Homework 3

PSTAT 115, Spring 2023

Due on May 26, 2022 at 11:59 pm

Problem 1. Posterior Credible Intervals

One way for us to learn more about posterior distributions is to find some credible intervals. Suppose we have a posterior density of a parameter λ, defined by $\lambda|y \sim \text{Gamma}(4,1)$.

a. (7pts) Plot this distribution. Construct the posterior middle 95% credible interval for λ. The middle 95% interval is such that 2.5% of the probability mass is left to the left and 2.5% to the right. Save the lower and upper endpoints in a vector of length 2, called `middle_95`. Using either line segments or shading, display this interval on the plot.

```
set.seed(1)
shape = 4
rate = 1
lambda <- rgamma(10000,shape,rate)

low_q <- qgamma(0.025, shape, rate)
upper_q <- qgamma(0.975, shape, rate)
middle_95 = c(low_q, upper_q)

ggplot(data = data.frame(x = lambda), aes(x = x)) +
    geom_density(col = "black") +
    geom_ribbon(data = data.frame(x = lambda[lambda >= low_q & lambda <= upper_q]), aes(x= x, ymin = 0,
            fill = "lightgray", alpha = 0.5) +
    geom_vline(xintercept = low_q, color = "red", linetype = "dashed")+
    geom_vline(xintercept = upper_q, color = "red", linetype = "dashed")
```
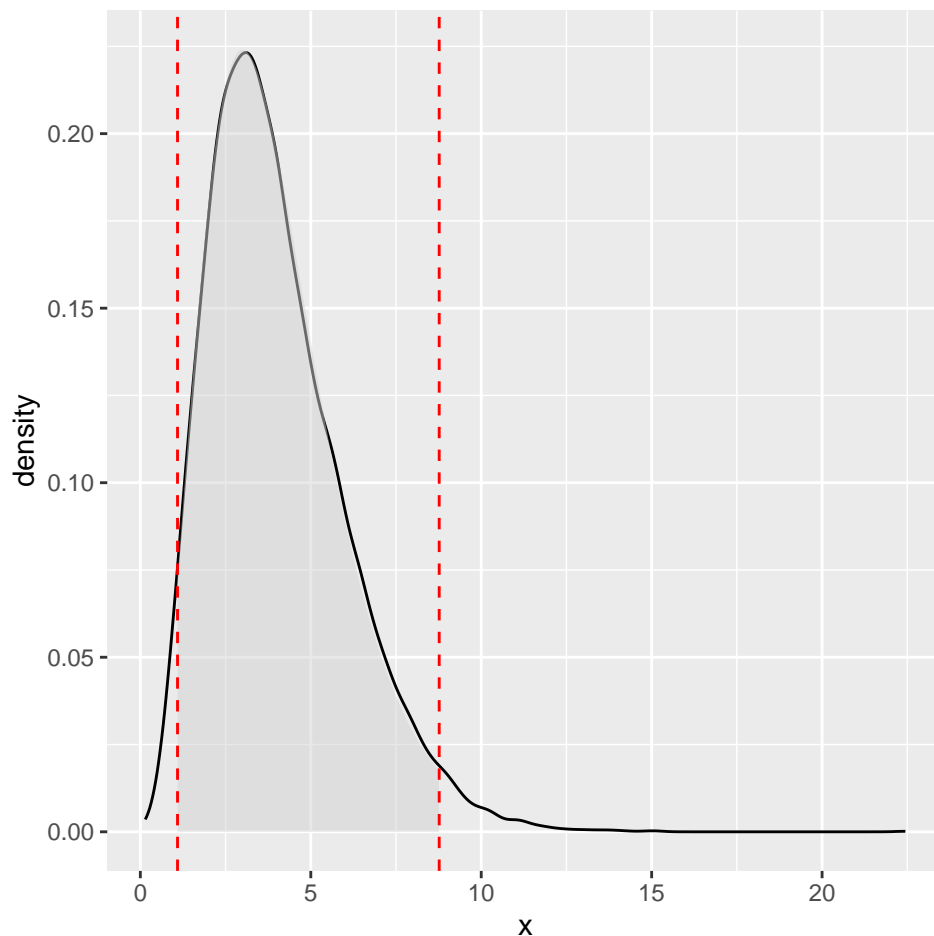
b. (5pts) Interpret the interval you found. How is it different from a frequentist confidence interval?

The credible interval suggest that there's a 95% probability that lambda lies within the interval (1.09, 8.77). It's an interval estimate that, given the data and the prior, contains the parameter with a certain probability. The Frequentist CI doesn't account for effect of the observed value, it produces intervals that contain the true parameter value approximately 95% of the time (for a 95% confidence interval).

c. (8pts) Besides the middle 95% credible interval, we could also find the 95% highest posterior density (HPD) region. This region contains the 95% of posterior values with the highest posterior densities. The HPD region will always be the shortest credible interval for a given probability, since it by definition contains the values of λ with the highest probability of occurring. Use `HDInterval::hdi()` to construct the HPD region. Save the lower and upper endpoints of this region in a variable called `hdi_region`. Add this interval to the plot you made in part (a), making sure that both intervals are distinguishable on the plot.

```
library(HDInterval)
hdi_region = hdi(lambda, credMass = 0.95)
ggplot(data = data.frame(x = lambda), aes(x = x)) +
    geom_density(col = "black") +
    geom_ribbon(data = data.frame(x = lambda[lambda >= low_q & lambda <= upper_q]), aes(x= x, ymin = 0,
            fill = "quantile"), alpha = 0.3) +
    geom_ribbon(data = data.frame(x = lambda[lambda >= hdi_region[1] & lambda <= hdi_region[2]]), aes(x=
            fill = "hpd"), alpha = 0.3) +
    geom_vline(xintercept = low_q, color = "red", linetype = "dashed")+
```

