Chapter 3

Theory

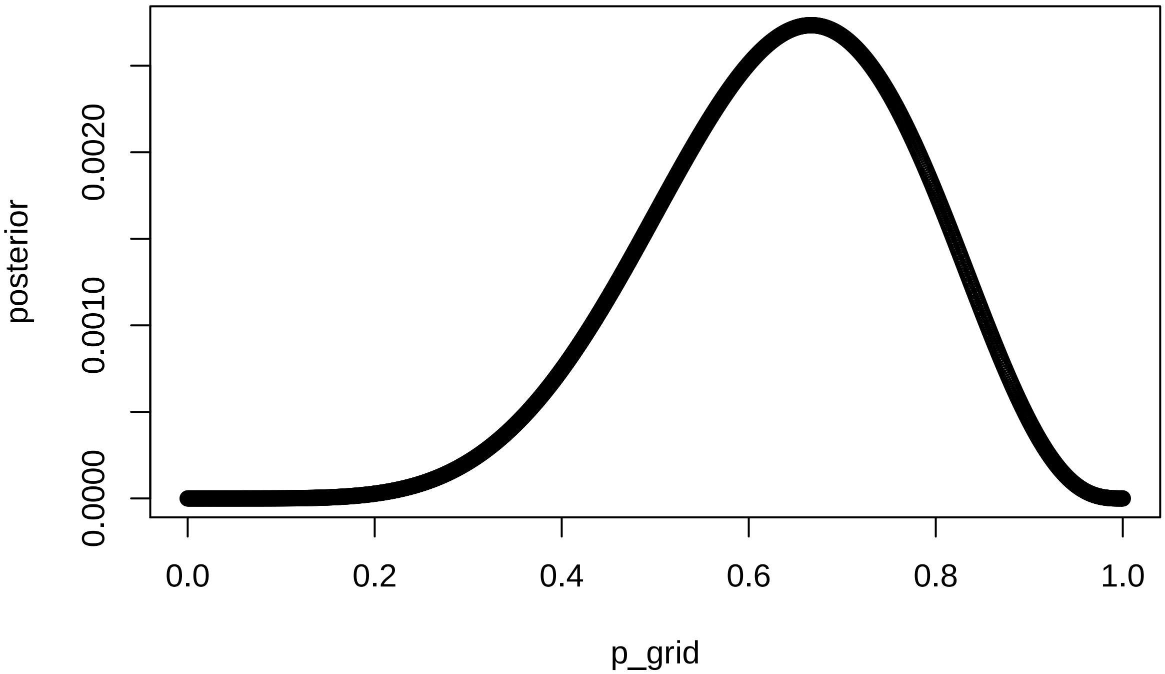
Using a simple grid approximation to calculate the posterior then sample the posterior

```
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)

samples <- sample(p_grid, prob = posterior, size = 10000, replace = TRUE)
```

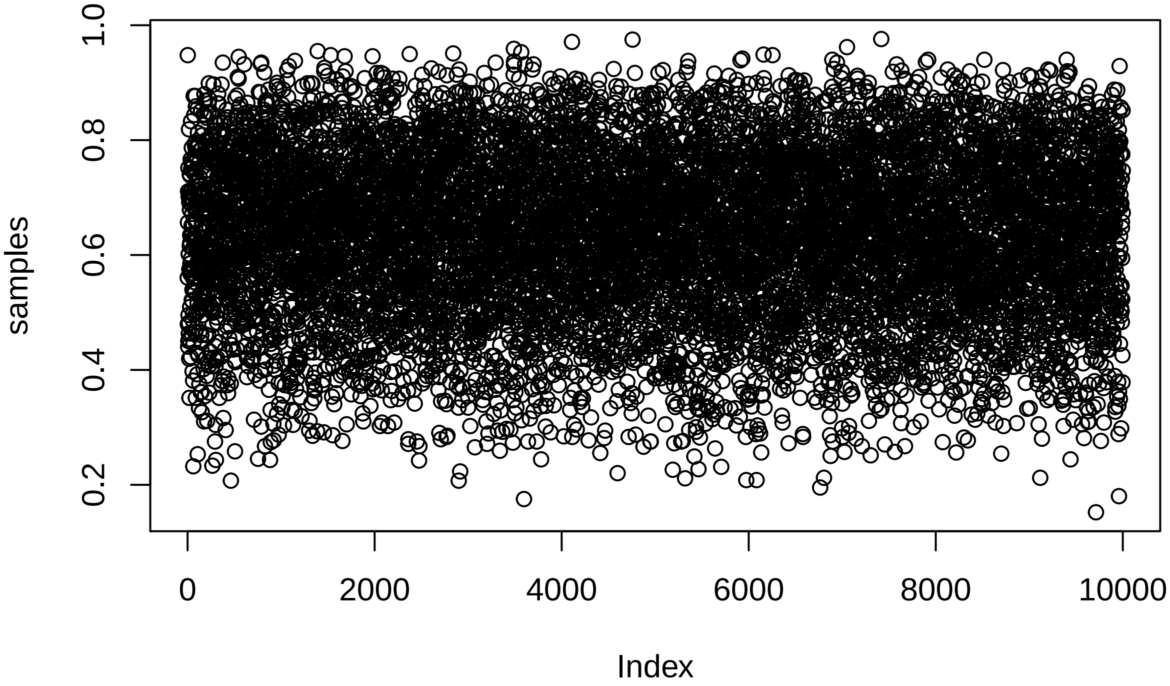
```
library(rethinking)
```

```
## Loading required package: rstan

## Loading required package: ggplot2

## Loading required package: StanHeaders

## Warning: package 'StanHeaders' was built under R version 3.5.2

## rstan (Version 2.18.2, GitRev: 2e1f913d3ca3)

## For execution on a local, multicore CPU with excess RAM we recommend calling
## options(mc.cores = parallel::detectCores()).
## To avoid recompilation of unchanged Stan programs, we recommend calling
## rstan_options(auto_write = TRUE)

## Loading required package: parallel

## rethinking (Version 1.59)
```

