BIOSTAT234 HW1

Name: Huiyu Hu

Problem 1: Normal data with a normal prior

1. Explain what your measurements will be.

- My resting heart rate measurements (per minutes) from iWatch

2. Before you collect the data, decide on your prior. Please use a normal density, and specify your prior mean μ_0 and standard deviation τ . So μ_0 is your best guess at the average of all your blood pressure (or other) measures, and τ is an estimate of the standard deviation that the true value may differ from your guess. (Note: you get better at this with practice, you won't be penalized for being too ridiculous in your guessing, within reason.) Explain your reasoning (1 or 2 sentences).

- I guess my heart rate most likely to be $\mu_0 = 75/min$ with standard deviation $\tau = 3$. Therefore the prior distribution will be $N(75, 3^2)$.

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

- The data (8 measurements) collected from my iWatch in last 10 minutes:
  `74, 68, 70, 76, 77, 74, 73, 68`

```
mydata <- c(74, 68, 70, 76, 77, 74, 73, 68)
mean <- mean(mydata)
var <- sum((mydata-mean)^2)/(length(mydata)-1)
mean
```

```
## [1] 72.5
```

```
var
```

```
## [1] 12
```

- Therefore, the sample mean and variance are 72.5 and 12 respectively.

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 σ that you use in all your calculations

- I use sample sd of my data set: $\sigma = sqrt(12) = 3.4641$

5. Calculate the posterior mean $\overline{\mu}$, 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, $\overline{y} \sim N(\mu, \sigma^2/n)$.

$$\overline{\mu} = \frac{\frac{n}{\sigma^2}}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}\bar{y} + \frac{\frac{1}{\tau^2}}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}\mu_0$$

$$\text{var} = \left(\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right)^{-1}$$

```
mean <- mean(mydata)
var <- sum((mydata-mean)^2)/(length(mydata)-1)
sd <- sqrt(var)
n <- length(mydata)
mu0 <- 75
tau <- 3

mu_bar <- (n/(var))/((n/var)+(1/tau^2)) * mean + (1/tau^2)/((n/var)+(1/tau^2)) * mu0
post.var <- ((n/(var))+(1/(tau^2)))^(-1)

mu_bar
```

```
## [1] 72.85714
```

```
post.var
```

```
## [1] 1.285714
```

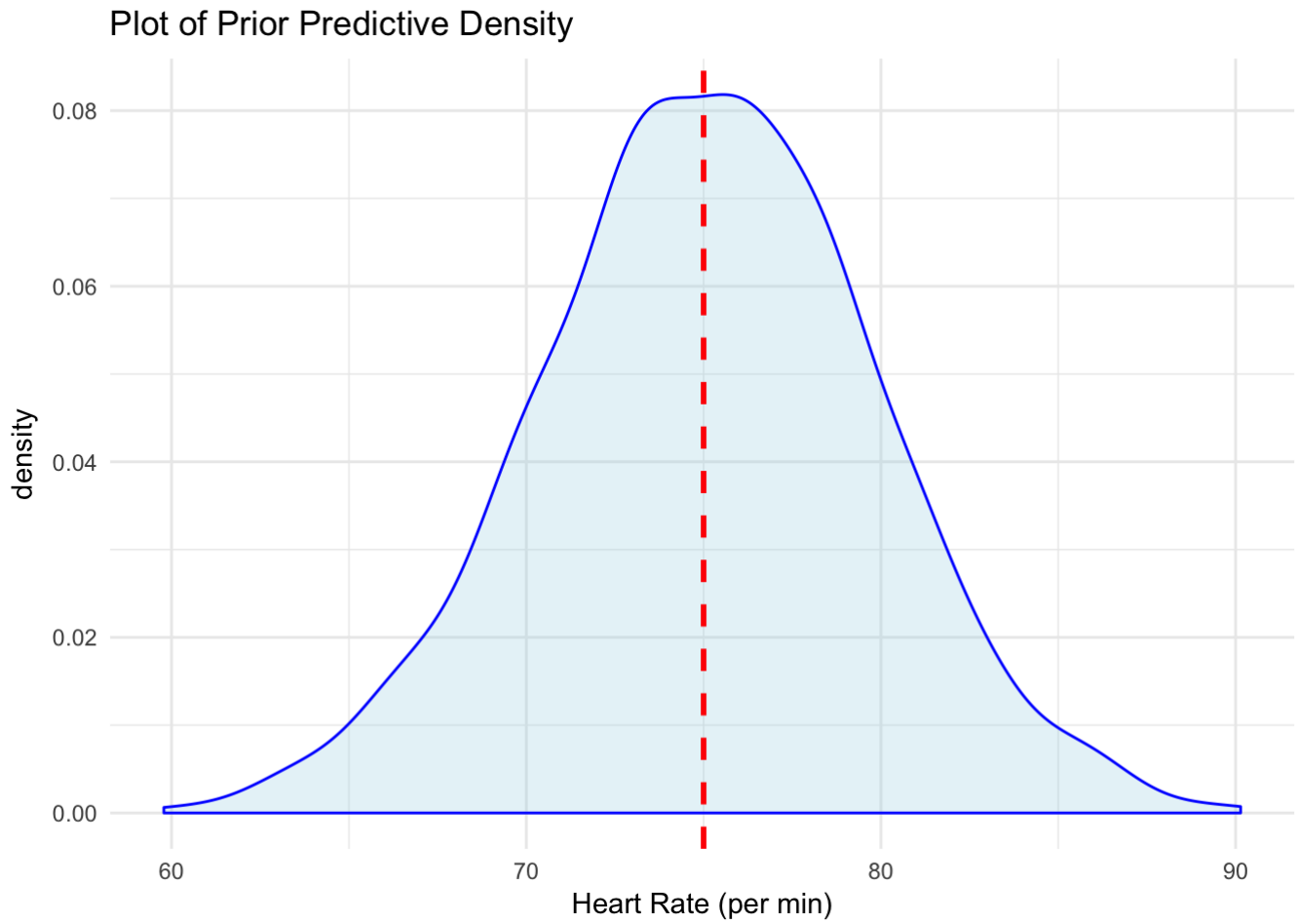
```
sqrt(post.var)
```

```
## [1] 1.133893
```

