Homework 1

Sam Bashevkin

2022-05-13

```
library(rethinking)
```

1. Suppose the globe tossing data (Chapter 2) had turned out to be 4 water and 11 land. Construct the posterior distribution, using grid approximation. Use the same flat prior as in the book.

Compute the posterior

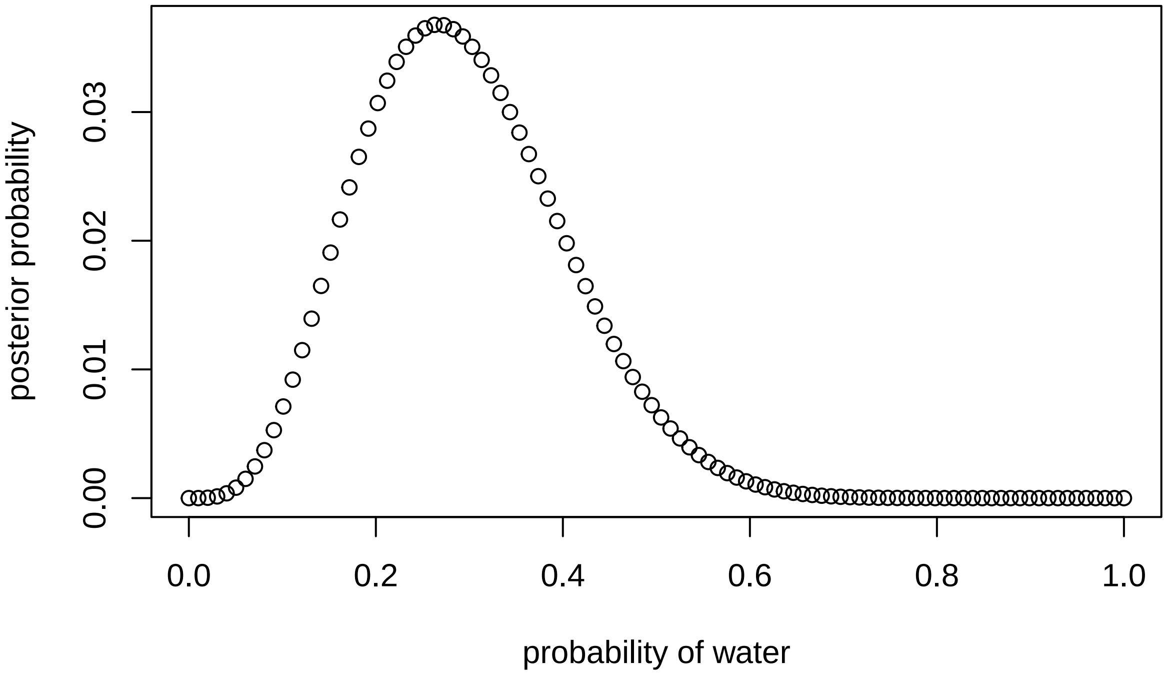
```
# define data
w<-4
l<-11

grid_post<-function(w, l, prior, grid_length=100){
    # define grid
    p_grid <- seq( from=0 , to=1 , length.out=grid_length )
    # define prior
    if(prior=="uniform"){
        prior<-rep(1, length(p_grid))
    }else{
        if(prior=="cutoff"){
            prior<-ifelse(p_grid<0.5, 0, 1)
        }else{
            stop("Prior must be either 'uniform' or 'cutoff'")
        }
    }
    # compute likelihood at each value in grid
    likelihood <- dbinom( w , size=w+l , prob=p_grid )
    # compute product of likelihood and prior
    unstd.posterior <- likelihood * prior
    # standardize the posterior, so it sums to 1
    posterior <- unstd.posterior / sum(unstd.posterior)

    out<-data.frame(p_grid=p_grid, posterior=posterior, prior=prior)
    return(out)
}
posterior<-grid_post(w, l, prior="uniform")
```

Plot the posterior

```
plot( posterior$p_grid , posterior$posterior , type="b" ,
      xlab="probability of water" , ylab="posterior probability" )
```



2. Now suppose the data are 4 water and 2 land. Compute the posterior again, but this time 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.

