

# Chapter-03-assignment.R

*Ruijuan*

*February 23, 2016*

just for practice

install packages

```
library(rmarkdown)
library(rstan)
```

```
## Loading required package: ggplot2
```

```
## rstan (Version 2.9.0-3, packaged: 2016-02-11 15:54:41 UTC, GitRev: 05c3d0058b6a)
```

```
## For execution on a local, multicore CPU with excess RAM we recommend calling
```

```
## rstan_options(auto_write = TRUE)
```

```
## options(mc.cores = parallel::detectCores())
```

```
library(rethinking)
```

```
## Loading required package: parallel
```

```
## rethinking (Version 1.58)
```

R code 3.27

```
p_grid <- seq(0, 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)
```

use the values in samples to answer the questions

3E1-3E3 the intervals of defined boundary

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

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

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

```
sum(samples<0.2)/1e4
```

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

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

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

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

```
sum(samples>0.8)/1e4
```

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

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

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

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

```
sum(samples>0.2 & samples<0.8)/1e4
```

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

The intervals of defined mass

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

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

```
##      20%
```

```
## 0.5195195
```

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

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

```
##      80%
```

```
## 0.7567568
```

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|
```

```
## 0.5205205 0.7847848
```

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)
```

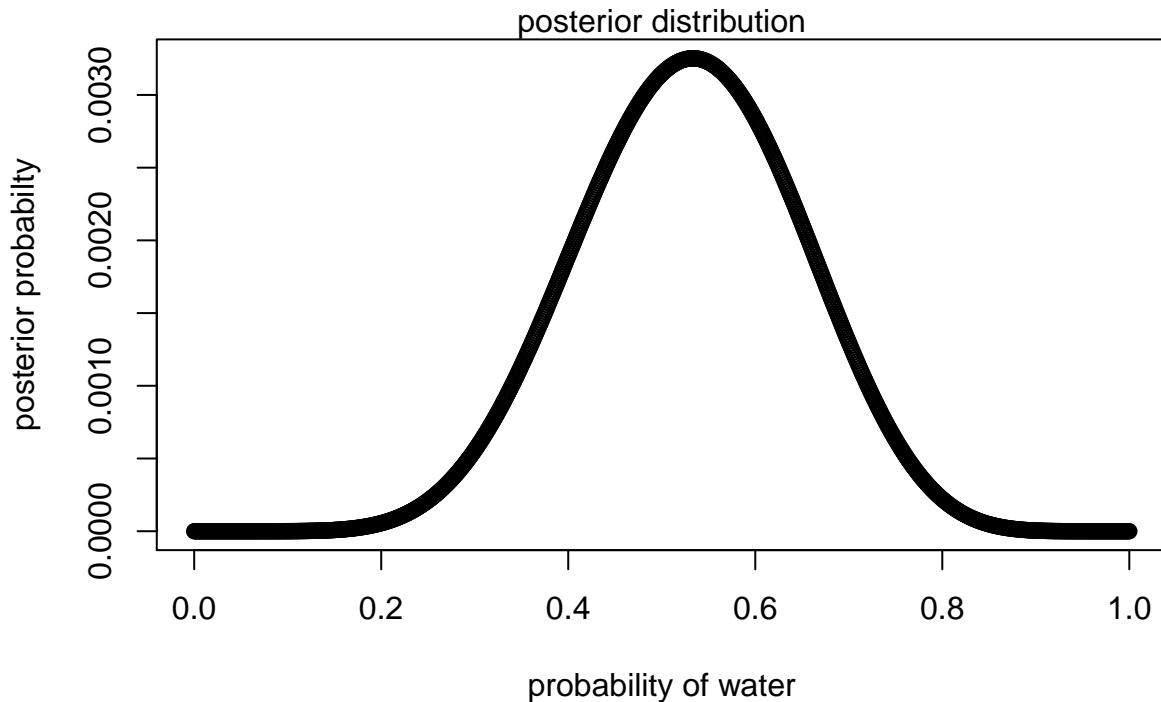
```
##      17%      83%
```

```
## 0.5005005 0.7687688
```

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

```
p_grid <- seq(0, 1, length.out = 1000)
prior <- rep(1, 1000)
likelihood <- dbinom(8, size = 15, prob = p_grid)
posterior <- likelihood * prior
posterior <- posterior/sum(posterior)
```



3M2. Draw 10,000 samples from the grid approximation from above.

Then use the samples to calculate the 90% HPDI() for p.

```
set.seed(8808)
samples <- sample(p_grid, prob = posterior, size = 1e4, replace = TRUE)
# the 90% highest posterior density (narrowest) interval
HPDI(samples, prob = 0.9)
```

```
##      |0.9      0.9|
## 0.3413413 0.7317317
```

*STOP AFTER 3M2 FOR 02/25 ASSIGNMENT*

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 does this mean).

