An Introduction to Bayesian Reasoning and Methods

Kevin Ross

2022-02-19

Contents

Preface

Statistics is the science of learning from data. Statistics involves

- Asking questions
- Formulating conjectures
- Designing studies
- Collecting data
- Wrangling data
- Summarizing data
- Visualizing data
- *Analyzing data*
- *Developing models*
- *Drawing conclusions*
- *Communicating results*

We will assume some familiarity with many of these aspects, and we will focus on the items in italics. That is, we will focus on **statistical inference**, the process of using data analysis to draw conclusions about a population or process beyond the existing data. "Traditional" hypothesis tests and confidence intervals that you are familiar with are components of *"frequestist" statistics.* This book will introduce aspects of *"Bayesian" statistics.* We will focus on analyzing data, developing models, drawing conclusions, and communicating results from a Bayesian perspective. We will also discuss some similarities and differences between frequentist and Bayesian approaches, and some advantages and disadvantages of each approach.

We want to make clear from the start: Bayesian versus frequentist is NOT a question of "right versus wrong". Both Bayesian and frequentist are valid approaches to statistical analyses, each with advantages and disadvantages. We'll address some of the issues along the way. But at no point in your career do you need to make a definitive decision to be a Bayesian or a frequentist; a good modern statistician is probably a bit of both.

While our focus will be on statistical inference, remember that the other parts of Statistics are equally important, if not more important. In particular, any statistical analysis is only as good as the data upon which it is based.

The exercises in this book are used to both motivate new topics and to help you practice your understanding of the material. You should attempt the exercises on your own before reading the solutions. To encourage you to do so, the solutions have been hidden. You can reveal the solution by clicking on the **Show/hide solution** button.

Show/hide solution

Here is where a solution would be, but be sure to think about the problem on your own first!

(Careful: in your browser, the triangle for the Show/hide solution button might be close to the back button, so clicking on Show/hide might take you to the previous page. To avoid this, click on the words **Show/hide**.)

Chapter 1

Introductory Example

Statistics is the science of learning from data. But what is "Bayesian" statistics? This chapter provides a relatively simple and brief example of a Bayesian statistical analysis. As you work through the example, think about: What aspects are familiar? What features are new or different? Think big picture for now; we'll fill in lots of details later.

Example 1.1. Suppose we're interested in **the proportion of all current Cal Poly students who have ever read at least one book in the *Harry Potter* series**. We'll refer to this proportion as the "population proportion".

1. What are some challenges to computing the population proportion? How could we *estimate* it?

2. What are the *possible* values of the population proportion?

3. Which one of the following do you think is the *most plausible* value of the population proportion? Record your value in the plot on the board.

   0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

4. Sketch the plot of guesses here. What seems to be the consensus? What do we think are the most plausible values of the population proportion? Somewhat plausible? Not plausible?

5. The plot just shows our *guesses* for the population proportion. How could we estimate the actual population proportion based on data?

6. We will treat the roughly 30 students in our class as a random sample from the population of current Cal Poly students. But before collecting the data, let's consider what *might* happen.

Suppose that the actual population proportion is 0.5. That is, suppose that 50% of current Cal Poly students have read at least one Harry Potter book. How many students in a random sample of 30 students would you expect to have read at least one Harry Potter book? Would it necessarily be 15 students? How could you use a coin to *simulate* how many students in a random sample of 30 students *might* have read at least one Harry Potter book?

7. Now suppose the actual population proportion is 0.1. How would the previous part change?

8. Using your choice for the most plausible value of the population proportion, simulate how many students in a random sample of 30 students *might* have read at least one Harry Potter book. Repeat to get a few hypothetical samples, using your guess for the most plausible value of the population proportion, and record your results in the plot. (Here is one applet you can use.)

9. Why are there more dots corresponding to a proportion of 0.4 than to a proportion of 0.9?

10. How could we get an even clearer picture of what *might* happen?

11. Sketch the plot that we created. The plot illustrates two sources of uncertainty or variability. What are these two sources?

12. So far, everything we've considered is what *might* happen in a class of 30 students. Now let's see what is actually true for our class. What proportion of students *in the class* have read at least one Harry Potter book? Is the proportion of *all current Cal Poly students* who have read at least one Harry Potter book necessarily equal to the sample proportion?

13. Remember that we started with *guesses* about which values of the population proportion were more plausible than others, and we used these guesses to get a picture of what might happen in samples. How can we reconsider the plausibility of the possible values of the population proportion in light of the sample data that we actually observed?

14. Given the observed sample proportion, what can we say about the plausible values of the *population* proportion? How has our assessment of plausibility changed from before observing the sample data?

15. What elements of the analysis are similar to the kinds of statistical analysis you have done before? What elements are new or different?

Solution. to Example 1.1

Show/hide solution

1. It would be extremely challenging to survey *all* Cal Poly students. Even if we were able to obtain contact information for all students, many students would not respond to the survey. Instead, we can take a sample of Cal Poly students, collect data for students in the sample, and use the proportion of students in the sample who have read at least one Harry Potter book as a starting point to estimate the population proportion.

2. The population proportion could possibly be any value in the interval [0, 1]. Between 0% and 100% of current Cal Poly students have read at least one Harry Potter book.

3. There is no right answer for what you think is most plausible. Maybe you have a lot of friends that have read at least one Harry Potter book, so you might think the population proportion is 0.8. Maybe you don't know anyone who has read at least one Harry Potter book, so you might think the population proportion is 0.1. Maybe you have no idea and you just guess that the population proportion is 0.5. Everyone has their own background information which influences their initial assessment of plausibility.

4. Results for the class will vary, but Figure 1.1 shows an example. The consensus for the class in Figure 1.1 is that values of 0.3, 0.4, and 0.5 are most plausible, 0.2 and 0.6 less so, and values close to 0 or 1 are not plausible.

5. We could use our class of 30 students as a sample, ask each student if they have read at least one Harry Potter book, and find the proportion of students in our class who have read at least one Harry Potter book.

6. If the actual population proportion is 0.5, we would expect around 15 students in a random sample of 30 students to have read at least one Harry Potter book. However, there would be natural sample-to-sample variability. To get a sense of this variability we could:

    - Flip a fair coin. Heads represents a student who has read at least one Harry Potter book; Tails, not.
    - A set of 30 flips represents one hypothetical random sample of 30 students.
    - The number of the 30 flips that land on Heads represents one hypothetical value of the number of students in a random sample of 30 students who have read at least one Harry Potter book.
    - Repeat the above process to get many hypothetical values of the number of students in a random sample of 30 students who have read at least one Harry Potter book, *assuming that the population proportion is 0.5.*

7. If the population proportion is 0.1 we would expect around 3 students in a random sample of 30 students to have read at least one Harry Potter

book. Again, there would be natural sample-to-sample variability. To get a sense of this variability we could:

- Roll a fair 10-sided die. A roll of 1 represents a student who has read at least one Harry Potter book; all other rolls, not.
- A set of 30 rolls represents one hypothetical random sample of 30 students.
- The number of the 30 rolls that land on 1 represents one hypothetical value of the number of students in a random sample of 30 students who have read at least one Harry Potter book.
- Repeat the above process to get many hypothetical values of the number of students in a random sample of 30 students who have read at least one Harry Potter book, *assuming that the population proportion is 0.1.*

8. Figure 1.2 shows the number of students who have read at least one Harry Potter book in 5 hypothetical samples assuming the population proportion is 0.5, and in 5 hypothetical samples assuming the population proportion is 0.1.

9. Results for the class will vary. In the scenario in Figure 1.1, a value of 0.4 was initially more plausible than a value of 0.9. There were more students who thought 0.4 was the most plausible value than 0.9. So the value 0.4 gets more "weight" in the simulation than 0.9. The plot on the left in Figure 1.3 reflects the results of a simulation where every student who plotted a dot in Figure 1.1 simulates 5 random samples of size 30, using their guess for the population proportion.

10. Repeat the simulation process to get many hypothetical samples for each value for the population proportion, reflecting differences in initial plausibility. Imagine each student simulated 10000 samples instead of 5. The plot on the right in Figure 1.3 displays the results.

11. The plot illustrates natural sample-to-sample variability in the sample proportion for a given value of the population proportion. The plot also illustrates *the uncertainty in the value of the population proportion.* That is, the population proportion has a *distribution of values determined by our relative initial plausibilities.*

12. Results will vary. We'll assume that 9 out of 30 students have read at least one Harry Potter book, for a sample proportion of $9/30 = 0.3$. While we hope that 0.3 is close to the proportion of all current Cal Poly students who have read at least one Harry Potter book, because of natural sample-to-sample variability the sample proportion is not necessarily equal to the population proportion.

13. The simulation demonstrated what might happen in a sample of size 30. Now we can zoom in on what actually did happen. Among the samples

in the simulation that resulted in 9 students having read a Harry Potter book, what were the corresponding population proportions?

14. Figure 1.4 displays the results based on the smaller scale simulation in the plot on the left in Figure 1.3, in which every initial guess for the sample proportion generated five hypothetical samples of size 30. Now we focus on samples that resulted in a sample proportion of 9/30, the observed sample proportion. The middle plot displays the population proportions correspoding to samples with a sample proportion of 9/30. The distribution of all the dots in the middle plot illustrates our initial plausibility. The plot on the right displays only the green dots, which correspond to samples with a sample proportion of 9/30. The distribution in the plot on the right reflects a reassessment of the plausibilities of possible values of the population proportion given the observed sample proportion of 9/30 and the simulation results. Among the simulated samples that resulted in a sample proportion of 9/30, the population proportion was much more likely to be 0.3 than to be 0.5.

    Figure 1.5 displays the same analsyis based on the full simulation from the plot on the right in Figure 1.3. The plot on the right in Figure 1.5 compares the initial plausibilities to the plausibilities revised upon observing a sample proportion of 9/30. Initially, the values 0.3, 0.4, and 0.5 were roughly equally plausible, and more plausible than any other value. After observing a sample proportion of 9/30:

    - 0.3 is the most plausible value of the population proportion
    - 0.3 is about two times more plausible than the next most plausible value, 0.4
    - 0.3 and 0.4 together account for the bulk of plausibility.
    - Initially, 0.5 was much more plausible than 0.2, but given the observed data 0.2 is now more plausible than 0.5 (though neither is very plausible)

15. Familiar elements include:

    - using sample statistics to make inference about population parameters
    - reflecting sample-to-sample variability of statistics, for a given value of the population parameter
    - using simulation to analyze data and understand ideas

    New elements include:

    - Quantifying the uncertainty of the population proportion with relative plausibilities of possible values
    - Treating the population proportion as a variable with a distribution determined by the relative plausibilities

- Conditioning on the observed data and revising our assessment of plausibilities

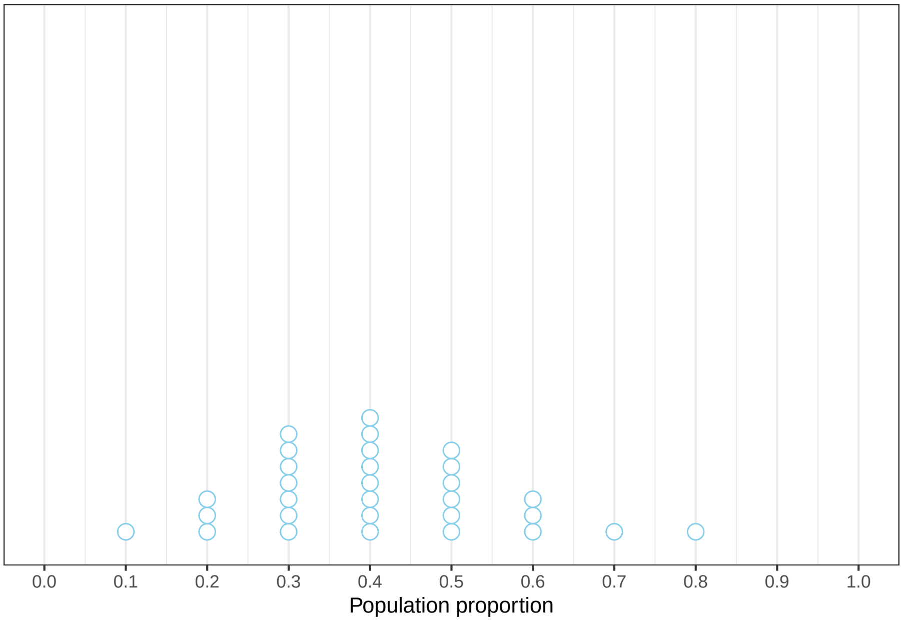


Figure 1.1: Example plot of the guesses of 30 students for the most plausible value of the proportion of current Cal Poly students who have read at least one Harry Potter book.

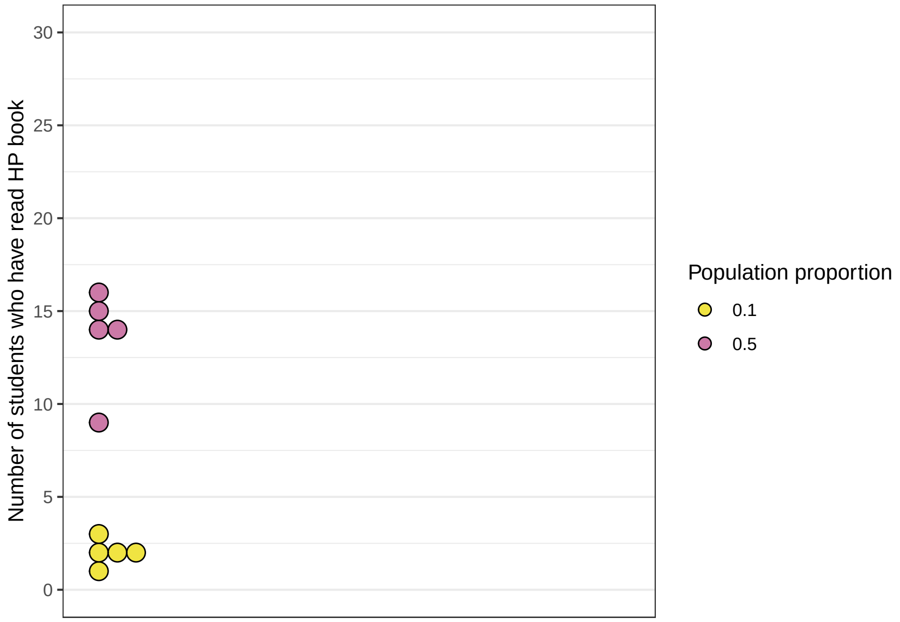


Figure 1.2: Number of students who have read at least one Harry Potter book in hypothetical samples of size 30. Five samples simulated assuming the population proportion is 0.1 (yellow), and five samples simulated assuming the population proportion is 0.5 (purple).

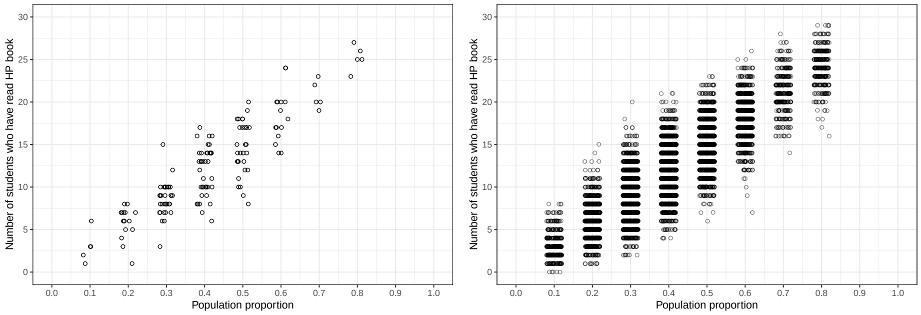


Figure 1.3: Simulation of the number of students who have read at least one Harry Potter book in hypothetical samples of size 30, reflecting initial plausibility of values of the population proportion from Figure 1.1. Left: 5 hypothetical samples for each guess for the population proportion. Right: 10000 hypothetical samples for each guess for the population proportion.

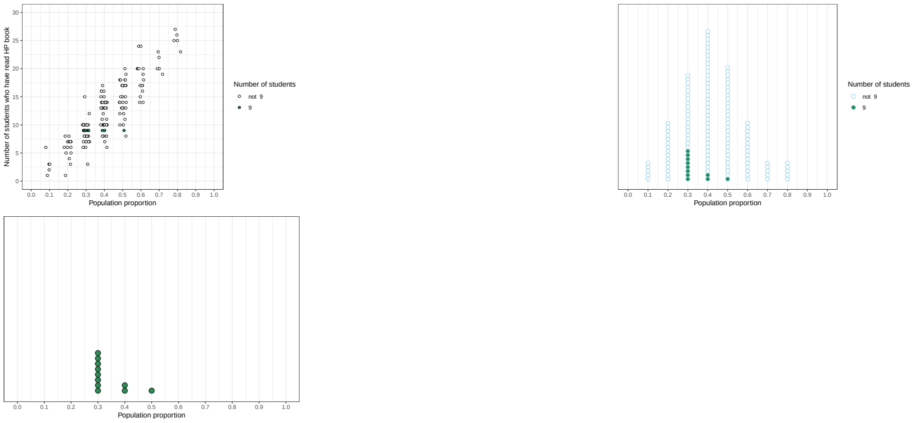


Figure 1.4: Left: Simulation results from the plot on the left in Figure 1.3 highlighting samples with a sample proportion of 9/30. Middle: Comparison of initial distribution of population proportion with conditional distribution of population proportion given a sample proportion of 9/30. Right: Distribution reflecting relative plausibility of possible values of the population proportion after observing a sample of 30 students in which 9 have read at least one Harry Potter book.

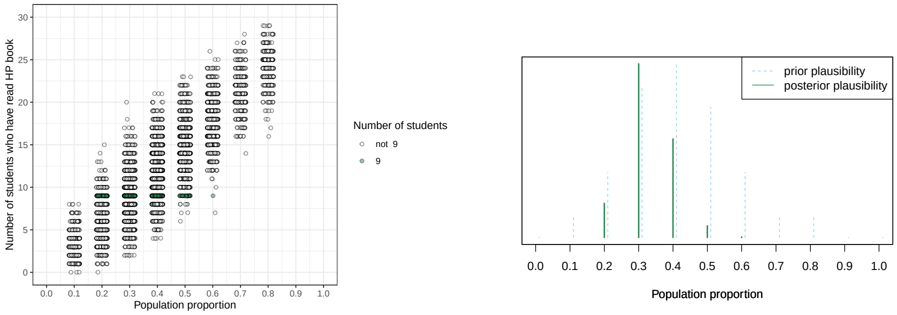


Figure 1.5: Left: Simulation results from the plot on the right in Figure 1.3 highlighting samples with a sample proportion of 9/30. Right: Distribution reflecting relative plausibility of possible values of the population proportion, both “prior” plausibility (blue) and “posterior” plausibility after observing a sample of 30 students in which 9 have read at least one Harry Potter book (green).

Chapter 2

Ideas of Bayesian Reasoning

In this section we'll look a closer look at the example from the previous section. In particular, we'll inspect the simulation process and results in more detail. We'll also consider a more "realistic" scenario.

In the previous section, each student identified a "most plausible" value. We collected these guesses to form a "wisdom of crowds" measure of initial plausibility.

However, your own initial assessment of the plausibility of the different values could involve much more than just identifying the most plausible value. What was your next most plausible value? How much less plausible was it? What about the other values and their relative plausibilities?

In the example below we'll start with an assessment of plausibility. We'll discuss later how you might obtain such an assessment. For now, focus on the big picture: we start with some initial assessment of plausibility before observing data, and we want to update that assessment upon observing some data.

Example 2.1. Suppose we're interested in **the proportion of all current Cal Poly students who have ever read at least one book in the *Harry Potter* series**. We'll refer to this proportion as the "population proportion" and denote it as θ (the Greek letter "theta").

1. We'll start by considering only the values 0.1, 0.2, 0.3, 0.4, 0.5, 0.6 as initially plausible for the population proportion θ. Suppose that before collecting sample data, our *prior* assessment is that:

    - 0.2 is four times more plausible than 0.1
    - 0.3 is two times more plausible than 0.2
    - 0.3 and 0.4 are equally plausible
    - 0.2 and 0.5 are equally plausible

- 0.1 and 0.6 are equally plausible

Construct a table to represent this *prior distribution* and sketch a plot of it.

2. Discuss what the prior distribution from the previous part represents.

3. If we simulate many values of the population proportion according to the above distribution, on what proportion of repetitions would we expect to see a value of 0.1? If we conduct 260000 repetitions of the simulation, on how many repetitions would we expect to see a value of 0.1? Repeat for the other plausible values to make a table of what we would expect the simulation results to look like. (For now, we're ignoring simulation variability; just consider expected values.)

4. The "prior" describes our initial assessment of plausibility of possible values of the population proportion prior to observing data. Now suppose we observe a sample of 30 students in which 9 students have read at least one HP book. We want to update our assessment of the population proportion in light of the observed data.

   Suppose that the actual population proportion is $\theta = 0.1$. How could we use simulation to determine the likelihood of observing 9 students who have read at least one HP book in a sample of 30 students? How could we use math? (Hint: what is the probability distribution of the number of "successes" in a sample of size 30 when $\theta = 0.1$?)

5. Recall the simulation with 260000 repetitions that we started above. Consider the 10000 repetitions in which the population proportion is 0.1. Suppose that for each of these repetitions we simulate the number of students in a class of 30 who have read at least one HP book. On what proportion of these repetitions would we expect to see a sample count of 9? On how many of these 10000 repetitions would we expect to see a sample count of 9?

6. Repeat the previous part for each of the possible values of θ: $0.1, \dots, 0.6$. Add two columns to the table: one column for the likelihood of observing a count of 9 in a sample of size 30 for each value of θ, and one column for the expected number of repetitions in the simulation which would result in the count of 9.

7. Consider just the repetitions that resulted in a simulated sample count of 9. What proportion of these repetitions correspond to a population proportion of 0.1? Of 0.2? Continue for the other possible values of θ to construct this *posterior distribution*, and sketch a plot of it.

8. After observing a sample of 30 Cal Poly students with a proportion of 9/30=0.3 who have read at least one Harry Potter book, what can we say about the plausibility of possible values of the *population* proportion?

How has our assessment of plausibility changed from before observing the sample data?

9. Prior to observing data, how many times more plausible is a value of 0.3 than 0.2 for the population proportion θ?

10. Recall that we observed a sample count of 9 in a sample of size 30. How many times more likely is a count of 9 in a sample of size 30 when the population proportion θ is 0.3 than when it is 0.2?

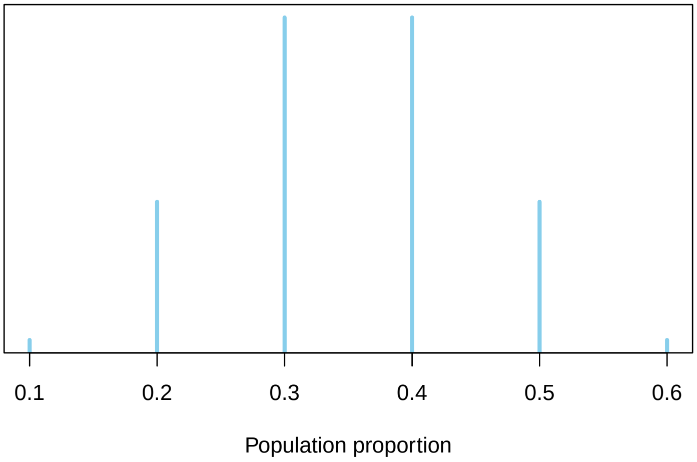
11. After observing data, how many times more plausible is 0.3 than 0.2 for the population proportion θ?

12. How are the values from the three previous parts related?

Solution. to Example 2.1

1. The relative plausibilities allow us to draw the shape of the plot below: the spike for 0.2 is four times as high as the one for 0.1; the spike for 0.3 is two times higher than the spike for 0.2, etc. To make a distribution we need to rescale the heights, maintaining the relative ratios, so that they add up to 1. It helps to consider one value as a "baseline"; we'll choose 0.1 (but it doesn't matter which value is the baseline). Assign 1 "unit" of plausibility to the value 0.1 Then 0.6 also gets 1 unit of plausibility. The values 0.2 and 0.5 each receive 4 units of plausibility, and the values 0.3 and 0.4 each receive 8 units of plausibility. The six values account for 26 total units of plausibility. Divide the units by 26 to obtain values that sum to 1. See the "Prior" column in the table below. Check that the relative ratios are maintained; for example, the rescaled prior plausibility of 0.308 for 0.3 is two times larger than the rescaled prior plausibility of 0.154 for 0.2.

Population proportion	Prior ”Units”	Prior
0.1	1	0.0385
0.2	4	0.1538
0.3	8	0.3077
0.4	8	0.3077
0.5	4	0.1538
0.6	1	0.0385
Total	26	1.0000



2. The parameter θ is an uncertain quantity. We are considering this quantity as a random variable, and using distributions to describe our degree of uncertainty. The prior distribution represents our degree of uncertainty, or our assessment of relative plausibility of possible values of the parameter, prior to observing data.

3. The relative prior plausibility of 0.1 is 1/26, so we would expect to see a value of 0.1 for θ on about 3.8% of repetitions. If we conduct 260000 repetitions of the simulation, we would expect to see a value of 0.1 for θ on about 10000 repetitions.

Population proportion	Prior ”Units”	Prior	Number of reps
0.1	1	0.0385	10000
0.2	4	0.1538	40000
0.3	8	0.3077	80000
0.4	8	0.3077	80000
0.5	4	0.1538	40000
0.6	1	0.0385	10000
Total	26	1.0000	260000

4. If the population proportion is 0.1 we would expect around 3 students in a random sample of 30 students to have read at least one Harry Potter book. Again, there would be natural sample-to-sample variability. To get a sense of this variability we could:

   - Roll a fair 10-sided die. A roll of 1 represents a student who has read at least one Harry Potter book; all other rolls, not.

- A set of 30 rolls represents one hypothetical random sample of 30 students.
- The number of the 30 rolls that land on 1 represents one hypothetical value of the number of students in a random sample of 30 students who have read at least one Harry Potter book.
- Repeat the above process to get many hypothetical values of the number of students in a random sample of 30 students who have read at least one Harry Potter book, *assuming that the population proportion is 0.1.*

If Y is the number of students in the same who have read at least one HP book, then Y has a Binomial(30, θ) distribution. If $\theta = 0.1$ then Y has a Binomial(30, 0.1) distribution. If $\theta = 0.1$ the probability that $Y = 9$ is $\binom{30}{9}(0.1^9)(0.9^{21}) = 0.0016$, which can be computed using `dbinom(9, 30, 0.1)` in R.

```
dbinom(9, 30, 0.1)
```

```
## [1] 0.001565
```

5. From the previous part, the likelihood of observing a count of 9 in a sample of size 30 when $\theta = 0.1$ is 0.0016. If $\theta = 0.1$ then we would expect to observe a sample count of 9 in about 0.16% of samples. In the 10000 repetitions with $\theta = 0.1$, we would expect to observe a count of 9 in about $10000 \times 0.0016 = 16$ repetitions.

6. See the table below. For example, if $\theta = 0.2$ then the likelihood of a sample count of 9 in a sample of size 30 is $\binom{30}{9}(0.2^9)(0.8^{21}) = 0.068$, which can be computed using `dbinom(9, 30, 0.2)`. If $\theta = 0.2$ then we would expect to observe a sample count of 9 in about 6.8% of samples. In the 40000 repetitions with $\theta = 0.2$, we would expect to observe a count of 9 in about $40000 \times 0.068 = 2703$ repetitions.

   In the table below, "Likelihood of 9" represents the probability of a sample count of 9 in a sample of size 30 computed for each possible value of θ. Note that this column does *not* sum to 1, as the values in this column do *not* comprise a probability distribution. Rather, the values in the likelihood column represent the probability of the same event (sample count of 9) computed under various different scenarios (different possible values of θ).

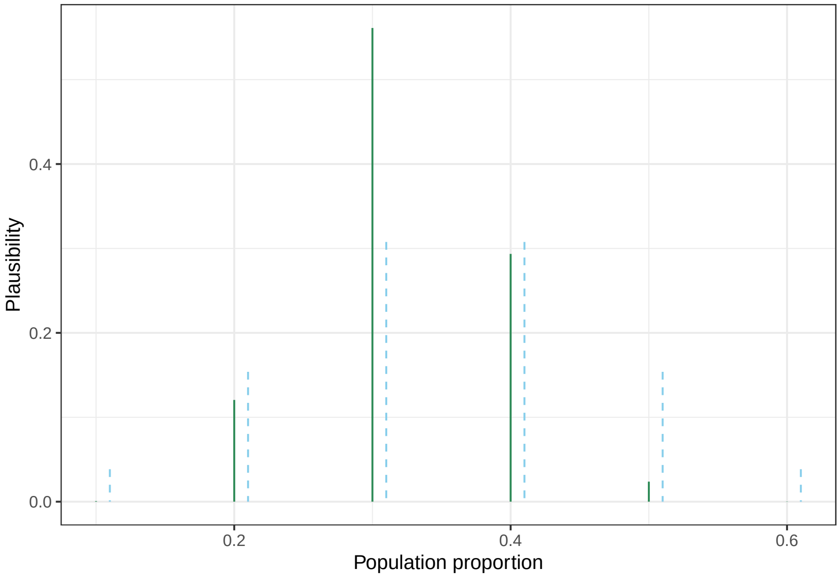
   The "Repetitions with a count of 9" column corresponds to the green dots in Figure 1.5. The prior plausibilities and total number of repetitions are different between the two examples, but the process is the same. (The overall "Number of reps" column corresponds to all the dots.)

Population proportion	Prior ”Units”	Prior	Number of reps	Likelihood of count of 9	R
0.1	1	0.0385	10000	0.0016	
0.2	4	0.1538	40000	0.0676	
0.3	8	0.3077	80000	0.1573	
0.4	8	0.3077	80000	0.0823	
0.5	4	0.1538	40000	0.0133	
0.6	1	0.0385	10000	0.0006	
Total	26	1.0000	260000	0.3227	

7. There were 22423 repetitions that resulted in a simulated sample count of 9. Of these, 16 correspond to a population proportion of 0.1. Therefore, the proportion of repetitions that resulted in a count of 9 that correspond to a proportion of 0.1 is 16 / 22423 = 0.0007. The proportion of repetitions that resulted in a count of 9 that correspond to a proportion of 0.2 is 2703 / 22423 = 0.1205. See the “Posterior” column in the table below.

Population proportion	Prior ”Units”	Prior	Number of reps	Likelihood of count of 9	R
0.1	1	0.0385	10000	0.0016	
0.2	4	0.1538	40000	0.0676	
0.3	8	0.3077	80000	0.1573	
0.4	8	0.3077	80000	0.0823	
0.5	4	0.1538	40000	0.0133	
0.6	1	0.0385	10000	0.0006	
Total	26	1.0000	260000	0.3227	

8. The plot below compares our prior plausibility (blue) and our posterior plausibility (green) after observing the data. The values 0.3 and 0.4 initially were equally plausible for θ, but after observing a sample proportion of 0.3 the value 0.3 is now almost 2 times more plauible than a value of 0.4. The values 0.2, 0.3, 0.4 together accounted for about 77% of our initial plausibility, but after observing the data these three values now account for over 97% of our plausibility.



9. Our prior assessment was that a value of 0.3 is 2 times more plausible than 0.2 for the population proportion θ.

10. A count of 9 in a sample of size 30 is 0.1573 / 0.0676 = 2.33 times mores likely when the population proportion θ is 0.3 than when it is 0.2.

11. After observing a count of 9 in a sample of size 30, a value of 0.3 is 0.5612 / 0.1205 = 4.66 times more plausible than 0.2 for the population proportion θ.

12. The ratio of the posterior plausibilities (4.66) is the product of the ratio of the prior plausibilities and the ratio of the likelihoods (2.33). In short, *posterior is proportional to product of prior and likelihood.*

Example 2.2. In the previous example we only considered the values 0.1, 0.2, …, 0.6 as plausible. Now we'll consider a more realistic scenario.

1. What is the problem with considering only the values 0.1, 0.2, …, 0.6 as plausible? How could we resolve this issue?

2. Suppose we have a prior distribution that assigns initial relative plausibility to a fine grid of possible values of the population proportion θ, e.g., 0, 0.0001, 0.0002, 0.0003, …, 0.9999, 1. We then observe a sample count of 9 in a sample of size 30. Explain the process we would use to construct a table like the one in the previous example to find the posterior distribution of the population proportion θ.

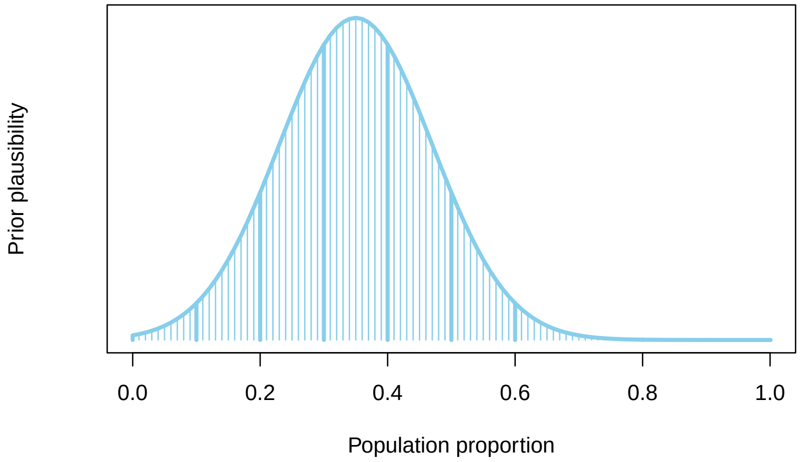
3. How could we use the posterior distribution to fill in the blank in the following: “There is a [blank] percent chance that fewer than 50 percent of current Cal Poly students have read at least one HP book.”

4. What are some other questions of interest regarding θ? How could you use the posterior distribution to answer them?

Solution. to Example 2.2

1. These six values are not the only possible values of θ. The parameter θ is a proportion, which could take any value in [0, 1]. We really want a prior that assigns relative plausibility to all values in the continuous interval [0, 1]. One way to bridge the gap is to consider a fine grid of values in [0, 1], rather than all possible values.

   We'll consider the possible values of θ to be $0, 0.0001, 0.0002, 0.0003, \ldots, 0.9998, 0.9999, 1$ and assign a relative plausibility to each of these values. We'll start with our assessment from the previous example: 0.1 and 0.6 are equally plausible, 0.2 is four times more plausible than 0.1, etc. We'll assign plausibility to in between values by “smoothly connecting the dots”. In the plot below this is achieved with a Normal distribution, but the details are not important for now. Just understand that (1) we have expanded our grid of possible values of θ, and (2) we have assigned a relative plausibility to each of the possible values.



2. The table has one row for each possible value of θ: $0, 0.0001, 0.0002, \ldots, 0.9999, 1$.

   - Prior: There would be a column for prior plausibility, say corresponding to the plot above.

- Likelihood: For each value of θ, we would compute the likelihood of observing a sample count of 9 in a sample of size 30: $\binom{30}{9}(\theta^9)(1-\theta)^{21}$ or `dbinom(9, 30, theta)`.
- Product: For each value of θ, compute the product of prior and likelihood. This is essentially what we did in the previous example in the "Reps with a count of 9" column. Here we're just not multiplying by the total number of repetitions.
- Posterior: The product column gives us the relative ratios. For example, the product column tells us that the posterior plausibility of 0.3 is 4.66 times greater than the posterior plausibility of 0.2. We simply need to rescale these values — by dividing by the sum of the product column — to obtain posterior plausibilities in the proper ratios that add up to one.

The following is some code; think of this as creating a spreadsheet. (Note that only a few select rows of the spreadsheet are displayed below.) We will explore code like this in much more detail as we go. For now, just notice that we can accomplish what we wanted in just a few lines of code.

```r
# Possible values of theta
theta = seq(0, 1, 0.0001)


# Prior distribution
# Smoothly connect the dots using a Normal distribution
# Then rescale to sum to 1

prior = dnorm(theta, 0.35, 0.12) # prior "units" - relative values
prior = prior / sum(prior) # recale to sum to 1


# Likelihood
# Likelihood of observing sample count of 9 out of 30
# for each theta

likelihood = dbinom(9, 30, theta)


# Posterior
# Product gives relative posterior plausibilities
# Then rescale to sum to 1

product = prior * likelihood
posterior = product / sum(product)

# Put the columns together
```

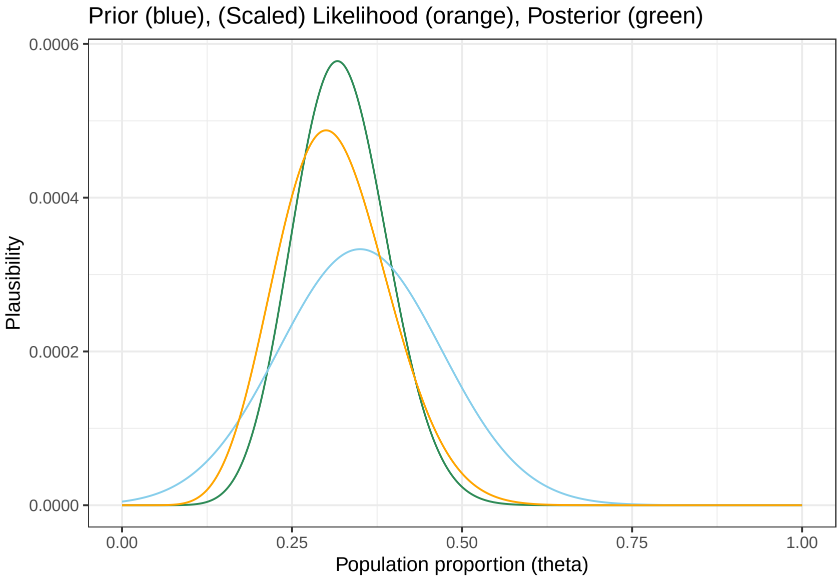
```
bayes_table = data.frame(theta,
                         prior,
                         likelihood,
                         product,
                         posterior)

# Display a portion of the table
bayes_table %>%
  slice(seq(2001, 4001, 250)) %>% # selects a few rows to display
  kable(digits = 8)
```

theta	prior	likelihood	product	posterior
0.200	0.0001525	0.06756	0.00001030	0.0001204
0.225	0.0001936	0.10012	0.00001938	0.0002265
0.250	0.0002353	0.12981	0.00003055	0.0003570
0.275	0.0002740	0.15019	0.00004115	0.0004808
0.300	0.0003054	0.15729	0.00004803	0.0005613
0.325	0.0003259	0.15062	0.00004909	0.0005736
0.350	0.0003330	0.13285	0.00004424	0.0005170
0.375	0.0003259	0.10847	0.00003535	0.0004131
0.400	0.0003054	0.08228	0.00002512	0.0002936

The plot below displays the prior, likelihood[1], and posterior. Notice that the likehood of the observed data is highest for θ near 0.3, so our plausibility has "moved" in the direction of θ near 0.3 after observing the data.

```
# The code below plots three curves
# One for each of prior, likelihood, posterior
# There are easier/better ways to do this

ggplot(bayes_table, aes(x = theta)) +
  geom_line(aes(y = posterior), col = "seagreen") +
  geom_line(aes(y = prior), col = "skyblue") +
  geom_line(aes(y = likelihood / sum(likelihood)), col = "orange") +
  labs(x = "Population proportion (theta)",
   y = "Plausibility",
   title = "Prior (blue), (Scaled) Likelihood (orange), Posterior (green)") +
  theme_bw()
```

[1] Prior and posterior are distributions which sum to 1, so prior and posterior are on the same scale. However, the likelihood does not sum to anything in particular. In order to plot the likehood on the same scale, it has been rescaled to sum to 1. Only the relative shape of the likelihood matters; not its absolute scale.



3. Sum the posterior plausibilities for θ values between 0.5. We can see from the plot that almost all our plausibility is placed on values of θ less than 0.5.

```
sum(posterior[theta < 0.5])
```

```
## [1] 0.9939
```

4. The posterior distribution gives us lots of information. We might be interested in questions like: "There is a [blank1] percent chance that between [blank2] and [blank3] percent of Cal Poly students have read a HP book." For example, for 80 percent in blank1, we might compute the 10th percentile for blank2 and the 90th percentile for blank3. Using the spreadsheet, start from $\theta = 0$ and go down the table summing the posterior probabilities until they reach 0.1; the corresponding θ value is the 10th percentile.

```
theta[max(which(cumsum(posterior) < 0.1))]
```

```
## [1] 0.235
```

We can find the 90th percentile similarly.

```
theta[max(which(cumsum(posterior) < 0.9))]
```

```
## [1] 0.411
```

There is an 80% chance that between 24% and 41% of Cal Poly students have read a HP book.

Chapter 3

Interpretations of Probability and Statistics

You have some familiarity with "probability" or "chance" or "odds". But what do we really mean when talk about "probability"? It turns out there are two main interpretations: relative frequency and "subjective" probability. These two interpretations provide the philosophical foundation for two schools of statistics: frequentist (hypothesis tests and confidence intervals that you've seen before) and Bayesian (what this book is about). This chapter introduces the two interpretations.

3.1 Instances of randomness

A wide variety of situations involve probability. Consider just a few examples.

1. The probability that you roll doubles in a turn of a board game.
2. The probability you win the next Powerball lottery if you purchase a single ticket, 4-8-15-16-42, plus the Powerball number, 23.
3. The probability that a "randomly selected" Cal Poly student is a California resident.
4. The probability that the high temperature in San Luis Obispo tomorrow is above 90 degrees F.
5. The probability that Hurricane Peter makes landfall in the U.S.
6. The probability that the San Francisco 49ers win the next Superbowl.
7. The probability that President Biden wins the 2024 U.S. Presidential Election.
8. The probability that extraterrestrial life currently exists somewhere in the universe.

9. The probability that Alexander Hamilton actually wrote 51 of the Federalist Papers. (The papers were published under a common pseudonym and authorship of some of the papers is disputed.)
10. The probability that you ate an apple on April 17, 2009.

Example 3.1. How are the situations above similar, and how are they different? What is one feature that all of the situations have in common? Is the interpretation of "probability" the same in all situations? Take some time to consider these questions before looking at the solution. The goal here is to just think about these questions, and not to compute any probabilities (or to even think about how you would).

Solution. to Example 3.1

Show/hide solution

This exercise is intended to motivate discussion, so you might have thought of some other ideas we don't address here. That's good! And some of the things you considered might come up later in the book. But here are a few thoughts we specifically want to mention now.

The one feature that all of the situations have in common is *uncertainty*. Sometimes the uncertainty arises from a repeatedable physical phenomenon that can result in multiple potential outcomes, like rolling dice or drawing the winning Powerball number. In other cases, there is uncertainty because the probability concerns the future, like tomorrow's high temperature or the result of the next Superbowl. But there can also be uncertainty about the past: there are some Federalist papers for which the author is unknown, and you probably don't know for sure whether or not you ate an apple on April 17, 2009.

Whenever there is uncertainty, it is reasonable to consider relative likelihoods of potential outcomes. For example, even though you don't know for certain whether you ate an apple on April 17, 2009, if you're usually an apple-a-day person (or were when you were younger) you might think the probability is high. We don't know for sure what team will win the next Superbowl, but we might think that the 49ers are more likely than the Eagles to be the winner.

While all of the situations in the example involve uncertainty, it seems that there are different "types" of uncertainty. Even though we don't know which side a die will land on, the notion of "fairness" implies that the sides are "equally likely". Likewise, there are some rules to how the Powerball drawing works, and it seems like these rules should determine the probability of drawing that particular winning number.

However, there aren't any specific "rules of uncertainty" that govern whether or not you ate an apple on April 17, 2009. You either did or you didn't, but that doesn't mean the two outcomes are necessarily equally likely. Regarding the Superbowl, of course there are rules that govern the NFL season and playoffs, but there are no "rules of uncertainty" that tell us precisely how likely any particular team is to win any particular game, let alone how likely a team is to advance to and win the Superbowl.

It also seems that there are different interpretations of probability. Given that a six-sided die is fair, we might all agree that the probability that it lands on any particular side is 1/6. Similarly, given the rules of the Powerball lottery, we might all agree on the probability that a drawing results in a particular winning number. However, there isn't necessarily consensus about what the high temperature will be in San Luis Obispo tomorrow. Different weather prediction models, forecasters, or websites might provide different values for the probability that the high temperature will be above 90 degrees Fahrenheit. Similarly, Superbowl odds might vary by source. Situations like tomorrow's weather or the Superbowl where there is no consensus about the "rules of uncertainty" require some subjectivity in determining probabilities.

Finally, some of these situations are repeatedable. We could (in principle) roll a pair of dice many times and see how often we get doubles, or repeat the Powerball drawing over and over to see how the winning numbers behave. However, many of these situations involve something that only happens once, like tomorrow or April, 17, 2009 or the next Superbowl. Even when the phenomenon happens only once in reality, we can still develop models of what might happen if we were to hypothetically repeat the phenomenon many times. For example, meteorologists use historical data and meteorological models to forecast potential paths of a hurricane.

The subject of probability concerns *random* phenomena. A phenomenon is **random**[1] if there are multiple potential outcomes, and there is **uncertainty** about which outcome will occur. Uncertainty is understood in broad terms, and in particular does not only concern future occurrences.

Some phenomena involve physical randomness[2], like flipping coins, rolling dice, drawing Powerballs at random from a bin, or randomly selecting Cal Poly students. In many other situations randomness just vaguely reflects uncertainty.

Contrary to colloquial uses of the word, random does *not* mean haphazard. In a random phenomenon, while individual outcomes are uncertain, we will see that there is a *regular distribution of outcomes over a large number of (hypothetical) repetitions.* For example,

[1] In this book, "random" and "uncertain" are synonyms; the opposite of "random" is "certain". (Later we will encounter random variables; "constant" is an antonym of "random variable".) The word "random" has many uses in everyday life, which have evolved over time. Unfortunately, some of the everyday meanings of "random", like "haphazard" or "unexpected", are contrary to what we mean by "random" in this book. For example, we would consider Steph Curry shooting a free throw to be a random phenomenon because we're not certain if he'll make it or miss it; but we would not consider this process to be haphazard or unexpected.

[2] We will refer to as "random" any scenario that involves a reasonable degree of uncertainty. We're avoiding philosophical questions about what is "true" randomness, like the following. Is a coin flip really random? If all factors that affect the trajectory of the coin were known precisely, then wouldn't the outcome be determined? Does true randomness only exist in quantum mechanics?

- In two flips of a fair coin we wouldn't necessarily see one head and one tail. But in 10000 flips of a fair coin, we might expect to see close to 5000 heads and 5000 tails.
- We don't know who will win the next Superbowl, but we can and should consider some teams as more likely to win than others. We could imagine a large number of hypothetical 2021-2022 seasons; how often would we expect the 49ers to win? The Eagles?

Random also does *not* necessarily mean equally likely. In a random phenomenon, certain outcomes or events might be more or less likely than others. For example,

- It's much more likely than not that a randomly selected Cal Poly student is a California resident.
- Not all NFL teams are equally likely to win the next Superbowl.

Finally, randomness is also not necessarily undesirable. In particular, many statistical applications often employ the planned use of randomness with the goal of collecting "good" data. For example,

- *Random selection* involves selecting a sample of individuals "at random" from a population (e.g., via random digit dialing), with the goal of selecting a representative sample.

- *Random assignment* involves assigning individuals at random to groups (e.g., in a randomized experiment), with the goal of constructing groups that are similar in all aspects so that the effect of a treatment (like a new vaccine) can be isolated.

The **probability** of an event associated with a random phenomenon is a number in the interval $[0, 1]$ measuring the event's likelihood or degree of uncertainty. A probability can take any values in the continuous scale from 0% to 100%[3]. In particular, a probability requires much more interpretation than "is the probability greater than, less than, or equal to 50%?" As Example 3.1 suggests, there can be different interpretations of "probability", which we'll start to explore in the next section.

3.2 Interpretations of probability

In the previous section we encountered a variety of scenarios which involved uncertainty, a.k.a. randomness. Just as there are a few "types" of randomness, there are a few ways of interpreting probability, most notably, *long run relative frequency* and *subjective probability.*

[3] Probabilities are usually defined as decimals, but are often colloquially referred to as percentages. We're not sticklers; we'll refer to probabilities both as decimals and as percentages.

3.2.1 Long run relative frequency

One of the oldest documented[4] problems in probability is the following: If three fair six-sided dice are rolled, what is more likely: a sum of 9 or a sum of 10? Let's try to answer this question by simply rolling dice and seeing if a sum of 9 or 10 happens more frequently. Roll three fair six-sided dice, find the sum, repeat many times, and see how often we get a sum of 9 versus a sum of 10. Of course, this would be a time consuming process by hand, but it's quick and easy on a computer. Figure 3.1 displays the result of one million repetitions of this process, each repetition resulting in the sum of three rolls. A sum of 9 occurred in 115384 repetitions and a sum of 10 occurred in 125005 repetitions. Comparing these frequencies, our results suggest that a sum of 10 is more likely than a sum of 9.

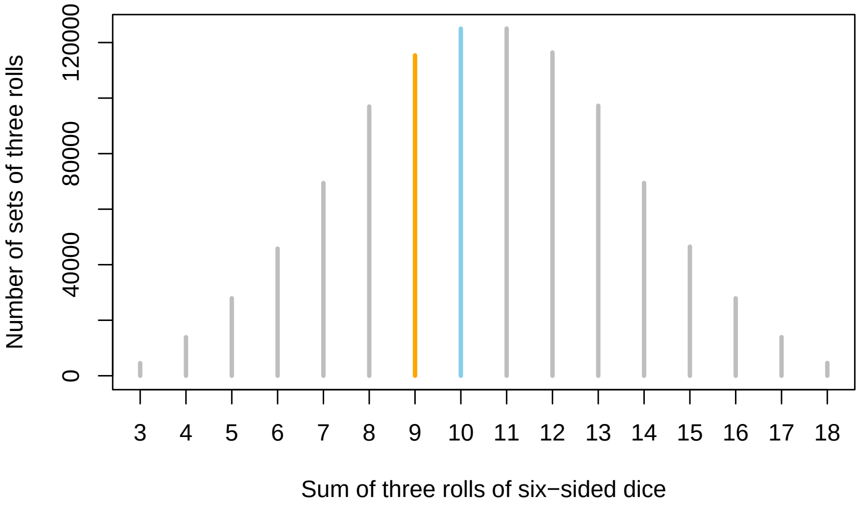


Figure 3.1: Results of one million sets of three rolls of fair six-sided dice. Sets in which the sum of the dice is 9 (10) are represented by orange (blue) spike.

In the previous problem we assessed relative likelihoods by repeating the process many times. This is the idea behind the relative frequency interpretation of probability. We'll investigate this idea further in the context of what is probably the most iconic random process: coin flipping.

[4] The Grand Duke of Tuscany posed this problem to Galileo, who published his solution in 1620. However, unbeknownst to Galileo, the same problem had been solved almost 100 years earlier by Gerolamo Cardano, one of the first mathematicians to study probability.

Table 3.1: Results and running proportion of H for 10 flips of a fair coin.

Flip	Result	Running count of H	Running proportion of H
1	T	0	0.000
2	H	1	0.500
3	T	1	0.333
4	H	2	0.500
5	H	3	0.600
6	H	4	0.667
7	H	5	0.714
8	T	5	0.625
9	T	5	0.556
10	T	5	0.500

We might all agree that the probability that a single flip of a fair coin lands on heads is 1/2, a.k.a., 0.5, a.k.a, 50%. After all, the notion of "fairness" implies that the two outcomes, heads and tails, should be equally likely, so we have a "50/50 chance" of heads. But how else can we interpret this 50%? As in the dice rolling problem, we can consider *what would happen if we flipped the coin main times*. Now, if we would flipped the coin twice, we wouldn't expect to necessarily see one head and one tail. But in many flips, we might expect to see heads on something close to 50% of flips.

Let's try this out. Table 3.1 displays the results of 10 flips of a fair coin. The first column is the flip number and the second column is the result of the flip. The third column displays the *running proportion of flips that result in H*. For example, the first flip results in T so the running proportion of H after 1 flip is 0/1; the first two flips result in (T, H) so the running proportion of H after 2 flips is 1/2; and so on. Figure 3.2 plots the running proportion of H by the number of flips. We see that with just a small number of flips, the proportion of H fluctuates considerably and is not guaranteed to be close to 0.5. Of course, the results depend on the particular sequence of coin flips. We encourage you to flip a coin 10 times and compare your results.

Now we'll flip the coin 90 more times for a total of 100 flips. The plot on the left in Figure 3.3 summarizes the results, while the plot on the right also displays the results for 3 additional sets of 100 flips. The running proportion fluctuates considerably in the early stages, but settles down and tends to get closer to 0.5 as the number of flips increases. However, each of the fours sets results in a different proportion of heads after 100 flips: 0.5 (blue), 0.44 (orange), 0.56 (green), 0.56 (purple). Even after 100 flips the proportion of flips that result in H isn't guaranteed to be very close to 0.5.

Now for each set of 100 flips, we'll flip the coin 900 more times for a total of 1000 flips in each of the four sets. The plot on the left in Figure 3.4 summarizes the

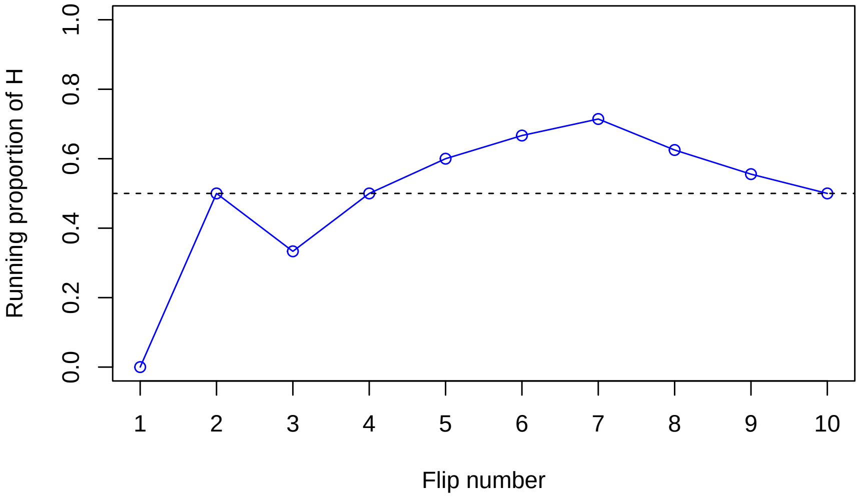


Figure 3.2: Running proportion of H versus number of flips for the 10 coin flips in Table 3.1.