2

```
    geom_vline(xintercept = upper_q, color = "red", linetype = "dashed")+
    geom_vline(xintercept = hdi_region[1], color = "blue", linetype = "dashed")+
    geom_vline(xintercept = hdi_region[2], color = "blue", linetype = "dashed")+
    scale_colour_manual("",values=c("hpd"="blue", "quantile"="red"))+
    scale_fill_manual("",values=c("hpd"="blue", "quantile"="red"))
```
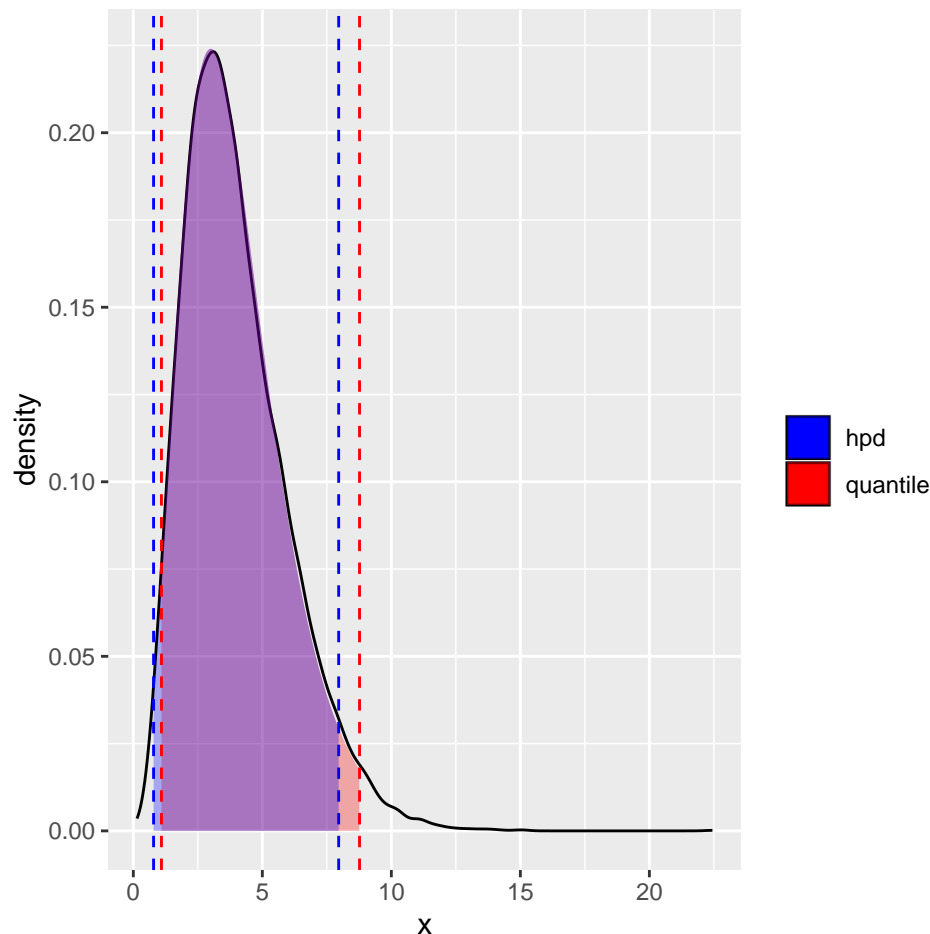

d. (5pts) Based on your plot, how do the two kinds of 95% credible intervals differ? How long is the middle interval? The HDI interval?

A quantile-based CI is calculated by finding the values below which a certain cumulative percentage of the data fall. quantile-based CI assumes that the probability mass is evenly distributed around the median, which is not the case in non-symmetrical distributions like gamma. HPD, defined as the shortest possible intervals containing a given portion of the probability density. HPD-based CI always includes the mode of the distribution and considers the shape of the distribution, making it more suitable for gamma

```
diff(middle_95)
```

```
## [1] 7.677408
```

```
diff(hdi_region)
```

```
##     upper
## 7.173509
```

3

Problem 2. Posterior predictive Beta-Binomial

As we saw in the lectures, it is sometimes possible to derive an exact posterior predictive model. In this exercise we will discuss one further example. To begin, suppose we observe $Y = y$ successes in n trials where $Y \mid \pi \sim Bin(n, \pi)$ and $\pi \sim Beta(\alpha, \beta)$.

a. (5pts) Identify the posterior probability density function (pdf) of π given the observed data $Y = y$, denoted by $f(\pi \mid y)$. This will depend on $(y, n, \alpha, \beta, \pi)$.