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

```
# 1e5 random results from size of 15 (15 tosses, based on the prob of samples generated above)
dummy_w_eight_fifteen <- rbinom(1e5, size = 15, prob = samples)
table(dummy_w_eight_fifteen)/1e5
```

```
## dummy_w_eight_fifteen
```

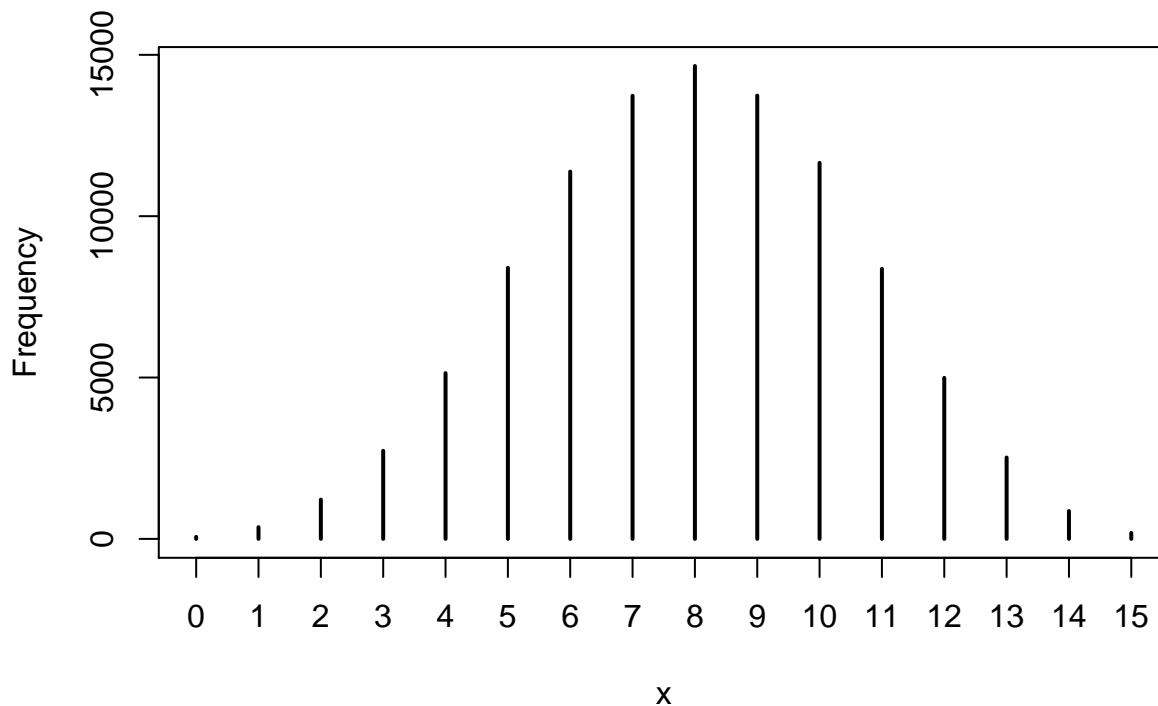
```
##      0      1      2      3      4      5      6      7      8
## 0.00062 0.00365 0.01218 0.02730 0.05138 0.08401 0.11381 0.13732 0.14654
##      9     10     11     12     13     14     15
## 0.13738 0.11653 0.08370 0.04989 0.02522 0.00863 0.00184
```

3M4. Using the posterior distribution constructed from the new (8/15) data, now calculate the probability of observing 6 water in 9 tosses. (don't understand the question, use this posterior distribution as what? draw samples from this posterior distribution? view this distribution as a model? )

```
dummy_w_six_nine_new <- rbinom(1e5, size = 9, prob = samples)
table(dummy_w_six_nine_new)/1e5
```

```
## dummy_w_six_nine_new
##      0      1      2      3      4      5      6      7      8
## 0.00556 0.02782 0.07549 0.13695 0.19100 0.20970 0.17677 0.11368 0.05056
##      9
## 0.01247
```

```
# check the new 8/15 distribution
simplehist(dummy_w_eight_fifteen)
```



```
# define a grid
p_grid <- seq(0, 1, length.out = 1e5)
# use new 8/15 as the posterior
posterior = dummy_w_eight_fifteen
posterior = dummy_w_eight_fifteen/sum(dummy_w_eight_fifteen)
# draw samples from posterior
```

```

set.seed(100)
samples <- sample(p_grid, prob = posterior, size = 1e4, replace = TRUE)
# sampling to simulate prediction
dummy_w_six_nine <- rbinom(1e5, size = 9, prob = samples)
table(dummy_w_six_nine)/1e5

## dummy_w_six_nine
##      0      1      2      3      4      5      6      7      8
## 0.09606 0.10079 0.10410 0.10208 0.09872 0.10010 0.10008 0.10012 0.09881
##      9
## 0.09914

