Homework 1

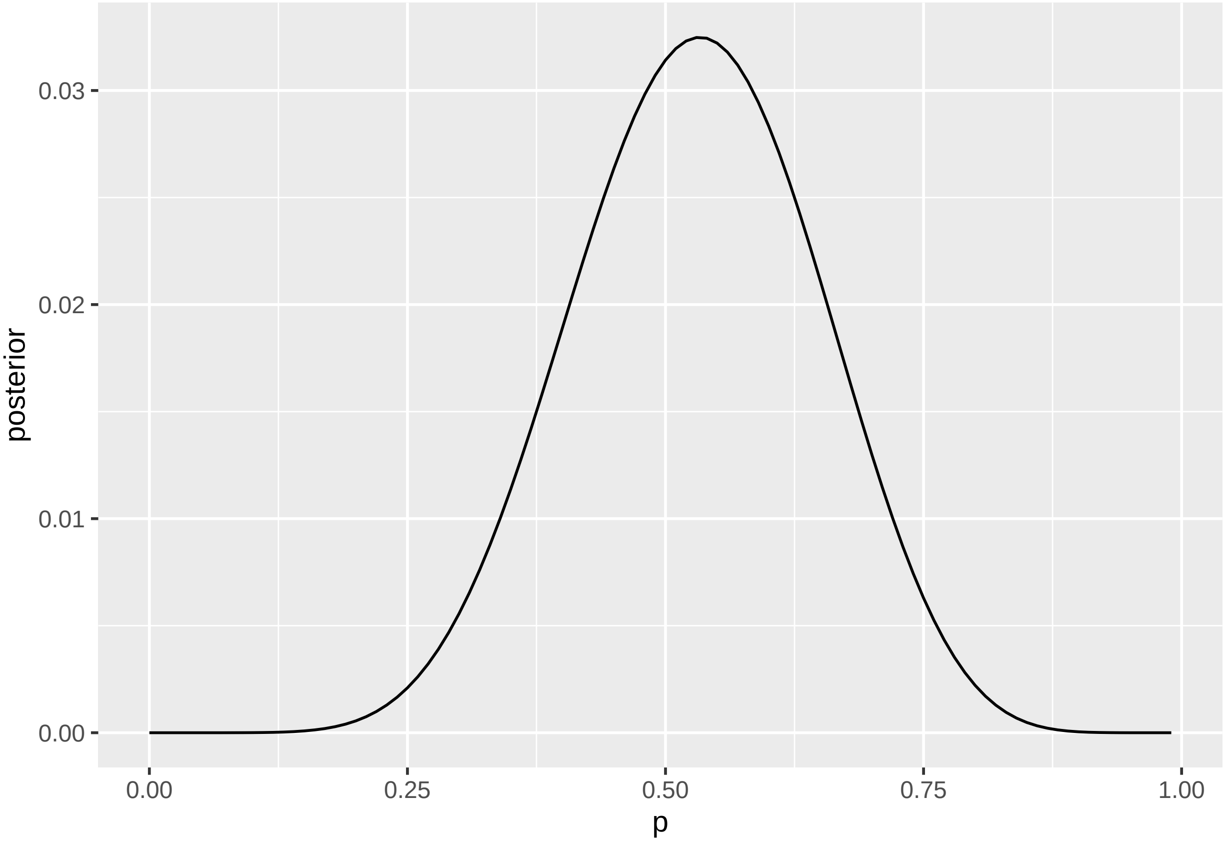
Used libraries are:

