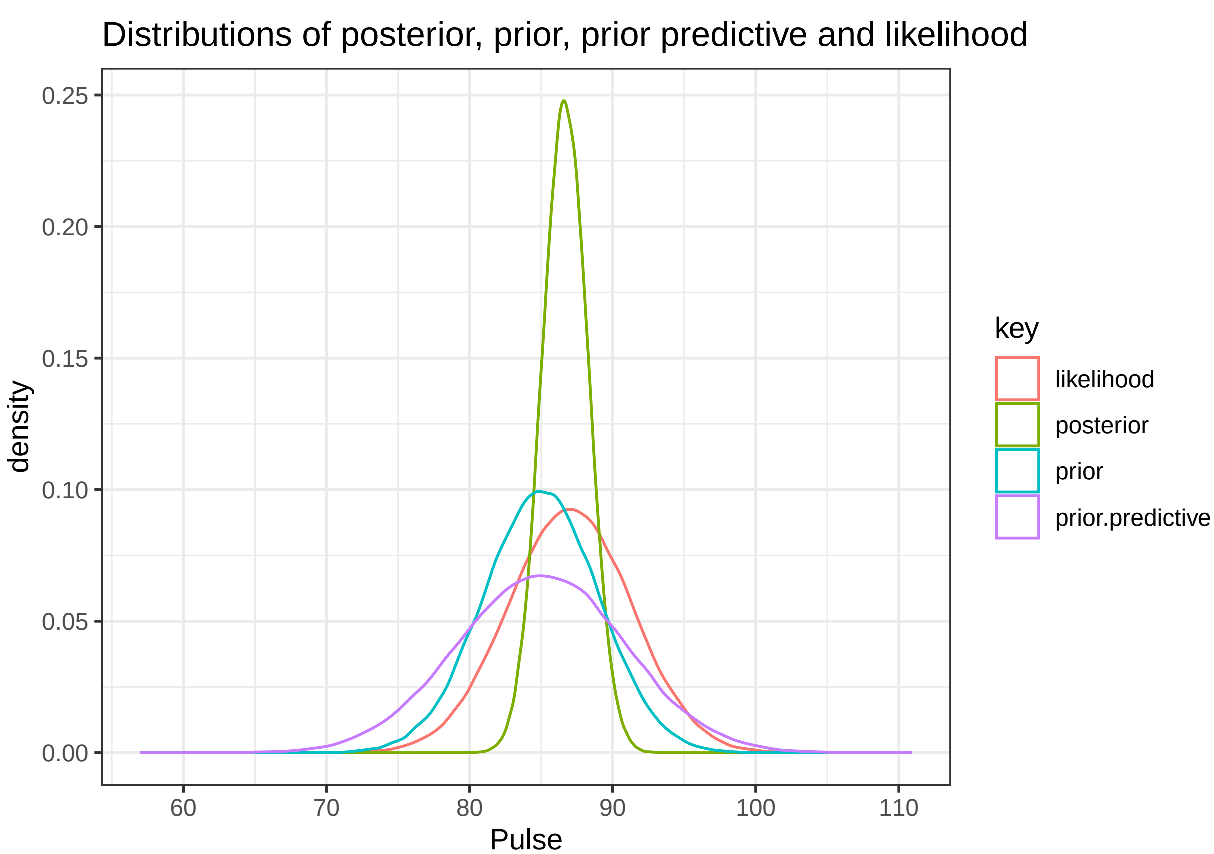
	posterior	prior	prior predictive	likelihood
Mean	86.678	85	85.000	87.00
Variance	2.573	16	34.400	18.40
Std. Deviation	1.604	4	5.865	4.29

8. Plot on a single plot the (i) posterior for μ, (ii) the prior for μ, (iii) the prior predictive for y, and (iv) the likelihood of μ (suitably normalized so it looks like a density, ie a normal with mean $\bar{y}$ and variance σ^2/n) all on the same graph. Interpret the plot.

```
set.seed(1997)
posterior_dist <- data.frame(posterior = rnorm(100000, mu_bar, posterior_sd))
prior_dist <- data.frame(prior = rnorm(100000, mu_0, tau))
predictive_dist <- data.frame(prior.predictive = rnorm(100000, mu_0, sqrt(pulse_var+tau**2)))
likelihood_dist <- data.frame(likelihood = rnorm(100000, pulse_mean, sqrt(pulse_var)))
combine_dist <- cbind(posterior_dist, prior_dist, predictive_dist, likelihood_dist)
plot_dist <- gather(combine_dist)
head(plot_dist)
```

```
##         key    value
## 1 posterior 85.44709
## 2 posterior 88.13026
## 3 posterior 84.90311
## 4 posterior 86.49855
## 5 posterior 83.89104
## 6 posterior 87.04662
```

```
ggplot(plot_dist, aes(x = value,col = key)) +
  geom_density(alpha = 0.4) +
  labs(title = "Distributions of posterior, prior, prior predictive and likelihood",
      x = "Pulse")  + theme_bw()
```