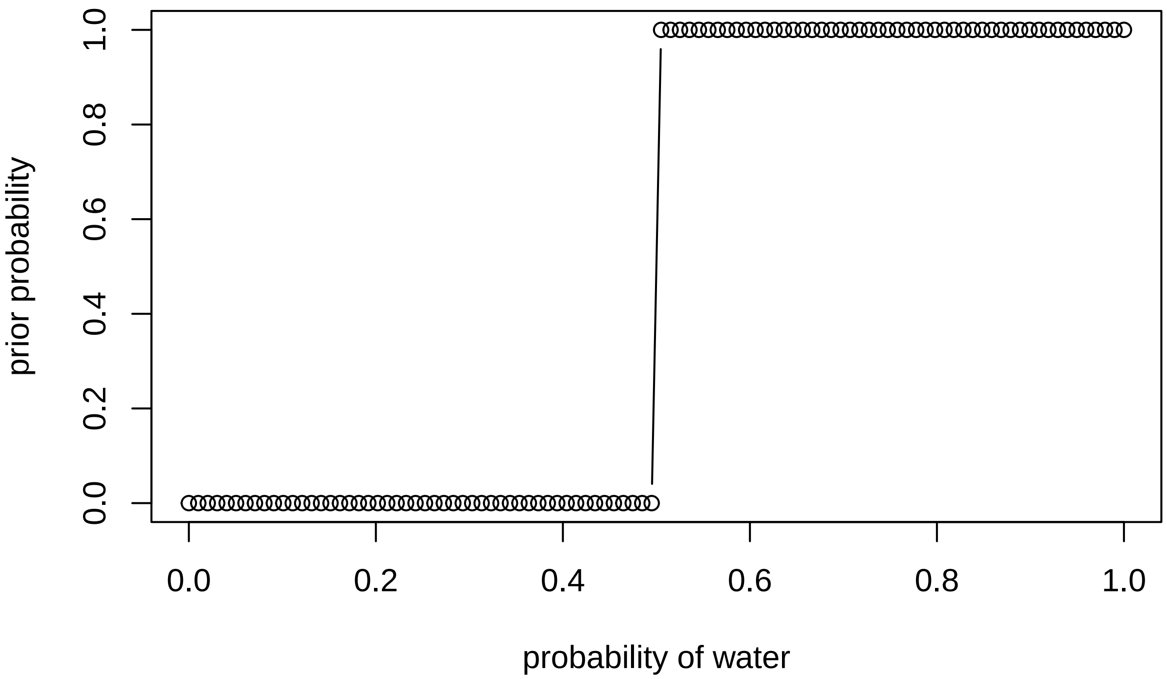
Compute the posterior

```
# define data
w<-4
l<-2

posterior<-grid_post(w, l, prior="cutoff")
```

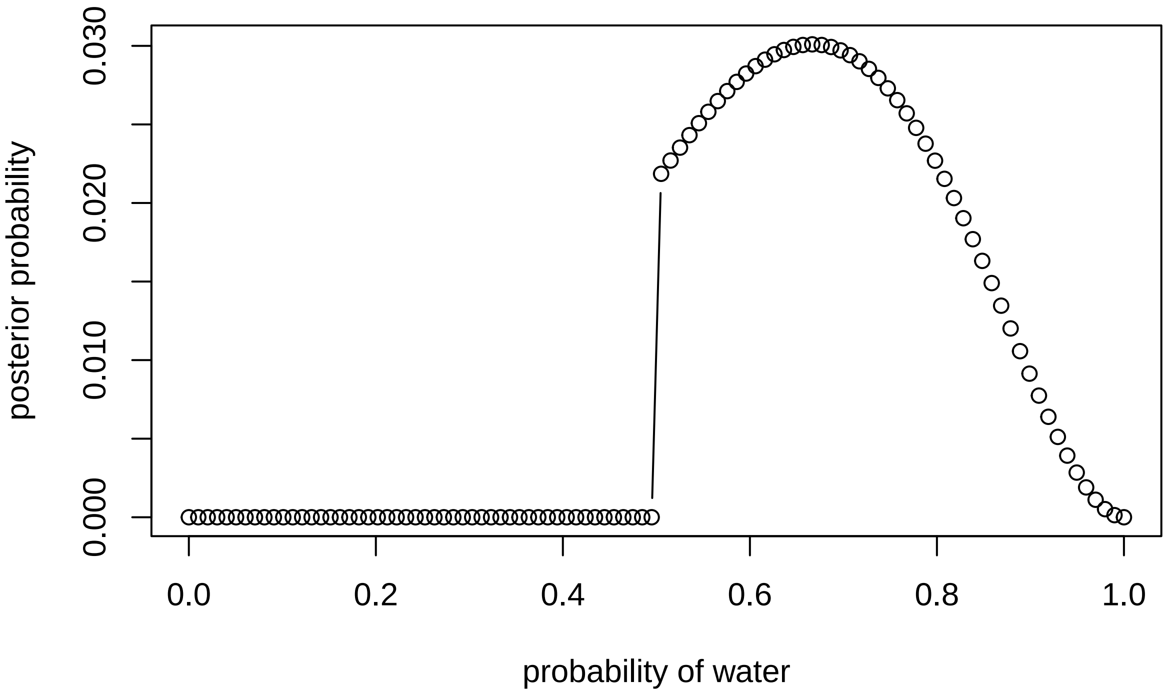
plot prior

```
plot( posterior$p_grid , posterior$prior , type="b" ,
      xlab="probability of water" , ylab="prior probability" )
```