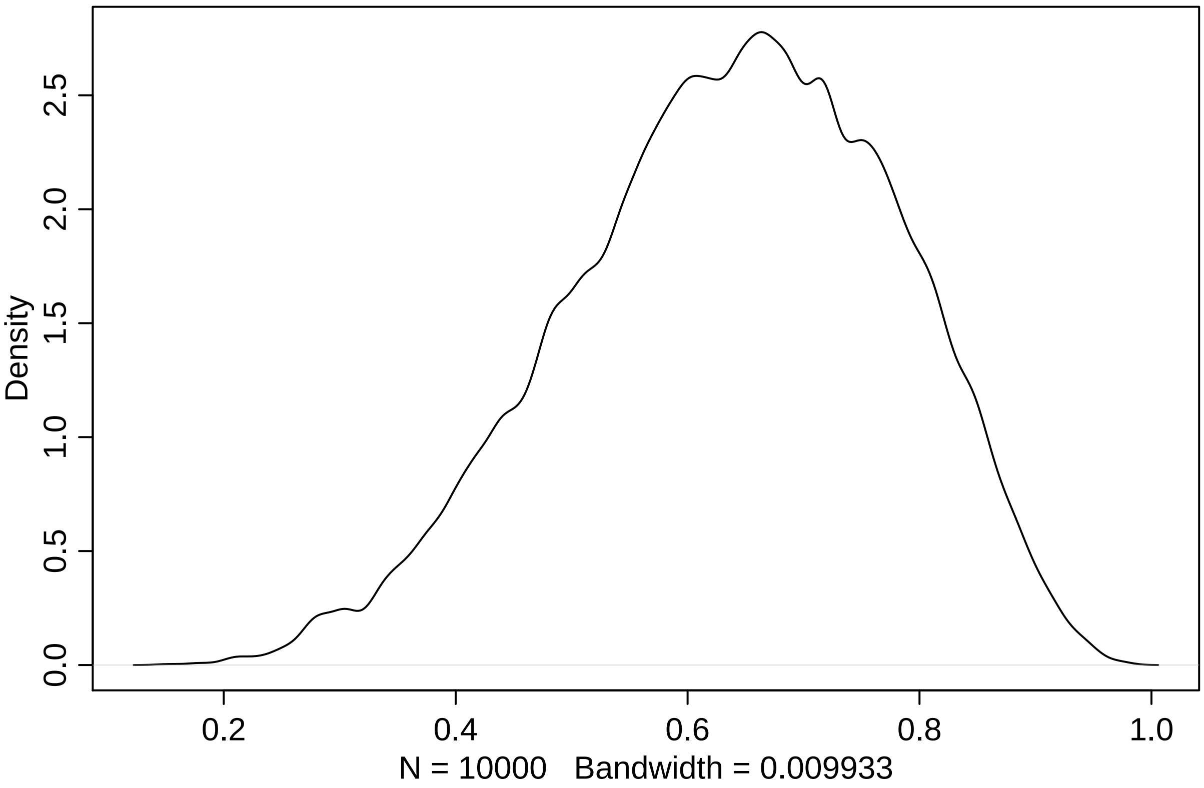
```
plot(p_grid,posterior)
```



```
plot(samples)
```



```
dens(samples)
```



Getting summary statistics from the samples

```
#less than 50% water
sum(samples < 0.5)/ 1e4
```

```
## [1] 0.174
```

```
#between 50 and 75 percent water
sum( samples > 0.5 & samples < 0.75)/ 1e4
```

```
## [1] 0.5986
```

```
# 80 percent of the calculated  percentages are below what value?
quantile( samples, c(0.8))
```

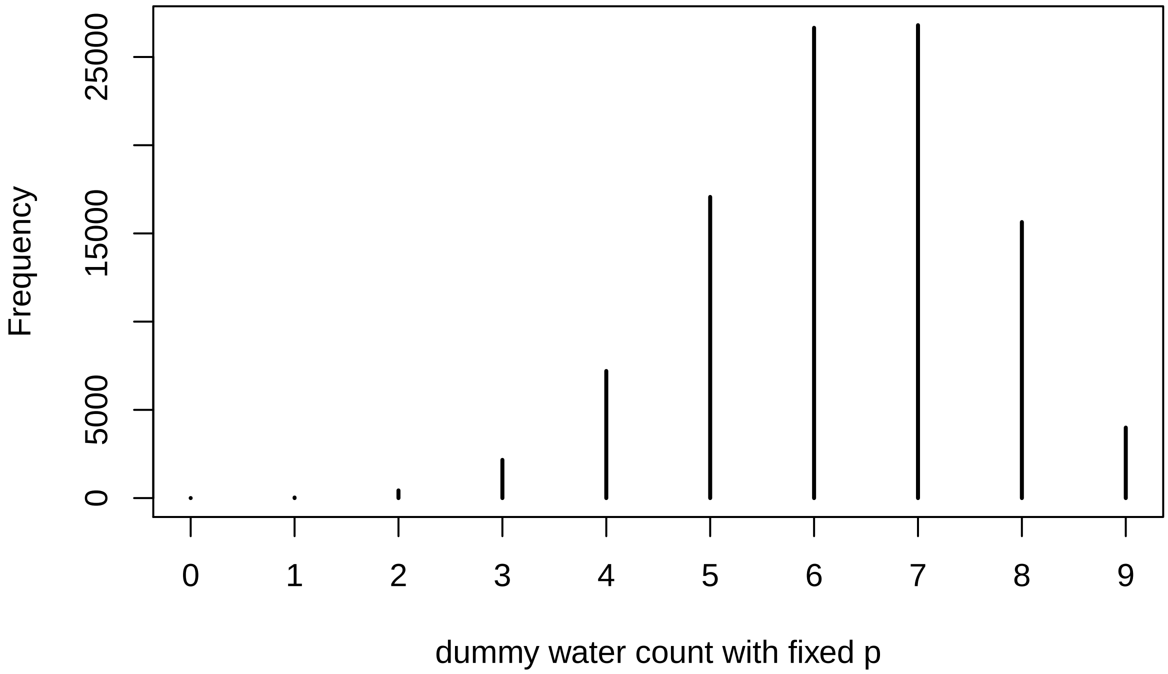
```
##       80%
## 0.7627628
```

```
# extend to 90 and 95 percent also
quantile( samples, c(0.8,0.9,0.95))
```

```
##       80%       90%       95%
## 0.7627628 0.8129129 0.8498498
```