- Therefore, The posterior mean, posterior variance and posterior sd are 72.86, 1.29 and 1.13, respectively.

6. The t vsvt vi hrgxzi hi r wx) 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 $y \sim N(\mu_0, \sigma^2 + \tau^2)$

```
library(ggplot2)
set.seed(100)
# dataset:
sample <- data.frame(heart.rate=rnorm(1000, mu0, sqrt(sd^2+tau^2)))
ggplot(sample, aes(x=heart.rate)) +
  geom_density(colour = "blue", fill = "lightblue", alpha=.3) +
  geom_vline(aes(xintercept=mu0),
               linetype="dashed", size=1, colour="red") +
  labs(title="Plot of Prior Predictive Density",
        x ="Heart Rate (per min)") + theme_minimal()
```



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.

```
mean <- mean(mydata) # Likelihood mean
var <- sum((mydata-mean)^2)/(length(mydata)-1) # Likelihood var
sd <- sqrt(var) # likelihood sd

mu0 <- 75 # Prior mean
tau <- 3 # Prior sd
tau2 <- 9 # Prior var

mu_bar <- (n/(var))/((n/var)+(1/tau^2)) * mean +
  (1/tau^2)/((n/var)+(1/tau^2)) * mu0 # Posterior mean
post.var <- ((n/(var))+(1/(tau^2)))^(-1) # Posterior variance
post.sd <- sqrt(post.var) # Posterior sd

pp.mean <- mu0  # prior predictive mean
pp.sd <- sqrt(sd^2+tau^2) # prior predictive sd
pp.var <- sd^2+tau^2 # prior predictive var

table <- rbind(c(mu_bar, post.sd, post.var),c( mu0, tau, tau2),
               c(pp.mean, pp.sd, pp.var), c(mean, sd, var))
rownames(table) <- c("Posterior", "Prior", "Prior Predictive", "Likelihood")
colnames(table) <- c("Mean", "SD", "Variance")

knitr::kable(table)
```

	Mean	SD	Variance
Posterior	72.85714	1.133893	1.285714
Prior	75.00000	3.000000	9.000000
Prior Predictive	75.00000	4.582576	21.000000
Likelihood	72.50000	3.464102	12.000000

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 $\overline{y}$ and variance σ^2/n) all on the same graph. Interpret the plot.

```
library(dplyr)
```