The mean of posterior is between the means of prior and likelihood. Prior and prior predictive distributions have the same mean, but the prior predictive has a wider distribution because of larger variance.

Problem 2: Count Data with a Gamma Prior

For $y_i \mid \lambda \sim \text{Poisson}(\lambda), i = 1, \ldots, n$, the conjugate prior is $\lambda \sim \text{Gamma}(a, b)$. The parameter b is the rate parameter and the mean of the Gamma(a, b) distribution is a/b and the variance is a/b^2. The posterior given a sample of size n will be Gamma $(a + \sum_i y_i, b + n)$. You can calculate a gamma (a, b) density using dgamma (x, shape = a, rate = b, log = FALSE), or by calculating the density yourself $f(x \mid a, b) = b^a * x^{(a-1)} \exp(-b * x)/\,\text{gamma}(a)$, where gamma (a) is the gamma function.

1. What is the support (place where density/function is non-negative) of: (i) prior, (ii) posterior, (iii) sampling density, (iv) likelihood?

- (i) Prior: $\lambda \sim Gamma(a, b)$, and the support is $\lambda \in (0, \infty)$
- (ii) Posterior: $\lambda | y \sim Gamma(a + \sum_i y_i, b + n)$, and the support is $\lambda \in (0, \infty)$
- (iii) sampling density: $y|\lambda \sim Poisson(\lambda)$, and the support is $y \in \mathbb{N}_0$
- (iv) likelihood: $y|\lambda \sim Poisson(n\lambda)$, and the support is $y \in \mathbb{N}_0$

2. In the prior gamma(a, b), which parameter acts like a prior sample size? (Hint: look at the posterior, how does n enter into the posterior density?) You will need this answer later.

b acts like a prior sample size.

3. You will go (soon, but not yet!) to your favorite store entrance and count the number of customers entering the store in a 5 minute period. Collect it as 5 separate observations $y_1, \ldots, y_5$ of 1 minute duration each, this allows you to blink and take a break if needed. This will give you 5 data points.

4. Name your store, and the date and time.

My store is Shake Shack at LAS Airport, NV at 6:43am to 6:47am on Jan 29th.

5. We are now going to specify the parameters a and b of the gamma prior density. We will do this in two different ways, giving two different priors. We designate one set of prior parameters as a_1 and b_1; the other set of prior parameters are a_2 and b_2.

(a) Before you visit the store, make a guess as to the mean number of customers entering the store in one minute. Call this m_0. This is the mean of your prior distribution for λ.

The time I plan to visit the store is early in the morning so I don't expect there are too many customers. I set $m_0 = 4$

(b) Make a guess s_0 of the prior sd associated with your estimate m_0. This s_0 is the standard deviation of the prior distribution for λ. Note: most people underestimate s_0.

This is a store behind the security check so the customer flow should be stable. Set $s_0 = 2$

(c) Separately from the previous question 5 b, estimate how many data points n_0 your prior guess is worth. That is, n_0 is the number (strictly greater than zero) of data points (counts of 5 minutes) you would just as soon have as have your prior guess of m_0.

$n_0 = 1$

(d) Solve for a_1 and b_1 based on m_0 and s_0.

$$E(\lambda) = \frac{a_1}{b_1} = 4$$
$$Var(\lambda) = \frac{a_1}{{b_1}^2} = 2$$

So $a_1 = 8, b_1 = 2$

(e) Separately solve for a_2 and b_2 using m_0 and n_0 only. You usually will not get the same answer each time. This is ok and is NOT wrong. (Note: if you do get the same answer, then please specify a second choice of a_2, b_2 to use with the remainder of this problem!)

$$E(\lambda) = \frac{a_2}{b_2} = 4$$

So $a_2 = 4, b_2 = 1$

6. Suppose we need to have a single prior, rather than two priors. Suggest 2 distinct methods to settle on a single prior.