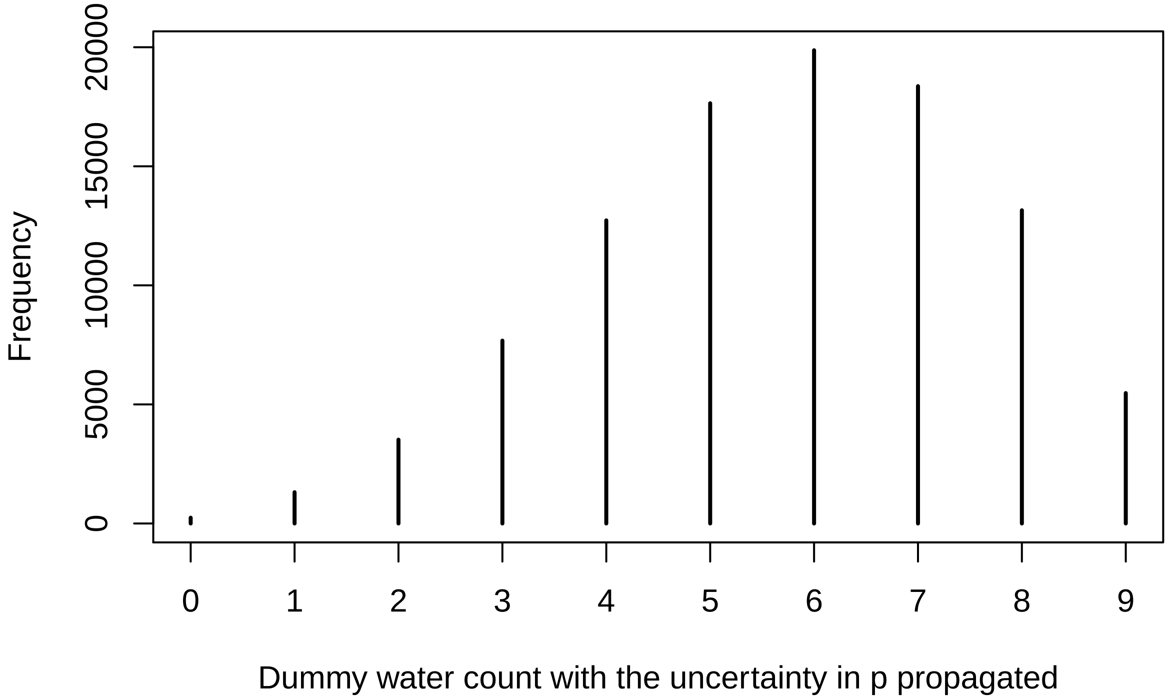
Dummy samples given a fixed known probability

```
dummy_w <- rbinom(1e5, size=9, prob=0.7)
simplehist( dummy_w, xlab="dummy water count with fixed p")
```



Dummy samples given the posterior distribution of the probability.

```
w <- rbinom( 1e5, size=9, prob=samples)
simplehist( w, xlab="Dummy water count with the uncertainty in p propagated")
```



Easy questions

```r
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=1e4, replace=TRUE)

#question 1
```

Question 1

How much posterior probability lies below $p = 0.2$*?*

```r
print("Direct from posterior")
```

```
## [1] "Direct from posterior"
```

```r
sum(posterior[p_grid < 0.2])
```

```
## [1] 0.0008560951
```

```r
print("From resample")
```

```
## [1] "From resample"
```

```r
sum(samples < 0.2)/length(samples)
```

```
## [1] 5e-04
```

Question 2

How much posterior probability lies above $p = 0.8$*?*

```r
print("Direct from posterior")
```

```
## [1] "Direct from posterior"
```

```r
sum(posterior[p_grid > 0.8])
```

```
## [1] 0.1203449
```

```r
print("From resample")
```

```
## [1] "From resample"
```

```r
sum(samples > 0.8)/length(samples)
```

```
## [1] 0.1117
```

Question 3

How much posterior probability lies between $p = 0.2$ *and* $p = 0.8$*?*

This could be calculated using the results from question 1 and 2

```r
print("Direct from posterior")
```

```
## [1] "Direct from posterior"
```

```r
sum(posterior[p_grid < 0.8 & p_grid > 0.2])
```

```
## [1] 0.878799
```

```r
print("From resample")
```

```
## [1] "From resample"
```

```r
sum(samples < 0.8 & samples > 0.2)/length(samples)
```

```
## [1] 0.8878
```

Question 4

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

```r
quantile(samples, probs = 0.2)
```

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

Question 5

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

```r
quantile(samples, probs = 0.8)
```

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

Question 6

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

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

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

Question 7

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

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

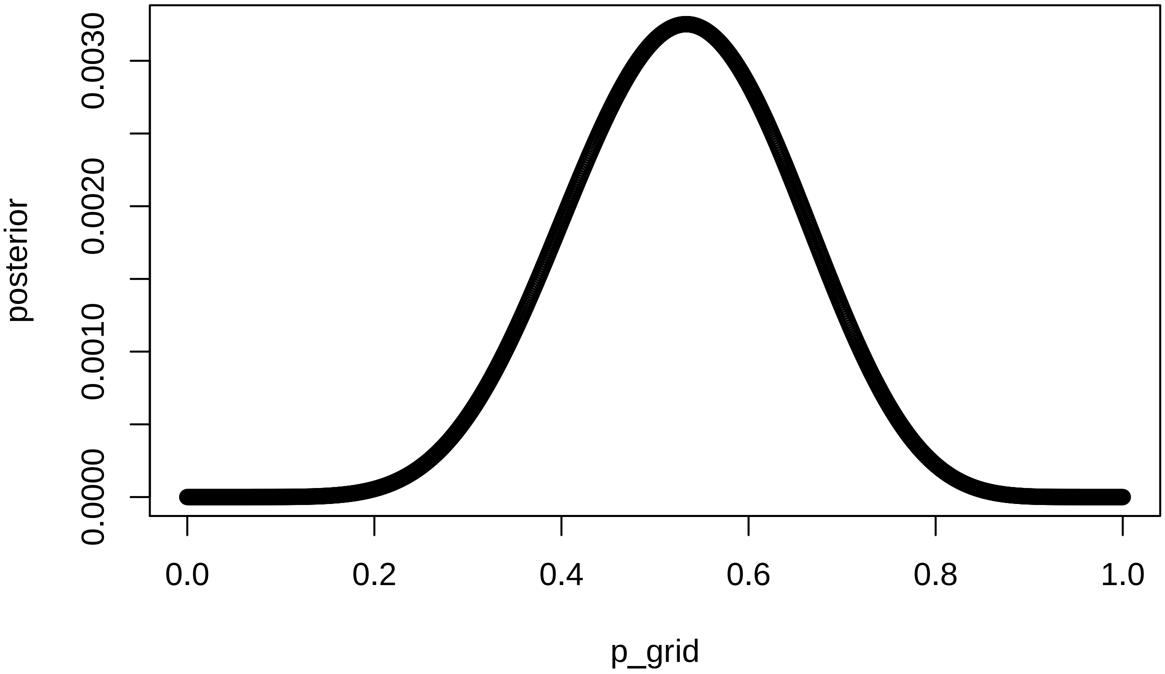
```
##       17%       83% 
## 0.5005005 0.7687688
```

Medium questions

Question 1

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.