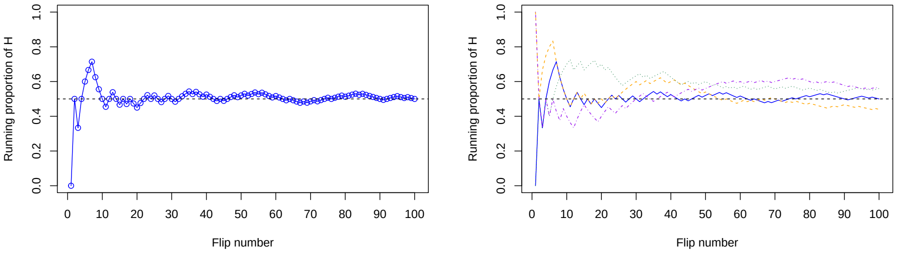


Figure 3.3: Running proportion of H versus number of flips for four sets of 100 coin flips.

results for our original set, while the plot on the right also displays the results for the three additional sets from Figure 3.4. Again, the running proportion fluctuates considerably in the early stages, but settles down and tends to get closer to 0.5 as the number of flips increases. Compared to the results after 100 flips, there is less variability between sets in the proportion of H after 1000 flips: 0.51 (blue), 0.488 (orange), 0.525 (green), 0.492 (purple). Now, even after 1000 flips the proportion of flips that result in H isn't guaranteed to be exactly 0.5, but we see a tendency for the proportion to get closer to 0.5 as the number of flips increases.

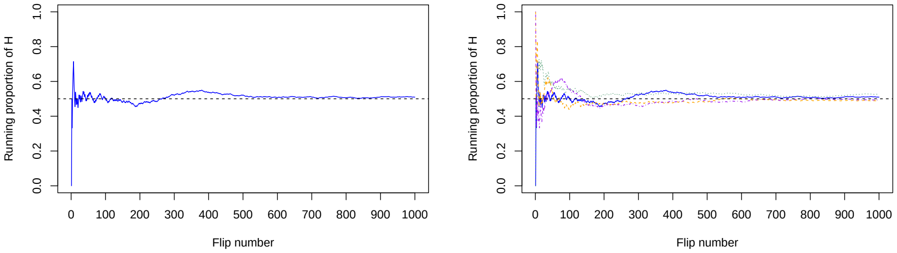


Figure 3.4: Running proportion of H versus number of flips for four sets of 1000 coin flips.

In summary, in a large number of flips of a fair coin we expect about 50% of flips to result in H. That is, the probability that a flip of a fair coin results in H can be interpreted as the *long run proportion of flips that result in H*, or in other words, the *long run relative frequency of H*.

In general, the probability of an event associated with a random phenomenon can be interpreted as a **long run proportion** or **long run relative frequency**: the probability of the event is the proportion of times that the event would occur in a very large number[5] of hypothetical repetitions of the random phenomenon.

The long run relative frequency interpretation of probability can be applied when a situation can be repeated numerous times, at least conceptually, and an outcome can be observed for each repetition. One benefit of the relative frequency interpretation is that the probability of an event can be *approximated by simulating* the random phenomenon a large number of times and determining the proportion of simulated repetitions on which the event occurred out of the total number of repetitions in the simulation. A **simulation** involves an artificial recreation of the random phenomenon, usually using a computer. After many repetitions the relative frequency of the event will settle down to a single constant value, and that value is the approximately the probability of the event.

[5] A natural question is: "how many repetitions are required to represent the long run?" We'll consider this question when we discuss MCMC methods.

Of course, the accuracy of simulation-based approximations of probabilities depends on how well the simulation represents the actual random phenomenon. Conducting a simulation can involve many assumptions which influence the results. Simulating many flips of a fair coin is one thing; simulating an entire NFL season and the winner of the Superbowl is an entirely different story.

3.2.2 Subjective probability

The long run relative frequency interpretation is natural in repeatable situations like flipping coins, rolling dice, drawing Powerballs, or randomly selecting Cal Poly students.

On the other hand, it is difficult to conceptualize some scenarios in the long run. The next Superbowl will only be played once, the 2024 U.S. Presidential Election will only be conducted once (we hope), and there was only one April 17, 2009 on which you either did or did not eat an apple. But while these situations are not naturally repeatable they still involve randomness (uncertainty) and it is still reasonable to assign probabilities. At this point in time we might think that the Kansas City Chiefs are more likely than the Philadelphia Eagles to win Superbowl 2022 and that President Biden is more likely than Dwayne Johnson to win the U.S. 2024 Presidential Election. If you've always been an apple-a-day person, you might think there's a good chance you ate one on April 17, 2009. It is still reasonable to assign probabilities to quantify such assessments even when an uncertain phenomenon is not repeated.

However, the *meaning* of probability does seem different in a physically repeatable situations like coin flips than in single occurrences like the 2022 Superbowl. For example, as of Dec 30, 2021,

- According to FiveThirtyEight, the Kansas City Chiefs have a 26% chance of winning the 2022 Superbowl, and the Green Bay Packers have a 24% chance.
- According to Football Outsiders, the Kansas City Chiefs have a 19.4% chance of winning the 2022 Superbowl, and the Green Bay Packers have a 14% chance.
- As reported by CBS Sports, the Kansas City Chiefs have a 20% chance of winning the 2022 Superbowl, and the Green Bay Packers have a 21% chance.

Each source, as well as many others, assigns different probabilities to the Chiefs and Packers winning. Which source, if any, is "correct"?

When the situation involves a fair coin flip, we could perform a simulation to see that the long run proportion of flips that land on H is 0.5, and so the probability that a fair coin flip lands on H is 0.5. Even though the actual 2022 Superbowl will only happen once, we could still perform a simulation involving hypothetical repetitions. However, simulating the Superbowl involves first simulating the 2021-2022 season to determine the playoff matchups, then simulating the playoffs to see which teams make the Superbowl, then simulating the Superbowl matchup itself. And simulating the season involves simulating all the individual games. Even just simulating a single game involves many assumptions; differences in opinions with regards to these assumptions can lead to different probabilities. For example, on Dec 30, according to FiveThirtyEight the Eagles had a 55% chance of beating the Washington Football Team in their game on Jan 2, but according to numberFire it was 65%. (Let's hope the Eagles won.) Even though the differences in probabilities between sources are often small, many small differences over the course of the season could result in large differences in predictions for the Superbowl champion.

Unlike physically repeatable situations such as flipping a coin, there is no single set of "rules" for conducting a simulation of a season of football games or the Superbowl champion. Therefore, there is no single long run relative frequency that determines the probability. Instead we consider *subjective probability.*

A **subjective (a.k.a. personal) probability** describes the degree of likelihood a given individual assigns to a certain event. As the name suggests, different individuals (or probabilistic models) might have different subjective probabilities for the same event. In contrast, in the long run relative frequency interpretation the probability is agreed to be defined as the long run relative frequency, a single number.

Think of subjective probabilities as measuring *relative degrees of likelihood, uncertainty, or plausibility* rather than long run relative frequencies. For example, in the FiveThirtyEight forecast (as of Dec 30), the Chiefs (26% chance) are about *3.25 times more likely* to win the 2022 Superbowl than the Cowboys (8% chance); 3.25 = 26/8. Relative likelihoods can also be compared across different forecasts or scenarios. For example, FiveThirtyEight believes that the Packers are about 1.7 times more likely to win the Superbowl than Football Outsiders does (24% versus 14%). Also, FiveThirtyEight believes that the likelihood that a fair coin lands on H is about 1.92 times larger than the likelihood that the Chiefs win the 2022 Superbowl.

The FiveThirtyEight NFL predictions are the output of a probabilistic forecast. A **probabilistic forecast** combines observed data and statistical models to make predictions. Rather than providing a single prediction (such as "the Chiefs will win the 2022 Superbowl"), probabilistic forecasts provide a range of scenarios and their relative likelihoods. Such forecasts are subjective in nature, relying upon the data used and assumptions of the model. Changing the data or assumptions can result in different forecasts and probabilities. In particular, probabilistic forecasts are usually revised over time as more data becomes available.

Simulations can also be based on subjective probabilities. If we were to conduct a simulation consistent with FiveThirtyEight's model (as of Dec 30), then in about 26% of repetitions the Chiefs would win the Superbowl, and in about 8% of repetitions the Cowboys would win. Of course, different sets of subjective probabilities correspond to different assumptions and different ways of conducting the simulation.

Subjective probabilities can be calibrated by weighing the relative favorability of different bets, as in the following example.

Example 3.2. What is your subjective probability that Professor Ross has a TikTok account? Consider the following two bets, and suppse you must choose only one[6].

A) You win $100 if Professor Ross has a TikTok account, and you win nothing otherwise.
B) A box contains 40 green and 60 gold marbles that are otherwise identical. The marbles are thoroughly mixed and one marble is selected at random. You win $100 if the selected marble is green, and you win nothing otherwise.

1. Which of the above bets would you prefer? Or are you completely indifferent? What does this say about your subjective probability that Professor Ross has a Tik Tok account?
2. If you preferred bet B to bet A, consider bet C which has a similar setup to B but now there are 20 green and 80 gold marbles. Do you prefer bet A or bet C? What does this say about your subjective probability that Professor Ross has a Tik Tok account?
3. If you preferred bet A to bet B, consider bet D which has a similar setup to B but now there are 60 green and 40 gold marbles. Do you prefer bet A or bet D? What does this say about your subjective probability that Professor Ross has a Tik Tok account?
4. Continue to consider different numbers of green and gold marbles. Can you zero in on your subjective probability?

Solution. to Example 3.2

Show/hide solution

1. Since the two bets have the same payouts, you should prefer the one that gives you a greater chance of winning! If you choose bet B you have a 40% chance of winning.
    - If you prefer bet B to bet A, then your subjective probability that Professor Ross has a TikTok account is less than 40%.
    - If you prefer bet A to bet B, then your subjective probability that Professor Ross has a TikTok account is greater than 40%.

[6] We do not advocate gambling. We merely use gambling contexts to motivate probability concepts.

- If you're indifferent between bets A and B, then your subjective probability that Professor Ross has a TikTok account is equal to 40%.

2. If you choose bet C you have a 20% chance of winning.
    - If you prefer bet C to bet A, then your subjective probability that Professor Ross has a TikTok account is less than 20%.
    - If you prefer bet A to bet C, then your subjective probability that Professor Ross has a TikTok account is greater than 20%.
    - If you're indifferent between bets A and C, then your subjective probability that Professor Ross has a TikTok account is equal to 20%.

3. If you choose bet D you have a 60% chance of winning.
    - If you prefer bet D to bet A, then your subjective probability that Professor Ross has a TikTok account is less than 60%.
    - If you prefer bet A to bet D, then your subjective probability that Professor Ross has a TikTok account is greater than 60%.
    - If you're indifferent between bets A and D, then your subjective probability that Professor Ross has a TikTok account is equal to 60%.

4. Continuing in this way you can narrow down your subjective probability. For example, if you prefer bet B to bet A and bet A to bet C, your subjective probability is between 20% and 40%. Then you might consider bet E corresponding to 30 gold marbles and 70 green to determine if you subjective probability is greater than or less than 30%. At some point it will be hard to choose, and you will be in the ballpark of your subjective probability. (Think of it like going to the eye doctor: "which is better: 1 or 2?" At some point you can't really see a difference.)

Of course, the strategy in the above example isn't an exact science, and there is a lot of behavioral psychology behind how people make choices in situations like this, especially when betting with real money. But the example provides a very rough idea of how you might discern a subjective probability of an event. The example also illustrates that probabilities can be "personal"; *your* information or assumptions will influence your assessment of the likelihood.

We close this section with some brief comments about subjectivity. Subjectivity is not bad; "subjective" is not a "dirty" word. Any probability model involves some subjectivity, even when probabilities can be interpreted naturally as long run relative frequencies. For example, assuming a die is fair does not codify an objective truth about the die. Instead, "fairness" reflects a reasonable and tractable mathematical model. In the real world, any "fair" six-sided die has small physical imperfections that cause the six faces to have different probabilities. However, the differences are usually small enough to be ignored for most practical purposes. Assuming that the probability that the die lands

40 green, 60 gold
20 green, 80 gold
60 green, 40 gold

Figure 3.5: The three marble bins in Example 3.2. Left: Bet A, 40% chance of selecting green. Middle: Bet B, 20% chance of selecting green. Left: Bet C, 60% chance of selecting green.

on each side is 1/6 is much more tractable than assuming the probability of a 1 is 0.1666666668, the probability of a 2 is 0.1666666665, etc. (Furthermore, measuring the probability of each side so precisely would be extremely difficult.) But assuming that the probability that the die lands on each side is 1/6 is also subjective. We might readily agree to assume that the probability that a six-sided die lands on 1 is 1/6, but we might not reach a concensus on the probability that the Chiefs win the Superbowl. But the fact that there cam be many reasonable probability models for a situation like the 2022 Superbowl does not make the corresponding subjective probabilities any less valid than long run relative frequencies.

3.3 Working with probabilities

In the previous section we encountered two interpretations of probability: long run relative frequency and subjective. We will use these interpretations interchangeably. With subjective probabilities it is often helpful to consider what might happen in a simulation. It is also useful to consider long run relative frequencies in terms of relative degrees of likelihood. Fortunately, the mathematics of probability work the same way regardless of the interpretation.

3.3.1 Consistency requirements

With either the long run relative frequency or subjective probability interpretation there are some basic logical consistency requirements which probabilities need to satisfy. Roughly, probabilities cannot be negative and the sum of probabilities over all possible outcomes must be 100%.

Example 3.3. As of Dec 30, FiveThirtyEight listed the following probabilities for who will win the 2022 Superbowl.

Team	Probability
Kansas City Chiefs	26%
Green Bay Packers	24%
Tampa Bay Buccaneers	9%
Dallas Cowboys	8%
Other	

According to FiveThirtyEight (as of Dec 30):

1. What would you expect the results of 10000 repetitions of a simulation of the Superbowl champion to look like? Construct a table summarizing what you expect. Is this necessarily what would happen?
2. What must be the probability that the Chiefs do *not* win the 2022 Superbowl?
3. What must be the probability that one of the above four teams is the Superbowl champion?
4. What must be the probability that a team other than the above four teams is the Superbowl champion? That is, what value goes in the "Other" row in the table?

Solution. to Example 3.3

Show/hide solution

1. While these particular probabilities are subjective, imagining probabilities as relative frequencies often helps our intuition. If we think of this as a simulation, each repetition results in a World Series champion and in the long run we would expect the Dodgers would be the champion in 22%, or 2200, of the 10000 repetitions. We would expect the simulation results to look like

Team	Repetitions as winner
Kansas City Chiefs	2600
Green Bay Packers	2400
Tampa Bay Buccaneers	900
Dallas Cowboys	800

Team	Repetitions as winner
Other	3300

Of course, there would be some variability from simulation to simulation, just like in the sets of 1000 coin flips in Figure 3.4. But the above counts represent about what we would expect.

2. 74%. Either the Chiefs win or they don't; if there's a 26% chance that the Chiefs win, there must be a 74% chance that they do not win. If we think of this as a simulation with 10000 repetitions, each repetition results in either the Chiefs winning or not, so if they win in 2600 of repetitions then they must not win in the other 7400.

3. 67%. There is only one Superbowl champion, so if say the Chiefs win then no other team can win. Thinking again of the simulation, the repetitions in which the Chiefs win are distinct from those in which the Cowboys win. So if the Chiefs win in 2600 repetitions and the Cowboys win in 800 repetitions, then on a total of 3400 repetitions either the Chiefs or Cowboys win. Adding the four probabilities, we see that the probability that one of the four teams above wins must be 67%.

4. 33%. Either one of the four teams above wins, or some other team wins. If one of the four teams above wins in 6700 repetitions, then in 3300 repetitions the winner is not one of these four teams.

Example 3.4. Suppose your subjective probabilities for the 2022 Superbowl champion satisfy the following conditions.

- The Cowboys and Buccaneers are equally likely to win
- The Packers are 1.5 times more likely than the Cowboys to win
- The Chiefs are 2 times more likely than the Packers to win
- The winner is as likely to be among these four teams — Chiefs, Packers, Buccaneers, Cowboys — as not

Construct a table of your subjective probabilities like the one in Example 3.3.

Solution. to Example 3.4

Show/hide solution

Here, probabilities are specified indirectly via relative likelihoods. We need to find probabilities that are in the given ratios and add up to 100%. It helps to designate one outcome as the "baseline". It doesn't matter which one; we'll choose the Cowboys.

- Suppose the Cowboys account for 1 "unit". It doesn't really matter what a unit is, but let's say it corresponds to 1000 repetitions of the simulation. That is, the Cowboys win in 1000 repetitions. Careful: we haven't yet specified how many total repetitions we have done, or how many units the entire simulation accounts for. We're just starting with a baseline of what happens for the Cowboys.
- The Cowboys and Buccaneers are equally like to win, so the Buccaneers also account for 1 unit.
- The Packers are 1.5 times more likely than the Cowboys to win, so the Packers account for 1.5 units. If 1 unit is 1000 repetitions, then the Packers win in 1500 repetitions, 1.5 times more often than the Cowboys.
- The Chiefs are 2 times more likely than the Packers to win, so the Chiefs account for $2 \times 1.5 = 3$ units. If 1 unit is 1000 repetitions, then the Chiefs win in 3000 repetitions.
- The four teams account for a total of $1 + 1 + 1.5 + 3 = 6.5$ units. Since the winner is as likely to among these four teams as not, then "Other" also accounts for 6.5 units.
- In total, there are 13 units which account for 100% of the probability. The Cowboys account for 1 unit, so their probability of winning is 1/13 or about 7.7%. Likewise, the probability that the Chiefs win is 3/13 or about 23.1%.

Team	Units	Repetitions	Probability
Kansas City Chiefs	3.0	3000	23.1%
Green Bay Packers	1.5	1500	11.5%
Tampa Bay Buccaneers	1.0	1000	7.7%
Dallas Cowboys	1.0	1000	7.7%
Other	6.5	6500	50.0%
Total	13.0	13000	100.0%

You should verify that all of the probabilities are in the specified ratios. For example, the Chiefs are 2 times more likely ($2 = 23.1/11.5$) than the Packers to win, and the Packers are 1.5 times more likely ($1.5 \approx 11.5/7.7$) than the Cowboys to win.

We could have also solved this problem using algebra. Let x be the probability, as a decimal, that the Cowboys are the winner. (Again, it doesn't matter which team is the baseline.) Then x is also the probability that the Buccaneers are the winner, $1.5x$ for the Packers, and $3x$ for the Chiefs. The probability that one of the four teams wins is $x + x + 1.5x + 3x = 6.5x$, so the probability of Other is also $6.5x$. The probabilities in decimal form must sum to 1 (that is, 100%), so $1 = x + x + 1.5x + 3x + 6.5x = 13x$. Solve for $x = 1/13$ and then plug in $x = 1/13$ to find the other probabilities.

Example 3.4 illustrates one way of formulating probabilities. We start by specifying probabilities in relative terms, and then “normalize” these probabilities so that they add up to 100% while maintaining the ratios. As in the example, it helps to consider one outcome as a “baseline” and to specify all likelihoods relative to the baseline.

Figure 3.6 provides a visual representation of Example 3.4. The ratios provided in the problem setup are enough to draw the shape of the plot, represented by the plot on the left without a scale on the vertical axis. The heights are equal for the Cowboys and Buccaneers, the height for the Packers is 1.5 times higher, etc. The plot on the right simply adds a probability axis to ensure the values add to 1. The plot on the right represents the “normalization” step, but it does not affect the shape of the plot or the relative heights of the bars.

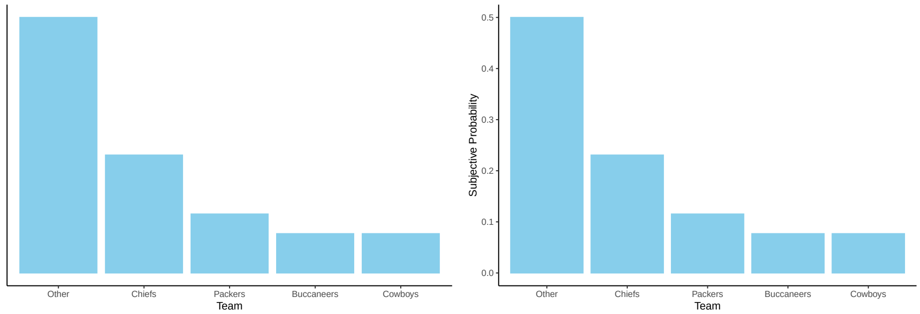


Figure 3.6: Bar chart representation of the subjective probabilities in Example 3.4. Left: Relative heights without absolute scale. Right: Heights scaled to sum to 1 to represent probabilities.

3.4 Interpretations of Statistics

In the previous sections we have seen two interpretations of statistics: relative frequency and subjective. These two interpretations provide the philosophical foundation for two schools of statistics: *frequentist* (hypothesis tests and confidence intervals that you’ve seen before) and *Bayesian*. This section provides a very brief introduction to some of the main ideas in Bayesian statistics. The examples in this section only motivate ideas. We will fill in lots more details throughout the book.

Example 3.5. How old do you think your instructor (Professor Ross) currently is[7]? Consider age on a continuous scale, e.g., you might be 20.73 or 21.36 or 19.50.

[7] You could probably get a pretty good idea by searching online, but don’t do that. Instead, answer the questions based on what you already know about me.

In this example, you will use probability to quantify your uncertainty about your instructor's age. You only need to give ballpark estimates of your subjective probabilities, but you might consider what kinds of bets you would be willing to accept like in Example 3.2. (This exercise just motivates some ideas. We'll fill in lots of details later.)

1. What is your subjective probability that your instructor is at most 30 years old? More than 30 years old? (What must be true about these two probabilities?)
2. What is your subjective probability that your instructor is at most 60 years old? More than 60 years old?
3. What is your subjective probability that your instructor is at most 40 years old? More than 40 years old?
4. What is your subjective probability that your instructor is at most 50 years old? More than 50 years old?
5. Fill in the blank: your subjective probability that your instructor is at most [blank] years old is equal to 50%.
6. Fill in the blanks: your subjective probability that your instructor is between [blank] and [blank] years old is equal to 95%.
7. Let θ represent your instructor's age at noon on Jan 6, 2022. Use your answers to the previous parts to sketch a continuous probability density function to represent your *subjective probability distribution* for θ.
8. If you ascribe a probability distribution to θ, then are you treating θ as a constant or a random variable?

Solution. to Example 3.5

Show/hide solution

Even though in reality your instructor's current age is a fixed number, its value is unknown or uncertain to you, and you can use probability to quantify this uncertainty. You would probably be willing to bet any amount of money that your instructor is over 20 years old, so you would assign a probability of 100% to that event, and 0% to the event that he's at most 20 years old. Let's say you're pretty sure that he's over 30, but you don't know that for a fact, so you assign a probability of 99% to that event (and 1% to the event that he's at most 30). You think he's over 40, but you're even less sure about that, so maybe you assign the event that he's over 40 a probability of 67% (say you'd accept a bet at 2 to 1 odds.) You think there's a 50/50 chance that he's over 50. You're 95% sure that he's between 35 and 60. And so on. Continuing in this way, you can start to determine a probability distribution to represent your beliefs about the instructor's age. Your distribution should correspond to your subjective probabilities. For example, the distribution should assign a probability of 67% to values over 40.

This is just one example. Different students will have different distributions depending upon (1) how much information you know about the instructor, and

(2) how that information informs your beliefs about the instructor's age. We'll see some example plots in the next exercise.

Regarding the last question, since we are using a probability distribution to quantify our uncertainty about θ, we are treating θ as a *random variable*.

Recall that a **random variable** is a numerical quantity whose value is determined by the outcome of a random or uncertain phenomenon. The random phenomenon might involve physically repeatable randomness, as in "flip a coin 10 times and count the number of heads." But remember that "random" just means "uncertain" and there are lots of different kinds of uncertainty. For example, the total number of points scored in the 2022 Superbowl will be one and only one number, but since we don't know what that number is we can treat it as a random variable. Treating the number of points as a random variable allows us to quantify our uncertainty about it through probability statements like "there is a 60% chance that fewer than 45 points will be scored in Superbowl 2022".

The **(probability) distribution** of a random variable specifies the possible values of the random variable and a way of determining corresponding probabilities. Like probabilities themselves, probability distributions of random variables can also be interpreted as:

- *relative frequency distributions*, e.g., what pattern would emerge if I simulated many values of the random variable? or as
- *subjective probability distributions*, e.g., which potential values of this uncertain quantity are relatively more plausible than others?

As the name suggests, different individuals might have different subjective probability distributions for the same random variable.

Example 3.6. Continuing Example 3.5, Figure 3.7 displays the subjective probability distribution of the instructor's age for four students.

1. Since age is treated as a continuous random variable, each of the above plots is a probability "density". Explain briefly what this means. How is probability represented in density plots like these?
2. Rank the students in terms of their subjective probability that the instructor is at most 40 years old.
3. Rank the students in terms of their answers to the question: your subjective probability that your instructor is at most [blank] years old is equal to 50%.
4. Rank the students in terms of their uncertainty about the instructor's age. Who is the most uncertain? The least?

Solution. to Example 3.6

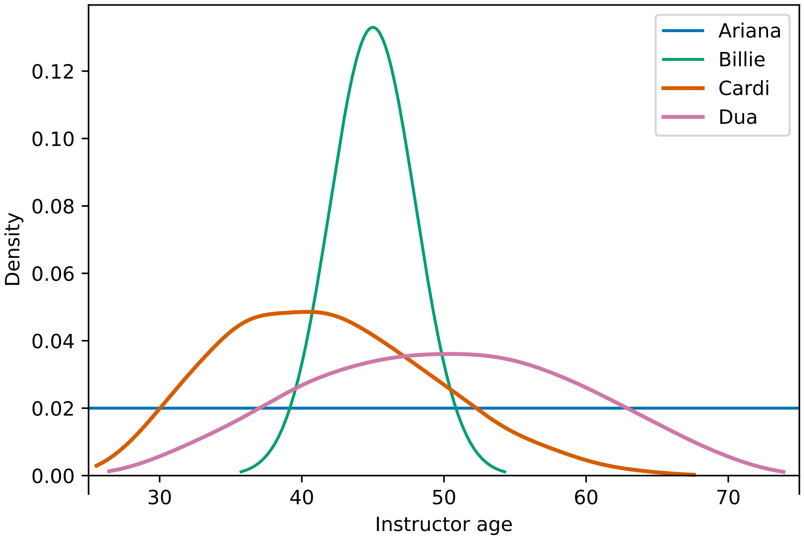


Figure 3.7: Subjective probability distributions of instructor age for four students in Example 3.6.

Show/hide solution

1. In a density plot, probability is represented by area under the curve. The total area under each curve is 1, corresponding to 100% probability. The density height at any particular value x represents the relative likelihood that the random variable takes a value "close to" x. (We'll consider densities in more detail later.)
2. Each student's subjective probability that the instructor is at most 40 is equal to the area under her subjective probability density over the range of values less than 40. Billie has the smallest probability, then Dua, then Ariana, then Cardi has the largest probability.
3. Now we want to find the "equal areas point" of each distribution. From smallest to largest: Cardi then Billie, and Ariana and Dua appear to be about the same. The equal areas point appears to be around 40 or so for Cardi. It's definitely less than 45, which appears to the equal areas point for Billie. The equal areas point for Ariana is 50 (halfway between 25 and 75), and Dua's appears to be about 50 also.
4. Ariana is most uncertain, then Dua, then Cardi, then Billie is the least uncertain. Each distribution represents 100% probability, but Ariana stretches this probability over the largest range of possible values, while Billie stretches this over the shortest. Ariana is basically saying the in-

structor can be any age between 25 and 75. Billie is fairly certain that the instructor is close to 45, and she's basically 100% certain that the instructor is between 35 and 55.

Example 3.7. Consider Ariana's subjective probability distribution in Figure 3.7. Ariana learns that her instructor received a Ph.D. in 2006. How would her subjective probability distribution change?

Solution. to Example 3.7

Show/hide solution

Ariana's original subjective probability distribution reflects very little knowledge about her instructor. Ariana now reasons that her instructor was probably between 25 and 35 when he received his Ph.D. in 2006, so she revises her subjective probability distribution to place almost 100% probability on ages between 40 and 50. Ariana's subjective probability distribution now looks more like Billie's in Figure 3.7.

The previous examples introduce how probability can be used to quantify uncertainty about unknown numbers. One key aspect of Bayesian analyses is applying a subjective probability distribution to a *parameter* in a statistical model.

Example 3.8. Let θ_b represent the proportion of current Cal Poly students who have ever read any of the books in the *Harry Potter* series. Let θ_m represent the proportion of current Cal Poly students who have ever seen any of the movies in the *Harry Potter* series.

1. Are θ_b and θ_m parameters or statistics? Why?
2. Are the values of θ_b and θ_m known or unknown, certain or uncertain?
3. What are the possible values of θ_b and θ_m?
4. Sketch a probability distribution representing what you think are more/less credible values of θ_b. Repeat for θ_m.
5. Are you more certain about the value of θ_b or θ_m? How is this reflected in your distributions?
6. Suppose that in a class of 35 Cal Poly students, 21 have read a Harry Potter book, and 30 have seen a Harry Potter movie. Now that we have observed some data, sketch a probability distribution representing what you think are more/less credible values of θ_b. Repeat for θ_m. How do your distributions after observing data compare to the distributions you sketched before?

Solution. to Example 3.8

Show/hide solution

1. The population of interest is current Cal Poly students, so θ_b and θ_m are *parameters*. We don't have relevant data for the entire population, but we could collect data on a sample.

2. Since we don't have data on the entire population, the values of θ_b and θ_m are unknown, uncertain.

3. θ_b and θ_m are proportions so they take values between 0 and 1. Any value on the continuous scale between 0 and 1 is theoretically possible, though the values are not equally plausible.

4. Results will vary, but here's my thought process. I think that a strong majority of Cal Poly students have seen at least one Harry Potter movie, maybe 80% or so. I wouldn't be that surprised if it were even close to 100%, but I would be pretty surprised if it were less than 60%.

   However, I'm less certain about θ_b. I suspect that fewer than 50% of students have read at least one Harry Potter book, but I'm not very sure and I wouldn't be too surprised if it were actually more than 50%.

   See Figure 3.8 for what my subjective probability distributions might look like.

5. I'm less certain about θ_b, so its density is "spread out" over a wider range of values.

6. The values of θ_b and θ_m are still unknown, but I am less uncertain about their values now that I have observed some data. The sample proportion who have watched a Harry Potter movie is $30/35 = 0.857$, which is pretty consistent with my initial beliefs. But now I update my subjective distribution to concentrate even more of my subjective probability on values in the 80 percent range.

   I had suspected that θ_b was less than 0.5, so the observed sample proportion of $21/35 = 0.6$ goes against my expectations. However, I was fairly uncertain about the value of θ_m prior to observing the data, so 0.6 is not too surprising to me. I update my subjective distribution so that it's centered closer to 0.6, while still allowing for my suspicion that θ_b is less than 0.5.

   See Figure 3.9 for what my subjective probability distributions might look like after observing the sample data. Of course, the sample proportions are not necessarily equal to the population proportions. But if the samples are reasonably representative, I would hope that the observed sample proportions are close to the respective population proportions. Even after observing data, there is still uncertainty about the parameters θ_b and θ_m, and my subjective distributions quantify this uncertainty.

Recall some statistical terminology.

- **Observational units** (a.k.a., cases, individuals, subjects) are the people, places, things, etc we collect information on.
- A **variable** is any characteristic of an observational unit that we can measure.

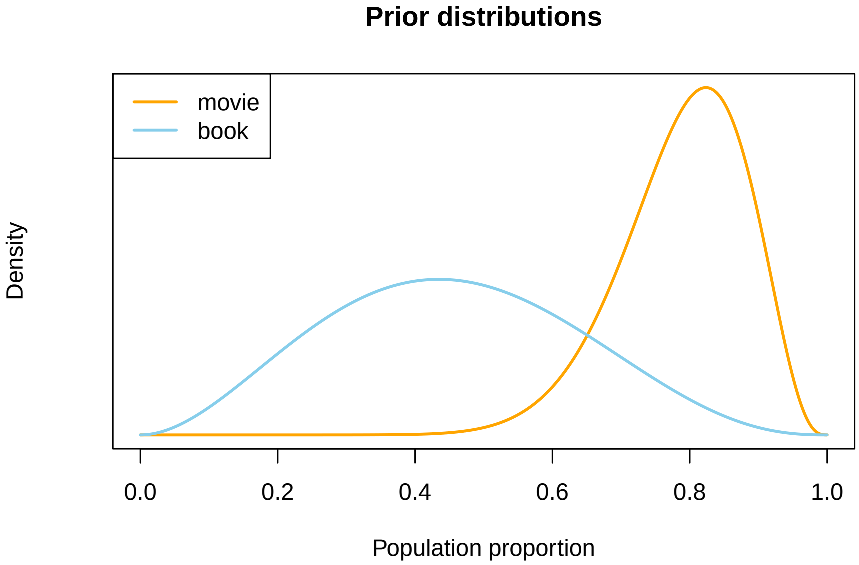


Figure 3.8: Example subjective distributions in Example 3.8, prior to observing sample data.

- **Statistical inference** involves using data collected on a *sample* to make conclusions about a *population*.

- Inference often concerns specific numerical summaries, using values of *statistics* to make conclusions about *parameters*.
- A **parameter** is a number that describes the **population**, e.g., *population mean*, *population proportion*. The actual value of a parameter is almost always *unknown*.
  - Parameters are often denoted with Greek letters. We'll often use the Greek letter θ ("theta") to denote a generic parameter.
- A **statistic** is a number that describes the **sample**, e.g., *sample mean*, *sample proportion*.

Parameters are unknown numbers. In "traditional", *frequentist* statistical analysis, parameters are treated as *fixed — that is, not random — constants*. Any randomness in a frequentist analysis arises from how the data were collected, e.g., via random sampling or random assignment. In a frequentist analysis, statistics are random variables; parameters are fixed numbers.

For example, a frequentist 95% confidence interval for θ_b in the previous example is [0.434, 0.766]. We estimate with 95% confidence that the proportion of Cal

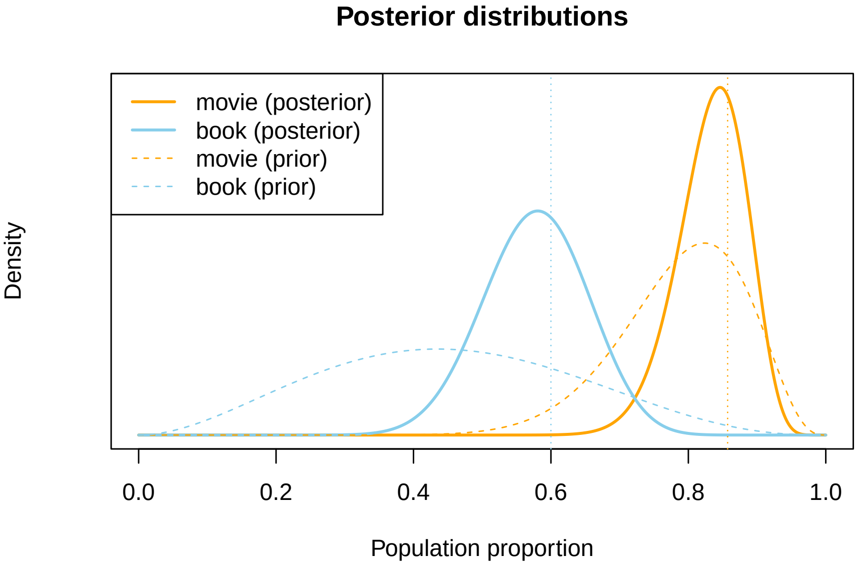


Figure 3.9: Example subjective distributions in Example 3.8, after observing sample data.

Poly students that have read any of the books in the Harry Potter series is between 0.434 and 0.766. Does this mean that there is a 95% probability that θ_b is between 0.434 and 0.766? No! In a frequentist analysis, the parameter θ_b is treated like a fixed constant. That constant is either between 0.434 and 0.766 or it's not; we don't know which it is, but there's no probability to it. In a frequentist analysis, it doesn't make sense to say "what is the probability that θ_b (a number) is between 0.434 and 0.766?" just like it doesn't make sense to say "what is the probability that 0.5 is between 0.434 and 0.766?" Remember that 95% confidence derives from the fact that for 95% *of samples* the procedure that was used to produce the interval [0.434, 0.766] will produce intervals that contain the true parameter θ_b. It is the samples and the intervals that are changing from sample to sample; θ_b stays constant at its fixed but unknown value. In a frequentist analysis, probability quantifies the *randomness in the sampling procedure.*

On the other hand, in a Bayesian statistical analysis, since a parameter θ is unknown — that is, it's value is *uncertain* to the observer — θ is treated as a *random variable.* That is, **in Bayesian statistical analyses unknown parameters are random variables that have probability distributions.** The probability distribution of a parameter quantifies the degree of uncertainty about the value of the parameter. Therefore, the Bayesian perspective allows for probability statements about parameters. For example, a Bayesian analysis of the previous example might conclude that there is a 95% chance that θ_b is between 0.426 and 0.721. Such a statement is valid in the Bayesian context, but nonsensical in the frequentist context.

In the previous example, we started with distributions that represented our uncertainty about θ_b and θ_m based on our "beliefs", then we revised these distributions after observing some data. If we were to observe more data, we could revise again. In this course we will see (among other things) (1) how to quantify uncertainty about parameters using probability distributions, and (2) how to update those distributions to reflect new data.

Throughout these notes we will focus on Bayesian statistical analyses. We will occasionally compare Bayesian and frequentist analyses and viewpoints. But we want to make clear from the start: Bayesian versus frequentist is NOT a question of right versus wrong. Both Bayesian and frequentist are valid approaches to statistical analyses, each with advantages and disadvantages. We'll address some of the issues along the way. But at no point in your career do you need to make a definitive decision to be a Bayesian or a frequentist; a good modern statistician is probably a bit of both.

Chapter 4

Bayes' Rule

The mechanism that underpins all of Bayesian statistical analysis is *Bayes' rule*[1], which describes how to update uncertainty in light of new information, evidence, or data.

Example 4.1. A recent survey of American adults asked: "Based on what you have heard or read, which of the following two statements best describes the scientific method?"

- 70% selected "The scientific method produces findings meant to be continually tested and updated over time". (We'll call this the "iterative" opinion.)
- 14% selected "The scientific method identifies unchanging core principles and truths". (We'll call this the "unchanging" opinion).
- 16% were not sure which of the two statements was best.

How does the response to this question change based on education level? Suppose education level is classified as: high school or less (HS), some college but no Bachelor's degree (college), Bachelor's degree (Bachelor's), or postgraduate degree (postgraduate). The education breakdown is

- Among those who agree with "iterative": 31.3% HS, 27.6% college, 22.9% Bachelor's, and 18.2% postgraduate.
- Among those who agree with "unchanging": 38.6% HS, 31.4% college, 19.7% Bachelor's, and 10.3% postgraduate.
- Among those "not sure": 57.3% HS, 27.2% college, 9.7% Bachelor's, and 5.8% postgraduate

[1] This section only covers Bayes' rule for events. We'll see Bayes' rule for distributions of random variables later. But the ideas are analogous.

1. Use the information to construct an appropriate two-way table.
2. Overall, what percentage of adults have a postgraduate degree? How is this related to the values 18.2%, 10.3%, and 5.8%?
3. What percent of those with a postgraduate degree agree that the scientific method is "iterative"? How is this related to the values provided?

Solution. to Example 4.1

Show/hide solution

1. Suppose there are 100000 hypothetical American adults. Of these 100000, $100000 \times 0.7 = 70000$ agree with the "iterative" statement. Of the 70000 who agree with the "iterative" statement, $70000 \times 0.182 = 12740$ also have a postgraduate degree. Continue in this way to complete the table below.
2. Overall 15.11% of adults have a postgraduate degree (15110/100000 in the table). The overall percentage is a weighted average of the three percentages; 18.2% gets the most weight in the average because the "iterative" statement has the highest percentage of people that agree with it compared to "unchanging" and "not sure".

$$0.1511 = (0.70)(0.182) + (0.14)(0.103) + (0.16)(0.058)$$

3. Of the 15110 who have a postgraduate degree 12740 agree with the "iterative" statement, and $12740/15110 = 0.843$. 84.3% of those with a graduate degree agree that the scientific method is "iterative". The value 0.843 is equal to the product of (1) 0.70, the overall proportion who agree with the "iterative" statement, and (2) 0.182, the proportion of those who agree with the "iterative" statement that have a postgraduate degree; divided by 0.1511, the overall proportion who have a postgraduate degree.

$$0.843 = \frac{0.182 \times 0.70}{0.1511}$$

	HS	college	Bachelors	postgrad	total
iterative	21910	19320	16030	12740	70000
unchanging	5404	4396	2758	1442	14000
not sure	9168	4352	1552	928	16000
total	36482	28068	20340	15110	100000

Bayes' rule for events specifies how a prior probability $P(H)$ of event H is updated in response to the evidence E to obtain the posterior probability $P(H|E)$.

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

- Event H represents a particular hypothesis[2] (or model or case)

[2] We're using "hypothesis" in the sense of a general scientific hypothesis, not necessarily a statistical null or alternative hypothesis.

- Event E represents observed evidence (or data or information)
- $P(H)$ is the unconditional or **prior probability** of H (prior to observing evidence E)
- $P(H|E)$ is the conditional or **posterior probability** of H after observing evidence E.
- $P(E|H)$ is the **likelihood** of evidence E given hypothesis (or model or case) H

Example 4.2. Continuing the previous example. Randomly select an American adult.

1. Consider the conditional probability that a randomly selected American adult agrees that the scientific method is "iterative" given that they have a postgraduate degree. Identify the prior probability, hypothesis, evidence, likelihood, and posterior probability, and use Bayes' rule to compute the posterior probability.
2. Find the conditional probability that a randomly selected American adult with a postgraduate degree agrees that the scientific method is "unchanging".
3. Find the conditional probability that a randomly selected American adult with a postgraduate degree is not sure about which statement is best.
4. How many times more likely is it for an *American adult* to have a postgraduate degree and agree with the "iterative" statement than to have a postgraduate degree and agree with the "unchanging" statement?
5. How many times more likely is it for an *American adult with a postgraduate degree* to agree with the "iterative" statement than to agree with the "unchanging" statement?
6. What do you notice about the answers to the two previous parts?
7. How many times more likely is it for an *American adult* to agree with the "iterative" statement than to agree with the "unchanging" statement?
8. How many times more likely is it for an American adult to have a postgraduate degree when the adult agrees with the iterative statement than when the adult agree with the unchanging statement?
9. How many times more likely is it for an *American adult with a postgraduate degree* to agree with the "iterative" statement than to agree with the "unchanging" statement?
10. How are the values in the three previous parts related?

Solution. to Example 4.2

Show/hide solution

1. This is essentially the same question as the last part of the previous problem, just with different terminology.

    - The hypothesis is H_1, the event that the randomly selected adult agrees with the "iterative" statement.

- The prior probability is $P(H_1) = 0.70$, the overall or unconditional probability that a randomly selected American adult agrees with the "iterative" statement.
- The given "evidence" E is the event that the randomly selected adult has a postgraduate degree. The marginal probability of the evidence is $P(E) = 0.1511$, which can be obtained by the law of total probability as in the previous problem.
- The likelihood is $P(E|H_1) = 0.182$, the conditional probability that the adult has a postgraduate degree (the evidence) given that the adult agrees with the "iterative" statement (the hypothesis).
- The posterior probability is $P(H_1|E) = 0.843$, the conditional probability that a randomly selected American adult agrees that the scientific method is "iterative" given that they have a postgraduate degree. By Bayes rule

$$P(H_1|E) = \frac{P(E|H_1)P(H_1)}{P(E)} = \frac{0.182 \times 0.70}{0.1511} = 0.843$$

2. Let H_2 be the event that the randomly selected adult agrees with the "unchanging" statement; the prior probability is $P(H_2) = 0.14$. The evidence E is still "postgraduate degree" but now the likelihood of this evidence is $P(E|H_2) = 0.103$ under the "unchanging" hypothesis. The conditional probability that a randomly selected adult with a postgraduate degree agrees that the scientific method is "unchanging" is

$$P(H_2|E) = \frac{P(E|H_2)P(H_2)}{P(E)} = \frac{0.103 \times 0.14}{0.1511} = 0.095$$

3. Let H_3 be the event that the randomly selected adult is "not sure"; the prior probability is $P(H_3) = 0.16$. The evidence E is still "postgraduate degree" but now the likelihood of this evidence is $P(E|H_3) = 0.058$ under the "not sure" hypothesis. The conditional probability that a randomly selected adult with a postgraduate degree is "not sure" is

$$P(H_3|E) = \frac{P(E|H_3)P(H_3)}{P(E)} = \frac{0.058 \times 0.16}{0.1511} = 0.061$$

4. The probability that an *American adult* has a postgraduate degree and agrees with the "iterative" statement is $P(E \cap H_1) = P(E|H_1)P(H_1) = 0.182 \times 0.70 = 0.1274$. The probability that an *American adult* has a postgraduate degree and agrees with the "unchanging" statement is $P(E \cap H_2) = P(E|H_2)P(H_2) = 0.103 \times 0.14 = 0.01442$. Since

$$\frac{P(E \cap H_1)}{P(E \cap H_2)} = \frac{0.182 \times 0.70}{0.103 \times 0.14} = \frac{0.1274}{0.01442} = 8.835$$

an *American adult* is 8.835 times more likely to have a postgraduate degree and agree with the "iterative" statement than to have a postgraduate degree and agree with the "unchanging" statement.

5. The conditional probability that an *American adult with a postgraduate degree* agrees with the "iterative" statement is $P(H_1|E) = P(E|H_1)P(H_1)/P(E) = 0.182 \times 0.70/0.1511 = 0.843$. The conditional probability that an *American adult with a postgraduate degree* agrees with the "unchanging" statement is $P(H_2|E) = P(E|H_2)P(H_2)/P(E) = 0.103 \times 0.14/0.1511 = 0.09543$. Since

$$\frac{P(H_1|E)}{P(H_2|E)} = \frac{0.182 \times 0.70/0.1511}{0.103 \times 0.14/0.1511} = \frac{0.84315}{0.09543} = 8.835$$

   an *American adult with a postgraduate degree* is 8.835 times more likely to agree with the "iterative" statement than to agree with the "unchanging" statement.
6. The ratios are the same! Conditioning on having a postgraduate degree just "slices" out the Americans who have a postgraduate degree. The ratios are determined by the overall probabilities for Americans. The conditional probabilities, given postgraduate degree, simply rescale the probabilities for Americans who have a postgraduate degree to add up to 1 (by dividing by 0.1511.)
7. This is a ratio of prior probabilities: 0.70 / 0.14 = 5. An *American adult* is 5 times more likely to agree with the "iterative" statement than to agree with the "unchanging" statement.
8. This is a ratio of likelihoods: 0.182 / 0.103 = 1.767. An American adult is 1.767 times more likely to have a postgraduate degree when the adult agrees with the iterative statement than when the adult agree with the unchanging statement.
9. This is a ratio of posterior probabilities: 0.8432 / 0.0954 = 8.835. An *American adult with a postgraduate degree* is 8.835 times more likely to agree with the "iterative" statement than to agree with the "unchanging" statement.
10. The ratio of the posterior probabilities is equal to the product of the ratio of the prior probabilities and the ratio of the likelihoods: $8.835 = 5 \times 1.767$. *Posterior is proportional to the product of prior and likelihood.*

Bayes rule is often used when there are multiple hypotheses or cases. Suppose $H_1, \ldots, H_k$ is a series of distinct hypotheses which together account for all possibilities[3], and E is any event (evidence). Then Bayes' rule implies that the posterior probability of any particular hypothesis H_j satisfies

$$P(H_j|E) = \frac{P(E|H_j)P(H_j)}{P(E)}$$

The marginal probability of the evidence, $P(E)$, in the denominator can be

[3] More formally, $H_1, \ldots, H_k$ is a *partition* which satisfies $P\left(\cup_{i=1}^{k} H_i\right) = 1$ and $H_1, \ldots, H_k$ are disjoint — $H_i \cap H_j = \emptyset, i \neq j$.

calculated using the *law of total probability*

$$P(E) = \sum_{i=1}^{k} P(E|H_i)P(H_i)$$

The law of total probability says that we can interpret the unconditional probability $P(E)$ as a probability-weighted average of the case-by-case conditional probabilities $P(E|H_i)$ where the weights $P(H_i)$ represent the probability of encountering each case.

Combining Bayes' rule with the law of total probability,

$$\begin{aligned} P(H_j|E) &= \frac{P(E|H_j)P(H_j)}{P(E)} \\ &= \frac{P(E|H_j)P(H_j)}{\sum_{i=1}^{k} P(E|H_i)P(H_i)} \end{aligned}$$

$$P(H_j|E) \propto P(E|H_j)P(H_j)$$

The symbol $\propto$ is read "is proportional to". The relative *ratios* of the posterior probabilities of different hypotheses are determined by the product of the prior probabilities and the likelihoods, $P(E|H_j)P(H_j)$. The marginal probability of the evidence, $P(E)$, in the denominator simply normalizes the numerators to ensure that the updated probabilities sum to 1 over all the distinct hypotheses.

In short, Bayes' rule says[4]

$$\textbf{posterior} \propto \textbf{likelihood} \times \textbf{prior}$$

In the previous examples, the prior probabilities for an American adult's perception of the scientific method are 0.70 for "iterative", 0.14 for "unchanging", and 0.16 for "not sure". After observing that the American has a postgraduate degree, the posterior probabilities for an American adult's perception of the scientific method become 0.8432 for "iterative", 0.0954 for "unchanging", and 0.0614 for "not sure". The following organizes the calculations in a **Bayes' table** which illustrates "posterior is proportional to likelihood times prior".

hypothesis	prior	likelihood	product	posterior
iterative	0.70	0.182	0.1274	0.8432
unchanging	0.14	0.103	0.0144	0.0954
not sure	0.16	0.058	0.0093	0.0614
sum	1.00	NA	0.1511	1.0000

The likelihood column depends on the evidence, in this case, observing that the American has a postgraduate degree. This column contains the probability of

[4] "Posterior is proportional to likelihood times prior" summarizes the whole course in a single sentence.

the same event, $E =$ "the American has a postgraduate degree", under each of the distinct hypotheses:

- $P(E|H_1) = 0.182$, given the American agrees with the "iterative" statement
- $P(E|H_2) = 0.103$, given the American agrees with the "unchanging" statement
- $P(E|H_3) = 0.058$, given the American is "not sure"

Since each of these probabilities is computed under a different case, these values do not need to add up to anything in particular. The sum of the likelihoods is meaningless, which is why we have listed a sum of "NA" for the likelihood column.

The "product" column contains the product of the values in the prior and likelihood columns. The product of prior and likelihood for "iterative" (0.1274) is 8.835 (0.1274/0.0144) times higher than the product of prior and likelihood for "unchanging" (0.0144). Therefore, Bayes rule implies that the conditional probability that an American with a postgraduate degree agrees with "iterative" should be 8.835 times higher than the conditional probability that an American with a postgraduate degree agrees with "unchanging". Similarly, the conditional probability that an American with a postgraduate degree agrees with "iterative" should be $0.1274/0.0093 = 13.73$ times higher than the conditional probability that an American with a postgraduate degree is "not sure", and the conditional probability that an American with a postgraduate degree agrees with "unchanging" should be $0.0144/0.0093 = 1.55$ times higher than the conditional probability that an American with a postgraduate degree is "not sure". The last column just translates these relative relationships into probabilities that sum to 1.

The sum of the "product" column is $P(E)$, the marginal probability of the evidence. The sum of the product column represents the result of the law of total probability calculation. However, for the purposes of determining the posterior probabilities, it isn't really important what $P(E)$ is. Rather, it is the *ratio* of the values in the "product" column that determine the posterior probabilities. $P(E)$ is whatever it needs to be to ensure that the posterior probabilities sum to 1 while maintaining the proper ratios.

The process of conditioning can be thought of as **"slicing and renormalizing".**

- Extract the "slice" corresponding to the event being conditioned on (and discard the rest). For example, a slice might correspond to a particular row or column of a two-way table.

- "Renormalize" the values in the slice so that corresponding probabilities add up to 1.

We will see that the "slicing and renormalizing" interpretation also applies when dealing with conditional distributions of random variables, and corresponding plots. Slicing determines the *shape*; renormalizing determines the *scale*. Slicing determines relative probabilities; renormalizing just makes sure they "add up" to 1 while maintaining the proper ratios.

Example 4.3. Now suppose we want to compute the posterior probabilities for an American adult's perception of the scientific method given that the randomly selected American adult has some college but no Bachelor's degree ("college").

1. Before computing, make an educated guess for the posterior probabilities. In particular, will the changes from prior to posterior be more or less extreme given the American has some college but no Bachelor's degree than when given the American has a postgraduate degree? Why?
2. Construct a Bayes table and compute the posterior probabilities. Compare to the posterior probabilities given postgraduate degree from the previous examples.

Solution. to Example 4.3

Show/hide solution

1. We start with the same prior probabilities as before: 0.70 for iterative, 0.14 for unchanging, 0.16 for not sure. Now the evidence is that the American has some college but no Bachelor's degree. The likelihood of the evidence ("college") is 0.276 under the iterative hypothesis, 0.314 under the unchanging hypothesis, and 0.272 under the not sure hypothesis. The likelihood of the evidence does not change as much across the different hypotheses when the evidence is "college" than when the evidence was "postgraduate degree". Therefore, the changes from prior to posterior should be less extreme when the evidence is "college" than when the evidence was "postgraduate degree". Furthermore, since the likelihood doesn't vary much across hypotheses when the evidence is "college" we expect the posterior probabilities to be close to the prior probabilities.
2. See the table below. As expected, the posterior probabilities are closer to the prior probabilities when the evidence is "college" than when the evidence is "postgraduate degree".

```
hypothesis = c("iterative", "unchanging", "not sure")

prior = c(0.70, 0.14, 0.16)

likelihood = c(0.276, 0.314, 0.272) # likelihood of college

product = prior * likelihood
```

```
posterior = product / sum(product)

bayes_table = data.frame(hypothesis,
                         prior,
                         likelihood,
                         product,
                         posterior) %>%
  add_row(hypothesis = "sum",
          prior = sum(prior),
          likelihood = NA,
          product = sum(product),
          posterior = sum(posterior))

kable(bayes_table, digits = 4, align = 'r')
```

hypothesis	prior	likelihood	product	posterior
iterative	0.70	0.276	0.1932	0.6883
unchanging	0.14	0.314	0.0440	0.1566
not sure	0.16	0.272	0.0435	0.1551
sum	1.00	NA	0.2807	1.0000

Like the scientific method, Bayesian analysis is often an iterative process. Posterior probabilities are updated after observing some information or data. These probabilities can then be used as prior probabilities before observing new data. Posterior probabilities can be sequentially updated as new data becomes available, with the posterior probabilities after the previous stage serving as the prior probabilities for the next stage. The final posterior probabilities only depend upon the cumulative data. It doesn't matter if we sequentially update the posterior after each new piece of data or only once after all the data is available; the final posterior probabilities will be the same either way. Also, the final posterior probabilities are not impacted by the order in which the data are observed.

Chapter 5

Introduction to Estimation

A parameter is a number that describes the population, e.g., population mean, population proportion. The actual value of a parameter is almost always unknown.

