Notes on and Solutions for Statistical Rethinking

Alexander Pastukhov

2020-11-21

Contents

1 Precis **5**

2 Loss functions **7**

2.1 Loss function, the concept 7

2.2 L0 (mode) 8

2.3 L1 (median) 9

2.4 L2 (mean) 12

2.5 L1 (median) vs. L2 (mean) 13

2.6 Choosing a likelihood 14

2.7 Gaussian in frenquentist versus Bayesian statistics 16

3 Solutions for Chapter 2 **17**

4 Solutions for Chapter 3 **29**

Chapter 1

Precis

This is a collection of solutions for exercises but also of notes that attempt to provide further details and intuition for some topics, such as information theory, information criteria, MCMC algorithms, etc.

Chapter 2

Loss functions

The purpose of this comment is to give you an intuition about loss functions, mentioned in chapter 3. In particular, I want you to understand why different loss functions (L0, L1, and L2) correspond to different point-estimates (mode, median, and mean). Plus, I want you to understand that you can view a choice of a likelihood function, as in picking Gaussian in chapter 4, as being analogous to picking a loss function.

I am afraid that the easiest way to explain why an *L2* loss results in *mean* is via a derivative. So, if you are not confident in your basic calculus skill, it might be useful for you to first watch a few episodes of Essense of Calculus series by Grant Sanderson, a.k.a. 3Blue1Brown. I would suggest watching at least the first three episodes (actually, I would recommend to watch the whole series) but if you are short on time watch only episode 2[1].

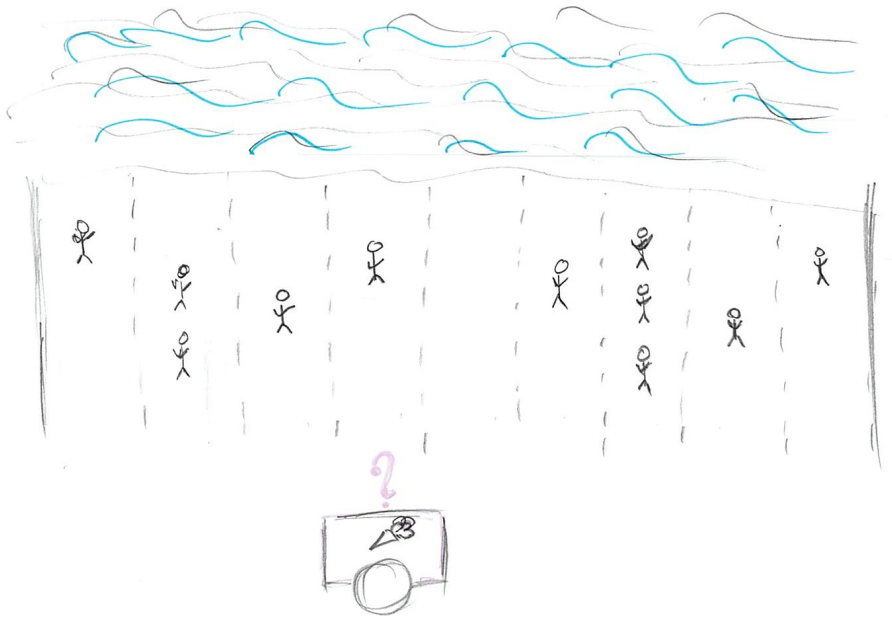
2.1 Loss function, the concept

Imagine that you are selling ice-cream on a beach, so we can assume it is a narrow strip of sand and, therefore, a one-dimensional problem. It is hot, so *everyone* wants ice-cream (obviously) and you want to maximize the number of ice-creams you sell (obviously). People are distributed in some random (not necessarily uniform or symmetric) way along the beach, so the question is: Where do you put your *single* ice-cream stand to maximize your profits? The answer depends on your choice of the *loss function* that describes how distance between a particular person and your stand influences whether person will buy your ice-cream. In other words, it describes the *cost* of getting to your stand, i.e. walking all-the-way through the sand in that heat. This *cost* clearly depends on the

[1] Although, if you skip episode 1, you won't know why it is *obvious* that area of a circle is $\pi \cdot r^2$

distance and in the simplest case, it is linearly proportional to the distance: If you need to walk twice the distance, your costs for getting an ice-cream are twice as high. However, the relationship between the distance and cost does not have to be so simple and linear and this is why we have many different *loss / cost* functions.

We can write a loss/cost function more formally as $L(stand, person_i)$ where `stand` is the location of your stand and `person_i` is a location of a particular ith person. The cost can be either zero or positive, i.e., we assume there is no benefit in walking all the way, only no or some cost. So, where should you put your ice-cream stand?



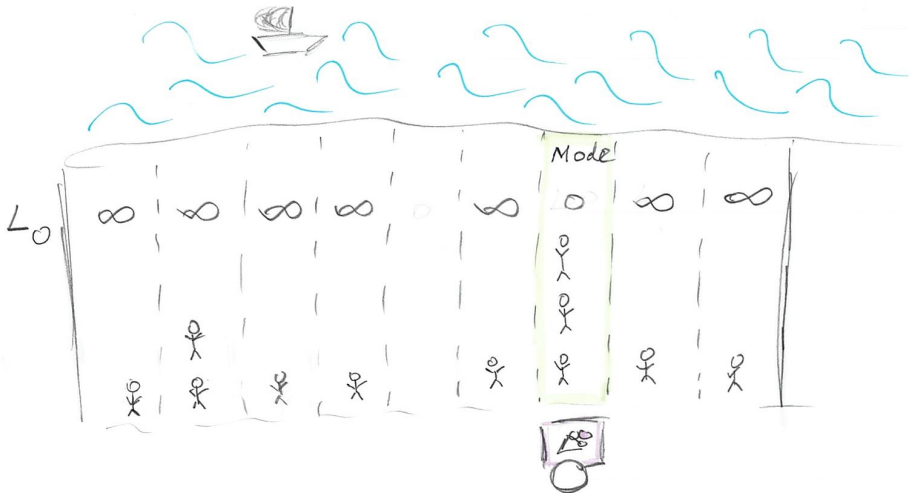
2.2 L0 (mode)

The simplest loss function is

$$L0(stand, person_i) = \begin{cases} 0, stand == person_i \\ \infty, stand \neq person_i \end{cases}$$

This function assumes that everybody hates walking so much, that *any* walk is unbearable and should be avoided. Thus, there is no cost for getting your ice-cream only for people who are positioned right next to your stand. For everybody else, even one meter away, the costs of walking are infinite, so they

won't bother and, therefore, won't buy your ice-cream. Still, we are in the business of selling one, so where do we put our stand given how lazy our customers are? Well, we just find the biggest group of people and put our stand next to them. No one else will come but at least you got the biggest group of customers you could. If you look at the *distribution* of your customers along the beach this is the highest peak (that you peak) and it is called the *mode* of the distribution.

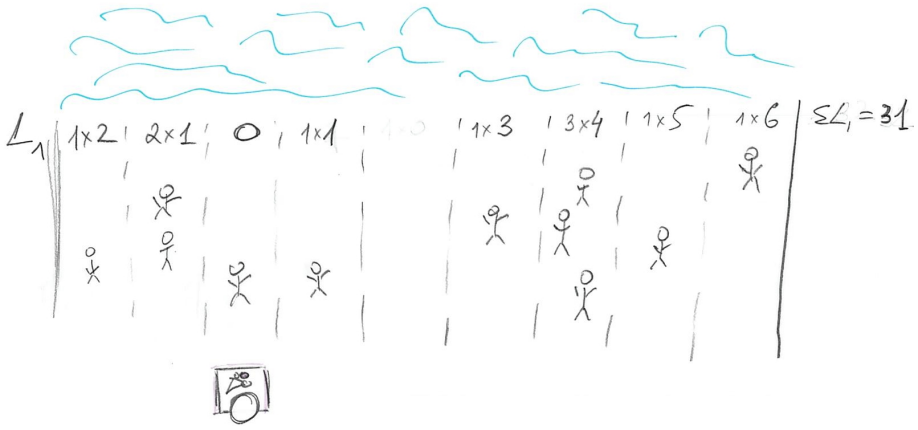


2.3 L1 (median)

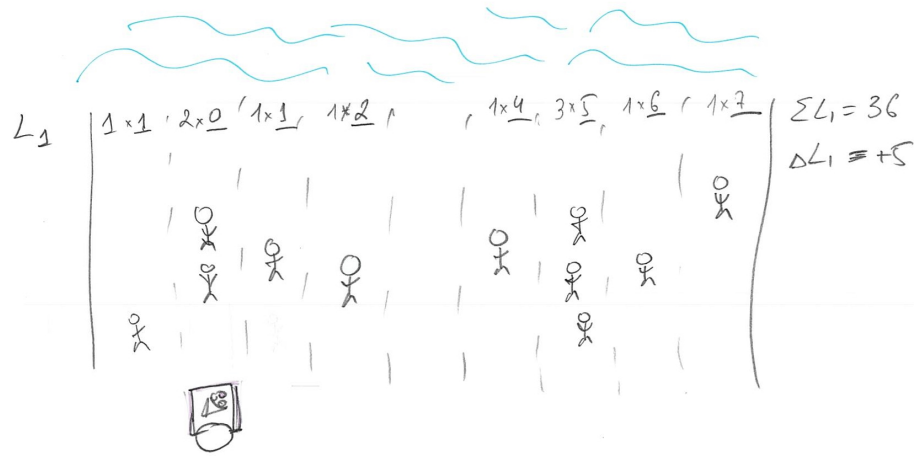
The next loss function, that I already mentioned, assumes a simple linear relationship between the distance and the cost

$$L1(stand, person_i) = |person_i - stand|$$

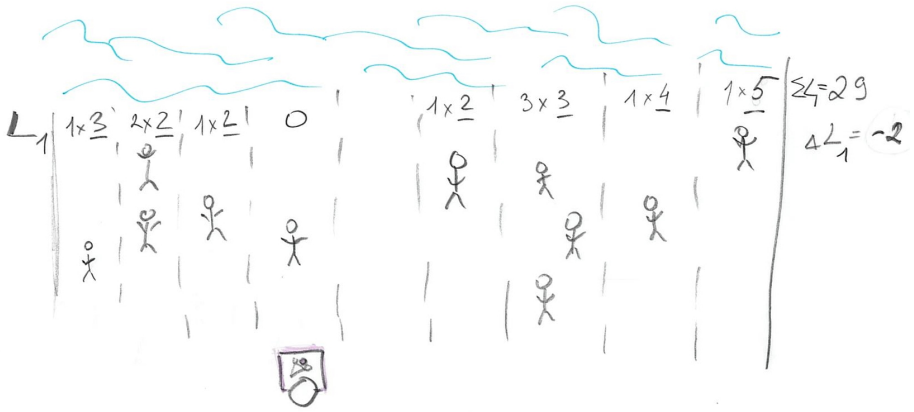
In other words, the cost is equal to distance (we need `|` `|` to get an absolute value, because the person could be "to the left of" of stand, in which case `person - stand` distance will be negative). So, where should we put our stand? Let us start at a fairly random location so that 3 of our customers are on the left and 7 are on the right.



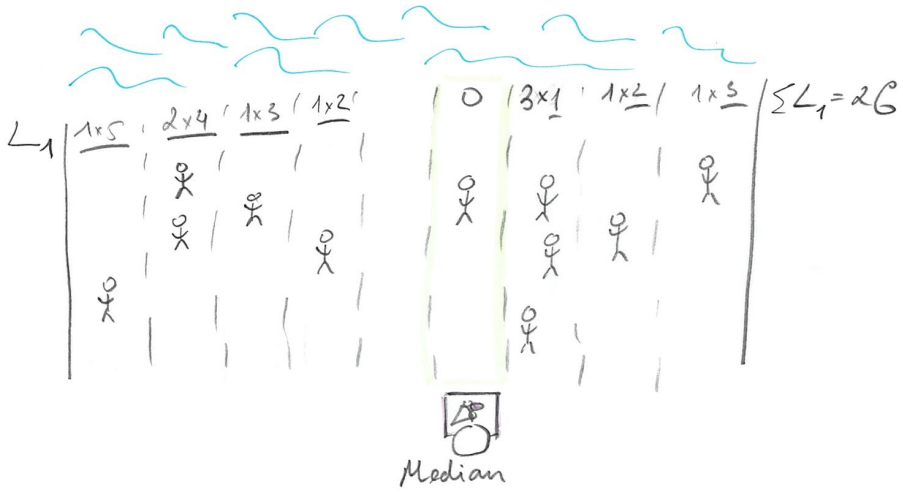
We can, in principle, compute the actual cost but it is simpler to ask the question of whether we can *improve* on that cost by moving somewhere else? Imagine that we move to the left where *minority* of our customers are. Now we have 1 on the left and 8 on the right (plus 2 more at our location).