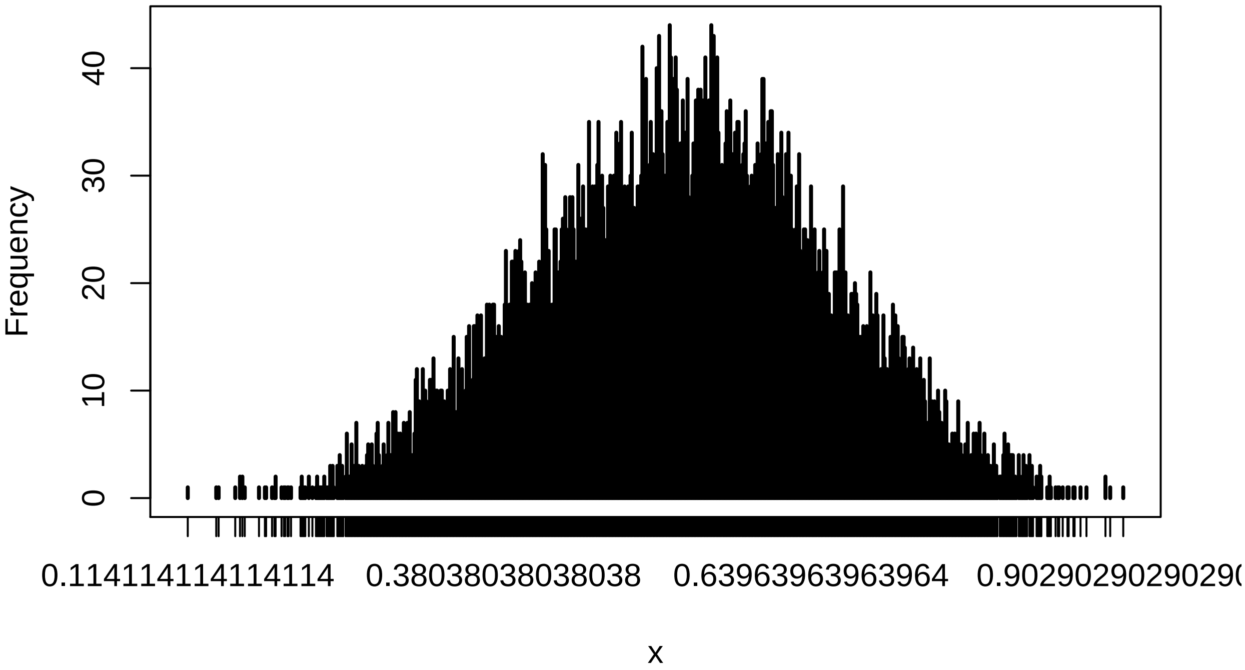
```
p_grid = seq(0,1, length.out = 1e3)
prior = rep(1, length(p_grid))

likelihood = dbinom(8, size=15, p_grid)
posterior = likelihood * prior # not needed if the prior is flat
posterior = posterior/sum(posterior)
plot(p_grid,posterior)
```



Question 2

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 = 1e4 , replace = TRUE, prob = posterior)
simplehist(samples, round=FALSE)
```



```r
HPDI(samples, prob = 0.9)
```

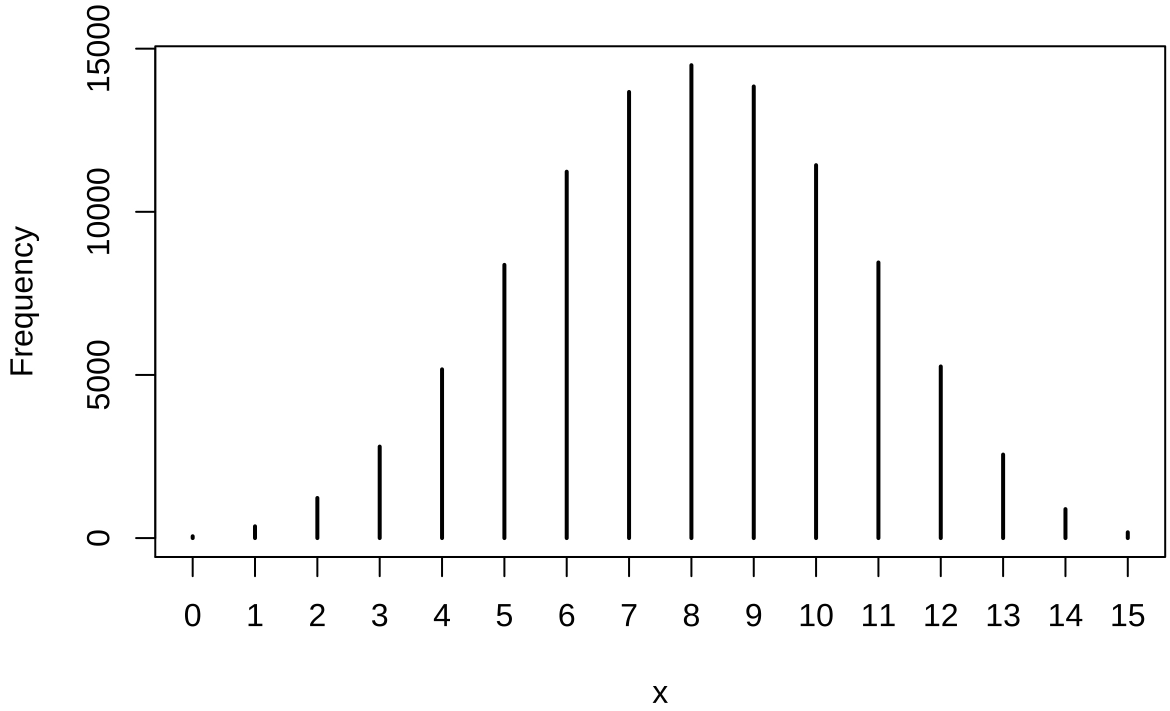
```
##      |0.9      0.9|
## 0.3383383 0.7317317
```

Question 3

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?

Rememeber to run with more one or more repetitions for each sample probability