1. Take the average of two sets of prior parameters.
2. choose a_2 and b_2 because most people underestimate s_0 but estimation of a_2 and b_2 did not use s_0.

7. Go to your store and collect your data as instructed in 3. Report it here.

```
q2data <- data.frame(
  minute = c(1,2,3,4,5),
  customer = c(3,5,6,2,7)
  )
knitr::kable(q2data,
             caption = "number of customers entering Shake Shack in a 5 minute period",
             col.names = c("Minute","Number of customers"),
             align = "cc")
```

Table 2: number of customers entering Shake Shack in a 5 minute period

Minute	Number of customers
1	3
2	5
3	6
4	2

Minute	Number of customers
5	7

8. Update both priors algebraically using your 5 data points. Give the two posteriors.

$$\text{Gamma}\left(a + \sum_i y_i, b + n\right)$$

For a_1 and b_1:

```
a1 = 8
b1 = 2
a_post1 = a1 + sum(q2data$customer)
b_post1 = b1 + length(q2data$customer)
c(a_post1,b_post1)
```

```
## [1] 31  7
```

$\text{posterior : Gamma}(31, 7)$

For a_2 and b_2:

```
a2 = 4
b2 = 1
a_post2 = a2 + sum(q2data$customer)
b_post2 = b2 + length(q2data$customer)
c(a_post2,b_post2)
```

```
## [1] 27  6
```

$\text{posterior : Gamma}(27, 6)$

9. Give the posterior mean and sd for your two posteriors.

$\text{Mean} = \frac{a_1}{b_1} = 4 \text{ sd} = \sqrt{\frac{a_1}{{b_1}^2}}$

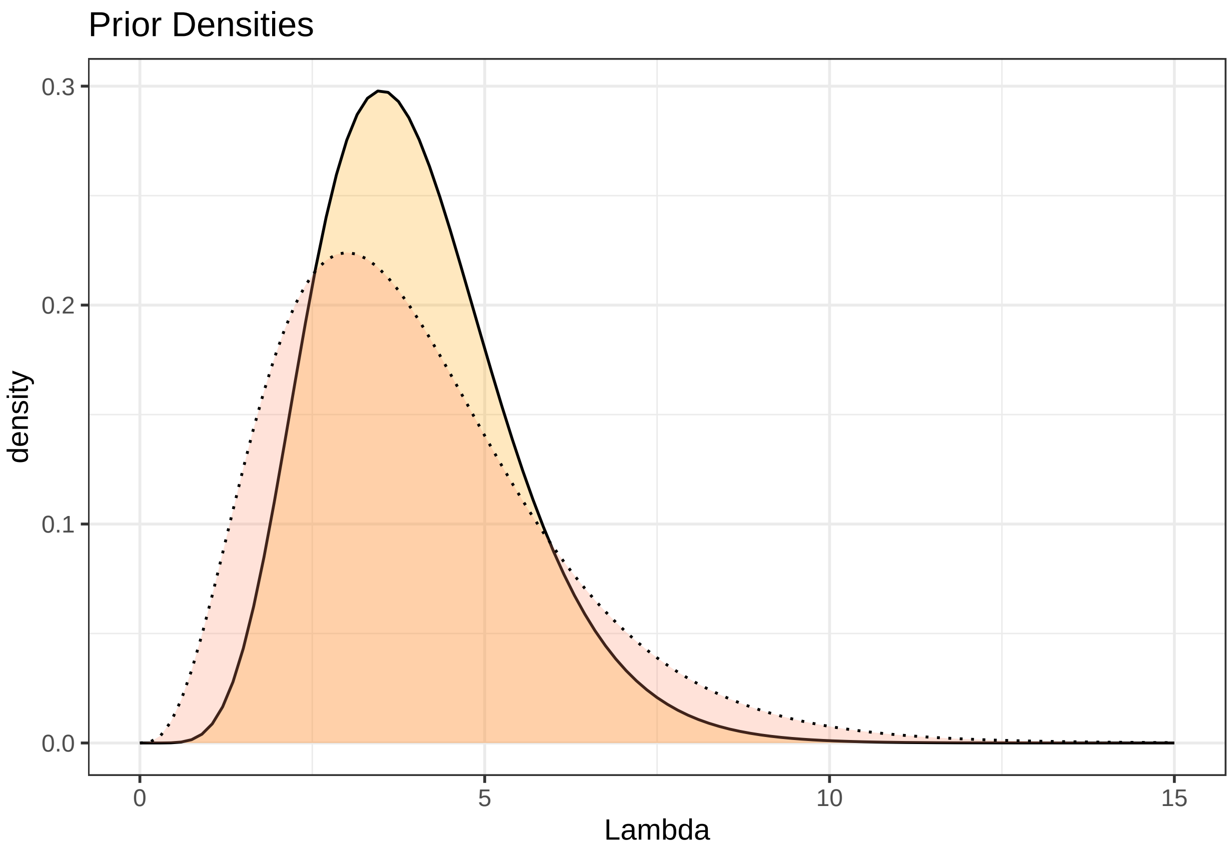
```
mean_post1 = a_post1 / b_post1
sd_post1 = sqrt(a_post1 / (b_post1)^2)
mean_post2 = a_post2 / b_post2
sd_post2 = sqrt(a_post2 / (b_post2)^2)
sum_post = data.frame(Mean = c(mean_post1,mean_post2),
                      sd = c(sd_post1,sd_post2),
                      row.names = c("posterior1","posterior2"))
knitr::kable(sum_post,
             caption = " posterior mean and sd for the two posteriors",
             align = "cc",
             col.names = c("Mean","Std. Deviation"),
             digits = 3)
```