```

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  $p=0.7$  (I feel there is very little difference made with different prior...)

```

# do p_grid approximation and draw samples
p_grid <- seq(0, 1, length.out = 1000)
prior <- rep(p_grid < 0.5, 0, 1)
likelihood <- dbinom(8, size = 15, prob = p_grid)
posterior <- likelihood * prior
posterior <- posterior/sum(posterior)
set.seed(100)
samples <- sample(p_grid, prob = posterior, size = 1e4, replace = TRUE)

# sampling to simulate prediction 8/15
dummy_w_eight_fifteen_new <- rbinom(1e5, size = 15, prob = samples)
table(dummy_w_eight_fifteen_new)/1e5

## dummy_w_eight_fifteen_new
##      0      1      2      3      4      5      6      7      8
## 0.00051 0.00361 0.01268 0.02866 0.05360 0.08486 0.11419 0.13528 0.14628
##      9     10     11     12     13     14     15
## 0.13758 0.11553 0.08269 0.05098 0.02378 0.00826 0.00151

```

```

# 6/9, with this original 8/15 posterior
dummy_w_six_nine_new <- rbinom(1e5, size = 9, prob = samples)
table(dummy_w_six_nine_new)/1e5

```

```

## dummy_w_six_nine_new
##      0      1      2      3      4      5      6      7      8
## 0.00577 0.02898 0.07638 0.13907 0.19231 0.20827 0.17435 0.11288 0.05000
##      9
## 0.01199

```

```

# with 8/15 posterior from simulated data
p_grid <- seq(0, 1, length.out = 1e5)
posterior <- dummy_w_eight_fifteen_new
posterior <- posterior/sum(posterior)
set.seed(100)
samples <- sample(p_grid, prob = posterior, size = 1e5, replace = TRUE)
dummy_w_six_nine_new_new <- rbinom(1e5, size = 9, prob = samples)
table(dummy_w_six_nine_new_new)/1e5

## dummy_w_six_nine_new_new
##      0      1      2      3      4      5      6      7      8
## 0.09847 0.10067 0.10052 0.09925 0.10082 0.10152 0.10030 0.09840 0.09927
##      9
## 0.10078