```r
dummy_w = rbinom(n=1e5, size = 15, prob = samples)
simplehist(dummy_w, round = FALSE)
```



```
sum(dummy_w==8)/length(dummy_w)
```

```
## [1] 0.14494
```

Question 4

Using the posterior distribution constructed from the new (8/15) data, now calculate the probability of observing 6 water in 9 tosses.

This can be done by another round of sampling

```
samples2 <- rbinom(1e5, 9, samples)
sum(samples2==6)/length(samples2)
```

```
## [1] 0.17613
```

But the question asks us to use the posterir distribution

```
prob_dist <-  dbinom(6, size=9, p_grid)
prob_dist <- prob_dist * posterior
sum(prob_dist)
```

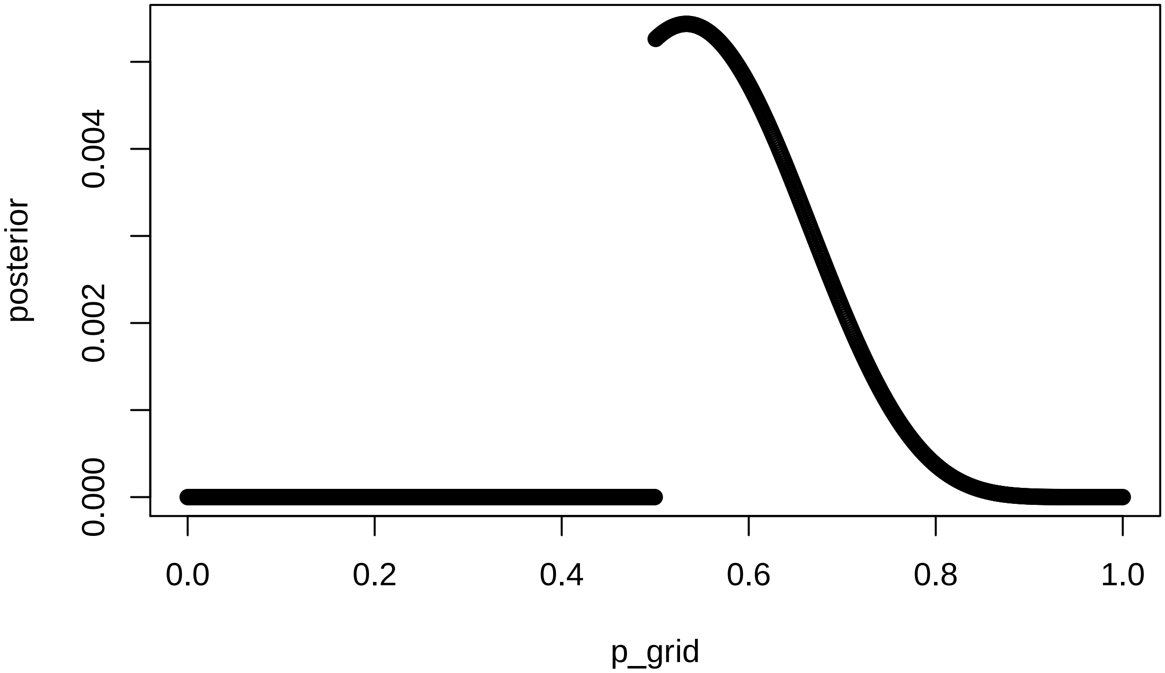
```
## [1] 0.1763898
```

Question 5

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.

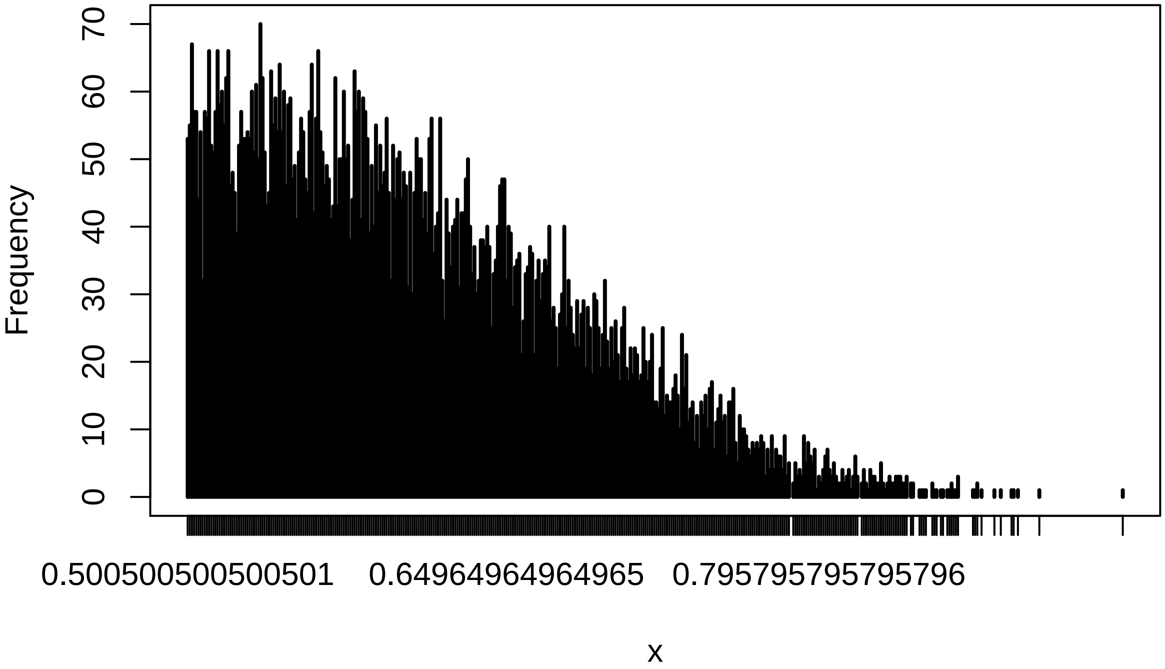
```
p_grid = seq(0,1, length.out = 1e3)
prior = ifelse(p_grid < 0.5 , 0, 1)
```

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



The posterior has zero probability because the prior put that probability to zero.