The problem is, we moved *away* from the majority of the people so our total cost is *original cost - 3 (improvement due to moving close to minority) + 8 (increase in loss due to moving away from majority)*, so $\Delta L1 = +5$. Oops, we made it worse! How about moving to the *right*?



Now that we move *towards* the majority of customers, we have four on the left and six on the right (plus one at our location). The change in cost is *original cost + 4 (loss due to moving away from minority) - 6 (improvement due to moving towards majority)*, so $\Delta L1 = -2$. Which gives us an idea: we should try to get even closer to that majority by keeping walking to the right! Eventually, you will get to point of the 50/50. Should you keep moving to the right? Should you move to the left? Should you move at all?



There is no point in moving to the left. You just came from where because moving to the right made things better. However, if you keep moving to the right, you will keep passing people, so that majority now will be on the left and you would be walking *away* from the majority, raising the costs (and your losses). So, once you get to point where half of your customers are on the left and half are on the right, you cannot do any better. Any movement that gets

you from 50/50 means there are more customers on one side (say left, if you moved to the right) and, as we already figured out, your best strategy is to move towards the majority, which gets you back where you started at 50/50 point. That 50/50 points split, when half of customers / probability mass is on one side and half is on the other, is called *median*.

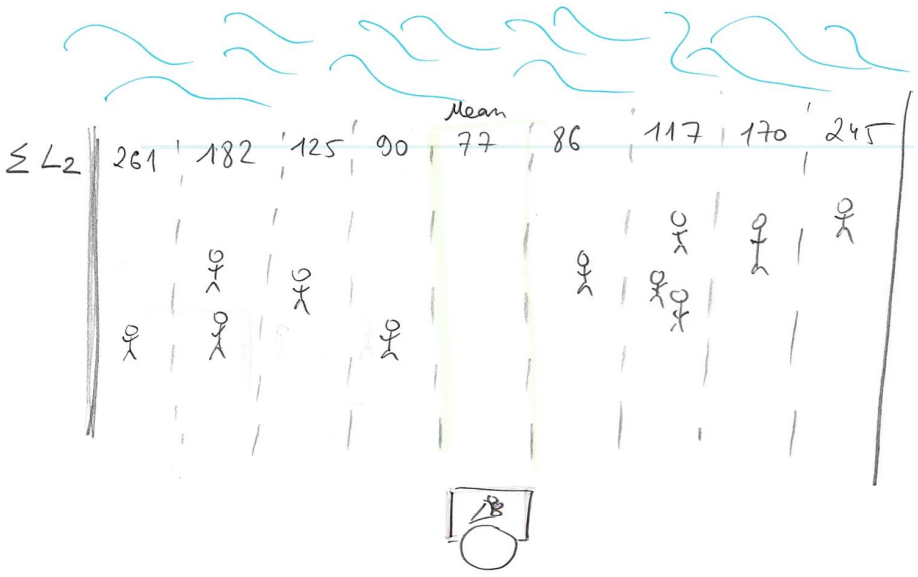
2.4 L2 (mean)

The classic loss function is Euclidean distance

$$L2(stand, person_i) = (person - stand)^2$$

Here, every next step becomes progressively harder for our customers. The cost of walking 1 meter is 1 (unit of effort). But walking 2 is $2^2 = 4$ and is $3^2 = 9$ for 3 meters. Thus, the penalty (cost/loss) for being further away from your stand increases as a power law. Still, one needs to sell ice-cream, so one needs to find the best spot where total cost is minimal

$$L2(stand, person) = \sum_{i=1}^{N} (person_i - stand)^2$$



Or, we can compute the minimal *average* cost by dividing the sum by the total number of customers `N`:

$$< L2(stand, person) >= \frac{1}{N} \sum_{i=1}^{N} (person_i - stand)^2$$

Conceptually, you find that minimum by walking along the beach in the direction that reduces the cost until you hit the point where it start going up again. This strategy is called *gradient descent* and, generally speaking, this is how computer finds minima computationally: They make steps in different directions to see which way is down and keep going until things start going up. However, in one-dimensional well-behaving case we have here things are even simpler as you can use calculus to figure out the solution analytically. If you watched the videos I advertised above, you'll know that the *derivative* of the function is zero at the extrema (minima or maxima), so we just need to differentiate our average *L2* over position of the stand and find where it is zero[2].

$$\frac{\partial L2}{\partial stand} = -\frac{2}{N}\sum_{i=1}^{N}(person_i - stand)$$

As we want $\frac{\partial L2}{\partial stand} = 0$, we state

$$-\frac{2}{N}\sum_{i=1}^{N}(person_i - stand) = 0$$

. Opening up brackets and rearranging we get

$$-\frac{2}{N}\sum_{i=1}^{N} person_i + \frac{2 \cdot N}{N} \cdot stand = 0 2 \cdot stand = \frac{2}{N}\sum_{i=1}^{N} person_i stand = \frac{1}{N}\sum_{i=1}^{N} person_i$$

So, the optimal location of your stand is the *mean*: an average location of all people on the beach.

2.5 L1 (median) vs. L2 (mean)

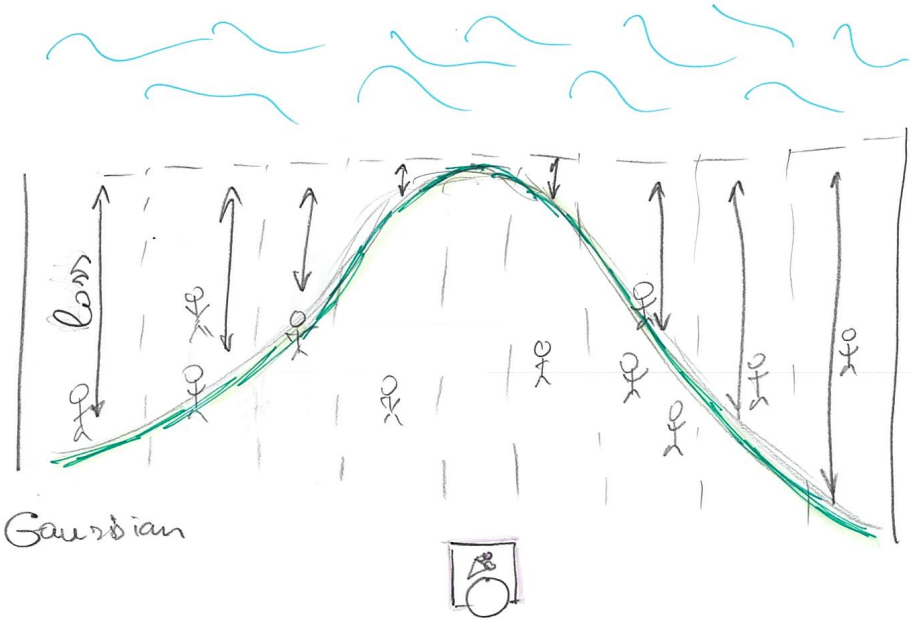
One problem about the *mean* is that it is sensitive to outliers. Because the costs grow as a power law, this approach favors a lot of medium-sized distances over lots of smalls ones plus one really large one. Thus, a single person at a far side of the beach would have a big influence on your stand's location (you already saw the difference in the example above). In data analysis, this means that those outliers will pull your estimates away from the majority of responses. Which is why it might be a good idea to consider using `median` rather than `mean`. If you distribution is symmetric, the difference will be negligible but in presence of outliers `median`, as a point-estimate, is more robust.

[2] I've nicked the derivations from [https://stats.stackexchange.com/a/312997]

2.6 Choosing a likelihood

So far we talked about selling ice-cream on the beach but same question of choosing your loss function applies when you are trying to fit a distribution or a regression line, as in chapter 4. Here, you also have a point-estimate (regression line at each point) and you try to put it in such a way as to minimize the costs of having data points off that line (the distance from the point-estimate of the line and each data point is called a *residual*). The classic way is to use *L2* distance and the approach is called *ordinary least squares*, as you try to minimize squared residuals.

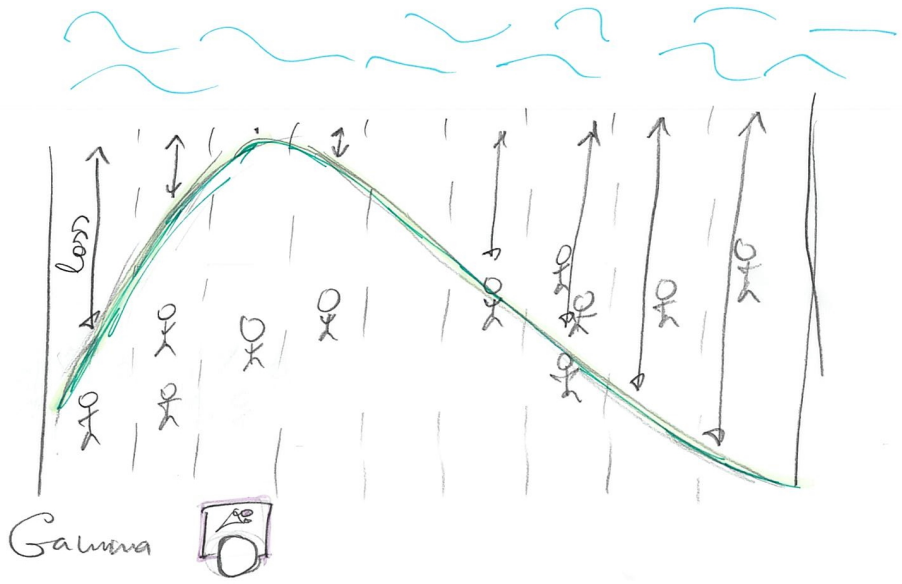
Alternatively, you can express same costs-of-being-off-the-line using a distribution, such as Gaussian. You put its peak (mean) at the (candidate) location of your point estimate (that point has highest probability, so lowest cost) and the loss is computed as a probability of the residual (distance-to-the-point). You can think about it in terms of the probability that a person will go and buy ice-cream from your stand.



The Gaussian is special because it uses L2 distance, see $(x-\mu)^2$ inside the exponential:

$$f(x) = \frac{1}{\sigma\sqrt{(2\pi)}} e^{\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)}$$

so using it is equivalent to fitting via ordinary least squares. However, as McElreath hinted, you can choose different priors that are different not only in the distance-to-loss formula (like *L1* is different from *L2*) but also in symmetry. Both *L1* and *L2* (and Gaussian) ignore the sign of the distance. It does not matter whether customers are on the left or on the right. Other distributions, such as Beta, Gamma, or Log Normal are not symmetric, so the same distance will cost differently depending on the side the customer is at.



This allows you to think about the choice of your likelihood distribution in terms of choosing a loss function. For example, a t-distribution has heavier tails than a Gaussian (if you want to sound like a real mathematician, you say "leptokurtic"), so its losses for outliers (penalty for larger residuals) are lower. Using it instead of a Gaussian would be similar to changing the loss function from L2 to be more like L1 (e.g. $|person_i - stand|^{1.5}$). Conversely, you can pick a symmetric distribution that is narrower than a Gaussian to make residuals penalty even higher (e.g. using $(person_i - stand)^4$). You can also consider other properties: Should it be symmetric? Should it operate only within certain range (1..7 for a Likert scale, 0..1 for proportions, positive values for Gamma)? Should it weight all points equally? As you saw in the examples above, picking a different function moves your cart (regression line), so you should keep in mind that using a different likelihood will move the regression line and produce different estimates and predictions.