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

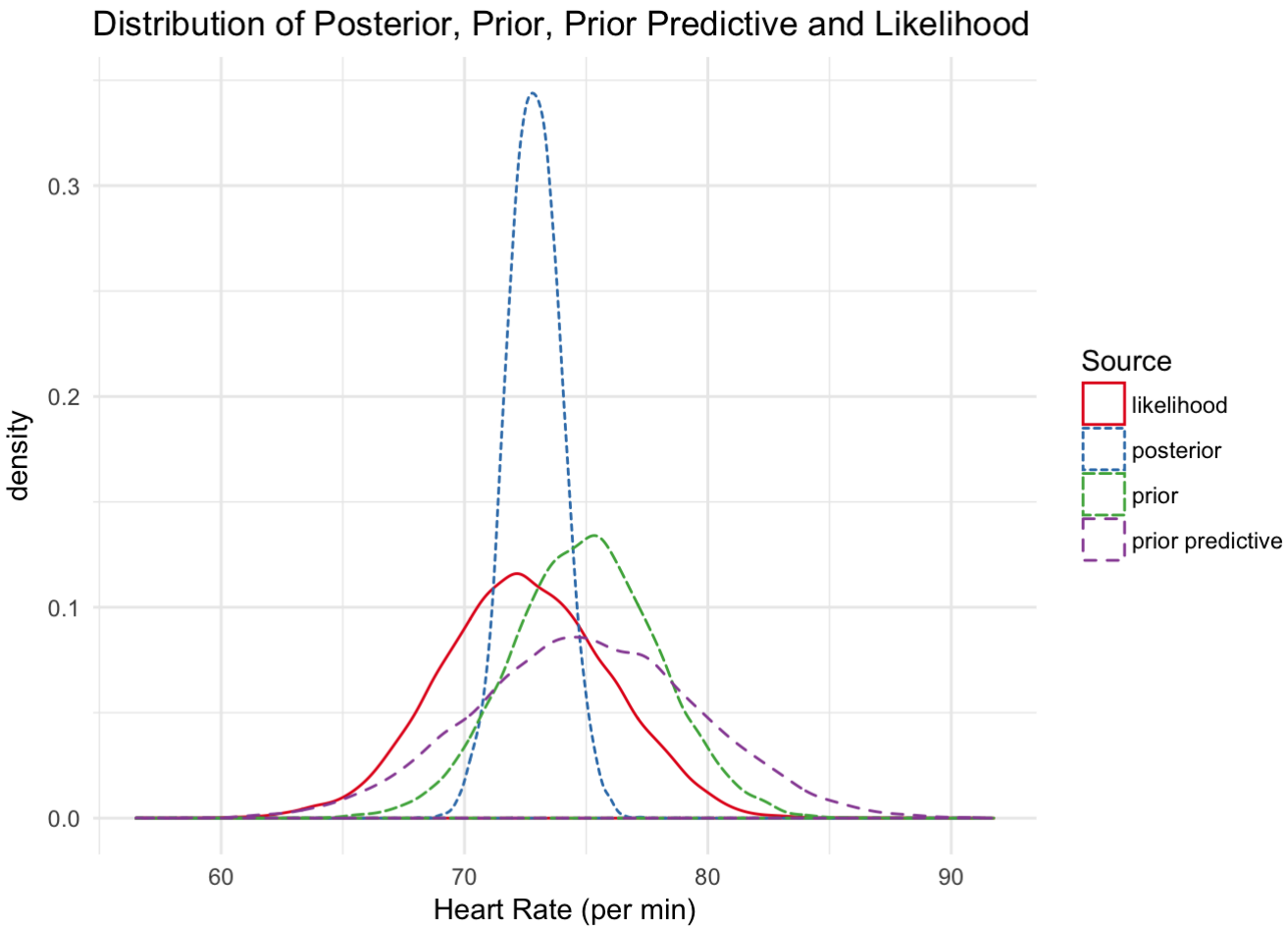
```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

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

