Maloney_HW1

Emily Maloney

January 22, 2019

Homework - Chapters 2 and 3

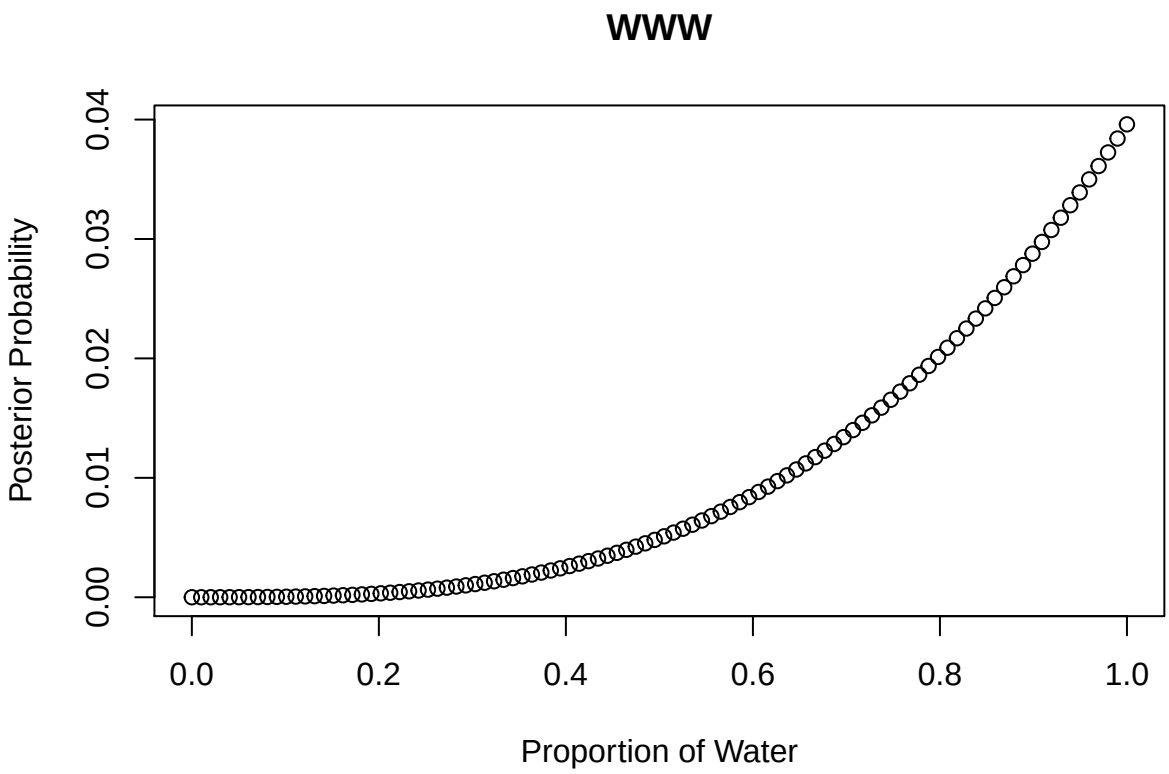
```
library(tidyverse)
library(rethinking)
library(tidybayes)
```

Chapter 2 Homework

Medium Problems

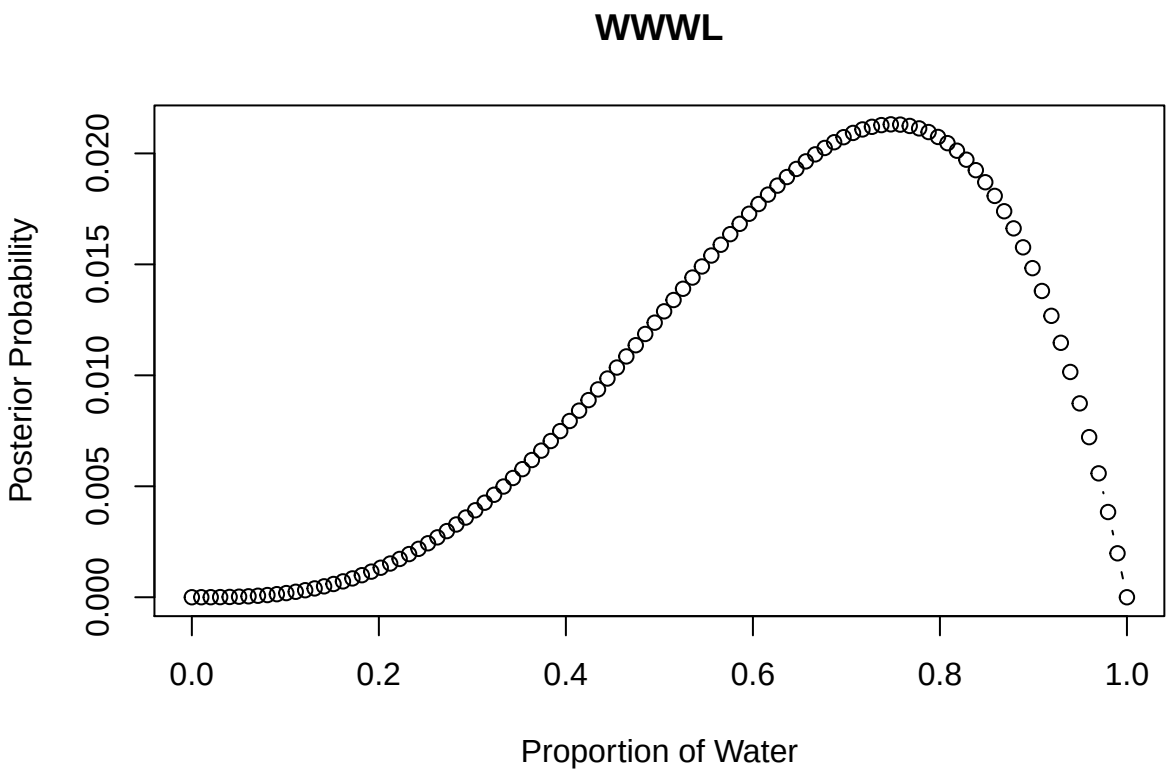
2M1

```
#WWW
              #define grid
              p_grid <- seq(from = 0, to = 1, length.out = 100)

              #define prior
              prior <- rep(1, 100)
              #prior <- exp(-5*abs(p_grid - 0.5))

              #compute likelihood at each value in the grid
              lh <- dbinom(3, size = 3, prob = p_grid)

              #compute product of lh & prior
              upost <- lh * prior

              #standardize post
              post <- upost/sum(upost)

              plot(p_grid, post, type = "b", main = "WWW", xlab = "Proportion of Water",
                   ylab = "Posterior Probability")
```