How do you pick a likelihood/loss function? It depends on the kind of data you have, on your knowledge about the process that generated the data, robustness of inferences in the presence of outliers, etc. However, most real-life cases you

are likely to encounter will be covered by the distributions described in the book (Gaussian, exponential, binomial, Poisson, Gamma, etc.). After finishing the book, you will have a basic understanding of which are most appropiate in typical cases. The atypical cases you'll have to research yourself!

2.7 Gaussian in frenquentist versus Bayesian statistics

Later on in the book McElreath will note that erroneously assuming normal distribution for residuals ruins your inferences in frequentist statistics but not in Bayesian. This is because of parametric nature of inferences in frequentist statistics. By assuming that residuals are normal, you can approximate significance by using t- or F-distribution with appropriate degrees of freedom. However, if your initial assumption is wrong, your inferences will be based on idealized not real residuals. If you residuals tend to be asymmetric, as in any proportion data, the numbers you see are about *symmetric* residuals with the observed variance, not your real *asymmetric* residuals. Numbers will be off and it won't be easy (although probably possible) to say by how much. In contrast, Bayesian statistics does not use approximations and builds the posterior distribution directly. Thus it treats Gaussian (or any other distribution) as a loss function and the asymmetry or abnormality of residuals is less critical (merely not optimal).

Chapter 3

Solutions for Chapter 2

```
library(tidyverse)
library(rethinking)
```

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 1973 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"?

He meant that, at least at macro level, there are no truly random events. At that scale, all events, such as motion of a globe throughout its flight, are described by deterministic laws of physics. Therefore, the outcome of globe tossing is deterministic and predictable if we have a complete knowledge about its initial momentum and about other forces at play (e.g., air movement). In most cases, we do not have the full knowledge and, therefore, our predictions are likely to diverge from the observed outcome and be mostly but not always correct. Hence, the concept of probability. At the end, noise is never truly random, noise is information that we did not include into our model explicitly.

2M1

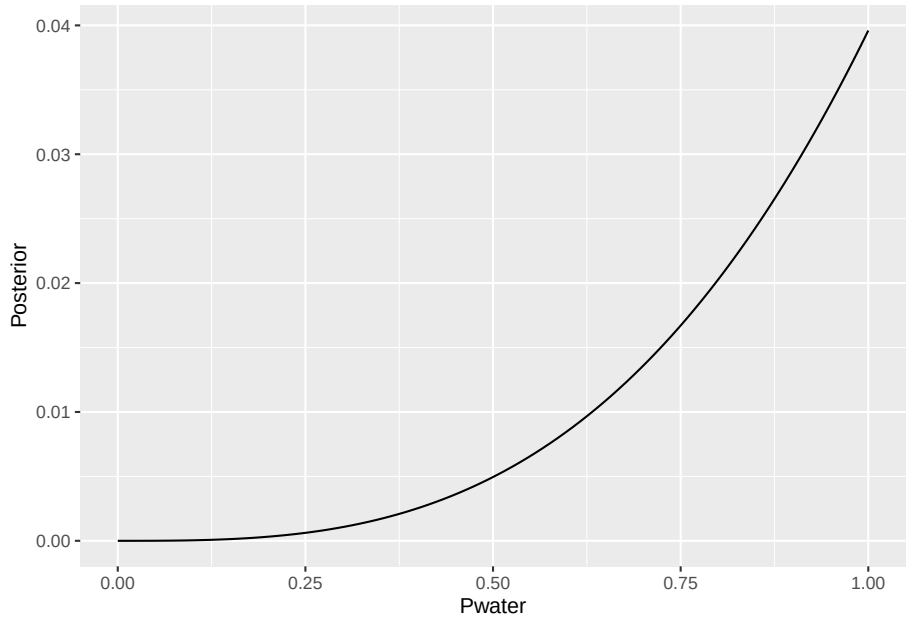
Recall the globe tossing model from the chapter. Compute and plot the grid approximate posterior distribution for each of the following sets of observations. In each case, assume a uniform prior for p.

```
#' Computes posterior for water probability assuming binomial likelihood
#'
#' @param observations vector of "W" and "L"
#' @param prior numeric vector, its length determines grid resolution
#'
#' @return tibble with Pwater (from 0 to 1), Posterior (probability of water for given
#'
#' @examples
#' probability_of_water(c("W", "W", "W"), rep(1, 100))
probability_of_water <- function(observations, prior){
```

```
  p_grid <- seq(from=0, to=1, length.out=length(prior))
  likelihood <- dbinom(sum(observations == "W"), size=length(observations), prob=p_grid)
  unstandardized_posterior <- likelihood * prior
  tibble(Pwater = p_grid,
         Posterior = unstandardized_posterior / sum(unstandardized_posterior),
         Prior = prior)
}
```

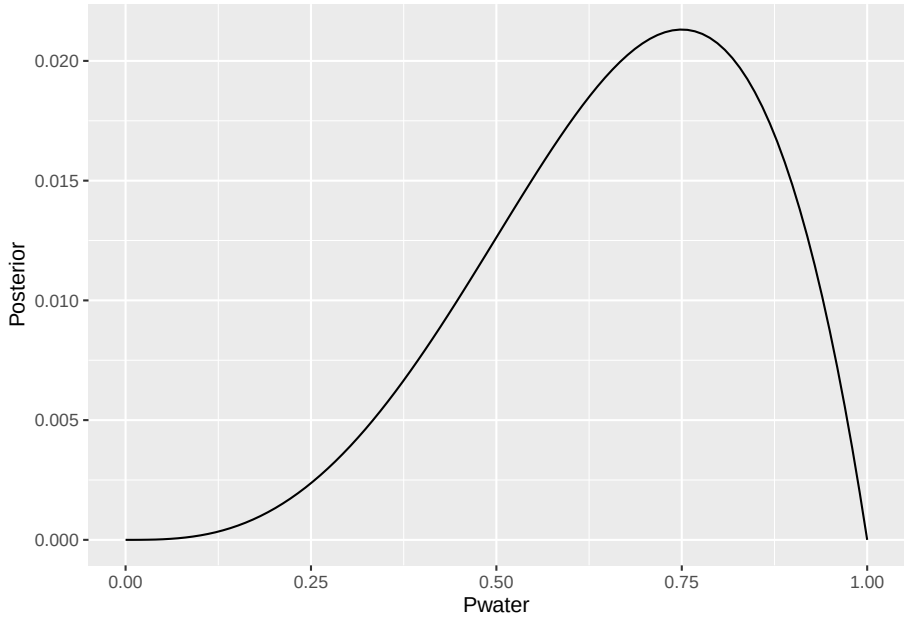
1. W, W, W

```
posterior2M1 <- probability_of_water(c("W", "W", "W"), rep(1, 100))
ggplot(posterior2M1, aes(x=Pwater, y=Posterior)) +
  geom_line()
```



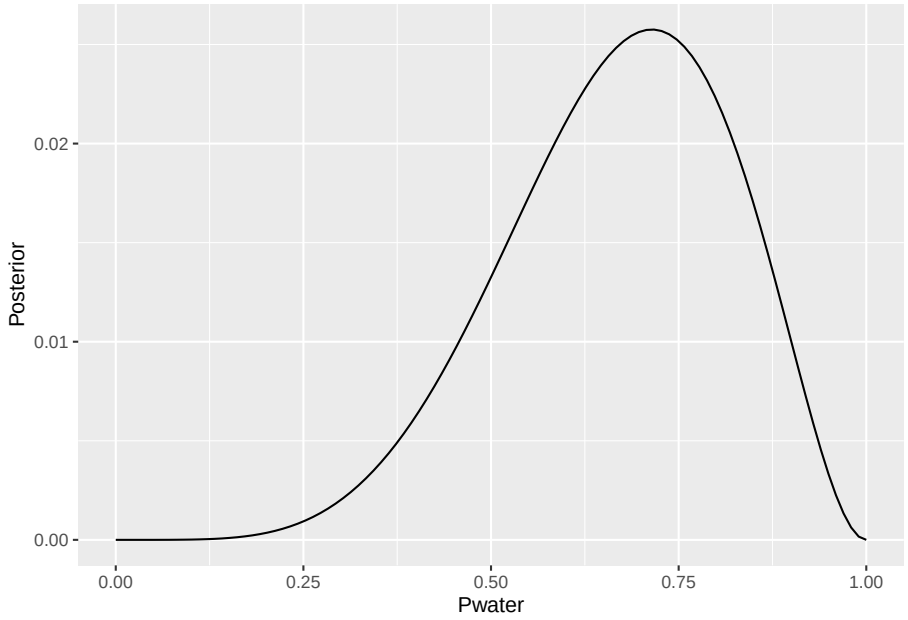
2. W, W, W, L

```
posterior2M2 <- probability_of_water(c("W", "W", "W", "L"), rep(1, 100))
ggplot(posterior2M2, aes(x=Pwater, y=Posterior)) +
  geom_line()
```



3. L, W, W, L, W, W, W

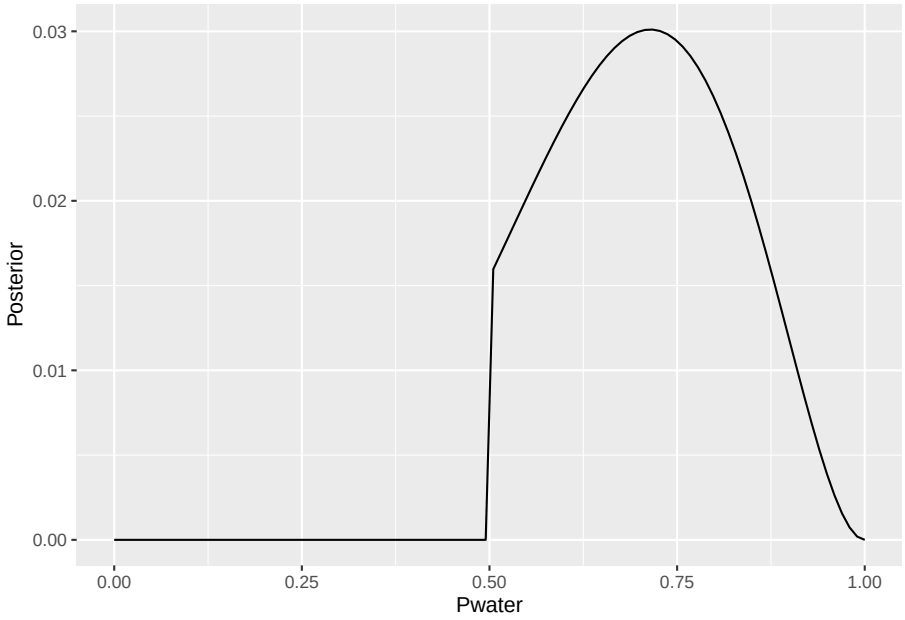
```
posterior2M3 <- probability_of_water(c("L", "W", "W", "L", "W", "W", "W"), rep(1, 100))
ggplot(posterior2M3, aes(x=Pwater, y=Posterior)) +
  geom_line()
```



2M2

Now assume a prior for p that is equal to zero when p<0.5 and is a positive constant when p 0.5. Again compute and plot the grid approximate posterior distribution for each of the sets of observations in the problem just above.

```
p_grid <- seq(from=0, to=1, length.out = 100)
priorM2 <- as.numeric(p_grid >= 0.5)

posterior2M3 <- probability_of_water(c("L", "W", "W", "L", "W", "W", "W"), priorM2)
ggplot(posterior2M3, aes(x=Pwater, y=Posterior)) +
  geom_line()
```



2M3

Suppose there are two globes, one for Earth and one for Mars. The Earth globe is 70% covered in water. The Mars globe is 100% land. Further suppose that one of these globes — you don't know which — was tossed in the air and produced a "land" observation. Assume that each globe was equally likely to be tossed. Show that the posterior probability that the globe was the Earth, conditional on seeing "land" (Pr (Earth | land)), is 0.23.

