Statistical Rethinking: Chapter 3 - Sampling the Imaginary

Chris Hayduk

December 29, 2018

1 Easy

These problems use the samples from the posterior distribution for the globe tossing example. This code will give you a specific set of samples, so that you can check your answers exactly.

```
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)
```

Use the values in *samples* to answer the questions that follow.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Problem 3E6.
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
```

Problem 3E7.
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
```

2 Medium

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

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

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

HPDI(samples, prob = 0.9)

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

Problem 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?

```
w <- rbinom(1e4, size = 15, prob = samples)

sum(w[w = 8]) / 1e4

## [1] 0.0012
```

Problem 3M4.
Using the posterior distribution constructed from the nw (8/15) data, now calculate the probability of observing 6 water in 9 tosses.

```
w <- rbinom(1e4, size = 9, prob = samples)

sum(w[w = 6]) / 1e4

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

Problem 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.

```
#Re-doing 3M1
p_grid <- seq(from = 0, to = 1, length.out = 1000)
prior <- ifelse(p_grid < 0.5, 0, 1)
likelihood <- dbinom(8, size = 15, prob = p_grid)
posterior <- likelihood*prior
posterior <- posterior / sum(posterior)

#Re-doing 3M2
samples <- sample(p_grid, prob = posterior, size = 1e4, replace = TRUE)

HPDI(samples, prob = 0.9)

##      |0.9      0.9|
## 0.5005005 0.7097097

#Re-doing 3M3
w <- rbinom(1e4, size = 15, prob = samples)

sum(w[w = 8]) / 1e4

## [1] 6e-04

#Re-doing 3M4
w <- rbinom(1e4, size = 9, prob = samples)

sum(w[w = 6]) / 1e4

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

The HPDI is now a much narrower interval.

Since we eliminated many values that are far from the true value p = 0.7 through the use of our prior, the probability of observing 8 instances of water in 15 tosses is now much lower. The expected instances of water in 15 tosses with p = 0.7 is 10.5.

The expected instances of water in 9 tosses with p = 0.7 is 6.3. Once again, since we have eliminated many p values that are far from the true value p = 0.7, our estimate of the probability of seeing 6 instances of water in 9 tosses has improved. Since it is very close to the true value of p, the probability of seeing 6 instances of water in 9 tosses has increased with the new prior.

3 Hard

The practice problems here all use the dataset homeworkch3 from the rethinking package. This data indiactes the gender (male = 1, female = 0) of officially reported first and second born children in 100 two-child families.

```
data(homeworkch3)
```