```
library(tidyr)
posterior <- as.vector(rnorm(10000, mu_bar, post.sd))
prior <- as.vector(rnorm(10000, mu0, tau))
prior.predictive <- as.vector(rnorm(10000, pp.mean, pp.sd))
likelihood<- as.vector(rnorm(10000, mean, sd))

df.wide <- bind_cols('posterior' = posterior,'prior' = prior,'prior predictive' = prior.
predictive,'likelihood'=likelihood)
df.long <- gather(df.wide, key = "Source", value)

ggplot(df.long, aes(x=value, color=Source, linetype = Source)) +
  geom_density(alpha=0.4) +
  labs(title="Distribution of Posterior, Prior, Prior Predictive and Likelihood",
       x ="Heart Rate (per min)") +
  scale_color_brewer(palette = "Set1") + theme_minimal()
```



Problem 2: Count Data with a Gamma Prior

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

i. Prior: $Gamma(a,b) $, support is $\lambda \in (0, \infty)$
ii. Posterior: $\lambda|data \sim Gamma(a + \sum y_i, b + n)$, support is $\lambda \in (0, \infty)$
iii. sampling density: $y|\lambda \sim Poisson(\lambda)$, support is $\lambda \in N$
iv. likelihood: $y|\lambda \sim Poisson(n\lambda)$ $\lambda \in N$

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.

- I think **f** acts lke 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 y1, . . . , y5 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.

- I went to boba time westwood on January 22, 4:40 - 4：45

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 a1 and b1; the other set of prior parameters are a2 and b2.

a. Before you visit the store, make a guess as to the mean number of customers entering the store in one minute. Call this m0. This is the mean of your prior distribution for $$.

- I guess $ m_0 = 5$

b. Make a guess s0 of the prior sd associated with your estimate m0. This s0 is the standard deviation of the prior distribution for $$.Note: most people underestimate s0.

- I guess $ s_0 = 3$

c. Separately from the previous question 5b, estimate how many data points n0 your prior guess is worth. That is, n0 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 m0.

- I guess $n_0 = 1$

d. Solve for a1 and b1 based on m0 and s0.

$$E[\lambda] = \alpha/\beta = 5; Var[\lambda] = \alpha/\beta^2 = 3$$

- By solving the equations above, $\alpha(a1) = 25/3, \beta(b1) = 5/3$

e. Separately solve for a2 and b2 using m0 and n0 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 a2; b2 to use with the remainder of this problem!)

- Set $n_o = \beta = 1$

$$E[\lambda] = \alpha/\beta = 5$$

- Solve for $\alpha = 5$

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

- Method1: Take average of the parameters in the priors to form a single prior.
- Method2: Set the prior with larger variance to allow more change in the prior.

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