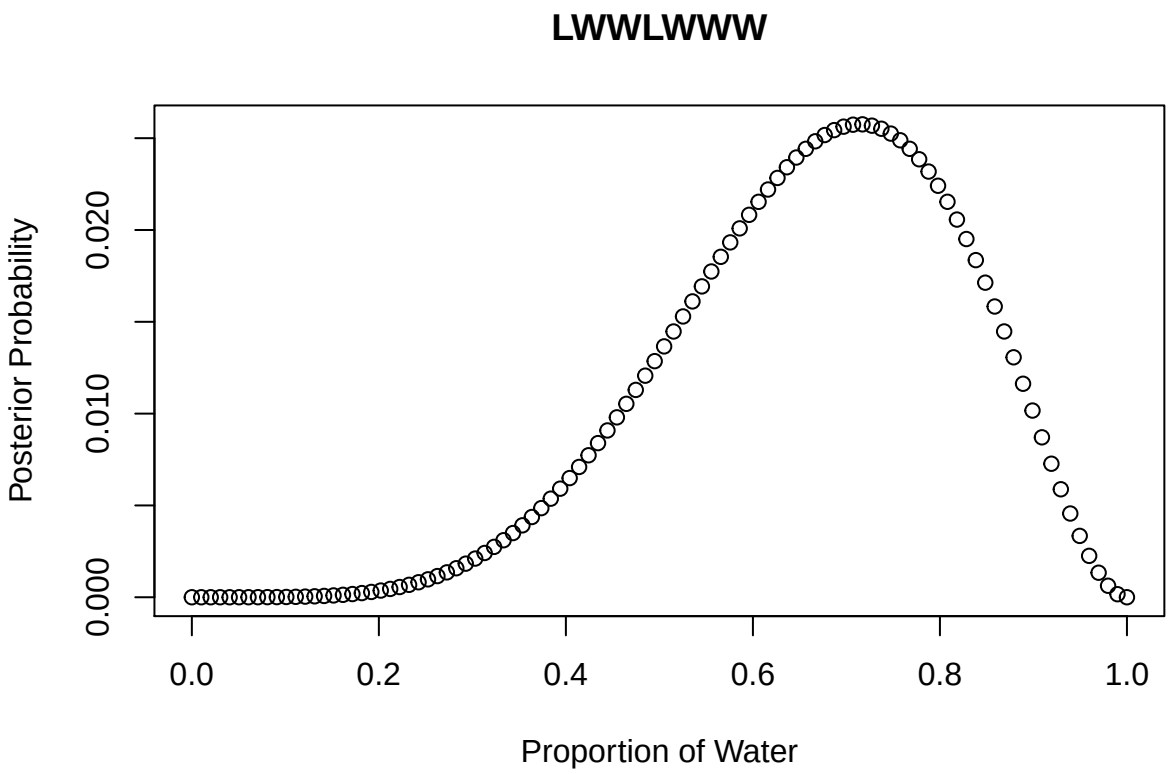
Assuming a flat prior, the grid approximate distribution of the observations WWW indicates that the most likely proportion of water on the globe is 1.

```
#WWWL
          #define grid
          p_grid <- seq(from = 0, to = 1, length.out = 100)

          #define prior
          prior <- rep(1, 100)
          #prior <- exp(-5*abs(p_grid - 0.5))

          #compute likelihood at each value in the grid
          lh <- dbinom(3, size = 4, prob = p_grid)

          #compute product of lh & prior
          upost <- lh * prior

          #standardize post
          post <- upost/sum(upost)

          plot(p_grid, post, type = "b", main = "WWWL", xlab = "Proportion of Water",
               ylab = "Posterior Probability")
```



Assuming a flat prior, the grid approximate distribution of the observations WWWL indicates that the most likely proportion of water on the globe is 0.75.

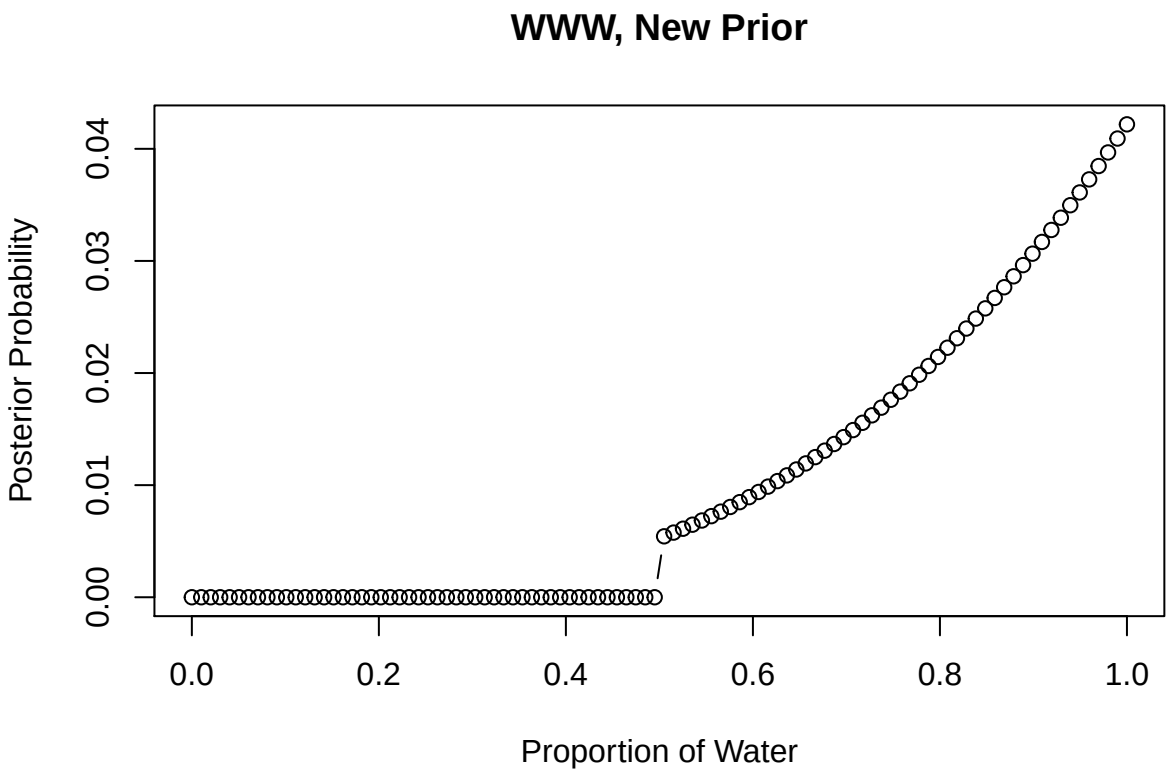
```
#LWWLWWW
          #define grid
          p_grid <- seq(from = 0, to = 1, length.out = 100)

          #define prior
          prior <- rep(1, 100)
          #prior <- exp(-5*abs(p_grid - 0.5))

          #compute likelihood at each value in the grid
          lh <- dbinom(5, size = 7, prob = p_grid)

          #compute product of lh & prior
          upost <- lh * prior

          #standardize post
          post <- upost/sum(upost)

          plot(p_grid, post, type = "b", main = "LWWLWWW", xlab = "Proportion of Water",
               ylab = "Posterior Probability")
```



Assuming a flat prior, the grid approximate distribution of the observations LWWLWWW indicates that the most likely proportion of water on the globe is 0.714.