Bayes Formula tells us that

$$Pr(Planet|land) = \frac{Pr(land|Planet)Pr(Planet)}{Pr(land)}$$

As $Pr(land)$ is a normalization constant, we can ignore it for a moment. Accordingly,

$$uPr(Earth|land) = 0.3 \cdot 0.5 = 0.15 uPr(Mars|land) = 1 \cdot 0.5 = 0.5$$

where $uPr()$ is unstandardized plausibility. Normalizing it, we get

$$Pr(Earth|land) = \frac{0.15}{0.15 + 0.5} = 0.2308$$

2M4

Suppose you have a deck with only three cards. Each card has two sides, and each side is either black or white. One card has two black sides. The second card has one black and one white side. The third card has two white sides. Now suppose all three cards are placed in a bag and shuffled. Someone reaches into the bag and pulls out a card and places it flat on a table. A black side is shown facing up, but you don't know the color of the side facing down. Show that the probability that the other side is also black is 2/3. Use the counting method (Section 2 of the chapter) to approach this problem. This means counting up the ways that each card could produce the observed data (a black side facing up on the table).

- B|B: B(1) $\rightarrow$ **B(2) : 1**
- B|B: B(2) $\rightarrow$ **B(1) : 1**
- B|W : B $\rightarrow$ W : 1
- W|B : 0
- W|W : 0
- W|W : 0

There are three possible outcomes, given the visible side is black and two out of three lead to a black back side: $\frac{2}{3}$.

2M5

Now suppose there are four cards: B/B, B/W, W/W, and another B/B. Again suppose a card is drawn from the bag and a black side appears face up. Again calculate the probability that the other side is black.

- B|B: B(1) $\rightarrow$ **B(2): 1**
- B|B: B(2) $\rightarrow$ **B(1) : 1**
- B|W : B $\rightarrow$ W : 1
- W|B : 0
- W|W : 0
- W|W : 0
- B|B: B(3) $\rightarrow$ **B(4): 1**
- B|B: B(4) $\rightarrow$ **B(3) : 1**

There are three possible outcomes, given the visible side is black and four out of five lead to a black back side: $\frac{4}{5}$

2M6

Imagine that black ink is heavy, and so cards with black sides are heavier than cards with white sides. As a result, it's less likely that a card with black sides is pulled from the bag. So again assume there are three cards: B/B, B/W, and W/W. After experimenting a number of times, you conclude that for every way to pull the B/B card from the bag, there are 2 ways to pull the B/W card and 3 ways to pull the W/W card. Again suppose that a card is pulled and a black side appears face up. Show that the probability the other side is black is now 0.5. Use the counting method, as before.

- B|B: B(1) $\rightarrow$ **B(2): 1 $\times$ 1 (prior) = 1**
- B|B: B(2) $\rightarrow$ **B(1) : 1 $\times$ 1 (prior) = 1**
- B|W : B $\rightarrow$ W : 1 $\times$ 2 (prior) = 2
- W|B : 0 $\times$ 2 (prior) = 0
- W|W : 0 $\times$ 3 (prior) = 0
- W|W : 0 $\times$ 3 (prior) = 0

Now the counts are two out of four that other side is black, i.e. 0.5.

2M7

Assume again the original card problem, with a single card showing a black side face up. Before looking at the other side, we draw another card from the bag and lay it face up on the table. The face that is shown on the new card is white. Show that the probability that the first card, the one showing a black side, has black on its other side is now 0.75. Use the counting method, if you can. Hint: Treat this like the sequence of globe tosses, counting all the ways to see each observation, for each possible first card.

Possible card sequences, bold means that back side of the first card is black:

- **B(1)|B(2) $\rightarrow$ W|B : 1**
- **B(2)|B(1) $\rightarrow$ W|B : 1**
- B(1)|B(2) $\rightarrow$ B|W : 0
- B(2)|B(1) $\rightarrow$ B|W : 0
- **B(1)|B(2) $\rightarrow$ W(1)|W(2) : 1**
- **B(2)|B(1) $\rightarrow$ W(1)|W(2) : 1**
- **B(1)|B(2) $\rightarrow$ W(2)|W(1) : 1**
- **B(2)|B(1) $\rightarrow$ W(2)|W(1) : 1**

- B|W → B(1)|B(2) : 0
- W|B → B(1)|B(2) : 0
- B|W → B(2)|B(1) : 0
- W|B → B(2)|B(1) : 0
- B|W → W(1)|W(2) : 1
- W|B → W(1)|W(2) : 0
- B|W → W(2)|W(1) : 1
- W|B → W(2)|W(1) : 0
- W(1)|W(2) → W|B : 0
- W(2)|W(1) → W|B : 0
- W(1)|W(2) → B|W : 0
- W(2)|W(1) → B|W : 0
- W(1)|W(2) → B(1)|B(2) : 0
- W(2)|W(1) → B(1)|B(2) : 0
- W(1)|W(2) → B(2)|B(1) : 0
- W(2)|W(1) → B(2)|B(1) : 0

Total of eight possible path, six of them have black back for the first card: 0.75.

2H1

Suppose there are two species of panda bear. Both are equally common in the wild and live in the same places. They look exactly alike and eat the same food, and there is yet no genetic assay capable of telling them apart. They differ however in their family sizes. Species A gives birth to twins 10% of the time, otherwise birthing a single infant. Species B births twins 20% of the time, otherwise birthing singleton infants. Assume these numbers are known with certainty, from many years of field research.

Now suppose you are managing a captive panda breeding program. You have a new female panda of unknown species, and she has just given birth to twins. What is the probability that her next birth will also be twins?

$uPr(A|twins) = Pr(twins|A)*Pr(A) = 0.1*0.5 = 0.05 uPr(B|twins) = Pr(twins|B)*Pr(B) = 0.2*0.5 = 0.1$

After normalization $Pr(A|twins) = 1/3$ and $Pr(B|twins) = 2/3$.

Probability that you will see twins again per species is

$Pr(twins|A, twins) = Pr(A|twings)*Pr(twins|A) = 1/3*0.1 = 0.1/3Pr(twins|B, twins) = Pr(B|twing$

The total probability is $\frac{0.1}{3} + \frac{0.4}{3} = \frac{0.5}{3} \approx 0.167$

2H2

Recall all the facts from the problem above. Now compute the probability that the panda we have is from species A, assuming we have observed only the first birth and that it was twins.

$uPr(A|twins) = Pr(twins|A)*Pr(A) = 0.1*0.5 = 0.05uPr(B|twins) = Pr(twins|B)*Pr(B) = 0.2*0.5 =$

After normalization $Pr(A|twins) = 1/3$

2H3

Continuing on from the previous problem, suppose the same panda mother has a second birth and that it is not twins, but a singleton infant. Compute the posterior probability that this panda is species A.

$uPr(A|twins) = Pr(twins|A)*Pr(A) = 0.1*0.5 = 0.05uPr(B|twins) = Pr(twins|B)*Pr(B) = 0.2*0.5 =$

After normalization $Pr(A|twins) = 1/3$ and $Pr(B|twins) = 2/3$.

Given that next birth is a singleton:

$uPr(A|twins, singleton) = Pr(A|twins)*(1-Pr(twins|A)) = 1/3 \cdot 0.9 = 0.9/3uPr(B|twins, singleton) =$

After normalization $Pr(A|twins, singleton) = 0.36$ (and $Pr(B|twins, singleton) = 0.64$).

2H4

A common boast of Bayesian statisticians is that Bayesian inference makes it easy to use all of the data, even if the data are of different types.

So suppose now that a veterinarian comes along who has a new genetic test that she claims can identify the species of our mother panda. But the test, like

all tests, is imperfect. This is the information you have about the test: • The probability it correctly identifies a species A panda is 0.8. • The probability it correctly identifies a species B panda is 0.65.

The vet administers the test to your panda and tells you that the test is positive for species A. First ignore your previous information from the births and compute the posterior probability that your panda is species A. Then redo your calculation, now using the birth data as well.

$uPr(A|test) = Pr(test|A)*Pr(A) = 0.8*0.5 uPr(B|test) = (1-Pr(test|B))*Pr(B) = (1-0.65)*0.5 = 0.35*0.5$

After normalization $Pr(A|test) \approx 0.7$ and $Pr(B|test) \approx 0.3$.

Given the twins, this becomes

$uPr(A|test, twins) = Pr(A|test)*Pr(twins|A) = 0.7*0.1 = 0.07 uPr(B|test, twins) = Pr(B|test)*Pr(twins|B) = 0.3$

After normalization $Pr(A|test, twins) \approx 0.533$ and $Pr(B|test, twins) \approx 0.467$.

Given the singleton, this becomes

$uPr(A|test, twins, singleton) = Pr(A|test, twins)*(1-Pr(twins|A)) = 0.533*0.9 = 0.48 uPr(B|test, twins, singleton$

After normalization $Pr(A|test, twins, singleton) \approx 0.5625$ and $Pr(B|test, twins, singleton) \approx 0.4375$.