```
time <- c('Minute 1','Minute 2','Minute 3','Minute 4','Minute 5')
count <- c(5, 3, 1, 7, 2)
table1 <- cbind(time, count)
knitr::kable(table1)
```

time	count
Minute 1	5
Minute 2	3
Minute 3	1
Minute 4	7
Minute 5	2

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

- **Tswi vsv' 5**:

$$\alpha_{posterior} = \alpha_{prior} + \sum_{i=1}^{n} y_i = \frac{25}{3} + 5 + 3 + 1 + 7 + 2 = \frac{79}{3}$$

$$\beta_{posterior} = \beta_{prior} + n = \frac{5}{3} + 5 = \frac{20}{3}$$

- The posterior for prior #1 is of the form $Gamma(\alpha = \frac{79}{3}, \beta = \frac{20}{3})$

- **Tswi vsv' 6**:

$$\alpha_{posterior} = \alpha_{prior} + \sum_{i=1}^{n} y_i = 5 + 5 + 3 + 1 + 7 + 2 = 23$$

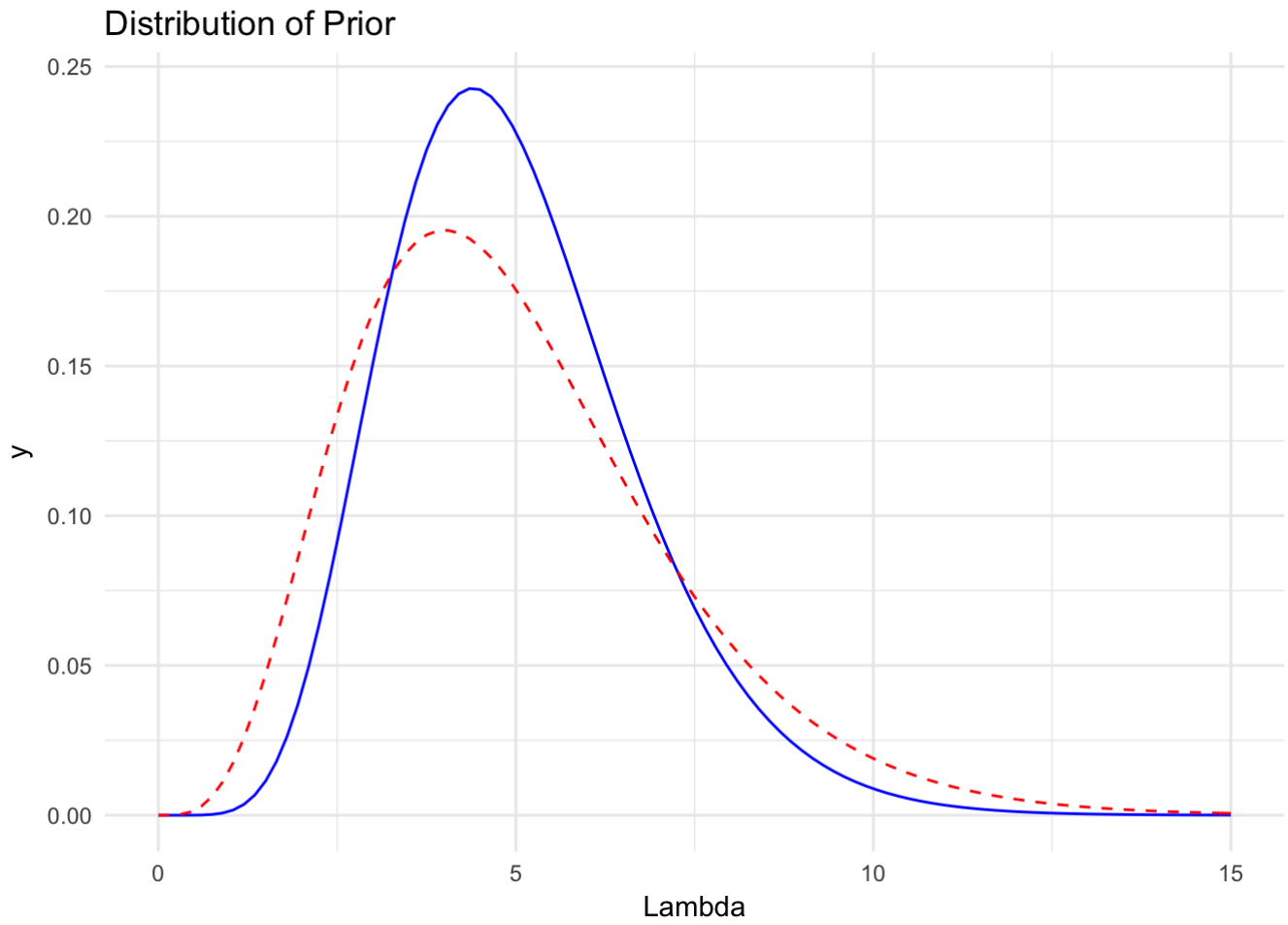
$$\beta_{posterior} = \beta_{prior} + n = 1 + 5 = 6$$

- The posterior for prior #1 is of the form $Gamma(\alpha = 23, \beta = 6)$

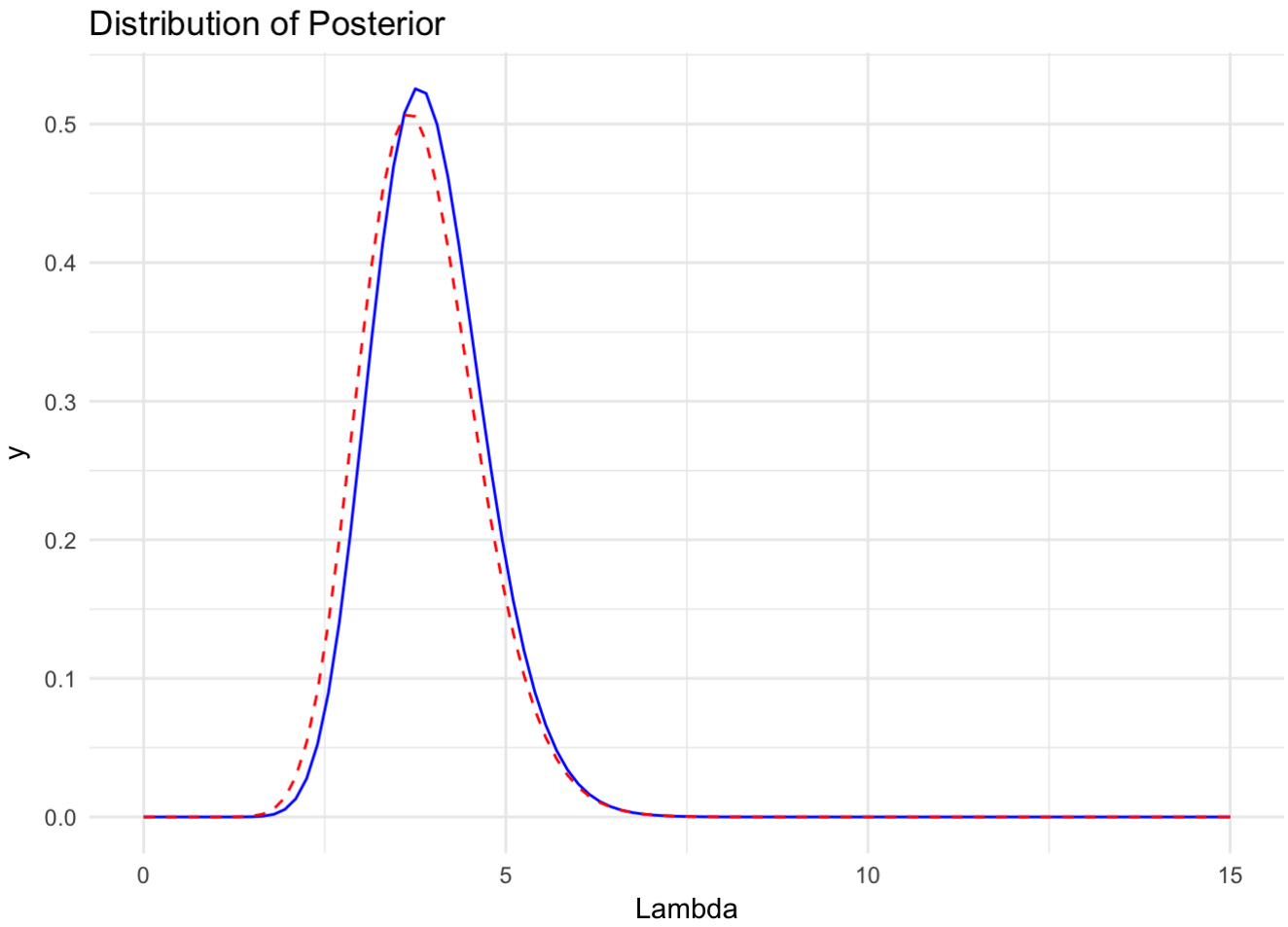
10. Plot your two prior densities on one graph. Plot your two posterior densities in another graph. (Use the algebraic formula, or you can use the dgamma function in R). In one sentence for each plot, compare the densities (talk about location, scale, shape and compare the two densities).