Plot the posterior

```
plot( posterior$p_grid , posterior$posterior , type="b" ,
      xlab="probability of water" , ylab="posterior probability" )
```



Interestingly, because it is standardized, the choice of the specific number for your prior (for p>0.5) doesn't matter

3. For the posterior distribution from 2, compute 89% percentile and HPDI intervals. Compare the widths of these intervals. Which is wider? Why? If you had only the information in the interval, what might you misunderstand about the shape of the posterior distribution?

```r
# First sample from the posterior
samples <- sample( posterior$p_grid , size=1e4 , replace=TRUE , prob=posterior$posterior )
# Percentile
(perc89<-PI(samples, 0.89))
```

```
##        5%       94%
## 0.5252525 0.8787879
```

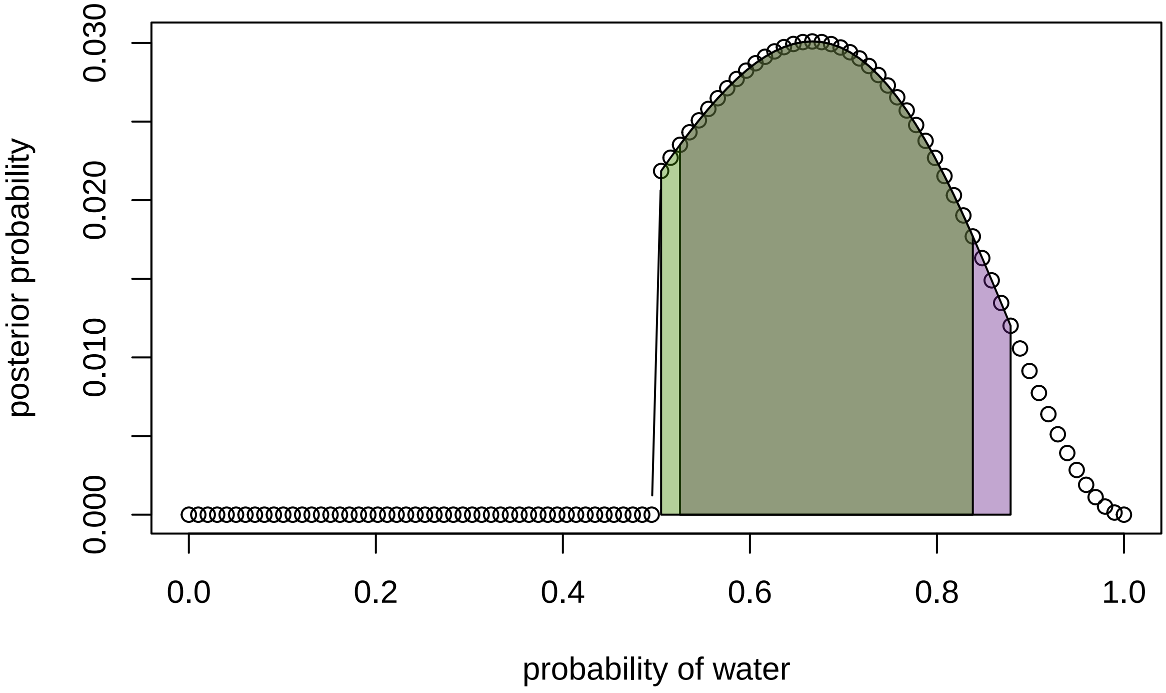
```r
# HPDI
(HPDI89<-HPDI(samples, 0.89))
```

```
##     |0.89     0.89|
## 0.5050505 0.8383838
```

The percentile must be higher because the HPDI is finding the narrowest possible interval.