```

3H1

## R code 3.29, load data into R

```

library(rethinking)
data(homeworkch3)

```

## R code 3.30, to compute the total number of boys born across all of these births

```

sum(birth1) + sum(birth2)

```

```

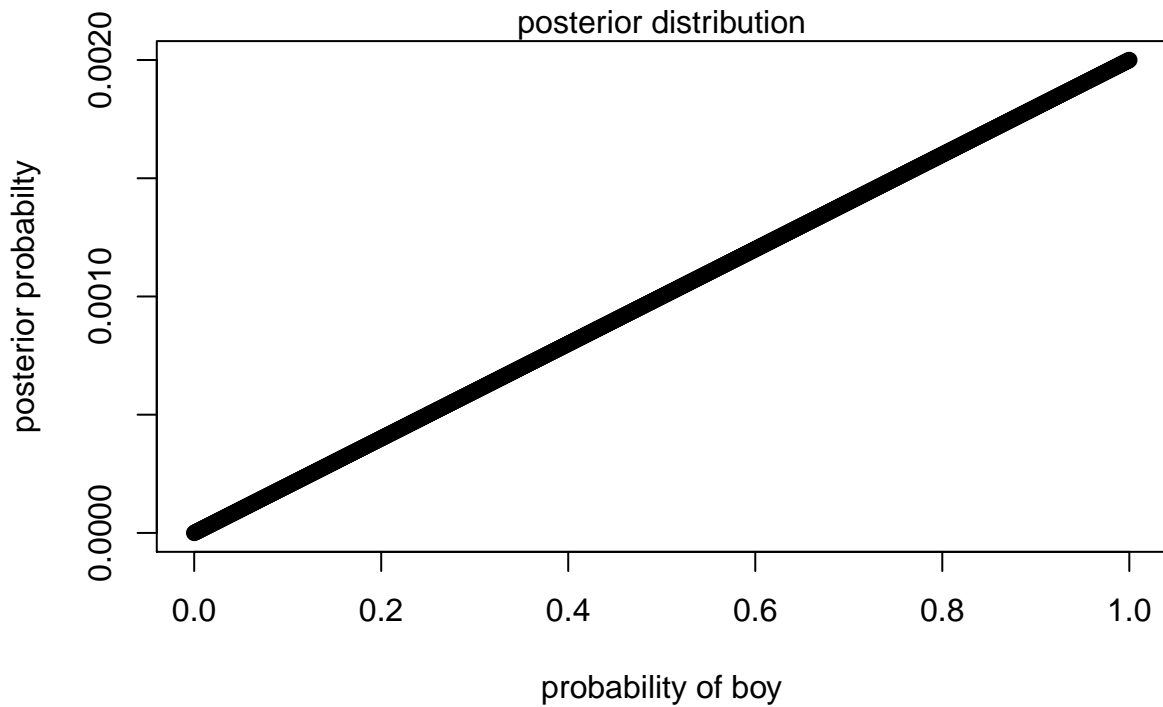
## [1] 111

```

```

# do p_grid approximation, to compute the posterior distribution of a birth being a boy, with
# uniform prior
p_grid <- seq(0, 1, length.out = 1000)
prior <- rep(1, 1000)
likelihood <- dbinom(1, size = 1, prob = p_grid) # also binomial distribution??? I don't think
# so!!! So this might be wrong!!!
posterior <- likelihood * prior
posterior <- posterior/sum(posterior)

```



3H2.

```
# randomly draw 1e4 samples and caculate 50%, 89%, and 97% highest posterior density intervals
set.seed(100)
samples <- sample(p_grid, prob = posterior, size = 1e4, replace = TRUE)
HPDI(samples, prob = 0.5)
```

```
##      |0.5      0.5|
## 0.7027027 0.9979980
```

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

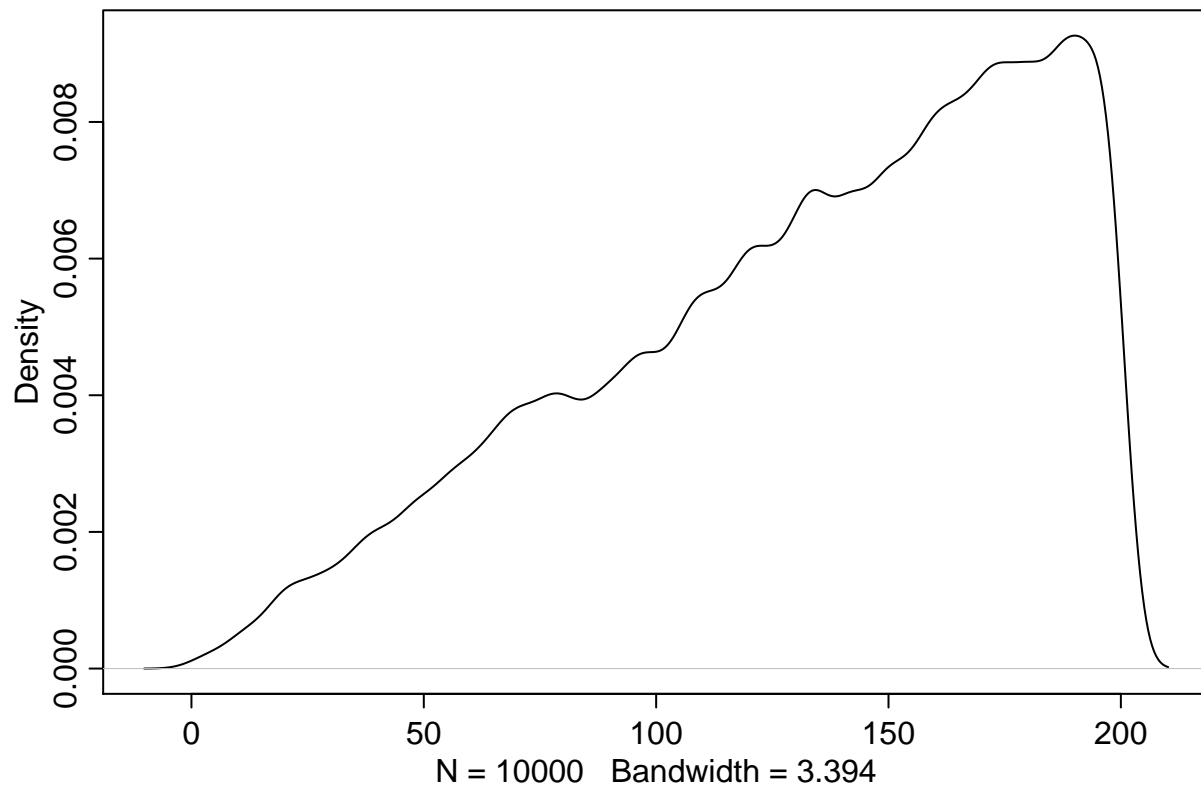
```
##      |0.89      0.89|
## 0.3253253 1.0000000
```

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

```
##      |0.97      0.97|
## 0.1731732 1.0000000
```

3H3.

```
# make simulated predictions
dummy_boy <- rbinom(1e4, size = 200, prob = samples)
dens(dummy_boy)
```



3H4. Now compare 10,000 counts of boys from 100 simulated first births only to the number of boys in the first births, birth1. How does the model look in this light? (don't understand this Q)

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 births. 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 births 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?