A statistic is a number that describes the sample, e.g., sample mean, sample proportion. We can use observed sample statistics to estimate unknown population parameters.

Example 5.1. Most people are right-handed, and even the right eye is dominant for most people. In a 2003 study reported in *Nature*, a German bio-psychologist conjectured that this preference for the right side manifests itself in other ways as well. In particular, he investigated if people have a tendency to lean their heads to the right when kissing. The researcher observed kissing couples in public places and recorded whether the couple leaned their heads to the right or left. (We'll assume this represents a randomly selected representative sample of kissing couples.)

The parameter of interest in this study is the population proportion of kissing couples who lean their heads to the right. Denote this unknown parameter θ; our goal is to estimate θ based on sample data.

Let Y be the number of couples in a random sample of n kissing couples that lean to right. Suppose that in a sample of $n = 12$ couples $y = 8$ leaned to the right. We'll start with a non-Bayesian analysis.

1. If you were to estimate θ with a single number based on this sample data alone, intuitively what number would you pick?
2. For a general n and θ, what is the distribution of Y?
3. For the next few parts suppose $n = 12$. For a moment we'll only consider these potential values for θ: $0.1, 0.3, 0.5, 0.7, 0.9$. If $\theta = 0.1$ what is the distribution of Y? Compute and interpret the probability that $Y = 8$ if $\theta = 0.1$.

4. If $\theta = 0.3$ what is the distribution of Y? Compute and interpret the probability that $Y = 8$ if $\theta = 0.3$.
5. If $\theta = 0.5$ what is the distribution of Y? Compute and interpret the probability that $Y = 8$ if $\theta = 0.5$.
6. If $\theta = 0.7$ what is the distribution of Y? Compute and interpret the probability that $Y = 8$ if $\theta = 0.7$.
7. If $\theta = 0.9$ what is the distribution of Y? Compute and interpret the probability that $Y = 8$ if $\theta = 0.9$.
8. Now remember that θ is unknown. If you had to choose your estimate of θ from the values $0.1, 0.3, 0.5, 0.7, 0.9$, which one of these values would you choose based on of observing $y = 8$ couples leaning to the right in a sample of 12 kissing couples? Why?
9. Obviously our choice is not restricted to those five values of θ. Describe in principle the process you would follow to find the estimate of θ based on of observing $y = 8$ couples leaning to the right in a sample of 12 kissing couples.
10. Let $f(y|\theta)$ denote the probability of observing y couples leaning to the right in a sample of 12 kissing couples. Determine $f(y = 8|\theta)$ and sketch a graph of it. What is this a function of? What is an appropriate name for this function?
11. From the previous part, what seems like a reasonable estimate of θ based solely on the data of observing $y = 8$ couples leaning to the right in a sample of 12 kissing couples?

Solution. to Example 5.1

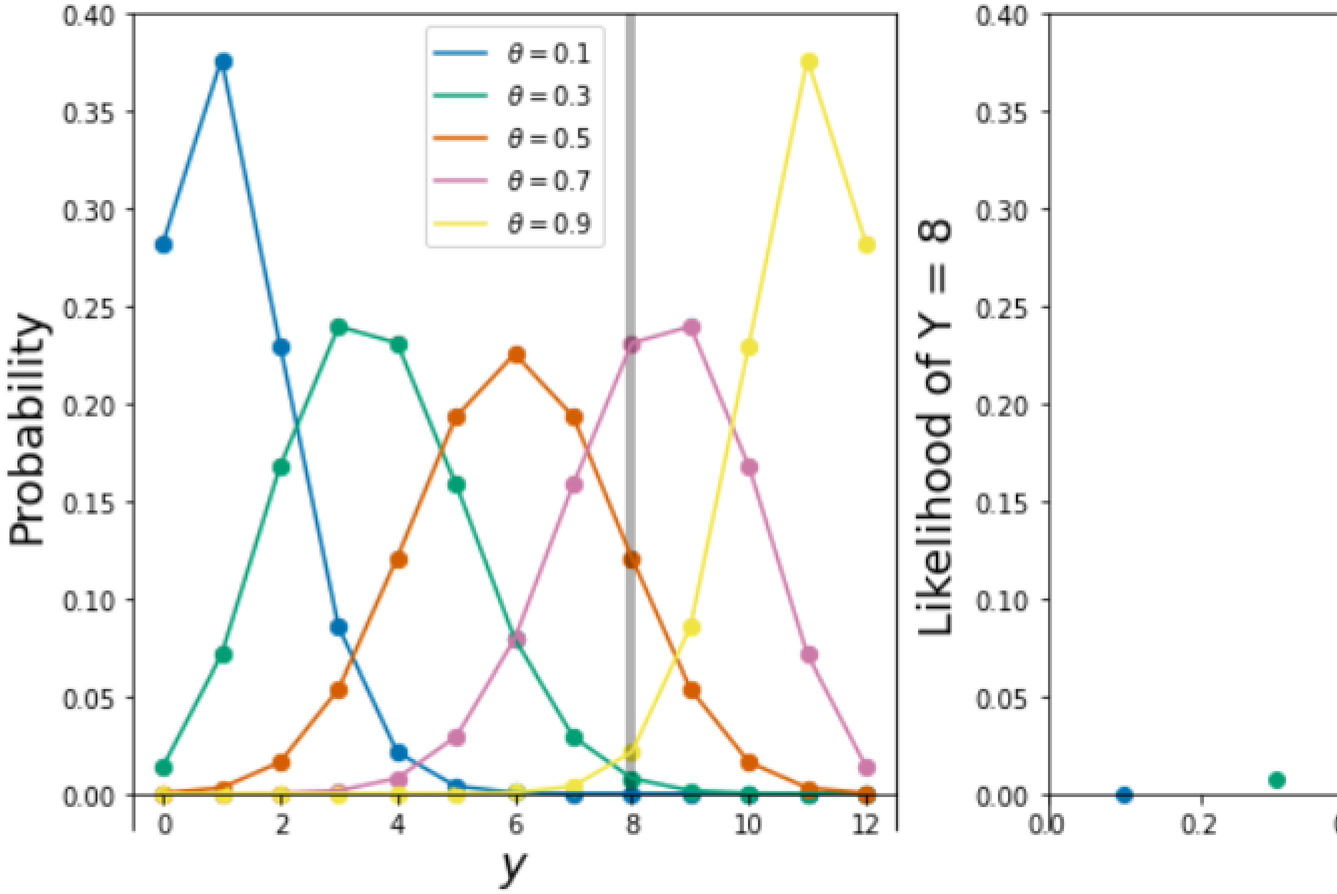
Show/hide solution

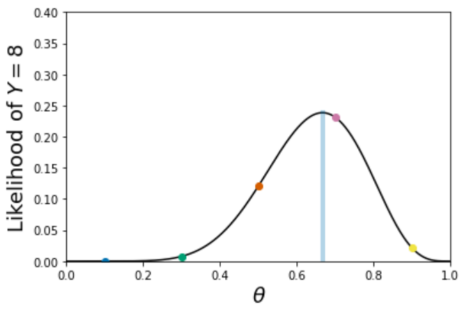
1. Seems reasonable to use the sample proportion $8/12 = 0.667$.
2. Y has a Binomial(n, θ) distribution.
3. If $\theta = 0.1$ then Y has a Binomial(12, 0.1) distribution and $P(Y = 8|\theta = 0.1) = \binom{12}{8}0.1^8(1-0.1)^{12-8} \approx 0.000$; `dbinom(8, 12, 0.1)`.
4. If $\theta = 0.3$ then Y has a Binomial(12, 0.3) distribution and $P(Y = 8|\theta = 0.3) = \binom{12}{8}0.3^8(1-0.3)^{12-8} \approx 0.008$; `dbinom(8, 12, 0.3)`.
5. If $\theta = 0.5$ then Y has a Binomial(12, 0.5) distribution and $P(Y = 8|\theta = 0.5) = \binom{12}{8}0.5^8(1-0.5)^{12-8} \approx 0.121$; `dbinom(8, 12, 0.5)`.
6. If $\theta = 0.7$ then Y has a Binomial(12, 0.7) distribution and $P(Y = 8|\theta = 0.7) = \binom{12}{8}0.7^8(1-0.7)^{12-8} \approx 0.231$; `dbinom(8, 12, 0.7)`.
7. If $\theta = 0.9$ then Y has a Binomial(12, 0.9) distribution and $P(Y = 8|\theta = 0.9) = \binom{12}{8}0.9^8(1-0.9)^{12-8} \approx 0.021$; `dbinom(8, 12, 0.9)`.
8. Compare the above values, and see the plots below. The probability of observing $y = 8$ is greatest when $\theta = 0.7$, so in some sense the data seems most "consistent" with $\theta = 0.7$. The sample that we actually observed has the greatest likelihood of occurring when $\theta = 0.7$ (among these choices for θ). From a different perspective, if $\theta = 0.1$ then the likelihood of observing 8 successes in a sample of size 12 is very small. Therefore, since

we actually did observed 8 successes in a sample of size 12, the data do not seem consistent with $\theta = 0.1$.

9. For each value of θ between 0 and 1 compute the probability of observing $y = 8$, $P(Y = 8|\theta)$, and find which value of θ maximizes this probability.

10. $f(y = 8|\theta) = P(Y = 8|\theta) = \binom{12}{8}\theta^8(1-\theta)^{12-8}$. This is a function of θ, with the data $y = 8$ fixed. Since this function computes the likelihood of observing the data (evidence) under different values of θ, "likelihood function" seems like an appropriate name. See the plots below.
11. The value which maximizes the likelihood of $y = 8$ is $8/12$. So the maximum likelihood estimate of θ is $8/12$.





- For given data y, the **likelihood function** $f(y|\theta)$ is the probability (or density for continuous data) of observing the sample data y viewed as a *function of the parameter* θ.
- In the likelihood function, the observed value of the data y is treated as a fixed constant.
- The value of a parameter that maximizes the likelihood function is called a **maximum likelihood estimate** (MLE).
- The MLE depends on the data y. For given data y, the MLE is the value of θ which gives the largest likelihood of having produced the observed data y.
- Maximum likelihood estimation is a common *frequentist* technique for estimating the value of a parameter based on data from a sample.

Example 5.2. We'll now take a Bayesian approach to estimating θ in Example 5.1. We treat the unknown parameter θ as a *random variable* and wish to find its posterior distribution after observing $y = 8$ couples leaning to the right in a sample of 12 kissing couples.

We will start with a very simplified, unrealistic prior distribution that assumes only five possible, equally likely values for θ: 0.1, 0.3, 0.5, 0.7, 0.9.

1. Sketch a plot of the prior distribution and fill in the prior column of the Bayes table.
2. Now suppose that $y = 8$ couples in a sample of size $n = 12$ lean right. Sketch a plot of the likelihood function and fill in the likelihood column in the Bayes table.
3. Complete the Bayes table and sketch a plot of the posterior distribution. What does the posterior distribution say about θ? How does it compare to the prior and the likelihood? If you had to estimate θ with a single number based on this posterior distribution, what number might you pick?
4. Now consider a prior distribution which places probability 1/9, 2/9, 3/9, 2/9, 1/9 on the values 0.1, 0.3, 0.5, 0.7, 0.9, respectively. Redo the previous parts. How does the posterior distribution change?
5. Now consider a prior distribution which places probability 5/15, 4/15, 3/15, 2/15, 1/15 on the values 0.1, 0.3, 0.5, 0.7, 0.9, respectively. Redo the previous parts. How does the posterior distribution change?

Solution. to Example 5.2

1. See plot below; the prior is "flat".

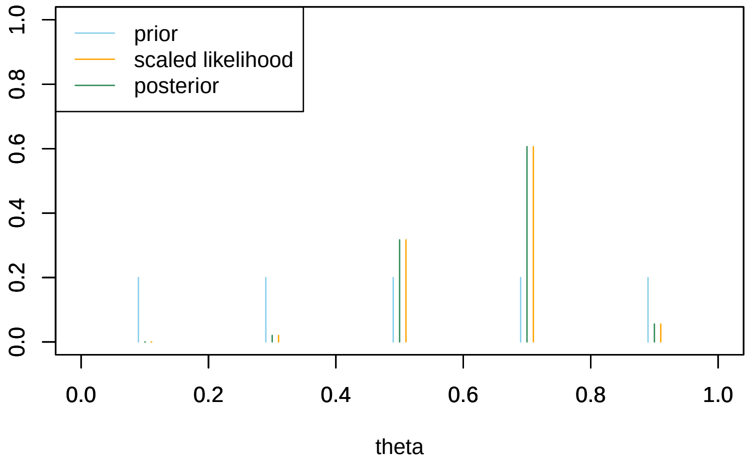
2. The likelihood is computed as in Example 5.1. See the plots above.

3. See the Bayes table below. Since the prior is flat, the posterior is proportional to the likelihood. The values 0.7 and 0.5 account for the bulk of posterior plausibility, and 0.7 is about twice as plausible as 0.5. If we had to estimate θ with a single number, we might pick 0.7 because that has the highest posterior probability.

```
theta = seq(0.1, 0.9, 0.2)

# prior

prior = rep(1, length(theta))
prior = prior / sum(prior)

# data
n = 12 # sample size
y = 8 # sample count of success

# likelihood, using binomial
likelihood = dbinom(y, n, theta) # function of theta

# posterior
product = likelihood * prior
posterior = product / sum(product)

# bayes table
bayes_table = data.frame(theta,
                         prior,
                         likelihood,
                         product,
                         posterior)

kable(bayes_table %>%
    adorn_totals("row"),
  digits = 4,
  align = 'r')
```

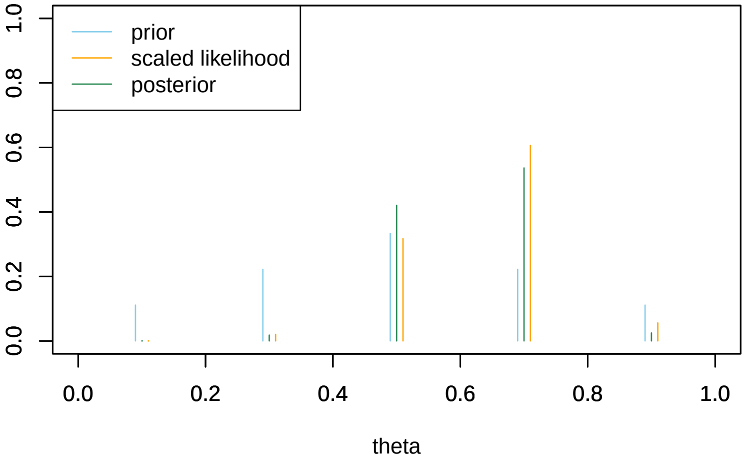
theta	prior	likelihood	product	posterior
0.1	0.2	0.0000	0.0000	0.0000
0.3	0.2	0.0078	0.0016	0.0205
0.5	0.2	0.1208	0.0242	0.3171
0.7	0.2	0.2311	0.0462	0.6065
0.9	0.2	0.0213	0.0043	0.0559
Total	1.0	0.3811	0.0762	1.0000

```
# plots
plot(theta-0.01, prior, type='h', xlim=c(0, 1), ylim=c(0, 1), col="skyblue", xlab=
par(new=T)
plot(theta+0.01, likelihood/sum(likelihood), type='h', xlim=c(0, 1), ylim=c(0, 1),
par(new=T)
plot(theta, posterior, type='h', xlim=c(0, 1), ylim=c(0, 1), col="seagreen", xlab=
legend("topleft", c("prior", "scaled likelihood", "posterior"), lty=1, col=c("skyb
```



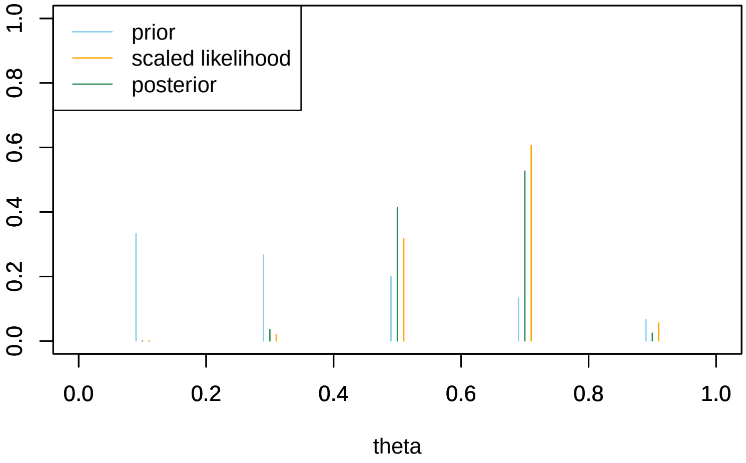
4. See table and plot below. Because the prior probability is now greater for 0.5 than for 0.7, the posterior probability of $\theta = 0.5$ is greater than in the previous part, and the posterior probability of $\theta = 0.7$ is less than in the previous part. The values 0.7 and 0.5 still account for the bulk of posterior plausibility, but now 0.7 is only about 1.3 times more plausible than 0.5. If we had to estimate θ with a single number, we might pick 0.7 because that has the highest posterior probability.

theta	prior	likelihood	product	posterior
0.1	0.1111	0.0000	0.0000	0.0000
0.3	0.2222	0.0078	0.0017	0.0181
0.5	0.3333	0.1208	0.0403	0.4207
0.7	0.2222	0.2311	0.0514	0.5365
0.9	0.1111	0.0213	0.0024	0.0247
Total	1.0000	0.3811	0.0957	1.0000



5. See the table and plot below. The prior probability is large for 0.1 and 0.3, but since the likelihood corresponding to these values is so small, the posterior probabilities are small. This posterior distribution is similar to the one from the previous part.

theta	prior	likelihood	product	posterior
0.1	0.3333	0.0000	0.0000	0.0000
0.3	0.2667	0.0078	0.0021	0.0356
0.5	0.2000	0.1208	0.0242	0.4132
0.7	0.1333	0.2311	0.0308	0.5269
0.9	0.0667	0.0213	0.0014	0.0243
Total	1.0000	0.3811	0.0585	1.0000



Bayesian estimation

- Regards parameters as *random variables* with probability distributions
- Assigns a subjective **prior distribution** to parameters
- Conditions on the observed data
- Applies Bayes' rule to produce a **posterior distribution** for parameters

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

- Determines parameter estimates from the posterior distribution

In a Bayesian analysis, the *posterior distribution* contains all relevant information about parameters. That is, all Bayesian inference is based on the posterior distribution. The posterior distribution is a compromise between

- prior "beliefs", as represented by the prior distribution
- data, as represented by the likelihood function

In contrast, a frequentist approach regards parameters as unknown but fixed (not random) quantities. Frequentist estimates are commonly determined by the likelihood function.

It is helpful to plot prior, likelihood, and posterior on the same plot. Since prior and likelihood are probability distributions, they are on the same scale. However, remember that the likelihood does not add up to anything in particular. To put the likelihood on the same scale as prior and posterior, it is helpful to rescale the likelihood so that it adds up to 1. Such a rescaling does not change the shape of the likelihood, it merely allows for easier comparison with prior and posterior.

Example 5.3. Continuing Example 5.2. While the previous exercise introduced the main ideas, it was unrealistic to consider only five possible values of θ.

1. What are the *possible* values of θ? Does the *parameter* θ take values on a continuous or discrete scale? (Careful: we're talking about the parameter and not the data.)
2. Let's assume that any multiple of 0.0001 is a possible value of θ: $0, 0.0001, 0.0002, \dots, 0.9999, 1$. Assume a discrete uniform prior distribution on these values. Suppose again that $y = 8$ couples in a sample of $n = 12$ kissing couples lean right. Use software to plot the prior distribution, the (scaled) likelihood function, and then find the posterior and plot it. Describe the posterior distribution. What does it say about θ? If you had to estimate θ with a single number based on this posterior distribution, what number might you pick?
3. Now assume a prior distribution which is proportional to $1 - 2|\theta - 0.5|$ for $\theta = 0, 0.0001, 0.0002, \dots, 0.9999, 1$. Use software to plot this prior; what does it say about θ? Then suppose again that $y = 8$ couples in a sample of $n = 12$ kissing couples lean right. Use software to plot the prior distribution, the (scaled) likelihood function, and then find the posterior and plot it. What does the posterior distribution say about θ? If you had to estimate θ with a single number based on this posterior distribution, what number might you pick?
4. Now assume a prior distribution which is proportional to $1 - \theta$ for $\theta = 0, 0.0001, 0.0002, \dots, 0.9999, 1$. Use software to plot this prior; what does it say about θ? Then suppose again that $y = 8$ couples in a sample of $n = 12$ kissing couples lean right. Use software to plot the prior distribution, the (scaled) likelihood function, and then find the posterior and plot it. What does the posterior distribution say about θ? If you had to estimate θ with a single number based on this posterior distribution, what number might you pick?
5. Compare the posterior distributions corresponding to the three different priors. How does each posterior distribution compare to the prior and the likelihood? Does the prior distribution influence the posterior distribution?

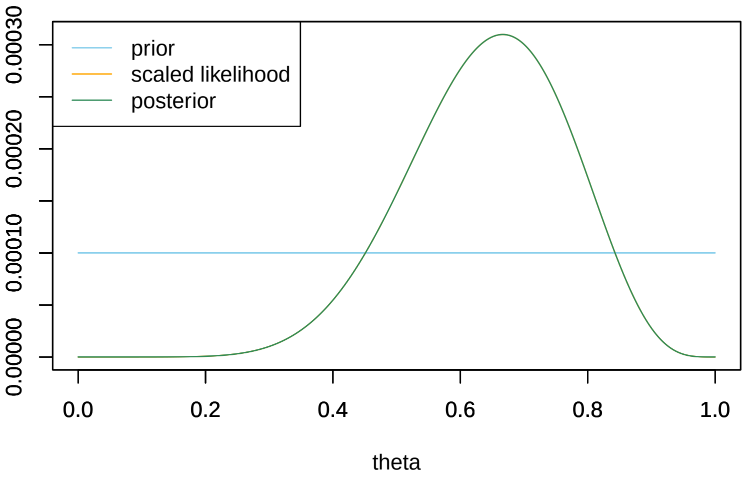
Solution. to Example 5.3

1. The *parameter* θ is a proportion, so it can possibly take any value in the continuous interval from 0 to 1.

2. See plot below. Since the prior is flat, the posterior is proportional to the likelihood. So the posterior distribution places highest posterior probability on values near the sample proportion 8/12. The interval of values

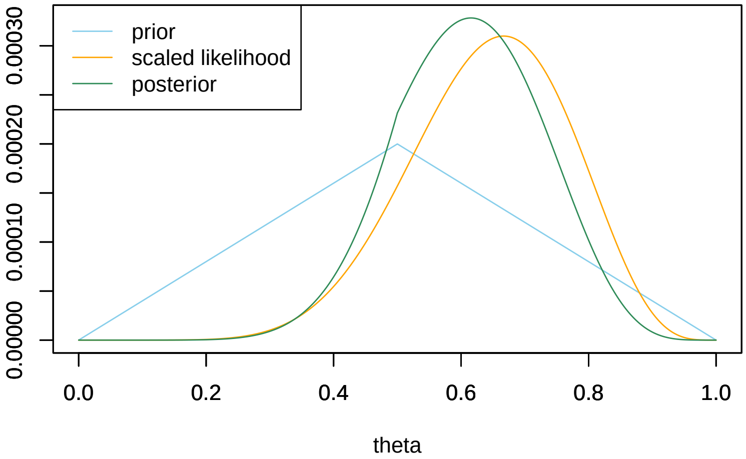
from about 0.4 to 0.9 accounts for almost all of the posterior plausibility. If we had to estimate θ with a single number, we might pick $8/12 = 0.667$ because that has the highest posterior probability.

```
theta = seq(0, 1, 0.0001)

# prior
prior = rep(1, length(theta))
prior = prior / sum(prior)

# data
n = 12 # sample size
y = 8 # sample count of success

# likelihood, using binomial
likelihood = dbinom(y, n, theta) # function of theta


# plots
plot_posterior <- function(theta, prior, likelihood){

  # posterior
  product = likelihood * prior
  posterior = product / sum(product)

  ylim = c(0, max(c(prior, posterior, likelihood / sum(likelihood))))
  plot(theta, prior, type='l', xlim=c(0, 1), ylim=ylim, col="skyblue", xlab='theta
  par(new=T)
  plot(theta, likelihood/sum(likelihood), type='l', xlim=c(0, 1), ylim=ylim, col="
  par(new=T)
  plot(theta, posterior, type='l', xlim=c(0, 1), ylim=ylim, col="seagreen", xlab="
  legend("topleft", c("prior", "scaled likelihood", "posterior"), lty=1, col=c("sk
}

plot_posterior(theta, prior, likelihood)
```



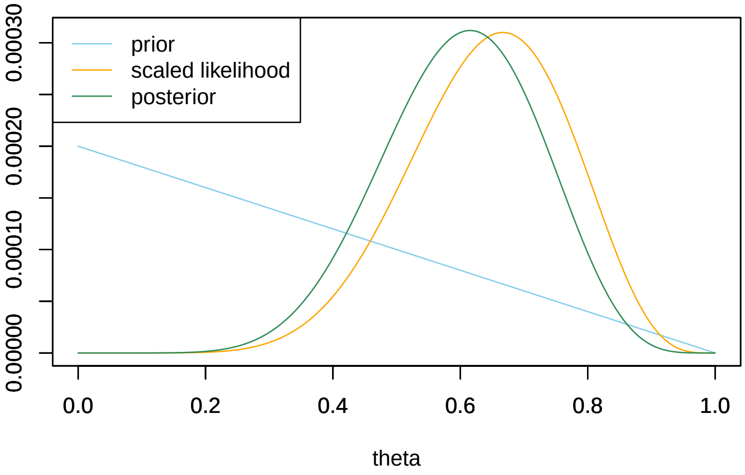
3. See plot below. The posterior is a compromise between the “triangular” prior which places highest prior probability near 0.5, and the likelihood. For this posterior, the posterior probability is greater near 0.5 than for the one in the previous part; the posterior here has been “shifted” towards 0.5 relative to the posterior from the previous part. If we had to estimate θ with a single number, we might pick 0.615 because that has the highest posterior probability.

```
# prior
theta = seq(0, 1, 0.0001)
prior = 1 - 2 * abs(theta - 0.5)
prior = prior / sum(prior)

# data
n = 12 # sample size
y = 8 # sample count of success

# likelihood, using binomial
likelihood = dbinom(y, n, theta) # function of theta


# plots
plot_posterior(theta, prior, likelihood)
```



4. Again the posterior is a compromise between prior and likelihood. The prior probabilities are greatest for values of θ near 0; however, the likelihood corresponding to these values is small, so the posterior probabilities are close to 0. As in the previous part, some of the posterior probability is shifted towards 0.5, as opposed to what happens with the uniform prior. The posterior here is fairly similar to the one from the previous part, and the maximum posterior probability still occurs around 0.61.

```
theta = seq(0, 1, 0.0001)

# prior
prior = 1 - theta
prior = prior / sum(prior)

# data
n = 12 # sample size
y = 8 # sample count of success

# likelihood, using binomial
likelihood = dbinom(y, n, theta) # function of theta

# plots
plot_posterior(theta, prior, likelihood)
```



5. For the “flat” prior, the posterior is proportional to the likelihood. For the other priors, the posterior is a compromise between prior and likelihood. The prior does have some influence. We do see three somewhat different posterior distributions corresponding to these three prior distributions.

- Even in situations where the data are discrete (e.g., binary success/failure data, count data), most statistical *parameters* take values on a *continuous* scale.
- Thus in a Bayesian analysis, parameters are usually *continuous random variables*, and have *continuous probability distributions*, a.k.a., *densities*.
- An alternative to dealing with continuous distributions is to use **grid approximation**: Treat the parameter as discrete, on a sufficiently fine grid of values, and use discrete distributions.

Example 5.4. Continuing Example 5.1. Now we’ll perform a Bayesian analysis on the actual study data in which 80 couples out of a sample of 124 leaned right. We’ll again use a grid approximation and assume that any multiple of 0.0001 between 0 and 1 is a possible value of θ: $0, 0.0001, 0.0002, \ldots, 0.9999, 1$.

1. Before performing the Bayesian analysis, use software to plot the likelihood when $y = 80$ couples in a sample of $n = 124$ kissing couples lean right, and compute the maximum likelihood estimate of θ based on this data. How does the likelihood for this sample compare to the likelihood based on the smaller sample (8/12) from previous exercises?
2. Now back to Bayesian analysis. Assume a discrete uniform prior distribution for θ. Suppose that $y = 80$ couples in a sample of $n = 124$ kissing

couples lean right. Use software to plot the prior distribution, the likelihood function, and then find the posterior and plot it. Describe the posterior distribution. What does it say about θ?

3. Now assume a prior distribution which is proportional to $1 - 2|\theta - 0.5|$ for $\theta = 0, 0.0001, 0.0002, \ldots, 0.9999, 1$. Then suppose again that $y = 80$ couples in a sample of $n = 124$ kissing couples lean right. Use software to plot the prior distribution, the likelihood function, and then find the posterior and plot it. What does the posterior distribution say about θ?
4. Now assume a prior distribution which is proportional to $1 - \theta$ for $\theta = 0, 0.0001, 0.0002, \ldots, 0.9999, 1$. Then suppose again that $y = 80$ couples in a sample of $n = 124$ kissing couples lean right. Use software to plot the prior distribution, the likelihood function, and then find the posterior and plot it. What does the posterior distribution say about θ?
5. Compare the posterior distributions corresponding to the three different priors. How does each posterior distribution compare to the prior and the likelihood? Comment on the influence that the prior distribution has. Does the Bayesian inference for these data appear to be highly sensitive to the choice of prior? How does this compare to the $n = 12$ situation?
6. If you had to produce a single number Bayesian estimate of θ based on the sample data, what number might you pick?

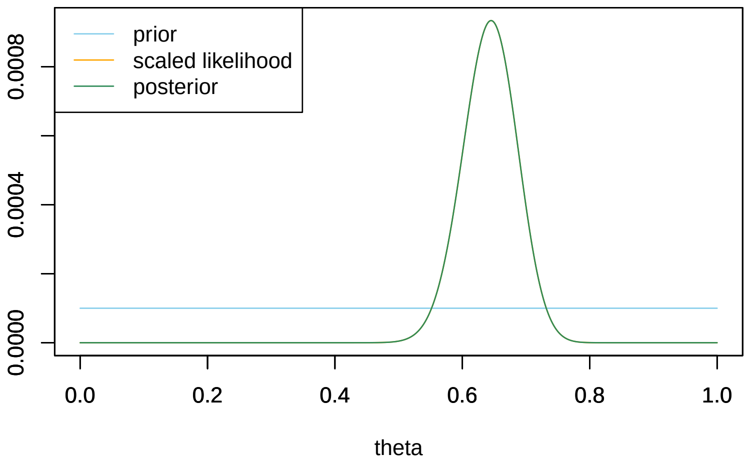
Solution. to Example 5.4

1. See plot below. The likelihood function is $f(y = 80|\theta) = \binom{124}{80}\theta^{80}(1-\theta)^{124-80}, 0 \leq \theta \leq 1$, the likelihood of observing a value of $y = 80$ from a Binomial(124, θ) distribution (`dbinom(80, 124, theta)`). The maximum likelihood estimate of θ is the sample proportion $80/124 = 0.645$. With the largest sample, the likelihood function is more "peaked" around its maximum.

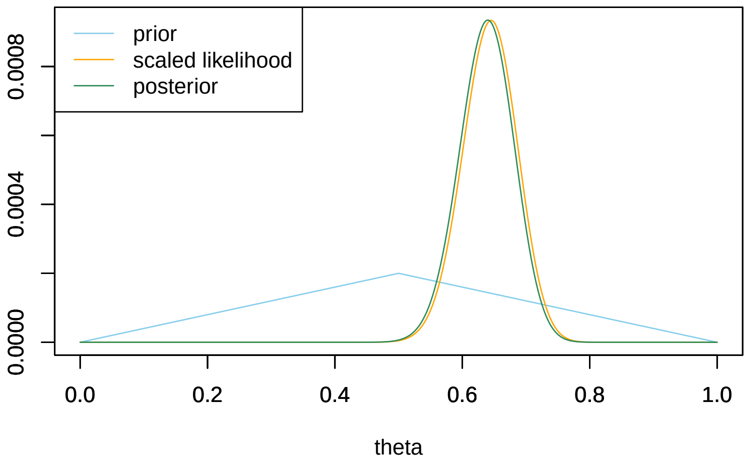
2. See plot below. Since the prior is flat, the posterior is proportional to the likelihood. The posterior places almost all of its probability on θ values between about 0.55 and 0.75, with the highest probability near the observed sample proportion of 0.645.

```
# prior
theta = seq(0, 1, 0.0001)
prior = rep(1, length(theta))
prior = prior / sum(prior)

# data
n = 124 # sample size
y = 80 # sample count of success

# likelihood, using binomial
likelihood = dbinom(y, n, theta) # function of theta
```

```
# plots
plot_posterior(theta, prior, likelihood)
```

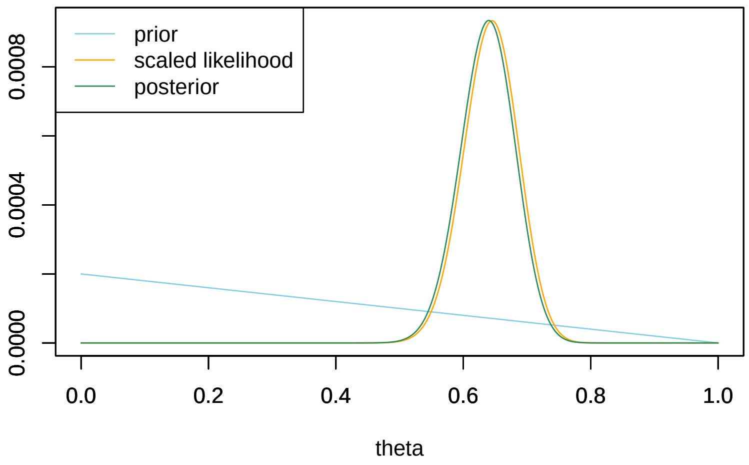


3. See the plot below. The posterior is very similar to the one from the previous part.



4. See the plot below. The posterior is very similar to the one from the

previous part.



5. Even though the priors are different, the posterior distributions are all similar to each other and all similar to the shape of the likelihood. Comparing these priors it does not appear that the posterior is highly sensitive to choice of prior. The data carry more weight when $n = 124$ than it did when $n = 12$. In other words, the prior has less influence when the sample size is larger. When the sample size is larger, the likelihood is more "peaked" and so the likelihood, and hence posterior, is small outside a narrower range of values than when the sample size is small.

6. It seems that regardless of the prior, the posterior distribution is about the same, yielding the highest posterior probability around the sample proportion 0.645. So if we had to estimate θ with a single number, we might just choose the sample proportion. In this case, we end up with the same numerical estimate of θ in both the Bayesian and frequentist analysis. But the process and interpretation differs between the two approaches; we'll discuss in more detail soon.

5.1 Point estimation

In a Bayesian analysis, the posterior distribution contains all relevant information about parameters after observing sample data. We often use certain summary characteristics of the posterior distribution to make inferences about parameters.

Example 5.5. Continuing the kissing study in Example 5.2 where θ can only take values 0.1, 0.3, 0.5, 0.7, 0.9. Consider a prior distribution which places probability 5/15, 4/15, 3/15, 2/15, 1/15 on the values 0.1, 0.3, 0.5, 0.7, 0.9, respectively. Suppose we want a single number *point estimate* of θ. What are some reasonable choices?

1. Suppose we want a single number point estimate of θ *before* observing sample data. Find the mode of the prior distribution of θ, a.k.a., the "prior mode".
2. Find the median of the prior distribution of θ, a.k.a., the "prior median".
3. Find the expected value of the prior distribution of θ, a.k.a., the "prior mean".

   Now suppose that $y = 8$ couples in a sample of size $n = 12$ lean right. Recall the Bayes table.

theta	prior	likelihood	product	posterior
0.1	0.3333	0.0000	0.0000	0.0000
0.3	0.2667	0.0078	0.0021	0.0356
0.5	0.2000	0.1208	0.0242	0.4132
0.7	0.1333	0.2311	0.0308	0.5269
0.9	0.0667	0.0213	0.0014	0.0243
Total	1.0000	0.3811	0.0585	1.0000

4. Find the mode of the posterior distribution of θ, a.k.a., the "posterior mode".
5. Find the median of the posterior distribution of θ, a.k.a., the "posterior median".
6. Find the expected value of the posterior distribution of θ, a.k.a., the "posterior mean".
7. How have the posterior values changed from the respective prior values?

Solution. to Example 5.5

Show/hide solution

1. The prior mode is 0.1, the value of θ with the greatest prior probability.
2. The prior median is 0.3. Start with the smallest possible value of θ and add up the prior probabilities until they go from below 0.5 to above 0.5. This happens when you add in the prior probability for $\theta = 0.3$.
3. The prior mean is 0.367. Remember that an expected value is a probability-weighted average value

$$0.1(5/15) + 0.3(4/15) + 0.5(3/15) + 0.7(2/15) + 0.9(1/15) = 0.367.$$

4. The posterior mode is 0.7, the value of θ with the greatest posterior probability.
5. The posterior median is 0.7. Start with the smallest possible value of θ and add up the posterior probabilities until they go from below 0.5 to above 0.5. This happens when you add in the posterior probability for $\theta = 0.7$.
6. The posterior mean is 0.608. Now the posterior probabilities are used in the probability-weighted average value

$$0.1(0.000) + 0.3(0.036) + 0.5(0.413) + 0.7(0.527) + 0.9(0.024) = 0.608.$$

7. The point estimates (mode, median, mean) shift from their prior values (0.1, 0.3, 0.367) towards the observed sample proportion of 8/12. However, the posterior distribution is not symmetric, and the posterior mean is less than the posterior median. In particular, note that the posterior mean (0.608) lies between the prior mean (0.367) and the sample proportion (0.667).

A **point estimate** of an unknown parameter is a single-number estimate of the parameter. Given a posterior distribution of a parameter θ, three possible Bayesian point estimates of θ are:

- the posterior mean
- the posterior median
- the posterior mode.

In particular, the **posterior mean** is the expected value of θ according to the posterior distribution.

Recall that the expected value, a.k.a., mean, of a discrete random variable U is its probability-weighted average value

$$\mathrm{E}(U) = \sum_u u\, P(U = u)$$

In the calculation of a posterior mean, the parameter θ plays the role of the random variable U and the posterior distribution provides the probability-weights.

In many situations, the posterior distribution will be roughly symmetric with a single peak, in which case posterior mean, median, and mode will all be about the same.

Reducing the posterior distribution to a single-number point estimate loses a lot of the information the posterior distribution provides. The entire posterior distribution quantifies the uncertainty about θ after observing sample data. We will soon see how to more fully use the posterior distribution in making inference about θ.

Example 5.6. Continuing the kissing study in Example 5.3. Now assume a prior distribution which is proportional to $1 - \theta$ for $\theta = 0, 0.0001, 0.0002, \dots, 0.9999, 1$. Use software to answer the following.

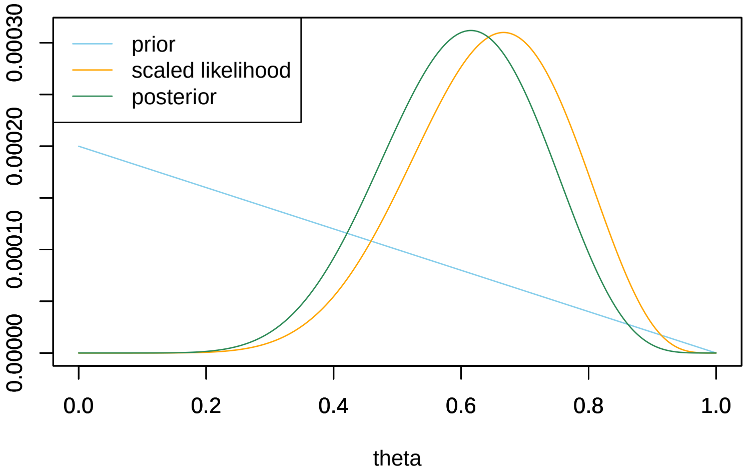
1. Find the mode of the prior distribution of θ, a.k.a., the "prior mode".

2. Find the median of the prior distribution of θ, a.k.a., the "prior median".

3. Find the expected value of the prior distribution of θ, a.k.a., the "prior mean".

Now suppose that $y = 8$ couples in a sample of size $n = 12$ lean right. Recall the prior, likelihood, and posterior.

```
theta = seq(0, 1, 0.0001)

# prior
prior = 1 - theta # shape of prior
prior = prior / sum(prior) # scales so that prior sums to 1

# data
n = 12 # sample size
y = 8 # sample count of success

# likelihood, using binomial
likelihood = dbinom(y, n, theta) # function of theta

# posterior
product = likelihood * prior
posterior = product / sum(product)
```



4. Find the mode of the posterior distribution of θ, a.k.a., the "posterior mode".

5. Find the median of the posterior distribution of θ, a.k.a., the "posterior median".

6. Find the expected value of the posterior distribution of θ, a.k.a., the "posterior mean".

7. How have the posterior values changed from the respective prior values?

Solution. to Example 5.6

1. See the code below. The prior mode is 0.

2. See the code below. The prior median is 0.293.

3. See the code below. The prior mean is 0.333.

```
## prior

# prior mode
theta[which.max(prior)]
```

```
## [1] 0
```

```
# prior median
min(theta[which(cumsum(prior) >= 0.5)])
```

```
## [1] 0.2929
```

```
# prior mean
sum(theta * prior)
```

```
## [1] 0.3333
```

4. See the code below. The posterior mode is 0.615.

5. See the code below. The posterior median is 0.605.

6. See the code below. The posterior mean is 0.6.

7. Each of the posterior point estimates has shifted from its prior value towards the sample proportion of 0.667. But note that each of the posterior point estimates is in between the prior point estimate and the sample proportion.

```
## posterior

# posterior mode
theta[which.max(posterior)]
```

```
## [1] 0.6154
```

```
# posterior median
min(theta[which(cumsum(posterior) >= 0.5)])
```

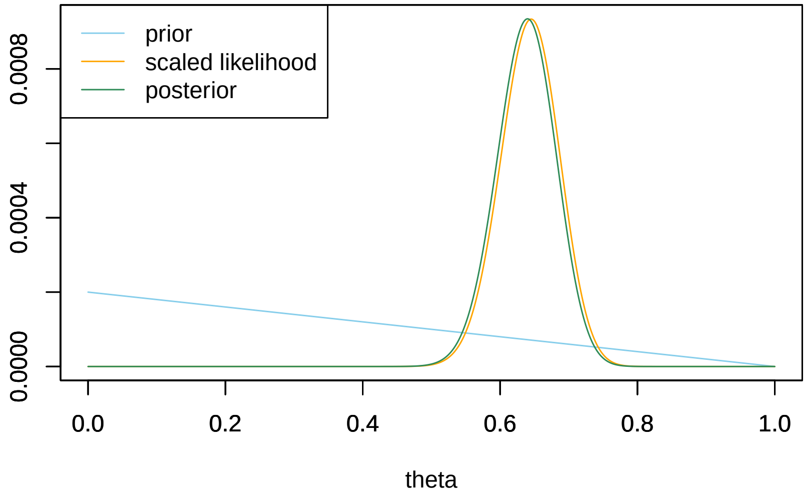
```
## [1] 0.6046
```

```
# posterior mean
sum(theta * posterior)
```

```
## [1] 0.6
```

Example 5.7. Continuing Example 5.6, now suppose that $y = 80$ couples in a sample of size $n = 124$ lean right (the actual study data). Recall the prior, likelihood, and posterior.

```
# data
n = 124 # sample size
y = 80 # sample count of success

# likelihood, using binomial
likelihood = dbinom(y, n, theta) # function of theta

# posterior
product = likelihood * prior
posterior = product / sum(product)
```



1. Find the mode of the posterior distribution of θ, a.k.a., the "posterior mode".
2. Find the median of the posterior distribution of θ, a.k.a., the "posterior median".
3. Find the expected value of the posterior distribution of θ, a.k.a., the "posterior mean".
4. How have the posterior values changed from the respective prior values? How does this compare to the smaller sample (8 out of 12)?

Solution. to Example 5.7

1. See the code below. The posterior mode is 0.64.

2. See the code below. The posterior median is 0.639.

3. See the code below. The posterior mean is 0.638.

4. The posterior distribution is roughly symmetric, and posterior mean, median, and mode are about the same. Each of the posterior point estimates has shifted from its prior value towards the sample proportion of 0.645. The posterior point estimates are now closer to the sample proportion than they were with the smaller sample size. With the larger sample size, the data carry more weight.

```r
## posterior

# posterior mode
theta[which.max(posterior)]
```

```
## [1] 0.64
```

```r
# posterior median
min(theta[which(cumsum(posterior) >= 0.5)])
```

```
## [1] 0.6385
```

```r
# posterior mean
sum(theta * posterior)
```

```
## [1] 0.6378
```

Chapter 6

Introduction to Inference

In a Bayesian analysis, the posterior distribution contains all relevant information about parameters after observing sample data. We can use the posterior distribution to make inferences about parameters.

Example 6.1. Suppose we want to estimate θ, the population proportion of American adults who have read a book in the last year.

1. Sketch your prior distribution for θ. Make a guess for your prior mode.
2. Suppose Henry formulates a Normal distribution prior for θ. Henry's prior mean is 0.4 and prior standard deviation is 0.1. What does Henry's prior say about θ?
3. Suppose Mudge formulates a Normal distribution prior for θ. Mudge's prior mean is 0.4 and prior standard deviation is 0.05. Who has more prior certainty about θ? Why?

Solution. to Example 6.1

Show/hide solution

1. Your prior distribution is whatever it is and represents your assessment of the degree of uncertainty of θ.
2. The posterior mean, median, and mode are all 0.4. A Normal distribution follows the empirical rule. In particular, the interval [0.3, 0.5] accounts for 68% of prior plausibility, [0.2, 0.6] for 95%, and [0.1, 0.7] for 99.7% of prior plausibility. Henry thinks θ is around 0.4, about twice as likely to lie inside the interval [0.3, 0.5] than to lie outside, and he would be pretty surprised if θ were outside of [0.1, 0.7].
3. Mudge has more prior certainty about θ due to the smaller prior standard deviation. The interval [0.3, 0.5] accounts for 95% of Mudge's plausibility, versus 68% for Henry.

The previous section discussed Bayesian point estimates of parameters, including the posterior mean, median, and mode. In some sense these values provide a single number "best guess" of the value of θ. However, reducing the posterior distribution to a single-number point estimate loses a lot of the information the posterior distribution provides. In particular, the posterior distribution quantifies the uncertainty about θ after observing sample data. The **posterior standard deviation** summarizes in a single number the degree of uncertainty about θ after observing sample data. The smaller the posterior standard deviation, the more certainty we have about the value of the parameter after observing sample data. (Similar considerations apply for the prior distribution. The **prior standard deviation** summarizes in a single number the degree of uncertainty about θ before observing sample data.)

Recall that the **variance** of a random variable U is its probability-weighted average squared distance from its expected value

$$\mathrm{Var}(U) = \mathrm{E}\left[(U - \mathrm{E}(U))^2\right]$$

The following is an equivalent "shortcut" formula for variance: "expected value of the square minus the square of the expected value."

$$\mathrm{Var}(U) = \mathrm{E}(U^2) - (\mathrm{E}(U))^2$$

The **standard deviation** of a random variable is the square root of its variance is $\mathrm{SD}(U) = \sqrt{\mathrm{Var}(U)}$. Standard deviation is measured in the same measurement units as the variable itself, while variance is measured in squared units.

In the calculation of a posterior standard deviation, θ plays the role of the variable U and the posterior distribution provides the probability-weights.

Example 6.2. Continuing Example 6.1, we'll assume a Normal prior distribution for θ with prior mean 0.4 and prior standard deviation 0.1.

1. Compute and interpret the prior probability that θ is greater than 0.7.
2. Find the 25th and 75th percentiles of the prior distribution. What is the prior probability that θ lies in the interval with these percentiles as endpoints? According to the prior, how plausible is it for θ to lie inside this interval relative to outside it? (Hint: use `qnorm`)
3. Repeat the previous part with the 10th and 90th percentiles of the prior distribution.
4. Repeat the previous part with the 1st and 99th percentiles of the prior distribution.

   In a sample of 150 American adults, 75% have read a book in the past year. (The 75% value is motivated by a real sample we'll see in a later example.)

5. Find the posterior distribution based on this data, and make a plot of prior, likelihood, and posterior.

6. Compute the posterior standard deviation. How does it compare to the prior standard deviation? Why?

7. Compute and interpret the posterior probability that θ is greater than 0.7. Compare to the prior probability.

8. Find the 25th and 75th percentiles of the posterior distribution. What is the posterior probability that θ lies in the interval with these percentiles as endpoints? According to the posterior, how plausible is it for θ to lie inside this interval relative to outside it? Compare to the prior interval.

9. Repeat the previous part with the 10th and 90th percentiles of the posterior distribution.

10. Repeat the previous part with the 1st and 99th percentiles of the posterior distribution.

Solution. to Example 6.2

Show/hide solution

1. We can use software (`1 - pnorm(0.7, 0.4, 0.1)`) but we can also use the empirical rule. For a Normal(0.4, 0.1) distribution, the value 0.7 is $\frac{0.7-0.4}{0.1} = 3$ SDs above the mean, so the probability is about 0.0015 (since about 99.7% of values are within 3 SDs of the mean). According to the prior, there is about a 0.1% chance that more than 70% of Americans adults have read a book in the last year.

2. We can use `qnorm(0.75)` $= 0.6745$ to find that the 75th percentile of a Normal distribution is about 0.67 SDs above the mean, so the 25th percentile is about 0.67 SDs below the mean. For the prior distribution, the 25th percentile is about 0.33 and the 75th percentile is about 0.47. The prior probability that θ lies in the interval [0.33, 0.47] is about 50%. According to the prior, it is equally plausible for θ to lie inside the interval [0.33, 0.47] as to lie outside this interval.

3. We can use `qnorm(0.9)` $= 1.28$ to find that the 90th percentile of a Normal distribution is about 1.28 SDs above the mean, so the 10th percentile is about 1.28 SDs below the mean. For the prior distribution, the 10th percentile is about 0.27 and the 90th percentile is about 0.53. The prior probability that θ lies in the interval [0.27, 0.53] is about 80%. According to the prior, it is four times more plausible for θ to lie inside the interval [0.27, 0.53] than to lie outside this interval.

4. We can use `qnorm(0.99)` $= 2.33$ to find that the 99th percentile of a Normal distribution is about 2.33 SDs above the mean, so the 1st percentile is about 2.33 SDs below the mean. For the prior distribution, the 1st percentile is about 0.167 and the 99th percentile is about 0.633. The prior probability that θ lies in the interval [0.167, 0.633] is about 98%. According to the prior, it is 49 times more plausible for θ to lie inside the interval [0.167, 0.633] than to lie outside this interval.

5. See below for a plot. Our prior gave very little plausibility to a sample like the one we actually observed. However, given our sample data, the likelihood corresponding to the values of θ we initially deemed most plausible is very low. Therefore, our posterior places most of the plausibility on values in the neighborhood of the observed sample proportion, even though the prior probability for many of these values was low. The prior does still have some influence; the posterior mean is 0.709 so we haven't shifted all the way towards the sample proportion yet.

6. Compute the posterior variance first using either the definition or the shortcut version, then take the square root; see code below. The posterior SD is 0.036, almost 3 times smaller than the prior SD. After observing data we have more certainty about the value of the parameter, resulting in a smaller posterior SD. The posterior distribution is approximately Normal with posterior mean 0.709 and posterior SD 0.036.

7. We can use the grid approximation; just sum the posterior probabilities for $\theta > 0.7$ to see that the posterior probability is about 0.603. Since the posterior distribution is approximately Normal, we can also use the empirical rule: the standardized value for 0.7 is $\frac{0.7-0.709}{0.036} = -0.24$, or 0.24 SDs below the mean. Using the empirical rule (or software `1 - pnorm(-0.24)`) gives about 0.596, similar to the grid calculation.

   We started with a very low prior probability that more than 70% of American adults have read at least one book in the last year. But after observing a sample in which more than 70% have read at least one book in the last year, we assign a much higher plausibility to more than 70% of *all American adults* having read at least one book in the last year. Seeing is believing.

8. See code below for calculations based on the grid approximation. But we can also use the fact the posterior distribution is approximately Normal; e.g., the 25th percentile is about 0.67 SDs below the mean: $0.709 - 0.67 \times 0.036 = 0.684$. For the posterior distibution, the 25th percentile is about 0.684 and the 75th percentile is about 0.733. The posterior probability that θ lies in the interval [0.684, 0.733] is about 50%. According to the posterior, it is equally plausible for θ to lie inside the interval [0.684, 0.733] as to lie outside this interval. This 50% interval is both (1) narrower than the prior interval, due to the smaller posterior SD, and (2) shifted towards

higher values of θ relative to the prior interval, due to the larger posterior mean.

9. For the posterior distibution, the 10th percentile is about 0.662 and the 90th percentile is about 0.754. The posterior probability that θ lies in the interval [0.662, 0.754] is about 80%. According to the posterior, it is four times more plausible for θ to lie inside the interval [0.662, 0.754] as to lie outside this interval. This 80% interval is both (1) narrower than the prior interval, due to the smaller posterior SD, and (2) shifted towards higher values of θ relative to the prior interval, due to the larger posterior mean.