b. (5pts) Suppose we conduct n' *new* trials, wheren n' might differ from our original number of trials n) and let $Y' = y'$ be the observed number of successes in these new trials. Identify the conditional probability mass function (pmf) of Y' given π, denoted by $f(y' \mid \pi)$. This will depend on (y', n', π).

c. (10pts) Idenfity the posterior predictive pmf of Y', denoted by $f(y' \mid y)$. This will depend on $(y, n, y', n', \alpha, \beta)$.

d. (7pts) Suppose now that your posterior model for π is based on a prior model with $\alpha = 4$ and $\beta = 6$ and an observed $y = 14$ successes in $n = 100$ original trials. We plan to conduct $n' = 20$ *new* trials. Specify the posterior predictive pmf of Y', the number of successes we might observe in these 20 trials.

e. (8pts) Continuing part d., suppose instead we plan to conduct $n' = 4$ *new* trials. Specify and sketch the posterior predictive pmf of Y', the number of successes we might observe in these 4 trials.

(a)

$$p(\pi \mid Y) = \frac{p(\pi) L(\pi \mid Y = y)}{\int p(\pi) L(\pi \mid Y = y)}$$

$$p(\pi) L(\pi \mid Y = y) = \frac{\Gamma(a+\beta)}{\Gamma(a)\Gamma(\beta)} \pi^{a-1} (1-\pi)^{\beta-1} \binom{n}{y} \pi^{y} (1-\pi)^{n-y}$$

$$= \frac{\Gamma(a+\beta)}{\Gamma(a)\Gamma(\beta)} \binom{n}{y} \pi^{a+y-1} (1-\pi)^{\beta+n-y-1}$$

$$\int \beta(a,\beta) \pi^{a-1} (1-\pi)^{\beta-1} = 1 \Rightarrow \frac{1}{\beta(a,\beta)} = \int x^{a-1} (1-\pi)^{\beta-1}$$

$$\int p(\pi) L(\pi \mid Y = y) = \frac{\Gamma(a+\beta)}{\Gamma(a)\Gamma(\beta)} \binom{n}{y} \frac{\Gamma(a+y)\Gamma(\beta+n-y)}{\Gamma(a+\beta+n)}$$

$$f(\pi \mid Y) = \frac{\Gamma(a+\beta+n)}{\Gamma(a+y)\Gamma(\beta+n-y)} \pi^{a+y-1} (1-\pi)^{\beta+n-y-1}$$

(b)

$$f(y' | \pi) = \binom{n'}{y'} \pi^{y'} (1-\pi)^{n'-y'}$$

(c)

$$f(y' | y) = \int f(y' | \pi) f(\pi | y) \, d\pi$$

$$= \int \binom{n'}{y'} \pi^{y'} (1-\pi)^{n'-y'} \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+y)\Gamma(\beta+n-y)} \pi^{\alpha+y-1} (1-\pi)^{\beta+n-y-1} \, d\pi$$

$$\binom{n'}{y'} \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+y)(\beta+n-y)} \int \pi^{\alpha+y+y'-1} (1-\pi)^{\beta+n-y+n'-y'-1} \, d\pi$$

$$= \binom{n'}{y'} \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+y)\Gamma(\beta+n-y)} \times \frac{\Gamma(\alpha+y+y')\,\Gamma(\beta+n-y+n'-y')}{\Gamma(\alpha+\beta+n+n')}$$

d

$$f(y'|y; a=4, \beta=6, \gamma=14, n=100, n'=20)$$

$$= \binom{20}{y'} \frac{\Gamma(110)}{\Gamma(18)\Gamma(92)} \frac{\Gamma(18+y')\Gamma(112-y'))}{\Gamma(130)}$$