- rethinking
- ggplot2

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, 0.99, 1/100)
prior_of_p <- rep(1/100, 100)

likelihood <- dbinom(8, 15, p_grid) * prior_of_p
posterior <- likelihood / sum(likelihood)

df <- data.frame(p=p_grid, posterior=posterior)

ggplot(df, aes(x=p, y=posterior)) + geom_line()
```



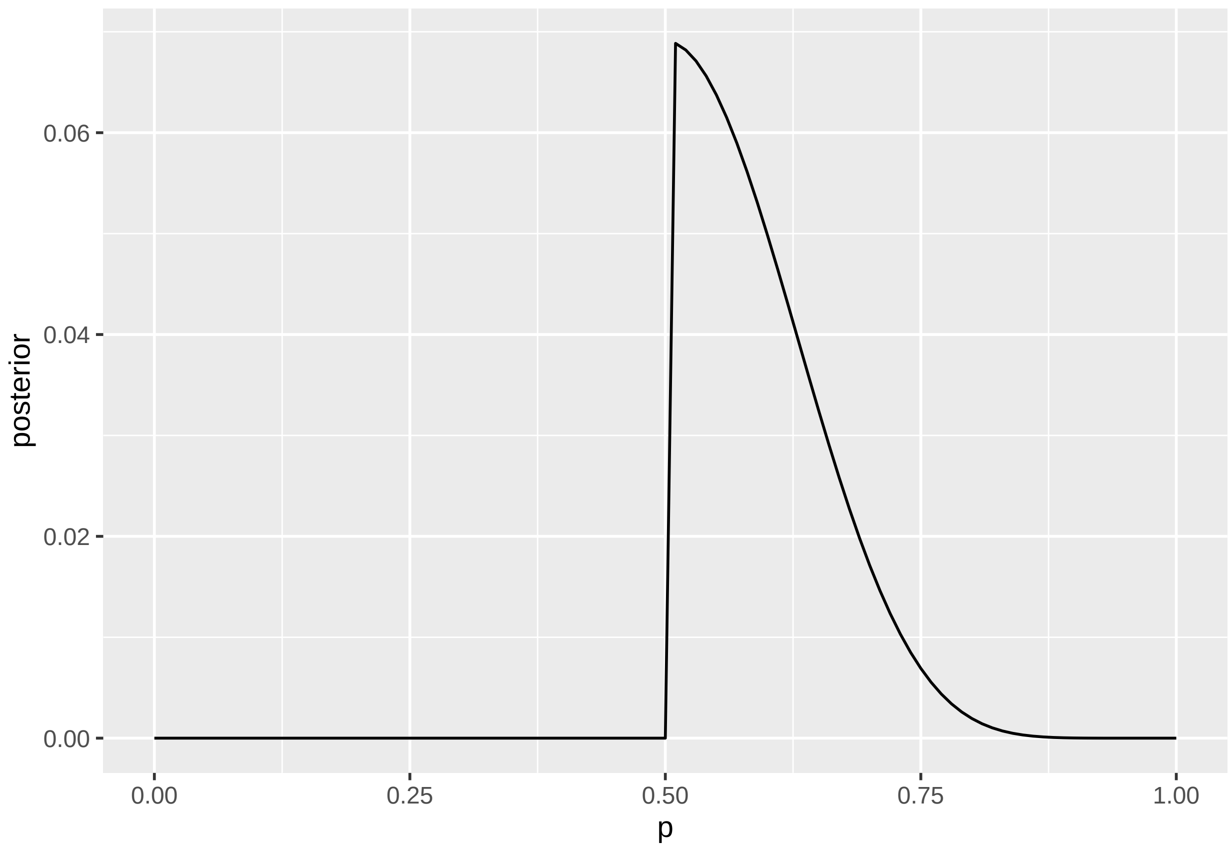
```
samples <- sample(p_grid, prob=posterior, size=1e4, replace=TRUE)