```
#1
a1 <- 25/3
b1 <- 5/3
a1.post <- 79/3
b1.post <- 20/3
#2
a2 <- 5
b2 <- 1
a2.post <- 23
b2.post <- 6
```

```
# Graph of prior
ggplot(data.frame(x=c(0, 15)), aes(x)) +
  stat_function(fun=dgamma, args=list(shape=a1, rate = b1), colour = "blue") +
  stat_function(fun=dgamma, args=list(shape=a2, rate = b2), colour = "red", linetype = 2
)  +
  labs(title="Distribution of Prior", x ="Lambda") + theme_minimal()
```



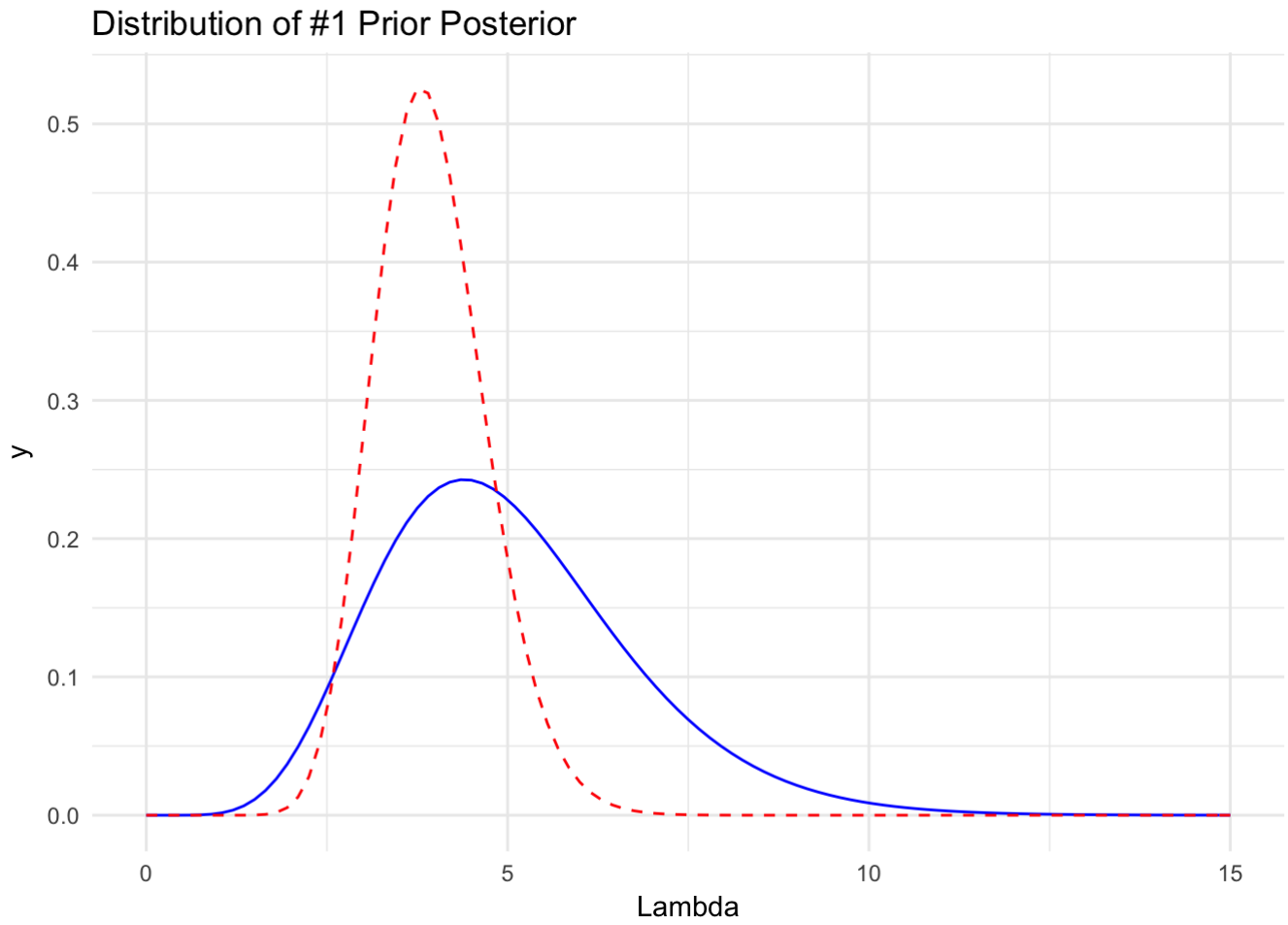

