Chapter 3

Easy

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.

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, prob = 0.2)
```

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

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

```
quantile(samples, prob = 0.8)
```

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

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

```
# find the HPDI
rethinking::HPDI(samples, prob = 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?

```
q <- (1 - 0.66)/2
quantile(samples, prob = c(q, 1-q))
```

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

Medium

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)
```

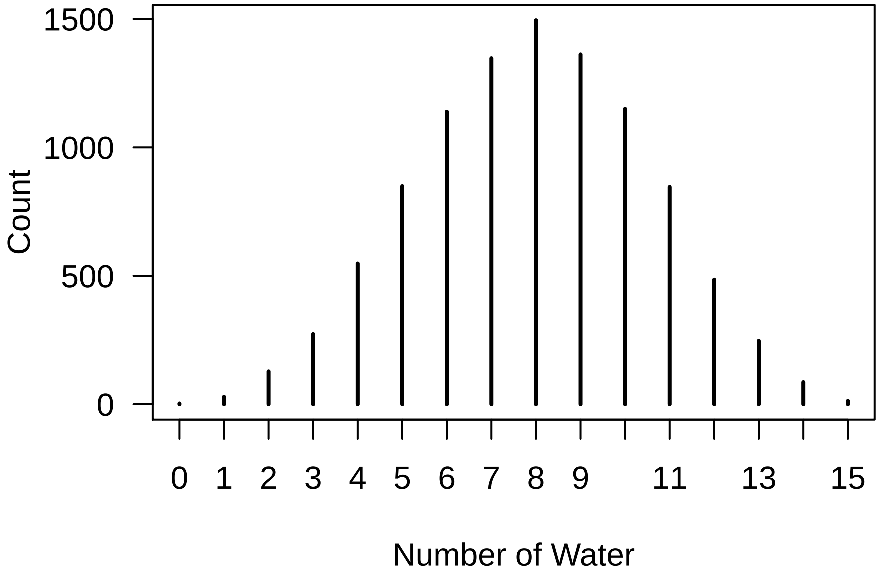
3M2. Draw 10,000 samples from the grid approximation from above. Then use the samples to calculate the 90% HPDI for p.

```
set.seed(101)
samples <- sample( p_grid , prob=posterior , size=1e4 , replace=TRUE )
rethinking::HPDI(samples, prob = 0.9)
```

```
##      |0.9      0.9|
## 0.3343343 0.7217217
```

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?

```
# borrowing from Code 3.26
set.seed(102)
w <- rbinom(1e4, size = 15, prob = samples)
plot(table(w), ylab = "Count", xlab = "Number of Water", las = 1)
```



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

```
## [1] 0.1495
```

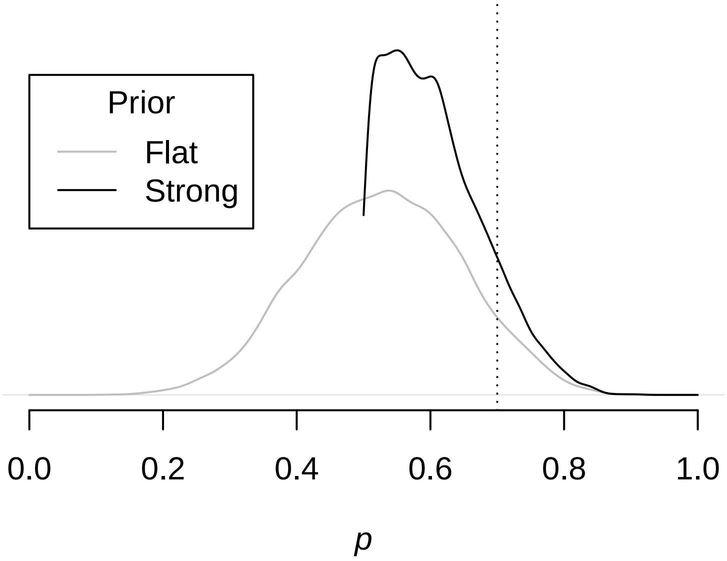
3M4. Using the posterior distribution constructed from the new (8/15) data, now calculate the probability of observing 6 water in 9 tosses.