$$E(Y'|y=14) = \sum_{y'=0}^{20} y' f(y'|y=14)$$

$$= 0 \cdot f(y'=0|y=14) + 1 \cdot f(y'=1|y=14) + \cdots$$

$$= 3.27$$

e. $\binom{4}{y'} \frac{\Gamma(110)}{\Gamma(18)\Gamma(92)} \frac{\Gamma(18+y')\Gamma(96-y')}{\Gamma(114)}$

```r
pmf <- function(y) {
    p = choose(4,y) * gamma(110)/(gamma(18)*gamma(92)) * gamma(18+y)*gamma(96 - y)/gamma(114)
    return(p)
}

y=0:4
prob = pmf(y)

ggplot(data=data.frame(x = y, y = prob), aes(x=x, y=y)) +
    geom_point() +
    geom_line()
```
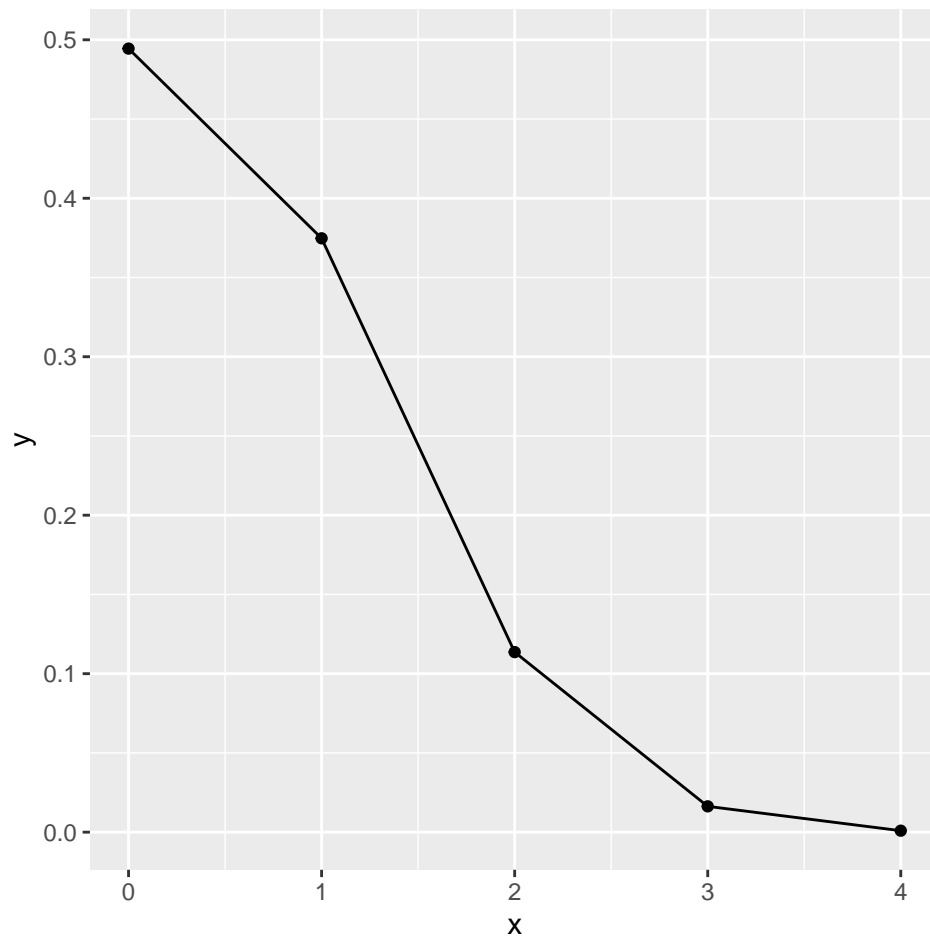


```r
paste("Expected Number of Success:", sum(y*prob))
```

```
## [1] "Expected Number of Success: 0.654545454545425"
```

Problem 3. Australian temperature

Let μ be the average 3 p.m. temperature in Perth, Australia. Not knowing much about Australian weather, your friend's prior understanding is that the average temperature is likely around 30 degrees Celsius, though

might be anywhere between 10 and 50 degrees Celsius. To learn about μ, they plan to analyze 1000 days of temperature data. Letting Y_i denote the 3 p.m. temperature on day i, they'll assume that daily temperatures vary Normally around μ with a standard deviation of 5 degrees:

$$Y_i \sim N(\mu, 5^2).$$

a. (5pts) Tune and plot a Normal prior for μ that reflects your friend's understanding.

```
mu = rnorm(10000, 30, sqrt(20))
ggplot(data.frame(x=mu), aes(x)) +
    geom_histogram(aes(y = ..density..), binwidth = 0.5, color = "black", fill = "lightblue") +
    geom_density(col="blue")
```

```
## Warning: The dot-dot notation ('..density..') was deprecated in ggplot2 3.4.0.
## i Please use 'after_stat(density)' instead.
## This warning is displayed once every 8 hours.
## Call 'lifecycle::last_lifecycle_warnings()' to see where this warning was
## generated.
```
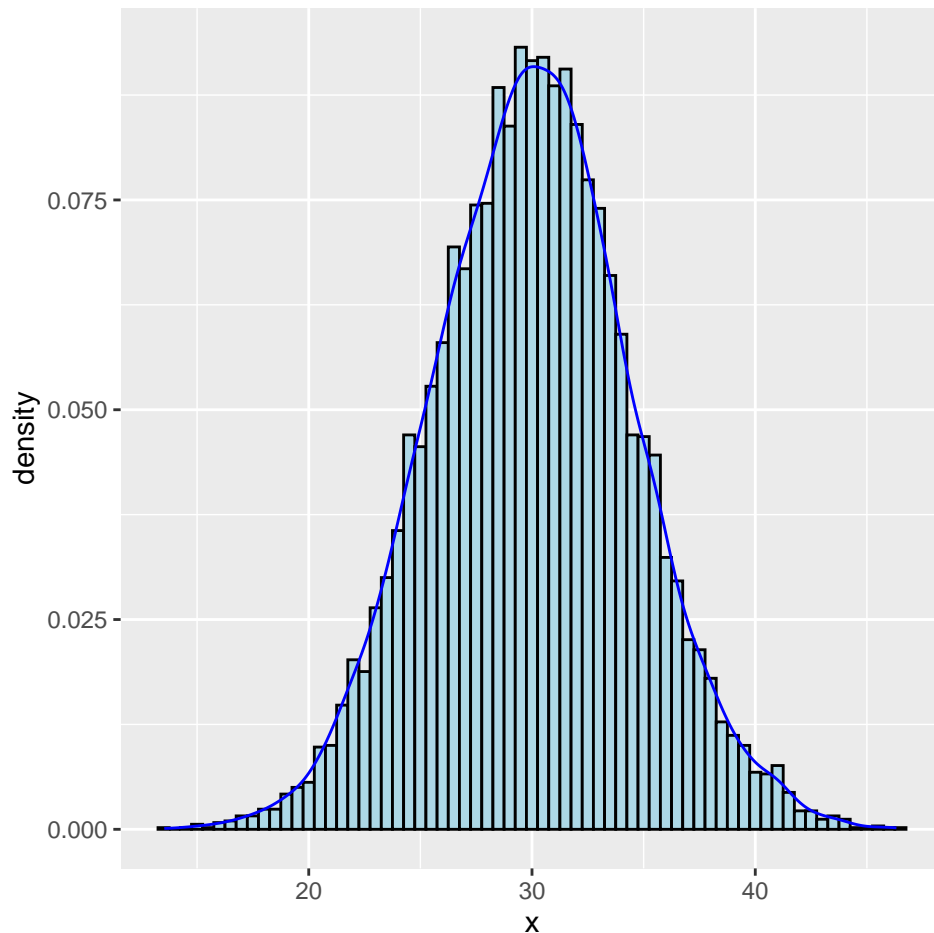