```
samples <- sample(p_grid, size = 1e4 , replace = TRUE, prob = posterior)
simplehist(samples, round=FALSE)
```

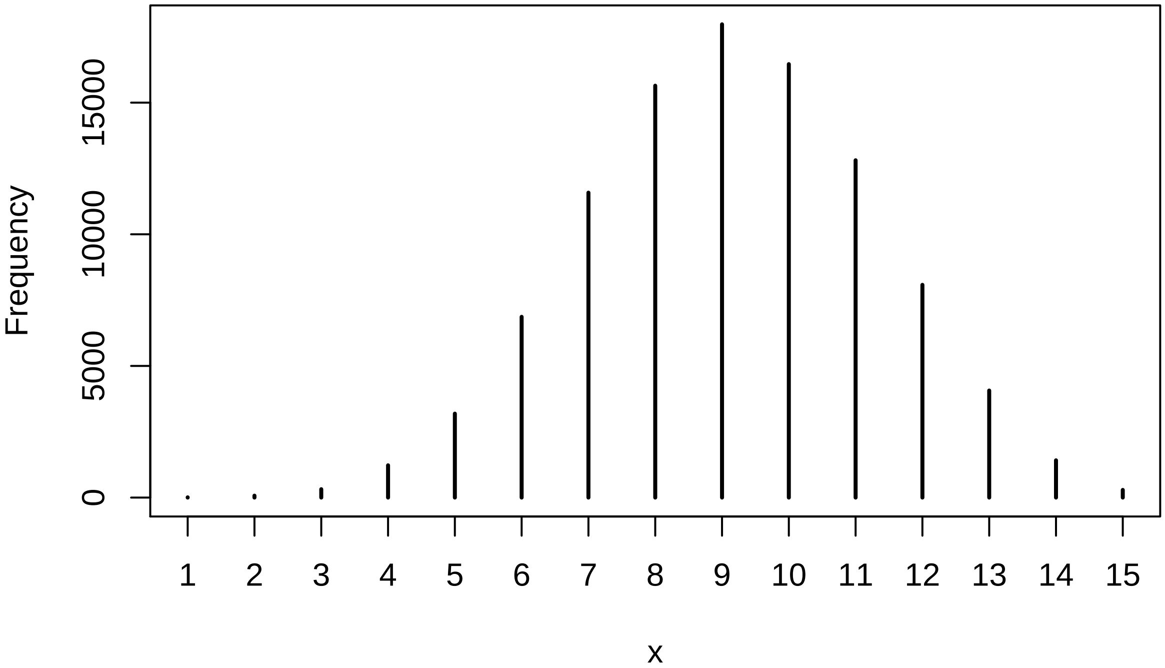


```
HPDI(samples, prob = 0.9)
```

```
##       |0.9      0.9|
## 0.5005005 0.7147147
```

This could seem strange. The new prior gives a smaller HDPI as expected. However, its max value is decresed which could seems strange. To see why it is important to realize that the amount of probability that are not in the intervall is the same in both cases. Since there is very little probability thats below the intervall more of the probability above must be outside the interval.

```
dummy_w = rbinom(n=1e5, size = 15, prob = samples)
simplehist(dummy_w, round = FALSE)
```



```
sum(dummy_w==8)/length(dummy_w)
```

```
## [1] 0.15643
```

As expected the probability of seeing values which are just above 50% should not be changed much by moving probability from far below 50% to far above 50%. When thte prior was flat, 8 was the most probable outcome. With the new prior 9 is most porbable outcome but the difference is small. The big difference is the probability of seeing 1 2 and 3. If the earth is mostly water and you sample 15 times to get an estimate of the amount, a p value above 50% is going to make that improbable.

Hard questions

Introduction. The practice problems here all use the data below. These data indicate the gender (male=1, female=0) of officially reported first and second born children in 100 two-child families.

```
# birth1 <- c(1,0,0,0,1,1,0,1,0,1,0,0,1,1,0,1,1,0,0,0,1,0,0,0,1,0,
#             0,0,0,1,1,1,0,1,0,1,1,1,0,1,0,1,1,0,1,0,0,1,1,0,1,0,0,0,0,0,0,0,
#             1,1,0,1,0,0,1,0,0,0,1,0,0,1,1,1,1,0,1,0,1,1,1,1,1,0,0,1,0,1,1,0,
#             1,0,1,1,1,0,1,1,1,1)
# birth2 <- c(0,1,0,1,0,1,1,1,0,0,1,1,1,1,1,0,0,1,1,1,0,0,1,1,1,0,
#             1,1,1,0,1,1,1,0,1,0,0,1,1,1,1,0,0,1,0,1,1,1,1,1,1,1,1,1,1,1,1,1,
#             1,1,1,0,1,1,0,1,1,0,1,1,1,0,0,0,0,0,0,1,0,0,0,1,1,0,0,1,0,0,1,1,
#              0,0,0,1,1,1,0,0,0,0)
```

So for example, the first family in the data reported a boy (1) and then a girl (0). The second family reported a girl (0) and then a boy (1). The third family reported two girls. You can load these two vectors into R's memory by typing:

```
library(rethinking)
data(homeworkch3)
```

So for example to compute the total number of boys born across all of these births, you could use:

