hw3

Christopher Huong

2024-01-28

```
library(rethinking)
```

```
## Warning: package 'posterior' was built under R version 4.2.3
```

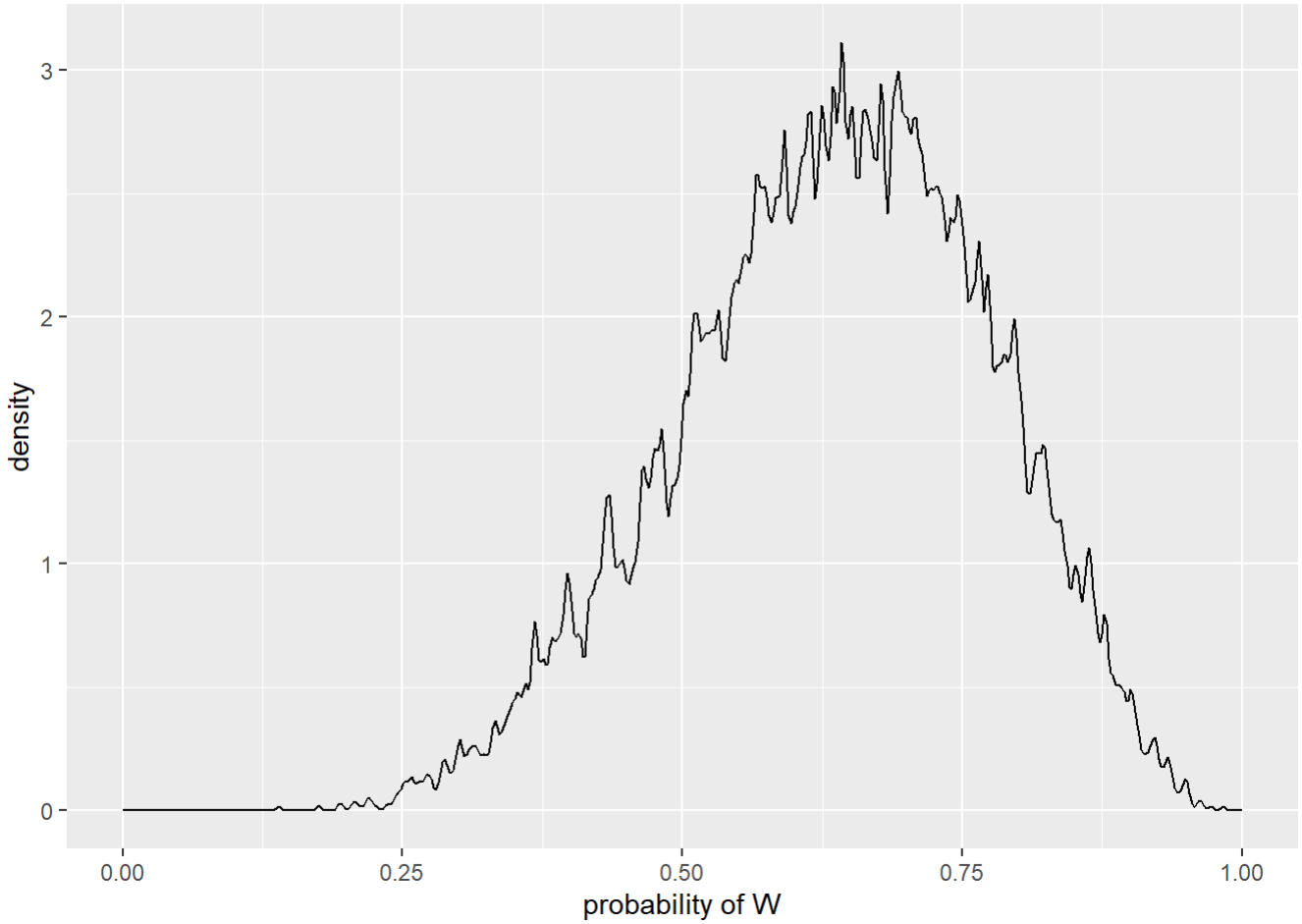
```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.2.3
```

```
p_grid <- seq(from=0, to=1, length.out=1000)
prior <- rep(1, 1000)
likelihood <- dbinom(6, size=9, prob=p_grid)
posterior <- likelihood * prior
posterior <- posterior / sum(posterior)

set.seed(100)
samples <- sample(p_grid, prob=posterior, size=10000, replace=T)
```

```
ggplot(data.frame(values=samples), aes(x=values)) +
  geom_density(adjust=.1) +
  xlim(0,1) + labs(x="probability of W")
```



EASY QUESTIONS

3E1. How much posterior probability lies below p = 0.2?

```
sum(samples < 0.2) / 10000
```

```
## [1] 4e-04
```

```
# sum the posterior distribution parameters for probabilities < 0.2
```

3E2. How much posterior probability lies above p = 0.8?

```
sum(samples > 0.8) / 10000
```

```
## [1] 0.1116
```

3E3. How much posterior probability lies between p = 0.2 and p = 0.8?

```
sum(samples < 0.8 & samples > 0.2) / 10000
```

```
## [1] 0.888
```

3E4. 20% of the posterior probability lies below which value of p?

```
quantile(samples, .20)
```

```
##       20%
## 0.5185185
```

3E5. 20% of the posterior probability lies above which value of p?

```
quantile(samples, 0.80)
```

```
##       80%
## 0.7557558
```

3E6. Which values of p contain the narrowest interval equal to 66% of the posterior probability?

```
HPDI(samples, prob=0.66)
```

```
##     |0.66     0.66|
## 0.5085085 0.7737738
```

3E7. Which values of p contain 66% of the posterior probability, assuming equal posterior probability both below and above the interval?

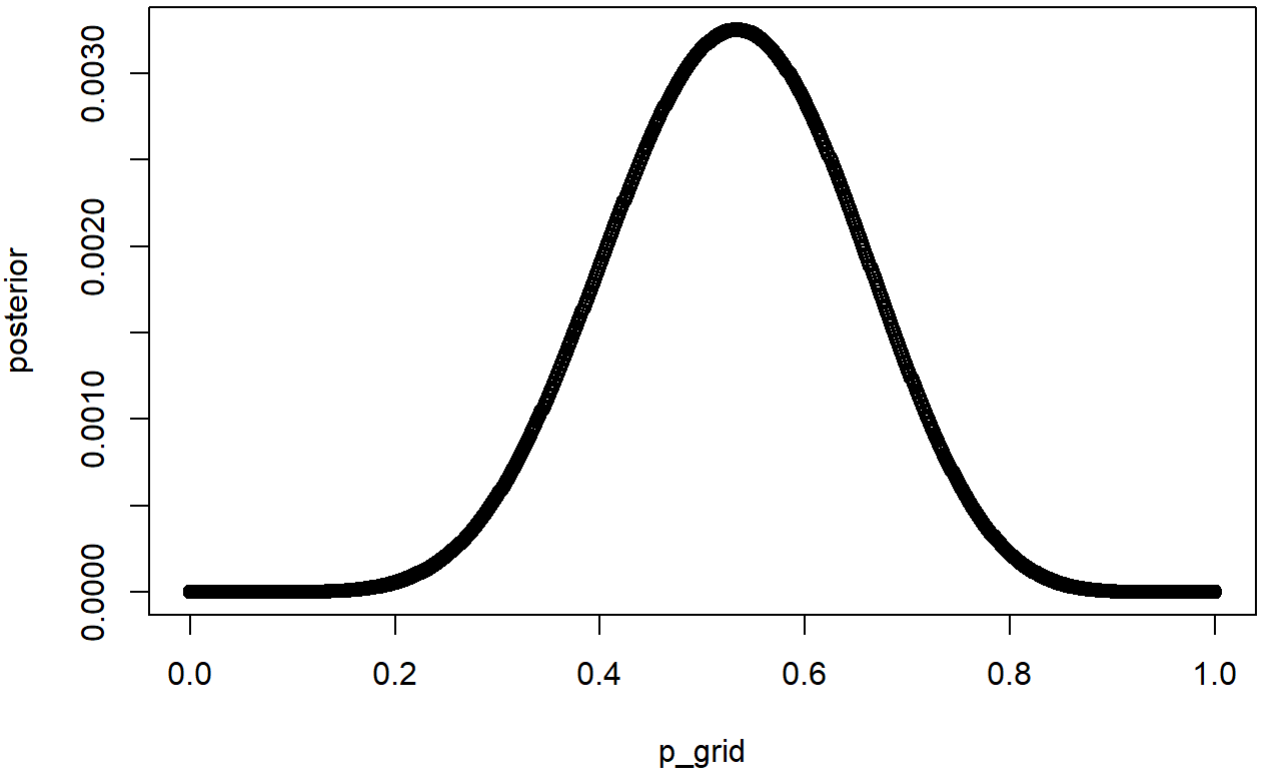
```
PI(samples, prob=0.66)
```

```
##       17%       83% 
## 0.5025025 0.7697698
```