b. (7pts) The `weather_perth` data set in the folder `homework4/data/` includes 1000 daily observations of 3 p.m. temperatures in Perth (`temp3pm` column). Plot this data and discuss whether it's reasonable to assume a Normal model for the temperature data.

```
load("data/weather_perth.RData")
ggplot(data = weather_perth, aes(x = temp3pm)) +
    geom_histogram()
```
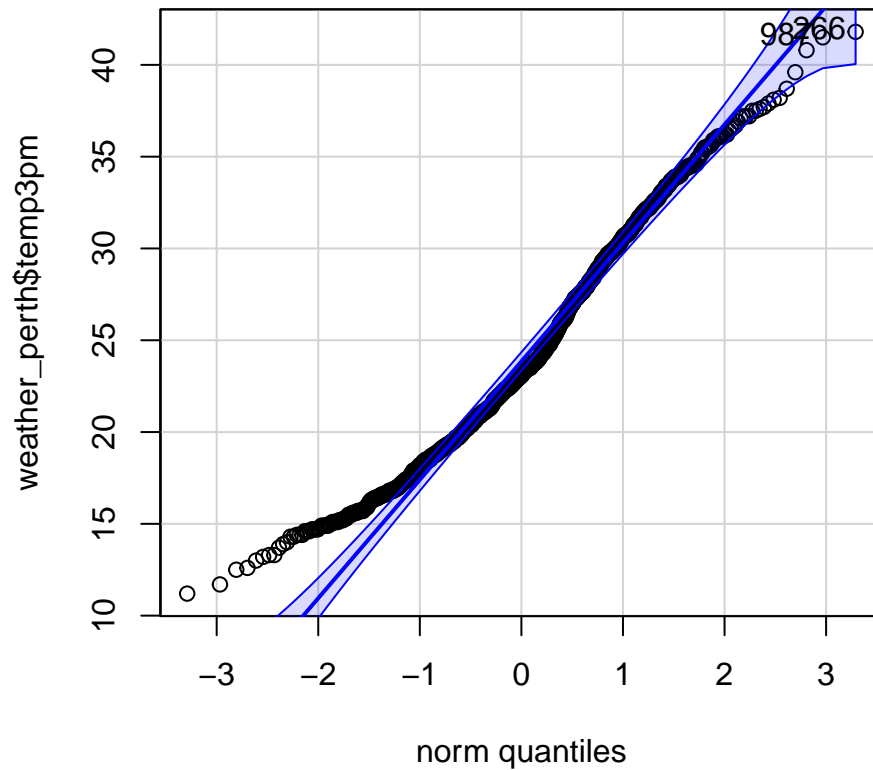
'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
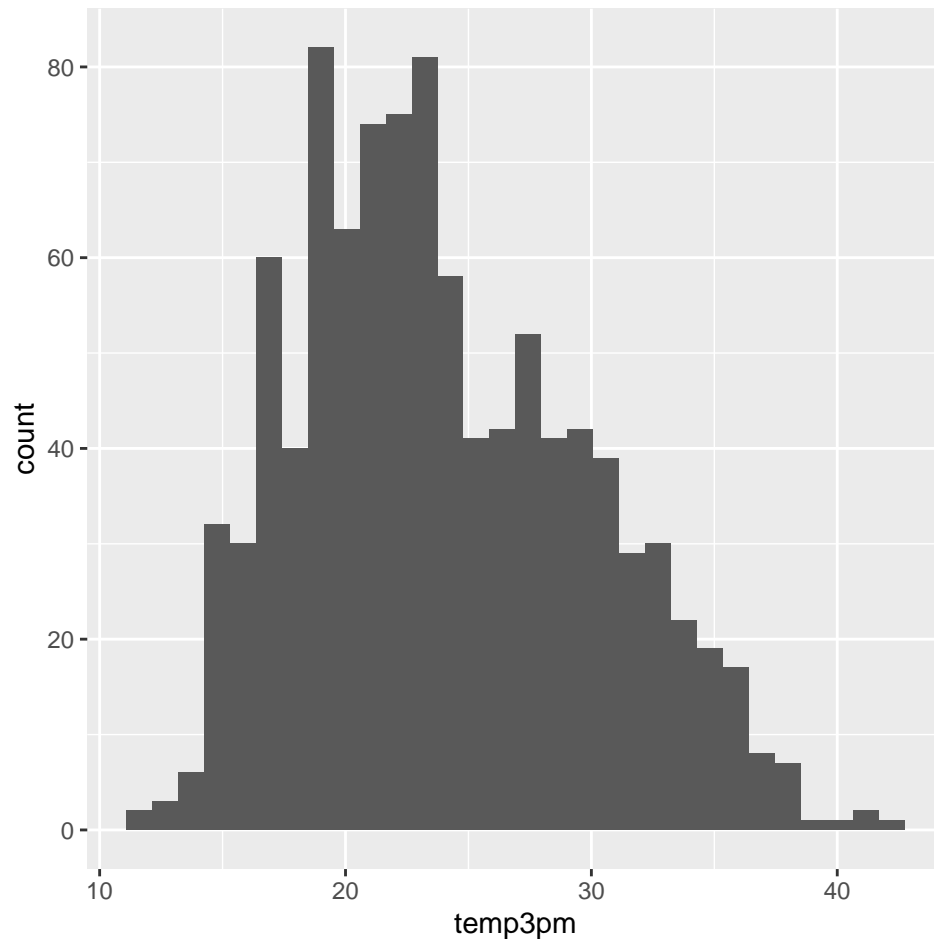

```
library(car)
```

```
## Loading required package: carData
##
## Attaching package: 'car'
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
## The following object is masked from 'package:purrr':
##
##     some
```

```
qqPlot(weather_perth$temp3pm)
```



```
## [1] 266 987
```

```
shapiro.test(weather_perth$temp3pm)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  weather_perth$temp3pm
## W = 0.97548, p-value = 5.949e-12
```

```
ks.test(weather_perth$temp3pm, "pnorm")
```

```
## Warning in ks.test.default(weather_perth$temp3pm, "pnorm"): ties should not be
## present for the Kolmogorov-Smirnov test
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
```

```
## data:  weather_perth$temp3pm
## D = 1, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
# The qqplot and Normality test suggest that the distribution of 3pm temperature is not likely to be no
```