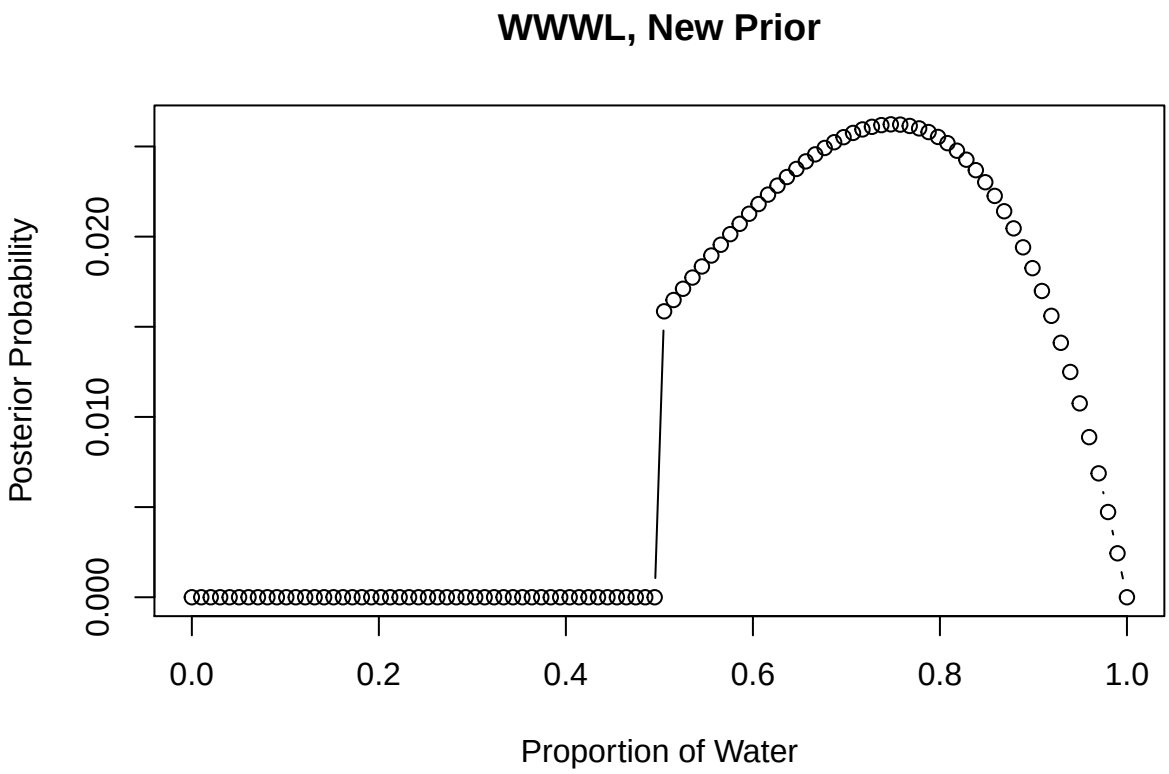
2M2

```
#WWW
              #define grid
              p_grid <- seq(from = 0, to = 1, length.out = 100)

              #define prior
              prior <- ifelse(p_grid < 0.5, 0, 0.75)

              #compute likelihood at each value in the grid
              lh <- dbinom(3, size = 3, prob = p_grid)

              #compute product of lh & prior
              upost <- lh * prior

              #standardize post
              post <- upost/sum(upost)

              plot(p_grid, post, type = "b", main = "WWW, New Prior", xlab = "Proportion of Water",
                   ylab = "Posterior Probability")
```



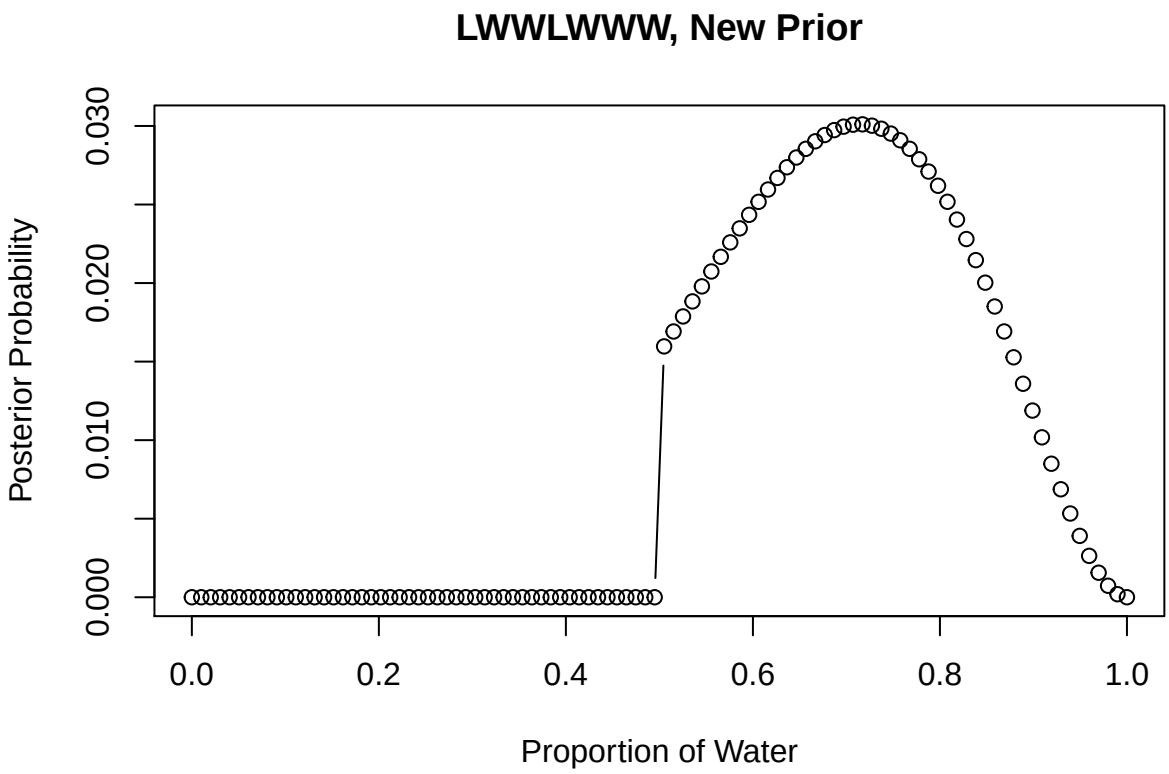
With a prior of 0 below 0.5 and 0.75 at or above 0.5, the most likely proportion of water is still 1, but now every proportion below 0.5 has a 0 posterior probability.

```
#WWWL
              #define grid
              p_grid <- seq(from = 0, to = 1, length.out = 100)

              #define prior
              prior <- ifelse(p_grid < 0.5, 0, 0.75)

              #compute likelihood at each value in the grid
              lh <- dbinom(3, size = 4, prob = p_grid)

              #compute product of lh & prior
              upost <- lh * prior

              #standardize post
              post <- upost/sum(upost)

              plot(p_grid, post, type = "b", main = "WWWL, New Prior",
                   xlab = "Proportion of Water",
                   ylab = "Posterior Probability")
```



With a prior of 0 below 0.5 and 0.75 at or above 0.5, the most likely proportion of water is still 0.75, but now every proportion below 0.5 has a 0 posterior probability.

```
#LWWLWWW
  #define grid
  p_grid <- seq(from = 0, to = 1, length.out = 100)

  #define prior
  prior <- ifelse(p_grid < 0.5, 0, 0.75)
  #prior <- exp(-5*abs(p_grid - 0.5))

  #compute likelihood at each value in the grid
  lh <- dbinom(5, size = 7, prob = p_grid)

  #compute product of lh & prior
  upost <- lh * prior

  #standardize post
  post <- upost/sum(upost)

  plot(p_grid, post, type = "b", main = "LWWLWWW, New Prior",
       xlab = "Proportion of Water",
       ylab = "Posterior Probability")
```



With a prior of 0 below 0.5 and 0.75 at or above 0.5, the most likely proportion of water is still 0.714, but now every proportion below 0.5 has a 0 posterior probability.