```
# Graph of posterior
ggplot(data.frame(x=c(0, 15)), aes(x)) +
  stat_function(fun=dgamma, args=list(shape=a1.post, rate=b1.post), colour = "blue") +
  stat_function(fun=dgamma, args=list(shape=a2.post, rate=b2.post), colour = "red", line
type = 2) +
  labs(title="Distribution of Posterior", x ="Lambda") + theme_minimal()
```



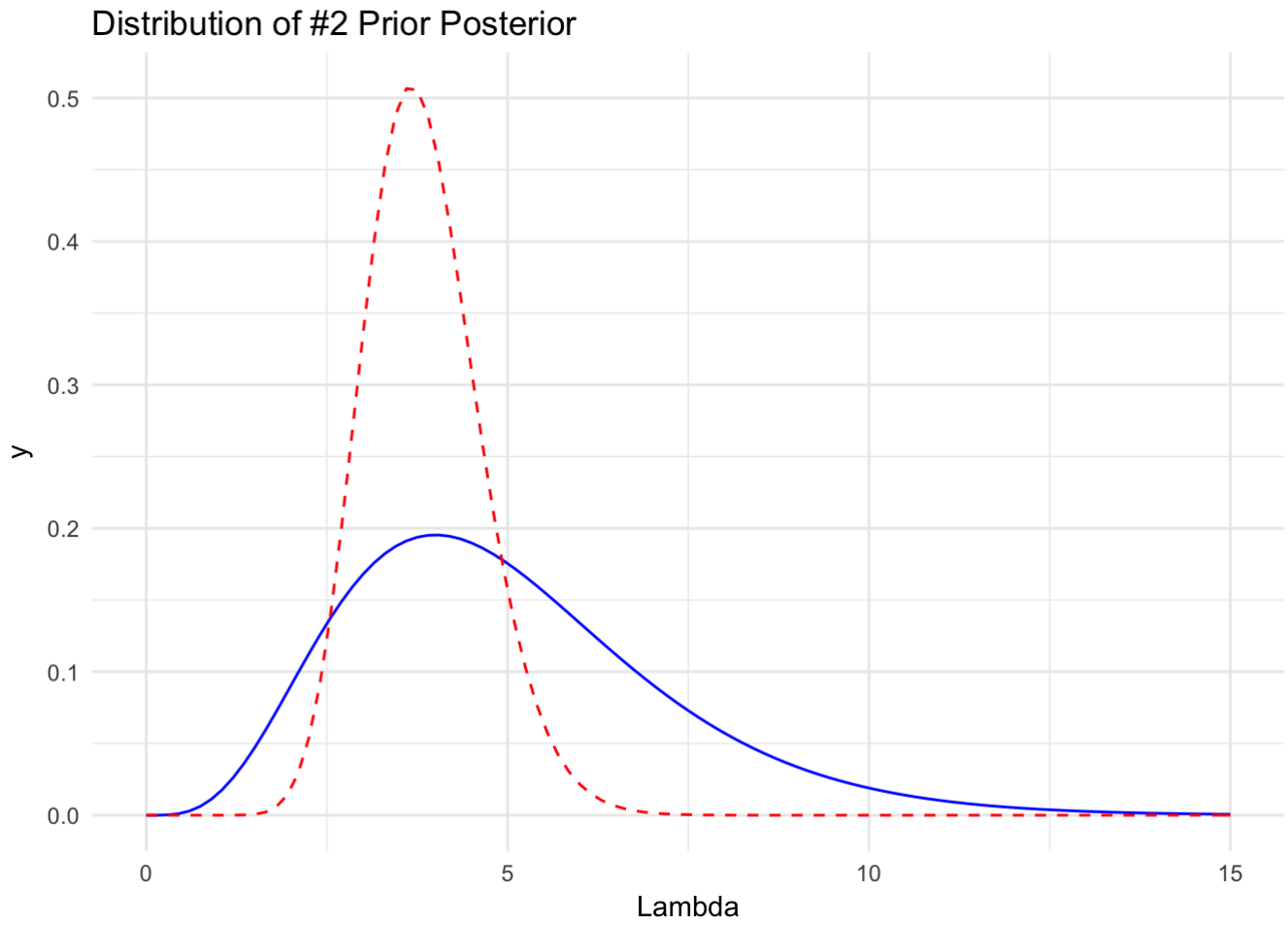
- Two priors have similar mean but #1 prior is sightly larger than #2 (dashed). Two posteriors have very similar distribution.

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

```
ggplot(data.frame(x=c(0, 15)), aes(x)) +
  stat_function(fun=dgamma, args=list(shape=a1, rate=b1), colour = "blue") +
  stat_function(fun=dgamma, args=list(shape=a1.post, rate=b1.post), colour = "red", line
type = 2)  +
  labs(title="Distribution of #1 Prior Posterior", x ="Lambda") + theme_minimal()
```




```
ggplot(data.frame(x=c(0, 15)), aes(x)) +
  stat_function(fun=dgamma, args=list(shape=a2, rate=b2), colour = "blue") +
  stat_function(fun=dgamma, args=list(shape=a2.post, rate=b2.post), colour = "red", line
type = 2)  +
  labs(title="Distribution of #2 Prior Posterior", x ="Lambda") + theme_minimal()
```



- Both graph showed that posteriors (dashed) have smaller mean and smaller variance. I might overestimated the count.

Extra Credit

(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.)

$$\lambda_5 = 5\lambda_1$$

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

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

$$SE(\lambda_5) = SE(\lambda_1)$$

$$\lambda_1 \sim Gamma(a_1, b_1)$$

$$\lambda_5 \sim Gamma(a_5, b_5)$$

Therefore, $a_1 = a_5, b_5 = b_1$