```
w2 <- rbinom(1e4, size = 9, prob = samples)
sum(w2 == 6) / length(w2)
```

```
## [1] 0.176
```

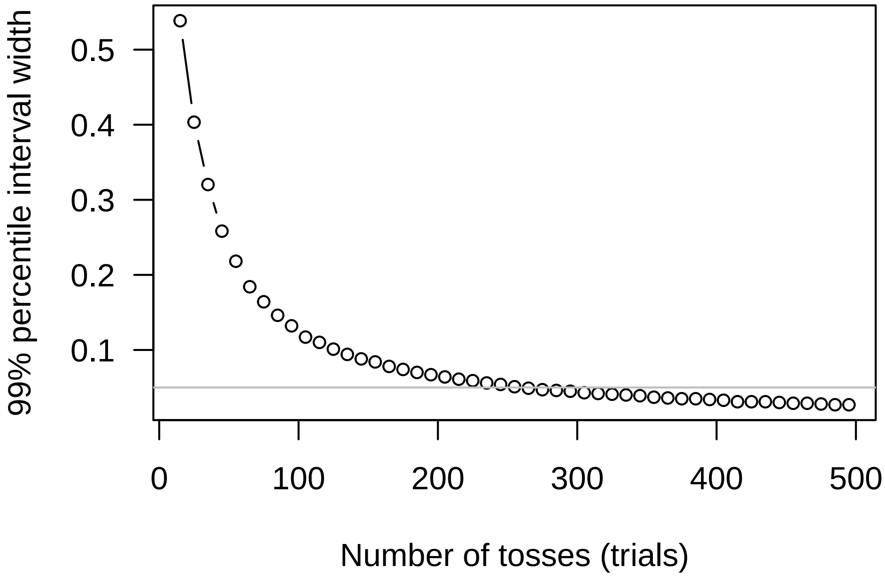
3M5. Start over at 3M1, but now use a prior that is zero below $p = 0.5$ and a constant above $p = 0.5$. This corresponds to prior information that a majority of the Earth's surface is water. Repeat each problem above and compare the inferences. What difference does the better prior make? If it helps, compare inferences (using both priors) to the true value $p = 0.7$.

```
prior2 <- ifelse(p_grid < 0.5, 0, 1)
posterior2 <- likelihood * prior2
posterior2 <- posterior2 / sum(posterior2)
set.seed(103)
samples2 <- sample( p_grid , prob=posterior2 , size=1e4 , replace=TRUE )

# compare the posterior distributions
plot(density(samples, from = 0, to = 1),
     xlim = c(0, 1), ylim = c(0, 6),
     xlab = expression(italic(p)),
     ylab = "", main = "", yaxt = "n", bty = "n",
     col = "grey")
lines(density(samples2, from = 0.5, to = 1))
abline(v = 0.7, lty = 3)
legend(0, 5, legend = c("Flat", "Strong"),
       col = c("grey", "black"), lty = c(1, 1),
       title = "Prior")
```



3M6. 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?

```r
n_toss <- seq(15, 500, 10)
width_99pi <-
    sapply(n_toss, function(toss) {
        p_grid <- seq( from=0 , to=1 , length.out=1000 )
        prior <- rep( 1 , 1000 )
        likelihood <- dbinom( 8 , size=toss , prob=p_grid )
        posterior <- likelihood * prior
        posterior <- posterior / sum(posterior)
        samples <- sample( p_grid , prob=posterior , size=1e4 , replace=TRUE )
        as.vector(diff(quantile(samples, prob = c(0.005, 0.985))))
    })