Plot both of the intervals, with PI in purple and HPDI in green

```
int_shade<-function(int, x, y, col){
    polygon(c(x[x>=int[1] & x<=int[2]], int[2], int[1]),
            c(y[x>=int[1] & x<=int[2]], 0, 0), col=adjustcolor(col, alpha.f = 0.4))
}
plot( posterior$p_grid , posterior$posterior , type="b" ,
      xlab="probability of water" , ylab="posterior probability" )
int_shade(perc89, posterior$p_grid, posterior$posterior, "darkorchid4")
int_shade(HPDI89, posterior$p_grid, posterior$posterior, "chartreuse4")
```



With just the intervals, you would not know that the posterior has 0 probability at values of p<0.5

4. OPTIONAL CHALLENGE. Suppose there is bias in sampling so that Land is more likely than Water to be recorded. Specifically, assume that 1-in-5 (20%) of Water samples are accidentally recorded instead as ”Land”. First, write a generative simulation of this sampling process. Assuming the true proportion of Water is 0.70, what proportion does your simulation tend to produce instead? Second, using a simulated sample of 20 tosses, compute the unbiased posterior distribution of the true proportion of water.

```
n<-20
N<-1e4
true_sample<-rbinom(N, n, 0.7)
biased_sample<-sapply(true_sample, function(x) sum(sample(c(FALSE, TRUE), size=x, prob=c(0.2, 0.8), repl
median(biased_sample/n)
```

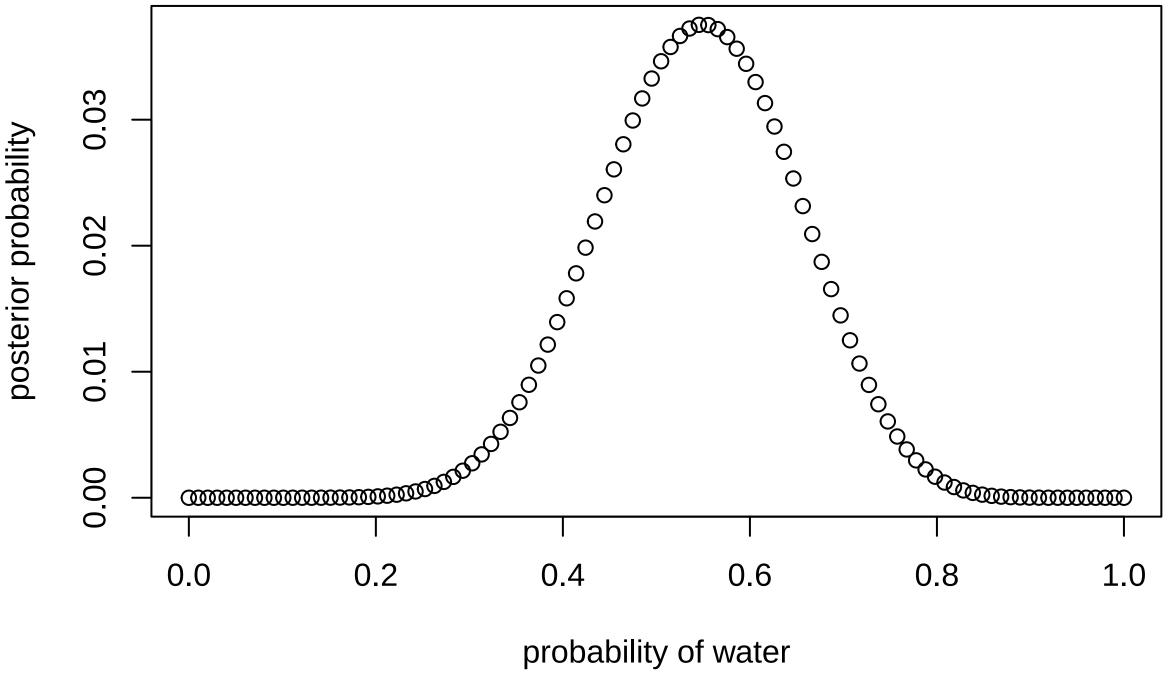
```
## [1] 0.55
```

```
w<-sample(biased_sample, 1)
l<-n-w
posterior_biased<-grid_post(w, l, prior="uniform")
```