Table 3: posterior mean and sd for the two posteriors

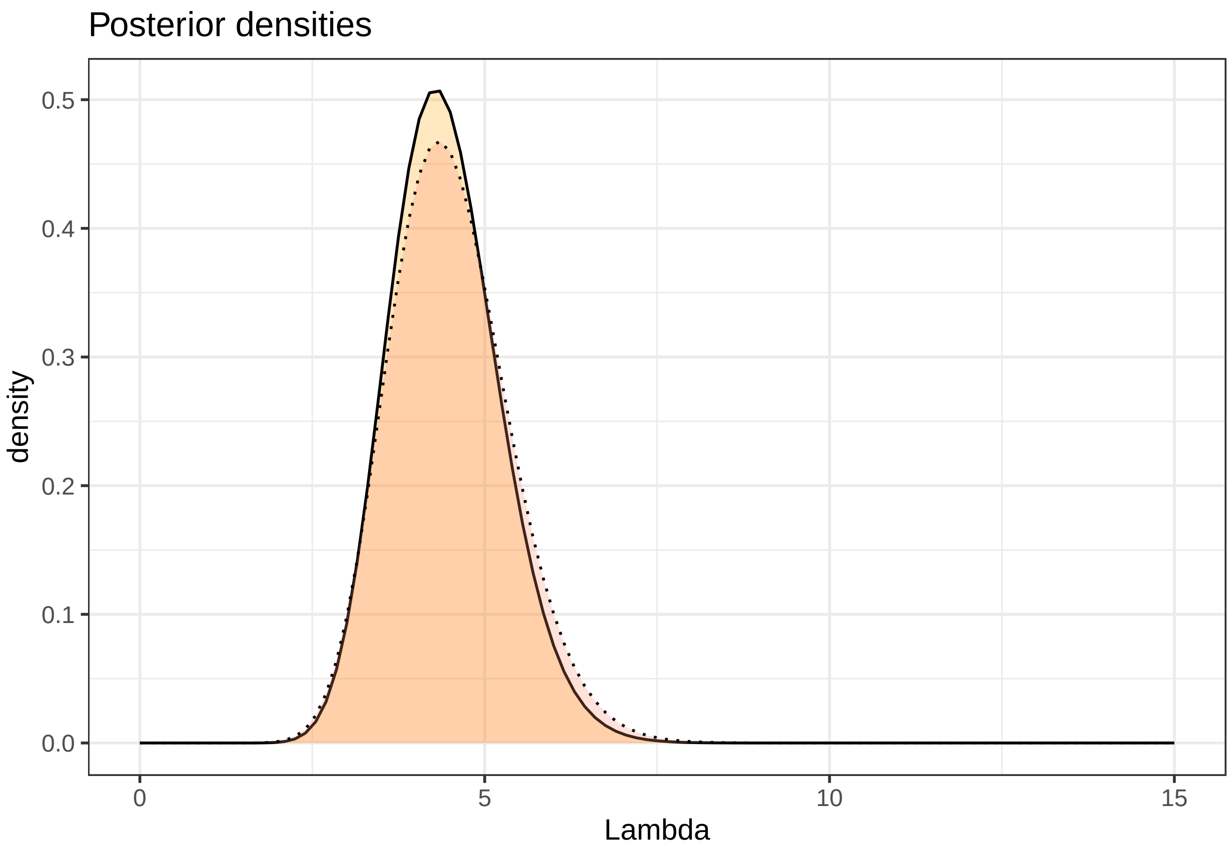
	Mean	Std. Deviation
posterior1	4.429	0.795
posterior2	4.500	0.866