Use these vectors as data. 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
```

Problem 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?

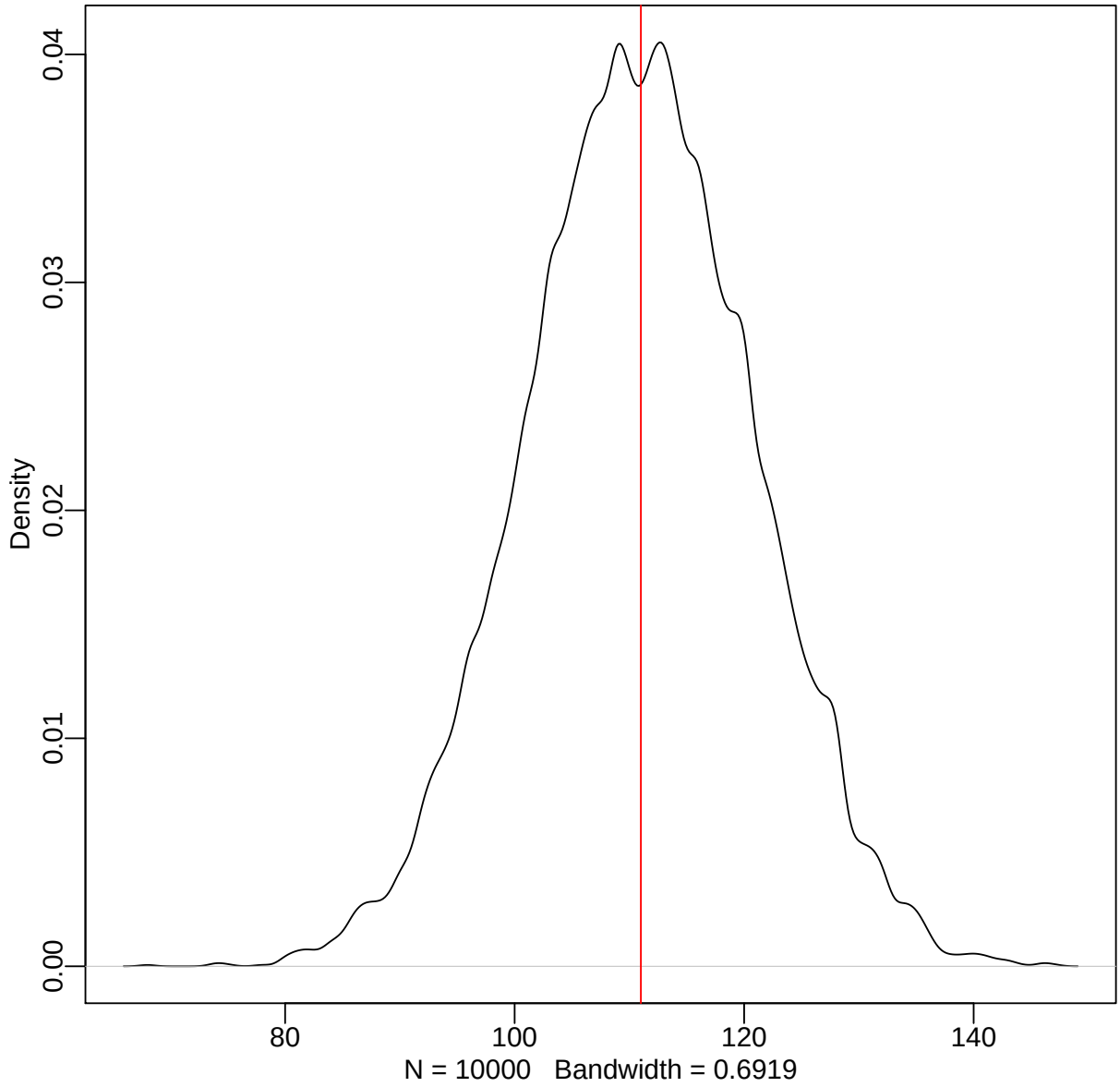
```
p_grid <- seq(from = 0, to = 1, length.out = 1000)
prior <- rep(1, 1000)

boys <- sum(birth1) + sum(birth2)
births <- length(birth1) + length(birth2)

likelihood <- dbinom(boys, size = births, prob = p_grid)
posterior <- likelihood*prior
posterior <- posterior / sum(posterior)

p_grid[which.max(posterior)]

## [1] 0.5545546
```

Problem 3H2.
Using the sample function, draw 10,000 random parameter values from the posterior distribution 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 = TRUE)

#50% HPDI

HPDI(samples, prob = 0.5)

##      |0.5      0.5|
## 0.5255255 0.5725726

#89% HPDI

HPDI(samples, prob = 0.89)

##     |0.89     0.89|
## 0.5015015 0.6116116

#97% HPDI

HPDI(samples, prob = 0.97)

##     |0.97     0.97|
## 0.4764765 0.6286286
```

Problem 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?