Chapter 4

Solutions for Chapter 3

```
library(tidyverse)
library(rethinking)
```

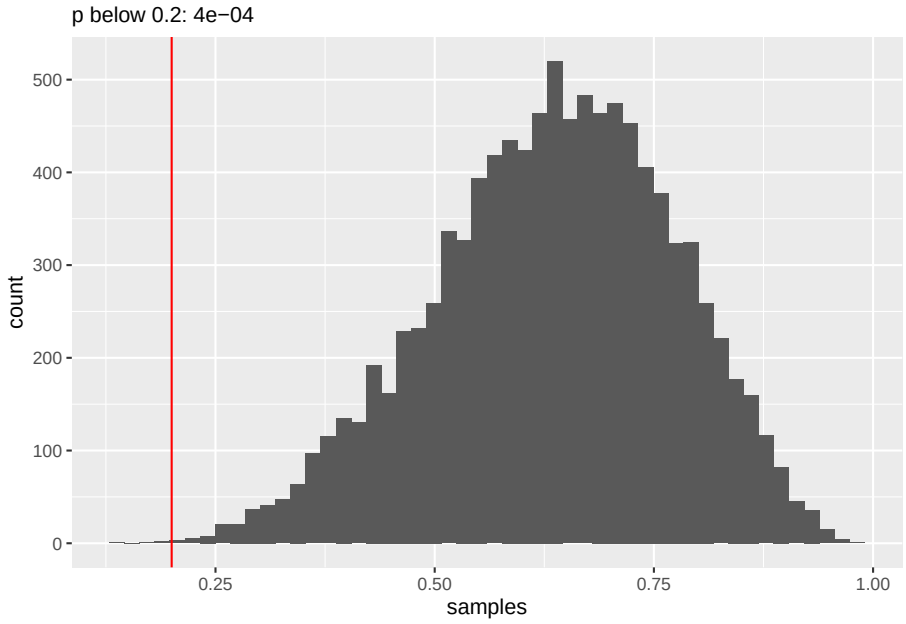
4.0.0.1 Initialization code for easy exercises

```
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?

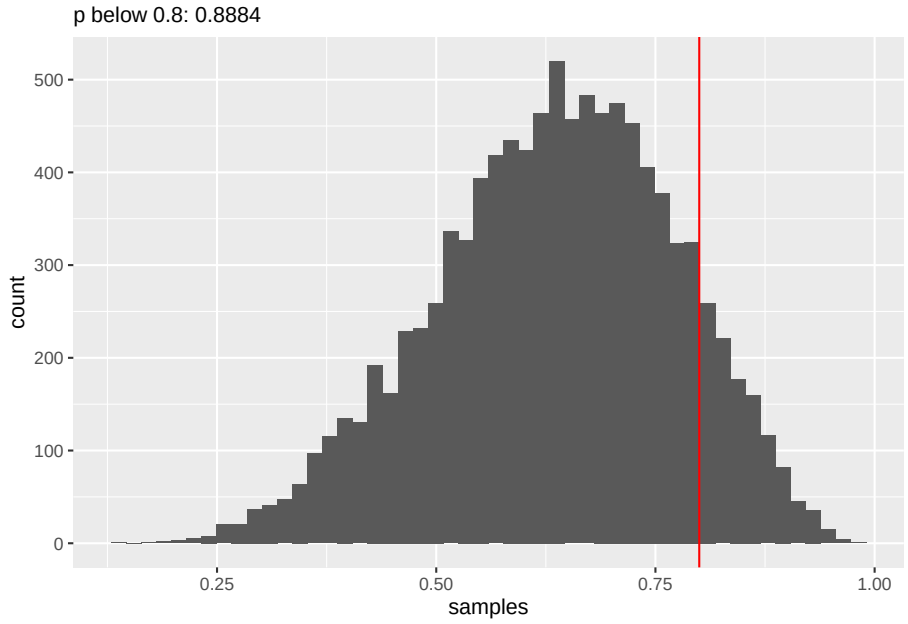
```
ggplot(data=NULL, aes(x=samples)) +
  geom_histogram(bins=50) +
  geom_vline(xintercept = 0.2, color="red") +
  labs(subtitle = glue::glue("p below 0.2: {mean(samples < 0.2)}"))
```



3E2

How much posterior probability lies below p = 0.8?

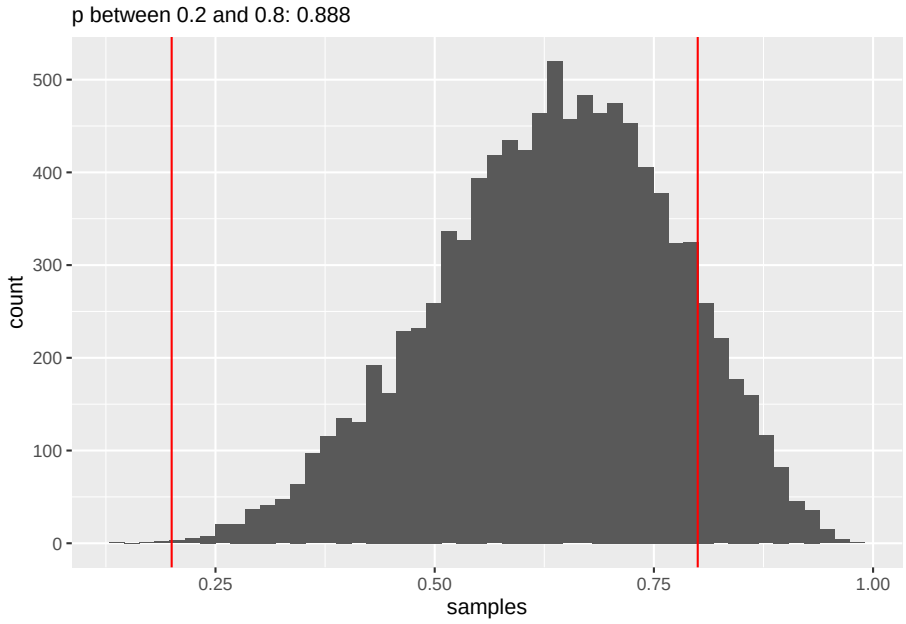
```
ggplot(data=NULL, aes(x=samples)) +
  geom_histogram(bins=50) +
  geom_vline(xintercept = 0.8, color="red") +
  labs(subtitle = glue::glue("p below 0.8: {mean(samples < 0.8)}"))
```



3E3

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

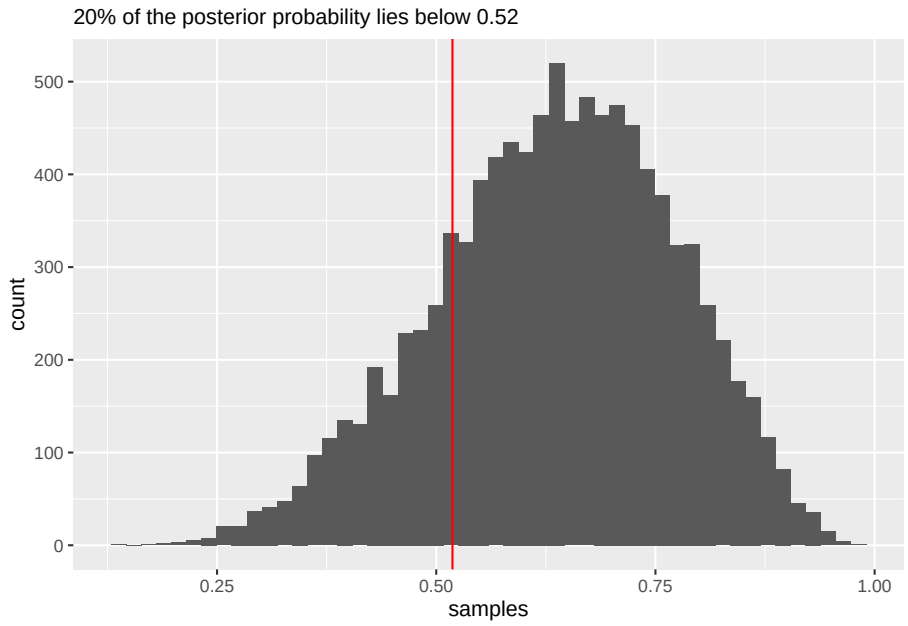
```
ggplot(data=NULL, aes(x=samples)) +
  geom_histogram(bins=50) +
  geom_vline(xintercept = c(0.2, 0.8), color="red") +
  labs(subtitle = glue::glue("p between 0.2 and 0.8: {mean(samples < 0.8 & samples > 0.2)}"))
```



3E4

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

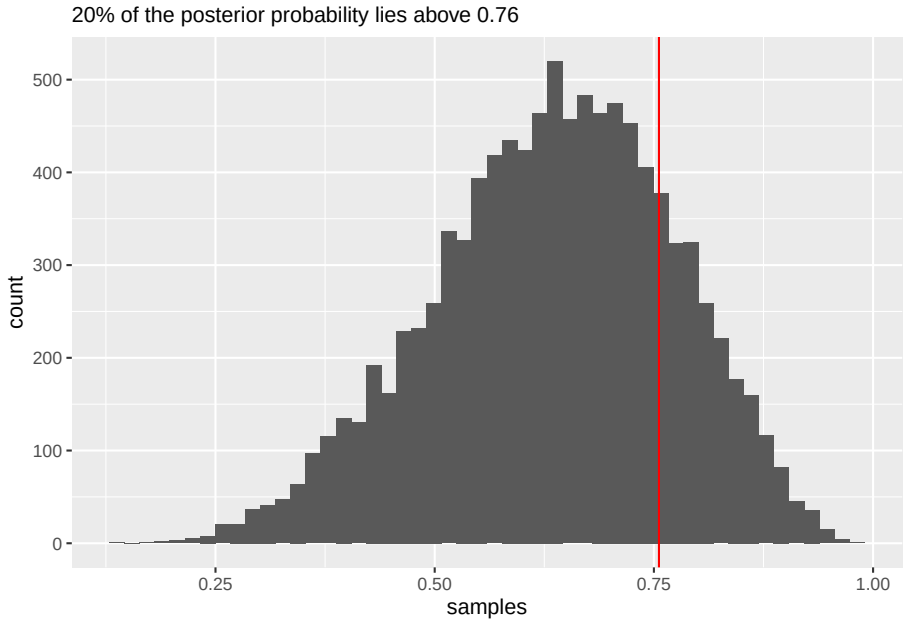
```
q20 <- quantile(samples, 0.2)
ggplot(data=NULL, aes(x=samples)) +
  geom_histogram(bins=50) +
  geom_vline(xintercept = q20, color="red") +
  labs(subtitle = glue::glue("20% of the posterior probability lies below {round(q20, 1
```



3E5

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

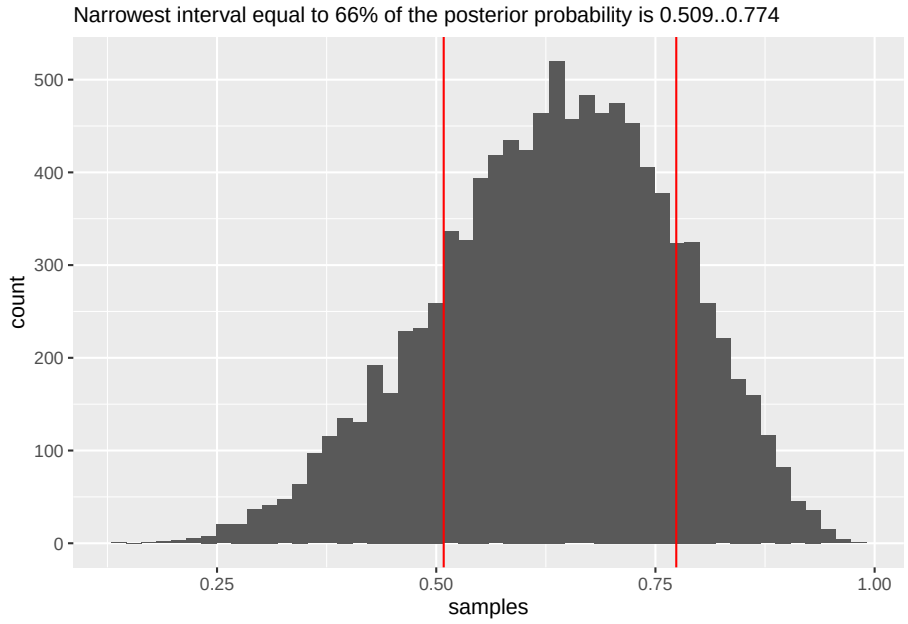
```
q80 <- quantile(samples, 0.8)
ggplot(data=NULL, aes(x=samples)) +
  geom_histogram(bins=50) +
  geom_vline(xintercept = q80, color="red") +
  labs(subtitle = glue::glue("20% of the posterior probability lies above {round(q80, 2)}"))
```



3E6

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

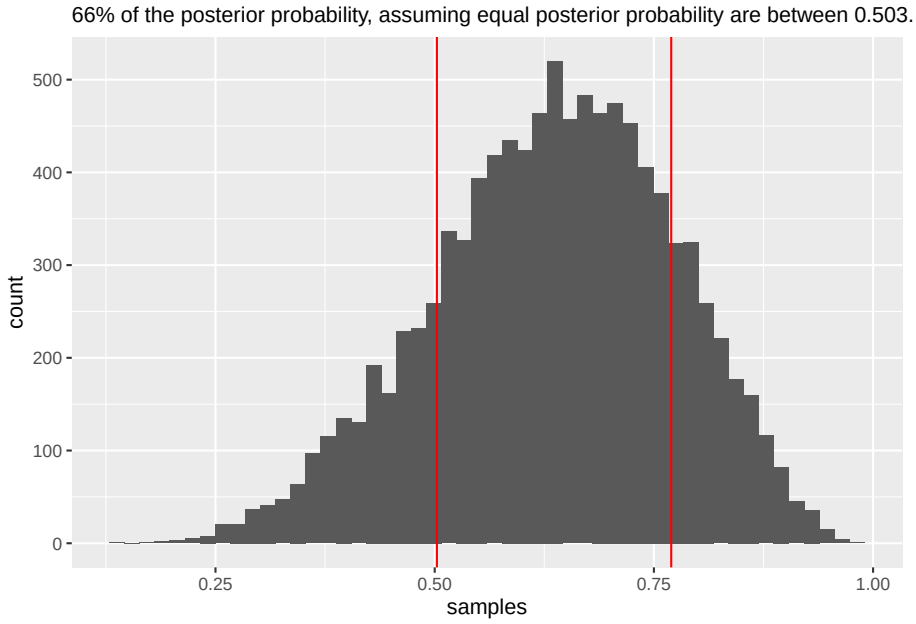
```
hpdi66 <- HPDI(samples, 0.66)
ggplot(data=NULL, aes(x=samples)) +
  geom_histogram(bins=50) +
  geom_vline(xintercept = hpdi66, color="red") +
  labs(subtitle = glue::glue("Narrowest interval equal to 66% of the posterior probabil
```



3E7

Which values of p contain 66% of the posterior probability, assuming equal posterior probability both below and above the interval?

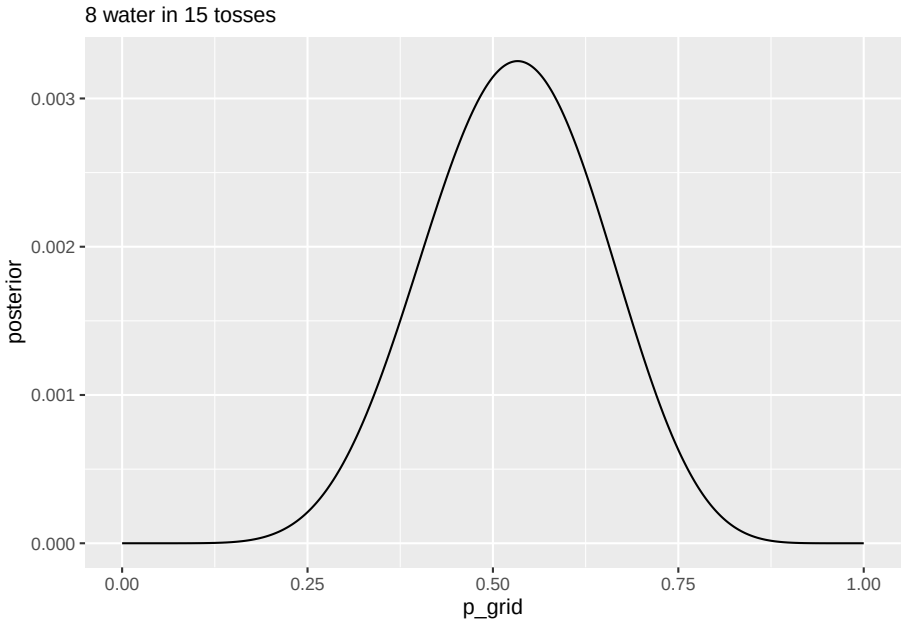
```
pi66 <- PI(samples, 0.66)
ggplot(data=NULL, aes(x=samples)) +
  geom_histogram(bins=50) +
  geom_vline(xintercept = pi66, color="red") +
  labs(subtitle = glue::glue("66% of the posterior probability, assuming equal posterior probabil
```



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.

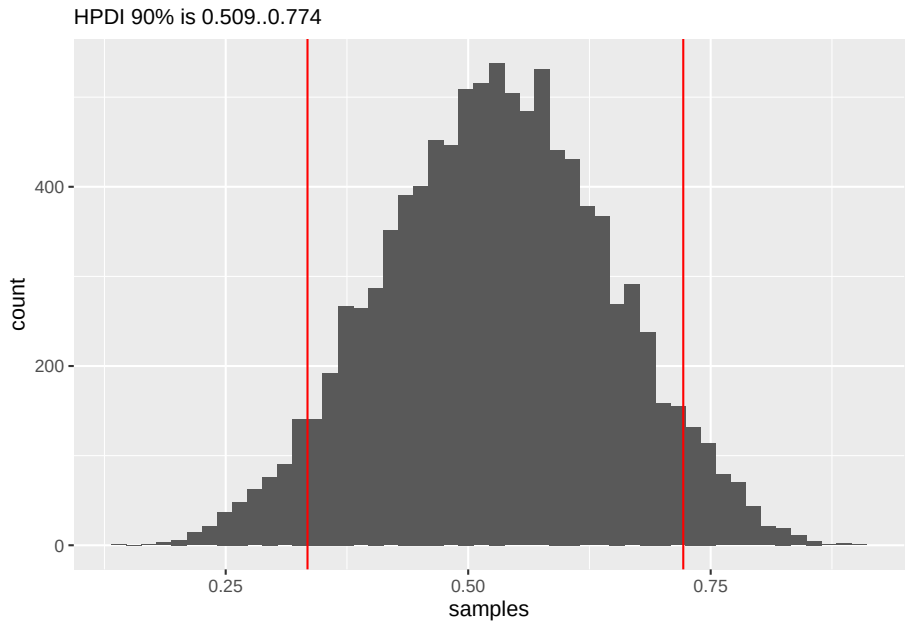
```
p_grid <- seq(from=0, to=1, length.out=1000)
prior <- rep(1,1000)
likelihood <- dbinom(8, size=15, prob=p_grid)
posterior <- likelihood * prior
posterior <- posterior / sum(posterior)