2M3

```
pwe <- 0.7
plm <- 1
lh <- 0.3/1
priorodds <- 1
postodds <- priorodds*lh
(post <- postodds/(postodds + 1))
```

```
## [1] 0.2307692
```

The posterior probability that the globe is the Earth, given seeing land is 0.2307692.

2M4

```
ww <- 0
bw <- 1
bb <- 2

(p <- bb/(bb + bw + ww))
```

```
## [1] 0.6666667
```

Given the fact that a card with two white sides cannot produce a black side facing up, a black and white

card can produce a black side facing up 1 way, and a card with two black sides can produce a black side facing up two ways, the probablity that the other side of a card with a black face up is 2/3.

2M5

```
ww <- 0
bw <- 1
bb <- 2*2

(p <- bb/(bb + bw + ww))
```

```
## [1] 0.8
```

If there are two cards that have black on both sides, the probability that a card with a black side facing up also has black on the other side is now 4/5.

2M6

```
ww <- 0*3
bw <- 1*2
bb <- 2

(p <- bb/(bb + bw + ww))
```

```
## [1] 0.5
```

If there are two ways to pull out a black and white card and 3 ways to pull out a white and white card for every way to pull out a black and black card, the probability that the other side of the drawn black card is black is now 1/2.

2M7

```
ww <- 0 * (1 + 0)
bb <- 2 * (2 + 1)
bw <- 1 * (0 + 2)

(p <- bb/(bb + bw + ww))
```

```
## [1] 0.75
```

If a second card is drawn with a white side face up, the probability that the first card with the black side facing up has black on the other side, is now 0.75.

Hard Problems

2H1

```
priorodds <- 1
pta <- 0.1
ptb <- 0.2

postpa <- (pta/ptb)/(pta/ptb + 1)
postpb <- (ptb/pta)/(ptb/pta + 1)

(ptwins <- postpa * .1 + postpb * .2)
```

```
## [1] 0.1666667
```

The probability of her next birth being twins is 0.1666667.

2H2

```
(postpa <- (pta/ptb)/(pta/ptb + 1))
```

```
## [1] 0.3333333
```

The probability that the panda is from species A, assuming that we have only observed the first birth and that is was twins is 0.3333333.

2H3

```
psa <- 0.9
psb <- 0.8

priorodds <- 1/2

postodds <- (0.9/0.8) * priorodds

(postprob <- postodds/(postodds + 1))
```

```
## [1] 0.36
```

The posterior probability that the same panda mother having a singleton infant is from species A is 0.36.

2H4

```
ppa <- 0.8
ppb <- 1-0.65

lh <- ppa/ppb

#no birth data, prior = 1
(pprob <- lh/(lh + 1))
```

```
## [1] 0.6956522
```

```
#with birth data as prior
priorodds <- postodds

p_odds <- priorodds * lh

(postprobability <- p_odds/(p_odds + 1))
```

```
## [1] 0.5625
```

Without using the prior birth data, the posterior probability that the panda is species A, given the positive test result, is 0.6956522.

When incorporating the prior birth data, the posterior probability that the panda is species A, given the positive test result, is 0.5625.

Chapter 3 Homework

Easy Problems

```
#drawing the data
p_grid <- seq(from = 0, to = 1, length.out = 1000)
prior <- rep(1, 1000)
lh <- dbinom(6, size = 9, prob = p_grid)
post <- lh * prior
post <- post/sum(post)
set.seed(100)
samples <- sample(p_grid, prob = post, size = 1e4, replace = T)
samples <- as_tibble(samples)
```

3E1

```
(less20 <- samples %>% filter(value < 0.2) %>%
            summarise(sum = n()/1e4))
```

```
## # A tibble: 1 x 1
##      sum
##    <dbl>
## 1 0.0005
```

There is .0005 posterior probability that lies below p = 0.2.

3E2

```
(more80 <- samples %>% filter(value > 0.8) %>%
            summarise(sum = n()/1e4))
```

```
## # A tibble: 1 x 1
##     sum
##   <dbl>
## 1 0.112
```

There is 0.1117 posterior probability that lies above p = 0.8.

3E3

```
(bw2080 <- samples %>% filter(value > 0.2 & samples < 0.8) %>%
            summarise(sum = n()/1e4))
```

```
## # A tibble: 1 x 1
##     sum
##   <dbl>
## 1 0.888
```

There is 0.8878 posterior probability that lies between p = 0.2 and p = 0.8.

3E4

```
(quant <- quantile(samples$value, p = 0.2))
```

```
##       20%
## 0.5195195
```

20% of the posterior probability lies below p = 0.5195195.

3E5

```
(quant <- quantile(samples$value, p = 0.8))
```

```
##       80%
## 0.7567568
```

20% of the posterior probability lies above p = 0.7567568.

3E6

```
HPDI(samples$value, p = 0.66)
```

```
##     |0.66     0.66|
## 0.5205205 0.7847848
```

The values of p which contain the narrowest interval equal to 66% of the posterior probability are 0.521 and 0.785.

3E7

```
quantile(samples$value, p = 0.83)
```

```
##       83%
## 0.7687688
```

```
quantile(samples$value, p = 0.17)
```

```
##       17%
## 0.5005005
```