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

dens(w)
abline(v = sum(birth1) + sum(birth2), col = "red")
```

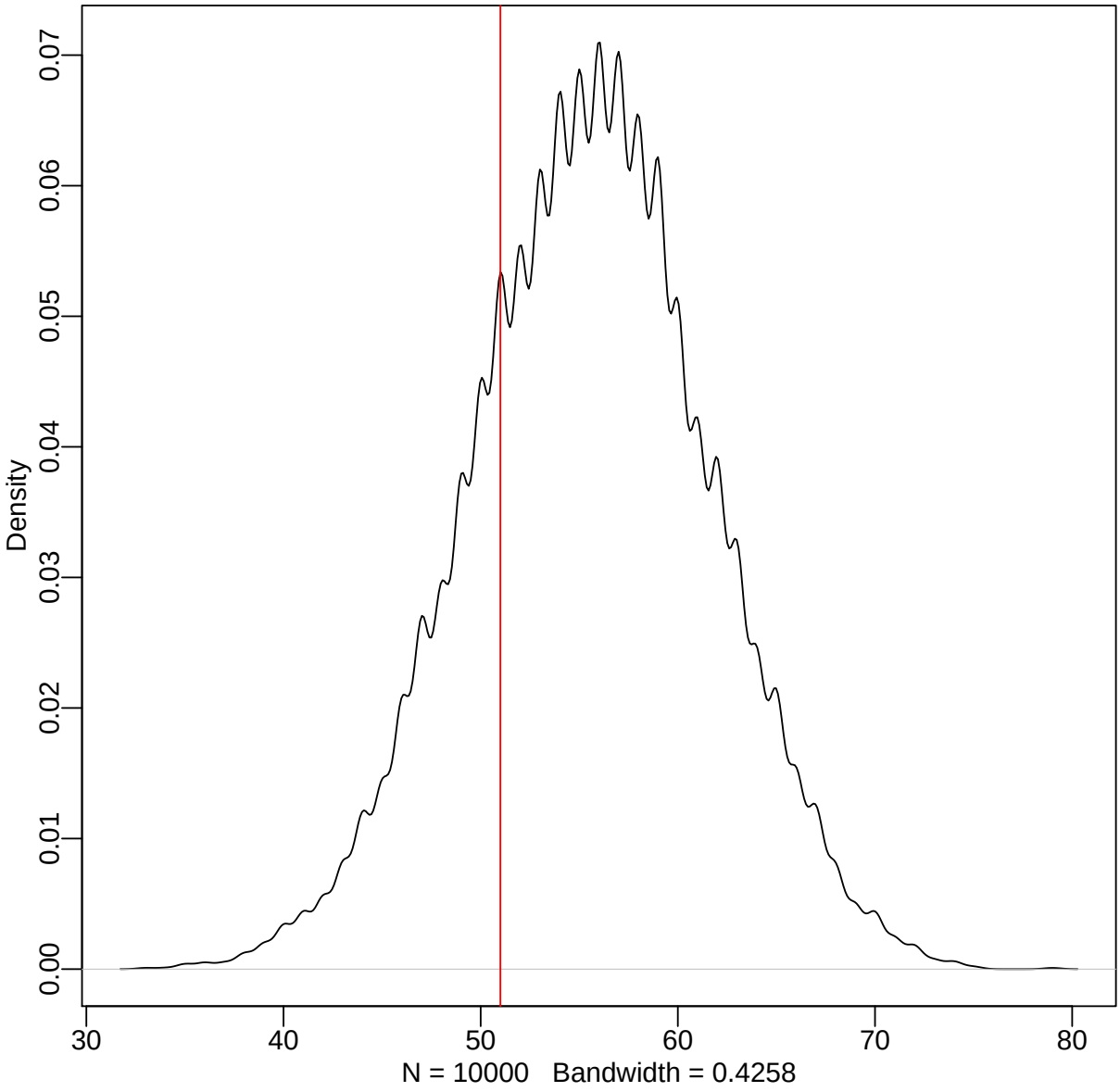


The peak of the distribution appears to be close to the true number of boys in the births (111 out of 200), so the distribution of predictions includes the actual observation as a central, likely outcome. Thus, it appears like the model fits the data well.

Problem 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?

```
w <- rbinom(1e4, size = 100, prob = samples)