pi <- PI(samples, 0.98)
posterior_mean <- mean(samples)
```

Posterior mean is around 0.528746 and 99 interval is between 0.26 and 0.78

Question 2

Start over in 1, 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. What difference does the better prior make? If it helps, compare posterior distributions (using both priors) to the true value p = 0.7.

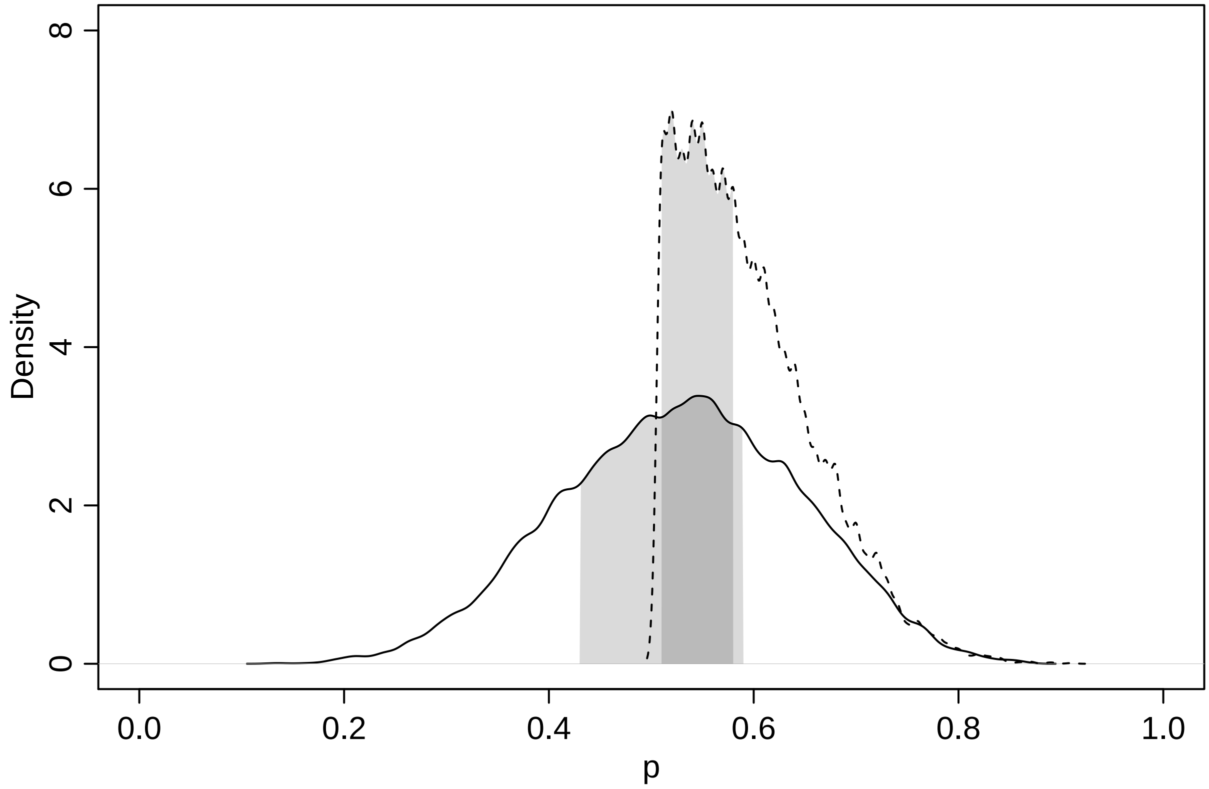
```
p_grid <- seq(0, 1, 1/100)
prior_of_p_blow_05 <- rep(0, 51)
prior_of_p_above_05 <- rep(1/50, 50)

prior_of_p <- c(prior_of_p_blow_05, prior_of_p_above_05)
names(prior_of_p) <- p_grid


likelihood <- dbinom(8, 16, p_grid) * prior_of_p
posterior <- likelihood / sum(likelihood)

df <- data.frame(p=p_grid, posterior=posterior)

ggplot(df, aes(x=p, y=posterior)) + geom_line()
```



```
samples_2 <- sample(p_grid, prob=posterior, size=1e4, replace=TRUE)

pi_2 <- PI(samples_2, 0.98)
posterior_mean_2 <- mean(samples_2)
```

Posterior mean is around 0.59511 and 99 interval is between 0.51 and 0.78