10. Plot your two prior densities on one graph. Plot your two posterior densities in another graph. (Use the algebraic formula, or you can use the `dgamma` function in R). In one sentence for each plot, compare the densities (talk about location, scale, shape and compare the two densities).

```
x_lower_g <- 0
x_upper_g <- 15
#Plot prior
ggplot(data.frame(x = c(x_lower_g , x_upper_g)), aes(x = x)) +
  xlim(c(x_lower_g , x_upper_g)) +
  stat_function(
    fun = dgamma,
    args = list(shape = a1, rate = b1),
    geom = "area",
    fill = "orange",
    alpha = 0.25
  ) +
  stat_function(fun = dgamma, args = list (shape = a1, rate = b1))  +
  stat_function(
    fun = dgamma,
    args = list(shape = a2, rate = b2),
    geom = "area",
    fill = "salmon1",
    alpha = 0.25
  ) +
  stat_function(fun = dgamma, args = list (shape = a2, rate = b2),linetype = 3) +
  labs(title = "Prior Densities", x = "Lambda", y = "density") + theme_bw()
```




```
# Plot posterior
ggplot(data.frame(x = c(x_lower_g , x_upper_g)), aes(x = x)) +
  xlim(c(x_lower_g , x_upper_g)) +
  stat_function(
    fun = dgamma,
    args = list(shape = a_post1, rate = b_post1),
    geom = "area",
    fill = "orange",
    alpha = 0.25
  ) +
  stat_function(fun = dgamma, args = list (shape = a_post1, rate = b_post1))  +
  stat_function(
    fun = dgamma,
    args = list(shape = a_post2, rate = b_post2),
    geom = "area",
    fill = "salmon1",
    alpha = 0.25
  ) +
  stat_function(fun = dgamma, args = list(shape = a_post2, rate = b_post2), linetype = 3) +
  labs(title = "Posterior densities", x = "Lambda", y = "density") + theme_bw()
```