ggplot(data=NULL, aes(x=p_grid, y=posterior)) +
  geom_line() +
  labs(subtitle="8 water in 15 tosses")
```



3M2

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

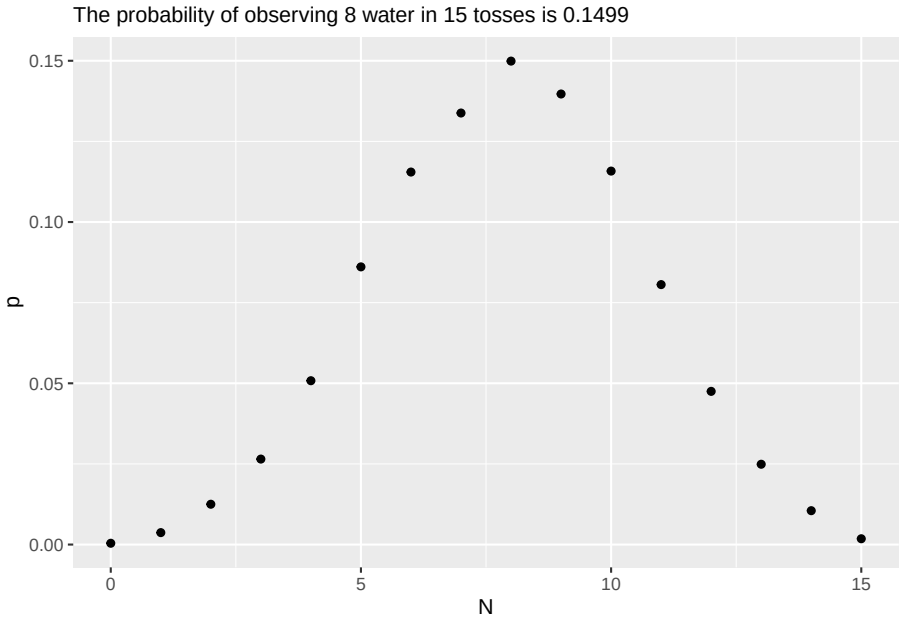
```
set.seed(100)
samples <- sample(p_grid, prob=posterior, size=1e4, replace=TRUE)

hpdi90 <- HPDI(samples, 0.9)
ggplot(data=NULL, aes(x=samples)) +
  geom_histogram(bins=50) +
  geom_vline(xintercept = hpdi90, color="red") +
  labs(subtitle = glue::glue("HPDI 90% is {round(hpdi66[1], 3)}..{round(hpdi66[2], 3)}"))
```



3M3

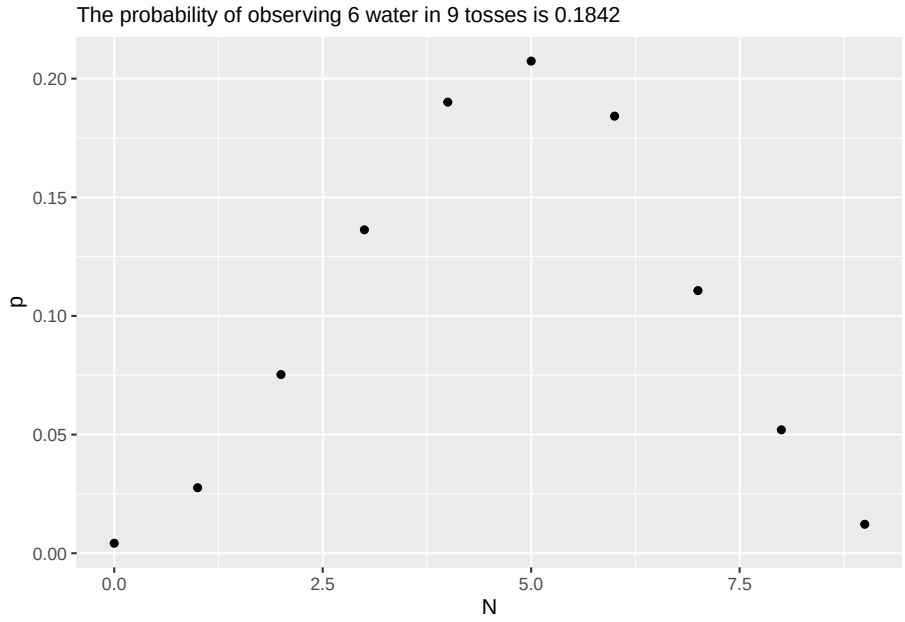
Construct a posterior predictive check for this model and data. This means simulate the distribution of samples, averaging over the posterior uncertainty in p . What is the probability of observing 8 water in 15 tosses?

```
predicted_water8_15 <-
  tibble(N = rbinom(length(samples), size=15, prob=samples)) %>%
  group_by(N) %>%
  summarize(count = n(), .groups = 'drop')  %>%
  ungroup() %>%
  mutate(p = count / sum(count))

p8 <-
  predicted_water8_15 %>%
  filter(N == 8) %>%
  pull(p)

ggplot(predicted_water8_15, aes(x=N, y= p)) +
    geom_point() +
  labs(subtitle = glue::glue("The probability of observing 8 water in 15 tosses is {p8}
```



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

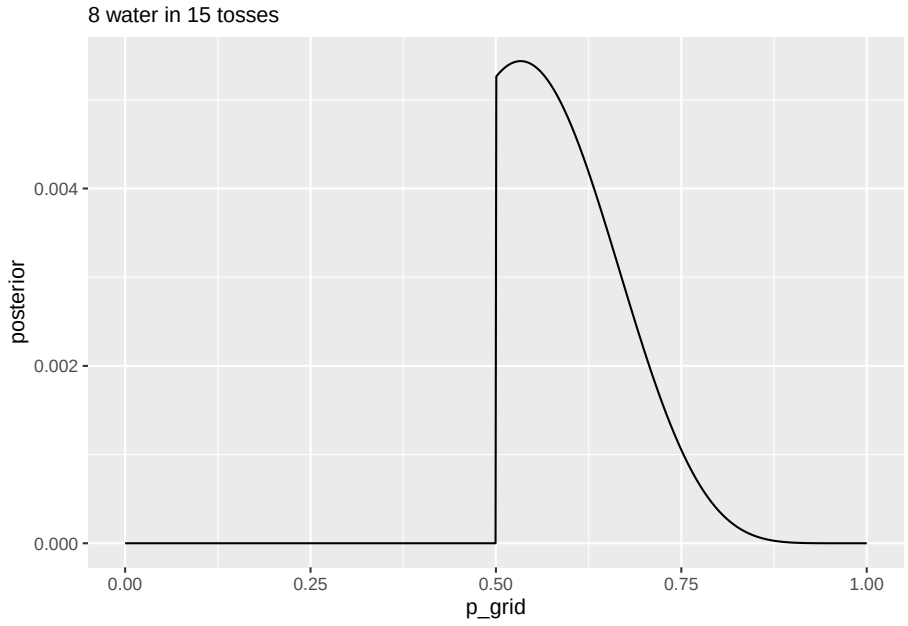
```
predicted_water6_9 <-
  tibble(N = rbinom(length(samples), size=9, prob=samples)) %>%
  group_by(N) %>%
  summarize(count = n(), .groups = 'drop')  %>%
  ungroup() %>%
  mutate(p = count / sum(count))

p6 <-
  predicted_water6_9 %>%
  filter(N == 6) %>%
  pull(p)