```
dens( samples , xlab="p" , xlim=c(0,1) , ylim=c(0,8) , show.HPDI=0.5)
dens( samples_2 , add=TRUE , lty=2 , show.HPDI=0.5)
```

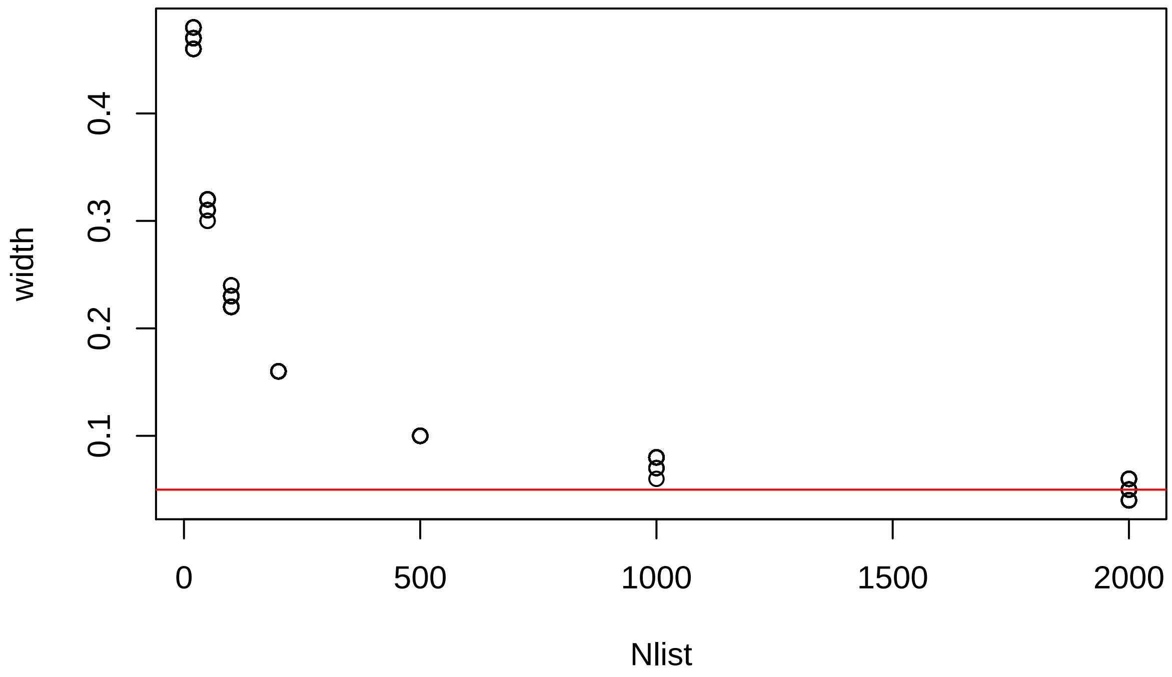


Question 3

This problem is more open-ended than the others. Feel free to collaborate on the solution. 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? I won't require a precise answer. I'm honestly more interested in your approach.

```r
f <- function(trial){

  p_grid <- seq(0, 1, 1/100)
  prior_of_p <- rep(1/101, 101)

  names(prior_of_p) <- p_grid


  likelihood <- dbinom(trial/2, trial, p_grid) * prior_of_p
  posterior <- likelihood / sum(likelihood)

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

  interval <- PI(samples, 0.98)

  return(as.numeric(interval['99%'] - interval['1%']))
}
```

```r
Nlist <- c( 20 , 50 , 100 , 200 , 500 , 1000 , 2000 )
Nlist <- rep( Nlist , each=100 )

width <- sapply( Nlist , f )

plot( Nlist , width )
abline( h=0.05 , col="red" )
```



```r
#ggplot(df, aes(x=trial, y=percentile_interval)) + geom_line()
```

```r
df <- data.frame(width=tapply(width, Nlist, FUN=mean))

knitr::kable(df, floating.environment="sidewaystable")
```

	width
20	0.470504
50	0.317603
100	0.224706
200	0.160000
500	0.100000
1000	0.079300
2000	0.051112