MEDIUM QUESTIONS

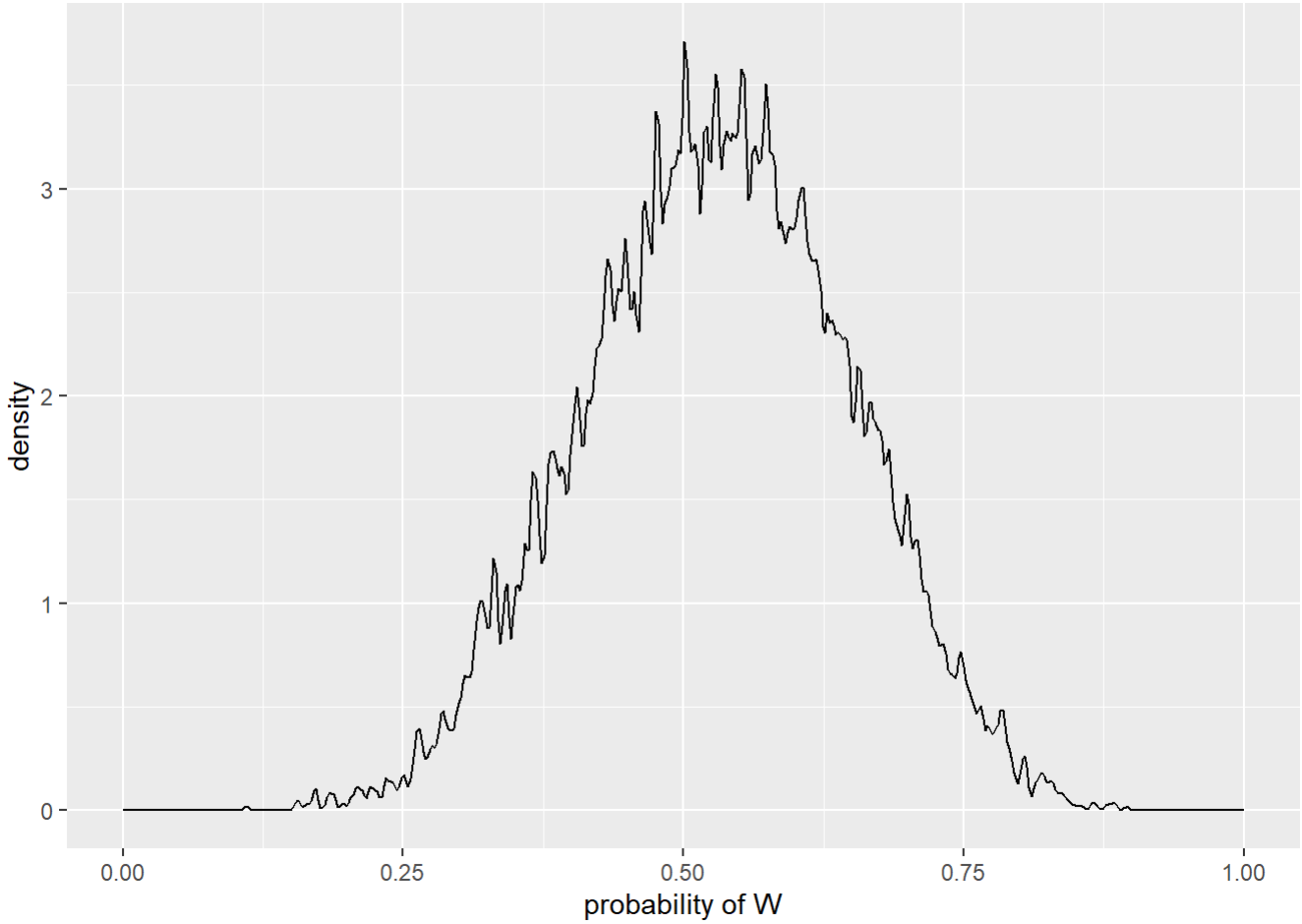
3M1. Suppose the globe tossing data had turned out to be 8 water in 15 tosses. Construct the posterior distribution, using grid approximation. Use the same flat prior as before.

```
likelihood <- dbinom(8, size=15, prob=p_grid)
posterior <- likelihood*prior
posterior <- posterior / sum(posterior)
plot(p_grid, posterior)
```



3M2. Draw 10,000 samples from the grid approximation from above. Then use the samples to calculate the 90% HPDI for p.

```
samples <- sample(p_grid, size=10000, prob=posterior, replace=T)
ggplot(data.frame(values=samples), aes(x=values)) +
  geom_density(adjust=.1) +
  xlim(0,1) + labs(x="probability of W")
```