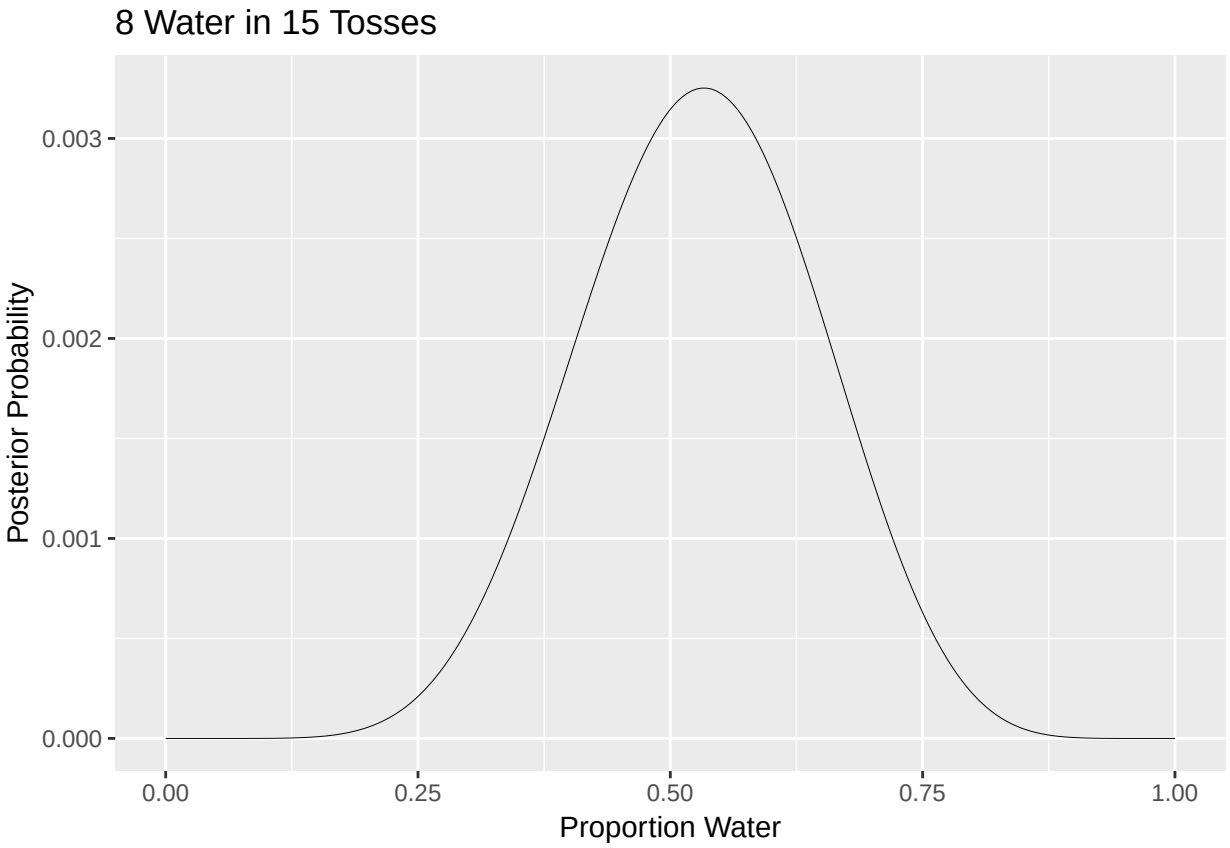
The values of p which contain 66% of the posterior probability, assuming equal posterior probability both below and above the interval, are 0.500 and 0.769.

Medium Problems

3M1

```
n <- 1000
n_success <- 8
n_trials <- 15

d <- tibble(p_grid = seq(from = 0, to = 1, length.out = n),
            prior = 1) %>%
     mutate(lh = dbinom(n_success, size = n_trials, prob = p_grid),
            post = lh * prior,
            post = post/sum(post))

d %>%
            ggplot(aes(x = p_grid, y = post)) +
            geom_line(size = 1/10) +
```

```
          labs(x = "Proportion Water",
               y = "Posterior Probability") + ggtitle("8 Water in 15 Tosses")
```



Assuming a flat prior and given an observation of 8 water out of 15 tosses, the posterior distribution for the proportion of water is centered around 0.533.

3M2

```
#drawing samples
samples <- tibble(samples = sample(d$p_grid, prob = d$post, size = 10000, replace = T)) %>%
           mutate(sample_n = 1:n())

#90% HPDI
HPDI(samples$samples, p = 0.9)
```

```
##      |0.9      0.9|
## 0.3383383 0.7317317
```

The 90% HPDI for p is 0.338-0.731.

3M3

```
ppc <- tibble(sample = rbinom(1e4, size = 15, prob = samples$samples))
(p8 <- ppc %>% filter(sample == 8) %>%
              summarise(sum = n()/1e4))
```

```
## # A tibble: 1 x 1
```

```
##     sum
##   <dbl>
## 1 0.143
```

Using data simulated from the model's posterior distribution, there is a 0.1428 probability of getting 8 tosses out of 15.

3M4

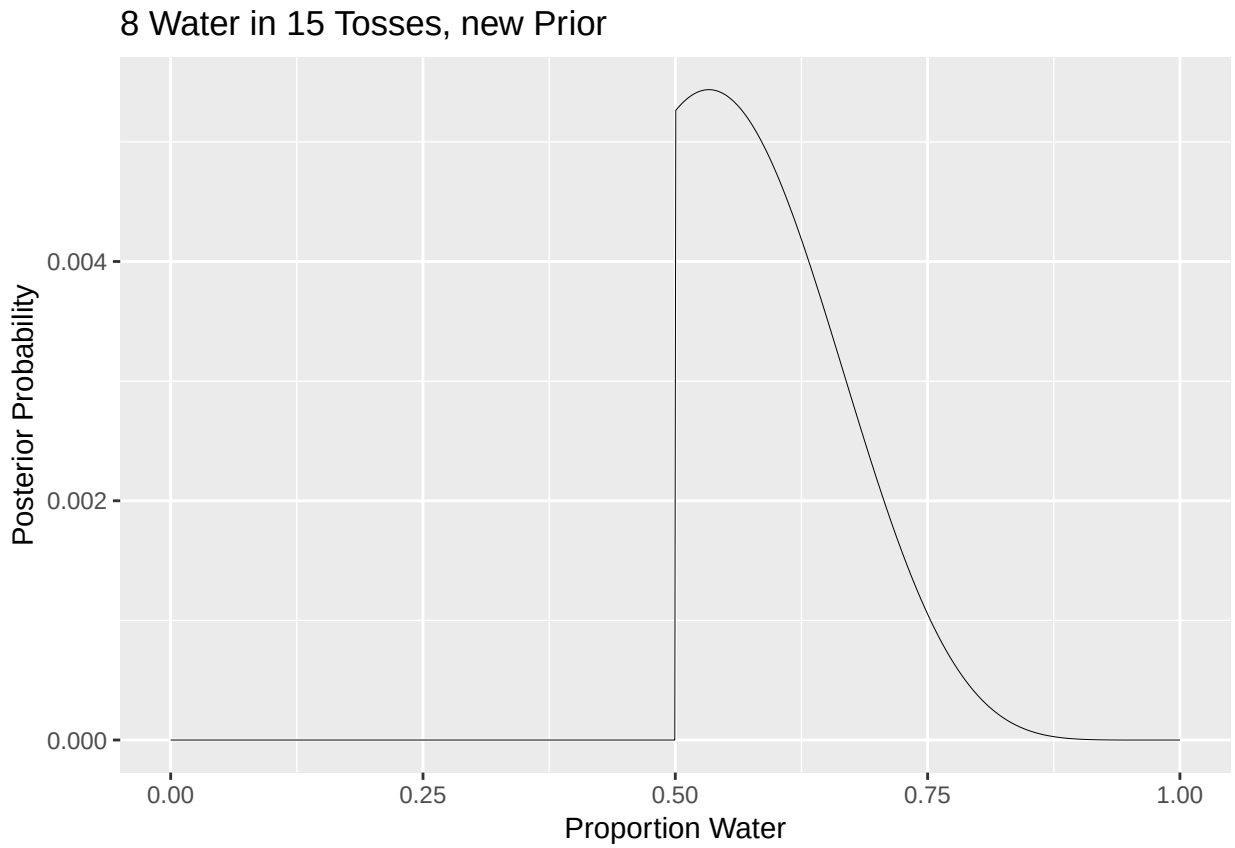
```
newsim <- tibble(sample = rbinom(1e4, size = 9, prob = samples$samples))
(p69 <- newsim %>% filter(sample == 6) %>%
                  summarise(sum = n()/1e4))
```

```
## # A tibble: 1 x 1
##     sum
##   <dbl>
## 1 0.170
```

There is a 0.1695 probability of getting 6 tosses out of 9.