plot(n_toss, width_99pi, type = "b", las = 1, cex = 0.8,
     xlab = "Number of tosses (trials)",
     ylab = "99% percentile interval width")
abline(h = 0.05, col = "grey")
```



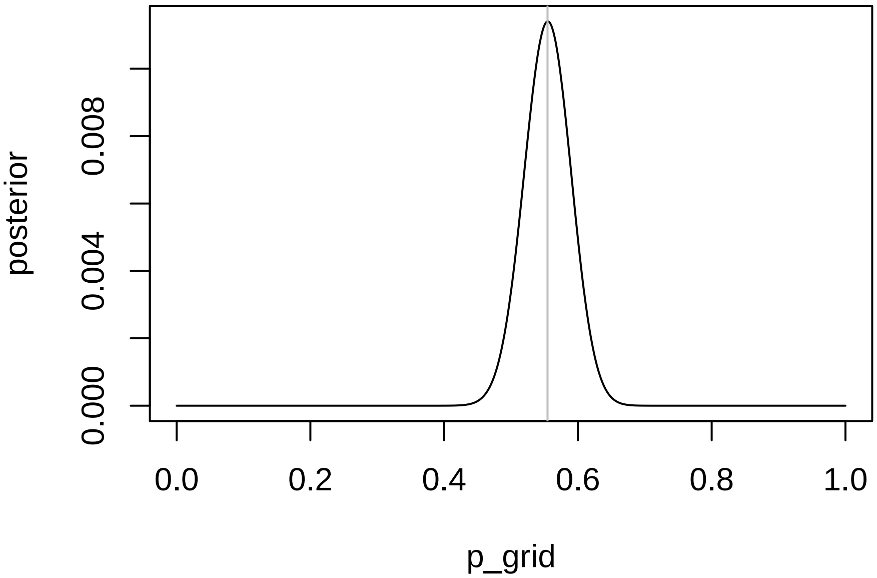
Hard

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.

```r
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)
```

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?

```r
n_boy <- sum(birth1) + sum(birth2)
n_birth <- length(birth1) * 2

p_grid <- seq( from=0 , to=1 , length.out=1000 )
prior <- rep( 1 , 1000 )
likelihood <- dbinom(n_boy, size = n_birth, prob = p_grid)
posterior <- likelihood * prior
posterior <- posterior / sum(posterior)

p_max <- p_grid[which(posterior == max(posterior))]
plot(p_grid, posterior, type = "l")
abline(v = p_max, col = "grey")
```



```r
p_max
```

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

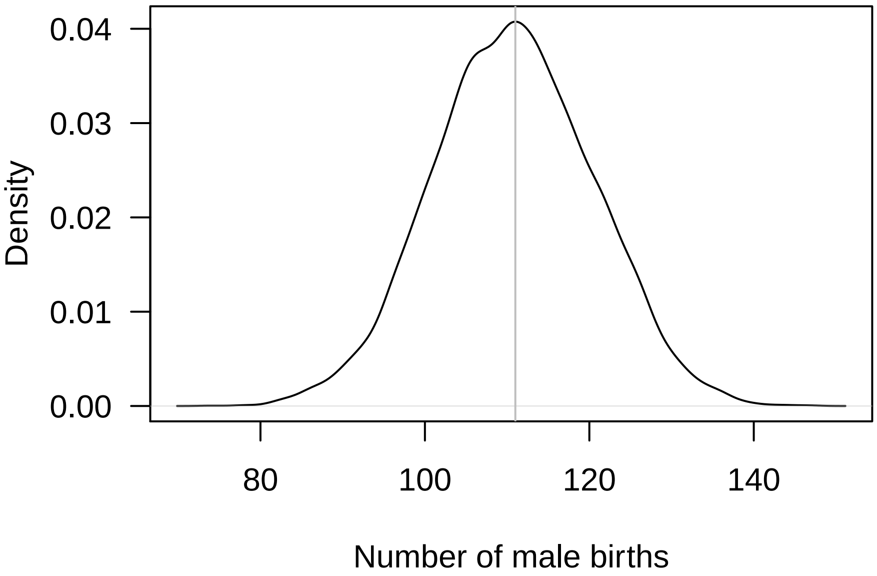
3H2. Using the sample function, draw 10,000 random parameter values from the posterior distribution you calculated above. Use these samples to estimate the 50%, 89%, and 97% highest posterior density intervals.

```r
set.seed(100)
samples <- sample(p_grid, prob = posterior, size = 1e4, replace = TRUE)
rethinking::HPDI(samples, prob = c(0.5, 0.89, 0.97))
```