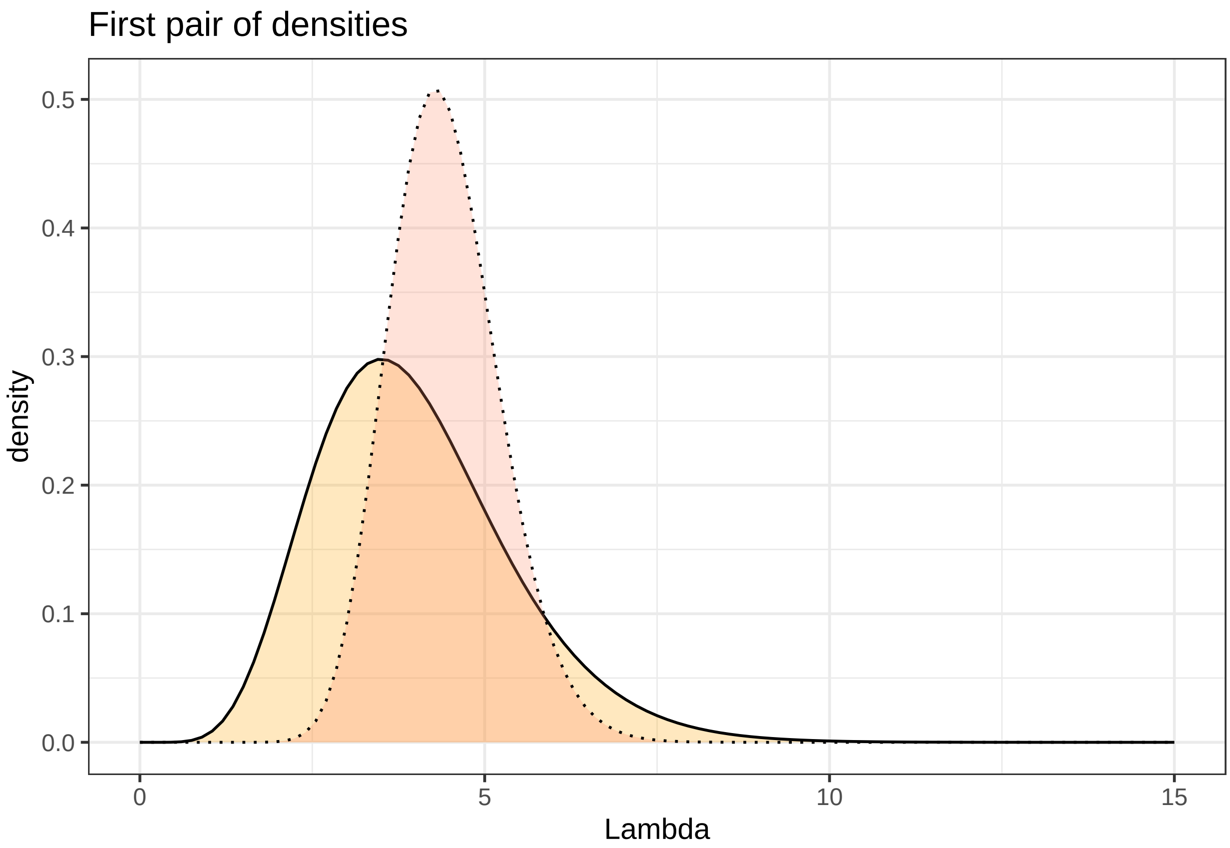
For prior, the first one has larger mean. For posterior, the two densities are very close.

11. Plot each prior density/posterior density pair on the same graph. For each plot, compare the two densities in one sentence.