3M5

```
#new simulation with a prior of 0 below 0.5 and 0.75 at or above 0.5.
n <- 1000
n_success <- 8
n_trials <- 15

d <- tibble(p_grid = seq(from = 0, to = 1, length.out = n),
            prior = ifelse(p_grid < 0.5, 0, 0.75)) %>%
     mutate(lh = dbinom(n_success, size = n_trials, prob = p_grid),
            post = lh * prior,
            post = post/sum(post))

d %>%
            ggplot(aes(x = p_grid, y = post)) +
            geom_line(size = 1/10) +
            labs(x = "Proportion Water",
                 y = "Posterior Probability") + ggtitle("8 Water in 15 Tosses, new Prior")
```




```
#drawing samples
samples <- tibble(samples = sample(d$p_grid, prob = d$post, size = 10000, replace = T)) %>%
           mutate(sample_n = 1:n())

#90% HPDI
hpdi90 <- HPDI(samples$samples, p = 0.9)
print(hpdi90)
```

```
##      |0.9      0.9|
## 0.5005005 0.7097097
```

```
#posterior predictive check
ppc <- tibble(sample = rbinom(1e4, size = 15, prob = samples$samples))
newp8 <- ppc %>% filter(sample == 8) %>%
              summarise(sum = n()/1e4)
print(newp8)
```

```
## # A tibble: 1 x 1
##     sum
##   <dbl>
## 1 0.159
```

```
#probability of observing 6 water in 9 tosses
newsim <- tibble(sample = rbinom(1e4, size = 9, prob = samples$samples))
newp69 <- newsim %>% filter(sample == 6) %>%
                  summarise(sum = n()/1e4)
print(newp69)
```

```
## # A tibble: 1 x 1
##     sum
##   <dbl>
## 1 0.236
```

Using a prior of 0 below 0.5 and 0.75 at or above 0.5, the new posterior distribution now follows the same general shape as before, but the curve has been cut at 0.5, with all posterior probability to the left of that point being 0.

The 90% HPDI for the new prior is 0.5005005, 0.7097097, which is smaller than the previous HPDI, which had been from 0.338 - 0.731. Both of the HPDIs contain the actual proportion of water, 0.7, but the narrower new HPDI range means that more posterior probability is contained in proportions of water closer to 0.7 than in the initial model based on a flat prior.

Running a posterior probability check on the data indicates that the likelihood of observing 8 water out of 15 tosses is 0.1592, which is higher than the probability calculated from the flat prior example, which had a probability of 0.1428.

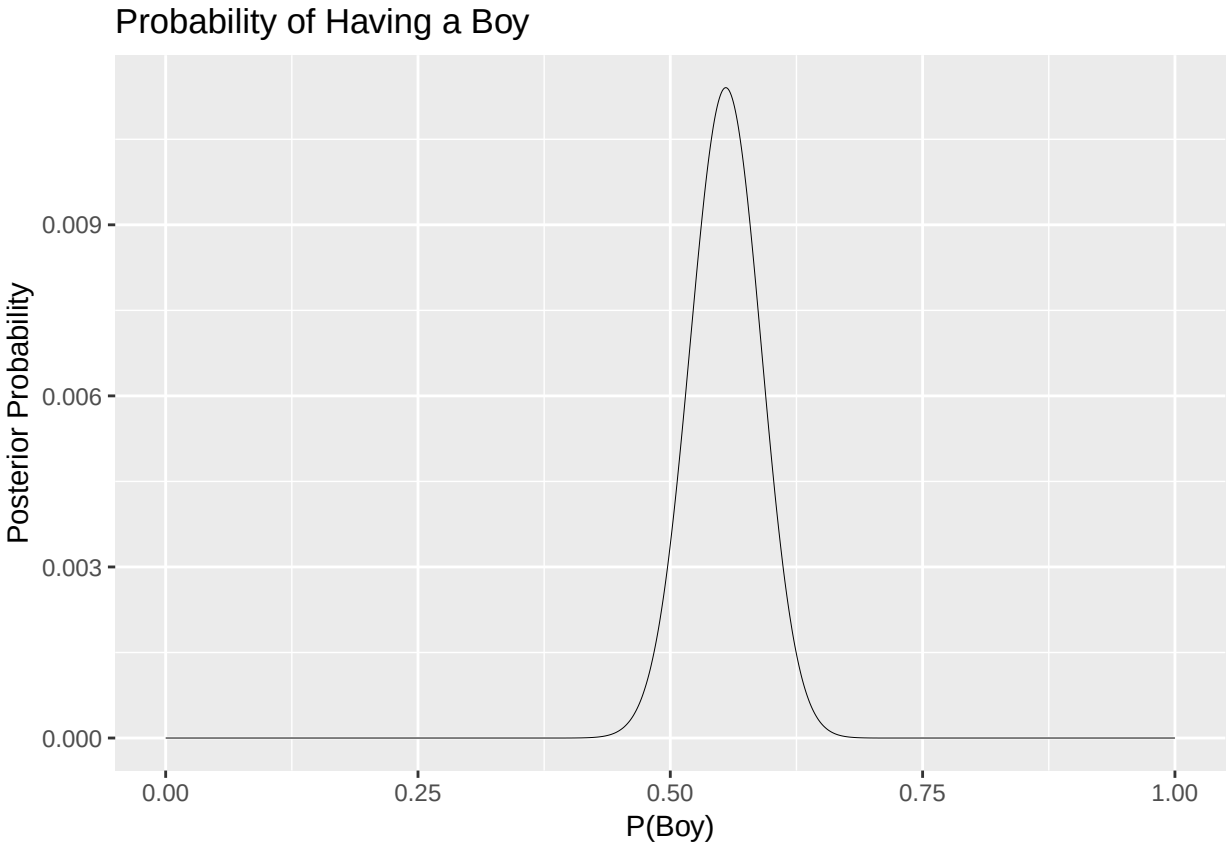
Similarly, using the same simulated data generated from the posterior distribution, the probability of observing 6 water in 9 tosses is 0.2357, which is much higher than in the simulated data from the flat prior, which was 0.1695.

Overall, including the prior in this case results in more accurate inferences and a better model of the actual world.