c. (8pts) Identify the posterior model of μ and verify your answer using the function `summarize_normal_normal()` saved in the folder `homework4/functions`.

$$\mu|y_1,\ldots,y_n \sim N(\theta\frac{\sigma^2}{n\tau^2+\sigma^2}+\bar{y}\frac{n\tau^2}{n\tau^2+\sigma^2},\frac{\tau^2}{n\tau^2+\sigma^2})$$

$$\mu|y_1,\ldots,y_n \sim N(30\cdot\frac{25}{1000\cdot20+25}+\bar{y}\frac{1000\cdot20}{1000\cdot20+25},\frac{20*25}{1000\cdot20+25})$$

$$\mu|y_1,\ldots,y_n \sim N(24,0.025)$$

```
y_bar = mean(weather_perth$temp3pm)
post_mu = (30*25 + y_bar*1000*20)/(1000*20+25)
post_var = (20*25)/(1000*sqrt(20)^2+25)
source("functions/summarize_normal_normal.R")
summarize_normal_normal(30, sqrt(20), 5, y_bar, 1000)
```

```
## [1] "25 20 1000"
```

```
##       model     mean     mode         var        sd
## 1     prior 30.00000 30.00000 20.00000000 4.4721360
## 2 posterior 24.00839 24.00839  0.02496879 0.1580152
```