```
##     |0.97     |0.89      |0.5      0.5|     0.89|     0.97|
## 0.4824825 0.4994995 0.5265265 0.5725726 0.6076076 0.6296296
```

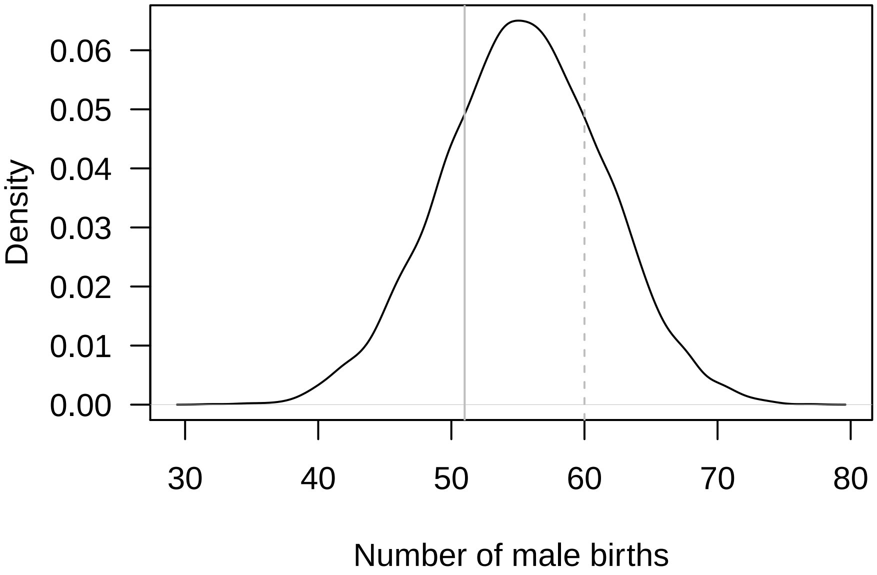
3H3. Use `rbinom` to simulate 10,000 replicates of 200 births. You should end up with 10,000 numbers, each one a count of boys out of 200 births. Compare the distribution of predicted numbers of boys to the actual count in the data (111 boys out of 200 births). There are many good ways to visualize the 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?

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

plot(density(n_boy_sim), main = "", xlab = "Number of male births", las = 1)
abline(v = 111, col = "grey")
```



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?

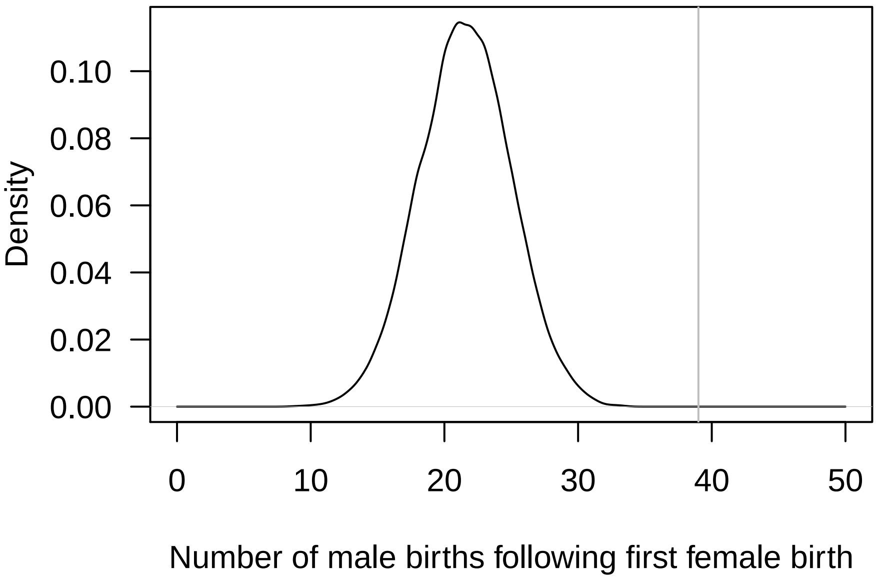
```
n_boy_sim <- rbinom(1e4, size = 100, prob = samples)
plot(density(n_boy_sim), main = "", xlab = "Number of male births", las = 1)
abline(v = sum(birth1), col = "grey")
abline(v = sum(birth2), col = "grey", lty = 2)
```



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 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?

```
gal_first <- birth2[birth1 == 0]
n_gal_then_guy <- sum(gal_first)
n_gal_then_guy_sim <- rbinom(1e4, size = n_gal_then_guy, prob = samples)
```

```
plot(density(n_gal_then_guy_sim, adjust = 1.2, from = 0, to = 50),
     main = "", xlab = "Number of male births following first female birth", las = 1,
     xlim = c(0, 50))
abline(v = n_gal_then_guy, col = "grey")
```


Density
Number of male births following first female birth