10. For the posterior distibution, the 1st percentile is about 0.622 and the 99th percentile is about 0.789. The posterior probability that θ lies in the interval [0.622, 0.789] is about 98%. According to the posterior, it is 49 times more plausible for θ to lie inside the interval [0.622, 0.789] as to lie outside this interval. This interval is both (1) narrower than the prior interval, due to the smaller posterior SD, and (2) shifted towards higher values of θ relative to the prior interval, due to the larger posterior mean.

```r
theta = seq(0, 1, 0.0001)

# prior
prior = dnorm(theta, 0.4, 0.1) # shape of prior
prior = prior / sum(prior) # scales so that prior sums to 1

# data
n = 150 # sample size
y = round(0.75 * n, 0) # sample count of success

# likelihood, using binomial
likelihood = dbinom(y, n, theta) # function of theta

# posterior
product = likelihood * prior
posterior = product / sum(product)

# posterior mean
posterior_mean = sum(theta * posterior)
posterior_mean
```

```
## [1] 0.7024
```

```r
# posterior_variance - "shortcut" formula
posterior_var = sum(theta ^ 2 * posterior) - posterior_mean ^ 2
posterior_sd = sqrt(posterior_var)
posterior_sd
```

```
## [1] 0.03597
```

```
# posterior probability that theta is greater than 0.7
sum(posterior[theta > 0.7])
```

```
## [1] 0.5345
```

```
# posterior percentiles - central 50% interval
theta[max(which(cumsum(posterior) < 0.25))]
```

```
## [1] 0.6783
```

```
theta[max(which(cumsum(posterior) < 0.75))]
```

```
## [1] 0.727
```

```
# posterior percentiles - central 80% interval
theta[max(which(cumsum(posterior) < 0.1))]
```

```
## [1] 0.6557
```

```
theta[max(which(cumsum(posterior) < 0.9))]
```

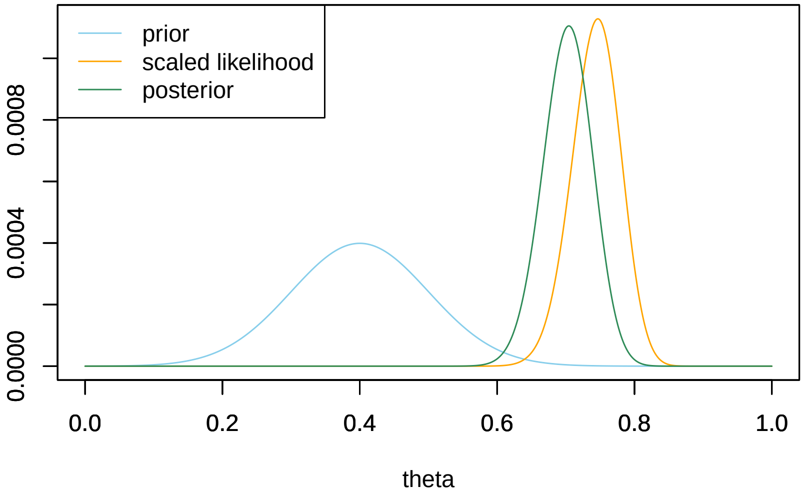
```
## [1] 0.7481
```

```
# posterior percentiles - central 98% interval
theta[max(which(cumsum(posterior) < 0.01))]
```

```
## [1] 0.6161
```

```
theta[max(which(cumsum(posterior) < 0.99))]
```

```
## [1] 0.7828
```



Bayesian inference for a parameter is based on its posterior distribution. Since a Bayesian analysis treats parameters as random variables, *it is possible to make probability statements about a parameter.*

A Bayesian **credible interval** is an interval of values for the parameter that has at least the specified probability, e.g., 50%, 80%, 98%. Credible intervals can be computed based on both the prior and the posterior distribution, though we are primarily interested in intervals based on the posterior distribution. For example,

- With a 50% credible interval, it is equally plausible that the parameter lies inside the interval as outside

- With an 80% credible interval, it is 4 times more plausible that the parameter lies inside the interval than outside
- With a 98% credible interval, it is 49 times more plausible that the parameter lies inside the interval than outside

Central credible intervals split the complementary probability evenly between the two tails. For example,

- The endpoints of a 50% central posterior credible interval are the 25th and the 75th percentiles of the posterior distribution.
- The endpoints of an 80% central posterior credible interval are the 10th and the 90th percentiles of the posterior distribution.

- The endpoints of a 98% central posterior credible interval are the 1st and the 99th percentiles of the posterior distribution.

There is nothing special about the values 50%, 80%, 98%. These are just a few convenient choices[1] whose endpoints correspond to "round number" percentiles (1st, 10th, 25th, 75th, 90th, 99th) and inside/outside ratios (1-to-1, 4-to-1, about 50-to-1). You could also throw in, say 70% (15th and 85th percentiles, about 2-to-1) or 90% (5th and 95th percentiles, about 10-to-1), if you wanted. As the previous example illustrates, it's not necessary to just select a single credible interval (e.g., 95%). Bayesian inference is based on the full posterior distribution. Credible intervals simply provide a summary of this distribution. Reporting a few credible intervals, rather than just one, provides a richer picture of how the posterior distribution represents the uncertainty in the parameter.

In many situations, the posterior distribution of a single parameter is approximately Normal, so posterior probabilities can be approximated with Normal distribution calculations — standardizing and using the empirical rule. In particular, an approximate central credible interval has endpoints

$$\text{posterior mean} \pm z^* \times \text{posterior SD}$$

where z^* is the appropriate multiple for a standard Normal distribution corresponding to the specified probability. For example,

Central credibility	50%	80%	95%	98%
Normal z^* multiple	0.67	1.28	1.96	2.33

Central credible intervals are easier to compute, but are not the only or most widely used credible intervals. A **highest posterior density interval** is the interval of values that has the specified posterior probability and is such that the posterior density within the interval is never lower than the posterior density outside the interval. If the posterior distribution is relatively symmetric with a single peak, central posterior credible intervals and highest posterior density intervals are similar.

Example 6.3. Continuing Example 6.1, we'll assume a Normal prior distribution for θ with prior mean 0.4 and prior standard deviation 0.1.

In a recent survey of 1502 American adults conducted by the Pew Research Center, 75% of those surveyed said thay have read a book in the past year.

[1] In Section 3.2.2 of *Statistical Rethinking* (McElreath (2020)), the author suggests 67%, 89%, and 97%: "a series of nested intervals may be more useful than any one interval. For example, why not present 67%, 89%, and 97% intervals, along with the median? Why these values? No reason. They are prime numbers, which makes them easy to remember. But all that matters is they be spaced enough to illustrate the shape of the posterior. And these values avoid 95%, since conventional 95% intervals encourage many readers to conduct unconscious hypothesis tests."

1. Find the posterior distribution based on this data, and make a plot of prior, likelihood, and posterior. Describe the posterior distribution. How does this posterior compare to the one based on the smaller sample size ($n = 150$)?
2. Compute and interpret the posterior probability that θ is greater than 0.7. Compare to the prior probability.
3. Compute and interpret in context 50%, 80%, and 98% central posterior credible intervals.
4. Here is how the survey question was worded: "During the past 12 months, about how many BOOKS did you read either all or part of the way through? Please include any print, electronic, or audiobooks you may have read or listened to." Does this change your conclusions? Explain.

Solution. to Example 6.3

Show/hide solution

1. See below for code and plots. The posterior distribution is approximately Normal with posterior mean 0.745 and posterior SD 0.011. Despite our prior beliefs that θ was in the 0.4 range, enough data has convinced us otherwise. With a large sample size, the prior has little influence on the posterior; much less than with the smaller sample size. Compared to the posterior based on the small sample size, the posterior now (1) has shifted to the neighborhood of the sample data, (2) exhibits a smaller degree of uncertainty about the parameter.

2. The posterior probability that θ is greater than 0.7 is about 0.9999. We started with only a 0.1% chance that more than 70% of American adults have read a book in the last year, but the large sample has convinced us otherwise.

3. There is a posterior probability of:

   - 50% that the population proportion of American adults who have read a book in the past year is between 0.737 and 0.753. We believe that the population proportion is as likely to be inside this interval as outside.
   - 80% that the population proportion of American adults who have read a book in the past year is between 0.730 and 0.759. We believe that the population proportion is four times more likely to be inside this interval than to be outside it.
   - 98% that the population proportion of American adults who have read a book in the past year is between 0.718 and 0.771. We believe that the population proportion is 49 times more likely to be inside this interval than to be outside it.

   In short, our conclusion is that somewhere-in-the-70s percent of American adults have read a book in the past year. But see the next part…

4. It depends on what our goal is. Do we want to only count completed books? Does there have to be a certain word count? Does it count if the adult read a children's book? Does listening to an audiobook count? Does it have to be for "fun" or does reading books for work count? If our goal is to estimate the population proportion of Americans who *have read completely a 100,000 word non-audiobook book in the last year*, then this particular sample data would be fairly biased from our perspective.

```
theta = seq(0, 1, 0.0001)

# prior
prior = dnorm(theta, 0.4, 0.1) # shape of prior
prior = prior / sum(prior) # scales so that prior sums to 1

# data
n = 1502 # sample size
y = round(0.75 * n, 0) # sample count of success

# likelihood, using binomial
likelihood = dbinom(y, n, theta) # function of theta

# posterior
product = likelihood * prior
posterior = product / sum(product)

# posterior mean
posterior_mean = sum(theta * posterior)
posterior_mean
```

```
## [1] 0.745
```

```
# posterior_variance - "shortcut" formula
posterior_var = sum(theta ^ 2 * posterior) - posterior_mean ^ 2
posterior_sd = sqrt(posterior_var)
posterior_sd
```

```
## [1] 0.01123
```

```
# posterior probability that theta is greater than 0.7
sum(posterior[theta > 0.7])
```

```
## [1] 0.9999
```

```
# posterior percentiles - central 50% interval
theta[max(which(cumsum(posterior) < 0.25))]
```

```
## [1] 0.7374
```

```
theta[max(which(cumsum(posterior) < 0.75))]
```

```
## [1] 0.7525
```

```
# posterior percentiles - central 80% interval
theta[max(which(cumsum(posterior) < 0.1))]
```

```
## [1] 0.7304
```

```
theta[max(which(cumsum(posterior) < 0.9))]
```

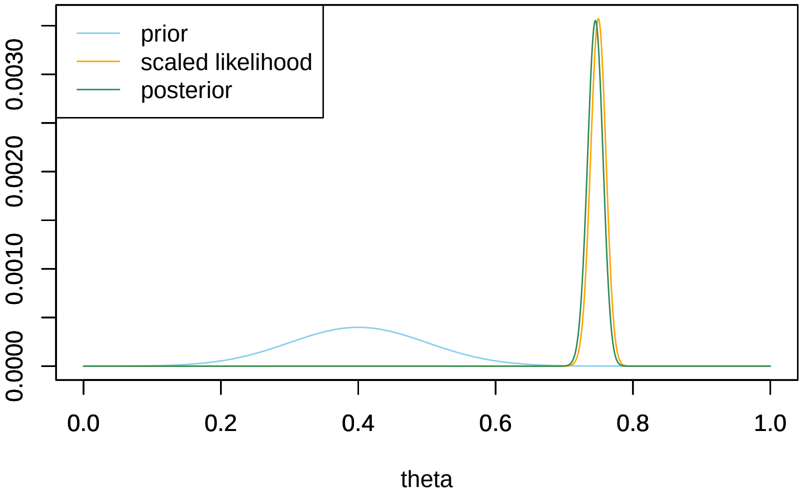
```
## [1] 0.7592
```

```
# posterior percentiles - central 98% interval
theta[max(which(cumsum(posterior) < 0.01))]
```

```
## [1] 0.7183
```

```
theta[max(which(cumsum(posterior) < 0.99))]
```

```
## [1] 0.7705
```



The quality of any statistical analysis depends very heavily on the quality of the data. Always investigate how the data were collected to determine what conclusions are appropriate. Is the sample reasonably representative of the population? Were the variables reliably measured?

Example 6.4. Continuing Example 6.3, we'll use the same sample data ($n = 1502$, 75%) but now we'll consider different priors.

For each of the priors below, plot prior, likelihood, and posterior, and compute the posterior probability that θ is greater than 0.7. Compare to Example 6.3.

1. Normal distribution prior with prior mean 0.4 and prior SD 0.05.
2. Uniform distribution prior on the interval [0, 0.7]

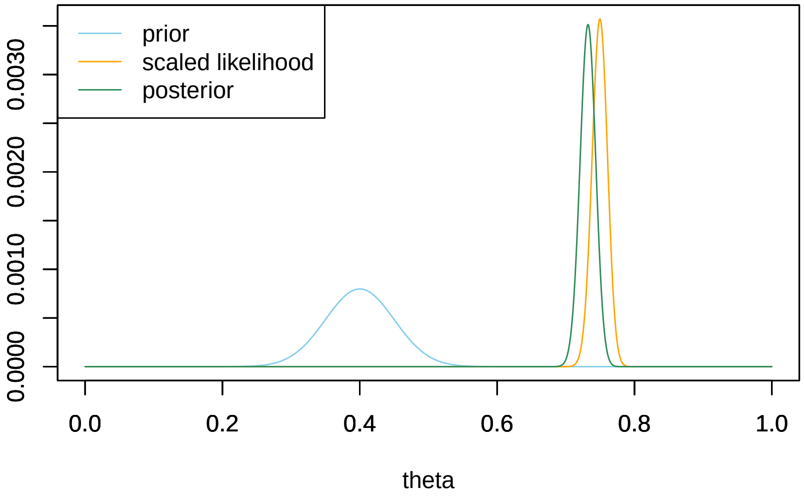
Solution. to Example 6.4

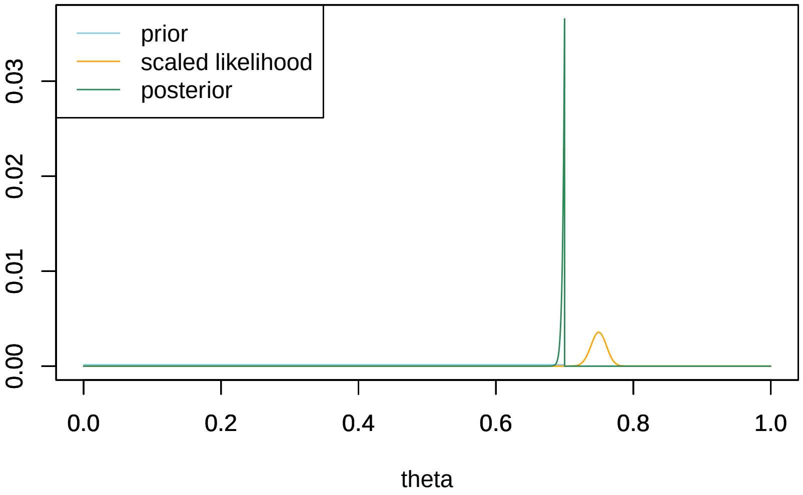
Show/hide solution

1. The Normal(0.4, 0.05) prior concentrates almost all prior plausibility in a fairly narrow range of values (0.25 to 0.55 or so) and represents more prior certainty about θ than the Normal(0.4, 0.1) prior. Even with the large sample size, we see that the Normal(0.4, 0.05) prior has more influence on the posterior than the Normal(0.4, 0.1). However, the two posterior distributions are not that different: Normal(0.73, 0.011) here compared with Normal(0.745, 0.011) from the previous problem. Both posteriors assign almost all posterior credibility to values in the low to mid 70s

percent. In particular, the posterior probability that θ is greater than 0.7 is 0.997 (compared with 0.9999 from the previous problem).

2. The Uniform prior distribution spreads prior plausibility over a fairly wide range of values, [0, 0.7]. However, the prior probability that θ is greater than 0.7 is 0. Even when we observed a large sample with a sample proportion greater than 0.7, the posterior probability that θ is greater than 0.7 remains 0. See the plot below; the posterior distribution is basically a spike that puts almost all of the posterior credibility on the value 0.7. Assigning 0 prior probability for θ values greater than 0.7 has essentially identified such θ values as impossible, and no amount of data can make the impossible possible.





You have a great deal of flexibility in choosing a prior, and there are many reasonable approaches. However, do NOT choose a prior that assigns 0 probability/density to possible values of the parameter regardless of how initially implausible the values are. Even very stubborn priors can be overturned with enough data, but no amount of data can turn a prior probability of 0 into a positive posterior probability. Always consider the range of *possible* values of the parameter, and be sure the prior density is non-zero over the entire range of possible values.

6.1 Comparing Bayesian and frequentist interval estimates

The most widely used elements of "traditional" frequentist inference are confidence intervals and hypothesis tests (a.k.a, null hypothesis significance tests). The numerical results of Bayesian and frequentist analysis are often similar. However, the interpretations are very different.

Example 6.5. We'll now compare the Bayesian credible intervals in Example 6.4 to frequentist confidence intervals. Recall the actual study data in which 75% of the 1502 American adults surveyed said they read a book in the last year.

1. Compute a 98% confidence interval for θ.

2. Write a clearly worded sentence reporting the confidence interval in context.
3. Explain what “98% confidence” means.
4. Compare the *numerical results* of the Bayesian and frequentist analysis. Are they similar or different?
5. How does the *interpretation* of these results differ between the two approaches?

Solution. to Example 6.5

Show/hide solution

1. The observed sample proportion is $\hat{p} = 0.75$ and its standard error is $\sqrt{\hat{p}(1-\hat{p})/n} = \sqrt{0.75(1-0.75)/1502} = 0.011$. The usual formula for a confidence interval for a population prportion is

   $$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

   where z^* is the multiple from a standard Normal distribution corresponding to the level of confidence (e.g., $z^* = 2.33$ for 98% confidence). A 98% confidence interval for θ is [0.724, 0.776].

2. We estimate with 98% confidence that the population proportion of American adults who have read a book in the last year is between 0.724 and 0.776.

3. Confidence is in the estimation procedure. Over many samples, 98% of samples will yield confidence intervals, computed using the above formula, that contain the true parameter value (a fixed number) The intervals change from sample to sample; the parameter is fixed.

4. The numerical results are similar: the 98% posterior credible interval is similar to the 98% confidence interval. Both reflect a conclusion that we think that somewhere-in-the-70s percent of American adults have read at least one book in the past year.

5. However, the *interpretation* of these results is very different between the two approaches.

   The Bayesian approach provides *probability statements about the parameter*: There is a 98% chance that θ is between 0.718 and 0.771; our assessment is that θ is 49 times more likely to lie inside the interval [0.718, 0.771] than outside.

   In the frequestist approach such a probability statement makes no sense. From the frequentist perspective, θ is an unknown *number*: either that number is in the interval [0.724, 0.776] or it’s not; there’s no probability to it. Rather, the frequentist approach develops procedures based on the probability of what might happen over many samples. Notice that in the interpretation of what 98% confidence means above, the actual numbers [0.724, 0.776] did not appear. The confidence is in the procedure that produced the interval, and not in the interval itself.

Example 6.6. Have more than 70% of Americans read a book in the last year? We'll now compare the Bayesian analysis in Example 6.4 to a frequentist (null) hypothesis (significance) test. Recall the actual study data in which 75% of the 1502 American adults surveyed said they read a book in the last year.

1. Conduct an appropriate hypothesis test.
2. Write a clearly worded sentence reporting the conclusion of the hypothesis test in context.
3. Write a clearly worded sentence interpreting the p-value in context.
4. Now back to the Bayesian analysis of Example 6.4. Compute the posterior probability that θ is less than or equal to 0.70.
5. Compare the *numerical values* of the posterior probability and the p-value. Are they similar or different?
6. How does the *interpretation* of these results differ between the two approaches?

Solution. to Example 6.6

Show/hide solution

1. The null hypothesis is $H_0 : \theta = 0.7$. The alternative hypothesis is $H_a : \theta > 0.7$. The standard deviation of the null distribution is $\sqrt{0.7(1-0.7)/1502} = 0.0118$. The standardized (test) statistic is $(0.75 - 0.7)/0.0118 = 4.23$. With such a large sample size, the null distribution of sample proportions is approximately Normal, so the p-value is approximately `1 - pnorm(4.23)` = 0.000012.

2. With a p-value of 0.000012 we have extremely strong evidence to reject the null hypothesis and conclude that more than 70% of Americans have read a book in the last year.

3. Interpreting the p-value

   - If the population proportion of Americans who have read a book in the last year is equal to 0.7,
   - Then we would observe a sample proportion of 0.75 or more in about 0.0012% (about 1 in 100,000) of random samples of size 1502.
   - Since we actually observed a sample proportion of 0.75, which would be extremely unlikely if the population proportion were 0.7,
   - The data provide evidence that the population proportion is not 0.7.

4. See Example 6.4 where we computed the posterior probability that θ is greater than 0.7. The posterior probability that θ is less than or equal to 0.7 is 0.000051.

   Note: in the frequentist hypothesis test, the null hypothesis $H_0 : \theta = 0.7$ is operationally the same as $H_0 : \theta \leq 0.7$; the test is conducted the same way and results in the same p-value. Computing the posterior probability that $\theta \leq 0.7$ is like computing the probability that the null hypothesis is true. Now, the p-value is *not* the probability that the null hypothesis is true, even though that is a common misinterpretation. But there is no direct Bayesian analog of a p-value, so this will have to do.

5. The numerical results are similar; both the p-value and the posterior probability are on the order of 1/100000. Both reflect a strong endorsement of the conclusion that more than 70% of Americans have read a book in the past year.

6. However, the *interpretation* of these results is very different between the two approaches.

   The Bayesian analysis computes a probability that $\theta < 0.7$: there's an extremely small probability that θ is less than 0.7, so we'd be willing to bet a very large amount of money that it's not.

   But such a probability make no sense from a frequentist perspective. From the frequentist perspective, the unknown parameter θ is a *number*: either than number is greater than 0.7 or it's not; there's no probability to it. The p-value is a probability referring to what would happen over many samples.

Since a Bayesian analysis treats parameters as random variables, it is possible to make probability statements about parameters. In contrast, a frequentist analysis treats unknown parameters as fixed — that is, not random — so probability statements do not apply to parameters. In a frequentist approach, probability statements (like "95% confidence") are based on how the sample data would behave over many hypothetical samples.

In a Bayesian approach

- Parameters are random variables and have distributions.
- Observed data are treated as fixed, not random.
- All inference is based on the posterior distribution of parameters which quantifies our uncertainty about the parameters.
- The posterior distribution quantifies our uncertainty in the parameters, after observing the sample data.
- The posterior (or prior) distribution can be used to make probability statements about parameters.

- For example, "95% credible" quantifies our assessment that the parameter is 19 times more likely to lie inside the credible interval than outside. (Roughly, we'd be willing to bet at 19-to-1 odds on whether θ is inside the interval [0.718, 0.771].)

In a frequentist approach

- Parameters are treated as fixed (not random), but unknown numbers
- Data are treated as random
- All inference is based on the sampling distribution of the data which quantifies how the data behaves over many hypothetical samples.
- For example, "95% confidence" is confidence in the procedure: confidence intervals vary from sample-to-sample; over many samples 95% of confidence intervals contain the parameter being estimated.

Chapter 7

Introduction to Prediction

A Bayesian analysis leads directly and naturally to making predictions about future observations from the random process that generated the data. Prediction is also useful for checking if model assumptions seem reasonable in light of observed data.

Example 7.1. Do people prefer to use the word "data" as singular or plural? Data journalists at FiveThirtyEight conducted a poll to address this question (and others). Rather than simply ask whether the respondent considered "data" to be singular or plural, they asked which of the following sentences they prefer:

a. Some experts say it's important to drink milk, but the data is inconclusive.
b. Some experts say it's important to drink milk, but the data are inconclusive.

Suppose we wish to study the opinions of students in Cal Poly statistics classes regarding this issue. That is, let θ represent the population proportion of students in Cal Poly statistics classes who prefer to consider data as a *singular* noun, as in option a) above.

To illustrate ideas, we'll start with a prior distribution which places probability 0.01, 0.05, 0.15, 0.30, 0.49 on the values 0.1, 0.3, 0.5, 0.7, 0.9, respectively.

1. Before observing any data, suppose we plan to randomly select a single Cal Poly statistics student. Consider the *unconditional* prior probability that the selected student prefers data as singular. (This is called a *prior predictive probability.*) Explain how you could use simulation to approximate this probability.

2. Compute the prior predictive probability from the previous part.

3. Before observing any data, suppose we plan to randomly select a sample of 35 Cal Poly statistics students. Consider the *unconditional* prior distribution of the number of students in the sample who prefer data as singular. (This is called a *prior predictive distribution.*) Explain how you could use simulation to approximate this distribution. In particular, how could you use simulation to approximate the prior predictive probability that at least 34 students in the sample prefer data as singular?

4. Compute and interpret the prior predictive probability that at least 34 students in a sample of size 35 prefer data as singular.

   For the remaining parts, suppose that 31 students in a sample of 35 Cal Poly statistics students prefer data as singular.

5. Find the posterior distribution of θ.

6. Now suppose we plan to randomly select an additional Cal Poly statistics student. Consider the *posterior predictive probability* that this student prefers data as singular. Explain how you could use simulation to estimate this probability.

7. Compute the posterior predictive probability from the previous part.

8. Suppose we plan to collect data on another sample of 35 Cal Poly statistics students. Consider the *posterior predictive distribution* of the number of students in the new sample who prefer data as singular. Explain how you could use simulation to approximate this distribution. In particular, how could you use simulation to approximate the prior predictive probability that at least 34 students in the sample prefer data as singular? (Of course, the sample size of the new sample does not have to be 35. However, we're keeping it the same so we can compare the prior and posterior predictions.)

9. Compute and interpret the posterior predictive probability that at least 34 students in a sample of size 35 prefer data as singular.

Solution. to Example 7.1

1. If we knew what θ was, this probability would just be θ. For example, if $\theta = 0.9$, then there is a probability of 0.9 that a randomly selected student prefers data singular. If θ were 0.9, we could approximate the probability by constructing a spinner with 90% of the area marked as "success", spinning it many times, and recording the proportion of spins that land on success, which should be roughly 90%. However, 0.9 is only one possible value of θ. Since we don't know what θ is, we need to first simulate a value of it, giving more weight to θ values with high prior probability. Therefore, we

   1. Simulate a value of θ from the prior distribution.

   2. Given the value of θ, construct a spinner that lands on success with probability θ. Spin the spinner once and record the result, success or not.
   3. Repeat steps 1 and 2 many times, and find the proportion of repetitions which result in success. This proportion approximates the unconditional prior probability of success.

2. Use the law of total probability, where the weights are given by the prior probabilities.

   $$0.1(0.01) + 0.3(0.05) + 0.5(0.15) + 0.7(0.30) + 0.9(0.49) = 0.742$$

   (This calculation is equivalent to the expected value of θ according to its prior distributon, that is, the prior mean.)

3. If we knew what θ was, we could construct a spinner than lands on success with probability θ, spin it 35 times, and count the number of successes. Since we don't know what θ is, we need to first simulate a value of it, giving more weight to θ values with high prior probability. Therefore, we

   1. Simulate a value of θ from the prior distribution.
   2. Given the value of θ, construct a spinner that lands on success with probability θ. Spin the spinner 35 times and count y, the number of spins that land on success.
   3. Repeat steps 1 and 2 many times, and record the number of successes (out of 35) for each repetition. Summarize the simulated y values to approximate the prior predictive distribution. To approximate the prior predictive probability that at least 34 students in a sample of size 35 prefer data as singular, count the number of simulated repetitions that result in at least 34 successes ($y \geq 34$) and divide by the total number of simulated repetitions.

4. If we knew θ, the probability of at least 34 (out of 35) successes is, from a Binomial distribution,

   $$35\theta^{34}(1-\theta) + \theta^{35}$$

   Use the law of total probability again.

   $$\begin{aligned}&\left(35(0.1)^{34}(1-0.1) + 0.1^{35}\right)(0.01) + \left(35(0.3)^{34}(1-0.3) + 0.3^{35}\right)(0.05)\\&+ \left(35(0.5)^{34}(1-0.5) + 0.5^{35}\right)(0.15) + \left(35(0.7)^{34}(1-0.7) + 0.7^{35}\right)(0.30)\\&+ \left(35(0.9)^{34}(1-0.9) + 0.9^{35}\right)(0.49) = 0.06\end{aligned}$$

   According to this model, about 6% of samples of size 35 would result in at least 34 successes. The value 0.06 accounts for both (1) our prior uncertainty about θ, (2) sample-to-sample variability in the number of successes y.

5. The likelihood is $\binom{35}{31}\theta^{31}(1-\theta)^4$, a function of θ; `dbinom(31, 35, theta)`. The posterior places almost all probability on $\theta = 0.9$, due to both its high prior probability and high likelihood.

```
theta = seq(0.1, 0.9, 0.2)

# prior
prior = c(0.01, 0.05, 0.15, 0.30, 0.49)

# data
n = 35 # sample size
y = 31 # sample count of success

# likelihood, using binomial
likelihood = dbinom(y, n, theta) # function of theta

# posterior
product = likelihood * prior
posterior = product / sum(product)

# bayes table
bayes_table = data.frame(theta,
                         prior,
                         likelihood,
                         product,
                         posterior)

bayes_table %>%
  adorn_totals("row") %>%
  kable(digits = 4, align = 'r')
```

theta	prior	likelihood	product	posterior
0.1	0.01	0.0000	0.0000	0.0000
0.3	0.05	0.0000	0.0000	0.0000
0.5	0.15	0.0000	0.0000	0.0000
0.7	0.30	0.0067	0.0020	0.0201
0.9	0.49	0.1998	0.0979	0.9799
Total	1.00	0.2065	0.0999	1.0000

6. The simulation would be similar to the prior simulation, but now we simulate θ from its posterior distribution rather than the prior distribution.

    1. Simulate a value of θ from the posterior distribution.
    2. Given the value of θ, construct a spinner that lands on success with probability θ. Spin the spinner once and record the result, success or not.

3. Repeat steps 1 and 2 many times, and find the proportion of repetitions which result in success. This proportion approximates the unconditional posterior probability of success.

7. Use the law of total probability, where the weights are given by the posterior probabilities.

   $$0.1(0.0000)+0.3(0.0000)+0.5(0.0000)+0.7(0.0201)+0.9(0.9799) = 0.8960$$

   (This calculation is equivalent to the expected value of θ according to its posterior distributon, that is, the posterior mean.)

8. The simulation would be similar to the prior simulation, but now we simulate θ from its posterior distribution rather than the prior distribution.

   1. Simulate a value of θ from the posterior distribution.
   2. Given the value of θ, construct a spinner that lands on success with probability θ. Spin the spinner 35 times and count the number of spins that land on success.
   3. Repeat steps 1 and 2 many times, and record the number of successes (out of 35) for each repetition. Summarize the simulated values to approximate the posterior predictive distribution. To approximate the posterior predictive probability that at least 34 students in a sample of size 35 prefer data as singular, count the number of simulated repetitions that result in at least 34 successes and divide by the total number of simulated repetitions.

   Since the posterior probability that θ equals 0.9 is close to 1, the posterior predictive distribution would be close to, but not quite, the Binomial(35, 0.9) distribution.

9. Use the law of total probability again, but with the posterior probabilities rather than the prior probabilities as the weights.

   $$\begin{aligned}
   &\left(35(0.1)^{34}(1-0.1)+0.1^{35}\right)(0.0000)+\left(35(0.3)^{34}(1-0.3)+0.3^{35}\right)(0.0000)\\
   &+\left(35(0.5)^{34}(1-0.5)+0.5^{35}\right)(0.0000)+\left(35(0.7)^{34}(1-0.7)+0.7^{35}\right)(0.0201)\\
   &+\left(35(0.9)^{34}(1-0.9)+0.9^{35}\right)(0.9799) = 0.1199
   \end{aligned}$$

   According to this posterior model, about 12% of samples of size 35 would result in at least 34 successes. The value 0.12 accounts for both (1) our posterior uncertainty about θ, after observing the sample data, (2) sample-to-sample variability in the number of successes y for the yet-to-be-observed sample.

The plots below illustrate the distributions from the previous example.

The first plot below illustrates the conditional distribution of Y given each value of θ.

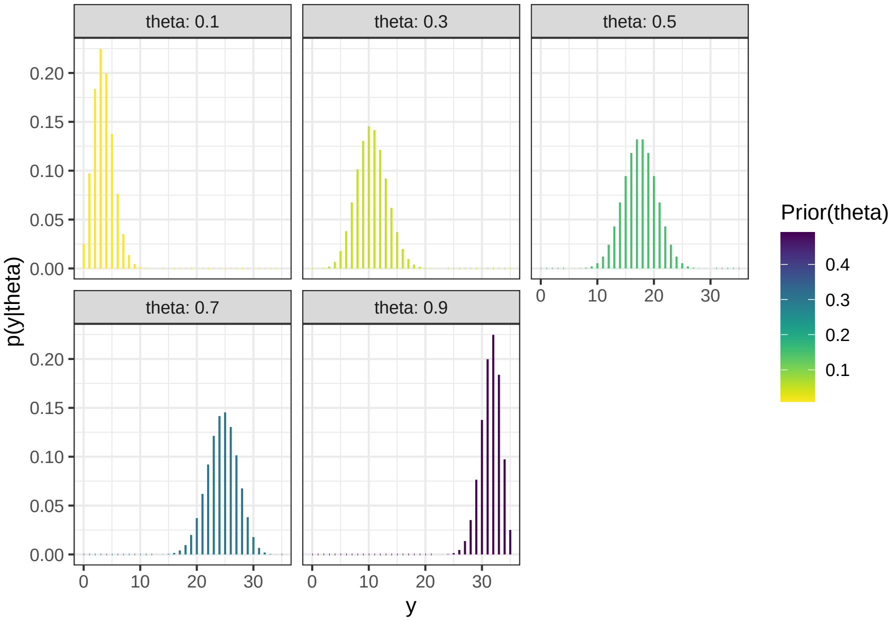


Figure 7.1: Sample-to-sample distribution of Y, the number of successes in samples of size $n = 35$ for different values of θ in Example 7.1. Color represents the prior probability of θ, with darker colors corresponding to higher prior probability.

In the plot above, the prior distribution of θ is represented by color. The prior predictive distribution of Y mixes the five conditional distributions in the previous plot, weighting by the prior distribution of θ, to obtain the unconditional prior predictive distribution of Y.

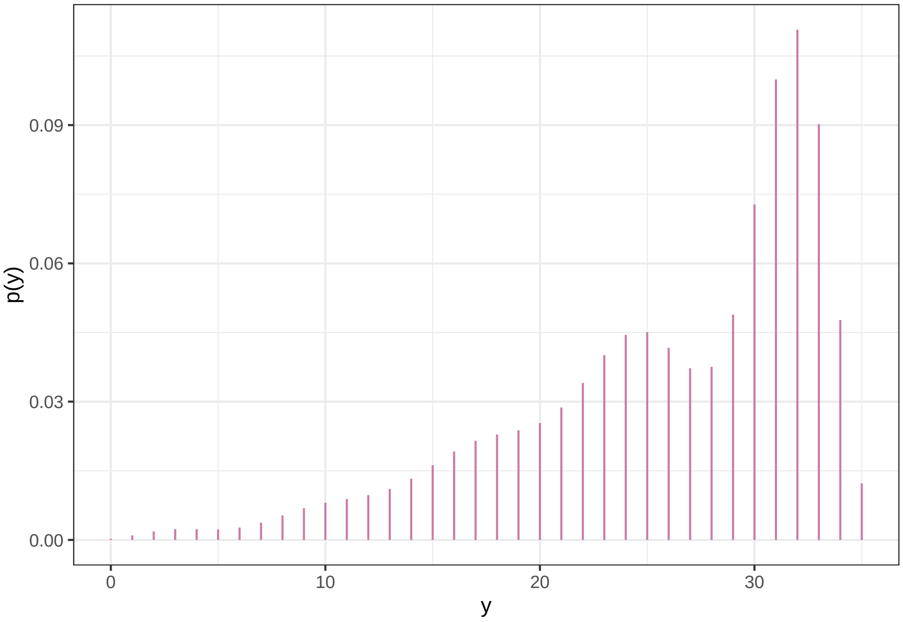


Figure 7.2: Prior predictive distribution of Y, the number of successes in samples of size $n = 35$, in Example 7.1.

The prior predictive distribution reflects the sample-to-sample variability of the number of successes over many samples of size $n = 35$, accounting for the prior uncertainty of θ.

After observing a sample, we compute the posterior distribution of θ as usual. The following plot is similar to Figure 7.1, with the colors revised to reflect the posterior probabilities of the values of θ.

The posterior predictive distribution of Y mixes the five conditional distributions in the previous plot, weighting by the posterior distribution of θ, to obtain the unconditional posterior predictive distribution of Y. Since the posterior distribution of θ places almost all posterior probability on the value 0.9, the posterior predictive distribution of Y is very similar to the conditional distribution of Y given $\theta = 0.9$.

The **predictive distribution** of a random variable is the marginal distribution (of the unobserved values) after accounting for the uncertainty in the parameters. A **prior predictive distribution** is calculated using the prior distribution

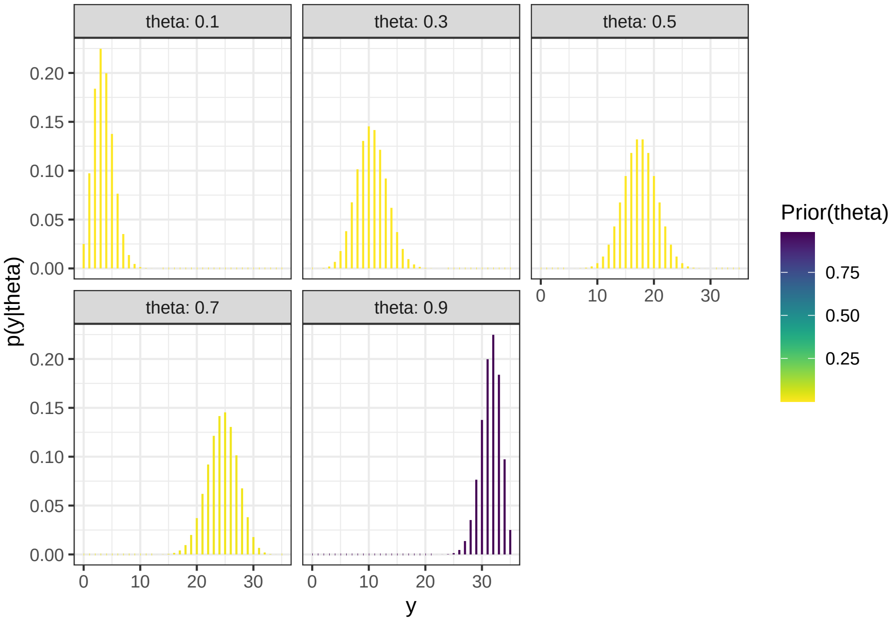


Figure 7.3: Sample-to-sample distribution of Y, the number of successes in samples of size $n = 35$ for different values of θ in Example 7.1. Color represents the posterior probability of θ, after observing data from a different sample of size 35, with darker colors corresponding to higher posterior probability.

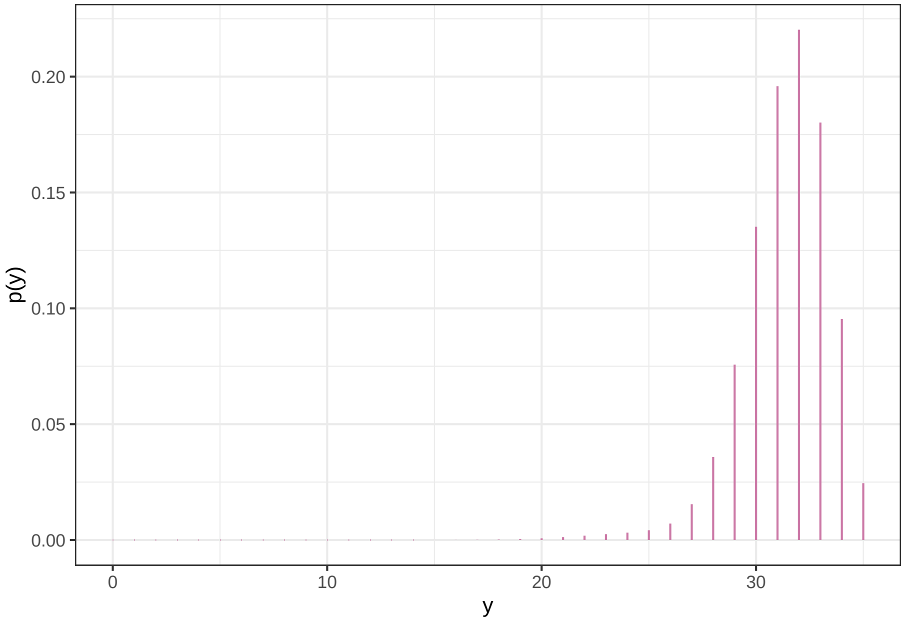


Figure 7.4: Posterior predictive distribution of Y, the number of successes in samples of size $n = 35$, in Example 7.1 after observing data from a different sample of size 35.

of the parameters. A **posterior predictive distribution** is calculated using the posterior distribution of the parameters, conditional on the observed data.

Be sure to carefully distinguish between posterior distributions and posterior predictive distributions (or between prior distributions and prior predictive distributions.)

Prior and posterior distributions are distributions on values of the *parameters*. These distributions quantify the degree of uncertainty about the unknown parameter θ (before and after observing data).

On the other hand, prior and posterior *predictive* distributions are distribution on potential values of the *data*. Predictive distributions reflect sample-to-sample variability of the sample data, while accounting for the uncertainty in the parameters.

Predictive probabilities can be computed via the law of total probability, as weighted averages over possible values of θ. However, even when conditional distributions of data given the parameters are well known (e.g., Binomial(n, θ)), marginal distributions of the data are often not. Simulation is an effective tool in approximating predictive distributions.

- Step 1: Simulate a value of θ from its posterior distribution (or prior distribution).
- Step 2: Given this value of θ simulate a value of y from $f(y|\theta)$, the data model conditional on θ.
- Repeat many times to simulate many (θ, y) pairs, and summarize the values of y to approximate the posterior predictive distribution (or prior predictive distribution).

Example 7.2. Continuing the previous example. We'll use a grid approximation and assume that any multiple of 0.0001 is a possible value of θ: $0, 0.0001, 0.0002, \ldots, 0.9999, 1$.

1. Consider the context of this problem and sketch your prior distribution for θ. What are the main features of your prior?

2. Assume the prior distribution for θ is proportional to θ^2. Plot this prior distribution and describe its main features.

3. Given the shape of the prior distribution, explain why we might not want to compute *central* prior credible intervals. Suggest an alternative approach, and compute and interpret 50%, 80%, and 98% prior credible intervals for θ.

4. Before observing any data, suppose we plan to randomly select a sample of 35 Cal Poly statistics students. Let Y represent the number of students in the selected sample who prefer data as singular. Use simulation to approximate the prior predictive distribution of Y and plot it.

5. Use software to compute the prior predictive distribution of Y. Compare to the simulation results.

6. Find a 95% prior *prediction* interval for Y. Write a clearly worded sentence interpreting this interval in context.

   For the remaining parts, suppose that 31 students in a sample of 35 Cal Poly statistics students prefer data as singular.

7. Use software to plot the prior distribution and the (scaled) likelihood, then find the posterior distribution of θ and plot it and describe its main features.

8. Find and interpret 50%, 80%, and 98% central posterior credible intervals for θ.

9. Suppose we plan to randomly select another sample of 35 Cal Poly statistics students. Let $\tilde{Y}$ represent the number of students in the selected sample who prefer data as singular. Use simulation to approximate the posterior predictive distribution of $\tilde{Y}$ and plot it. (Of course, the sample size of the new sample does not have to be 35. However, we're keeping it the same so we can compare the prior and posterior predictions.)

10. Use software to compute the posterior predictive distribution of $\tilde{Y}$. Compare to the simulation results.

11. Find a 95% posterior *prediction* interval for $\tilde{Y}$. Write a clearly worded sentence interpreting this interval in context.

12. Now suppose instead of using the Cal Poly sample data (31/35) to form the posterior distribution of θ, we had used the data from the FiveThirtyEight study in which 865 out of 1093 respondents preferred data as singular. Use software to plot the prior distribution and the (scaled) likelihood, then find the posterior distribution of θ and plot it and describe its main features. In particular, find and interpret a 98% central posterior credible interval for θ. How does the posterior based on the FiveThirtyEight data compare to the posterior distribution based on the Cal Poly sample data (31/35)? Why?

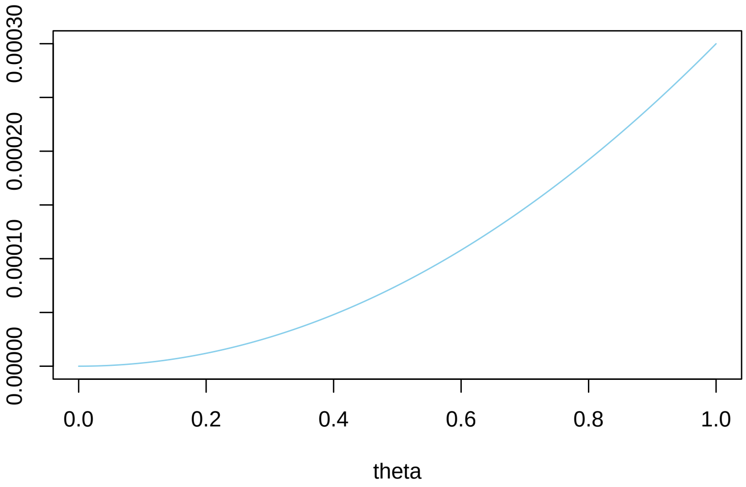
13. Again, suppose we use the FiveThirtyEight data to form the posterior distribution of θ. Suppose we plan to randomly select a sample of 35 Cal Poly statistics students. Let $\tilde{Y}$ represent the number of students in the selected sample who prefer data as singular. Use simulation to approximate the posterior predictive distribution of $\tilde{Y}$ and plot it. In particular, find and interpret a 95% posterior *prediction* interval for $\tilde{Y}$. How does the predictive distribution which uses the posterior distribution based on the FiveThirtyEight data compare to the one based on the Cal Poly sample data (31/35)? Why?

Solution. to Example 7.2

1. Results will of course vary, but do consider what your prior would look like.

2. We believe a majority, and probably a strong majority, of students will prefer data as singular. The prior mode is 1, the prior mean is 0.75, and the prior standard deviation is 0.19.

```
theta = seq(0, 1, 0.0001)

# prior
prior = theta ^ 2
prior = prior / sum(prior)

ylim = c(0, max(prior))
plot(theta, prior, type='l', xlim=c(0, 1), ylim=ylim, col="skyblue", xlab='theta',
```



```
# prior mean
prior_ev = sum(theta * prior)
prior_ev
```

```
## [1] 0.75
```

```
# prior variance
prior_var = sum(theta ^ 2 * prior) - prior_ev ^ 2

# prior sd
sqrt(prior_var)
```

```
## [1] 0.1937
```

3. Central credibles would exclude θ values near 1, but these are the values with highest prior probability. For example, a central 50% prior credible interval is [0.630, 0.909], but this excludes values of θ with the highest prior probability. An alternative is to use highest prior probability intervals. For this prior, it seems reasonable to just fix the upper endpoint of the credible intervals to be 1, and to find the lower endpoint corresponding to the desired probability. The lower bound of such a 50% credible interval is the 50th percentile; of an 80% credible interval is the 20th percentile; of a 98% credible interval is the 2nd percentile. There is a prior probability of 50% that at least 79.4% of Cal Poly students prefer data as singular; it's equally plausible that θ is above 0.794 as below.
   There is a prior probability of 80% that at least 58.5% of Cal Poly students prefer data as singular; it's four times more plausible that θ is above 0.585 than below. There is a prior probability of 98% that at least 27.1% of Cal Poly students prefer data as singular; it's 49 times more plausible that θ is above 0.271 than below.

```
prior_cdf = cumsum(prior)

# 50th percentile
theta[max(which(prior_cdf <= 0.5))]
```

```
## [1] 0.7936
```

```
# 10th percentile
theta[max(which(prior_cdf <= 0.2))]
```

```
## [1] 0.5847
```

```
# 2nd percentile
theta[max(which(prior_cdf <= 0.02))]
```

```
## [1] 0.2714
```

4. We use the `sample` function with the `prob` argument to simulate a value of θ from its prior distribution, and then use `rbinom` to simulate a sample. The table below displays the results of a few repetitions of the simulation.

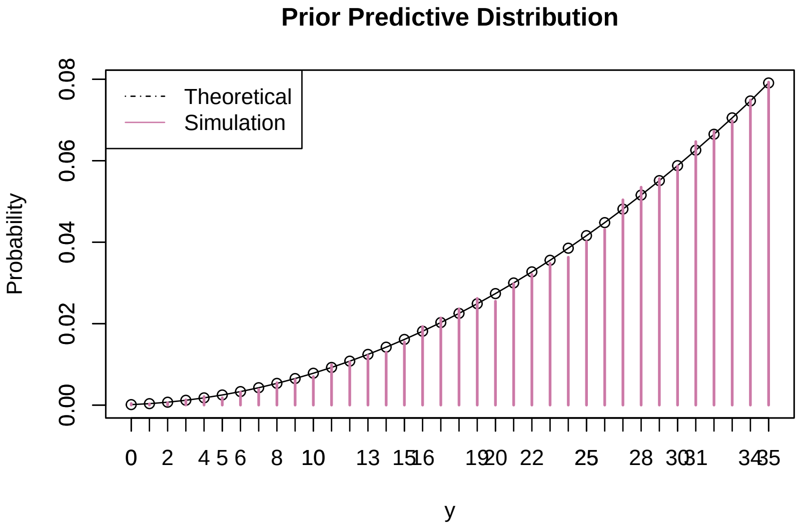
```
n = 35

n_sim = 10000
```

```
theta_sim = sample(theta, n_sim, replace = TRUE, prob = prior)

y_sim = rbinom(n_sim, n, theta_sim)

data.frame(theta_sim, y_sim) %>%
  head(10) %>%
  kable(digits = 5)
```

theta_sim	y_sim
0.6711	26
0.7621	26
0.8947	33
0.8719	30
0.5030	16
0.9272	34
0.8771	31
0.6362	25
0.8140	29
0.1598	3

5. We program the law of total probability calculation for each possible value of y. (There are better ways of doing this than a for loop, but it's good enough.)

```
# Predictive distribution
y_predict = 0:n

py_predict = rep(NA, length(y_predict))

for (i in 1:length(y_predict)) {
  py_predict[i] = sum(dbinom(y_predict[i], n, theta) * prior) # prior
}
```



```
# Prediction interval
py_predict_cdf = cumsum(py_predict)
c(y_predict[max(which(py_predict_cdf <= 0.025))], y_predict[min(which(py_predict_cdf >= 0.97
```

```
## [1]  8 35
```

6. There is prior predictive probability of 95% that between 8 and 35 students in a sample of 35 students will prefer data as singular.

   For the remaining parts, suppose that 31 students in a sample of 35 Cal Poly statistics students prefer data as singular.

7. The observed sample proportion is 31/35=0.886. The posterior distribution is slightly skewed to the left with a posterior mean of 0.872 and a posterior standard deviation of 0.053.

```
# data
n = 35 # sample size
y = 31 # sample count of success

# likelihood, using binomial
likelihood = dbinom(y, n, theta) # function of theta

# posterior
product = likelihood * prior
posterior = product / sum(product)
```

```
# posterior mean
posterior_ev = sum(theta * posterior)
posterior_ev
```

```
## [1] 0.8718
```

```
# posterior variance
posterior_var = sum(theta ^ 2 * posterior) - posterior_ev ^ 2

# posterior sd
sqrt(posterior_var)
```

```
## [1] 0.05286
```

```
# posterior cdf
posterior_cdf = cumsum(posterior)

# posterior 50% central credible interval
c(theta[max(which(posterior_cdf <= 0.25))], theta[min(which(posterior_cdf >= 0.75)
```

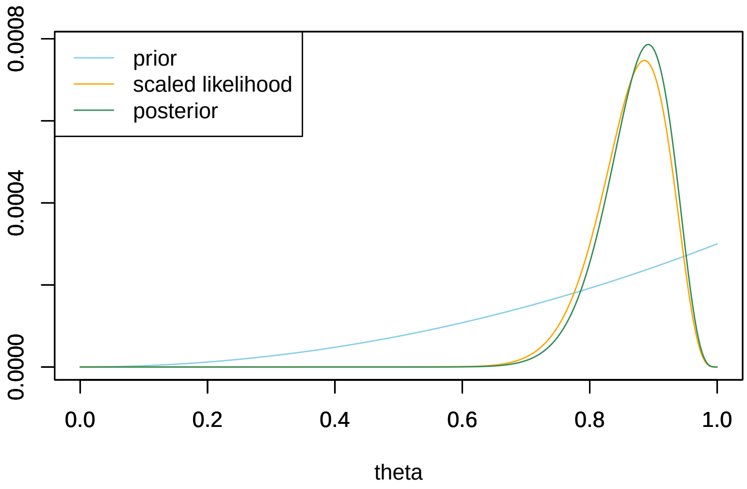
```
## [1] 0.8397 0.9106
```

```
# posterior 80% central credible interval
c(theta[max(which(posterior_cdf <= 0.1))], theta[min(which(posterior_cdf >= 0.9))]
```

```
## [1] 0.8005 0.9346
```

```
# posterior 98% central credible interval
c(theta[max(which(posterior_cdf <= 0.01))], theta[min(which(posterior_cdf >= 0.99)
```

```
## [1] 0.7238 0.9651
```



8. There is a posterior probability of 50% that between 84.0% and 91.1% of Cal Poly students prefer data as singular; after observing the sample data, it's equally plausible that θ is inside [0.840, 0.911] as outside. There is a posterior probability of 80% that between 80.0% and 93.5% of Cal Poly students prefer data as singular; after observing the sample data, it's four times mores plausible that θ is inside [0.800, 0.935] as outside. There is a posterior probability of 98% that between 72.4% and 96.5% of Cal Poly students prefer data as singular; after observing the sample data, it's 49 times mores plausible that θ is inside [0.724, 0.965] as outside.

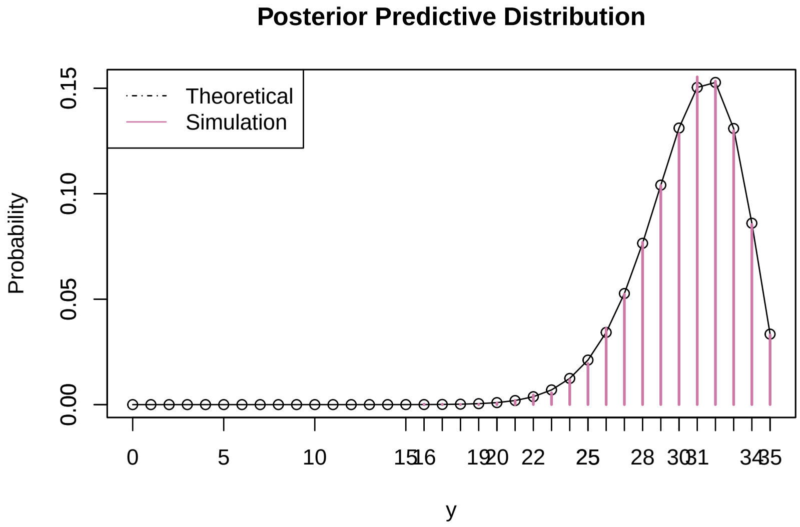
9. Similar to the prior simulation, but now we simulate θ based on its posterior distribution. The table below displays the results of a few repetitions of the simulation.

```
theta_sim = sample(theta, n_sim, replace = TRUE, prob = posterior)

y_sim = rbinom(n_sim, n, theta_sim)

  data.frame(theta_sim, y_sim) %>%
  head(10) %>%
  kable(digits = 5)
```

theta_sim	y_sim
0.8304	30
0.8823	32
0.9251	30
0.9442	35
0.9069	32
0.9083	32
0.9564	34
0.8123	32
0.7888	30
0.8705	29

10. Similar to the prior calculation, but now we use the posterior probabilities as the weights in the law of total probability calculation.

```
# Predictive distribution
y_predict = 0:n

py_predict = rep(NA, length(y_predict))

for (i in 1:length(y_predict)) {
  py_predict[i] = sum(dbinom(y_predict[i], n, theta) * posterior) # posterior
}
```



```
# Prediction interval
py_predict_cdf = cumsum(py_predict)
c(y_predict[max(which(py_predict_cdf <= 0.025))], y_predict[min(which(py_predict_c
## [1] 23 35
```

11. There is posterior predictive probability of 95% that between 23 and 35 students in a sample of 35 students will prefer data as singular.