d. (10pts) Plot the prior pdf, likelihood function, and posterior pdf of μ. Describe the evolution in your understanding of μ from the prior to the posterior.
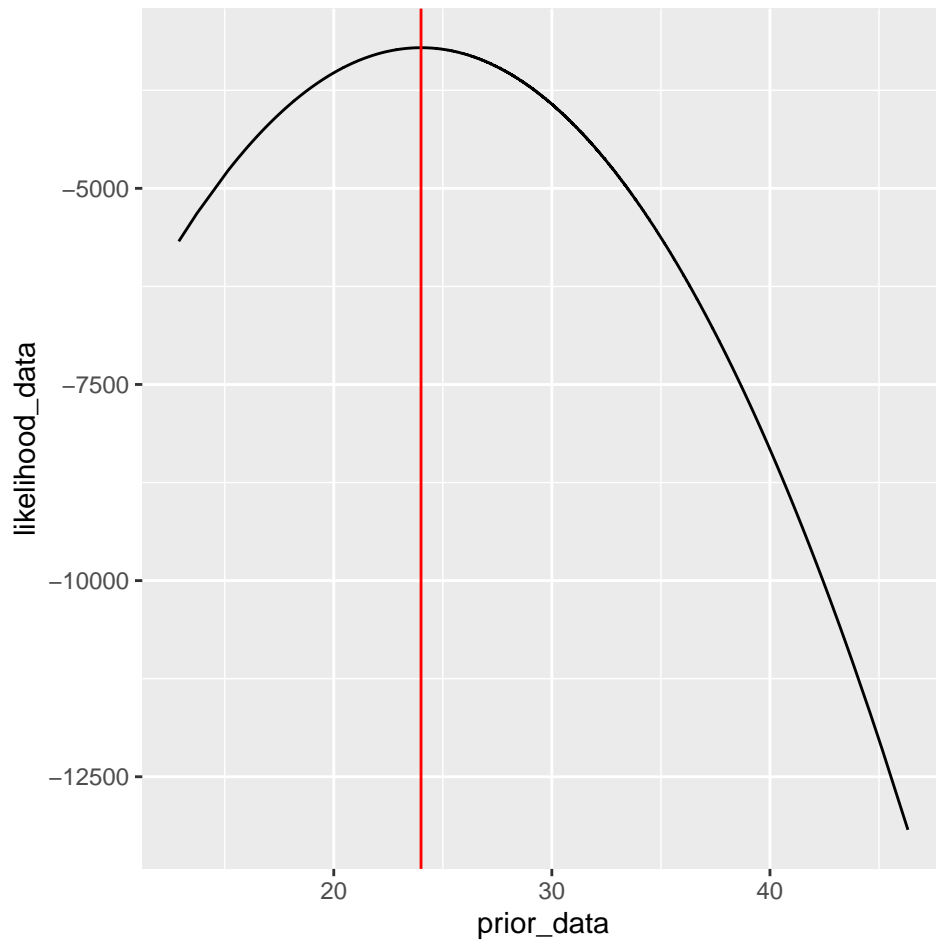
```
sigma = 5

prior_data = rnorm(10000, 30, sqrt(20))

ggplot(data=data.frame(x=prior_data), aes(x=x)) +
    geom_density()
```
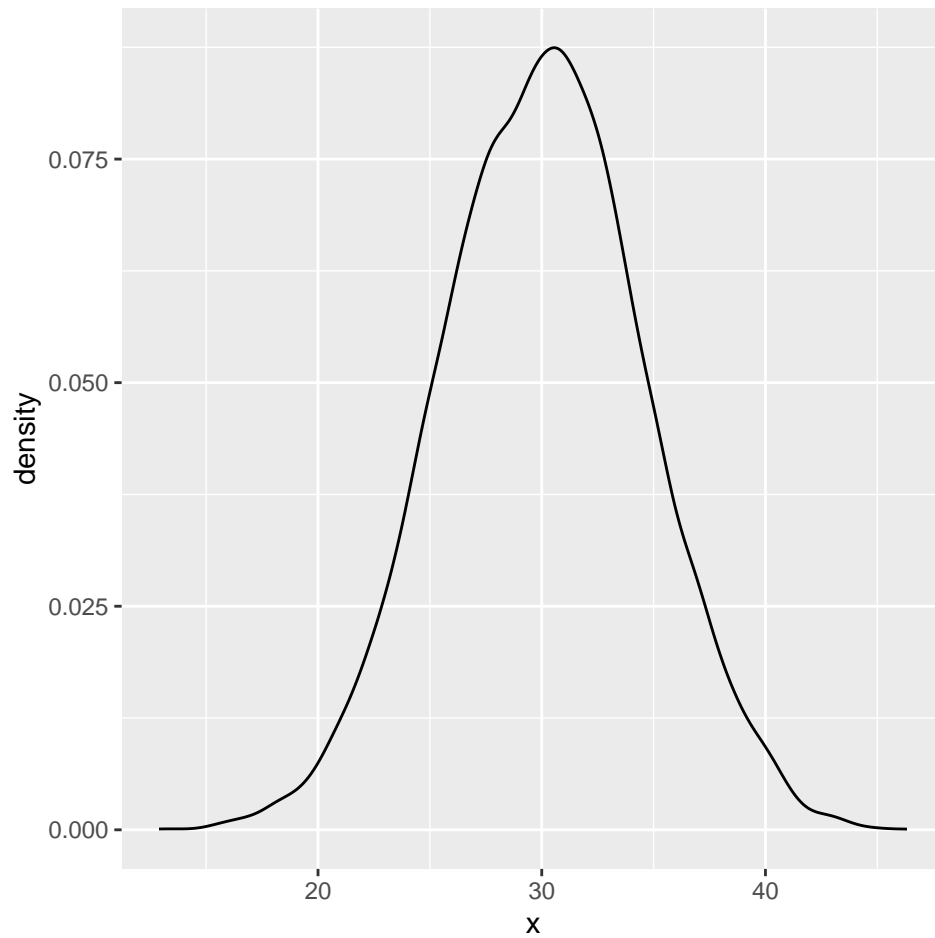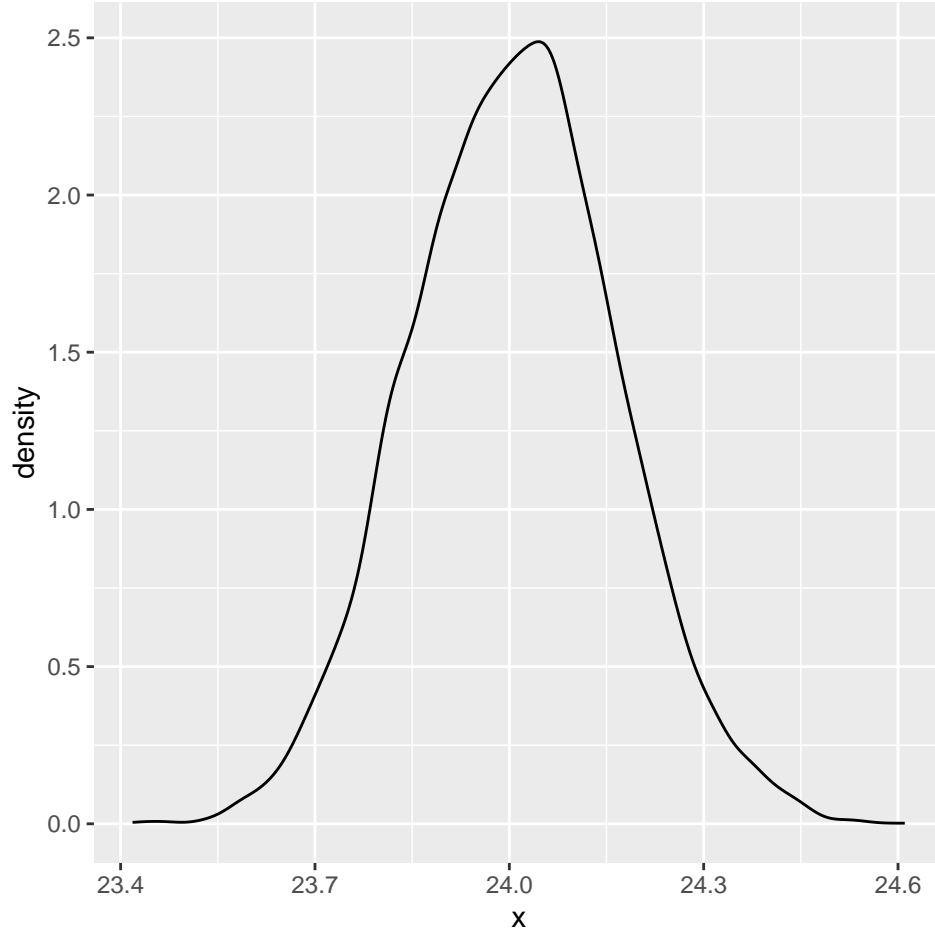
```
likelihood_data = c()
for (i in prior_data) {
    likelihood_data = c(likelihood_data, sum(log(dnorm(weather_perth$temp3pm, i, 5))))
}

ggplot(data=data.frame(x = prior_data, y = likelihood_data), aes(x=prior_data, y=likelihood_data)) +
    geom_line() +
    geom_vline(xintercept = y_bar, col = "red")
```

```
post_data = rnorm(10000, post_mu, sqrt(post_var))

ggplot(data=data.frame(x=post_data), aes(x=x)) +
    geom_density()
```