ggplot(predicted_water6_9, aes(x=N, y= p)) +
    geom_point() +
  labs(subtitle = glue::glue("The probability of observing 6 water in 9 tosses is {p6}"))
```



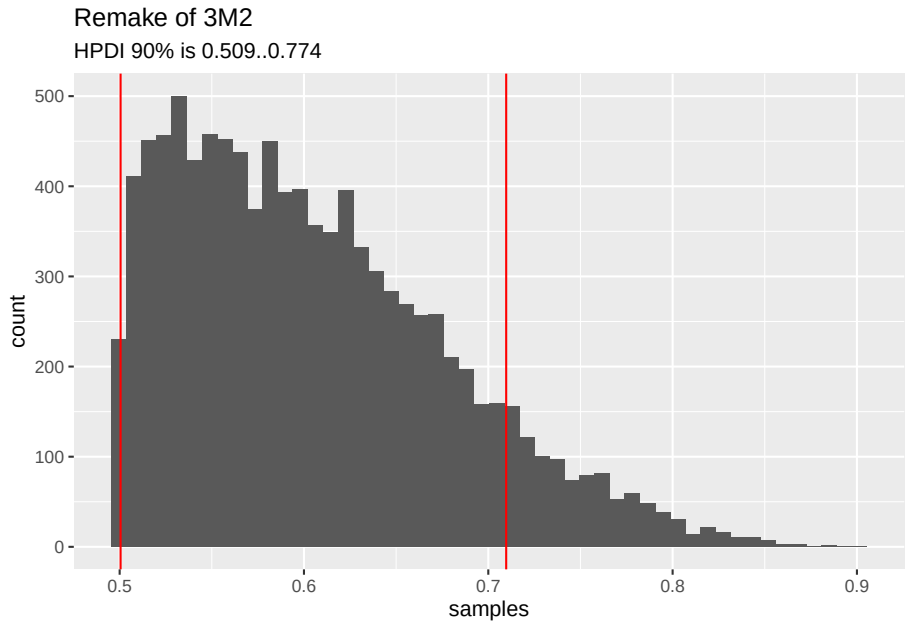
3M5

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.

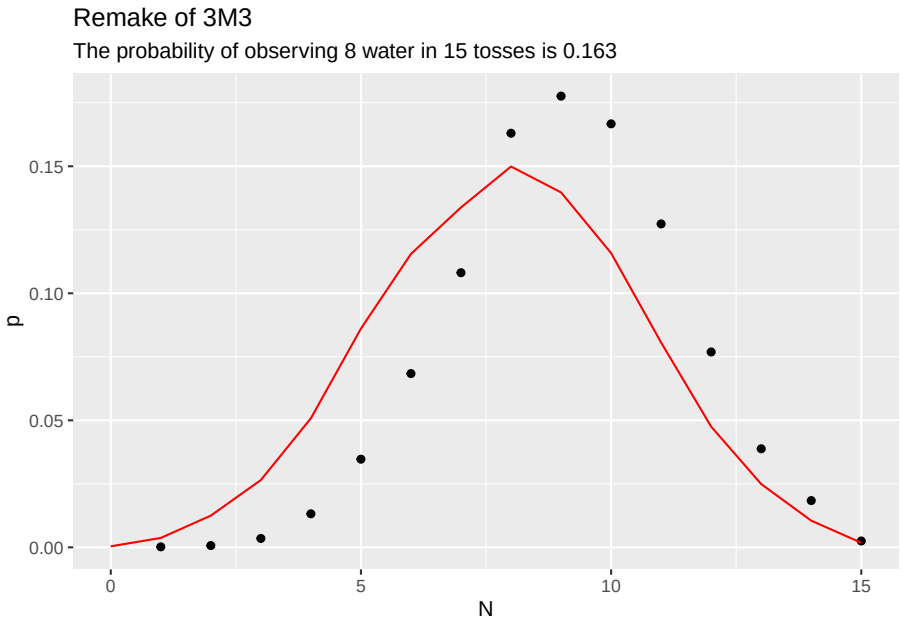
```
p_grid <- seq(from=0, to=1, length.out=1000)
prior <- as.numeric(p_grid >= 0.5)
likelihood <- dbinom(8, size=15, prob=p_grid)
posterior <- likelihood * prior
posterior <- posterior / sum(posterior)

ggplot(data=NULL, aes(x=p_grid, y=posterior)) +
  geom_line() +
  labs(subtitle="8 water in 15 tosses")
```



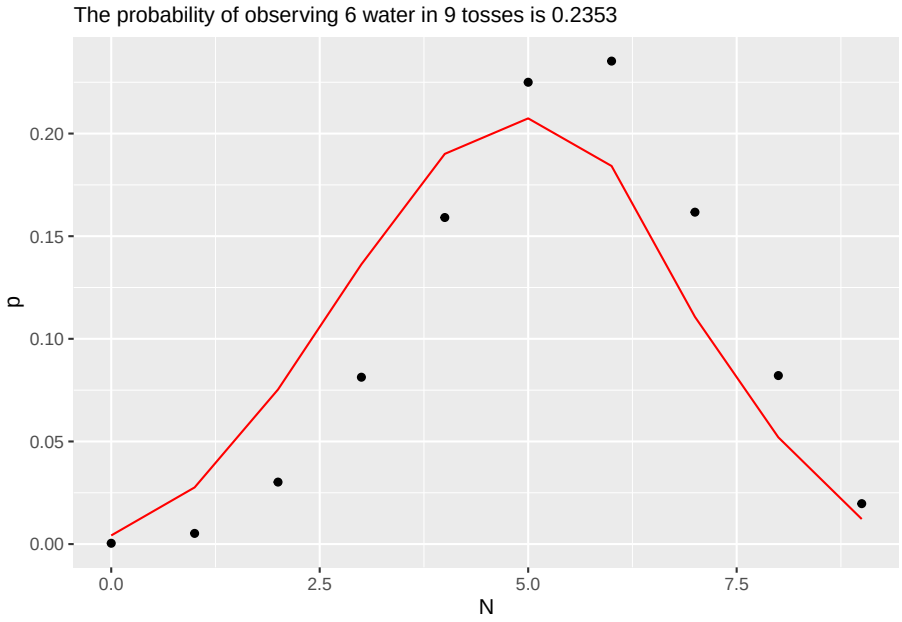

```
set.seed(100)
samples <- sample(p_grid, prob=posterior, size=1e4, replace=TRUE)

hpdi90 <- HPDI(samples, 0.9)
ggplot(data=NULL, aes(x=samples)) +
  geom_histogram(bins=50) +
  geom_vline(xintercept = hpdi90, color="red") +
  labs(title = "Remake of 3M2",
       subtitle = glue::glue("HPDI 90% is {round(hpdi66[1], 3)}..{round(hpdi66[2], 3)}"))
```




```
predicted_water8_15_remake <-
  tibble(N = rbinom(length(samples), size=15, prob=samples)) %>%
  group_by(N) %>%
  summarize(count = n(), .groups = 'drop')  %>%
  ungroup() %>%
  mutate(p = count / sum(count))

p8 <-
  predicted_water8_15_remake %>%
  filter(N == 8) %>%
  pull(p)

ggplot(predicted_water8_15_remake, aes(x=N, y= p)) +
  geom_point() +
  geom_line(data=predicted_water8_15, color="red") +
  labs(title = "Remake of 3M3",
       subtitle = glue::glue("The probability of observing 8 water in 15 tosses is {p8}
```




```
predicted_water6_9_remake <-
  tibble(N = rbinom(length(samples), size=9, prob=samples)) %>%
  group_by(N) %>%
  summarize(count = n(), .groups = 'drop')  %>%
  ungroup() %>%
  mutate(p = count / sum(count))

p6 <-
  predicted_water6_9_remake %>%
  filter(N == 6) %>%
  pull(p)

ggplot(predicted_water6_9_remake, aes(x=N, y= p)) +
  geom_point() +
  geom_line(data=predicted_water6_9, color="red") +
  labs(subtitle = glue::glue("The probability of observing 6 water in 9 tosses is {p6}"))
```

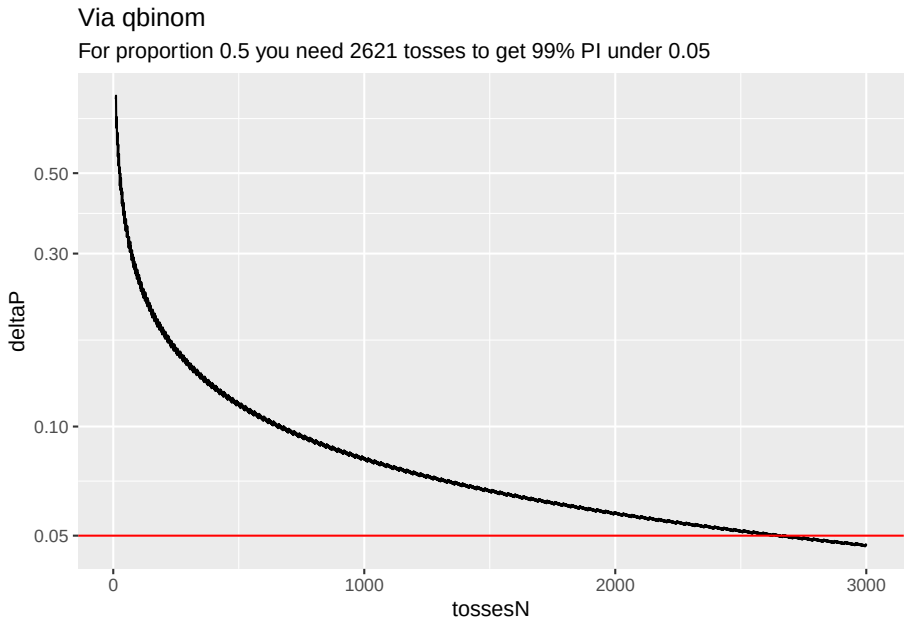


3M6

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?

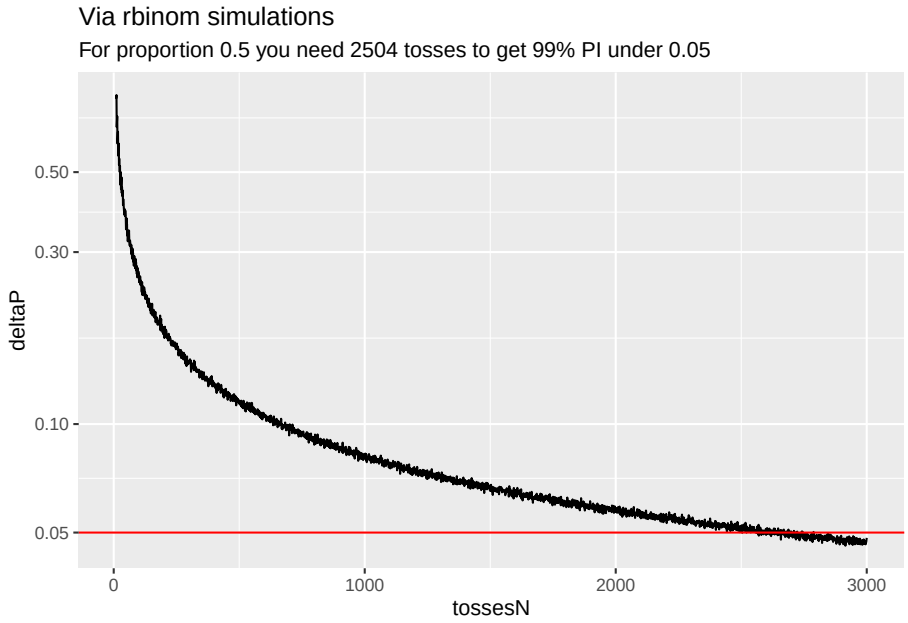
```
true_proportion <- 0.5
toss_quantiles <-
  tibble(tossesN = seq(10, 3000, by=1)) %>%
  mutate(lowerN = qbinom(0.005, tossesN, true_proportion),
         upperN = qbinom(0.995, tossesN, true_proportion),
         lowerP = lowerN / tossesN,
         upperP = upperN / tossesN,
         deltaP = upperP - lowerP)

minimal_N <-
  toss_quantiles %>%
  filter(deltaP < 0.05) %>%
  slice(1) %>%
  pull(tossesN)

ggplot(toss_quantiles, aes(x=tossesN, y=deltaP)) +
```

```
  geom_line() +
  geom_hline(yintercept = 0.05, color="red") +
  scale_y_log10() +
  labs(title = "Via qbinom",
       subtitle = glue::glue("For proportion {true_proportion} you need {minimal_N} tosses to get
```