12. The observed sample proportion is 865/1093=0.791. The posterior mean is 0.791, and the posterior standard deviation is 0.012. There is a posterior probability of 98% that between 76.2% and 81.9% of Cal Poly students prefer data as singular. The posterior SD is much smaller and the 98% credible interval is narrower based on the FiveThirtyEight due to the much larger sample size. (The posterior means and location of the credible intervals are also different due to the difference in sample proportions.)

```
# data
n = 1093 # sample size
y = 865 # sample count of success

# likelihood, using binomial
likelihood = dbinom(y, n, theta) # function of theta

# posterior
product = likelihood * prior
posterior = product / sum(product)

# posterior mean
posterior_ev = sum(theta * posterior)
posterior_ev
```

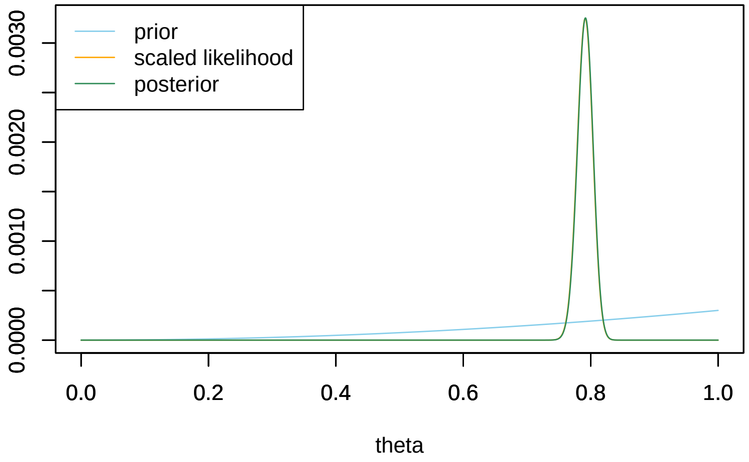
```
## [1] 0.7912
```

```
# posterior variance
posterior_var = sum(theta ^ 2 * posterior) - posterior_ev ^ 2

# posterior sd
sqrt(posterior_var)
```

```
## [1] 0.01227
```

```
# posterior 98% credible interval
posterior_cdf = cumsum(posterior)
c(theta[max(which(posterior_cdf <= 0.01))], theta[min(which(posterior_cdf >= 0.99))])
```

```
## [1] 0.7619 0.8190
```



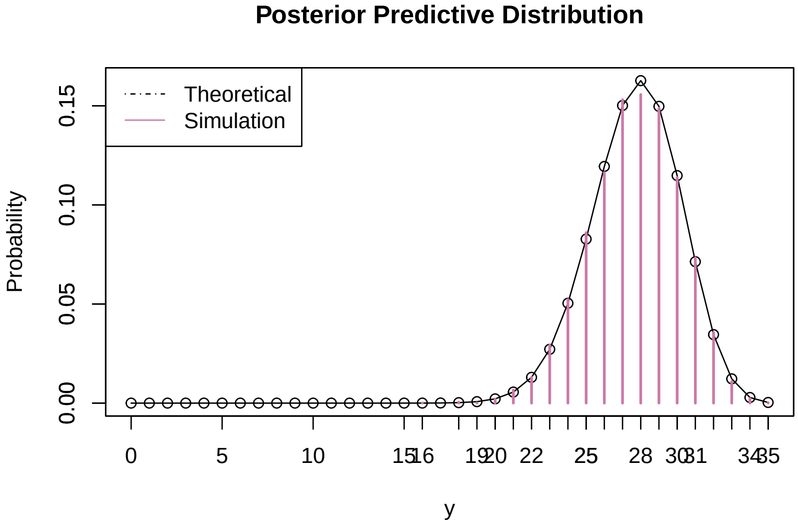
13. There is posterior predictive probability of 95% that between 23 and 35 students in a sample of 35 students will prefer data as singular. Despite the fact that the posterior distributions of θ are different in the two scenarios, the posterior predictive distributions are fairly similar. Even though there is less uncertainty about θ in the FiveThirtyEight case, the predictive distribution reflects the sample-to-sample variability of the number of students who prefer data as singular, which is mainly impacted by the size of the sample being “predicted”. The table below displays a few repetitions of the posterior predictive simulation. Notice that all the θ values are around 0.79 or so, but there is still sample-to-sample variability in the Y values.

```
n = 35

# Predictive simulation
theta_sim = sample(theta, n_sim, replace = TRUE, prob = posterior)

y_sim = rbinom(n_sim, n, theta_sim)

  data.frame(theta_sim, y_sim) %>%
  head(10) %>%
  kable(digits = 5)
```

theta_sim	y_sim
0.7979	32
0.7768	27
0.7559	26
0.8000	33
0.7844	28
0.7823	30
0.7811	25
0.7660	23
0.8057	32
0.7910	27

```
# Predictive distribution
y_predict = 0:n

py_predict = rep(NA, length(y_predict))

for (i in 1:length(y_predict)) {
  py_predict[i] = sum(dbinom(y_predict[i], n, theta) * posterior) # posterior
}

# Prediction interval
py_predict_cdf = cumsum(py_predict)
c(y_predict[max(which(py_predict_cdf <= 0.025))], y_predict[min(which(py_predict_cdf >= 0.97
```

```
## [1] 22 32
```



Be sure to distinguish between a prior/posterior *distribution* and a prior/posterior *predictive distribution*.

- A prior/posterior *distribution* is a distribution on potential values of the *parameters* θ. These distributions quantify the degree of uncertainty about the unknown parameter θ (before and after observing data).
- A prior/posterior *predictive distribution* is a distribution on potential values of the *data* y. Predictive distributions reflect sample-to-sample variability of the sample data, while accounting for the uncertainty in the parameters.

Even if parameters are essentially "known" — that is, even if the prior/posterior variance of parameters is small — there will still be sample-to-sample variability reflected in the predictive distribution of the data, mainly influenced by the size n of the sample being "predicted".

7.1 Posterior predictive checking

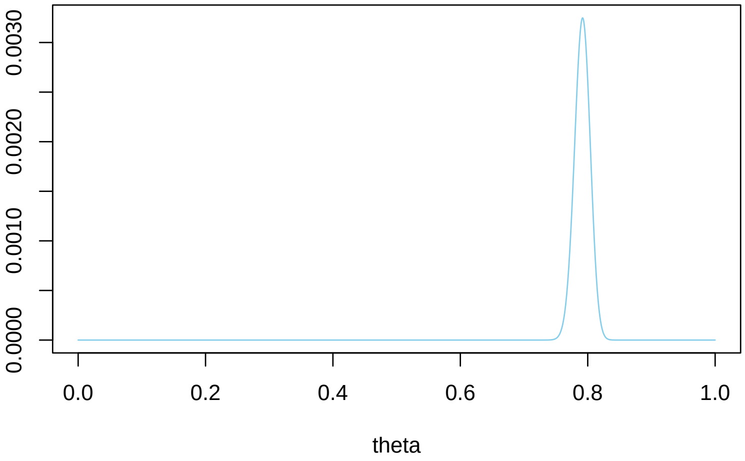
Example 7.3. Continuing the previous example, suppose that before collecting data for our sample of Cal Poly students, we had based our prior distribution off the FiveThirtyEight data. Suppose we assume a prior distribution that is proportional to $\theta^{864}(1-\theta)^{227}$ for θ values in the grid. (We will see where such a distribution might come from later.)

1. Plot the prior distribution. What does this say about our prior beliefs?
2. Now suppose we randomly select a sample of 35 Cal Poly students and 21 students prefer data as singular. Plot the prior and likelihood, and find the posterior distribution and plot it. Have our beliefs about θ changed? Why?
3. Find the posterior predictive distribution corresponding to samples of size 35. Compare the observed sample value of 21/35 with the posterior predictive distribution. What do you notice? Does this indicate problems with the model?

Solution. to Example 7.3

1. We have a very strong prior belief that θ is close to 0.79; the prior SD is only 0.012. There is a prior probability of 98% that between 76% and 82% of Cal Poly students prefer data as singular.

```
# prior
theta = seq(0, 1, 0.0001)
prior = theta ^ 864 * (1 - theta) ^ 227
prior = prior / sum(prior)

ylim = c(0, max(prior))
plot(theta, prior, type='l', xlim=c(0, 1), ylim=ylim, col="skyblue", xlab='theta', ylab='')
```



```
# prior mean
prior_ev = sum(theta * prior)
prior_ev
```

```
## [1] 0.7914
```

```
# prior variance
prior_var = sum(theta ^ 2 * prior) - prior_ev ^ 2

# prior sd
sqrt(prior_var)
```

```
## [1] 0.01228
```

```
# prior 98% credible interval
prior_cdf = cumsum(prior)
c(theta[max(which(prior_cdf <= 0.01))], theta[min(which(prior_cdf >= 0.99))])
```

```
## [1] 0.7620 0.8192
```

2. Our posterior distribution has barely changed from the prior. Even though the sample proportion is $21/35 = 0.61$, our prior beliefs were so strong (represented by the small prior SD) that a sample of size 35 isn't very convincing.

```
# data
n = 35 # sample size
y = 21 # sample count of success

# likelihood, using binomial
likelihood = dbinom(y, n, theta) # function of theta

# posterior
product = likelihood * prior
posterior = product / sum(product)

# posterior mean
posterior_ev = sum(theta * posterior)
posterior_ev
```

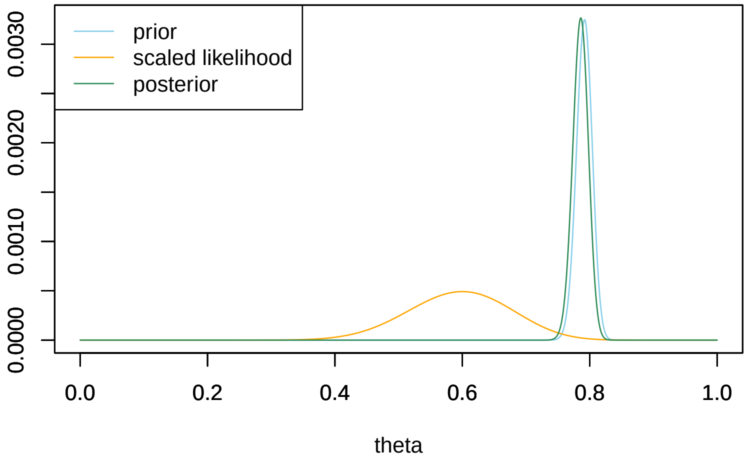
```
## [1] 0.7855
```

```
# posterior variance
posterior_var = sum(theta ^ 2 * posterior) - posterior_ev ^ 2

# posterior sd
sqrt(posterior_var)
```

```
## [1] 0.01222
```

```
# posterior 98% credible interval
posterior_cdf = cumsum(posterior)
c(theta[max(which(posterior_cdf <= 0.01))], theta[min(which(posterior_cdf >= 0.99)
```

```
## [1] 0.7562 0.8131
```

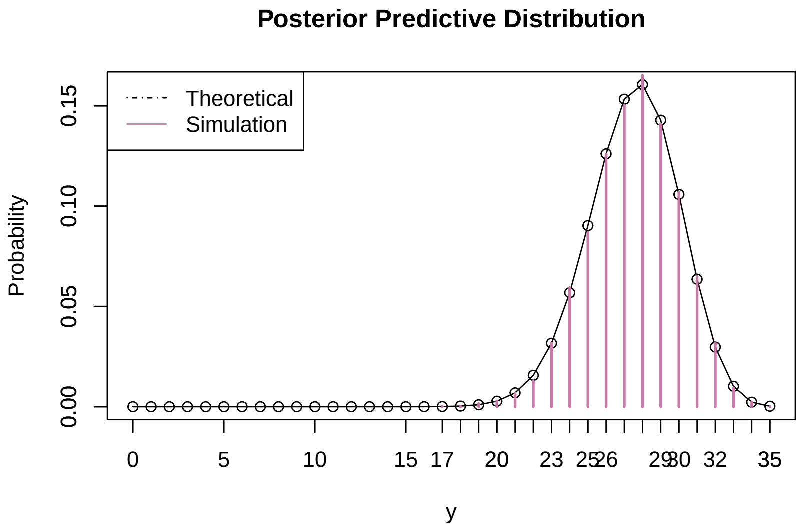


3. According to the posterior predictive distribution, it is very unlikely to observe a sample with only 21 students preferring data as singular; only about 1% of examples are this extreme. However, remember that the posterior predictive distribution is based on the observed data. So we're saying that based on the fact that we observed 21 students in a sample of 35 preferring data as singular it would be unlikely to observe 21 students in a sample of 35 preferring data as singular????? Seems problematic. In this case, the problem is that the prior is way too strict, and it doesn't give the data enough say.

```
n = 35

# Predictive simulation
theta_sim = sample(theta, n_sim, replace = TRUE, prob = posterior)

y_sim = rbinom(n_sim, n, theta_sim)

# Predictive distribution
y_predict = 0:n

py_predict = rep(NA, length(y_predict))

for (i in 1:length(y_predict)) {
  py_predict[i] = sum(dbinom(y_predict[i], n, theta) * posterior) # posterior
}
```

```
    # Prediction interval
sum(py_predict[y_predict <= y])
```

```
## [1] 0.01105
```



A Bayesian model is composed of both a model for the data (likelihood) and a prior distribution on model parameters.

Predictive distributions can be used as tools in model checking. **Posterior predictive checking** involves comparing the observed data to simulated samples (or some summary statistics) generated from the posterior predictive distribution. We'll focus on graphical checks: Compare plots for the observed data with those for simulated samples. Systematic differences between simulated samples and observed data indicate potential shortcomings of the model.

If the model fits the data, then replicated data generated under the model should look similar to the observed data. If the observed data is not plausible under the posterior predictive distribution, then this could indicate that the model is not a good fit for the data. ("Based on the data we observed, we conclude that it would be unlikely to observe the data we observed???")

However, a problematic model isn't necessarily due to the prior. Remember that a Bayesian model consists of both a prior and a likelihood, so model misspecification can occur in the prior or likelihood or both. The form of the likelihood is also based on subjective assumptions about the variables being measured and how the data are collected. Posterior predictive checking can help assess whether these assumptions are reasonable in light of the observed data.

Example 7.4. A basketball player will attempt a sequence of 20 free throws. Our model assumes

- The probability that the player successfully makes any particular free throw attempt is θ.
- A Uniform prior distribution for θ values in a grid from 0 to 1.
- Conditional on θ, the number of successfully made attempts has a Binomial(20, θ) distribution. (This determines the likelihood.)

1. Suppose the player misses her first 10 attempts and makes her second 10 attempts. Does this data seem consistent with the model?
2. Explain how you could use posterior predictive checking to check the fit of the model.

Solution. to Example 7.4

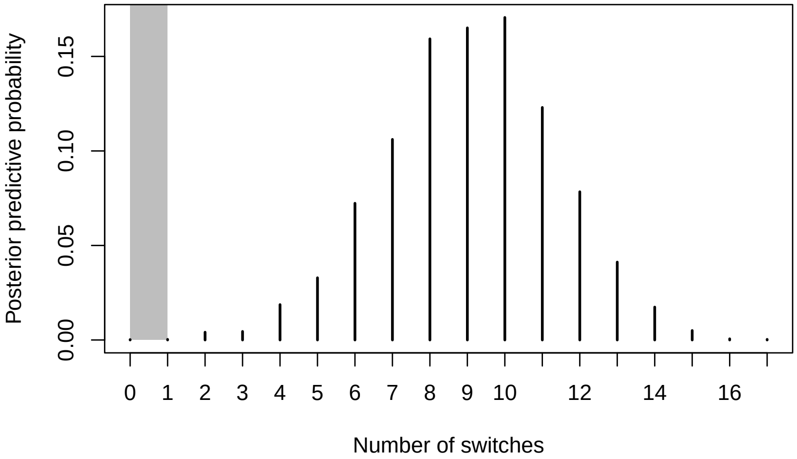
1. Remember that one condition of a Binomial model is independence of trials: the probability of success on a shot should not depend on the results of previous shots. However, independence seems to be violated here, since the shooter has a long hot streak followed by a long cold streak. So a Binomial model might not be appropriate.

2. We're particularly concerned about the independence assumption, so how could we check that? For example, the data seems consistent with a value of $\theta = 0.5$, but if the trials were independent, you would expect to see more alterations between makes and misses. So one way to measure degree of dependence is to count the number of "switches" between makes and misses. For the observed data there is only 1 switch. We can use simulation to approximate the posterior predictive distribution of the number of switches assuming the model is true, and then we can see if a value of 1 (the observed number of switches) would be consistent with the model.

    1. Find the posterior distribution of θ. Simulate a value of θ from its posterior distribution.
    2. Given θ, simulate a sequence of 20 independent success/failure trials with probability of success θ on each trial. Compute the number of switches for the sequence. (Since we're interested in the number of switches, we have to generate the individual success/failure results, and not just the total number of successes).
    3. Repeat many times, recording the total number of switches each time. Summarize the values to approximate the posterior predictive distribution of the number of switches.

See the simulation results below. It would be very unlikely to observe only 1 switch in 20 independent trials. Therefore, the proposed model does not fit the observed data well. There is evidence that the assumption of independence is violated.

```r
theta = seq(0, 1, 0.0001)

# prior
prior = rep(1, length(theta))
prior = prior / sum(prior)

# data
n = 20 # sample size
y = 10 # sample count of success

# likelihood, using binomial
likelihood = dbinom(y, n, theta) # function of theta

# posterior
product = likelihood * prior
posterior = product / sum(product)

# predictive simulation

n_sim = 10000

switches = rep(NA, n_sim)

for (r in 1:n_sim){
  theta_sim = sample(theta, 1, replace = TRUE, prob = posterior)
  trials_sim = rbinom(n, 1, theta_sim)
  switches[r] = length(rle(trials_sim)$lengths) - 1 # built in function
}

plot(table(switches) / n_sim,
 xlab = "Number of switches",
 ylab = "Posterior predictive probability",
 panel.first = rect(0, 0, 1, 1, col='gray', border=NA))
```



```
sum(switches <= 1) / n_sim
```

```
## [1] 0.0005
```

7.2 Prior predictive tuning

Prior distributions of parameters quantify uncertainty about parameters before observing data. Considering prior predictive distributions of possible samples under the proposed model can help tune prior distributions of parameters.

Example 7.5. Suppose we want to estimate θ, the population mean hours of sleep on a typical night for Cal Poly students. Assume that sleep hours for individual students follow a Normal distribution with unknown mean θ and known standard deviation 1.5 hours. (Known population SD is an unrealistic assumption that we use for simplicity here.)

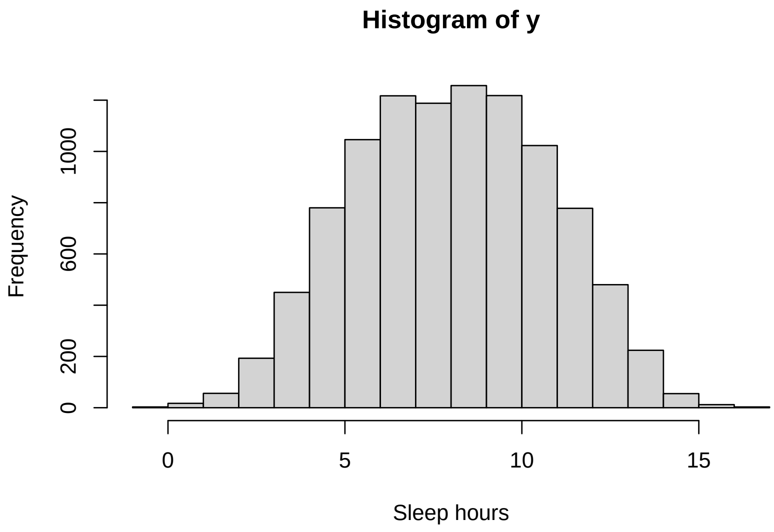
Suppose we want to use a fairly uninformative prior for θ, so we choose a Uniform distribution on the interval [4, 12].

1. Simulate sleep hours for 10000 Cal Poly students under this model and make a histogram of the simulated values.
2. According to this model, (approximately) what percent of students sleep less than 5 hours a night? More than 11? Do these values seem reasonable?

Solution. to Example 7.5

1. First simulate a value θ from the Uniform(4, 12) distribution. Then given θ simulate a value y from a Normal(θ, 1.5) distribution. Repeat many times to get many (θ, y) pairs and summarize the y values.

```
N_sim = 10000

theta = runif(N_sim, 4, 12)

sigma = 1.5
y = rnorm(N_sim, theta, sigma)

hist(y, xlab = "Sleep hours")
```



```
sum(y < 5) / N_sim
```

```
## [1] 0.1499
```

```
sum(y > 11) / N_sim
```

```
## [1] 0.1552
```

2. According to this model, about 15 percent of students sleep fewer than 5 hours on a typical night, and about 16 percent of students sleep more than 11 hours on a typical night. These values seem to be overestimates, indicating that perhaps the model isn't the greatest.

> It could be that the prior distribution is too uninformative. But it could also be that the assumptions of the data model are inadequate; perhaps a Normal distribution isn't appropriate for sleep times. (Of course, the value σ could be also wrong, but here we're assuming it's known.)

In the previous example, it was helpful to think about the distribution of sleep hours for individual students when formulating prior beliefs about the population mean. In general, it is often easier to think in terms of the scale of the data (individual sleep hours) rather than the scale of the parameters (*mean* sleep hours).

Prior predictive distributions "live" on the scale of the data, and are sometimes easier to interpret than prior distributions themselves. It is often helpful to tune prior distributions indirectly via prior predictive distributions rather than directly. We can choose a prior distribution for parameters, simulate a prior predictive distribution for the data given this prior, and consider if the distribution of possible data values seems reasonable given our background knowledge about the variable and context. If not, we can choose another prior and repeat the process until we have suitably "tuned" the prior.

Remember, the prior does not have to be perfect; there is no perfect prior. However, if a particular prior gives rise to obviously unreasonable data values (e.g., negative sleep hours) we should try to improve it. It's always a good idea to consider prior predictive distributions when formulating a prior distribution for parameters.

Chapter 8

Introduction to Continuous Prior and Posterior Distributions

Bayesian analysis is based on the posterior distribution of parameters θ given data y. The data y might be discrete (e.g., count data) or continuous (e.g., measurement data). However, *parameters* θ almost always take values on a *continuous* scale, even when the data are discrete. For example, in a Binomial situation, the number of successes y takes values on a discrete scale, but the probability of success on any single trial θ can potentially take any value in the continuous interval (0, 1).

Recall that the posterior distribution is proportional to the product of the prior distribution and the likelihood. Thus, there are two probability distributions which will influence the posterior distribution.

- The (unconditional) prior distribution of parameters θ, which is (almost always) a continuous distribution
- The conditional distribution of the data y given parameters θ, which determines the likelihood function. Viewed as a conditional distribution of y given θ, the distribution can be discrete or continuous, corresponding to the data type of y. However, the likelihood function treats the data y as fixed and the parameters θ as varying, and therefore the likelihood function is (almost always) a function of continuous θ.

This section provides an introduction to using continuous prior and posterior distributions to quantify uncertainty about parameters. Some general notation:

- θ represents[1] parameters of interest usually taking values on a continuous scale
- y denotes observed sample data (discrete or continuous)
- $\pi(\theta)$ denotes the prior distribution of θ, usually a probability density function (pdf) over possible values of θ
- $f(y|\theta)$ denotes the likelihood function, a function of continuous θ for fixed y
- $\pi(\theta|y)$ denotes the posterior distribution of θ, the conditional distribution of θ given the data y.

Bayes rule works analogously for a continuous parameter θ, given data y

$$\pi(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{f_Y(y)}$$

$$\pi(\theta|y) \propto f(y|\theta)\pi(\theta)$$
$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

The continuous analog of the law of total probability is

$$f_Y(y) = \int_{-\infty}^{\infty} f(y|\theta)\pi(\theta)d\theta$$

8.1 A brief review of continuous distributions

This section provides a brief review of continuous probability distributions. Throughout, U represents a continuous random variable that takes values denoted u. In a Bayesian framework, u can represent either values of parameters θ or values of data y.

The probability distribution of a *continuous* random variable is (usually) specified by its **probability density function (pdf)** (a.k.a., density), usually denoted f or f_U. A pdf f must satisfy:

$$f(u) \geq 0 \qquad \text{for all } u$$
$$\int_{-\infty}^{\infty} f(u)du = 1$$

For a continuous random variable U with pdf f the probability that the random variable falls between any two values a and b is given by the *area* under the density between those two values.

$$P(a \leq U \leq b) = \int_a^b f(u)du$$

[1] θ is used to denote both: (1) the actual parameter (i.e., the random variable) θ itself, and (2) possible values of θ.

A pdf will assign zero probability to intervals where the density is 0. A pdf is usually defined for all real values, but is often nonzero only for some subset of values, the possible values of the random variable. Given a specific pdf, the generic bounds $(-\infty, \infty)$ should be replaced by the range of possible values, that is, those values u for which $f(u) > 0$.

For example, if U can only take positive values we can write its pdf as

$$f(u) = \begin{cases} \text{some function of } u, & u > 0, \\ 0, & \text{otherwise} \end{cases}$$

The "0 otherwise" part is often omitted, but be sure to specify the range of values where f is positive.

The expected value of a continuous random variable U with pdf f is

$$E(U) = \int_{-\infty}^{\infty} u\, f(u)\, du$$

The probability that a continuous random variable U equals any particular value is 0: $P(U = u) = 0$ for all u. A continuous random variable can take uncountably many distinct values, e.g. 0.500000000... is different than 0.50000000010... is different than 0.500000000000001..., etc. Simulating values of a continuous random variable corresponds to an idealized spinner with an infinitely precise needle which can land on any value in a continuous scale.

A density is an idealized mathematical model for the entire population distribution of infinitely many distinct values of the random variable. In practical applications, there is some acceptable degree of precision, and events like "X, rounded to 4 decimal places, equals 0.5" correspond to intervals that do have positive probability. For continuous random variables, it doesn't really make sense to talk about the probability that the random value equals a particular value. However, we can consider the probability that a random variable is *close to* a particular value.

The density $f(u)$ at value u is *not* a probability But the density $f(u)$ at value u is related to the probability that the random variable U takes a value "close to u" in the following sense

$$P\left(u - \frac{\epsilon}{2} \le U \le u + \frac{\epsilon}{2}\right) \approx f(u)\epsilon, \qquad \text{for small } \epsilon$$

So a random variable U is more likely to take values close to those with greater density.

In general, a pdf is often defined only up to some multiplicative constant c, for example

$$\begin{aligned} &f(u) = c \times \text{some function of } u, \quad \text{or} \\ &f(u) \propto \text{some function of } u \end{aligned}$$

The constant c does not affect the shape of the density as a function of u, only the scale on the density (vertical) axis. The absolute scaling on the density axis is somewhat irrelevant; it is whatever it needs to be to provide the proper area. In particular, the total area under the pdf must be 1. The scaling constant is determined by the requirement that $\int_{-\infty}^{\infty} f(u)du = 1$. (Remember to replace the generic $(-\infty, \infty)$ bounds with the range of possible values.)

What is important about the pdf is *relative* height. For example, if two values u and $\tilde{u}$ satisfy $f(\tilde{u}) = 2f(u)$ then U is roughly "twice as likely to be near $\tilde{u}$ than u"

$$2 = \frac{f(\tilde{u})}{f(u)} = \frac{f(\tilde{u})\epsilon}{f(u)\epsilon} \approx \frac{P\left(\tilde{u} - \frac{\epsilon}{2} \le U \le \tilde{u} + \frac{\epsilon}{2}\right)}{P\left(u - \frac{\epsilon}{2} \le U \le u + \frac{\epsilon}{2}\right)}$$

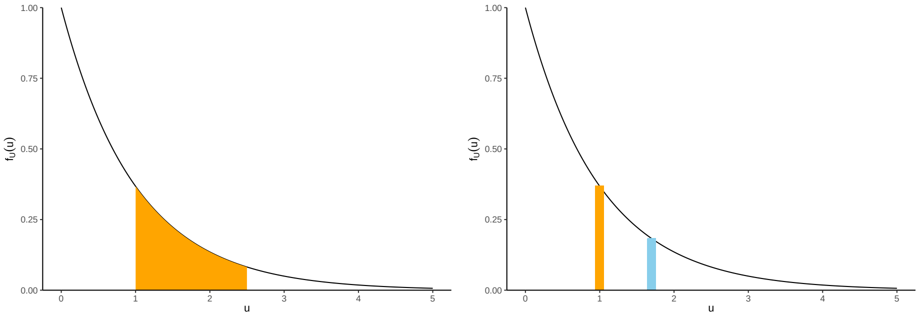
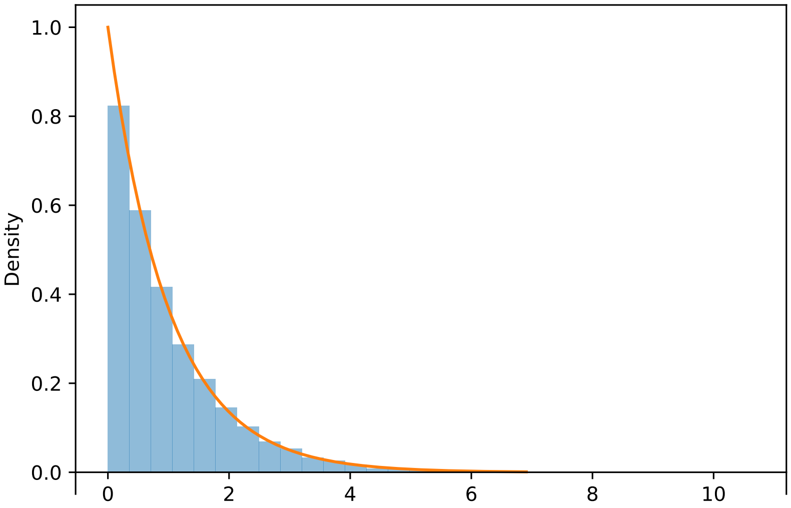


Figure 8.1: Illustration of $P(1 < U < 2.5)$ (left) and $P(0.995 < U < 1.005)$ and $P(1.695 < U < 1.705)$ (right) for U with an Exponential(1) distribution, with pdf $f_U(u) = e^{-u}, u > 0$. The plot on the left displays the true area under the curve over (1, 2.5). The plot on the right illustrates how the probability that U is "close to" u can be approximated by the area of a rectangle with height equal to the density at u, $f_U(u)$. The density height at $u = 1$ is twice as large than the density height at $u = 1.7$, so the probability that U is "close to" 1 is (roughly) twice as large as the probability that U is "close to" 1.7.

A sample of values of a continuous random variable is often displayed in a **histogram** which displays the frequencies of values falling in interval "bins". The vertical axis of a histogram is typically on the density scale, so that *areas* of the bars correspond to relative frequencies.



8.2 Continuous distributions for a population proportion

We have seen a few examples where we used Normal distributions as prior distributions for a population proportion θ. Normal distributions are commonly used as priors, but they do not allow for asymmetric prior distributions. We'll now consider Beta distributions, a family of distributions that are commonly used as prior distributions for population proportions.

Example 8.1. Continuing Example 7.1 where θ represents the population proportion of students in Cal Poly statistics classes who prefer to consider data as a singular noun.

1. Assume a continuous prior distribution for θ which is proportional to $\theta^2,\ 0 < \theta < 1$. Sketch this distribution.
2. The previous part implies that $\pi(\theta) = c\theta^2,\ 0 < \theta < 1$, for an appropriate constant c. Find c.
3. Compute the prior mean of θ.
4. Now we'll consider a few more prior distributions. Sketch each of the following priors. How do they compare?

   a. proportional to $\theta^2,\ 0 < \theta < 1$. (from previous)
   b. proportional to $\theta^5,\ 0 < \theta < 1$.

c. proportional to $(1-\theta)^2,\ 0 < \theta < 1$.
d. proportional to $\theta^2(1-\theta)^2,\ 0 < \theta < 1$.
e. proportional to $\theta^5(1-\theta)^2,\ 0 < \theta < 1$.

Solution. to Example 8.1

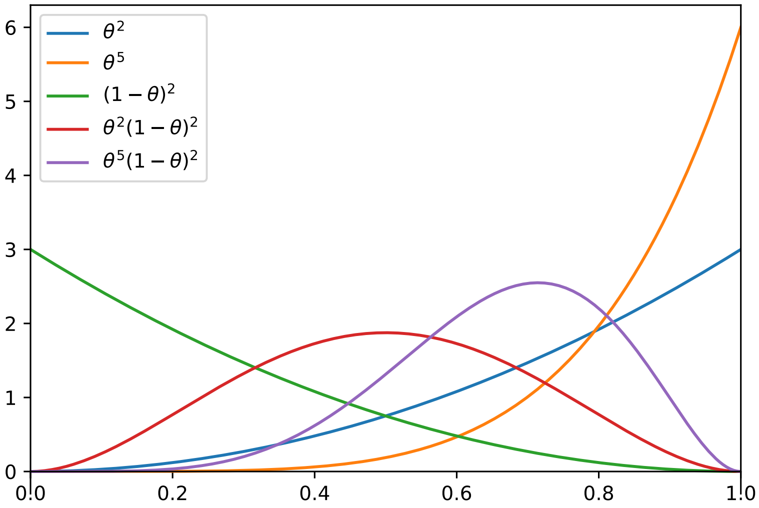
1. See the plot below. The distribution is similar to the discrete grid approximation in Example 7.2.
2. Set the total area under the curve equal to 1 and solve for $c = 3$

$$1 = \int_0^1 c\theta^2 d\theta = c\int_0^1 \theta^2 d\theta = c(1/3) \Rightarrow c = 3$$

3. Since θ is continuous we use calculus

$$E(\theta) = \int_0^1 \theta\,\pi(\theta) d\theta = \int_0^1 \theta(3\theta^2) d\theta = 3/4$$

4. See the plot below. The prior proportional to $(1-\theta)^2$ is the mirror image of the prior proportional to θ^2, reflected about 0.5. As the exponent on θ increases, more density is shifted towards 1. As the exponent on $1-\theta$ increases, more density is shifted towards 0. When the exponents are the same, the density is symmetric about 0.5



A continuous random variable U has a **Beta distribution** with *shape parameters* $\alpha > 0$ and $\beta > 0$ if its density satisfies[2]

$$f(u) \propto u^{\alpha-1}(1-u)^{\beta-1}, \quad 0 < u < 1,$$

and $f(u) = 0$ otherwise.

- If $\alpha = \beta$ the distribution is symmetric about 0.5
- If $\alpha > \beta$ the distribution is skewed to the left (with greater density above 0.5 than below)
- If $\alpha < \beta$ the distribution is skewed to the right (with greater density below 0.5 than above)
- If $\alpha = 1$ and $\beta = 1$, the Beta(1, 1) distribution is the Uniform distribution on (0, 1).

It can be shown that a Beta(α, β) density has

$$\begin{aligned}
\text{Mean (EV):} \quad & \frac{\alpha}{\alpha+\beta} \\
\text{Variance:} \quad & \frac{\left(\frac{\alpha}{\alpha+\beta}\right)\left(1-\frac{\alpha}{\alpha+\beta}\right)}{\alpha+\beta+1} \\
\text{Mode:} \quad & \frac{\alpha-1}{\alpha+\beta-2}, \qquad (\text{if } \alpha > 1, \beta \geq 1 \text{ or } \alpha \geq 1, \beta > 1)
\end{aligned}$$

Example 8.2. Continuing Example 8.1

1. Each of the previous distributions in the previous example was a Beta distribution. For each distribution, identify the shape parameters and the prior mean and standard deviation.

   a. proportional to $\theta^2,\ 0 < \theta < 1$.
   b. proportional to $\theta^5,\ 0 < \theta < 1$.
   c. proportional to $(1-\theta)^2,\ 0 < \theta < 1$.
   d. proportional to $\theta^2(1-\theta)^2,\ 0 < \theta < 1$.
   e. proportional to $\theta^5(1-\theta)^2,\ 0 < \theta < 1$.

2. Now suppose that 31 students in a sample of 35 Cal Poly statistics students prefer data as singular. Specify the shape of the likelihood as a function of $\theta, 0 < \theta < 1$.

[2] The expression defines the shape of the Beta density. All that's missing is the scaling constant which ensures that the total area under the density is 1. The actual Beta density formula, including the normalizing constant, is

$$f(u) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, u^{\alpha-1}(1-u)^{\beta-1}, \quad 0 < u < 1,$$

where $\Gamma(\alpha) = \int_0^\infty e^{-v}v^{\alpha-1}dv$ is the *Gamma function*. For a positive integer k, $\Gamma(k) = (k-1)!$. Also, $\Gamma(1/2) = \sqrt{\pi}$.

3. Starting with each of the prior distributions from the first part, find the posterior distribution of θ based on this sample, and identify it as a Beta distribution by specifying the shape parameters α and β

   a. proportional to $\theta^2,\ 0 < \theta < 1$.
   b. proportional to $\theta^5,\ 0 < \theta < 1$.
   c. proportional to $(1-\theta)^2,\ 0 < \theta < 1$.
   d. proportional to $\theta^2(1-\theta)^2,\ 0 < \theta < 1$.
   e. proportional to $\theta^5(1-\theta)^2,\ 0 < \theta < 1$.

4. For each of the posterior distributions in the previous part, compute the posterior mean and standard deviation. How does each posterior distribution compare to its respective prior distribution?

Solution. to Example 8.2

1. Careful with the exponents. For example, $\theta^2 = \theta^2(1-\theta)^0 = \theta^{3-1}(1-\theta)^{1-1}$, which corresponds to a Beta(3, 1) distribution.

	Distribution	α	β	Proportional to	Mean	SD
a	Beta(3, 1)	3	1	$\theta^2, 0 < \theta < 1$	0.750	0.194
b	Beta(6, 1)	6	1	$\theta^5, 0 < \theta < 1$	0.857	0.124
c	Beta(1, 3)	1	3	$(1-\theta)^2, 0 < \theta < 1$	0.250	0.194
d	Beta(3, 3)	3	3	$\theta^2(1-\theta)^2, 0 < \theta < 1$	0.500	0.189
e	Beta(6, 3)	6	3	$\theta^5(1-\theta)^2, 0 < \theta < 1$	0.667	0.149

2. Given θ, the number of students in the sample who prefer data as singular, Y, follows a Binomial(35, θ) distribution. The likelihood is the probability of observing $Y = 31$ viewed as a function of θ.

$$\begin{aligned} f(31|\theta) &= \binom{35}{31}\theta^{31}(1-\theta)^4, \qquad 0 < \theta < 1 \\ &\propto \theta^{31}(1-\theta)^4, \qquad 0 < \theta < 1 \end{aligned}$$

   The constant $\binom{35}{31}$ does not affect the *shape* of the likelihood as a function of θ.

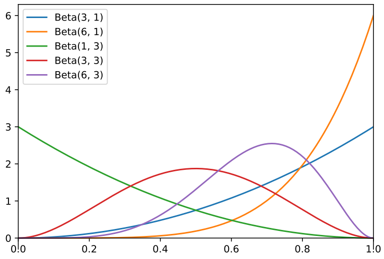
3. As always, the posterior distribution is proportional to the product of the prior distribution and the likelihood. For the Beta(3, 1) prior, the prior density is proportional to θ^2, $0 < \theta < 1$, and for the observed data $y = 31$ with $n = 35$, the likelihood is proportional to $\theta^{31}(1-\theta)^4$, $0 < \theta < 1$. Therefore, the posterior density, as a function of θ, is proportional to

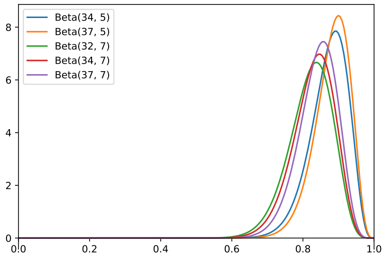
$$\begin{aligned} \pi(\theta|y=31) &\propto \left(\theta^2\right)\left(\theta^{31}(1-\theta)^4\right), \qquad 0 < \theta < 1 \\ &\propto \theta^{33}(1-\theta)^4, \qquad 0 < \theta < 1 \\ &\propto \theta^{34-1}(1-\theta)^{5-1}, \qquad 0 < \theta < 1 \end{aligned}$$

Therefore, the posterior distribution of θ is the Beta(3 + 31, 1 + 35 - 31), that is, the Beta(34, 5) distribution. The other situations are similar. The prior changes but the likelihood stays the same, based on a sample with 31 successes and $35 - 31 = 4$ failures. If the prior distribution is Beta(α, β) then the posterior distribution is Beta($\alpha + 31$, $\beta + 35 - 31$).

	Prior Distribution	Posterior proportional to	Posterior Distribution	Posterior Mean	Posterior SD
a	Beta(3, 1)	$\theta^{2+31}(1-\theta)^{0+4}, 0 < \theta < 1$	Beta(34, 5)	0.872	0.053
b	Beta(6, 1)	$\theta^{5+31}(1-\theta)^{0+4}, 0 < \theta < 1$	Beta(37, 5)	0.881	0.049
c	Beta(1, 3)	$\theta^{0+31}(1-\theta)^{2+4}, 0 < \theta < 1$	Beta(32, 7)	0.821	0.061
d	Beta(3, 3)	$\theta^{2+31}(1-\theta)^{2+4}, 0 < \theta < 1$	Beta(34, 7)	0.829	0.058
e	Beta(6, 3)	$\theta^{5+31}(1-\theta)^{2+4}, 0 < \theta < 1$	Beta(37, 7)	0.841	0.055

4. See the table above. Each posterior distribution concentrates more probability towards the observed sample proportion $31/35 = 0.886$, though there are some small differences due to the prior. The posterior SD is less than the prior SD; there is less uncertainty about θ after observing some data.





Beta distributions are often used in Bayesian models involving population proportions. Consider some binary ("success/failure") variable and let θ be the population proportion of success. Select a random sample of size n from the population and let Y count the number of successes in the sample.

Beta-Binomial model. If θ has a Beta(α, β) prior distribution and the conditional distribution of Y given θ is the Binomial(n, θ) distribution, then the posterior distribution of θ given y is the Beta$(\alpha + y, \beta + n - y)$ distribution.

$$\text{prior:} \qquad \pi(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}, \quad 0 < \theta < 1,$$

$$\text{likelihood:} \qquad f(y|\theta) \propto \theta^{y}(1-\theta)^{n-y}, \quad 0 < \theta < 1,$$

$$\text{posterior:} \qquad \pi(\theta|y) \propto \theta^{\alpha+y-1}(1-\theta)^{\beta+n-y-1}, \quad 0 < \theta < 1.$$

Try this applet which illustrates the Beta-Binomial model.

In a sense, you can interpret α as "prior successes" and β as "prior failures", but these are only "pseudo-observations". Also, α and β are not necessarily integers.

	Prior	Data	Posterior
Successes	α	y	$\alpha + y$
Failures	β	$n - y$	$\beta + n - y$
Total	$\alpha + \beta$	n	$\alpha + \beta + n$

When the prior and posterior distribution belong to the same family, that family is called a **conjugate** prior distribution for the likelihood. So, the Beta distributions form a conjugate prior family for Binomial distributions.

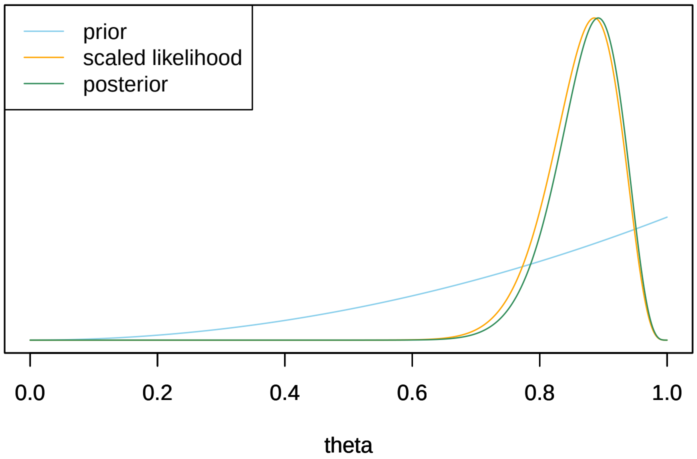
Example 8.3. In Example 7.2 we used a grid approximation to the prior distribution of θ. Now we will assume a continuous prior distributions. Assume that θ has a Beta(3, 1) prior distribution and that 31 students in a sample of 35 Cal Poly statistics students prefer data as singular.

1. Plot the prior distribution, (scaled) likelihood, and posterior distribution.
2. Use software to find 50%, 80%, and 98% central posterior credible intervals.
3. Compare the results to those using the grid approximation in Example 7.2.
4. Express the posterior mean as a weighted average of the prior mean and sample proportion. Describe what the weights are, and explain why they make sense.

Solution. to Example 8.3

1. See plot below. The posterior distribution is the Beta(34, 5) distribution. Note that the grid in the code is just to plot things in R. In particular, the posterior is computed using the Beta-Binomial model, not the grid.

```
theta = seq(0, 1, 0.0001) # the grid is just for plotting

# prior
alpha_prior = 3
beta_prior = 1
prior = dbeta(theta, alpha_prior, beta_prior)

# data
n = 35
y = 31

# likelihood
likelihood = dbinom(y, n, theta)

# posterior
alpha_post = alpha_prior + y
beta_post = beta_prior + n - y
posterior = dbeta(theta, alpha_post, beta_post)

# plot
ymax = max(c(prior, posterior))
scaled_likelihood = likelihood * ymax / max(likelihood)

plot(theta, prior, type='l', col='skyblue', xlim=c(0, 1), ylim=c(0, ymax), ylab='', yaxt='n'
par(new=T)
plot(theta, scaled_likelihood, type='l', col='orange', xlim=c(0, 1), ylim=c(0, ymax), ylab='
par(new=T)
plot(theta, posterior, type='l', col='seagreen', xlim=c(0, 1), ylim=c(0, ymax), ylab='', yax
legend("topleft", c("prior", "scaled likelihood", "posterior"), lty=1, col=c("skyblue", "ora
```



```
# 50% posterior credible interval
qbeta(c(0.25, 0.75), alpha_post, beta_post)
```

```
## [1] 0.8398 0.9106
```

```
# 80% posterior credible interval
qbeta(c(0.1, 0.9), alpha_post, beta_post)
```

```
## [1] 0.8006 0.9346
```

```
# 98% posterior credible interval
qbeta(c(0.01, 0.99), alpha_post, beta_post)
```

```
## [1] 0.7239 0.9651
```

2. We can use `qbeta` to compute quantiles (a.k.a. percentiles). The posterior mean is 0.872, and the prior standard deviation is 0.053. There is a posterior probability of 50% that between 84.0% and 91.1% of Cal Poly students prefer data as singular; after observing the sample data, it's equally plausible that θ is inside [0.840, 0.911] as outside. There is a posterior probability of 80% that between 80.0% and 93.5% of Cal Poly students prefer data as singular; after observing the sample data, it's four times mores plausible that θ is inside [0.800, 0.935] as outside. There is a posterior probability of 98% that between 72.4% and 96.5% of Cal Poly students prefer data as singular; after observing the sample data, it's 49 times mores plausible that θ is inside [0.724, 0.965] as outside.

3. The results based on continuous distributions are the same as those for the grid approximation. The grid is just an approximation of the "true" Beta-Binomial theory.

4. The prior mean is $\frac{3}{3+1} = 0.75$. The sample proportion is $\frac{31}{35} = 0.886$. The posterior mean is $\frac{34}{39} = 0.872$. We can write

$$\begin{aligned}\frac{34}{39} &= \left(\frac{3}{4}\right) \times \left(\frac{4}{39}\right) + \left(\frac{31}{35}\right) \times \left(\frac{35}{39}\right)\\ &= \left(\frac{3}{4}\right) \times \left(\frac{4}{4+35}\right) + \left(\frac{31}{35}\right) \times \left(\frac{35}{4+35}\right)\end{aligned}$$

   The posterior mean is a weighted average of the prior mean and the sample proportion where the weights are given by the relative "samples sizes". The "prior sample size" is $3+1=4$. The actual observed sample size is 35.

In the Beta-Binomial model, the posterior mean $E(\theta|y)$ can be expressed as a *weighted average* of the prior mean $E(\theta) = \frac{\alpha}{\alpha+\beta}$ and the sample proportion $\hat{p} = y/n$.

$$E(\theta|y) = \frac{\alpha+\beta}{\alpha+\beta+n}E(\theta) + \frac{n}{\alpha+\beta+n}\hat{p}$$

As more data are collected, more weight is given to the sample proportion (and less weight to the prior mean). The prior "weight" is detemined by $\alpha+\beta$, which is sometimes called the *concentration* and measured in "pseudo-observations". Larger values of $\alpha+\beta$ indicate stronger prior beliefs, due to smaller prior variance, and give more weight to the prior mean.

The posterior variance generally gets smaller as more data are collected

$$\text{Var}(\theta|y) = \frac{E(\theta|y)(1-E(\theta|y))}{\alpha+\beta+n+1}$$

Example 8.4. Now let's reconsider the posterior prediction parts of Example 7.2, treating θ as continuous. Assume that θ has a Beta(3, 1) prior distribution and that 31 students in a sample of 35 Cal Poly statistics students prefer data as singular, so that the posterior distribution of θ is the Beta(34, 5) distribution.

1. Suppose we plan to randomly select another sample of 35 Cal Poly statistics students. Let $\tilde{Y}$ represent the number of students in the selected sample who prefer data as singular. How could we use simulation to approximate the posterior predictive distribution of $\tilde{Y}$?
2. Use software to run the simulation and plot the posterior predictive distribution[3]. Compare to Example 7.2.

[3] The posterior predictive distribution can be found analytically in the Beta-Binomial sit-

3. Use the simulation results to approximate a 95% posterior *prediction* interval for $\tilde{Y}$. Write a clearly worded sentence interpreting this interval in context.

Solution. to Example 8.3

1. Simulate a value of θ from the posterior Beta(34, 5) distribution. Given this value of θ, simulate a value $\tilde{y}$ from a Binomial(35, θ) distribution. Repeat many times, simulating many $(\theta, \tilde{y})$ pairs. The simulated distribution of $\tilde{y}$ values will approximate the posterior predictive distribution.

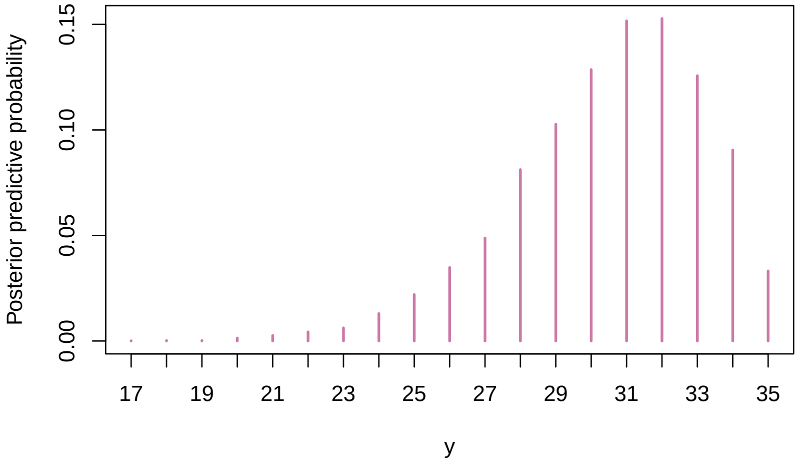
2. We can use `rbeta` to simulate from a Beta distribution. The simulation results are similar to those from the grid approximation.

```
n_sim = 10000

theta_sim = rbeta(n_sim, 34, 5)

y_sim = rbinom(n_sim, 35, theta_sim)

plot(table(y_sim) / n_sim, xlab = "y", ylab = "Posterior predictive probability",
```

uation. If $\theta \sim$ Beta(α, β) and $(Y|\theta) \sim$ Binomial(n, θ) then the marginal distribution of Y is the Beta-Binomial distribution with

$$P(Y = y) = \binom{n}{y} \frac{B(\alpha + y, \beta + n - y)}{B(\alpha, \beta)}, \qquad y = 0, 1, \dots, n,$$

$B(\alpha, \beta)$ is the *beta function*, for which $B(\alpha, \beta) = \frac{(\alpha-1)!(\beta-1)!}{(\alpha+\beta-1)!}$ if α, β are positive integers. (For general $\alpha, \beta > 0$, $B(\alpha, \beta) = \int_0^1 u^{\alpha-1}(1-u)^{\beta-1} du = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$.) The mean is $n\left(\frac{\alpha}{\alpha+\beta}\right)$. In R: `dbbinom, rbbinom, pbbinom` in `extraDistr` package



```
quantile(y_sim, c(0.025, 0.975))
```

```
##  2.5% 97.5%
##    24    35
```

3. The interval is similar to the one from the grid approximation, and the interpretation is the same. There is posterior predictive probability of 95% that between 24 and 35 students in a sample of 35 students will prefer data as singular.

You can tune the shape parameters — α (like "prior successes") and β (like "prior failures") — of a Beta distribution to your prior beliefs in a few ways. Recall that $\kappa = \alpha + \beta$ is the "concentration" or "equivalent prior sample size".

- If prior mean μ and prior concentration κ are specified then

$$\begin{aligned}\alpha &= \mu\kappa \\ \beta &= (1-\mu)\kappa\end{aligned}$$

- If prior mode ω and prior concentration κ (with $\kappa > 2$) are specified then

$$\begin{aligned}\alpha &= \omega(\kappa-2)+1 \\ \beta &= (1-\omega)(\kappa-2)+1\end{aligned}$$

- If prior mean μ and prior sd σ are specified then

$$\alpha = \mu\left(\frac{\mu(1-\mu)}{\sigma^2} - 1\right)$$
$$\beta = (1-\mu)\left(\frac{\mu(1-\mu)}{\sigma^2} - 1\right)$$

- You can also specify two percentiles and use software to find α and β. For example, you could specify the endpoints of a prior 98% credible interval.

Example 8.5. Suppose we want to estimate θ, the proportion of Cal Poly students that are left-handed.

1. Sketch your Beta prior distribution for θ. Describe its main features and your reasoning. Then translate your prior into a Beta distribution by specifying the shape parameters α and β.
2. Assume a prior Beta distribution for θ with prior mean 0.15 and prior SD is 0.08. Find α and β, and a prior 98% credible interval for θ.