Hard Problems

3H1

```
#load data
library(rethinking)
data(homeworkch3)

#how many boys / how many children in total
totalboy <- sum(birth1) + sum(birth2)
totalchild <- length(birth1) + length(birth2)
pboy <- totalboy/totalchild

#generating posterior probablity
n <- 1000
n_success <- 111
n_trials <- 200

d <- tibble(p_grid = seq(from = 0, to = 1, length.out = n),
            prior = 1) %>%
     mutate(lh = dbinom(n_success, size = n_trials, prob = p_grid),
            post = lh * prior,
            post = post/sum(post))

#making plot
d %>%
            ggplot(aes(x = p_grid, y = post)) +
            geom_line(size = 1/10) +
            labs(x = "P(Boy)",
                 y = "Posterior Probability") + ggtitle("Probability of Having a Boy")
```




```
mdp <- max(d$post)

(pmax <- d$p_grid[d$post == mdp])
```

```
## [1] 0.5545546
```

Assuming a uniform prior, the posterior distribution for the probability fo a birth being a boy is centered around 0.55, which is the proportion of boys seen in the data. Indeed, the parameter value that maximizes the posterior probability is 0.5545546.

3H2

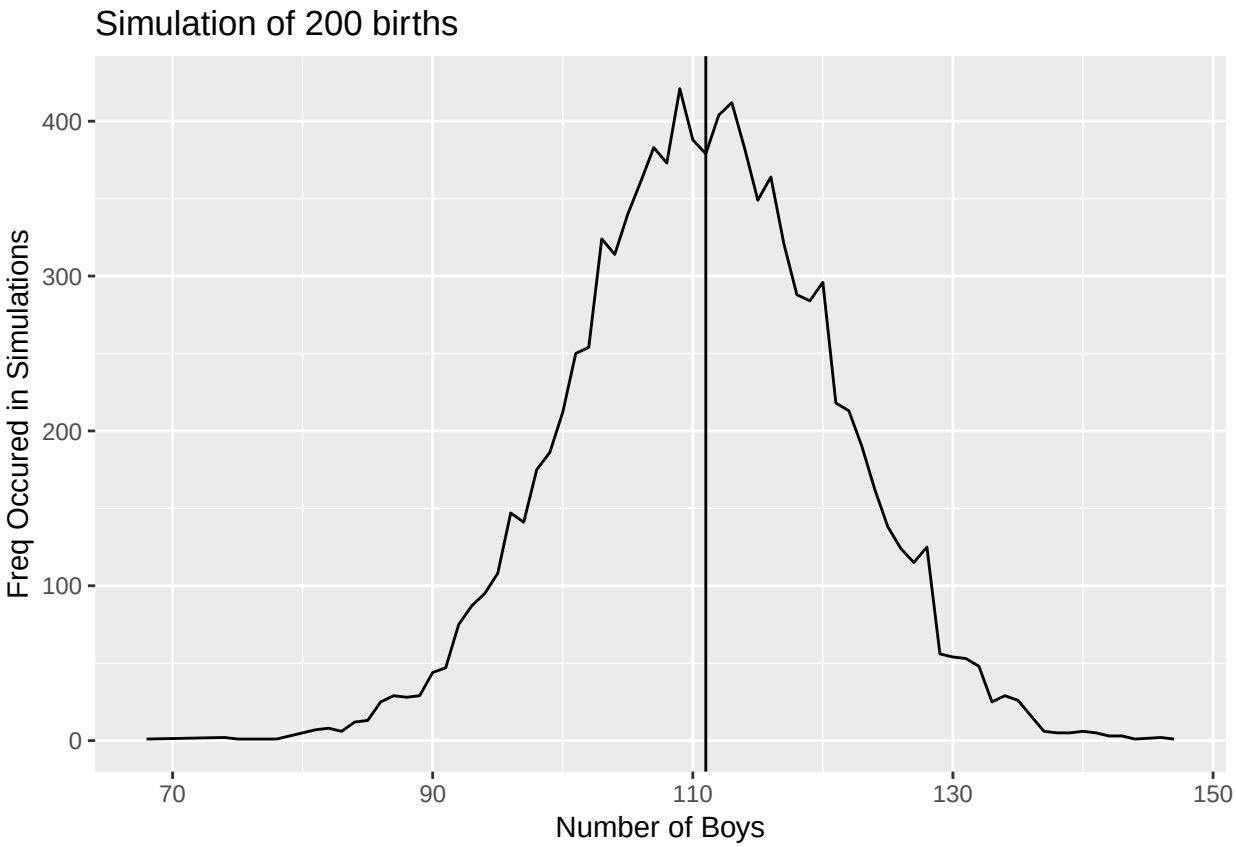
```
#drawing samples
samples <- tibble(samples = sample(d$p_grid, prob = d$post, size = 10000, replace = T)) %>%
           mutate(sample_n = 1:n())

#90% HPDI
(hpdisample <- HPDI(samples$samples, p = c(0.50,0.89, 0.97)))
```

```
##     |0.97     |0.89      |0.5      0.5|     0.89|     0.97|
## 0.4764765 0.5015015 0.5255255 0.5725726 0.6116116 0.6286286
```

The 50% HPDI ranges from 0.5255255 to 0.5725726. The 89% HPDI ranges from 0.5015015 to 0.6116116. The 97% HPDI ranges from 0.4764765 to 0.6286286.

3H3

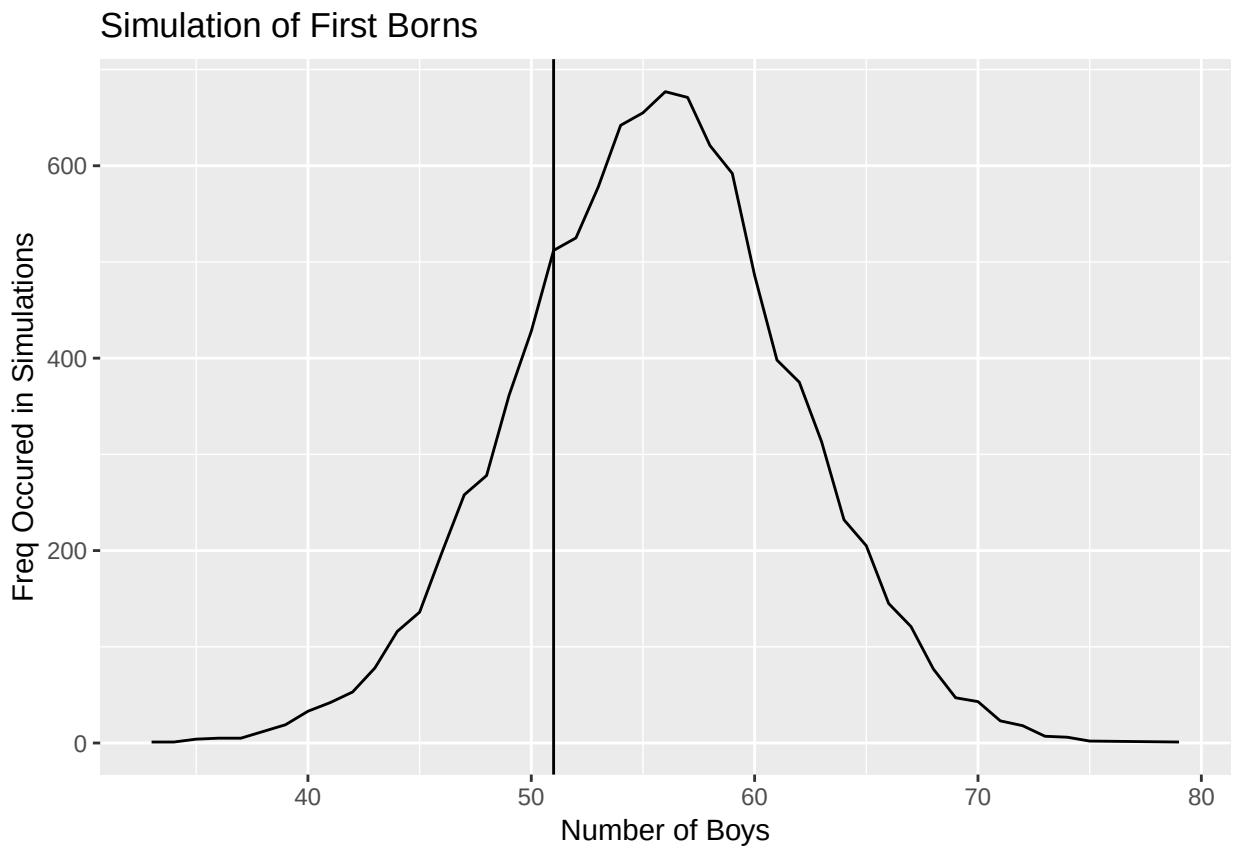
```
#simulation
boysim <- tibble(simbirths = rbinom(1e4, size = 200, prob = samples$samples))
p111 <- boysim %>% filter(simbirths == 111) %>%
              summarise(sum = n()/1e4)

#visualizing the sim
ggplot(data = boysim, mapping = aes(x = simbirths)) + geom_line(stat = "count") +
  geom_vline(xintercept = 111) + ggtitle("Simulation of 200 births") + xlab("Number of Boys") +
  ylab("Freq Occured in Simulations")
```