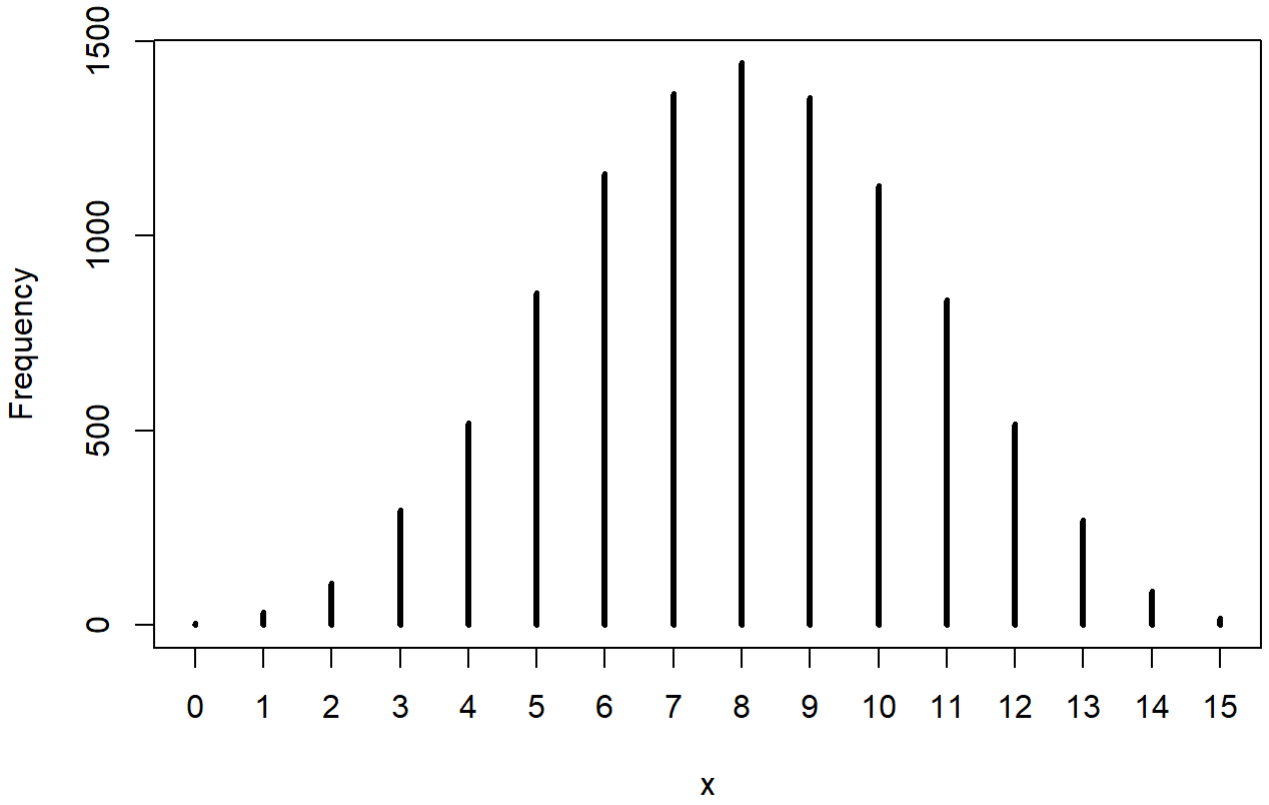

```
HPDI(samples, prob=0.90)
```

```
##      |0.9      0.9|
## 0.3293293 0.7167167
```

3M3. Construct a posterior predictive check for this model and data. This means simulate the distribution of samples, averaging over the posterior uncertainty

in p. What is the probability of observing 8 water in 15 tosses?

```
dummy_w <- rbinom(10000, size=15, prob=samples)
simplehist(dummy_w)
```




```
table(dummy_w)/10000
```

```
## dummy_w
##      0      1      2      3      4      5      6      7      8      9     10
## 0.0005 0.0034 0.0108 0.0295 0.0520 0.0853 0.1160 0.1366 0.1444 0.1355 0.1129
##     11     12     13     14     15
## 0.0837 0.0518 0.0270 0.0088 0.0018
```

```
#p(8/15|Data) = 0.15
```

3M4. Using the posterior distribution constructed from the new (8/15) data, now calculate the

probability of observing 6 water in 9 tosses.

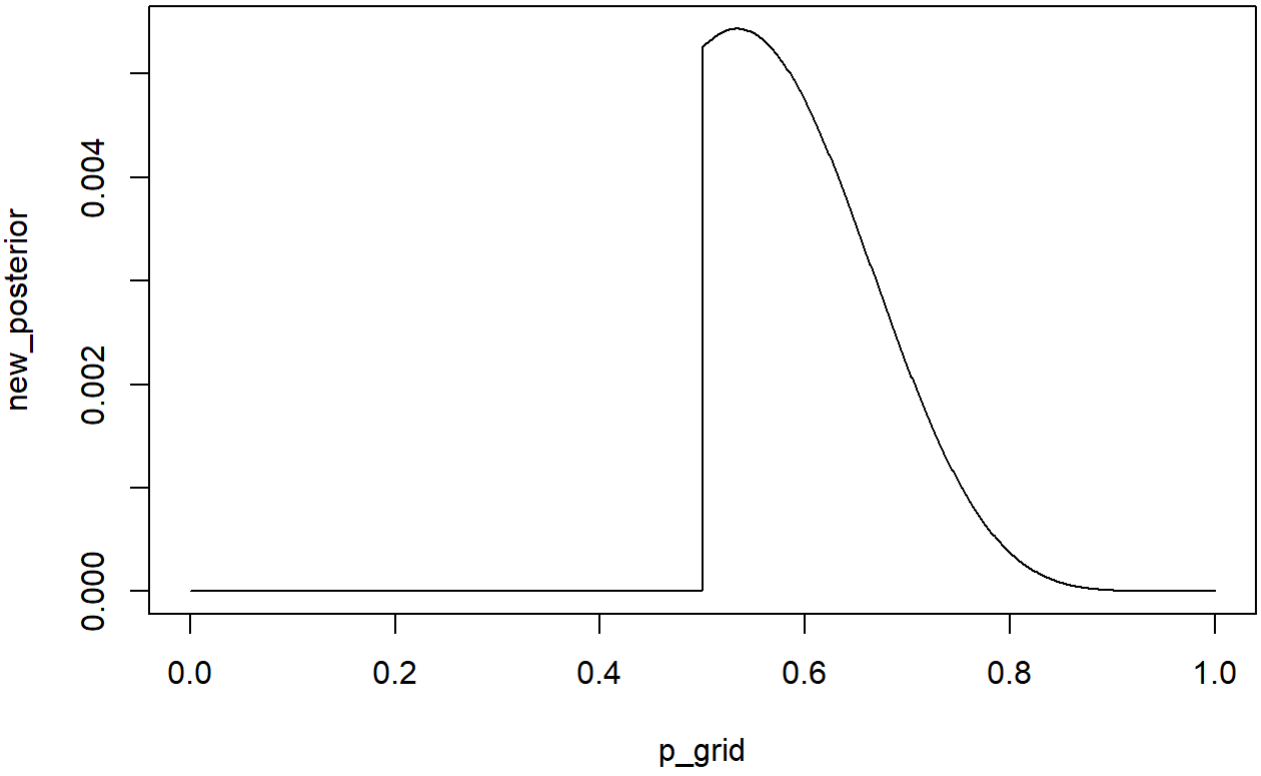
```
dummy_w <- rbinom(10000, size=9, prob=samples)
table(dummy_w) / 10000
```

```
## dummy_w
##      0      1      2      3      4      5      6      7      8      9
## 0.0060 0.0290 0.0758 0.1312 0.1930 0.2118 0.1751 0.1132 0.0527 0.0122
```