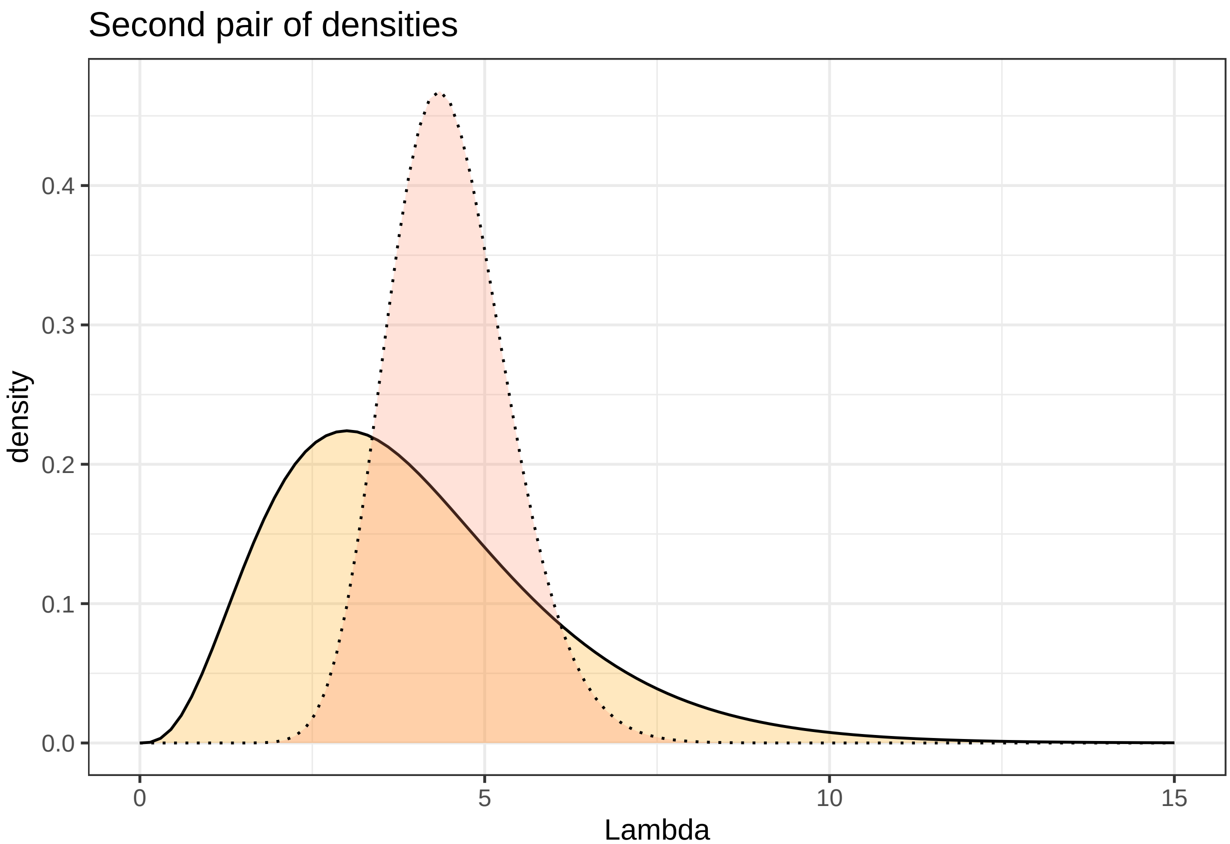
```
x_lower_g <- 0
x_upper_g <- 15

ggplot(data.frame(x = c(x_lower_g , x_upper_g)), aes(x = x)) +
  xlim(c(x_lower_g , x_upper_g)) +
  stat_function(
    fun = dgamma,
    args = list(shape = a1, rate = b1),
    geom = "area",
    fill = "orange",
    alpha = 0.25
  ) +
  stat_function(fun = dgamma, args = list (shape = a1, rate = b1))  +
  stat_function(
    fun = dgamma,
    args = list(shape = a_post1, rate = b_post1),
    geom = "area",
    fill = "salmon1",
    alpha = 0.25
  ) +
  stat_function(fun = dgamma, args = list (shape = a_post1, rate = b_post1),linetype = 3) +
```

```
  labs(title = "First pair of densities", x = "Lambda", y = "density") + theme_bw()
```




```
ggplot(data.frame(x = c(x_lower_g , x_upper_g)), aes(x = x)) +
  xlim(c(x_lower_g , x_upper_g)) +
  stat_function(
    fun = dgamma,
    args = list(shape = a2, rate = b2),
    geom = "area",
    fill = "orange",
    alpha = 0.25
  ) +
  stat_function(fun = dgamma, args = list (shape = a2, rate = b2))  +
  stat_function(
    fun = dgamma,
    args = list(shape = a_post2, rate = b_post2),
    geom = "area",
    fill = "salmon1",
    alpha = 0.25
  ) +
  stat_function(fun = dgamma, args = list(shape = a_post2, rate = b_post2), linetype = 3) +
  labs(title = "Second pair of densities", x = "Lambda", y = "density") + theme_bw()
```



Posterior has larger mean and smaller variance for both pairs.

12. Extra Credit.

(a) For this problem, treat the data as a single count y of customers that entered the store in 5 minutes. Define λ_1 as the 1 minute mean which you worked with previously. Define λ_5 as the 5 minute mean which you will work with now. Let a_5 and b_5 be the 5 minute prior parameters for λ_1 and similarly let a_1 and b_1 be 1 minute prior parameters from above.

(b) Give algebraic formulas for the relationships between (i) λ_5 and λ_1, (ii) the prior mean of λ_5 and λ_1, (iii) prior variances, (iv) prior standard deviations, (v) prior a-parameters, and (vi) b-parameters. (Hint: Transformation-of-variables.)

(c) λ_5 and λ_1

$$\lambda_5 = 5\lambda_1$$

(ii) the prior mean of λ_5 and λ_1

$$E(\lambda_5) = 5E(\lambda_1)$$

(iii) prior variances

$$Var(\lambda_5) = 25Var(\lambda_1)$$

(iv) prior standard deviations

$$SD(\lambda_5) = 5SD(\lambda_1)$$

(v) prior a-parameters, and (vi) b-parameters.

$$\lambda_1 \sim Gamma(a_1, b_1) \text{ and } \lambda_5 \sim Gamma(a_5, b_5)$$

$$\frac{a_5}{b_5} = 5 \cdot \frac{a_1}{b_1} \text{ and } \frac{a_5}{b_5^2} = 25 \cdot \frac{a_1}{b_1^2}$$

So $\frac{b_5}{b_1} = 5$, $b_1 = 5b_5$ and $a_1 = a_5$