The model does look like it fits the data well, considering that the distribution of predictions does include the actual observation (indicated by the vertical line at 110) as a central, highly likely outcome.

3H4

```
#simulation with only first born

#simulation
boysim <- tibble(simbirths = rbinom(1e4, size = 100, prob = samples$samples))
p111 <- boysim %>% filter(simbirths == 51) %>%
              summarise(sum = n()/1e4)

#visualizing the sim
ggplot(data = boysim, mapping = aes(x = simbirths)) + geom_line(stat = "count") +
  geom_vline(xintercept = 51) + ggtitle("Simulation of First Borns") + xlab("Number of Boys") +
  ylab("Freq Occured in Simulations")
```

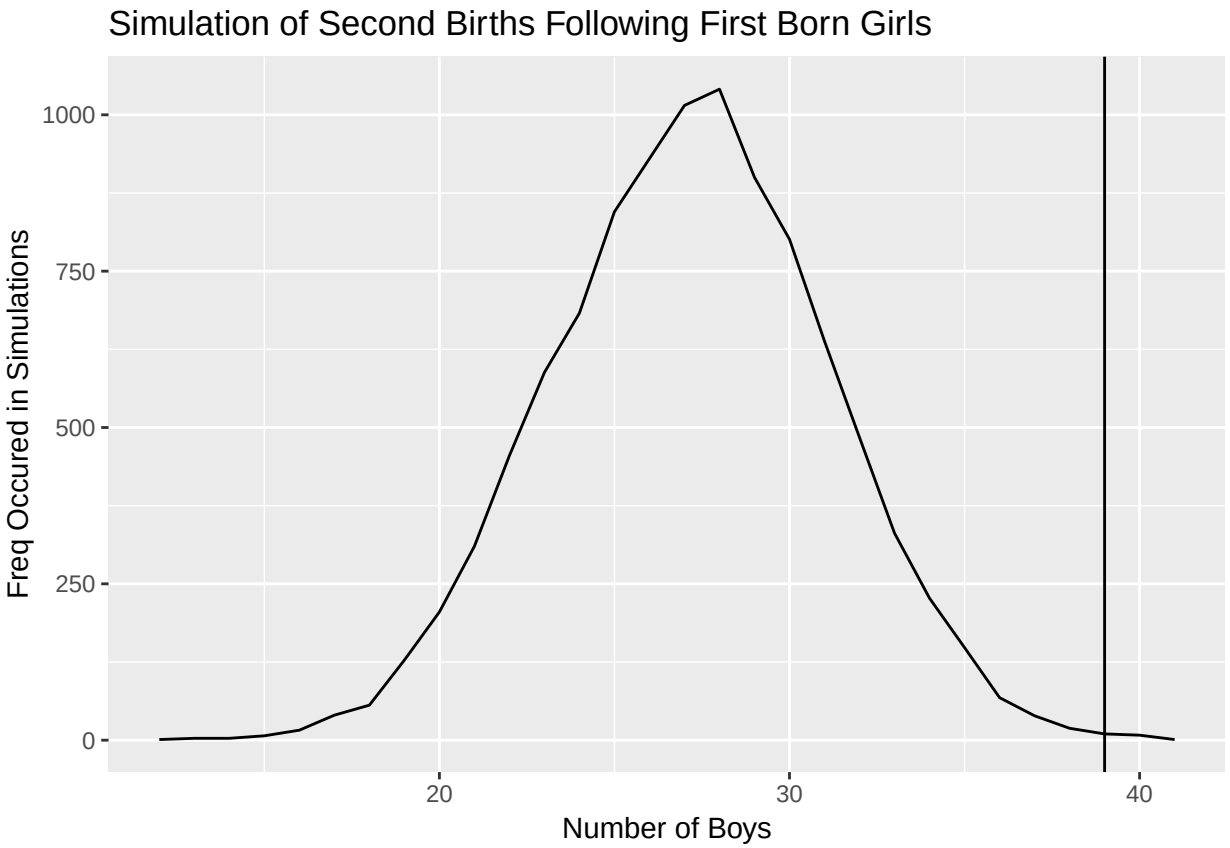


When only simulating first borns, the model does a less accurate job, considering that while the distribution of predictions does include the actual observation (indicated by the vertical line at 51), it is not the most central or likely outcome. It falls a bit to the left of the most likley outcome in the predicted distribution. This means that there are fewer first born boys in the data than our model predicts, indicating that there may be some element not captured by our model.

3H5

```
#simulation with number of first born girls

#simulation
boysim <- tibble(simbirths = rbinom(1e4, size = 49, prob = samples$samples))

#how many boys followed first born girls
allb <- as.data.frame(cbind(birth1, birth2))

bfg <- allb %>%
            filter(birth1 == 0) %>%
            summarise(sum(birth2))

#visualizing the sim
ggplot(data = boysim, mapping = aes(x = simbirths)) + geom_line(stat = "count") +
  geom_vline(xintercept = 39) + ggtitle("Simulation of Second Births Following First Born Girls") +
  xlab("Number of Boys") +
```

```
  ylab("Freq Occured in Simulations")
```



By simulating only second births following first born girls, we can begin to see a possible explanation for why the model works less well when first born and second born gender observations are considered separately. In this case, the actual number of boys following first born girls (39), falls significantly to the right of the central, most likely outcome generated by the model (around 27 boys). This indicates that the sex of first and second births are not independent, and namely, that when people have a female first born, they are likely to want to have a boy next and act accordingly.