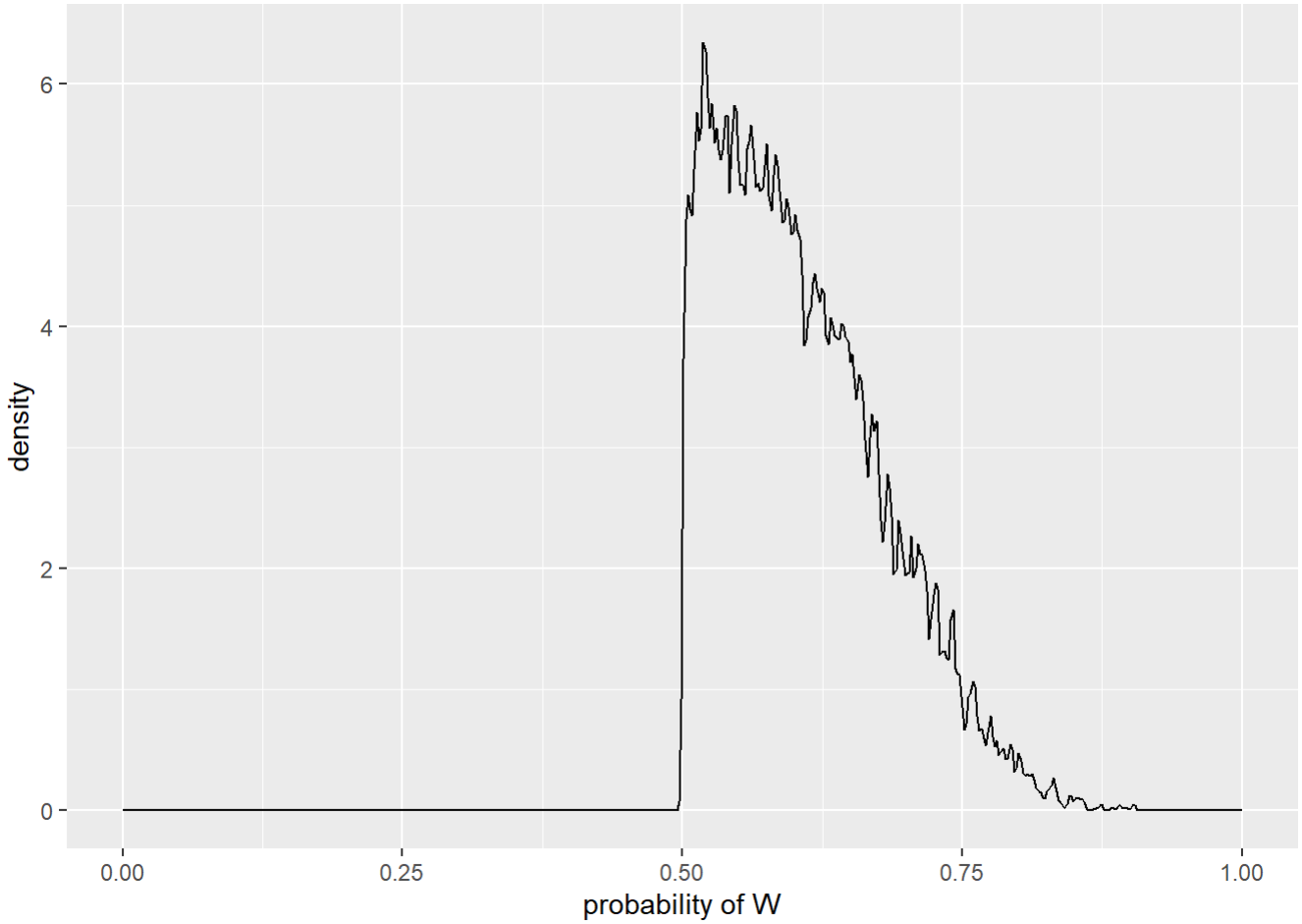
```
# p(6/9|Data) = 0.18
```

3M5. Start over at 3M1, but now use a prior that is zero below p = 0.5 and a constant above p = 0.5. This corresponds to prior information that a majority of the Earth's surface is water. Repeat each problem above and compare the inferences. What difference does the better prior make? If it helps, compare inferences (using both priors) to the true value p = 0.7.

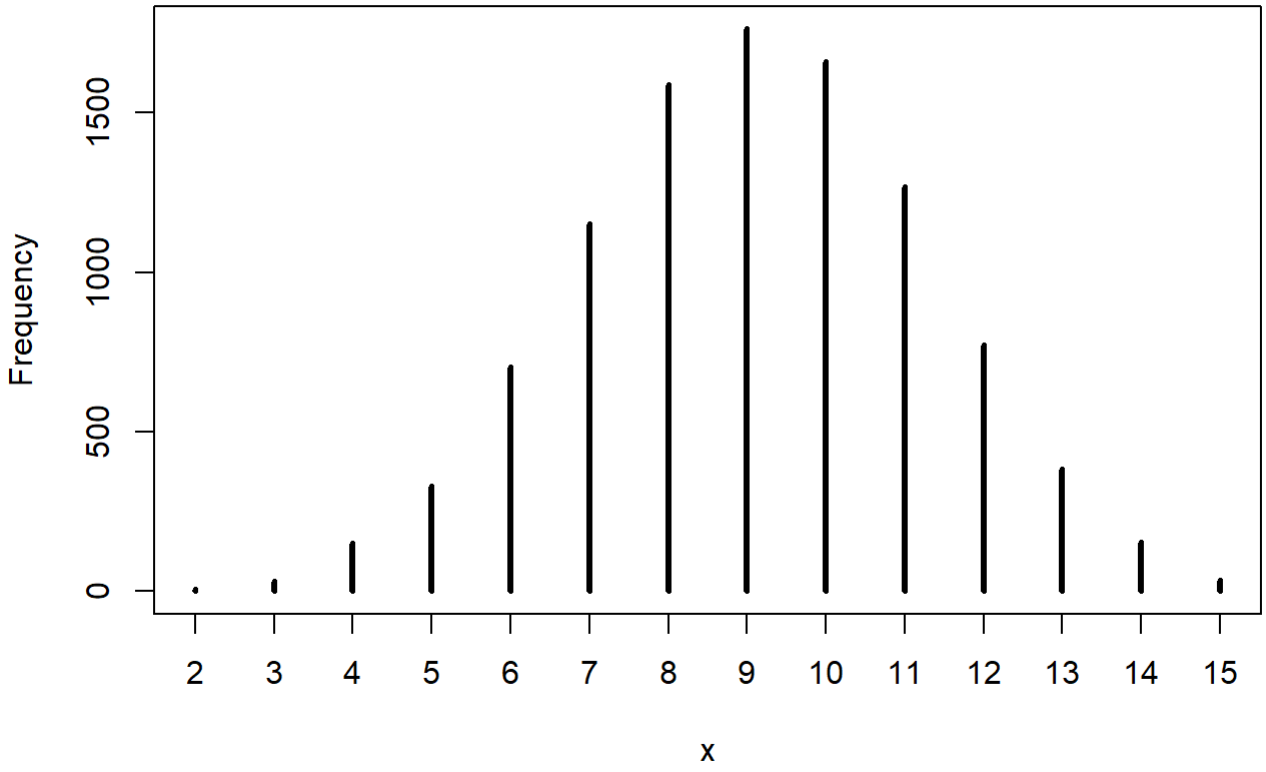
```
p_grid <- seq(from=0, to=1, length.out=1000)
new_prior <- ifelse(p_grid < 0.5, 0, 1)
new_likelihood <- dbinom(8, size=15, prob=p_grid)
new_posterior <- new_likelihood*new_prior
new_posterior <- new_posterior / sum(new_posterior)
plot(p_grid, new_posterior, type="l")
```




```
new_samples <- sample(p_grid, size=10000, prob=new_posterior, replace=T)
ggplot(data.frame(values=new_samples), aes(x=values)) +
  geom_density(adjust=.1) +
  xlim(0,1) + labs(x="probability of W")
```




```
dummy_w <- rbinom(10000, size=15, prob=new_samples)
simplehist(dummy_w)
```