The prior is normal distribution

The Normal-Normal conjugacy framework assumes that our observations $X_1, X_2, ..., X_n$ are normally distributed around an unknown mean μ with a known standard deviation σ. We then choose a prior distribution for μ that is also normal with sd τ.

then we update our belief about μ based on the observed data using Bayes' theorem, resulting in a posterior distribution for μ, which is normal because of the conjugacy.

The posterior mean and variance can be derived by

$\mu_n = \frac{\tau^2 \sum_{i=1}^{n} X_i + \sigma^2 \mu_0}{n\tau^2 + \sigma^2}$

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

The posterior mean is a weighted average of the prior mean and the sample mean, where the weights are inversely proportional to their respective variances. The posterior variance decreases as the number of observations n increases, reflecting increased certainty about the value of μ.

Problem 4. Normal-Normal calculation

Prof. Abebe and Prof. Morales both recently finished their PhDs and are teaching their first statistics classes at Bayesian University. Their colleagues told them that the average final exam score across all students, μ, varies Normally from year to year with a mean of 80 points and a standard deviation of 4. Further, individual students' scores Y vary Normally around μ with a known standard deviation of 3 points.

a. (5pts) Prof. Abebe conducts the final exam and observes that his 32 students scored an average of 86 points. Calculate the posterior mean and variance of μ using the data from Prof. Abebe's class.

```
abebe <- summarize_normal_normal(80, 4, 3, 86, 32) %>% filter(model == "posterior")
```

```
## [1] "9 16 32"
```

```
print(abebe)
```

```
##       model     mean     mode      var        sd
## 1 posterior 85.89635 85.89635 0.2763916 0.5257295
```

b. (5pts) Prof. Morales conducts the final exam and observes that her 32 students scored an average of 82 points. Calculate the posterior mean and variance of μ using the data from Prof. Morales' class.

```
morales <- summarize_normal_normal(80, 4, 3, 82, 32) %>% filter(model == "posterior")
```

```
## [1] "9 16 32"
```

```
print(morales)
```

```
##       model     mean     mode      var        sd
## 1 posterior 81.96545 81.96545 0.2763916 0.5257295
```

c. (10pts) Next, use Prof. Abebe and Prof. Morales' combined exams to calculate the posterior mean and variance of μ. Interpret your results.

```
combined <- summarize_normal_normal(80, 4, 3, (86+82)/2, 64) %>% filter(model == "posterior")
```

```
## [1] "9 16 64"
```

```
print(combined)
```

```
##       model     mean     mode      var        sd
## 1 posterior 83.96515 83.96515 0.1393998 0.3733628
```

The combined data accounts for a total of 64 students and a y_bar of 84. Thus the posterior mean is between the posterior mean of Abebe's class and Morales's class. The variance is lower than both because of the doubled n.