Solution. to Example 8.5

1. Of course, choices will vary, based on what you know about left-handedness. But do think about what your prior might look like, and use one of the methods to translate it to a Beta distribution.
2. Let's say we've heard that about 15% of people in general are left-handed, but we've also heard 10% so we're not super sure, and we also don't know how Cal Poly students compare to the general population. So we'll assume a prior Beta distribution for θ with prior mean 0.15 (our "best guess") and a prior SD of 0.08 to reflect our degree of uncertainty. This translates to a Beta(2.8, 16.1) prior, with a central 98% prior credible interval for θ that between 2.2% and 38.1% of Cal Poly students are left-handed. We could probably go with more prior certainty than this, but it seems at least like a reasonable starting place before observing data. We can (and should) use prior predictive tuning to aid in choosing α and β for our Beta distribution prior.

```
mu = 0.15
sigma = 0.08

alpha = mu ^ 2 * ((1 - mu) / sigma ^ 2 - 1 / mu); alpha
```

```
## [1] 2.838
```

```
beta <- alpha * (1 / mu - 1); beta
```

```
## [1] 16.08
```

```
qbeta(c(0.01, 0.99), alpha, beta)
```

```
## [1] 0.02222 0.38104
```

Chapter 9

Considering Prior Distributions

One of the most commonly asked questions when one first encounters Bayesian statistics is "how do we choose a prior?" While there is never one "perfect" prior in any situation, we'll discuss in this chapter some issues to consider when choosing a prior. But first, here are a few big picture ideas to keep in mind.

- Bayesian inference is based on the *posterior* distribution, not the prior. Therefore, the posterior requires much more attention than the prior.
- The prior is only one part of the Bayesian model. The likelihood is the other part. And there is the data that is used to fit the model. Choice of prior is just one of many modeling assumptions that should be evaluated and checked.
- In many situations, the posterior distribution is not too sensitive to reasonable changes in prior. In these situations, the important question isn't "what is the prior?" but rather "is there a prior at all"? That is, are you adopting a Bayesian approach, treating parameters as random variables, and quantifying uncertainty about parameters with probability distributions?
- One criticism of Bayesian statistics in general and priors in particular is that they are subjective. However, any statistical analysis is inherently subjective, filled with many assumptions and decisions along the way. Except in the simplest situations, if you ask five statisticians how to approach a particular problem, you will likely get five different answers. Priors and Bayesian data analysis are no more inherently subjective than any of the myriad other assumptions made in statistical analysis.

Subjectivity is OK, and often beneficial. Choosing a subjective prior allows us to explicitly incorporate a wealth of past experience into our analysis.

Example 9.1. Xiomara claims that she can predict which way a coin flip will land. Rogelio claims that he can taste the difference between Coke and Pepsi.

Before reading further, stop to consider: whose claim - Xiomara's or Rogelio's - is initially more convincing? Or are you equally convinced? Why? To put it another way, whose claim are you initially more skeptical of? Or are you equally skeptical? To put it one more way, whose claim would require more data to convince you?[1]

To test Xiomara's claim, you flip a fair coin 10 times, and she correctly predicts the result of 9 of the 10 flips. (You can assume the coin is fair, the flips are independent, and there is no funny business in data collection.)

To test Rogelio's claim, you give him a blind taste test of 10 cups, flipping a coin for each cup to determine whether to serve Coke or Pespi. Rogelio correctly identifies 9 of the 10 cups. (You can assume the coin is fair, the flips are independent, and there is no funny business in data collection.)

Let θ_X be the probability that Xiomara correctly guesses the result of a fair coin flip. Let θ_R be the probability that Rogelio correctly guesses the soda (Coke or Pepsi) in a randomly selected cup.

1. How might a frequentist address this situation? What would the conclusion be?
2. Consider a Bayesian approach. Describe, in general terms, your prior distributions for the two parameters. How do they compare? How would this impact your conclusions?

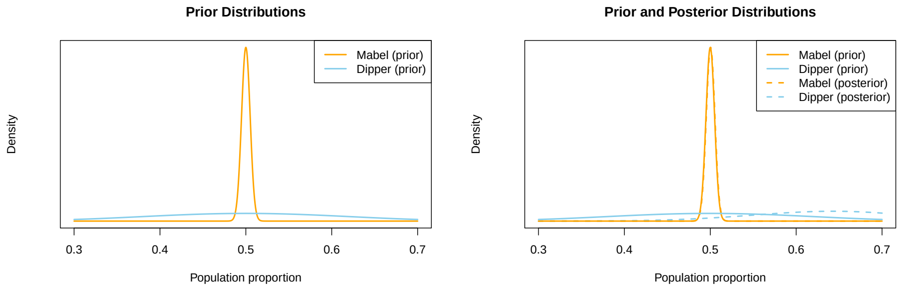
Solution. to Example 9.1

1. For Xiomara, a frequentist might conduct a hypothesis test of the null hypothesis $H_0 : \theta_X = 0.5$ versus the alternative hypothesis: $H_a : \theta_X > 0.5$. The p-value would be about 0.01, the probability of observing at least 9 out of 10 successes from a Binomial distribution with parameters 10 and 0.5 (`1 - pbinom(8, 10, 0.5)`). Rogelio's set up would be similar and would yield the same p-value. So a strict frequentist would be equally convinced of the two claims.
2. Prior to observing data, we are probably more skeptical of Xiomara's claim than Rogelio's. Since coin flips are unpredictable, we would have a strong prior belief that θ_X is close to 0.5 (what it would be if she were just guessing). Our prior for θ_X would have a mean of 0.5 and a small prior SD, to reflect that only values close to 0.5 seem plausible. Therefore, it would require a lot of evidence to sway our prior beliefs.

   On the other hand, we might be familiar with people who can tell the difference between Coke and Pepsi; maybe we even can ourselves. Our prior for θ_R would have a smaller prior SD than that of θ_X to allow for

[1] This example is motivated by an example in Section 1.1 of Dogucu et al. (2022).

a wider range of plausible values. We might even have a prior mean for θ_R above 0.5 if we have experience with a lot of people who can tell the difference between Coke and Pepsi. Given the sample data, our posterior probability that $\theta_R > 0.5$ would be larger than the posterior probability that $\theta_X > 0.5$, and we would be more convinced by Rogelio's claim than by Xiomara's.



Even if a prior does not represent strong prior beliefs, just having a prior distribution at all allows for Bayesian analysis. Remember, both Bayesian and frequentist are valid approaches to statistical analyses, each with advantages and disadvantages. That said, there are some issues with frequentist approaches that incorporating a prior distribution and adopting a Bayesian approach alleviates. (To be fair, an upcoming investigation will address some disadvantages of the Bayesian approach compared with the frequentist approach.)

Example 9.2. Tamika is a basketball player who throughout her career has had a probability of 0.5 of making any three point attempt. However, her coach is afraid that her three point shooting has gotten worse. To check this, the coach has Tamika shoot a series of three pointers; she makes 7 out of 24. Does the coach have evidence that Tamika has gotten worse?

Let θ be the probability that Tamika successfully makes any three point attempt. Assume attempts are independent.

1. Prior to collecting data, the coach decides that he'll have convincing evidence that Tamika has gotten worse if the p-value is less than 0.025. Suppose the coach told Tamika to *shoot 24 attempts and then stop* and count the number of successful attempts. Use software to compute the p-value. Is the coach convinced that Tamika has gotten worse?
2. Prior to collecting data, the coach decides that he'll have convincing evidence that Tamika has gotten worse if the p-value is less than 0.025. Suppose the coach told Tamika to *shoot until she makes 7 three pointers and then stop* and count the number of total attempts. Use software to compute the p-value. Is the coach convinced that Tamika has gotten worse? (Hint: the total number of attempts has a Negative Binomial distribution.)

3. Now suppose the coach takes a Bayesian approach and assumes a Beta(α, β) prior distribution for θ. Suppose the coach told Tamika to *shoot 24 attempts and then stop* and count the number of successful attempts. Identify the likelihood function and the posterior distribution of θ.
4. Now suppose the coach takes a Bayesian approach and assumes a Beta(α, β) prior distribution for θ. Suppose the coach told Tamika to *shoot until she makes 7 three pointers and then stop* and count the number of total attempts. Identify the likelihood function and the posterior distribution of θ.
5. Compare the Bayesian and frequentist approaches in this example. Does the "strength of the evidence" depend on how the data were collected?

Solution. to Example 9.2

1. The null hypothesis is $H_0 : \theta = 0.5$ and the alternative hypothesis is $H_a : \theta < 0.5$. If the null hypothesis is true and Tamika has not gotten worse, then Y, the number of successful attempts, has a Binomial(24, 0.5) distribution. The p-value is $P(Y \leq 7) = 0.032$ from `pbinom(7, 24, 0.5)`. Using a strict threshold of 0.025, the coach has NOT been convinced that Tamika has gotten worse.
2. The null hypothesis is $H_0 : \theta = 0.5$ and the alternative hypothesis is $H_a : \theta < 0.5$. If the null hypothesis is true and Tamika has not gotten worse, then N, the number of total attempts required to achieve 7 successful attempts, has a Negative Binomial(7, 0.5) distribution. The p-value is $P(N \geq 24) = 0.017$ from `1 - pnbinom(23 - 7, 7, 0.5)`. (In R, `nbinom` only counts the total number of failures, not the total number of trials.) Using a strict threshold of 0.025, the coach has been convinced that Tamika has gotten worse.
3. The data is Y, the number of successful attempts in 24 attempts, which follows a Binomial(24, θ) distribution. The likelihood is $P(Y = 7|\theta)$

$$f(y = 7|\theta) = \binom{24}{7}\theta^7(1-\theta)^{17} \propto \theta^7(1-\theta)^{17}, \qquad 0 < \theta < 1.$$

   The posterior distribution is the Beta($\alpha + 7$, $\beta + 17$) distribution.
4. The data is N, the number of total attempts required to achieve 7 successful attempts, which follows a Negative Binomial(7, θ) distribution. The likelihood is $P(N = 24|\theta)$

$$f(n = 24|\theta) = \binom{24-1}{7-1}\theta^7(1-\theta)^{17} \propto \theta^7(1-\theta)^{17}, \qquad 0 < \theta < 1.$$

   (The $\binom{24-1}{7-1}$ follows from the fact that the last attempt has to be success.) Note that the shape of the likelihood as a function of θ is the same as in the previous part. Therefore, the posterior distribution is the Beta($\alpha + 7$, $\beta + 17$) distribution.

5. Even though both frequentist scenario involves 7 successes in 24 attempts, the p-value measuring the strength of the evidence to reject the null hypothesis differed depending on how the data were collected. Using a strict cutoff of 0.025 led the coach to reject the null hypothesis in one scenario but not the other. However, the Bayesian analysis is the same in either scenario since the posterior distributions were the same. For the Bayesian analysis, all that mattered about the data was that there were 7 successes in 24 attempts.

Bayesian data analysis treats parameters as random variables with probability distributions. The prior distribution quantifies the researcher's uncertainty about parameters *before* observing data. Some issues to consider when choosing a prior include, in no particular order:

- The researcher's prior beliefs! A prior distribution is part of a statistical model, and should be consistent with knowledge about the underlying scientific problem. Researchers are often experts with a wealth of past experience that can be explicitly incorporated into the analysis via the prior distribution. Such a prior is called an informative or weakly informative prior.
- A regularizing prior. A prior which, when tuned properly, reduces overfitting or "overreacting" to the data.
- Noninformative prior a.k.a., (reference, vague, flat prior). A prior is sought that plays a minimal role in inference so that "the data can speak for itself".
- Mathematical convenience. The prior is chosen so that computation of the posterior is simplified, as in the case of conjugate priors.
- Interpretation. The posterior is a compromise between the data and prior. Some priors allow for easy interpretation of the relative contributions of data and prior to the posterior. For example, think of the "prior successes and prior failures" interpretation in the Beta-Binomial model.
- Prior based on *past* data. Bayesian updating can be viewed as an iterative process. The posterior distribution obtained from one round of data collection can inform the prior distribution for another round.

For those initially skeptical of prior distributions at all, the strategy of always choosing an noninformative or flat prior might be appealing. Flat priors are common, but are rarely ever the best choices from a modeling perspective. Just like you would not want to assume a Normal distribution for the likelihood in every problem, you would not to use a flat prior in every problem.

Furthermore, there are some subtle issues that arise when attempting to choose a noninformative prior.

Example 9.3. Suppose we want to estimate θ, the population proportion of Cal Poly students who wore socks at any point yesterday.

1. What are the possible values for θ? What prior distribution might you consider a noninformative prior distribution?
2. You might choose a Uniform(0, 1) prior, a.k.a., a Beta(1, 1) prior. Recall how we interpreted the parameters α and β in the Beta-Binomial model. Does the Beta(1, 1) distribution represent "no prior information"?
3. Suppose in a sample of 20 students, 4 wore socks yesterday. How would you estimate θ with a single number based only on the data?
4. Assume a Beta(1, 1) prior and the 4/20 sample data. Identify the posterior distribution. Recall that one Bayesian point estimate of θ is the posterior mean. Find the posterior mean of θ. Does this estimate let the "data speak entirely for itself"?
5. How could you change α and β in the Beta distribution prior to represent no prior information? Sketch the prior. Do you see any potential problems?
6. Assume a Beta(0, 0) prior for θ and the 4/20 sample data. Identify the posterior distribution. Find the posterior *mode* of θ. Does this estimate let the "data speak entirely for itself"?
7. Now suppose the parameter you want to estimate is the *odds* that a student wore socks yesterday, $\phi = \frac{\theta}{1-\theta}$. What are the possible values of ϕ? What might a non-informative prior look like? Is this a proper prior?
8. Assume a Beta(1, 1) prior for θ. Use simulation to approximate the prior distribution of the odds ϕ. Would you say this is a noninformative prior for ϕ?

Solution. to Example 9.3

1. θ takes values in (0, 1). We might assume a flat prior on (0, 1), that is a Uniform(0, 1) prior.

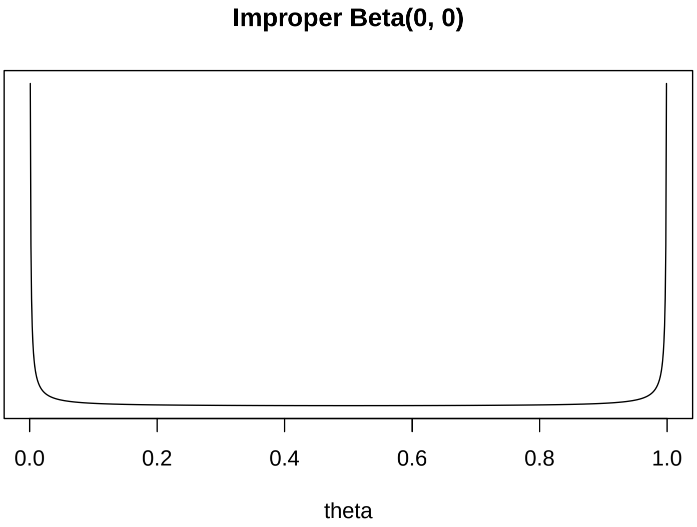
2. We interpreted α as "prior successes" and β as "prior failures". So a Beta(1, 1) is in some some equivalent to a "prior sample size" of 2. Certainly not a lot of prior information, but it's not "no prior information" either.

3. The sample proportion, $4/20 = 0.2$.

4. With a Beta(1, 1) prior and the 4/20 sample data, the posterior distribution is Beta(5, 17). The posterior mean of θ is $5/22 = 0.227$. The posterior mean is a weighted average of the prior mean and the sample proportion: $0.227 = (0.5)(2/22) + (0.2)(20/22)$. The "noninformative" prior does have influence; the data does not "speak entirely for itself".

5. If $\alpha + \beta$ represents "prior sample size", we could try a Beta(0, 0) prior. Unfortunately, such a probability distribution does not actually exist. For a Beta distribution, the parameters α and β have to be strictly positive in order to have a valid pdf. The Beta(0, 0) density would be proportional to
$$\pi(\theta) \propto \theta^{-1}(1-\theta)^{-1}, \qquad 0 < \theta < 1.$$

However, this is not a valid pdf since $\int_0^1 \theta^{-1}(1-\theta)^{-1} d\theta = \infty$, so there is no constant that can normalize it to integrate to 1. Even so, here is a plot of the "density".



Would you say this is a "noninformative" prior? It seems to concentrate almost all prior "density" near 0 and 1.

6. Beta(0, 0) is an "improper" prior. It's not a proper prior distribution, but it can lead to a proper posterior distribution. The likelihood is $f(y = 4|\theta) \propto \theta^4(1-\theta)^{16}, 0 < \theta < 1$. If we assume the prior is $\pi(\theta) \propto \theta^{-1}(1-\theta)^{-1}, 0 < \theta < 1$, then the posterior is

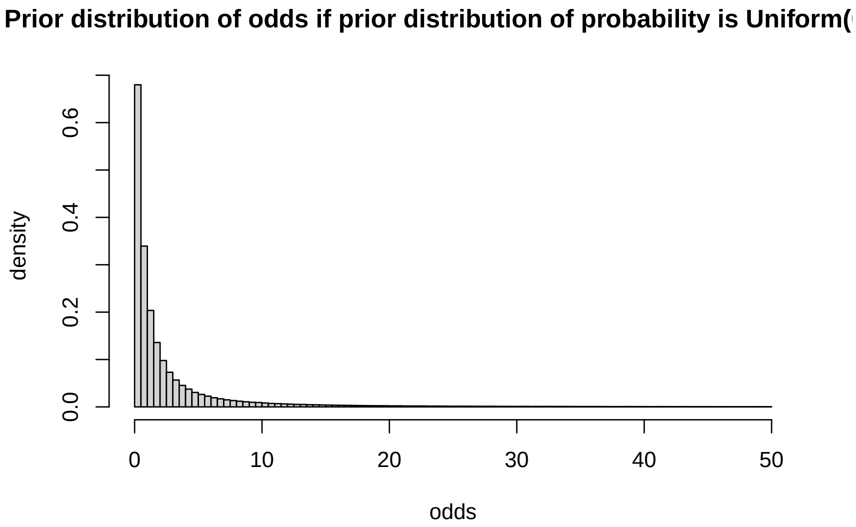
$$\pi(\theta|y=4) \propto \left(\theta^{-1}(1-\theta)^{-1}\right)\left(\theta^4(1-\theta)^{16}\right) = \theta^{4-1}(1-\theta)^{16-1}, \qquad 0 < \theta < 1$$

That is, the posterior distribution is the Beta(4, 16) distribution. The posterior mean is 4/20=0.2, the sample proportion. Hoever, the posterior *mode* is $\frac{4-1}{4+16-2} = \frac{3}{18} = 0.167$. So the posterior mode does not let the "data speak entirely for itself".

7. If $\theta = 0$ then $\phi = 0$; if $\theta = 1$ then $\phi = \infty$. So ϕ takes values in $(0, \infty)$. We might choose a flat prior on $(0, \infty)$, $\pi(\phi) \propto 1, \phi > 0$. However, this would be an improper prior.

8. Simulate a value of θ from a Beta(1, 1) distribution, compute $\phi = \frac{\theta}{1-\theta}$, and repeat many times. The simulation results are below. (The distribution is extremely skewed to the right, so we're only plotting values in (0, 50).)

```
theta = rbeta(1000000, 1, 1)
odds = theta / (1 - theta)
hist(odds[odds<50], breaks = 100, xlab = "odds", freq = FALSE,
 ylab = "density",
 main = "Prior distribution of odds if prior distribution of probability is Unifor
```



Even though the prior for θ was flat, the prior for a transformation of θ is not.

An *improper* prior distribution is a prior distribution that does not integrate to 1, so is not a proper probability density. However, an improper proper often results in a proper posterior distribution. Thus, improper prior distributions are sometimes used in practice.

Flat priors are common choices in some situations, but are rarely ever the best choices from a modeling perspective. Furthermore, flat priors are generally not preserved under transformations of parameters. So a prior that is flat under one parametrization of the problem will generally not be flat under another. For example, when trying to estimate a population SD σ, assuming a flat prior for σ will result in a non-flat prior for the population variance σ^2, and vice versa.

Example 9.4. Suppse that θ represents the population proportion of adults who have a particular rare disease.

1. Explain why you might not want to use a flat Uniform(0, 1) prior for θ.
2. Assume a Uniform(0, 1) prior. Suppose you will test $n = 100$ suspected cases. Use simulation to approximate the prior predictive distribution of

the number in the sample who have the disease. Does this seem reasonable?

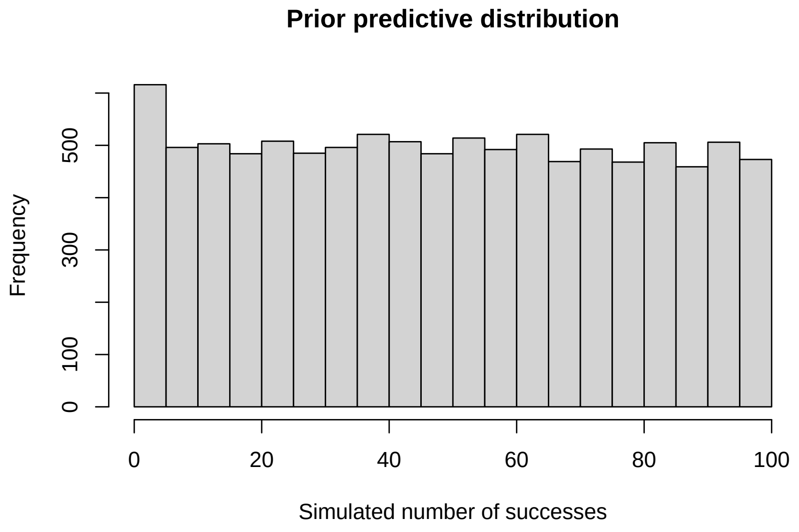
3. Assume a Uniform(0, 1) prior. Suppose that in $n = 100$ suspected cases, none actually has the disease. Find and interpret the posterior median. Does this seem reasonable?

Solution. to Example 9.4

1. We know it's a rare disease! We want to concentrate most of our prior probability for θ near 0.

2. If the disease is rare, we might not expect any actual cases in a sample of 100, maybe 1 or 2. However, the prior predictive distribution says that any value between 0 and 100 actual cases is equally likely! This seems very unreasonable given that the disease is rare.

```
theta_sim = runif(10000)
y_sim = rbinom(10000, 100, theta_sim)
hist(y_sim,
     xlab = "Simulated number of successes",
     main = "Prior predictive distribution")
```



3. The posterior distribution is the Beta(1, 101) distribution. The posterior median is 0.007 (`qbeta(0.5, 1, 101)`). Based on a sample of 100 suspected cases with no actual cases, there is a posterior probability of 50% that more than 0.7% of people have the disease. A rate of 7 actual cases in 1000 is not a very rare disease, and we think there's a 50% chance that the

> rate is even greater than this? Again, this does not seem very reasonable based on our knowledge that the disease is rare.

Prior predictive distributions can be used to check the reasonableness of a prior for a given situation before observing sample data. Do the simulated samples seem consistent with what you might expect of the data based on your background knowledge of the situation? If not, another prior might be more reasonable.

9.1 What NOT to do when considering priors

You have a great deal of flexibility in choosing a prior, and there are many reasonable approaches. However, there are a few things that you should NOT do.

Do NOT choose a prior that assigns 0 probability/density to *possible* values of the parameter regardless of how initially implausible the values are. Even very stubborn priors can be overturned with enough data, but no amount of data can turn a prior probability of 0 into a positive posterior probability. Always consider the range of possible values of the parameter, and be sure the prior density is non-zero over that range of values.

Do NOT base the *prior* on the observed data. The prior reflects the degree of uncertainty about parameters *before* observing data. Adjusting the *prior* to reflect observed data to achieve some desired result is akin to "data snooping" or "p-hacking" and is bad statistics. (Of course, the *posterior* is based on the observed data. But not the prior.)

Do NOT feel like you have to find that one, perfect prior. The prior is just one assumption of the model and should be considered just like other assumptions. In practice, no assumption of a statistical model is ever satisfied exactly. We only hope that our set of assumptions provides a reasonable model for reality. No one prior will ever be just right for a situation, but some might be more reasonable than others. You are not only allowed but encouraged to try different priors to see how sensitive the results are to the choice of prior. (Remember, you should check the other assumptions too!) There is also no requirement that you have to choose a single prior. It's possible to consider several models, each consisting of its own prior, and average over these models. (We'll see a little more detail about model averaging later.)

Do NOT worry too much about the prior! In general, in Bayesian estimation the larger the sample size the smaller the role that the prior plays. But it is often desirable for the prior to play some role. You should not feel the need to apologize for priors when significant prior knowledge is available.

Chapter 10

Introduction to Posterior Simulation and JAGS

In the Beta-Binomial model there is a simple expression for the posterior distribution. However, in most problems it is not possible to find the posterior distribution analytically, and therefore we must approximate it.

Example 10.1. Consider Example 8.5 again, in which we wanted to estimate the proportion of Cal Poly students that are left-handed. In that example we specifed a prior by first specifying a prior mean of 0.15 and a prior SD of 0.08 and then we found the corresponding Beta prior. However, when dealing with means and SDs, it is natural — but by no means necessary — to work with Normal distributions. Suppose we want to assume a Normal distribution prior for θ with mean 0.15 and SD 0.08. Also suppose that in a sample of 25 Cal Poly students 5 are left-handed. We want to find the posterior distribution.

Note: the Normal distribution prior assigns positive (but small) density outside of (0, 1). So we can either truncate the prior to 0 outside of (0, 1) or just rely on the fact that the likelihood will be 0 for θ outside of (0, 1) to assign 0 posterior density outside (0, 1).

1. Write an expression for the shape of the posterior density. Is this a recognizable probability distribution?
2. We have seen one method for approximating a posterior distribution. How could you employ it here?

Solution. to Example 10.1

1. As always, the posterior density is proportional to the product of the prior

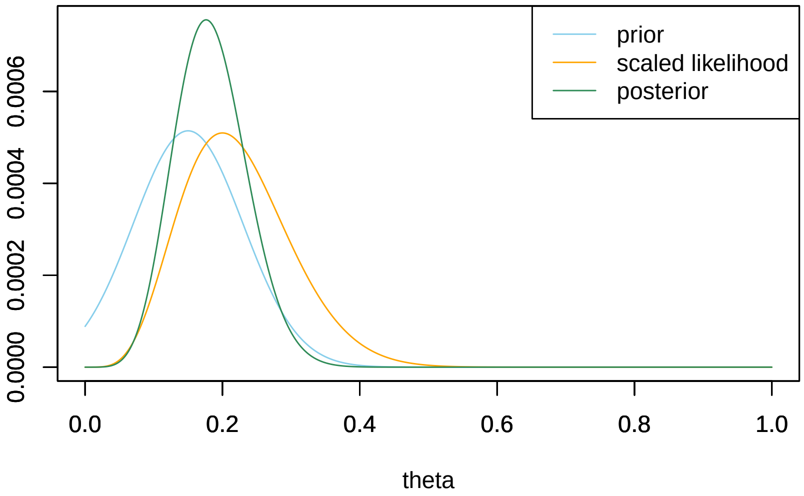
density and the likelihood function.

$$\begin{aligned}
\text{Prior:} \quad & \pi(\theta) \propto \frac{1}{0.08}\exp\left(-\frac{(\theta-0.15)^2}{2(0.08^2)}\right)\\
\text{Likelihood:} \quad & f(y|\theta) \propto \theta^5(1-\theta)^{20}\\
\text{Posterior:} \quad & \pi(\theta|y) \propto \left(\theta^5(1-\theta)^{20}\right)\left(\frac{1}{0.08}\exp\left(-\frac{(\theta-0.15)^2}{2(0.08^2)}\right)\right)
\end{aligned}$$

This is not a recognizable probability density.

2. We can use grid approximation and treat the continuous parameter θ as discrete.

```
theta = seq(0, 1, 0.0001)

# prior
prior = dnorm(theta, 0.15, 0.08)
prior = prior / sum(prior)

# data
n = 25 # sample size
y = 5 # sample count of success

# likelihood, using binomial
likelihood = dbinom(y, n, theta) # function of theta

# posterior
product = likelihood * prior
posterior = product / sum(product)

# plot
ylim = c(0, max(c(prior, posterior, likelihood / sum(likelihood))))
plot(theta, prior, type='l', xlim=c(0, 1), ylim=ylim, col="skyblue", xlab='theta', ylal
par(new=T)
plot(theta, likelihood / sum(likelihood), type='l', xlim=c(0, 1), ylim=ylim, col="orang
par(new=T)
plot(theta, posterior, type='l', xlim=c(0, 1), ylim=ylim, col="seagreen", xlab='', ylal
legend("topright", c("prior", "scaled likelihood", "posterior"), lty=1, col=c("skyblue'
```



Grid approximation is one method for approximating a posterior distribution. However, finding a sufficiently fine grid approximation suffers from the "curse of dimensionality" and does not work well in multi-parameter problems. For example, suppose you use a grid of 1000 points to approximate the distribution of any single parameter. Then you would need a grid of 1000^2 points to approximate the joint distribution of any two parameters, 1000^3 points for three parameters, and so on. The size of the grid increases exponentially with the number of parameters and becomes computationally infeasible in problems with more than a few parameters. (And later we'll see some examples that include *hundreds* of parameters.) Furthermore, if the posterior density changes very quickly over certain regions, then even finer grids might be needed to provide reliable approximations of the posterior in these regions. (Though if the posterior density is relative smooth over some regions, then we might be able to get away with a coarser grid in these regions.)

The most common way to approximate a posterior distribution is via simulation. The inputs to the simulation are

- Observed data y
- Model for the data, $f(y|\theta)$ which depends on parameters θ. (This model determines the likelihood function.)
- Prior distribution for parameters $\pi(\theta)$

We then employ some simulation algorithm to approximate the posterior distribution of θ given the observed data y, $\pi(\theta|y)$, *without computing the posterior distribution.*

Careful: we have already used simulation to approximate predictive distributions. Here we are primarily focusing on using simulation to approximate the posterior distribution of parameters.

Let's consider a discrete example first.

Example 10.2. Continuing the kissing study in Example 5.2 where θ can only take values 0.1, 0.3, 0.5, 0.7, 0.9. Consider a prior distribution which places probability 1/9, 2/9, 3/9, 2/9, 1/9 on the values 0.1, 0.3, 0.5, 0.7, 0.9, respectively. Suppose that $y = 8$ couples in a sample of size $n = 12$ lean right.

1. Describe in detail how you could use simulation to approximate the posterior distribution of θ, without first computing the posterior distribution.
2. Code and run the simulation. Compare the simulation-based approximation to the true posterior distribution from Example 5.2.
3. How would the simulation/code change if θ had a Beta prior distribution, say Beta(3, 3)?
4. Suppose that $n = 1200$ and $y = 800$. What would be the problem with running the above simulation in this situation? Hint: compute the probability that Y equals 800 for a Binomial distribution with parameters 1200 and 0.667.

Solution. to Example 10.2

1. Remember that the posterior distribution is the *conditional distribution of parameters θ given the observed data y*. Therefore, we need to approximate the conditional distribution of θ given $y = 8$ successes in a sample of size $n = 12$.

    - Simulate a value of θ from the prior distribution.
    - Given θ, simulate a value of Y from the Binomial distribution with parameters $n = 12$ and θ.
    - Repeat the above steps many times, generating many (θ, Y) pairs.
    - To condition on $y = 8$, discard any (θ, Y) pairs for which Y is not 8. Summarize the θ values for the remaining pairs to approximate the posterior distribution of θ. For example, to approximate the posterior probability that θ equals 0.7, count the number of repetitions in which θ equals 0.7 and Y equals 8 and divide by the count of repetitions in which Y equals 8.

2. See code below. The simulation approximates the posterior distribution fairly well in this case. Notice that we simulate 100,000 (θ, Y) pairs, but only around 10,000 or so yield a value of Y equal to 8. Therefore, the posterior approximation is based on roughly 10,000 values, not 100,000.

```
n_sim = 100000

theta_prior_sim = sample(c(0.1, 0.3, 0.5, 0.7, 0.9),
                         size = n_sim,
                         replace = TRUE,
                         prob = c(1, 2, 3, 2, 1) / 9)

y_sim = rbinom(n_sim, 12, theta_prior_sim)

kable(head(data.frame(theta_prior_sim, y_sim), 20))
```

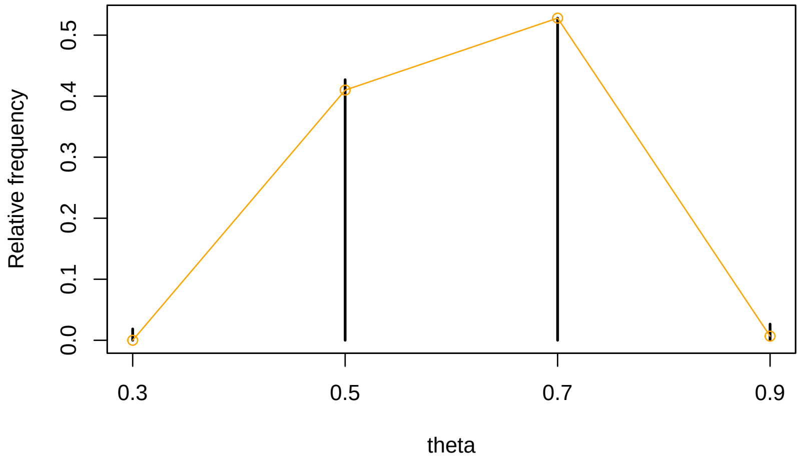
theta_prior_sim	y_sim
0.1	0
0.5	2
0.5	5
0.7	10
0.1	2
0.5	6
0.5	8
0.1	1
0.7	7
0.5	7
0.7	10
0.3	3
0.5	5
0.5	7
0.9	11
0.7	12
0.3	6
0.5	8
0.3	5
0.9	10

```
theta_post_sim = theta_prior_sim[y_sim == 8]

table(theta_post_sim)
```

```
## theta_post_sim
##  0.3  0.5  0.7  0.9
##  177 4070 5031  253
```

```
plot(table(theta_post_sim) / length(theta_post_sim),
 xlab = "theta",
 ylab = "Relative frequency")
```

```
# true posterior for comparison
par(new = T)
plot(c(0.3, 0.5, 0.7, 0.9), c(0.0181, 0.4207, 0.5365, 0.0247),
 col = "orange", type = "o",
 xaxt = 'n', yaxt = 'n', xlab = "", ylab = "")
```

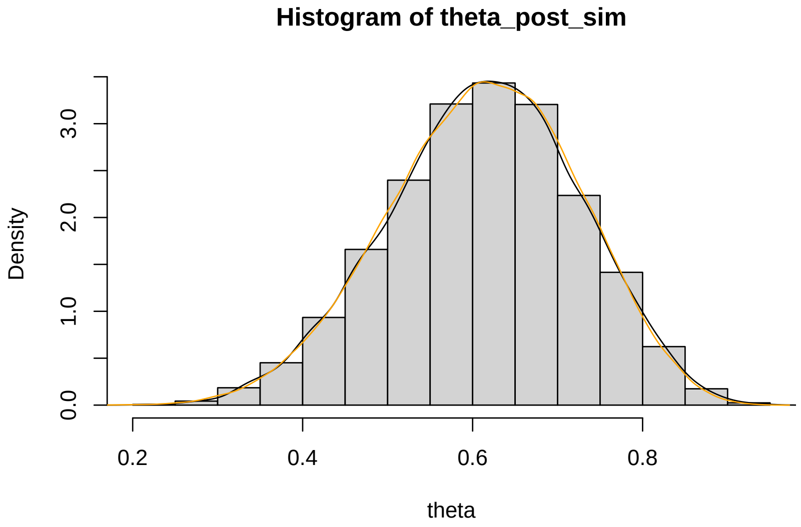


3. The only difference is that we would first simulate a value of θ from its Beta(3, 3) prior distribution (using `rbeta`). Now any value between 0 and 1 is a possible value of θ. But we would still approximate the posterior distribution by discarding any (θ, Y) pairs for which Y is not equal to 8. Since θ is continuous, we could summarize the simulated values with a histogram or density plot.

```
n_sim = 100000

theta_prior_sim = rbeta(n_sim, 3, 3)

y_sim = rbinom(n_sim, 12, theta_prior_sim)

kable(head(data.frame(theta_prior_sim, y_sim), 20))
```

theta_prior_sim	y_sim
0.4965	6
0.3985	6
0.4446	6
0.4233	7
0.4427	5
0.5821	6
0.1518	4
0.6136	6
0.7128	6
0.3245	5
0.7393	8
0.3007	4
0.9726	12
0.3001	3
0.2044	1
0.3401	2
0.6289	6
0.7696	11
0.7605	10
0.3823	4

```r
theta_post_sim = theta_prior_sim[y_sim == 8]

hist(theta_post_sim, freq = FALSE,
 xlab = "theta",
 ylab = "Density")
lines(density(theta_post_sim))

# true posterior for comparison
lines(density(rbeta(100000, 3 + 8, 3 + 4)), col = "orange")
```



4. Now we need to approximate the conditional distribution of θ given 800 successes in a sample of size $n = 1200$. The probability that Y equals 800 for a Binomial distribution with parameters 1200 and 2/3 is about 0.024 (`dbinom(800, 1200, 2 / 3)`). Since the sample proportion $800/1200 = 2/3$ maximizes the likelihood of $y = 800$, the probability is even smaller for the other values of θ.
   Therefore, if we generate 100,000 (θ, Y) pairs, only a few hundred or so of them would yield $y = 800$ and so the posterior approximation would be unreliable. If we wanted the posterior approximation to be based on 10,000 simulated values from the conditional distribution of θ given $y = 8$, we would first have to general about 10 million (θ, Y) pairs.

In principle, the posterior distribution $\pi(\theta|y)$ given observed data y can be found by

- simulating many θ values from the prior distribution
- simulating, for each simulated value of θ, a Y value from the corresponding conditional distribution of Y given θ (Y could be a sample or the value of a sample statistic)
- discarding (θ, Y) pairs for which the simulated Y value is not equal to the observed y value
- Summarizing the simulated θ values for the remaining pairs with $Y = y$.

However, this is a very computationally inefficient way of approximating the posterior distribution. Unless the sample size is really small, the simulated sample statistic Y will only match the observed y value in relatively few samples, simply because in large samples there are just many more possibilities. For example, in 1000 flips of a fair coin, the most likely value of the number of heads is 500, but the probability of *exactly* 500 heads in 1000 flips is only 0.025. When there are many possibilities, the probability gets stretched fairly thin. Therefore, if we want say 10000 simulated values of θ given y, we would have first simulate many, many more values.

The situation is even more extreme when the data is continuous, where the probably of replicating the observed sample is essentially 0.

Therefore, we need more efficient simulation algorithms for approximating posterior distributions. **Markov chain Monte Carlo (MCMC)** methods[1] provide powerful and widely applicable algorithms for simulating from probability distributions, including complex and high-dimensional distributions. These algorithms include Metropolis-Hastings, Gibbs sampling, Hamiltonian Monte Carlo, among others. We will see later some of the ideas behind MCMC algorithms. However, we will rely on software to carry out MCMC simulations.

10.1 Introduction to JAGS

JAGS[2] ("Just Another Gibbs Sampler") is a stand alone program for performing MCMC simulations. JAGS takes as input a Bayesian model description — prior plus likelihood — and data and returns an MCMC sample from the posterior distribution. JAGS uses a combination of Metropolis sampling, Gibbs sampling, and other MCMC algorithms.

A few JAGS resources:

- JAGS User Manual
- JAGS documentation
- Some notes about JAGS error messages
- *Doing Bayesian Data Analysis* textbook website

The basic steps of a JAGS program are:

1. Load the data
2. Define the model: likelihood and prior
3. Compile the model in JAGS
4. Simulate values from the posterior distribution
5. Summarize simulated values and check diagnostics

[1] For some history, and an origin of the use of "Monte Carlo", see Wikipedia.

[2] If you've ever heard of BUGS (or WinBUGS) JAGS is very similar but with a few nicer features.

This section introduces a brief introduction to JAGS in some relatively simple situations.

Using the `rjags` package, one can interact with JAGS entirely within R.

```
library(rjags)
```

10.1.1 Load the data

We'll use the "data is singular" context as an example. Compare the results of JAGS simulations to the results in Chapter 7.

The data could be loaded from a file, or specified via sufficient summary statistics. Here we'll just load the summary statistics and in later examples we'll show how to load individual values.

```
n = 35 # sample size
y = 31 # number of successes
```

10.1.2 Specify the model: likelihood and prior

A JAGS model specification starts with `model`. The model provides a *textual* description of likelihood and prior. This text string will then be passed to JAGS for translation.

Recall that for the Beta-Binomial model, the prior distribution is $\theta \sim \text{Beta}(\alpha, \beta)$ and the likelihood for the total number of successes Y in a sample of size n corresponds to $(Y|\theta) \sim \text{Binomial}(n, \theta)$. Notice how the following text reflects the model (prior & likelihood).

Note: JAGS syntax is similar to, but not the same, as R syntax. For example, compare `dbinom(y, n, theta)` in R versus `y ~ dbinom(theta, n)` in JAGS. See the JAGS user manual for more details. You can use comments with # in JAGS models, similar to R.

```
model_string <- "model{

  # Likelihood
  y ~ dbinom(theta, n)

  # Prior
  theta ~ dbeta(alpha, beta)
  alpha <- 3 # prior successes
  beta <- 1 # prior failures

}"
```

Again, the above is just a text string, which we'll pass to JAGS for translation.

10.1.3 Compile in JAGS

We pass the model (which is just a text string) and the data to JAGS to be compiled via `jags.model`. The model is defined by the text string via the `textConnection` function. The model can also be saved in a separate file, with the file name being passed to JAGS. The data is passed to JAGS in a list. In `dataList` below `y = y, n = n` maps the data defined by `y=31, n=35` to the terms `y, n` specified in the `model_string`.

```
dataList = list(y = y, n = n)

model <- jags.model(file = textConnection(model_string),
                    data = dataList)
```

```
## Compiling model graph
##    Resolving undeclared variables
##    Allocating nodes
## Graph information:
##    Observed stochastic nodes: 1
##    Unobserved stochastic nodes: 1
##    Total graph size: 5
##
## Initializing model
```

10.1.4 Simulate values from the posterior distribution

Simulating values in JAGS is completed in essentially two steps. The `update` command runs the simulation for a "burn-in" period. The `update` function merely "warms-up" the simulation, and the values sampled during the update phase are not recorded. (We will discuss "burn-in" in more detail later.)

```
update(model, n.iter = 1000)
```

After the update phase, we simulate values from the posterior distribution that we'll actually keep using `coda.samples`. Using `coda.samples` arranges the output in a format conducive to using `coda`, a package which contains helpful functions for summarizing and diagnosing MCMC simulations. The variables to record simulated values for are specified with the `variable.names` argument. Here there is only a single parameter theta, but we'll see multi-parameter examples later.

```
Nrep = 10000 # number of values to simulate

posterior_sample <- coda.samples(model,
                       variable.names = c("theta"),
                       n.iter = Nrep)
```

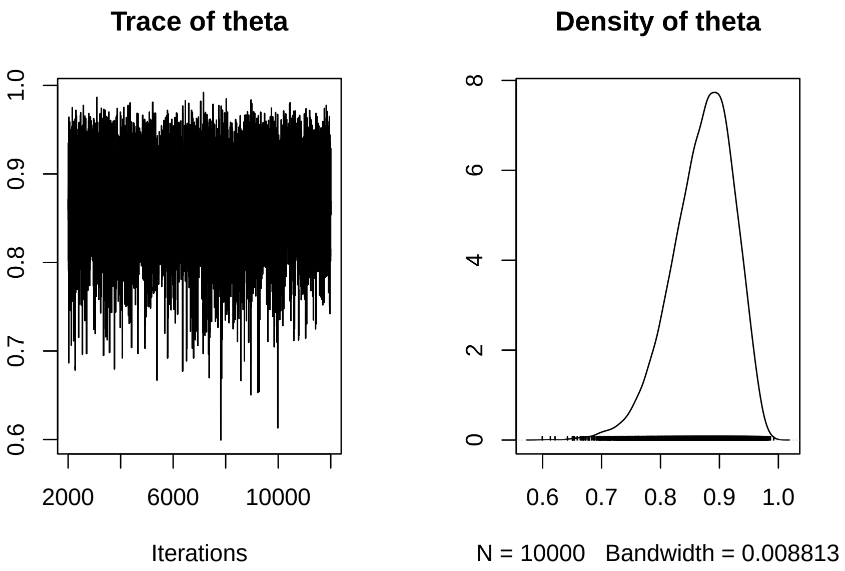
10.1.5 Summarizing simulated values and diagnostic checking

Standard R functions like `summary` and `plot` can be used to summarize results from `coda.samples`. We can summarize the simulated values of theta to approximate the posterior distribution.

```
summary(posterior_sample)
```

```
##
## Iterations = 2001:12000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##           Mean             SD       Naive SE Time-series SE
##       0.871382       0.052462       0.000525       0.000777
##
## 2. Quantiles for each variable:
##
##  2.5%   25%   50%   75% 97.5%
## 0.756 0.839 0.878 0.909 0.956
```

```
plot(posterior_sample)
```



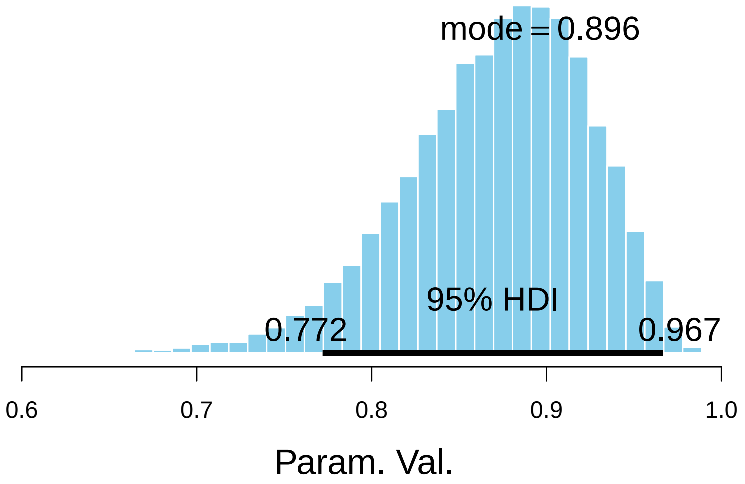
The **Doing Bayesian Data Analysis (DBDA2E)** textbook package also has some nice functions built in, in particular in the `DBD2AE-utilities.R` file.

For example, the `plotPost` functions creates an annotated plot of the posterior distribution along with some summary statistics. (See the DBDA2E documentation for additional arguments.)

```
source("DBDA2E-utilities.R")
```

```
##
## *********************************************************************
## Kruschke, J. K. (2015). Doing Bayesian Data Analysis, Second Edition:
## A Tutorial with R, JAGS, and Stan. Academic Press / Elsevier.
## *********************************************************************
```

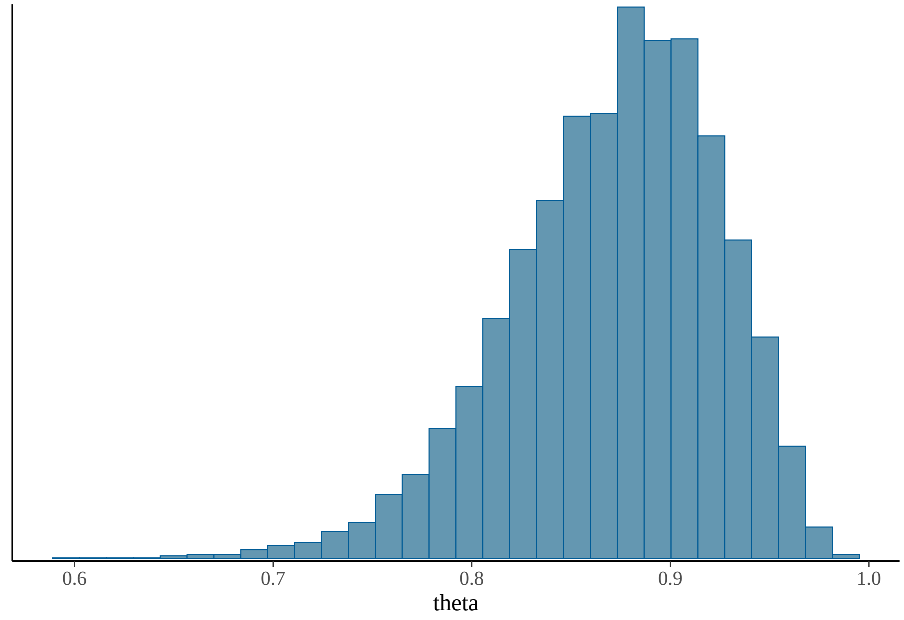
```
plotPost(posterior_sample)
```



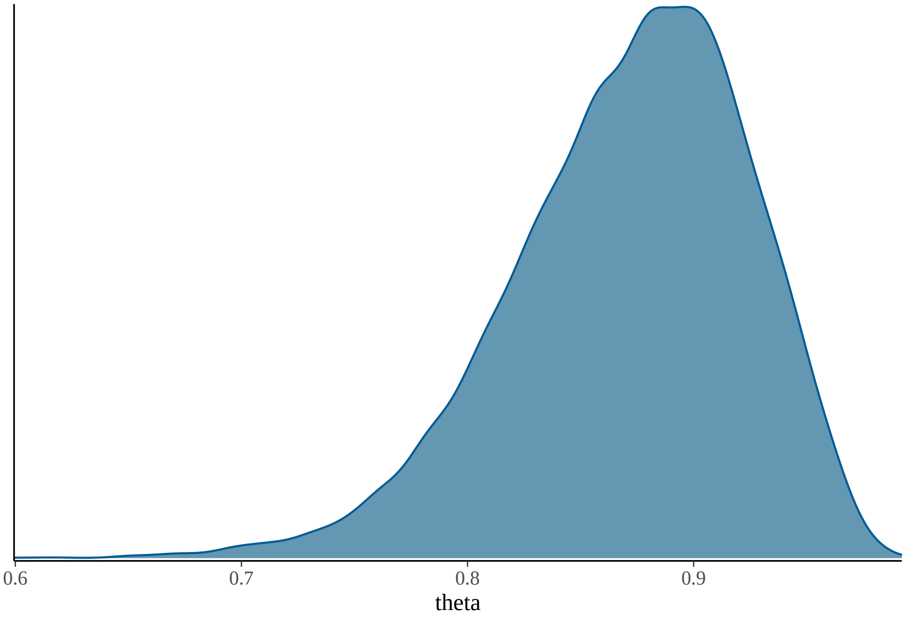
```
##                ESS   mean median   mode hdiMass hdiLow hdiHigh compVal pGtCompVal
## Param. Val. 4563 0.8714 0.8776 0.8964    0.95  0.772  0.9666      NA         NA
##             ROPElow ROPEhigh pLtROPE pInROPE pGtROPE
## Param. Val.      NA       NA      NA      NA      NA
```

The `bayesplot` R package also provides lots of nice plotting functionality.

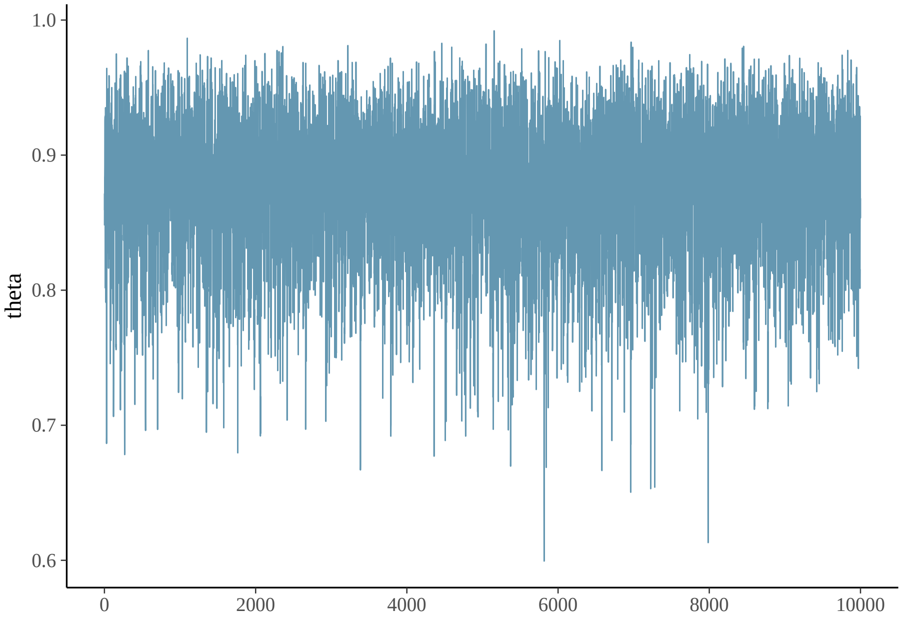
```
library(bayesplot)
mcmc_hist(posterior_sample)
```



```
mcmc_dens(posterior_sample)
```



```
mcmc_trace(posterior_sample)
```



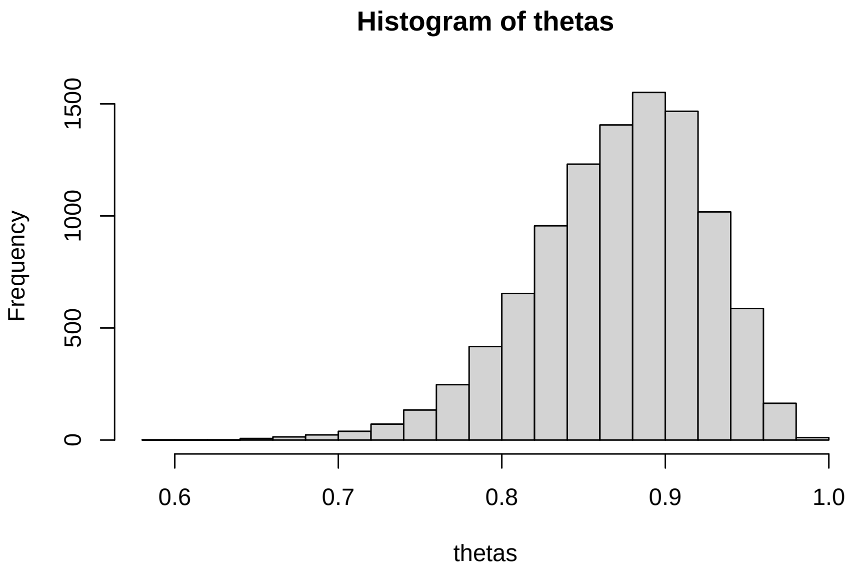
10.1.6 Posterior prediction

The output from `coda.samples` is stored as an mcmc.list format. The simulated values of the variables identified in the `variable.names` argument can be extracted as a matrix (or array) and then manipulated as usual R objects.

```
thetas = as.matrix(posterior_sample)
head(thetas)
```

```
##       theta
## [1,] 0.8479
## [2,] 0.8715
## [3,] 0.8592
## [4,] 0.8828
## [5,] 0.8846
## [6,] 0.8927
```

```
hist(thetas)
```

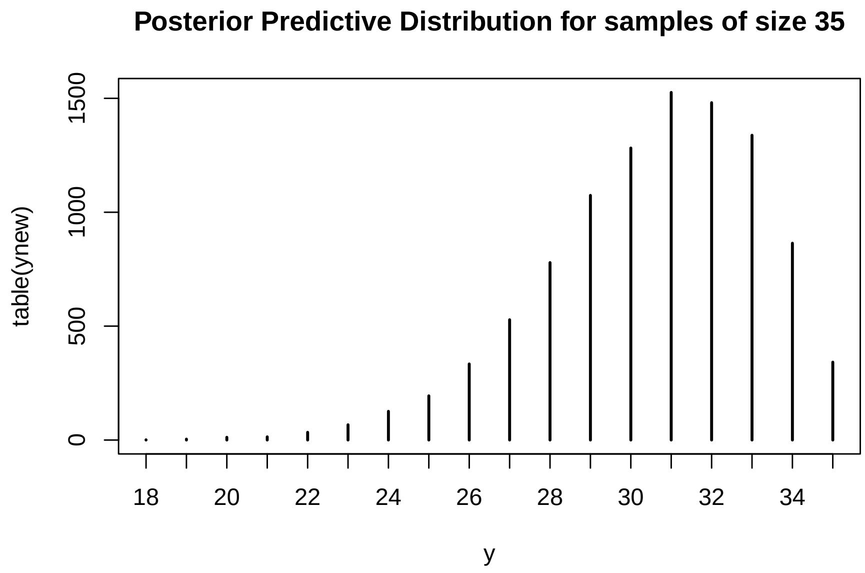


The matrix would have one column for each variable named in `variable.names`; in this case, there is only one column corresponding to the simulated values of theta.