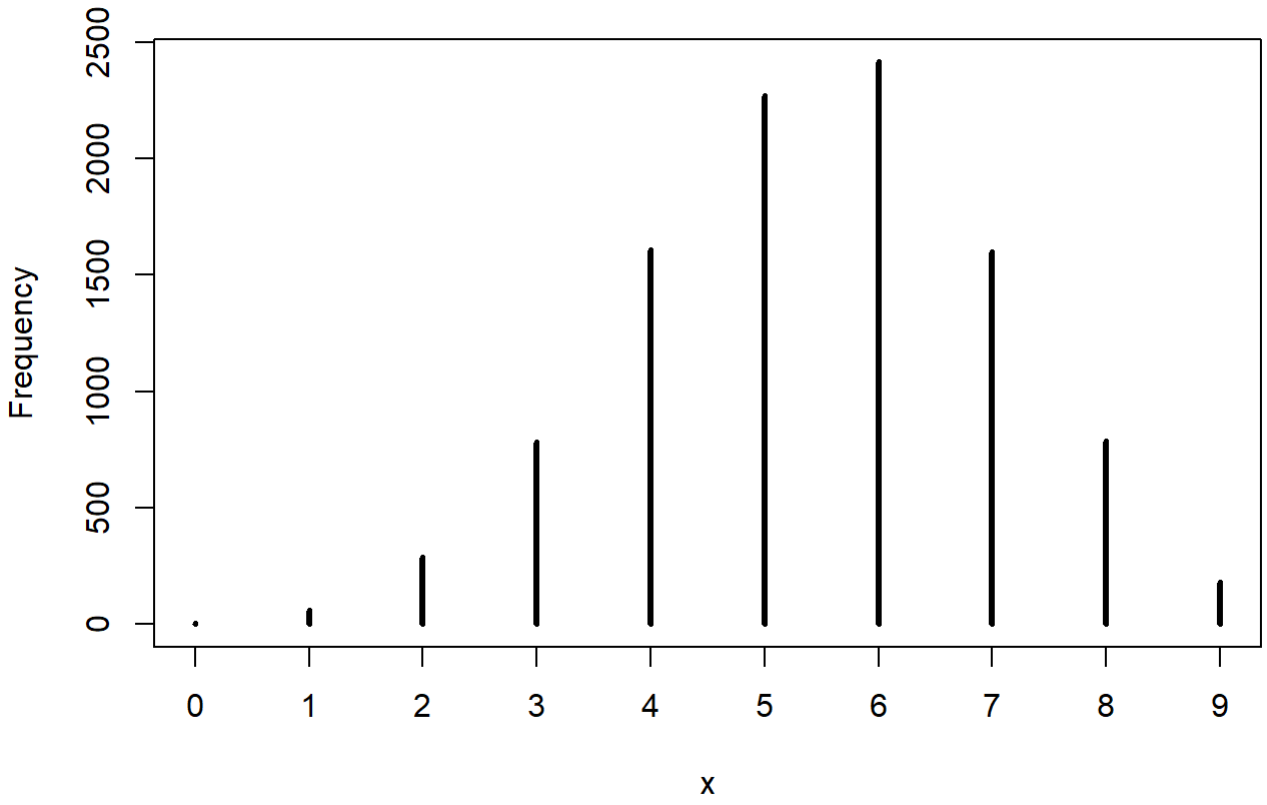



```
table(dummy_w) / 10000
```

```
## dummy_w
##      2      3      4      5      6      7      8      9     10     11     12
## 0.0006 0.0032 0.0152 0.0331 0.0703 0.1152 0.1589 0.1763 0.1662 0.1268 0.0772
##     13     14     15
## 0.0382 0.0154 0.0034
```

```
# p(8/15|Model) = 0.15
```

```
dummy_w <- rbinom(10000, size=9, prob=new_samples)
simplehist(dummy_w)
```




```
table(dummy_w) / 10000
```

```
## dummy_w
##      0      1      2      3      4      5      6      7      8      9
## 0.0005 0.0061 0.0287 0.0784 0.1610 0.2269 0.2415 0.1600 0.0786 0.0183
```

3M6. Suppose you want to estimate the Earth's proportion of water very precisely. Specifically, you want the 99% percentile interval of the posterior distribution of p to be only 0.05 wide. This means the

distance between the upper and lower bound of the interval should be 0.05. How many times will you have to toss the globe to do this?

```
p_grid <- seq(from=0, to=1, length.out=1000)
prior <- rep(1, 1000)

# function that computes the width of the 99th percentile interval of posterior when true p
toss <- function(N){
  likelihood <- dbinom(N*0.7, size=N, prob=p_grid)
  posterior <- likelihood*prior
  posterior <- posterior / sum(posterior)
  samples <- sample(p_grid, prob=posterior, size=10000, replace=T)
  interval <- PI(samples, prob=0.99)
  names(interval) <- NULL
  interval[2] - interval[1]
}
```

Test different numbers of globe tosses

```
toss(10)
```

```
## [1] 0.6206306
```

```
toss(100)
```

```
## [1] 0.2322372
```

```
toss(1000)
```

```
## [1] 0.07507508
```

```
toss(10000)
```

```
## [1] 0.02302302
```

Somewhere between 1000 and 10000 globe tosses to get a percentile interval of posterior distribution 0.05

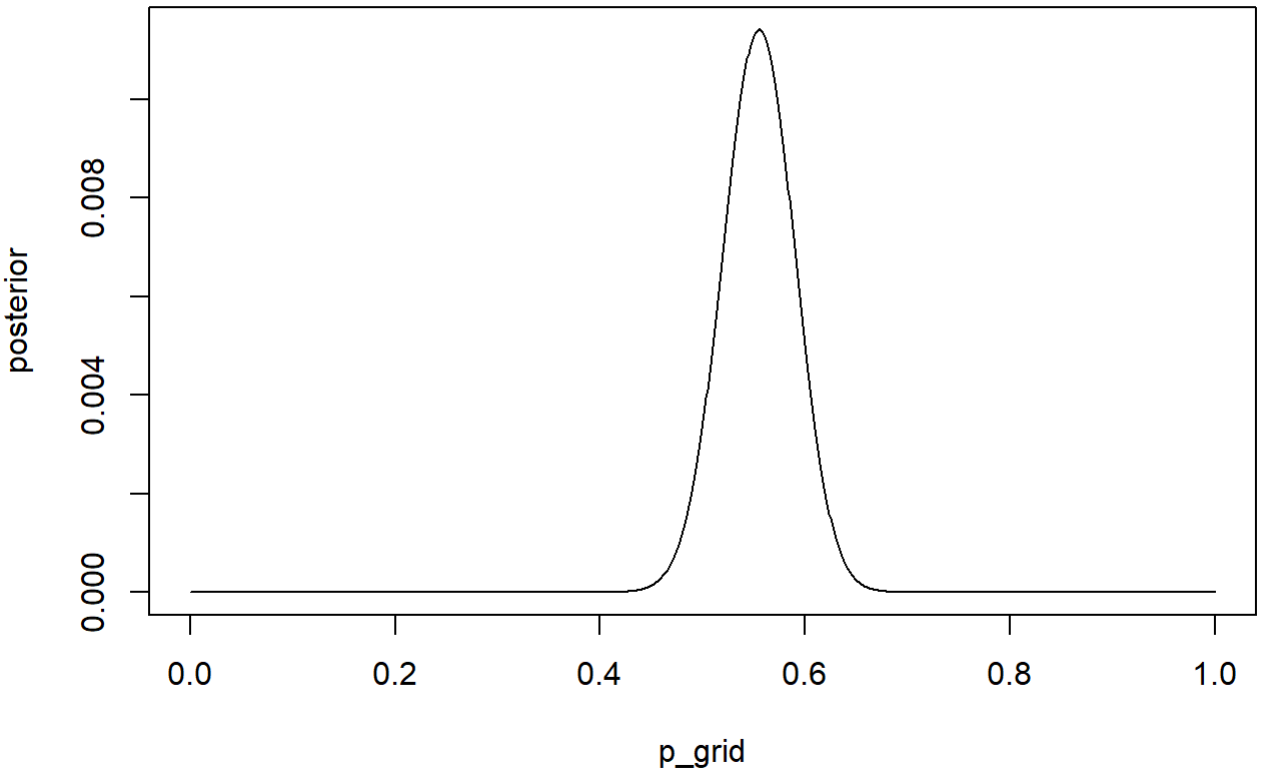
HARD QUESTIONS

```
rm(list=ls())
```

```
data(homeworkch3)
boys <- sum(birth1) + sum(birth2)
```