```
 sum(birth1) + sum(birth2)
```

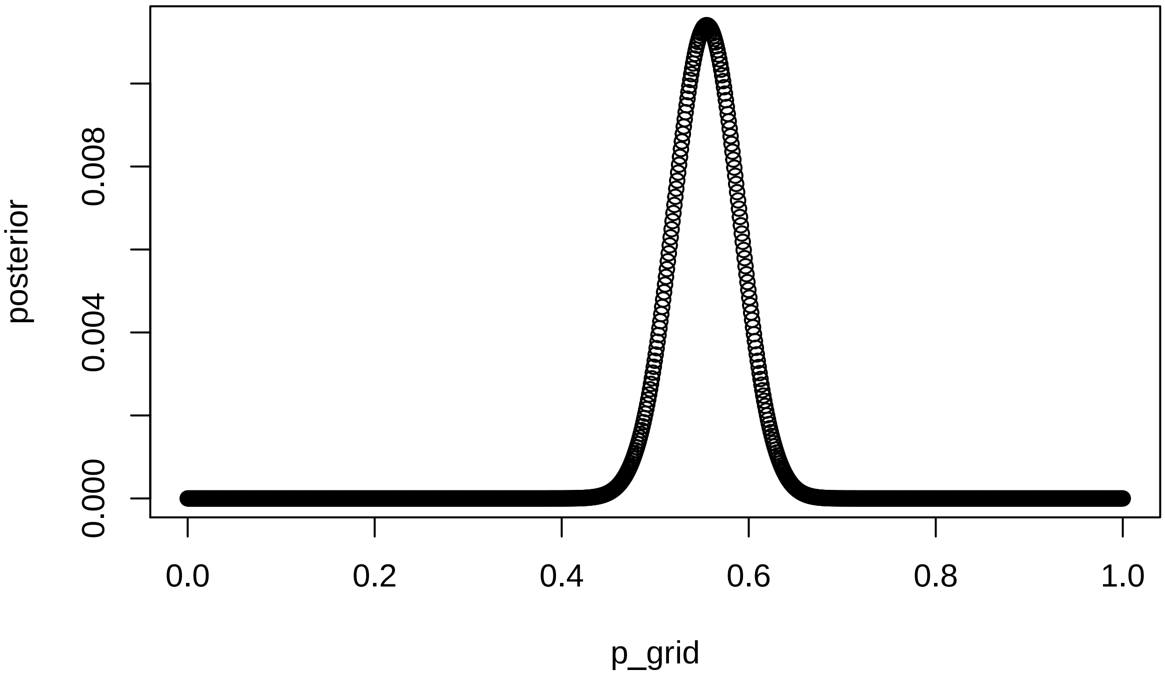
```
## [1] 111
```

Question 1

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?

Assume the sex is random with equal probability and that sex of the two children are indepedent. This is not true but we know its a good approximation.

```
p_grid = seq(0,1,length.out = 1e3)

prior = rep(1,length(p_grid))
prior = prior/sum(prior)

likelihood = dbinom(sum(birth1) + sum(birth2), size =  length(birth1) + length(birth2), prob = p_grid)

posterior = likelihood*prior
posterior = posterior/sum(posterior)

plot(p_grid,posterior)
```



```
p_grid[ which.max(posterior)]
```

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

Question 2

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, size = 1e4,  prob = posterior, replace = TRUE)
HPDI(samples,0.5)
```

```
##      |0.5      0.5|
## 0.5255255 0.5725726
```

```
HPDI(samples,0.89)
```

```
##     |0.89     0.89|
## 0.5005005 0.6096096
```

```
HPDI(samples,0.97)
```

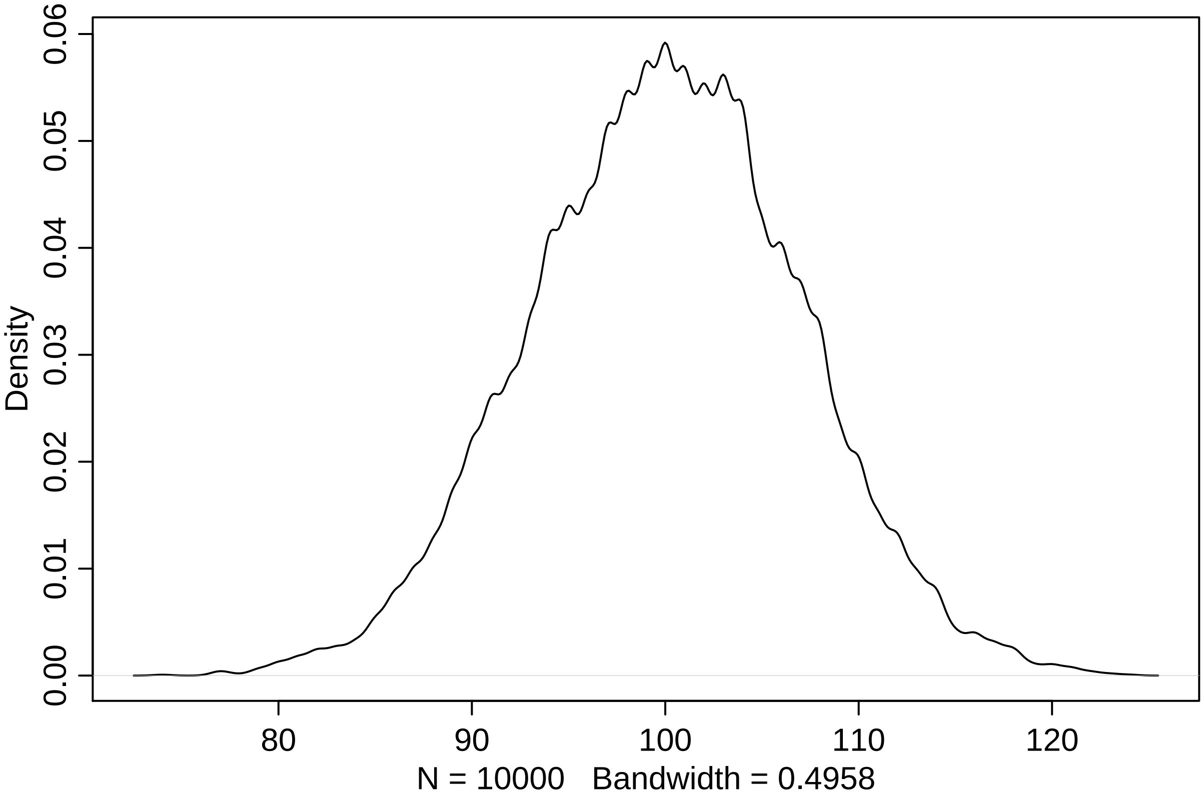
```
##     |0.97     0.97|
## 0.4774775 0.6266266
```

It is clear that 50% is not an unreasonalbe parameter

Question 3

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?*

```
simulated = rbinom(1e4, size = 200, prob = 0.5)
dens(simulated)
```



Yes, but 111 sum(birth1) + sum(birth2) is a bit high. Between 90 and a 110 would make me more comfertable saying the probaility of a boy is 0.5.