#Number of boys in birth1
sum(birth1)
```

```
## [1] 51

dens(w)
abline(v = sum(birth1), col = "red")
```

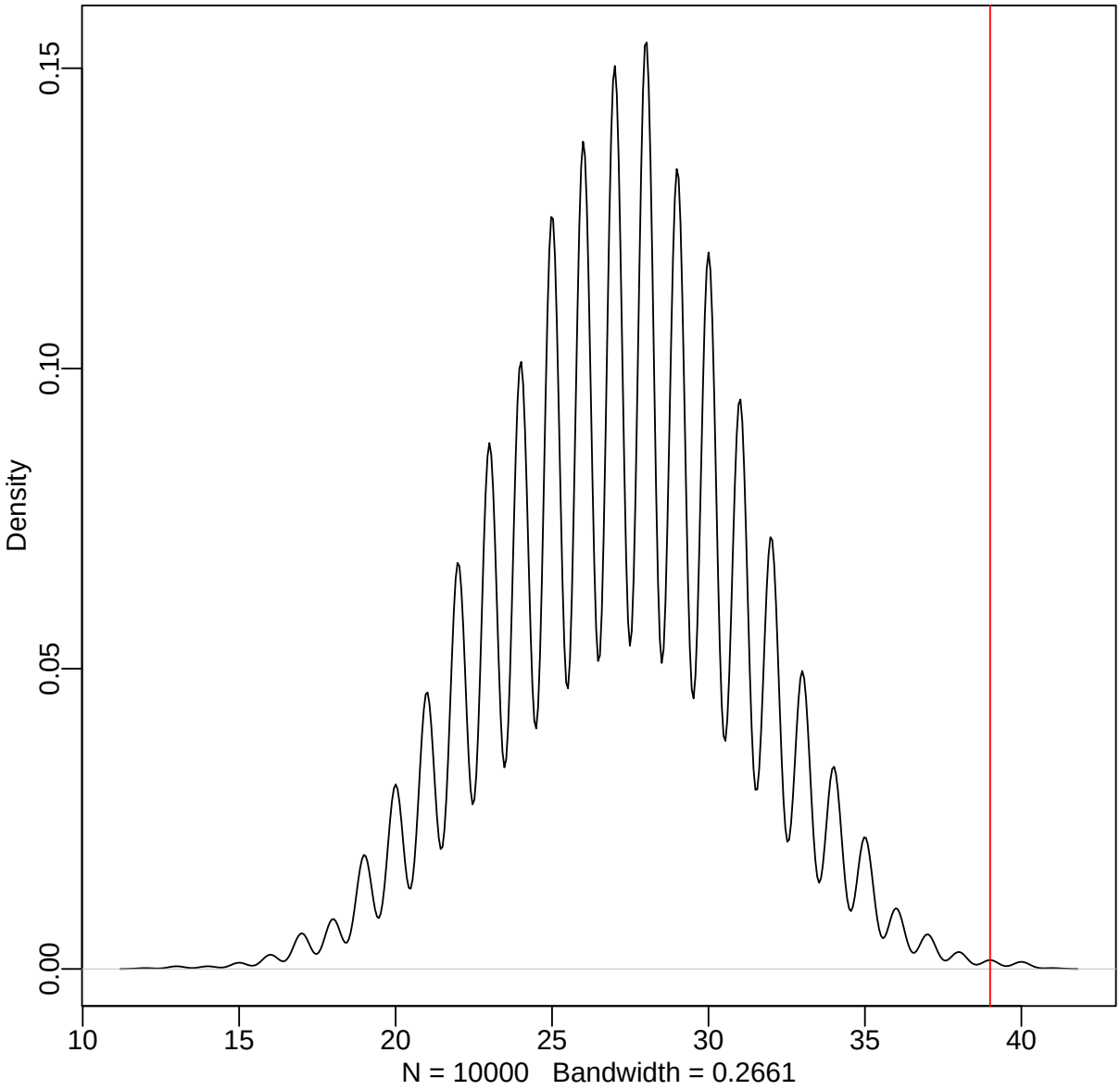


The model looks worse when comparing only to birth1. The peak of the distribution appears to be closer to 60, while there were 51 boys in birth1.

Problem 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 count 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 the data?

```
#Number of first borns who were girls
num_first_born_girls <- length(birth1) - sum(birth1)

w <- rbinom(1e4, size = num_first_born_girls, prob = samples)

boys_following_girls <- birth2[birth1 == 0 & birth2 == 1]

#Number of boys born after a girl
sum(boys_following_girls)

## [1] 39

dens(w)
abline(v = sum(boys_following_girls), col = "red")
```



The model appears to fail spectacularly in this instance. The true number of boys born after girls is 39, yet the density value is miniscule for this number. It is possible that the model is assuming that the first and second births are independent, while they may not be in reality.