3H1. Using grid approximation, compute the posterior distribution for the probability of a birth being a boy. Assume a uniform prior probability. Which parameter value maximizes the posterior probability?

```
p_grid <- seq(from=0, to=1, length.out=1000)
prior <- rep(1, 1000)
likelihood <- dbinom(boys, size=200, prob=p_grid)
posterior <- likelihood * prior
posterior <- posterior / sum(posterior)
plot(p_grid, posterior, type="l")
```



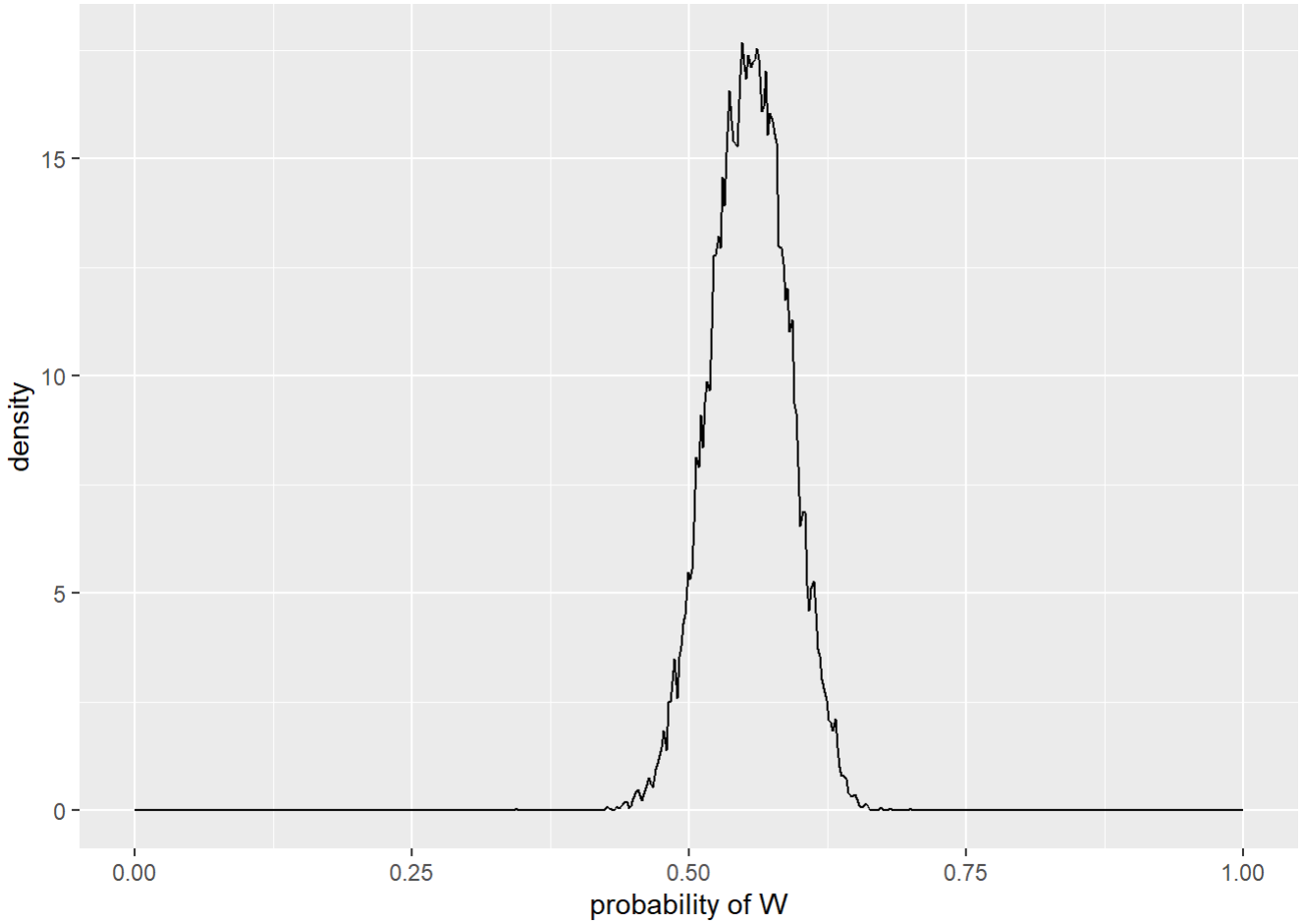

```
#find which prior probability has the highest posterior probability
p_grid[which.max(posterior)]
```

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

0.55 is the highest parameter value for the posterior probability. That is, after updating the prior with new data, the most probable probability for having a boy is p=0.55

3H2. Using the sample function, draw 10,000 random parameter values from the posterior distribution you calculated above. Use these samples to estimate the 50%, 89%, and 97% highest posterior density intervals.

```
samples <- sample(p_grid, prob=posterior, size=1e4, replace=T)
#sample probability values from the posterior distribution
ggplot(data.frame(values=samples), aes(x=values)) +
  geom_density(adjust=.1) +
  xlim(0,1) + labs(x="probability of W")
```




```
HPDI(samples, p=0.50)
```

```
##      |0.5      0.5|
## 0.5315315 0.5785786
```

```
HPDI(samples, p=0.89)
```

```
##     |0.89     0.89|
## 0.5005005 0.6126126
```