We can now use the simulated values of theta to simulate replicated samples to approximate the posterior predictive distribution. To be clear, the code below is running R commands within R (not JAGS).

(There is a way to simulate predictive values within JAGS itself, but I think it's more straightforward in R. Just use JAGS to get a simulated sample from the posterior distribution. On the other hand, if you're using Stan there are functions for simulating and summarizing posterior predicted values.)

```
ynew = rbinom(Nrep, n, thetas)

plot(table(ynew),
     main = "Posterior Predictive Distribution for samples of size 35",
     xlab = "y")
```



10.1.7 Loading data as individual values rather than summary statistics

Instead of the total count (modeled by a Binomial likelihood), the individual data values ($1/0 = \mathrm{S/F}$) can be provided, which could be modeled by a Bernoulli (i.e. Binomial(trials=1)) likelihood. That is, $(Y_1, \dots, Y_n|\theta) \sim$ i.i.d. Bernoulli(θ), rather than $(Y|\theta) \sim$ Binomial(n, θ). The vector y below represents the data in this format. Notice how the likelihood in the model specification changes in response; the n observations are specified via a `for` loop.

```
# Load the data
y = c(rep(1, 31), rep(0, 4)) # vector of 31 1s and 4 0s
n = length(y)

model_string <- "model{

  # Likelihood
  for (i in 1:n){
    y[i] ~ dbern(theta)
  }

  # Prior
  theta ~ dbeta(alpha, beta)
  alpha <- 3
  beta <- 1

}"
```

10.1.8 Simulating multiple chains

The Bernoulli model can be passed to JAGS similar to the Binomial model above. Below, we have also introduced the `n.chains` argument, which simulates multiple Markov chains and allows for some additional diagnostic checks. Simulating multiple chains helps assess convergence of the Markov chain to the target distribution. (We'll discuss more details later.) Initial values for the chains can be provided in a list with the `inits` argument; otherwise initial values are generated automatically.

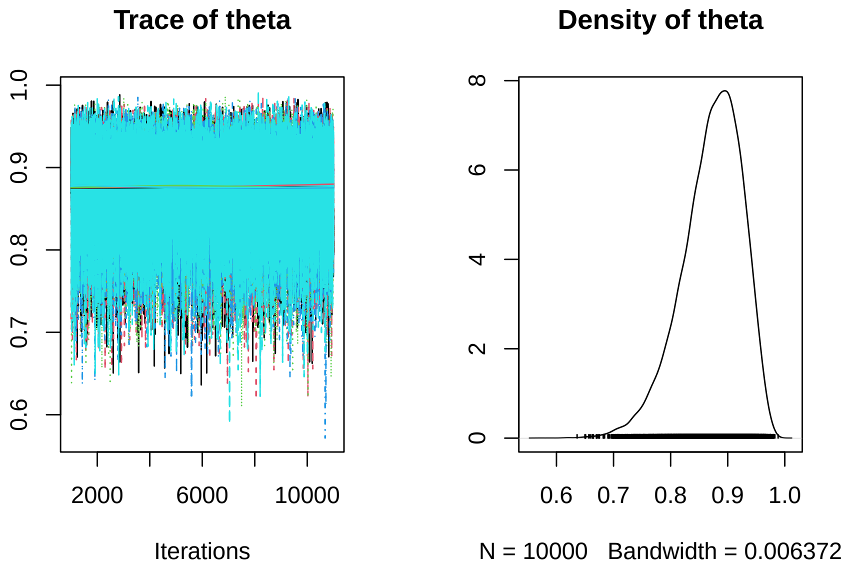
```
# Compile the model
dataList = list(y = y, n = n)

model <- jags.model(textConnection(model_string),
                    data = dataList,
                    n.chains = 5)
```

```
## Compiling model graph
##    Resolving undeclared variables
##    Allocating nodes
## Graph information:
##    Observed stochastic nodes: 35
##    Unobserved stochastic nodes: 1
##    Total graph size: 39
##
## Initializing model
```

```
# Simulate
update(model, 1000, progress.bar = "none")

Nrep = 10000

posterior_sample <- coda.samples(model,
                                 variable.names = c("theta"),
                                 n.iter = Nrep,
                                 progress.bar = "none")

# Summarize and check diagnostics
summary(posterior_sample)
```

```
##
## Iterations = 1001:11000
## Thinning interval = 1
## Number of chains = 5
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##           Mean             SD       Naive SE Time-series SE
##       0.872245       0.052547       0.000235       0.000238
##
## 2. Quantiles for each variable:
##
##  2.5%   25%   50%   75% 97.5%
## 0.753 0.841 0.878 0.911 0.956
```

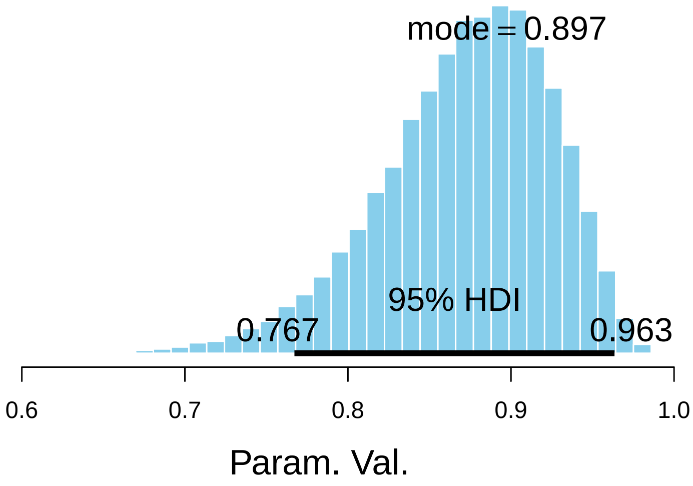
```
plot(posterior_sample)
```



If multiple chains are simulated, the DBDA2E function `diagMCMC` can be used for diagnostics.

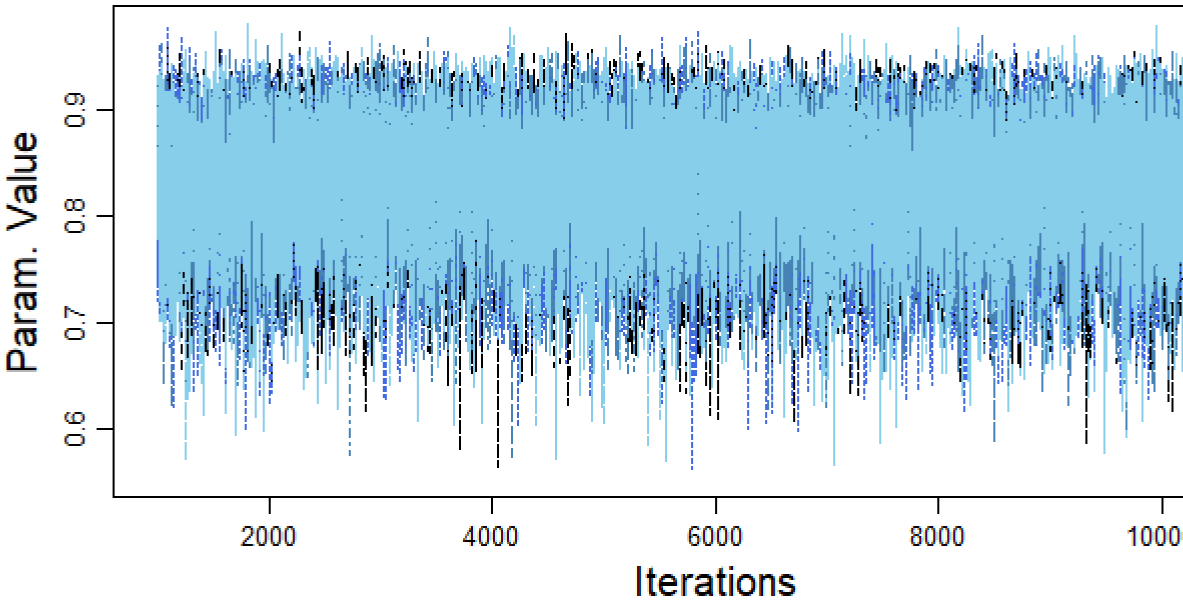
Note: Some of the DBDA2E output, in particular from `diagMCMC`, isn't always displayed when the RMarkdown file is knit. You might need to manually run these cells within RStudio. I'm not sure why; please let me know if you figure it out.

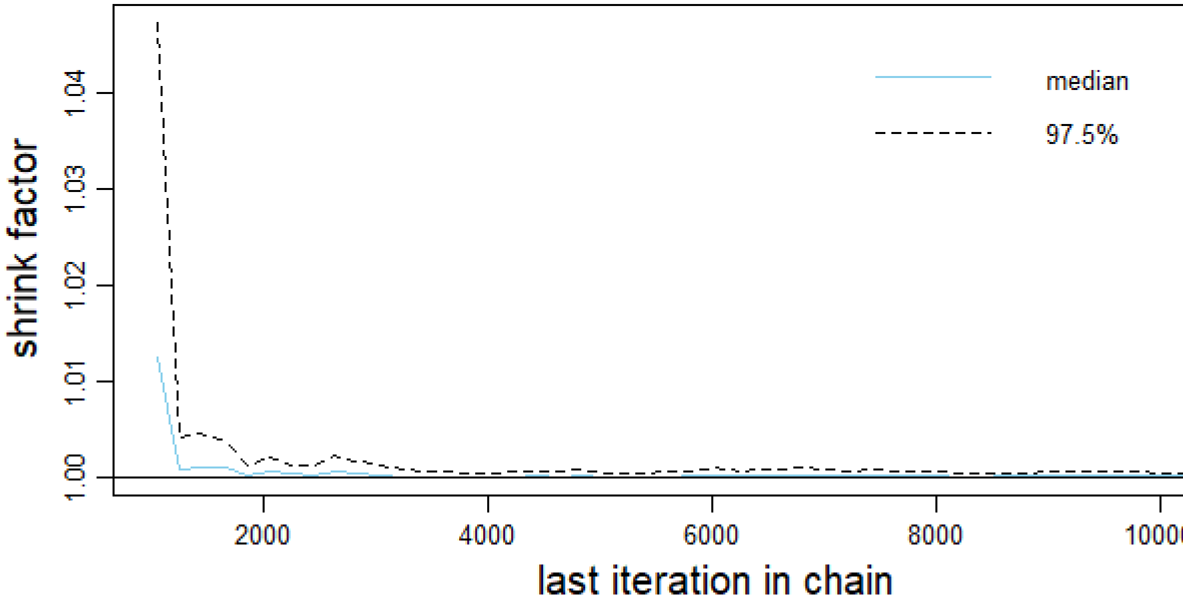
```
plotPost(posterior_sample)
```



```
##                ESS   mean median   mode hdiMass hdiLow hdiHigh compVal
## Param. Val. 50000 0.8722 0.8782 0.8974    0.95 0.7672  0.9635      NA
##             pGtCompVal ROPElow ROPEhigh pLtROPE pInROPE pGtROPE
## Param. Val.         NA      NA       NA      NA      NA      NA
```

```
diagMCMC(posterior_sample)
```


Param. Value
0.6
0.7
0.8
0.9
2000
4000
6000
8000
1000
Iterations


shrink factor
1.00
1.01
1.02
1.03
1.04
median
97.5%
2000
4000
6000
8000
1000
last iteration in chain

10.1.9 ShinyStan

We can use regular R functionality for plotting, or functions from packages likes DBDA2E or bayesplot. Another nice tool is ShinyStan, which provides an interactive utility for exploring the results of MCMC simulations. While ShinyStan was developed for the Stan package, it can use output from JAGS and other MCMC packages. You'll need to install the `shinystan` package and its dependencies.

The code below will launch in a browser the ShinyStan GUI for exploring the results of the JAGS simulation. The `as.shinystan` command takes coda.samples output (stored as an mcmc-list) and puts it in the proper format for ShinyStan. (Note: this code won't display anything in the notes. You'll have to actually run it to see what happens.)

```
library(shinystan)
my_sso <- launch_shinystan(as.shinystan(posterior_sample,
                                        model_name = "Bortles!!!"))
```

10.1.10 Back to the left-handed problem

Let's return again to the problem in Example 10.1, in which we wanted to estimate the proportion of Cal Poly students that are left-handed. Assume a Normal distribution prior for θ with mean 0.15 and SD 0.08. Also suppose that in a sample of 25 Cal Poly students 5 are left-handed. We will use JAGS to find the (approximate) posterior distribution.

Important note: in JAGS a Normal distribution is parametrized by its *precision*, which is the reciprocal of the variance: `dnorm(mean, precision)`. That is, for a $N(\mu, \sigma)$ distribution, the precision, often denoted τ, is $\tau = 1/\sigma^2$. For example, in JAGS `dnorm(0, 1 / 4)` corresponds to a precision of 1/4, a variance of 4, and a standard deviation of 2.

```
# Data
n = 25
y = 5

# Model
model_string <- "model{

  # Likelihood
  y ~ dbinom(theta, n)

  # Prior
  theta ~ dnorm(mu, tau)
  mu <- 0.15 # prior mean
  tau <- 1 / 0.08 ^ 2 # prior precision; prior SD = 0.08

}"

dataList = list(y = y, n = n)

# Compile
model <- jags.model(file = textConnection(model_string),
                    data = dataList)
```

```
## Compiling model graph
##    Resolving undeclared variables
##    Allocating nodes
## Graph information:
##    Observed stochastic nodes: 1
##    Unobserved stochastic nodes: 1
##    Total graph size: 9
##
## Initializing model
```

```
model <- jags.model(textConnection(model_string),
                    data = dataList,
                    n.chains = 5)
```

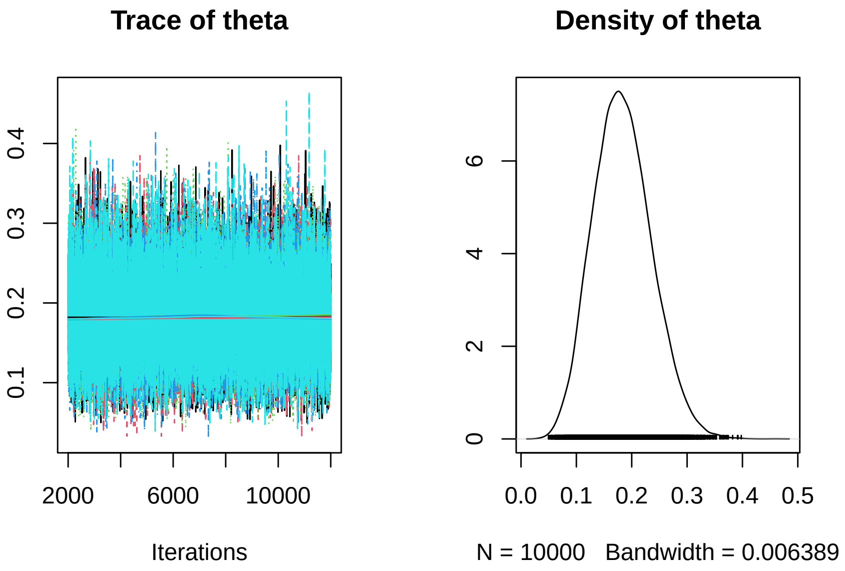
```
## Compiling model graph
##    Resolving undeclared variables
##    Allocating nodes
## Graph information:
##    Observed stochastic nodes: 1
##    Unobserved stochastic nodes: 1
##    Total graph size: 9
##
## Initializing model
```

```
# Simulate
update(model, 1000, progress.bar = "none")

Nrep = 10000

posterior_sample <- coda.samples(model,
                                 variable.names = c("theta"),
                                 n.iter = Nrep,
                                 progress.bar = "none")
```

```
# Summarize and check diagnostics
summary(posterior_sample)
```

```
##
## Iterations = 2001:12000
## Thinning interval = 1
## Number of chains = 5
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##           Mean             SD       Naive SE Time-series SE
##       0.183882       0.052470       0.000235       0.000308
##
## 2. Quantiles for each variable:
##
##   2.5%    25%    50%    75%  97.5%
## 0.0894 0.1468 0.1812 0.2181 0.2930
```

```
plot(posterior_sample)
```



The posterior density is similar to what we computed with the grid approximation.

Chapter 11

Odds and Bayes Factors

Example 11.1. The ELISA test for HIV was widely used in the mid-1990s for screening blood donations. As with most medical diagnostic tests, the ELISA test is not perfect. If a person actually carries the HIV virus, experts estimate that this test gives a positive result 97.7% of the time. (This number is called the *sensitivity* of the test.) If a person does not carry the HIV virus, ELISA gives a negative (correct) result 92.6% of the time (the *specificity* of the test). Estimates at the time were that 0.5% of the American public carried the HIV virus (the *base rate*).

Suppose that a randomly selected American tests positive; we are interested in the conditional probability that the person actually carries the virus.

1. Before proceeding, make a guess for the probability in question.

   $$0-20\% \qquad 20-40\% \qquad 40-60\% \qquad 60-80\% \qquad 80-100\%$$

2. Denote the probabilities provided in the setup using proper notation
3. Construct an appropriate two-way table and use it to compute the probability of interest.
4. Construct a Bayes table and use it to compute the probability of interest.
5. Explain why this probability is small, compared to the sensitivity and specificity.
6. By what factor has the probability of carrying HIV increased, given a positive test result, as compared to before the test?

Solution. to Example 11.1

Show/hide solution

1. We don't know what you guessed, but from experience many people guess 80-100%. Afterall, the test is correct for most of people who carry HIV,

and also correct for most people who don't carry HIV, so it seems like the test is correct most of the time. But this argument ignores one important piece of information that has a huge impact on the results: most people do not carry HIV.

2. Let H denote the event that the person carries HIV (hypothesis), and let E denote the event that the test is positive (evidence). Therefore, H^c is the event that the person does not carry HIV, another hypothesis. We are given

   - prior probability: $P(H) = 0.005$
   - likelihood of testing positive, if the person carries HIV: $P(E|H) = 0.977$
   - $P(E^c|H^c) = 0.926$
   - likelihood of testing positive, if the person does not carry HIV: $P(E|H^c) = 1 - P(E^c|H^c) = 1 - 0.926 = 0.074$
   - We want to find the posterior probability $P(H|E)$.

3. Considering a hypothetical population of Americans (at the time)

   - 0.5% *of Americans* carry HIV
   - 97.7% *of Americans who carry HIV* test positive
   - 92.6% *of Americans who do not carry HIV* test negative
   - We want to find the percentage *of Americans who test positive* that carry HIV.

4. Assuming 1000000 Americans

	Tests positive	Does not test positive	Total
Carries HIV	4885	115	5000
Does not carry HIV	73630	921370	995000
Total	78515	921485	1000000

   Among the 78515 who test positive, 4885 carry HIV, so the probability that an American who tests positive actually carries HIV is 4885/78515 = 0.062.

5. See the Bayes table below.

6. The result says that only 6.2% *of Americans who test positive* actually carry HIV. It is true that the test is correct for most Americans with HIV (4885 out of 5000) and incorrect only for a small proportion of Americans who do not carry HIV (73630 out of 995000). But since so few Americans carry HIV, the sheer *number* of false positives (73630) swamps the *number* of true positives (4885).

7. Prior to observing the test result, the prior probability that an American carries HIV is $P(H) = 0.005$. The posterior probability that an American carries HIV given a positive test result is $P(H|E) = 0.062$.

$$\frac{P(H|E)}{P(H)} = \frac{0.062}{0.005} = 12.44$$

An American who tests positive is about 12.4 times more likely to carry HIV than an American whom the test result is not known. So while 0.067 is still small in absolute terms, the posterior probability is much larger relative to the prior probability.

hypothesis	prior	likelihood	product	posterior
Carries HIV	0.005	0.977	0.0049	0.0622
Does not carry HIV	0.995	0.074	0.0736	0.9378
sum	1.000	NA	0.0785	1.0000

Remember, the conditional probability of H given E, $P(E|H)$, is not the same as the conditional probability of E given H, $P(E|H)$, and they can be vastly different. It is helpful to think of probabilities as percentages and ask “percent of what?” For example, the percentage of *people who carry HIV* that test positive is a very different quantity than the percentage of *people who test positive* that carry HIV. Make sure to properly identify the “denominator” or baseline group the percentages apply to.

Posterior probabilities can be highly influenced by the original prior probabilities, sometimes called the **base rates**. The example illustrates that when the base rate for a condition is very low and the test for the condition is less than perfect there will be a relatively high probability that a positive test is a *false positive.* Don’t neglect the base rates when evaluating posterior probabilities

Example 11.2. True story: On a camping trip in 2003, my wife and I were driving in Vermont when, suddenly, a very large, hairy, black animal lumbered across the road in front of us and into the woods on the other side. It happened very quickly, and at first I said “It’s a gorilla!” But then after some thought, and much derision from my wife, I said “it was probably a bear.”

I think this story provides an anecdote about Bayesian reasoning, albeit bad reasoning at first but then good. Put the story in a Bayesian context by identifying hypotheses, evidence, prior, and likelihood. What was the mistake I made initially?

Show/hide solution

- “Type of animal” is playing the role of the hypothesis: gorilla, bear, dog, squirrel, rabbit, etc.
- That the animal is very large, hairy, and black is the evidence.

- The likelihood value for the animal being very large, hairy, and black is close to 1 for both a bear and gorilla, maybe more middling for a dog, but close to 0 for a squirrel, rabbit, etc.

The mistake I made initially was to neglect the base rates and not consider my prior probabilities. Let's say the likelihood is 1 for both gorilla and bear and 0 for all other animals. Then based solely on the likelihoods, the posterior probability would be 50/50 for gorilla and bear, which maybe is why I guessed gorilla.

After my initial reaction, I paused to formulate my prior probabilities, which considering I was in Vermont, gave much higher probability to a bear than a gorilla. (My prior probabilities should also have given even higher probability to animals such as dogs, squirrels, and rabbits.)

By combining prior and likelihood in the appropriate way, the posterior probability is

- very high for a bear, due to high likelihood and not-too-small prior,
- close to 0 for a gorilla, due to the very small prior,
- and very low for a squirrel or rabbit or other small animals because of the close-to-zero likelihood, even if the prior is large.

Recall that the odds of an event is a ratio involving the probability that the event occurs and the probability that the event does not occur

$$\text{odds}(A) = \frac{P(A)}{P(A^c)} = \frac{P(A)}{1 - P(A)}$$

In many situations (e.g. gambling) odds are reported as odds *against* A, that is, the odds in favor of A *not* occuring, a.k.a., the odds of A^c: $P(A^c)/P(A)$.

The probability of an even can be obtained from odds

$$P(A) = \frac{\text{odds}(A)}{1 + \text{odds}(A)}$$

Example 11.3. Continuing Example 11.1

1. In symbols and words, what does one minus the answer to the probability in question in Example 11.1 represent?
2. Calculate the *prior odds* of a randomly selected American having the HIV virus, before taking an ELISA test.
3. Calculate the *posterior odds* of a randomly selected American having the HIV virus, given a positive test result.
4. By what factor has the *odds* of carrying HIV increased, given a positive test result, as compared to before the test? This is called the **Bayes factor**.

5. Suppose you were given the prior odds and the Bayes factor. How could you compute the posterior odds?
6. Compute the ratio of the likelihoods of testing positive, for those who carry HIV and for those who do not carry HIV. What do you notice?

Solution. to Example 11.3

Show/hide solution

1. $1 - P(H|E) = P(H^c|E) = 0.938$ is the posterior probability that an American who has a positive test does not carry HIV.
2. The prior probability of carrying HIV is $P(H) = 0.005$ and the prior probability of not carrying HIV is $P(H^c) = 1 - 0.005 = 0.995$

   $$\frac{P(H)}{P(H^c)} = \frac{0.005}{0.995} = \frac{1}{199} \approx 0.005025$$

   These are the prior odds in favor of carrying HIV. The prior odds against carrying HIV are

   $$\frac{P(H^c)}{P(H)} = \frac{0.995}{0.005} = 199$$

   That is, prior to taking the test, an American is 199 times more likely to not carry HIV than to carry HIV.
3. The posterior probability of carrying HIV given a positive test is $P(H|E) = 0.062$ and the posterior probability of not carrying HIV given a positive test is $P(H^c|E) = 1 - 0.062 = 0.938$.

   $$\frac{P(H|E)}{P(H^c|E)} = \frac{0.062}{0.938} \approx 0.066$$

   These are the posterior odds in favor of carrying HIV given a positive test. The posterior odds against carrying HIV given a positive test are

   $$\frac{P(H^c|E)}{P(H|E)} = \frac{0.938}{0.062} \approx 15.1$$

   That is, given a positive test, an American is 15.1 times more likely to not carry HIV than to carry HIV.
4. Comparing the prior and posterior odds in favor of carrying HIV,

   $$BF = \frac{\text{posterior odds}}{\text{prior odds}} = \frac{0.066}{0.005025} = 13.2$$

   The *odds* of carrying HIV are 13.2 times greater given a positive test result than prior to taking the test. The Bayes Factor is $BF = 13.2$.
5. By definition

   $$BF = \frac{\text{posterior odds}}{\text{prior odds}}$$

   Rearranging yields

   $$\text{posterior odds} = \text{prior odds} \times BF$$

6. The likelihood of testing positive given HIV is $P(E|H) = 0.977$ and the likelihood of testing positive given no HIV is $P(E|H^c) = 1 - 0.926 = 0.074$.

$$\frac{P(E|H)}{P(E|H^c)} = \frac{0.977}{0.074} = 13.2$$

   This value is the Bayes factor! So we could have computed the Bayes factor without first computing the posterior probabilities or odds.

- If $P(H)$ is the prior probability of H, the prior odds (in favor) of H are $P(H)/P(H^c)$
- If $P(H|E)$ is the posterior probability of H given E, the posterior odds (in favor) of H given E are $P(H|E)/P(H^c|E)$
- The **Bayes factor (BF)** is defined to be the ratio of the posterior odds to the prior odds

$$BF = \frac{\text{posterior odds}}{\text{prior odds}} = \frac{P(H|E)/P(H^c|E)}{P(H)/P(H^c)}$$

- The odds form of Bayes rule says

$$\begin{aligned}\text{posterior odds} &= \text{prior odds} \times \text{Bayes factor}\\ \frac{P(H|E)}{P(H^c|E)} &= \frac{P(H)}{P(H^c)} \times BF\end{aligned}$$

- Apply Bayes rule to $P(H|E)$ and $P(H^c|E)$

$$\begin{aligned}\frac{P(H|E)}{P(H^c|E)} &= \frac{P(E|H)P(H)/P(E)}{P(E|H^c)P(H^c)/P(E)}\\ &= \frac{P(H)}{P(H^c)} \times \frac{P(E|H)}{P(E|H^c)}\\ \text{posterior odds} &= \text{prior odds} \times \frac{P(E|H)}{P(E|H^c)}\end{aligned}$$

- Therefore, the Bayes factor for hypothesis H given evidence E can be calculated as the *ratio of the likelihoods*

$$BF = \frac{P(E|H)}{P(E|H^c)}$$

- That is, the Bayes factor can be computed without first computing posterior probabilities or odds.
- **Odds form of Bayes rule**

$$\begin{aligned}\frac{P(H|E)}{P(H^c|E)} &= \frac{P(H)}{P(H^c)} \times \frac{P(E|H)}{P(E|H^c)}\\ \text{posterior odds} &= \text{prior odds} \times \text{Bayes factor}\end{aligned}$$

Example 11.4. Continuing Example 11.1. Now suppose that 5% of individuals in a high-risk group carry the HIV virus. Consider a randomly selectd person from this group who takes the test. Suppose the sensitivity and specificity of the test are the same as in Example 11.1.

1. Compute and interpret the prior odds that a person carries HIV.
2. Use the odds form of Bayes rule to compute the posterior odds that the person carries HIV given a positive test, and interpret the posterior odds.
3. Use the posterior odds to compute the posterior probability that the person carries HIV given a positive test.

Solution. to Example 11.4

1. $P(H)/P(H^c) = 0.05/0.95 = 1/19 \approx 0.0526$. A person in this group is 19 times more likely to not carry HIV than to carry HIV.
2. The posterior odds are the product of the prior odds and the Bayes factor. The Bayes factor is the ratio of the likelihoods. Since the sensitivity and specificity are the same as in the previous example, the likelihoods are the same, and the Bayes factor is the same.

   $$\frac{P(E|H)}{P(E|H^c)} = \frac{0.977}{0.074} = 13.2$$

   Therefore

   $$\text{posterior odds} = \text{prior odds} \times \text{Bayes factor} = \frac{1}{19} \times 13.2 \approx \frac{1}{1.44} \approx 0.695$$

   Given a positive test, a person in this group is 1.44 times more likely to not carry HIV than to carry HIV.
3. The odds is the ratios of the posterior probabilities, and we basically just rescale so they add to 1. The posterior probability is

   $$P(H|E) = \frac{0.695}{1+0.695} = \frac{1}{1+1.44} \approx 0.410$$

   The Bayes table is below; we have added a row for the ratios to illustrate the odds calculations.

hypothesis	prior	likelihood	product	posterior
Carries HIV	0.0500	0.977	0.0489	0.4100
Does not carry HIV	0.9500	0.074	0.0703	0.5900
sum	1.0000	NA	0.1191	1.0000
ratio	0.0526	13.203	0.6949	0.6949

Chapter 12

Introduction to Bayesian Model Comparison

A Bayesian model is composed of both a model for the data (likelihood) and a prior distribution on model parameters. **Model selection** usually refers to choosing between different models for the data (likelihoods). But it can also concern choosing between models with the same likelihood but different priors.

In Bayesian model comparison, prior probabilities are assigned to each of the models, and these probabilities are updated given the data according to Bayes rule. Bayesian model comparison can be viewed as Bayesian estimation in a *hierarchical* model with an extra level for "model". (We'll cover hiearchical models in more detail later.)

Example 12.1. Suppose I have some trick coins, some of which are biased in favor of landing on heads, and some of which are biased in favor of landing on tails.[1] I will select a trick coin at random; let θ be the probability that the selected coin lands on heads in any single flip. I will flip the coin n times and use the data to decide about the direction of its bias. This can be viewed as a choice between two models

- Model 1: the coin is biased in favor of landing on heads
- Model 2: the coin is biased in favor of landing on tails

1. Assume that in model 1 the prior distribution for θ is Beta(7.5, 2.5). Suppose in $n = 10$ flips there are 6 heads. Use simulation to approximate the probability of observing 6 heads in 10 flips *given that model 1 is correct.*

[1] The examples in the section are motivated by examples in Kruschke (2015).

2. Assume that in model 2 the prior distribution for θ is Beta(2.5, 7.5). Suppose in $n = 10$ flips there are 6 heads. Use simulation to approximate the probability of observing 6 heads in 10 flips *given that model 2 is correct.*

3. Use the simulation results to approximate and interpret the Bayes Factor in favor of model 1 given 6 heads in 10 flips.

4. Suppose our prior probability for each model was 0.5. Find the posterior probability of each model given 6 heads in 10 flips.

5. Suppose I know I have a lot more tail biased coins, so my prior probability for model 1 was 0.1. Find the posterior probability of each model given 6 heads in 10 flips.

   Now suppose I want to predict the number of heads in the next 10 flips of the selected coin.

6. Use simulation to approximate the posterior predictive distribution of the number of heads in the next 10 flips given 6 heads in the first 10 flips *given that model 1 is the correct model.* In particular, approximate the posterior predictive probability that there are 7 heads in the next 10 flips given then model 1 is the correct model.

7. Repeat the previous part assuming model 2 is the correct model.

8. Suppose our prior probability for each model was 0.5. Use simulation to approximate the posterior predictive distribution of the number of heads in the next 10 flips given 6 heads in the first 10 flips. In particular, approximate the posterior predictive probability that there are 7 heads in the next 10 flips.

Solution. to Example 12.1

1. Given that model 1 is correct, simulate a value of θ from a Beta(7.5, 2.5) prior, and then given θ simulate a value of y from a Binomial(10, θ) distribution. Repeat many times. The proportion of simulated repetitions that yield a y value of 6 approximates the probability of observing 6 heads in 10 flips given that model 1 is correct. The probability is 0.124.

   ```
   Nrep = 1000000
   theta = rbeta(Nrep, 7.5, 2.5)
   y = rbinom(Nrep, 10, theta)
   sum(y == 6) / Nrep
   ```

   ```
   ## [1] 0.1243
   ```

2. Similar to he previous part, with the model 2 prior. The probability is 0.042.

```
Nrep = 1000000
theta = rbeta(Nrep, 2.5, 7.5)
y = rbinom(Nrep, 10, theta)
sum(y == 6) / Nrep
```

```
## [1] 0.04191
```

3. The Bayes factor is the ratio of the likelihoods. The likelihood of 6 heads in 10 flips under model 1 is 0.124, and under model 2 is 0.042. The Bayes factor in favor of model 1 is $0.124/0.042 = 2.95$. Observing 6 heads in 10 flips is 2.95 more likely under model 1 than under model 2. Also, the posterior odds in favor of model 1 given 6 heads in 10 flips are 2.95 times greater than the prior odds in favor of model 1.

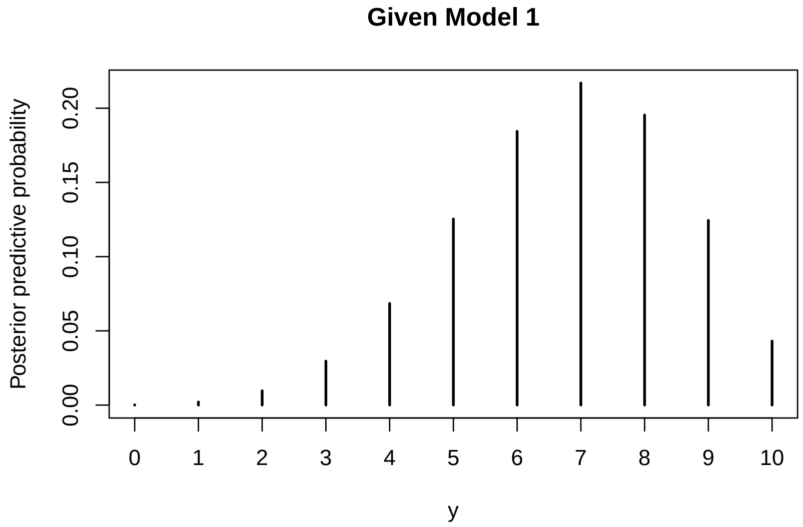
4. In this case, the prior odds are 1, so the posterior odds in favor of model 1 are 2.95. The posterior probability of model 1 is 0.747, and the posterior probability of model 2 is 0.253.

5. Now the prior odds in favor of model 1 are 1/9. So the posterior odds in favor of model 1 given 6 heads in 10 flips are (1/9)(2.95)=0.328. The posterior probability of model 1 is 0.247, and the posterior probability of model 2 is 0.753.

   Now suppose I want to predict the number of heads in the next 10 flips of the selected coin.

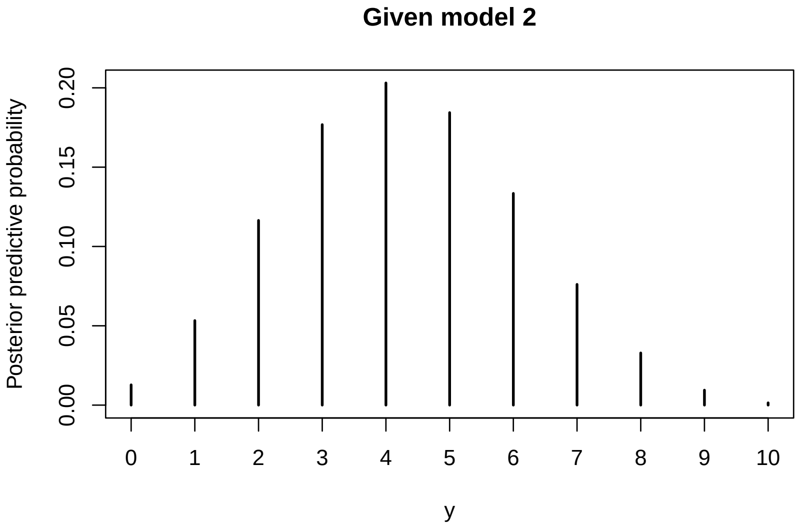
6. If model 1 is correct the prior is Beta(7.5, 2.5) so the posterior after observing 6 heads in 10 flips is Beta(13.5, 6.5). Simulate a value of θ from a Beta(13.5, 6.5) distribution and given θ simulate a value of y from a Binomial(10, θ) distribution. Repeat many times. Approximate the posterior predictive probability of 7 heads in the 10 flips flips, given model 1 is correct and 6 heads in the first 10 flips, with the proportion of simulated repetitions that yield a y value of 7; the probability is 0.216.

```
Nrep = 1000000
theta = rbeta(Nrep, 7.5 + 6, 2.5 + 4)
y = rbinom(Nrep, 10, theta)
plot(table(y) / Nrep,
 ylab = "Posterior predictive probability",
 main = "Given Model 1")
```



7. The simulation is similar, just use the prior in model 2. The posterior predictive probability of 7 heads in the 10 flips flips, given model 2 is correct and 6 heads in the first 10 flips, is 0.076.

```
Nrep = 1000000
theta = rbeta(Nrep, 2.5 + 6, 7.5 + 4)
y = rbinom(Nrep, 10, theta)
plot(table(y) / Nrep,
 ylab = "Posterior predictive probability",
 main = "Given model 2")
```

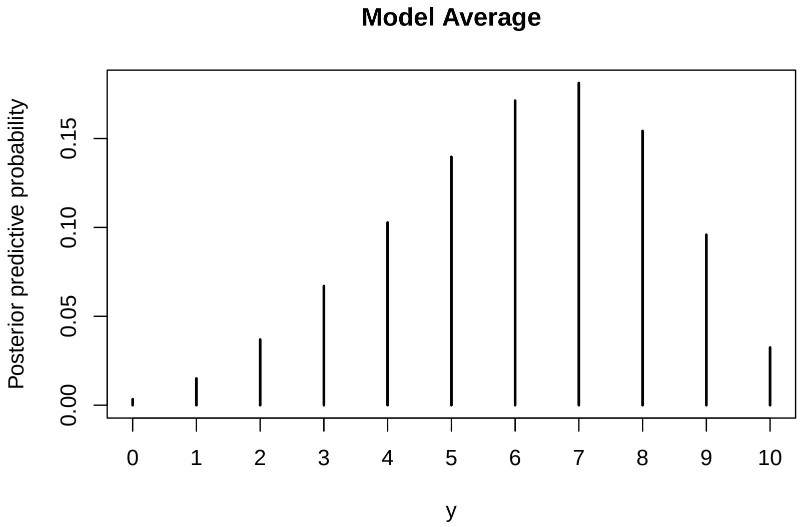


8. We saw in a previous part that with a 0.5/0.5 prior on model and 6 heads in 10 flips, the posterior probability of model 1 is 0.747 and of model 2 is 0.253. We now add another stage to our simulation

   - Simulate a model: model 1 with probability 0.747 and model 2 with probability 0.253
   - Given the model simulate a value of θ from its posterior distribution: Beta(13.5, 6.5) if model 1, Beta(8.5, 11.5) if model 2.
   - Given θ simulate a value of y from a Binomial(10, θ) distribtution

   The simulation results are below. We can also find the posterior predictive probability of 7 heads in the next 10 flips using the law of total probability to combine the results from the two previous parts: $(0.747)(0.216) + (0.253)(0.076) = 0.18$

```r
Nrep = 1000000
alpha = c(7.5, 2.5) + 6
beta = c(2.5, 7.5) + 4

model = sample(1:2, size = Nrep, replace = TRUE, prob = c(0.747, 0.253))

theta = rbeta(Nrep, alpha[model], beta[model])

y = rbinom(Nrep, 10, theta)

plot(table(y) / Nrep,
 ylab = "Posterior predictive probability",
 main = "Model Average")
```



When several models are under consideration, the Bayesian model is the full hierarchical structure which spans all models being compared. Thus, the most complete posterior prediction takes into account all models, weighted by their posterior probabilities. That is, prediction is accomplished by taking a weighted average across the models, with weights equal to the posterior probabilities of the models. This is called **model averaging**.

Example 12.2. Suppose again I select a coin, but now the decision is whether the coin is fair. Suppose we consider the two models

- "Must be fair" model: prior distribution for θ is Beta(500, 500)
- "Anything is possible" model: prior distribution for θ is Beta(1, 1)

1. Suppose we observe 15 heads in 20 flips. Use simulation to approximate the Bayes factor in favor of the "must be fair" model given 15 heads in 20 flips. Which model does the Bayes factor favor?
2. Suppose we observe 11 heads in 20 flips. Use simulation to approximate the Bayes factor in favor of the "must be fair" model given 11 heads in 20 flips. Which model does the Bayes factor favor?
3. The "anything is possible" model has any value available to it, including 0.5 and the sample proportion 0.55. Why then is the "must be fair" option favored in the previous part?

Solution. to Example 12.2

1. A sample proportion of 15/20 = 0.75 does not seem consistent with the “must be fair” model, so we expect the Bayes Factor to favor the “anything in possible” model.

   The Bayes Factor is the ratio of the likelihoods (of 15/20). To approximate the likelihood of 15 heads in 20 flips for the “must be fair” model

   - Simulate a value θ from a Beta(500, 500) distribution
   - Given θ, simulate a value y from a Binomial(20, θ) distribution
   - Repeat many times and the proportion of simulated repetitions that yield a y of 15.

   Approximate the likelihood of 15 heads in 20 flips for the “anything is possible” model similarly. The Bayes factor is the ratio of the likelihoods, about 0.323 in favor of the “must be fair” model. That is, the Bayes factor favors the “anything is possible” model.

   ```
   Nrep = 1000000

   theta1 = rbeta(Nrep, 500, 500)
   y1 = rbinom(Nrep, 20, theta1)

   theta2 = rbeta(Nrep, 1, 1)
   y2 = rbinom(Nrep, 20, theta2)

   sum(y1 == 15) / sum(y2 == 15)
   ```

   ```
   ## [1] 0.3197
   ```

2. Similar to the previous part but with different data; now we compute the likelihood of 11 heads in 20 flips. The Bayes factor is about 3.34. Thus, the Bayes factor favors the “must be fair” model.

   ```
   sum(y1 == 11) / sum(y2 == 11)
   ```

   ```
   ## [1] 3.328
   ```

3. A central 99% prior credible interval for θ based on the “must be fair” model is (0.459, 0.541), which does not include the sample proportion of 0.55. So you might think that the data would favor the “anything is possible” model. However, the numerator and denominator in the Bayes factor are *average* likelihoods: the likelihood of the data averaged over each possible value of θ. The “must be fair” model only gives initial plausibility

to θ values that are close to 0.5, and for such θ values the likelihood of 11 heads in 20 flips is not so small. Values of θ that are far from 0.5 are effectively not included in the average, due to their low prior probability, so the average likelihood is not so small.

In contrast, the "anything is possible" model stretches the prior probability over all values in (0, 1). For many θ values in (0, 1) the likelihood of observing 11 heads in 20 flips is close to 0, and with the Uniform(0, 1) prior, each of these θ values contributes equally to the average likelihood. Thus, the average likelihood is smaller for the "anything is possible" model than for the "must be fair" model.

Complex models generally have an inherent advantage over simpler models because complex models have many more options available, and one of those options is likely to fit the data better than any of the fewer options in the simpler model. However, we don't always want to just choose the more complex model. Always choosing the more complex model overfits the data.

Bayesian model comparison naturally compensates for discrepancies in model complexity. In more complex models, prior probabilities are diluted over the many options available. Even if a complex model has some particular combination of parameters that fit the data well, the prior probability of that particular combination is likely to be small because the prior is spread more thinly than for a simpler model. Thus, in Bayesian model comparison, a simpler model can "win" if the data are consistent with it, even if the complex model fits well.

Example 12.3. Continuing Example 12.2 where we considered the two models

- "Must be fair" model: prior distribution for θ is Beta(500, 500)
- "Anything is possible" model: prior distribution for θ is Beta(1, 1)

Suppose we observe 65 heads in 100 flips.

1. Use simulation to approximate the Bayes factor in favor of the "must be fair" model given 65 heads in 100 flips. Which model does the Bayes factor favor?

2. We have discussed different notions of a "non-informative/vague" prior. We often think of Beta(1, 1) = Uniform(0, 1) as a non-informative prior, but there are other considerations. In particular, a Beta(0.01, 0.01) is often used a non-informative prior in this context. Think of a Beta(0.01, 0.01) prior like an approximation to the improper Beta(0, 0) prior based on "no prior successes or failures".

   Suppose now that the "anything is possible" model corresponds to a Beta(0.01, 0.01) prior distribution for θ. Use simulation to approximate the Bayes factor in favor of the "must be fair" model given 65 heads in

100 flips. Which model does the Bayes factor favor? Is the *choice of model* sensitive to the change of prior distribution within the “anything is possible” model?

3. For each of the two “anything is possible” priors, find the posterior distribution of θ and a 98% posterior credible interval for θ given 65 heads in 100 flips. Is *estimation of θ* within the “anything is possible” model sensitive to the change in the prior distribution for θ?

Solution. to Example 12.3

1. The simulation is similar to the ones in the previous example, just with different data. The Bayes Factor is about 0.126 in favor of the “must be fair” model. So the Bayes Factor favors the “anything is possible” model.

```r
Nrep = 1000000

theta1 = rbeta(Nrep, 500, 500)
y1 = rbinom(Nrep, 100, theta1)

theta2 = rbeta(Nrep, 1, 1)
y2 = rbinom(Nrep, 100, theta2)

sum(y1 == 65) / sum(y2 == 65)
```

```
## [1] 0.1325
```

2. The simulation is similar to the one in the previous part, just with a different prior. The Bayes Factor is about 5.73 in favor of the “must be fair” model. So the Bayes Factor favors the “must be fair” model. Even though there both non-informative priors, the Beta(1, 1) and Beta(0.01, 0.01) priors leads to very different Bayes factors and decisions. The *choice of model* does appear to be sensitive to the choice of prior distribution.

```r
Nrep = 1000000

theta1 = rbeta(Nrep, 500, 500)
y1 = rbinom(Nrep, 100, theta1)

theta2 = rbeta(Nrep, 0.01, 0.01)
y2 = rbinom(Nrep, 100, theta2)

sum(y1 == 65) / sum(y2 == 65)
```

```
## [1] 5.272
```

3. For a Beta(1, 1) prior, the posterior of θ given 65 heads in 100 flips is the Beta(66, 36) distribution, and a central 98% posterior credible interval for θ is (0.534, 0.752). For a Beta(0.01, 0.01) prior, the posterior of θ given 65 heads in 100 flips is the Beta(65.01, 35.01) distribution, and a central 98% posterior credible interval for θ is (0.536, 0.755). The Beta(66, 36) and Beta(65.01, 35.01) distributions are virtually identical, and the 98% credible intervals are practically the same. At least in this case, the *estimation of* θ within the "anything is possible" model does not appear to be sensitive to the choice of prior.

   ```r
   qbeta(c(0.01, 0.99), 1 + 65, 1 + 35)
   ```

   ```
   ## [1] 0.5340 0.7517
   ```

   ```r
   qbeta(c(0.01, 0.99), 0.01 + 65, 0.01 + 35)
   ```

   ```
   ## [1] 0.5359 0.7553
   ```

In Bayesian *estimation of continuous parameters within a model*, the posterior distribution is typically not too sensitive to changes in prior (provided that there is a reasonable amount of data and the prior is not too strict).

In contrast, in Bayesian *model comparison*, the posterior probabilities of the models and the Bayes factors can be extremely sensitive to the choice of prior distribution within each model.

When comparing different models, prior distributions on parameters within each model should be equally informed. One strategy is to use a small set of "training data" to inform the prior of each model before comparing.

Example 12.4. Continuing Example 12.3 where we considered two priors in the "anything is possible model": Beta(1, 1) and Beta(0.1, 0.1). We will again compare the "anything is possible model" to the "must be fair" model which corresponds to a Beta(500, 500) prior.

Suppose we observe 65 heads in 100 flips.

1. Assume the "anything is possible" model corresponds to the Beta(1, 1) prior. Suppose that in the first 10 flips there were 6 heads. Compute the posterior distribution of θ in each of the models after the first 10 flips. Then use simulation to approximate the Bayes factor in favor of the "must be fair" model given 65 heads in 100 flips, using the posterior distribution of θ after the first 10 flips as the prior distribution in the simulation. Which model does the Bayes factor favor?
2. Repeat the previous part assuming the "anything is possible" model corresponds to the Beta(0.01, 0.01) prior. Compare with the previous part.

Solution. to Example 12.4

1. With the Beta(1, 1) prior in the "anything is possible" model, the posterior distribution of θ after 6 heads in the first 10 flips is the Beta(7, 5) distribution. With the Beta(500, 500) prior in the "must be fair" model, the posterior distribution of θ after 6 heads in the first 10 flips is the Beta(506, 504) distribution. The simulation to approximate the likelihood in each model is similar to before, but now we simulate θ from its posterior distribution after the first 10 flips, and evaluate the likelihood of observing 59 heads in the remaining 90 flips. The Bayes factor is about 0.056 in favor of the "must be fair" model. So the Bayes Factor favors the "anything is possible" model.

   ```
   Nrep = 1000000

   theta1 = rbeta(Nrep, 500 + 6, 500 + 4)
   y1 = rbinom(Nrep, 90, theta1)

   theta2 = rbeta(Nrep, 1 + 6, 1 + 4)
   y2 = rbinom(Nrep, 90, theta2)

   sum(y1 == 59) / sum(y2 == 59)
   ```

   ```
   ## [1] 0.05509
   ```

2. With the Beta(0.01, 0.01) prior in the "anything is possible" model, the posterior distribution of θ after 6 heads in the first 10 flips is the Beta(6.01, 4.01) distribution. The simulation is similar to the previous part, just with the different distribution for θ in the "anything is possible" model. The Bayes factor is about 0.057 in favor of the "must be fair" model, about the same as in the previous part. So the Bayes Factor favors the "anything is possible" model. Notice that after "training" the models on the first 10 observations, the model comparison is no longer so sensitive to the choice of prior within the "anything is possible" model.

   ```
   Nrep = 1000000

   theta1 = rbeta(Nrep, 500 + 6, 500 + 4)
   y1 = rbinom(Nrep, 90, theta1)

   theta2 = rbeta(Nrep, 0.01 + 6, 0.01 + 4)
   y2 = rbinom(Nrep, 90, theta2)

   sum(y1 == 59) / sum(y2 == 59)
   ```

```
## [1] 0.05941
```

Example 12.5. Consider a null hypothesis significance test of $H_0 : \theta = 0.5$ versus $H_1 : \theta \neq 0.5$. How does this situation resemble the previous problem?

Solution. to Example 12.5

We could treat this as a problem of Bayesian model comparison. The null hypothesis corresponds to a prior distribution which places all prior probability on the null hypothesized value of 0.5. The alternative hypothesis corresponds to a prior distribution over the full range of possible values of θ. Given data, we could compute the posterior probability of each model and use that to make a decision regarding the hypotheses. However, there are infinitely many choices for the prior that corresponds to the alternative hypothesis, and we have already seen that Bayesian model comparison can be very sensitive to the choice of prior within in model.

A null hypothesis significance test can be viewed as a problem of Bayesian model selection in which one model has a prior distribution that places all its credibility on the null hypothesized value. However, is it really plausible that the parameter is *exactly* equal to the hypothesized value?

Unfortunately, this model-comparison (Bayes factor) approach to testing can be extremely sensitive to the choice of prior corresponding to the alternative hypothesis.

An alternative Bayesian approach to testing involves choosing a **region of practical equivalence (ROPE).** A ROPE indicates a small range of parameter values that are considered to be practically equivalent to the null hypothesized value.

- A hypothesized value is rejected — that is, declared to be not credible — if its ROPE lies outside a posterior credible interval (e.g., 99%) for the parameter.
- A hypothesized value is accepted for practical purposes if its ROPE contains the posterior credible interval (e.g., 99%) for the parameter.

How do you choose the ROPE? That determines on the practical application.

In general, traditional testing of point null hypotheses (that is, "*no* effect/difference") is not a primary concern in Bayesian statistics. Rather, the *posterior distribution* provides all relevant information to make decisions about practically meaningful issues. Ask research questions that are important in the context of the problem and use the posterior distribution to answer them.

Chapter 13

Bayesian Analysis of Poisson Count Data

In this chapter we'll consider Bayesian analysis for count data.

We have covered in some detail the problem of estimating a population proportion for a binary categorical variable. In these situations we assumed a Binomial likelihood for the count of "successes" in the sample. However, a Binomial model has several restrictive assumptions that might not be satisfied in practice. *Poisson models* are more flexible models for count data.

Example 13.1. Let Y be the number of home runs hit (in total by both teams) in a randomly selected Major League Baseball game.

1. In what ways is this like the Binomial situation? (What is a trial? What is "success"?)
2. In what ways is this NOT like the Binomial situation?

Solution. to Example 13.1

Show/hide solution

1. Each pitch is a trial, and on each trial either a home run is hit ("success") or not. The random variable Y counts the number of home runs (successes) over all the trials
2. Even though Y is counting successes, this is not the Binomial situation.
    - The number of trials is not fixed. The total number of pitches varies from game to game. (The average is around 300 pitches per game).
    - The probability of success is not the same on each trial. Different batters have different probabilities of hitting home runs. Also, different pitch counts or game situations lead to different probabilities of home runs.

- The trials might not be independent, though this is a little more questionable. Make sure you distinguish independence from the previous assumption of unequal probabilities of success; you need to consider conditional probabilities to assess independence. Maybe if a pitcher gives up a home run on one pitch, then the pitcher is "rattled" so the probability that he also gives up a home run on the next pitch increases, or the pitcher gets pulled for a new pitcher which changes the probability of a home run on the next pitch.