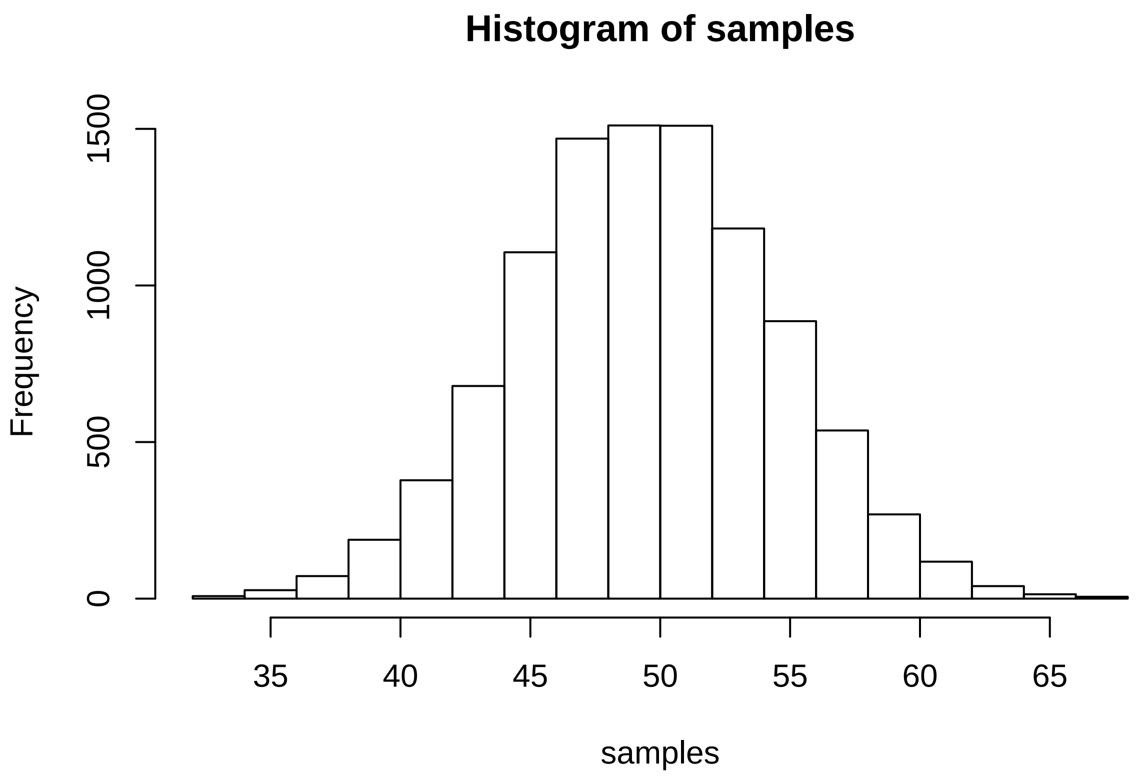
Question 4

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?

```
sum(birth1)
```

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

```
samples = rbinom(1e4, size = 100, prob = 0.5)
hist(samples)
```



This is a much more credible which leads to concern for child. There has to be 60 boys amongs the 100 which are child number 2. That is a high number when looking at model.

Question 5

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?

```
secondChild_firstGirl = birth2[birth1==0]
length(secondChild_firstGirl)
```

```
## [1] 49
```

```
sum(secondChild_firstGirl)
```

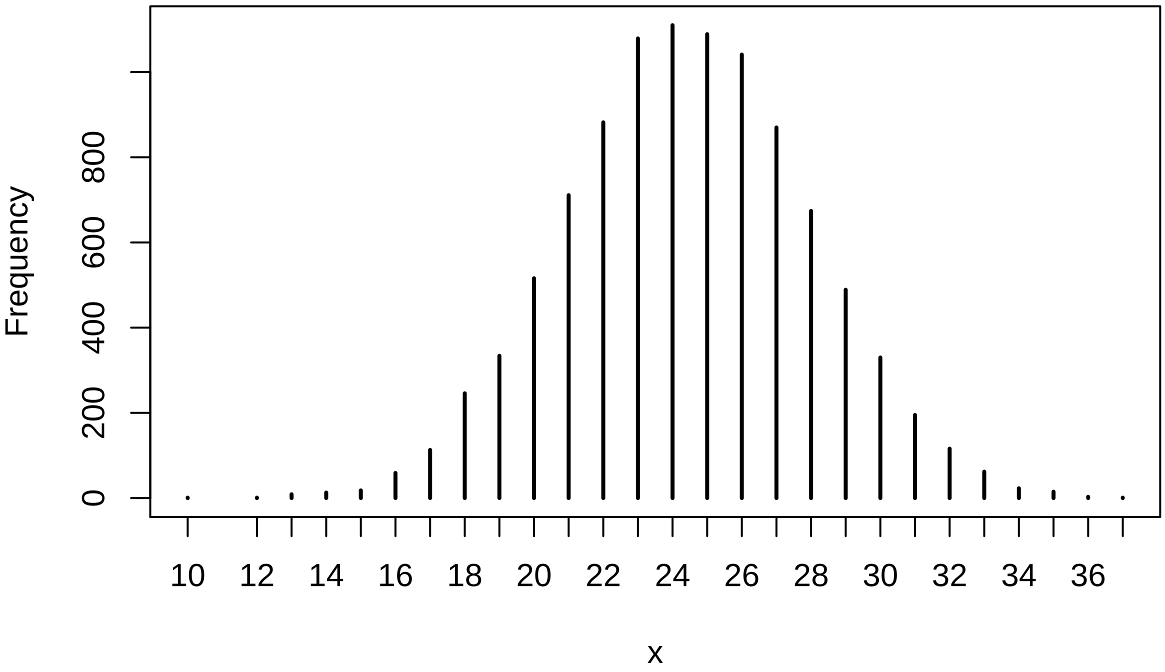
```
## [1] 39
```

```
sum(secondChild_firstGirl)/length(secondChild_firstGirl)
```

```
## [1] 0.7959184
```

80% boys for a group that is close to 50 children. There is no way that is a coinsidence.

```
happySample = rbinom(1e4, size = length(secondChild_firstGirl), prob = 0.5)
simplehist(happySample)
```



With 10000 simulations the highest observed value is 37. 39 is two larger than the higest observed simulated value amongst 10000 simulations. It is clear that parents which have a girl as their first child are likely to chose abortion if they learn that their second child is also a girl.