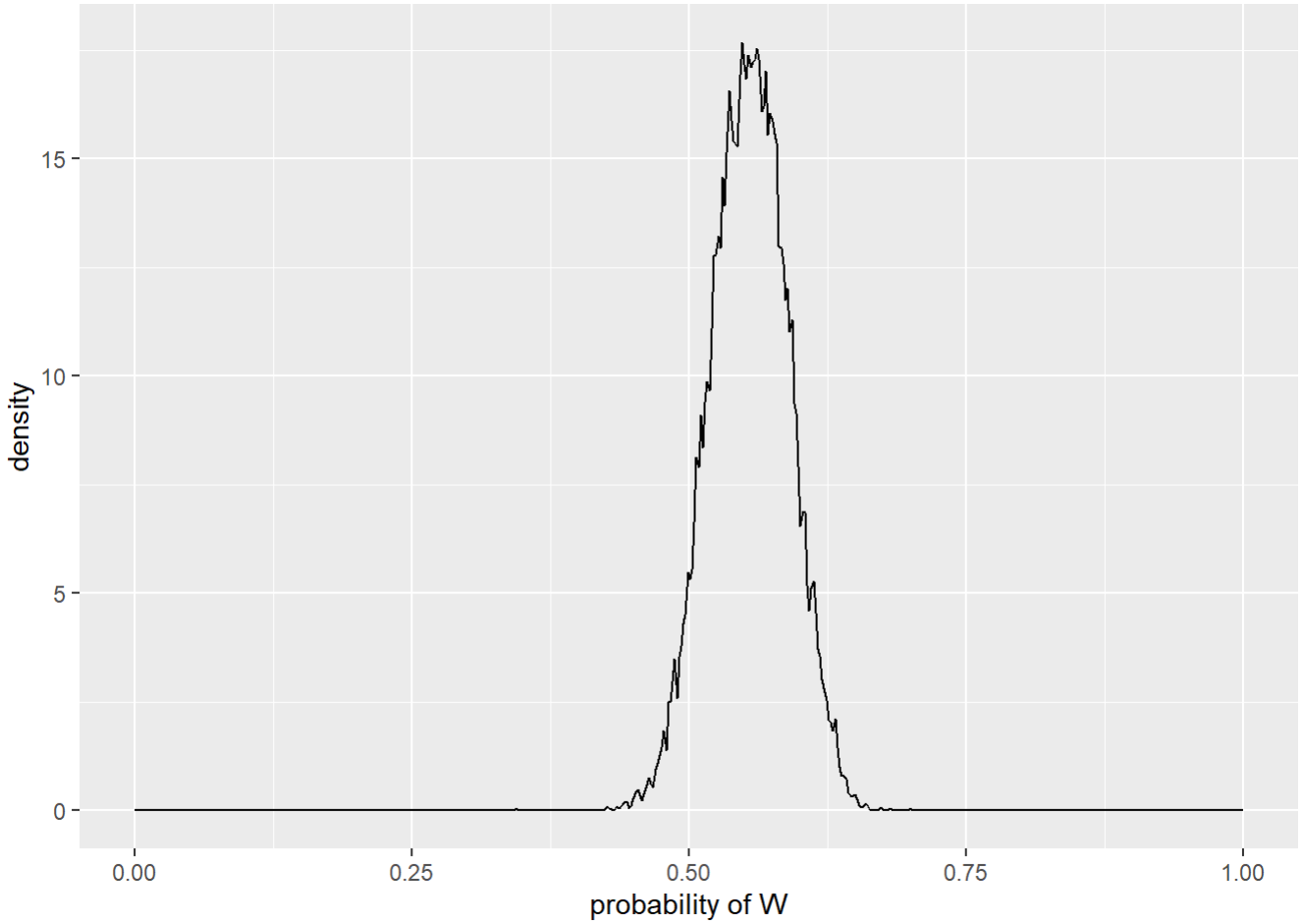
```
HPDI(samples, p=0.97)
```

```
##     |0.97     0.97|
## 0.4804805 0.6296296
```

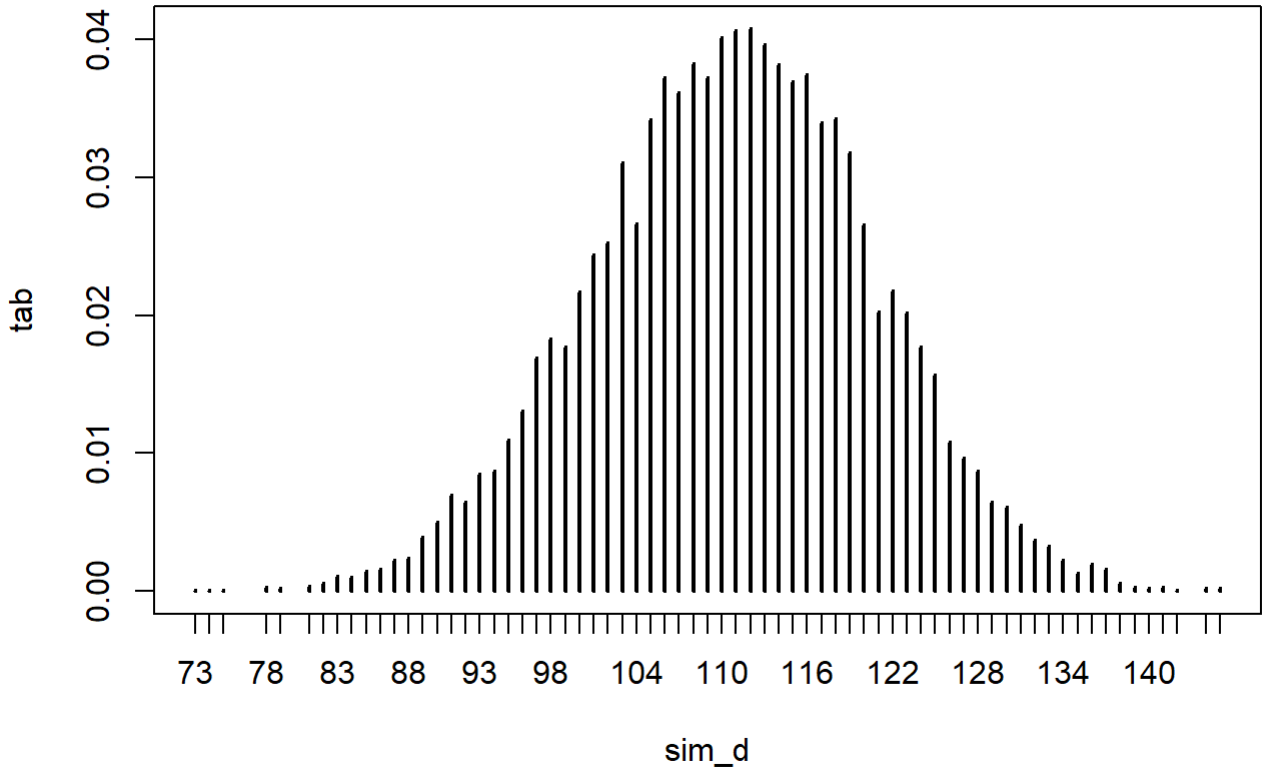
3H3. Use rbinom to simulate 10,000 replicates of 200 births. You should end up with 10,000 numbers, each one a count of boys out of 200 births. Compare the distribution of predicted numbers of boys to the actual count in the data (111 boys out of 200 births). There are many good ways to visualize the simulations, but the dens command (part of the rethinking package) is probably the easiest way in this case. Does it look like the model fits the data well? That is, does the distribution of predictions include the actual observation as a central, likely outcome?

```
sim_d <- rbinom(1e4, size=200, prob=samples)
```

```
ggplot(data.frame(values=samples), aes(x=values)) +
  geom_density(adjust=.1) +
  xlim(0,1) + labs(x="probability of W")
```