```
simulate_and_compute_PI99 <- function(tossN, true_proportion){
  simulated_N <- rbinom(1e4, tossN, true_proportion)
  simulated_P <- simulated_N / tossN
  diff(rethinking::PI(simulated_P, prob=0.99))
}

true_proportion <- 0.5
toss_sims <-
  tibble(tossesN = seq(10, 3000, by=1)) %>%
  rowwise() %>%
  mutate(deltaP =simulate_and_compute_PI99(tossesN, true_proportion))

minimal_N <-
  toss_sims %>%
  filter(deltaP < 0.05) %>%
  slice(1) %>%
  pull(tossesN)
```

```
ggplot(toss_sims, aes(x=tossesN, y=deltaP)) +
  geom_line() +
  geom_hline(yintercept = 0.05, color="red") +
  scale_y_log10() +
  labs(title = "Via rbinom simulations",
       subtitle = glue::glue("For proportion {true_proportion} you need {minimal_N} tos
```



Hard

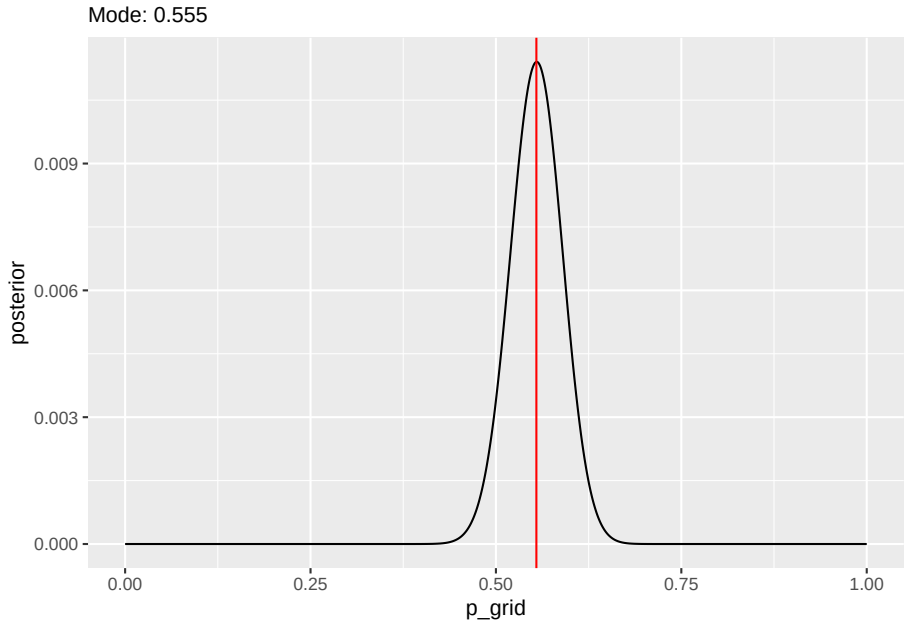
```
data(homeworkch3)
```

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?

```
birth <- c(birth1, birth2)
birthN <- length(birth)
boyN <- sum(birth)
```

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

boyP <- p_grid[which.max(posterior)]

ggplot(data=NULL, aes(x=p_grid, y=posterior)) +
  geom_line()+
  geom_vline(xintercept = boyP, color="red") +
  labs(subtitle = glue::glue("Mode: {round(boyP, 3)}"))
```

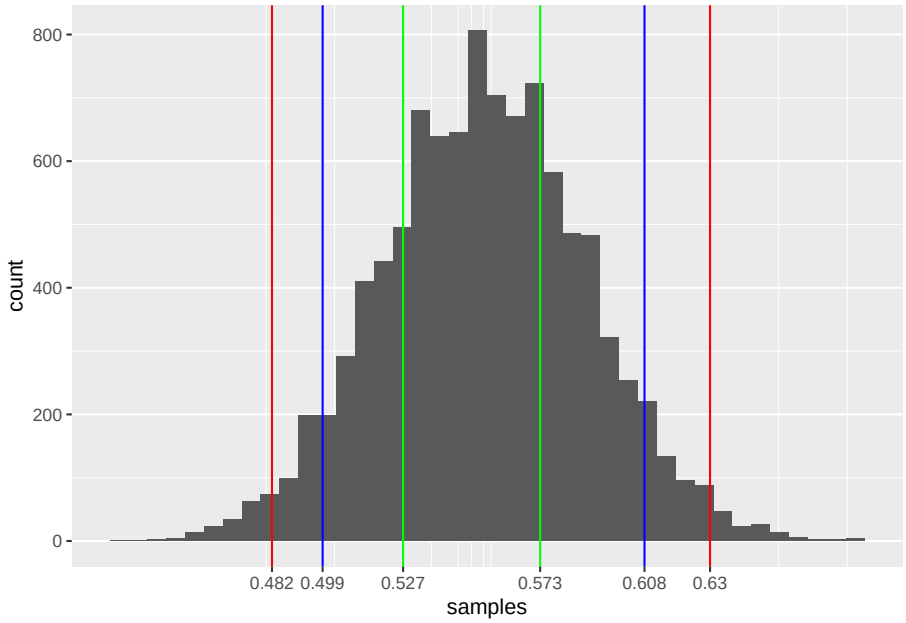


3H2

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.

```
set.seed(100)
samples <- sample(p_grid, prob=posterior, size=1e4, replace=TRUE)

HPDI50 <- rethinking::HPDI(samples, prob = 0.5)
```

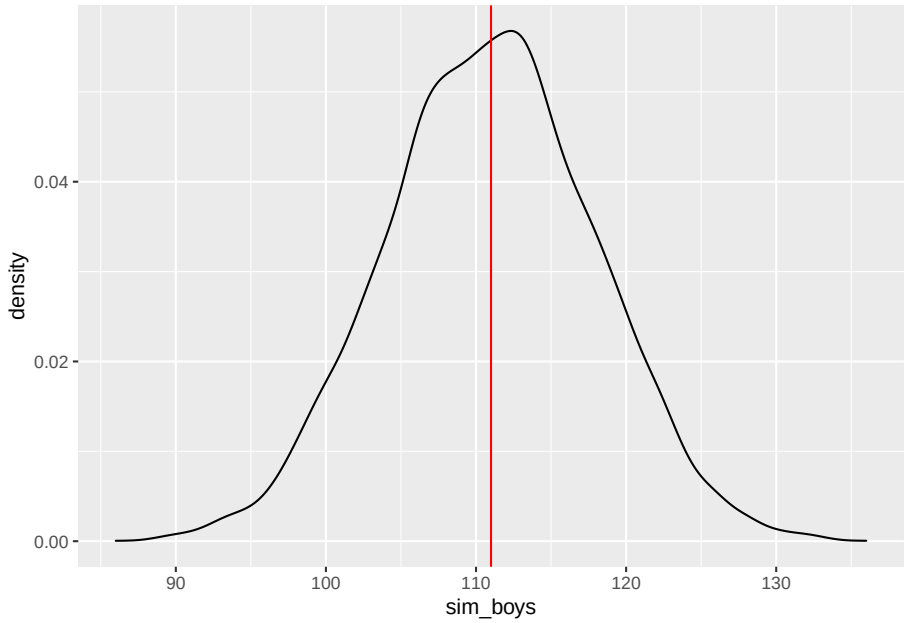
```
HPDI89 <- rethinking::HPDI(samples, prob = 0.89)
HPDI97 <- rethinking::HPDI(samples, prob = 0.97)

all_HPDI <- c(HPDI50, HPDI89, HPDI97)
names(all_HPDI) <- NULL


ggplot(data=NULL, aes(x=samples)) +
  geom_histogram(bins=40) +
  geom_vline(xintercept = HPDI50, color="green") +
  geom_vline(xintercept = HPDI89, color="blue") +
  geom_vline(xintercept = HPDI97, color="red") +
  scale_x_continuous(breaks = all_HPDI, labels = round(all_HPDI, 3))
```



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?

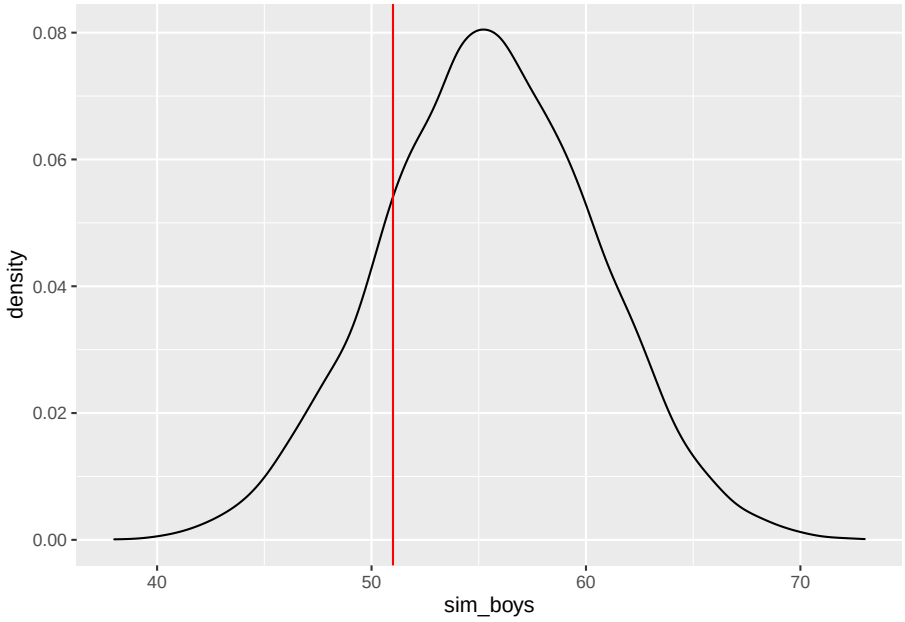
```
sim_boys <- rbinom(1e4, 200, boyN / birthN)

ggplot(data=NULL, aes(x=sim_boys)) +
  geom_density() +
  geom_vline(xintercept = boyN, color="red")
```



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?

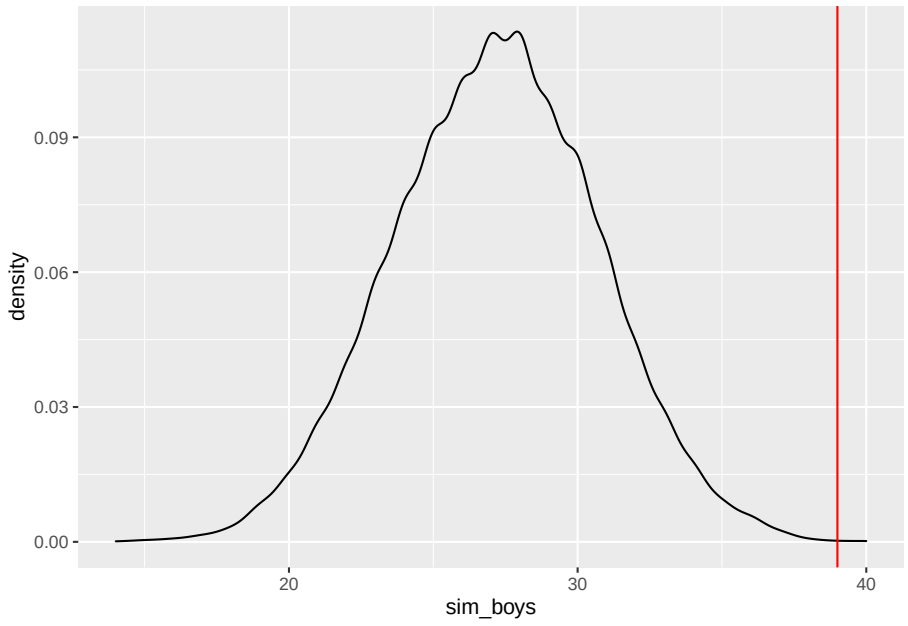
```
sim_boys <- rbinom(1e4, 100, boyN / birthN)

ggplot(data=NULL, aes(x=sim_boys)) +
  geom_density() +
  geom_vline(xintercept = sum(birth1), color="red")
```



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?

```
birth_after_girl <- birth2[birth1 == 0]
sim_boys <- rbinom(1e4, length(birth_after_girl), boyN / birthN)

ggplot(data=NULL, aes(x=sim_boys)) +
  geom_density() +
  geom_vline(xintercept = sum(birth_after_girl), color="red")
```


density
0.09
0.06
0.03
0.00
20
30
40
sim_boys