It tends to produce a proprtion around 0.55

Plot the posterior

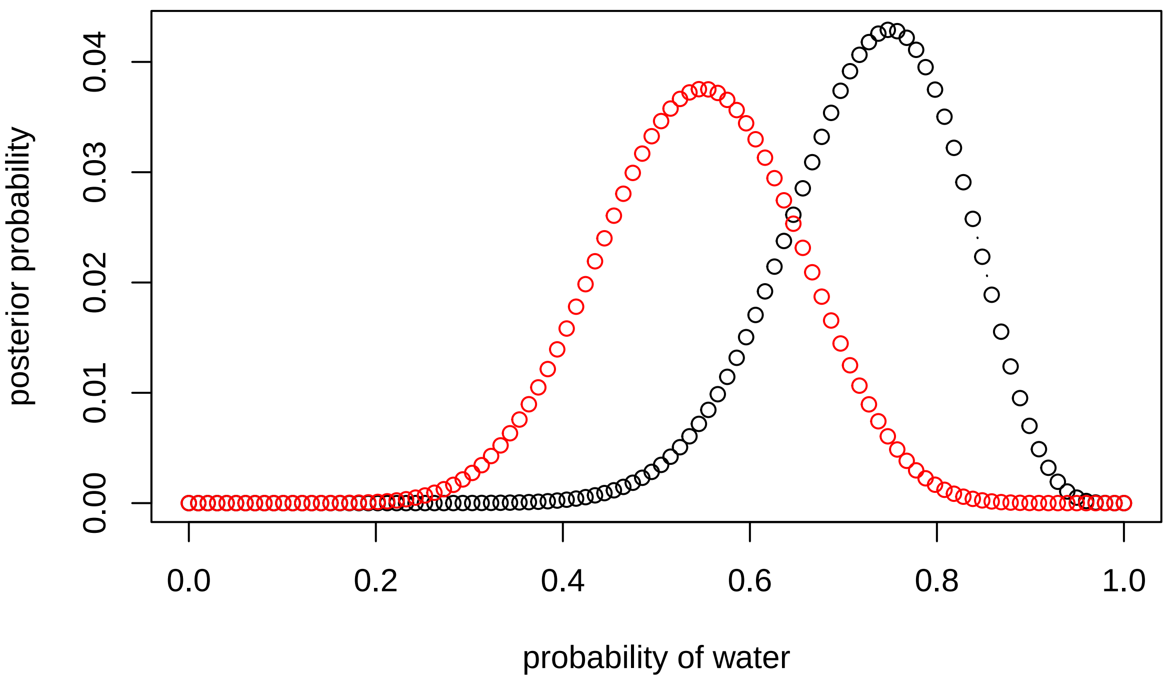
```
plot( posterior_biased$p_grid , posterior_biased$posterior , type="b" ,
      xlab="probability of water" , ylab="posterior probability" )
```



```
# define grid
p_grid <- seq( from=0 , to=1 , length.out=100 )
# define prior
prior <- rep(1,20)
# compute likelihood at each value in grid
likelihood <- dbinom(w+sum(sample(c(FALSE, TRUE), size=20-w, prob=c(0.8, 0.2), replace=T)) , size=w+l ,
# compute product of likelihood and prior
unstd.posterior <- likelihood * prior
# standardize the posterior, so it sums to 1
posterior <- unstd.posterior / sum(unstd.posterior)
```

Plot the posterior (unbiased in black, biased in red)

```
plot(p_grid , posterior , type="b" ,
      xlab="probability of water" , ylab="posterior probability" )
points(posterior_biased$p_grid , posterior_biased$posterior, col="red", type="b")
```



Easy questions from each chapter

Chapter 2

2E1. Which of the expressions below correspond to the statement: the probability of rain on Monday?

1. Pr(rain)
2. **Pr(rain|Monday)**
3. Pr(Monday|rain)
4. Pr(rain, Monday)/ Pr(Monday)

2E2. Which of the following statements corresponds to the expression: Pr(Monday|rain)?

1. The probability of rain on Monday.
2. The probability of rain, given that it is Monday.
3. **The probability that it is Monday, given that it is raining.**
4. The probability that it is Monday and that it is raining.

2E3. Which of the expressions below correspond to the statement: the probability that it is Monday, given that it is raining?

1. **Pr(Monday|rain)**
2. Pr(rain|Monday)
3. Pr(rain|Monday) Pr(Monday)
4. Pr(rain|Monday) Pr(Monday)/ Pr(rain)

5. Pr(Monday|rain) Pr(rain)/ Pr(Monday)

2E4. The Bayesian statistician Bruno de Finetti (1906–1985) began his book on probability theory with the declaration: "PROBABILITY DOES NOT EXIST." The capitals appeared in the original, so I imagine de Finetti wanted us to shout this statement. What he meant is that probability is a device for describing uncertainty from the perspective of an observer with limited knowledge; it has no objective reality. Discuss the globe tossing example from the chapter, in light of this statement. What does it mean to say "the probability of water is 0.7"?

On average, in 70% of tosses the result will be water.

Chapter 3

Run the code

```
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 )
```

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

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

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

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

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

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

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

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

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

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

```
quantile(samples, 0.2)
```

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

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

```
quantile(samples, 0.8)
```

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

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

```
HPDI(samples, 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, 0.66)
```

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