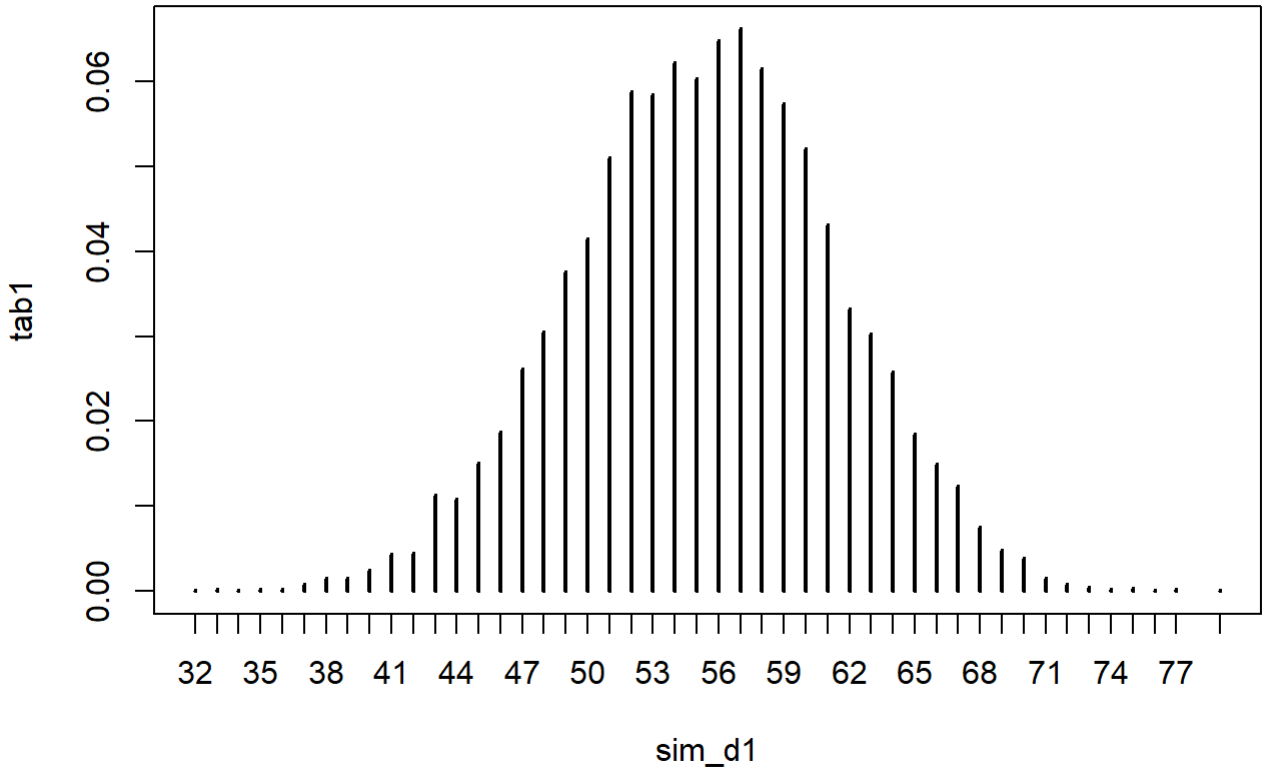

```
tab <- table(sim_d) / 1e4
plot(tab)
```



The posterior predictive distribution seems to center around the data (p=0.55, and boys=111)

3H4. Now compare 10,000 counts of boys from 100 simulated first borns only to the number of boys in the first births, birth1. How does the model look in this light?

```
sim_d1 <- rbinom(1e4, size=100, prob=samples)
tab1 <- table(sim_d1) / 1e4
plot(tab1)
```




```
sum(birth1)
```

```
## [1] 51
```

The distribution of simulated first-borns centers around 55 boys, whereas the data are 51 first-born as boys.

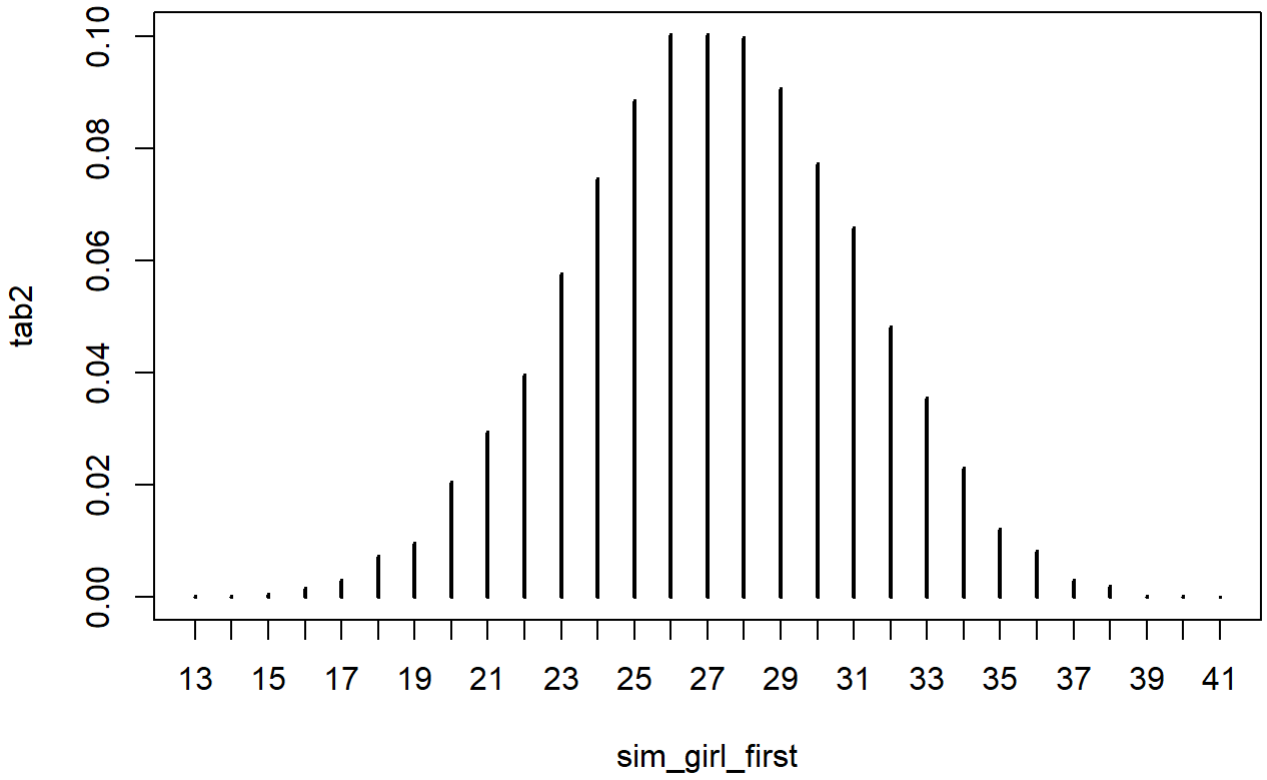
3H5. The model assumes that sex of first and second births are independent. To check this assumption, focus now on second births that followed female first borns. Compare 10,000 simulated counts of boys to only those second births that followed girls. To do this correctly, you need to count the number of first borns who were girls and simulate that many births, 10,000 times. Compare the counts of boys in your

simulations to the actual observed count of boys following girls. How does the model look in this light? Any guesses what is going on in these data?

```
girl_first <- birth2[birth1==0]

sim_girl_first <-rbinom(1e4, size=length(girl_first), prob=samples)
```

```
tab2 <- table(sim_girl_first) / 1e4
plot(tab2)
```




```
sum(girl_first==1)
```

```
## [1] 39
```

Of the 49 births that followed girls, 39 of those were boys. Yet, simulating 49 births under the assumption that births are independent of each other, the posterior predictive distribution centers around 27 boys, and gives very little probability to the observed data. Thus, something is wrong with our model. The probability of a 2nd born boy after a 1st born female is higher than the total probability of a boy birth.