Example 13.2. Let Y be the number of automobiles that get in accidents on Highway 101 in San Luis Obispo on a randomly selected day.

1. In what ways is this like the Binomial situation? (What is a trial? What is "success"?)
2. In what ways is this NOT like the Binomial situation?

Solution. to Example 13.2

Show/hide solution

1. Each automobile on the road in the day is a trial, and on each automobile either gets in an accident ("success") or not. The random variable Y counts the number of automobiles that get into accidents (successes). (Remember "success" is just a generic label for the event you're interested in; "success" is not necessarily good.)
2. Even though Y is counting successes, this is not the Binomial situation.
   - The number of trials is not fixed. The total number of automobiles on the road varies from day to day.
   - The probability of success is not the same on each trial. Different drivers have different probabilities of getting into accidents; some drivers are safer than others. Also, different conditions increase the probability of an accident, like driving at night.
   - The trials are plausibly not independent. Make sure you distinguish independence from the previous assumption of unequal probabilities of success; you need to consider conditional probabilities to assess independence. If an automobile gets into an accident, then the probability of getting into an accident increases for the automobiles that are driving near it.

Poisson models are models for counts that have more flexibility than Binomial models. Poisson models are parameterized by a single parameter (the mean) and do not require all the assumptions of a Binomial model. Poisson distributions are often used to model the distribution of variables that count the number of "relatively rare" events that occur over a certain interval of time or in a certain location (e.g., number of accidents on a highway in a day, number of car insurance policies that have claims in a week, number of bank loans that go into default, number of mutations in a DNA sequence, number of earthquakes that occur in SoCal in an hour, etc.)

A discrete random variable Y has a **Poisson distribution** with parameter $\theta > 0$ if its probability mass function satisfies

$$f(y|\theta) = \frac{e^{-\theta}\theta^{y}}{y!}, \quad y = 0, 1, 2, ...$$

If Y has a Poisson(θ) distribution then

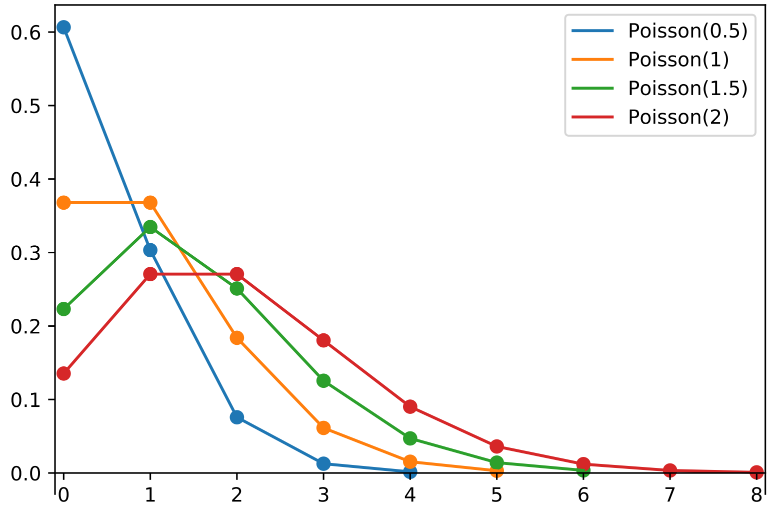
$$\begin{aligned} E(Y) &= \theta \\ Var(Y) &= \theta \end{aligned}$$

For a Poisson distribution, both the mean and variance are equal to θ, but remember that the mean is measured in the count units (e.g., home runs) but the variance is measured in squared units (e.g., (home runs)^2).

Poisson distributions have many nice properties, including the following.

Poisson aggregation. If Y_1 and Y_2 are independent, Y_1 has a Poisson(θ_1) distribution, and Y_2 has a Poisson(θ_2) distribution, then $Y_1 + Y_2$ has a Poisson($\theta_1 + \theta_2$) distribution[1]. That is, if independent component counts each follow a Poisson distribution then the total count also follows a Poisson distribution. Poisson aggregation extends naturally to more than two components. For example, if the number of babies born each day at a certain hospital follows a Poisson distribution — perhaps with different daily rates (e.g., higher for Friday than Saturday) — independently from day to day, then the number of babies born each week at the hospital also follows a Poisson distribution.

[1] If Y_1 has mean θ_1 and Y_2 has mean θ_2 then linearity of expected value implies that $Y_1 + Y_2$ has mean $\theta_1 + \theta_2$. If Y_1 has variance θ_1 and Y_2 has variance θ_2 then independence of Y_1 and Y_2 implies that $Y_1 + Y_2$ has variance $\theta_1 + \theta_2$. What Poisson aggregation says is that if component counts are independent and each with a Poisson *shape*, then the total count also has a Poisson *shape*.



Example 13.3. Suppose the number of home runs hit per game (by both teams in total) at a particular Major League Baseball park follows a Poisson distribution with parameter θ.

1. Sketch your prior distribution for θ and describe its features. What are the possible values of θ? Does θ take values on a discrete or continuous scale?

2. Suppose Y represents a home run count for a single game. What are the possible values of Y? Does Y take values on a discrete or continuous scale?

3. We'll start with a discrete prior for θ to illustrate ideas.

θ	0.5	1.5	2.5	3.5	4.5
Probability	0.13	0.45	0.28	0.11	0.03

Suppose a single game with 1 home run is observed. Find the posterior distribution of θ. In particular, how do you determine the likelihood column?

4. Now suppose a second game, with 3 home runs, is observed, independently of the first. Find the posterior distribution of θ after observing these two games, using the posterior distribution from the previous part as the prior distribution in this part.

5. Now consider the original prior again. Find the posterior distribution of θ after observing 1 home run in the first game and 3 home runs in the second, without the intermediate updating of the posterior after the first game. How does the likelihood column relate to the likelihood columns from the previous parts? How does the posterior distribution compare with the posterior distribution from the previous part?

6. Now consider the original prior again. Suppose that instead of observing the two individual values, we only observe that there is a total of 4 home runs in 2 games. Find the posterior distribution of θ. In particular, you do you determine the likelihood column? How does the likelihood column compare to the one from the previous part? How does posterior compare to the previous part?

7. Suppose we'll observe a third game tomorrow. How could you find — both analytically and via simulation —the posterior predictive probability that this game has 0 home runs?

8. Now let's consider a continuous prior distribution for θ which satisfies

$$\pi(\theta) \propto \theta^{4-1}e^{-2\theta}, \qquad \theta > 0$$

Use grid approximation to compute the posterior distribution of θ given 1 home run in a single game. Plot the prior, (scaled) likelihood, and posterior. (Note: you will need to cut the grid off at some point. While θ can take any value greater than 0, the interval [0, 8] accounts for 99.99% of the prior probability.)

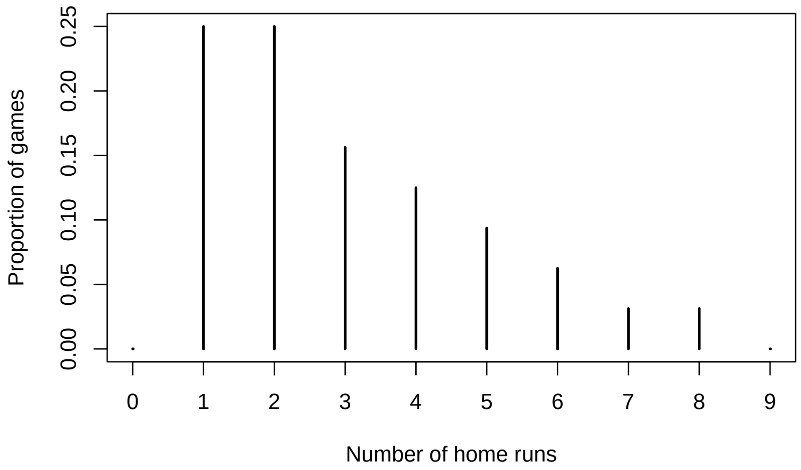
9. Now let's consider some real data. Assume home runs per game at Citizens Bank Park (Phillies!) follow a Poisson distribution with parameter θ. Assume that the prior distribution for θ satisfies

$$\pi(\theta) \propto \theta^{4-1}e^{-2\theta}, \qquad \theta > 0$$

The following summarizes data for the 2020 season[2]. There were 97 home runs in 32 games. Use grid approximation to compute the posterior distribution of θ given the data. Be sure to specify the likelihood. Plot the prior, (scaled) likelihood, and posterior.

[2] Source: https://www.baseball-reference.com/teams/PHI/2020.shtml

Home runs	Number of games
0	0
1	8
2	8
3	5
4	4
5	3
6	2
7	1
8	1
9	0



Solution. to Example 13.3

1. Your prior is whatever it is. We'll discuss how we chose a prior in a later part. Even though each data value is an integer, the mean number of home runs per game θ can be any value greater than 0. That is, the *parameter* θ takes values on a continuous scale.

2. Y can be 0, 1, 2, and so on, taking values on a discrete scale. Technically, there is no fixed upper bound on what Y can be.

3. The likelihood is the Poisson probability of 1 home run in a game computed for each value of θ.
$$f(y = 1|\theta) = \frac{e^{-\theta}\theta^1}{1!}$$
For example, the likelihood of 1 home run in a game given $\theta = 0.5$ is $f(y = 1|\theta = 0.5) = \frac{e^{-0.5}0.5^1}{1!} = 0.3033$. If on average there are 0.5 home

runs per game, then about 30% of games would have exactly 1 home run. As always posterior is proportional to the product of prior and likelihood. We see that the posterior distibution puts even greater probability on $\theta = 1.5$ than the prior.

theta	prior	likelihood	product	posterior
0.5	0.13	0.3033	0.0394	0.1513
1.5	0.45	0.3347	0.1506	0.5779
2.5	0.28	0.2052	0.0575	0.2205
3.5	0.11	0.1057	0.0116	0.0446
4.5	0.03	0.0500	0.0015	0.0058

4. The likelihood is the Poisson probability of 3 home runs in a game computed for each value of θ.

$$f(y = 3|\theta) = \frac{e^{-\theta}\theta^3}{3!}$$

The posterior places about 90% of the probability on θ being either 1.5 or 2.5.

theta	prior	likelihood	product	posterior
0.5	0.1513	0.0126	0.0019	0.0145
1.5	0.5779	0.1255	0.0725	0.5488
2.5	0.2205	0.2138	0.0471	0.3566
3.5	0.0446	0.2158	0.0096	0.0728
4.5	0.0058	0.1687	0.0010	0.0073

5. Since the games are independent[3] the likelihood is the product of the likelihoods from the two previous parts

$$f(y = (1,3)|\theta) = \left(\frac{e^{-\theta}\theta^1}{1!}\right)\left(\frac{e^{-\theta}\theta^3}{3!}\right)$$

Unsuprisingly, the posterior distribution is the same as in the previous part.

theta	prior	likelihood	product	posterior
0.5	0.13	0.0038	0.0005	0.0145
1.5	0.45	0.0420	0.0189	0.5488
2.5	0.28	0.0439	0.0123	0.3566
3.5	0.11	0.0228	0.0025	0.0728
4.5	0.03	0.0084	0.0003	0.0073

6. By Poisson aggregation, the total number of home runs in 2 games follows a Poisson(2θ) distribution. The likelihood is the probability of a value of

[3] I keep meaning to say this, but technically the Y values are not independent. Rather, they are *conditionally independent given θ*. This is a somewhat subtle distinction, so I've glossed over the details.

4 (home runs in 2 games) computed using a Poisson(2θ) for each value of θ.

$$f(\bar{y} = 2|\theta) = \frac{e^{-2\theta}(2\theta)^4}{4!}$$

For example, the likelihood of 4 home runs in 2 games given $\theta = 0.5$ is $f(\bar{y} = 2|\theta = 0.5) = \frac{e^{-2\times 0.5}(2\times 0.5)^4}{4!} = 0.0153$. If on average there are 0.5 home runs per game, then about 1.5% of samples of 2 games would have exactly 4 home runs.

The likelihood is not the same as in the previous part because there are more samples of two games that yield a total of 4 home runs than those that yield 1 home run in the first game and 3 in the second. However, the likelihoods are *proportionally* the same. For example, the likelihood for $\theta = 2.5$ is about 1.92 times greater than the likelihood for $\theta = 3.5$ in both this part and the previous part. Therefore, the posterior distribution is the same as in the previous part.

theta	prior	likelihood	product	posterior
0.5	0.13	0.0153	0.0020	0.0145
1.5	0.45	0.1680	0.0756	0.5488
2.5	0.28	0.1755	0.0491	0.3566
3.5	0.11	0.0912	0.0100	0.0728
4.5	0.03	0.0337	0.0010	0.0073

7. Simulate a value of θ from its posterior distribution and then given θ simulate a value of Y from a Poisson(θ) distribution, and repeat many times. Approximate the probability of 0 home runs by finding the proportion of repetitions that yield a Y value of 0. (We'll see some code a little later.)

   We can compute the probability using the law of total probability. Find the probability of 0 home runs for each value of θ, that is $e^{-\theta}\theta^0/0! = e^{-\theta}$, and then weight these values by their posterior probabilities to find the predictive probability of 0 home runs, which is 0.163.

$$\begin{aligned} & e^{-0.5}(0.0145) + e^{-1.5}(0.5488) + e^{-2.5}(0.3566) + e^{-3.5}(0.0728) + e^{-4.5}(0.0073) \\ = & (0.6065)(0.0145) + (0.2231)(0.5488) + (0.0821)(0.3566) + (0.0302)(0.0728) + (0.0111)(0.0073) = \end{aligned}$$

   According to this model, we predict that about 16% of games would have 0 home runs.

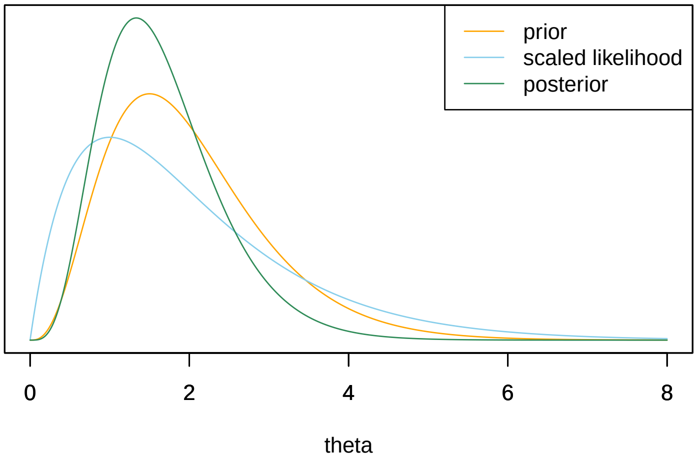
8. Now let's consider a continuous prior distribution for θ which satisfies

$$\pi(\theta) \propto \theta^{4-1}e^{-2\theta}, \qquad \theta > 0$$

   Use grid approximation to compute the posterior distribution of θ given 1 home run in a single game. Plot the prior, (scaled) likelihood, and posterior. (Note: you will need to cut the grid off at some point. While θ

can take any value greater than 0, the interval [0, 8] accounts for 99.99% of the prior probability.)

```
# prior
theta = seq(0, 8, 0.001)

prior = theta ^ (4 - 1) * exp(-2 * theta)
prior = prior / sum(prior)

# data
n = 1 # sample size
y = 1 # sample mean

# likelihood
likelihood = dpois(y, theta)

# posterior
product = likelihood * prior
posterior = product / sum(product)

ylim = c(0, max(c(prior, posterior, likelihood / sum(likelihood))))
xlim = range(theta)
plot(theta, prior, type='l', xlim=xlim, ylim=ylim, col="orange", xlab='theta', ylab='', yaxt
par(new=T)
plot(theta, likelihood/sum(likelihood), type='l', xlim=xlim, ylim=ylim, col="skyblue", xlab=
par(new=T)
plot(theta, posterior, type='l', xlim=xlim, ylim=ylim, col="seagreen", xlab='', ylab='', yax
legend("topright", c("prior", "scaled likelihood", "posterior"), lty=1, col=c("orange", "sky
```



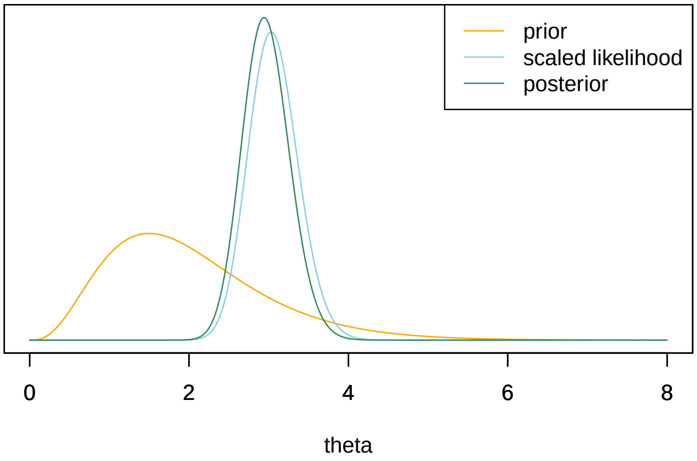
9. By Poisson aggregation, the total number of home runs in 32 games follows a Poisson(32θ) distribution. The likelihood is the probability of observing a value of 97 (for the total number of home runs in 32 games) from a Poisson(32θ) distribution.

$$\begin{aligned} f(\bar{y} = 97/32|\theta) &= e^{-32\theta}(32\theta)^{97}/97!, \qquad \theta > 0 \\ &\propto e^{-32\theta}\theta^{97}, \qquad \theta > 0 \end{aligned}$$

The likelihood is centered at the sample mean of $97/32 = 3.03$. The posterior distribution follows the likelihood fairly closely, but the prior still has a little influence.

```r
# prior
theta = seq(0, 8, 0.001)

prior = theta ^ (4 - 1) * exp(-2 * theta)
prior = prior / sum(prior)

# data
n = 32 # sample size
y = 97 / 32 # sample mean

# likelihood - for total count
likelihood = dpois(n * y, n * theta)
```

```
# posterior
product = likelihood * prior
posterior = product / sum(product)

ylim = c(0, max(c(prior, posterior, likelihood / sum(likelihood))))
xlim = range(theta)
plot(theta, prior, type='l', xlim=xlim, ylim=ylim, col="orange", xlab='theta', ylab='', yaxt
par(new=T)
plot(theta, likelihood/sum(likelihood), type='l', xlim=xlim, ylim=ylim, col="skyblue", xlab=
par(new=T)
plot(theta, posterior, type='l', xlim=xlim, ylim=ylim, col="seagreen", xlab='', ylab='', yax
legend("topright", c("prior", "scaled likelihood", "posterior"), lty=1, col=c("orange", "sky
```



Gamma distributions are commonly used as prior distributions for parameters that take positive values, $\theta > 0$.

A continuous RV U has a **Gamma distribution** with *shape parameter* $\alpha > 0$

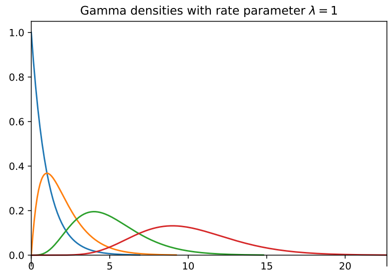
and *rate parameter*[4] $\lambda > 0$ if its density satisfies[5]

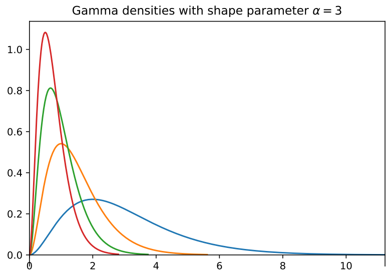
$$f(u) \propto u^{\alpha-1}e^{-\lambda u}, \quad u > 0,$$

In R: `dgamma(u, shape, rate)` for density, `rgamma` to simulate, `qgamma` for quantiles, etc.

It can be shown that a Gamma(α, λ) density has

$$\begin{aligned}\text{Mean (EV)} &= \frac{\alpha}{\lambda}\\ \text{Variance} &= \frac{\alpha}{\lambda^2}\\ \text{Mode} &= \frac{\alpha-1}{\lambda}, \qquad \text{if } \alpha \geq 1\end{aligned}$$





Example 13.4. The plots above show a few examples of Gamma distributions.

1. The plot on the left above contains a few different Gamma densities, all with rate parameter $\lambda = 1$. Match each density to its shape parameter α; the choices are 1, 2, 5, 10.
2. The plot on the right above contains a few different Gamma densities, all with shape parameter $\alpha = 3$. Match each density to its rate parameter λ; the choices are 1, 2, 3, 4.

Solution. to Example 13.4

[4] Sometimes Gamma densities are parametrized in terms of the *scale parameter* $1/\lambda$, so that the mean is $\alpha\lambda$.

[5] The expression defines the shape of a Gamma density. All that's missing is the scaling constant which ensures that the total area under the density is 1. The actual Gamma density formula, including the normalizing constant, is

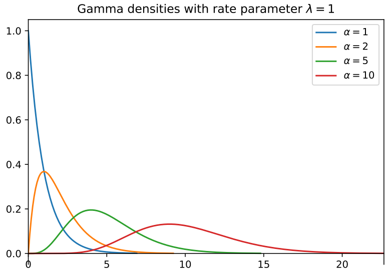
$$f(u) = \frac{\lambda^\alpha}{\Gamma(\alpha)}\, u^{\alpha-1}e^{-\lambda u}, \quad u > 0,$$

where $\Gamma(\alpha) = \int_0^\infty e^{-u}u^{\alpha-1}du$ is the *Gamma function*. For a positive integer k, $\Gamma(k) = (k-1)!$. Also, $\Gamma(1/2) = \sqrt{\pi}$.

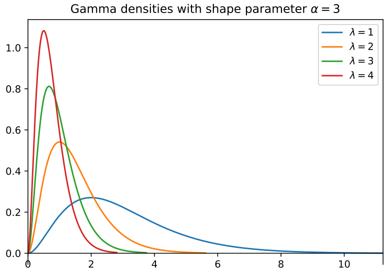
1. For a fixed λ, as the shape parameter α increases, both the mean and the standard deviation increase.

2. For a fixed α, as the rate parameter λ increases, both the mean and the standard deviation decrease.

   Observe that changing λ doesn't change the overall shape of the curve, just the scale of values that it covers. However, changing α does change the shape of the curve; notice the changes in concavity in the plot on the left.



Gamma densities with rate parameter $\lambda = 1$



Gamma densities with shape parameter $\alpha = 3$

Example 13.5. Assume home runs per game at Citizens Bank Park follow a Poisson distribution with parameter θ. Assume for θ a Gamma prior distribution with shape parameter $\alpha = 4$ and rate parameter $\lambda = 2$.

1. Write an expression for the prior density $\pi(\theta)$. Plot the prior distribution. Find the prior mean, prior SD, and prior 50%, 80%, and 98% credible intervals for θ.
2. Suppose a single game with 1 home run is observed. Write the likelihood function.
3. Write an expression for the posterior distribution of θ given a single game with 1 home run. Identify by the name the posterior distribution and the values of relevant parameters. Plot the prior distribution, (scaled) likelihood, and posterior distribution. Find the posterior mean, posterior SD, and posterior 50%, 80%, and 98% credible intervals for θ.
4. Now consider the original prior again. Determine the likelihood of observing 1 home run in game 1 and 3 home runs in game 2 in a sample of 2 games, and the posterior distribution of θ given this sample. Identify by the name the posterior distribution and the values of relevant parameters. Plot the prior distribution, (scaled) likelihood, and posterior distribution. Find the posterior mean, posterior SD, and posterior 50%, 80%, and 98% credible intervals for θ.
5. Consider the original prior again. Determine the likelihood of observing a total of 4 home runs in a sample of 2 games, and the posterior distribution of θ given this sample. Identify by the name the posterior distribution and

the values of relevant parameters. How does this compare to the previous part?

6. Consider the 2020 data in which there were 97 home runs in 32 games. Determine the likelihood function, and the posterior distribution of θ given this sample. Identify by the name the posterior distribution and the values of relevant parameters. Plot the prior distribution, (scaled) likelihood, and posterior distribution. Find the posterior mean, posterior SD, and posterior 50%, 80%, and 98% credible intervals for θ.
7. Interpret the credible interval from the previous part in context.
8. Express the posterior mean of θ based on the 2020 data as a weighted average of the prior mean and the sample mean.
9. While the main parameter is θ, there are other parameters of interest. For example, $\eta = e^{-\theta}$ is the population proportion of games in which there are 0 home runs. Assuming that you already have the posterior distribution of θ (or a simulation-based approximation), explain how you could use simulation to approximate the posterior distribution of η. Run the simulation and plot the posterior distribution, and find and interpret 50%, 80%, and 98% posterior credible intervals for η.
10. Use JAGS to approximate the posterior distribution of θ given this sample. Compare with the results from the previous example.

Solution. to Example 13.5

1. Remember that in the Gamma(4,2) prior distribution θ is treated as the variable.

$$\pi(\theta) \propto \theta^{4-1}e^{-2\theta}, \qquad \theta > 0.$$

This is the same prior we used in the grid approximation in Example 13.3. See below for a plot.

$$\begin{aligned} \text{Prior mean } &= \frac{\alpha}{\lambda} & \frac{4}{2} &= 2 \\ \text{Prior SD} &= \sqrt{\frac{\alpha}{\lambda^2}} & \sqrt{\frac{4}{2^2}} &= 1 \end{aligned}$$

Use `qgamma` for find the endpoints of the credible intervals.

```
qgamma(c(0.25, 0.75), shape = 4, rate = 2)
```

```
## [1] 1.268 2.555
```

```
qgamma(c(0.10, 0.90), shape = 4, rate = 2)
```

```
## [1] 0.8724 3.3404
```

```r
qgamma(c(0.01, 0.99), shape = 4, rate = 2)
```

```
## [1] 0.4116 5.0226
```

2. The likelihood is the Poisson probability of 1 home run in a game computed for each value of $\theta > 0$.

$$f(y = 1|\theta) = \frac{e^{-\theta}\theta^{1}}{1!} \propto e^{-\theta}\theta, \qquad \theta > 0.$$

3. Posterior is proportional to likelihood times prior

$$\begin{aligned} \pi(\theta|y=1) &\propto \left(e^{-\theta}\theta\right)\left(\theta^{4-1}e^{-2\theta}\right), \qquad \theta > 0, \\ &\propto \theta^{(4+1)-1}e^{-(2+1)\theta}, \qquad \theta > 0. \end{aligned}$$

We recognize the above as the Gamma density with shape parameter $\alpha = 4 + 1$ and rate parameter $\lambda = 2 + 1$.

$$\begin{aligned} \text{Posterior mean } &= \frac{\alpha}{\lambda} && \frac{5}{3} = 1.667 \\ \text{Posterior SD} &= \sqrt{\frac{\alpha}{\lambda^2}} && \sqrt{\frac{5}{3^2}} = 0.745 \end{aligned}$$

```r
qgamma(c(0.25, 0.75), shape = 4 + 1, rate = 2 + 1)
```

```
## [1] 1.123 2.091
```

```r
qgamma(c(0.10, 0.90), shape = 4 + 1, rate = 2 + 1)
```

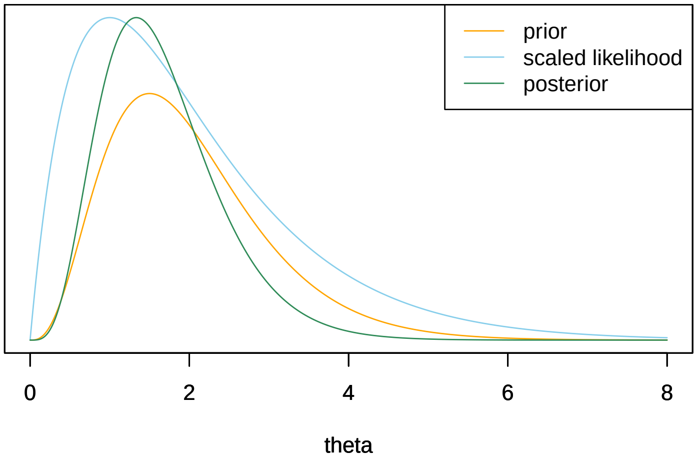
```
## [1] 0.8109 2.6645
```

```r
qgamma(c(0.01, 0.99), shape = 4 + 1, rate = 2 + 1)
```

```
## [1] 0.4264 3.8682
```

```r
theta = seq(0, 8, 0.001) # the grid is just for plotting

# prior
alpha = 4
lambda = 2
prior = dgamma(theta, shape = alpha, rate = lambda)
```

```
# likelihood
n = 1 # sample size
y = 1 # sample mean
likelihood = dpois(n * y, n * theta)

# posterior
posterior = dgamma(theta, alpha + n * y, lambda + n)

# plot
plot_continuous_posterior <- function(theta, prior, likelihood, posterior) {

  ymax = max(c(prior, posterior))

  scaled_likelihood = likelihood * ymax / max(likelihood)

  plot(theta, prior, type='l', col='orange', xlim= range(theta), ylim=c(0, ymax),
  par(new=T)
  plot(theta, scaled_likelihood, type='l', col='skyblue', xlim=range(theta), ylim=
  par(new=T)
  plot(theta, posterior, type='l', col='seagreen', xlim=range(theta), ylim=c(0, ym
  legend("topright", c("prior", "scaled likelihood", "posterior"), lty=1, col=c("c
}

plot_continuous_posterior(theta, prior, likelihood, posterior)
```



4. The likelihood is the product of the likelihoods of $y = 1$ and $y = 3$.

$$f(y = (1,3)|\theta) = \left(\frac{e^{-\theta}\theta^1}{1!}\right)\left(\frac{e^{-\theta}\theta^3}{3!}\right) \propto e^{-2\theta}\theta^4, \qquad \theta > 0.$$

The posterior satisfies

$$\begin{aligned}\pi(\theta|y = (1,3)) &\propto \left(e^{-2\theta}\theta^4\right)\left(\theta^{4-1}e^{-2\theta}\right), \qquad \theta > 0,\\ &\propto \theta^{(4+4)-1}e^{-(2+2)\theta}, \qquad \theta > 0.\end{aligned}$$

We recognize the above as the Gamma density with shape parameter $\alpha = 4 + 4$ and rate parameter $\lambda = 2 + 2$.

$$\begin{aligned}\text{Posterior mean } &= \frac{\alpha}{\lambda} & \frac{8}{4} &= 2\\ \text{Posterior SD} &= \sqrt{\frac{\alpha}{\lambda^2}} & \sqrt{\frac{8}{4^2}} &= 0.707\end{aligned}$$

```r
qgamma(c(0.25, 0.75), shape = 4 + 4, rate = 2 + 2)
```

```
## [1] 1.489 2.421
```

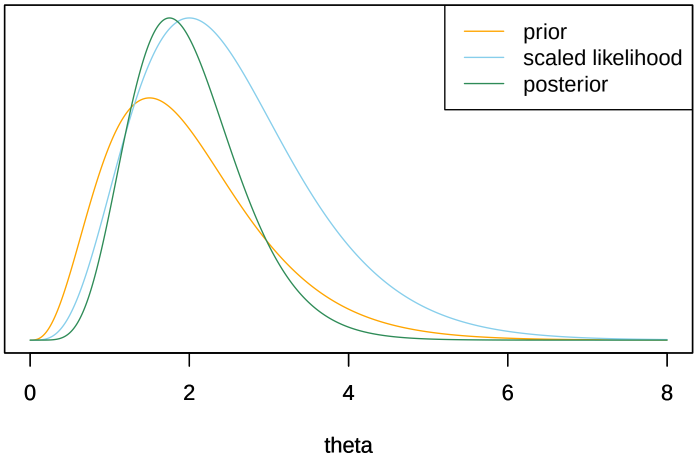
```r
qgamma(c(0.10, 0.90), shape = 4 + 4, rate = 2 + 2)
```

```
## [1] 1.164 2.943
```

```r
qgamma(c(0.01, 0.99), shape = 4 + 4, rate = 2 + 2)
```

```
## [1] 0.7265 4.0000
```

```r
n  = 2 # sample size
y = 2 # sample mean

# likelihood
likelihood = dpois(1, theta) * dpois(3, theta)

# posterior
posterior = dgamma(theta, alpha + n * y, lambda + n)

# plot
plot_continuous_posterior(theta, prior, likelihood, posterior)
```

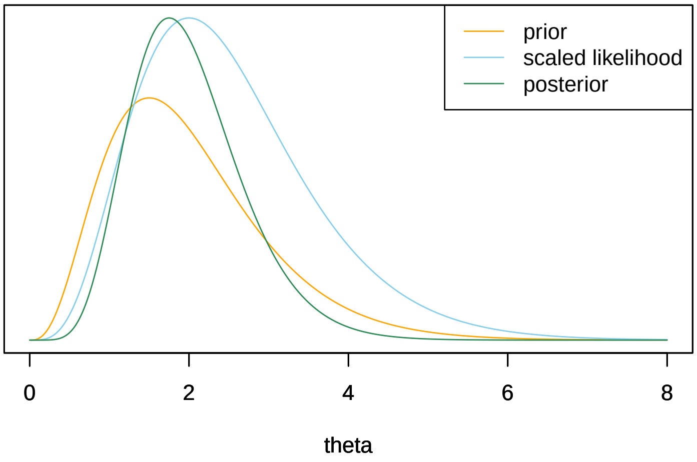


5. By Poisson aggregation, the total number of home runs in 2 games follows a Poisson(2θ) distribution. The likelihood is the probability of a value of 4 (home runs in 2 games) computed using a Poisson(2θ) for each value of θ.

$$f(\bar{y} = 2|\theta) = \frac{e^{-2\theta}(2\theta)^4}{4!} \propto e^{-2\theta}\theta^4, \qquad \theta > 0$$

The shape of the likelihood as a function of θ is the same as in the previous part; the likelihood functions are proportionally the same regardless of whether you observe the individual values or just the total count. Therefore, the posterior distribution is the same as in the previous part.

```
# likelihood
n = 2 # sample size
y = 2 # sample mean
likelihood = dpois(n * y, n * theta)

# posterior
posterior = dgamma(theta, alpha + n * y, lambda + n)

# plot
plot_continuous_posterior(theta, prior, likelihood, posterior)
```



6. By Poisson aggregation, the total number of home runs in 32 games follows a Poisson(32θ) distribution. The likelihood is the probability of observing a value of 97 (for the total number of home runs in 32 games) from a Poisson(32θ) distribution.

$$\begin{aligned} f(\bar{y} = 97/32|\theta) &= e^{-32\theta}(32\theta)^{97}/97!, \qquad \theta > 0 \\ &\propto e^{-32\theta}\theta^{97}, \qquad \theta > 0 \end{aligned}$$

The posterior satisfies

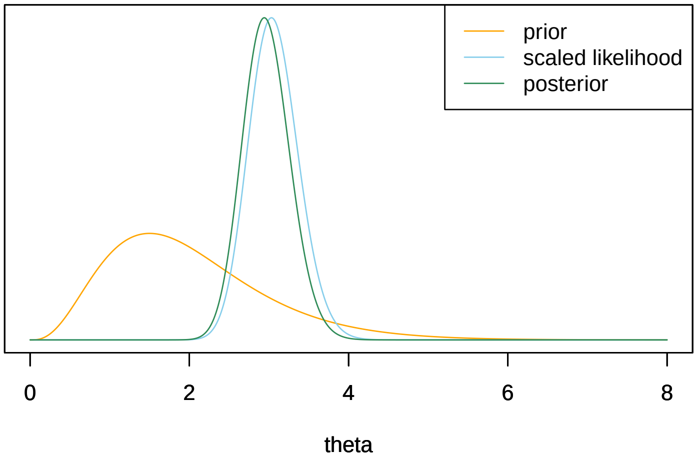
$$\begin{aligned} \pi(\theta|\bar{y} = 97/32) &\propto \left(e^{-32\theta}\theta^{97}\right)\left(\theta^{4-1}e^{-2\theta}\right), \qquad \theta > 0, \\ &\propto \theta^{(4+97)-1}e^{-(2+32)\theta}, \qquad \theta > 0. \end{aligned}$$

We recognize the above as the Gamma density with shape parameter $\alpha = 4 + 97$ and rate parameter $\lambda = 2 + 32$.

$$\begin{aligned} \text{Posterior mean } &= \frac{\alpha}{\lambda} \qquad \frac{101}{34} = 2.97 \\ \text{Posterior SD} &= \sqrt{\frac{\alpha}{\lambda^2}} \qquad \sqrt{\frac{101}{34^2}} = 0.296 \end{aligned}$$

The likelihood is centered at the sample mean of $97/32 = 3.03$. The posterior distribution follows the likelihood fairly closely, but the prior still has a little influence. The posterior is essentially identical to the one we computed via grid approximation in Example 13.3.

```
# likelihood
n = 32 # sample size
y = 97 / 32 # sample mean
likelihood = dpois(n * y, n * theta)

# posterior
posterior = dgamma(theta, alpha + n * y, lambda + n)

# plot
plot_continuous_posterior(theta, prior, likelihood, posterior)
```



```
qgamma(c(0.25, 0.75), alpha + n * y, lambda + n)
```

```
## [1] 2.766 3.164
```

```
qgamma(c(0.10, 0.90), alpha + n * y, lambda + n)
```

```
## [1] 2.599 3.355
```

```
qgamma(c(0.01, 0.99), alpha + n * y, lambda + n)
```

```
## [1] 2.326 3.701
```

7. The credible intervals represent conclusions about θ, the mean number of home runs per game at Citizen Bank Park.

   There is a posterior probability of 50% that the mean number of home runs per games at Citizen Bank Park is between 2.77 and 3.16. It is equally plausible that θ is inside this interval as outside.

   There is a posterior probability of 80% that the mean number of home runs per games at Citizen Bank Park is between 2.6 and 3.36. It is four times more plausible that θ is inside this interval than outside.

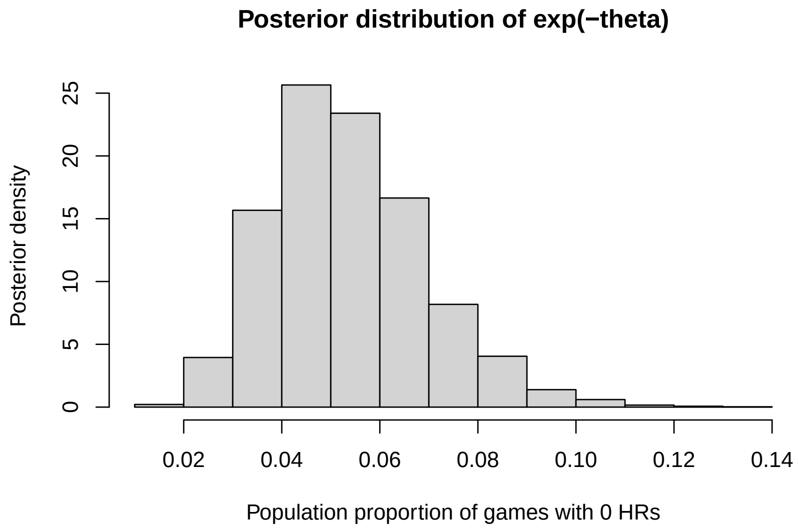
   There is a posterior probability of 98% that the mean number of home runs per games at Citizen Bank Park is between 2.33 and 3.7. It is 49 times more plausible that θ is inside this interval than outside.

8. The prior mean is 4/2=2, based on a "prior sample size" of 2. The sample mean is $97/32 = 3.03$, based on a sample size of 32. The posterior mean is $(4 + 97)/(2 + 32) = 2.97$. The posterior mean is a weighted average of the prior mean and the sample mean with the weights based on the "sample sizes"

$$2.97 = \frac{4+97}{2+32} = \left(\frac{2}{2+32}\right)\left(\frac{4}{2}\right)+\left(\frac{32}{2+32}\right)\left(\frac{97}{32}\right) = (0.0589)(2)+(0.941)(3.03)$$

9. Simulate a value of θ from its posterior distribution, compute $\eta = e^{-\theta}$, repeat many times, and summarize the simulated values of η. We can use the `quantile` function to find the endpoints of credible intervals.

```
theta_sim = rgamma(10000, alpha + n * y, lambda + n)

eta_sim = exp(-theta_sim)

hist(eta_sim, freq = FALSE,
 xlab = "Population proportion of games with 0 HRs",
 ylab = "Posterior density",
 main = "Posterior distribution of exp(-theta)")
```



```
quantile(eta_sim, c(0.25, 0.75))
```

```
##     25%     75% 
## 0.04220 0.06315
```

```
quantile(eta_sim, c(0.10, 0.90))
```

```
##     10%     90% 
## 0.03481 0.07455
```

```
quantile(eta_sim, c(0.01, 0.99))
```

```
##      1%     99% 
## 0.02488 0.09849
```

There is a posterior probability of 98% that the population proportion of games with 0 home runs is between 0.025 and 0.098.

10. The JAGS code is below. The results are very similar to the theoretical results from previous parts.

Here is the JAGS code. Note

- The data has been loaded as individual values, number of home runs in each of the 32 games

- Likelihood is defined as a loop. For each `y[i]` value, the likelihood is computing according to a Poisson(θ) distribution
- Prior distribution is a Gamma distribution. (Remember, JAGS syntax for `dgamma`, `dpois`, etc, is not the same as in R.)

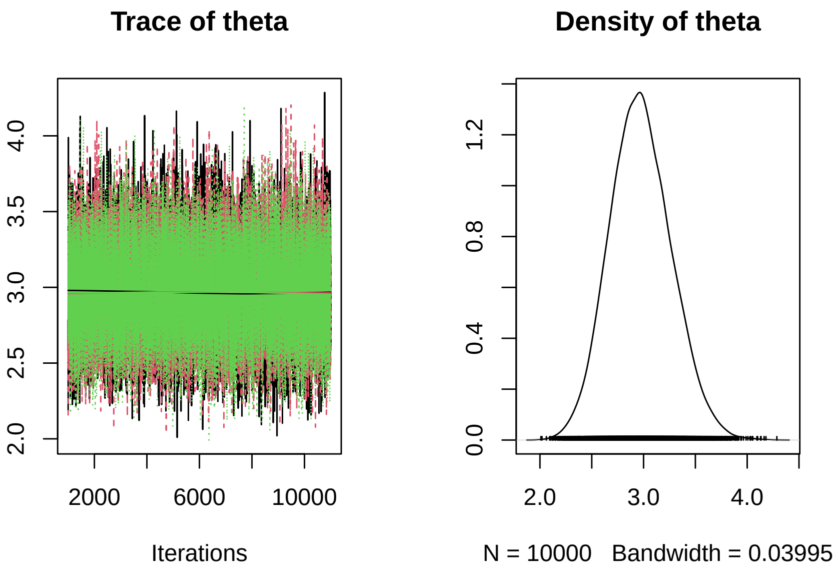
```
# data
df = read.csv("_data/citizens-bank-hr-2020.csv")
y = df$hr
n = length(y)

# model
model_string <- "model{

  # Likelihood
  for (i in 1:n){
    y[i] ~ dpois(theta)
  }

  # Prior
  theta ~ dgamma(alpha, lambda)
  alpha <- 4
  lambda <- 2

}"

# Compile the model
dataList = list(y=y, n=n)

Nrep = 10000
Nchains = 3

model <- jags.model(textConnection(model_string),
                    data=dataList,
                    n.chains=Nchains)
```

```
## Compiling model graph
##    Resolving undeclared variables
##    Allocating nodes
## Graph information:
##    Observed stochastic nodes: 32
##    Unobserved stochastic nodes: 1
##    Total graph size: 36
##
## Initializing model
```

```
update(model, 1000, progress.bar="none")

posterior_sample <- coda.samples(model,
                                 variable.names=c("theta"),
                                 n.iter=Nrep,
                                 progress.bar="none")

# Summarize and check diagnostics
summary(posterior_sample)
```

```
##
## Iterations = 1001:11000
## Thinning interval = 1
## Number of chains = 3
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##           Mean             SD       Naive SE Time-series SE
##        2.97290        0.29624        0.00171        0.00173
##
## 2. Quantiles for each variable:
##
##  2.5%   25%   50%   75% 97.5%
##  2.42  2.77  2.96  3.17  3.58
```

```
plot(posterior_sample)
```



In the previous example we saw that if the values of the measured variable follow a Poisson distribution with parameter θ and the prior for θ follows a Gamma distribution, then the posterior distribution for θ given the data also follows a Gamma distribution.

Gamma-Poisson model.[6] Consider a measured variable Y which, given θ, follows a Poisson(θ) distribution. Let $\bar{y}$ be the sample mean for a random sample of size n. Suppose θ has a Gamma(α, λ) prior distribution. Then the posterior distribution of θ given $\bar{y}$ is the Gamma($\alpha + n\bar{y}, \lambda + n$) distribution.

That is, Gamma distributions form a *conjugate prior* family for a Poisson likelihood.

The posterior distribution is a compromise between prior and likelihood. For the Gamma-Poisson model, there is an intuitive interpretation of this compromise. In a sense, you can interpret α as "prior total count" and λ as "prior sample size", but these are only "pseudo-observations". Also, α and λ are not necessarily integers.

Note that if $\bar{y}$ is the sample mean count is then $n\bar{y} = \sum_{i=1}^{n} y_i$ is the sample total count.

[6] I've been naming these models in the form "Prior-Likelihood", e.g. Gamma prior and Poisson likelihood. I would rather do it as "Likelihood-Prior". In modeling, the likelihood comes first; what is an appropriate distributional model for the observed data? This likelihood depends on some parameters, and then a prior distribution is placed on these parameters. So in modeling the order is likelihood then prior, and it would be nice if the names followed that pattern. But "Beta-Binomial" is the canonical example, and no one calls that "Binomial-Beta". To be consistent, we'll stick with the "Prior-Likelihood" naming convention.

	Prior	Data	Posterior
Total count	α	$n\bar{y}$	$\alpha + n\bar{y}$
Sample size	λ	n	$\lambda + n$
Mean	$\frac{\alpha}{\lambda}$	$\bar{y}$	$\frac{\alpha+n\bar{y}}{\lambda+n}$

- The posterior total count is the sum of the "prior total count" α and the sample total count $n\bar{y}$.
- The posterior sample size is the sum of the "prior sample size" λ and the observed sample size n.
- The posterior mean is a weighted average of the prior mean and the sample mean, with weights proportional to the "sample sizes".

$$\frac{\alpha + n\bar{y}}{\lambda + n} = \frac{\lambda}{\lambda + n}\left(\frac{\alpha}{\lambda}\right) + \frac{n}{\lambda + n}\bar{y}$$

- As more data are collected, more weight is given to the sample mean (and less weight to the prior mean)
- Larger values of λ indicate stronger prior beliefs, due to smaller prior variance (and larger "prior sample size"), and give more weight to the prior mean

Try this applet which illustrates the Gamma-Poisson model.

Rather than specifying α and β, a Gamma distribution prior can be specified by its prior mean and SD directly. If the prior mean is μ and the prior SD is σ, then

$$\lambda = \frac{\mu}{\sigma^2}$$
$$\alpha = \mu\lambda$$

Example 13.6. Continuing the previous example, assume home runs per game at Citizens Bank Park follow a Poisson distribution with parameter θ. Assume for θ a Gamma prior distribution with shape parameter $\alpha = 4$ and rate parameter $\lambda = 2$. Consider the 2020 data in which there were 97 home runs in 32 games.

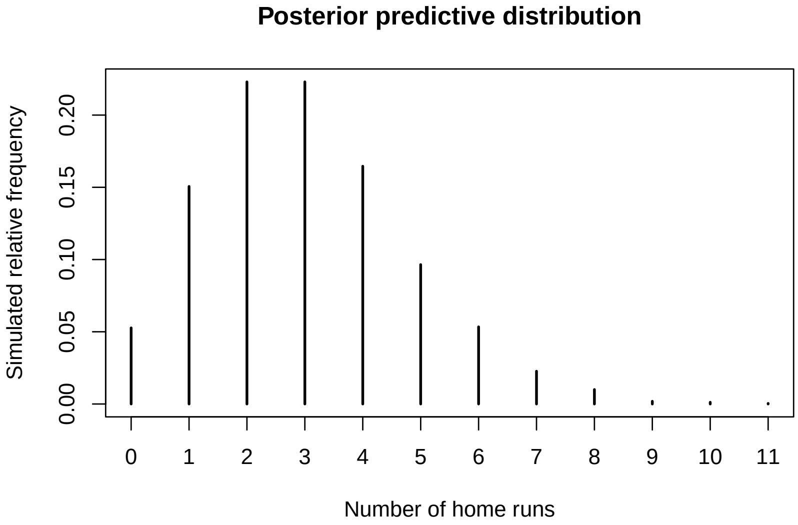
1. How could you use simulation (not JAGS) to approximate the posterior predictive distribution of home runs in a game?
2. Use the simulation from the previous part to find and interpret a 95% posterior prediction interval with a lower bound of 0.
3. Is a Poisson model a reasonable model for the data? How could you use posterior predictive simulation to simulate what a sample of 32 games might look like under this model. Simulate many such samples. Does the observed sample seem consistent with the model?

4. Regarding the appropriateness of a Poisson model, we might be concerned that there are no games in the sample with 0 home runs. Use simulation to approximate the posterior predictive distribution of the number of games in a sample of 32 with 0 home runs. From this perspective, does the observed value of the statistic seem consistent with the Gamma-Poisson model?

Solution. to Example 13.6

1. Simulate a value of θ from its Gamma(101, 34) posterior distribution, then given θ simulate a value of y from a Poisson(θ) distribution. Repeat many times and summarize the y values to approximate the posterior predictive distribution.

```r
Nrep = 10000
theta_sim = rgamma(Nrep, 101, 34)

y_sim = rpois(Nrep, theta_sim)

plot(table(y_sim) / Nrep, type = "h",
 xlab = "Number of home runs",
 ylab = "Simulated relative frequency",
 main = "Posterior predictive distribution")
```



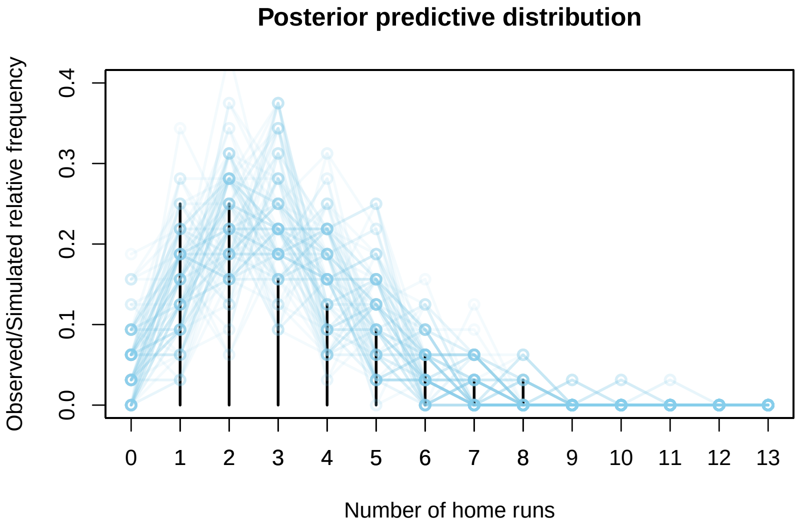
2. There is a posterior predictive probability of 95% of between 0 and 6 home runs in a game. Very roughly, about 95% of games have between 0 and 6 home runs.

```
quantile(y_sim, 0.95)
```

```
## 95%
##   6
```

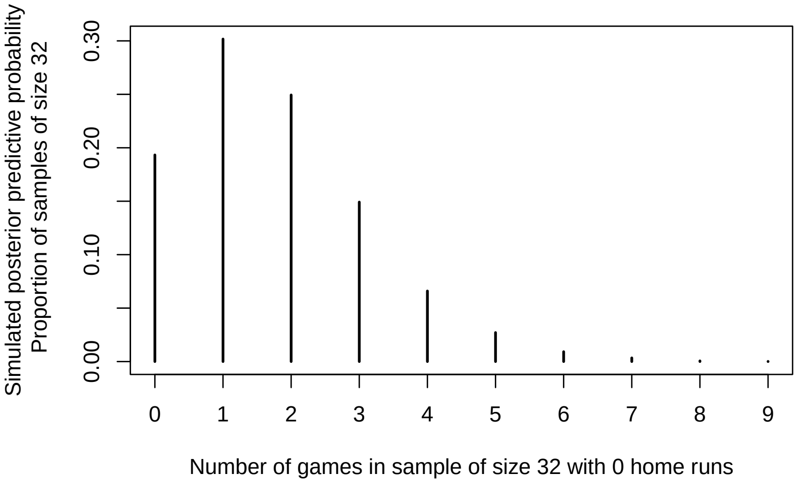
3. Simulate a value of θ from its Gamma(101, 34) posterior distribution, then given θ simulate 32 values of y from a Poisson(θ) distribution. Summarize each sample. Repeat many times to simulate many samples of size 32. Compare the observed sample with the simulated samples. Aside from the fact there the sample has no games with 0 home runs, the model seems reasonable.

```
df = read.csv("_data/citizens-bank-hr-2020.csv")
y = df$hr
n = length(y)

plot(table(y) / n, type = "h", xlim = c(0, 13), ylim = c(0, 0.4),
 xlab = "Number of home runs",
 ylab = "Observed/Simulated relative frequency",
 main = "Posterior predictive distribution")
axis(1, 0:13)

n_samples = 100


# simulate samples
for (r in 1:n_samples){

  # simulate theta from posterior distribution
  theta_sim = rgamma(1, 101, 34)

  # simulate values from Poisson(theta) distribution
  y_sim = rpois(n, theta_sim)

  # add plot of simulated sample to histogram
  par(new = T)
  plot(table(factor(y_sim, levels = 0:13)) / n, type = "o", xlim = c(0, 13), ylim
  xlab = "", ylab = "", xaxt='n', yaxt='n',
    col = rgb(135, 206, 235, max = 255, alpha = 25))
}
```



4. Continuing with the simulation from the previous part, now for each simulated sample we record the number of games with 0 home runs. Each "dot" in the plot below represents a sample of size 32 for which we measure the number of games in the sample with 0 home runs. While we see that it's less likely to have 0 home runs in 32 games than not, it would not be too surprising to see 0 home runs in a sample of 32 games. Therefore, the fact that there are 0 home runs in the observed sample alone does not invalidate the model.

```
n_samples = 10000

zero_count = rep(NA, n_samples)

# simulate samples
for (r in 1:n_samples){

  # simulate theta from posterior distribution
  theta_sim = rgamma(1, 101, 34)

  # simulate values from Poisson(theta) distribution
  y_sim = rpois(n, theta_sim)
  zero_count[r] = sum(y_sim == 0)
}

par(mar = c(5, 5, 4, 2) + 0.1)
plot(table(zero_count) / n_samples, type = "h",
  xlab = "Number of games in sample of size 32 with 0 home runs",
  ylab = "Simulated posterior predictive probability\n Proportion of samples of si
```



Chapter 14

Introduction to Multi-Parameter Models

So far we have considered situations with a just a single unknown parameter θ. However, most interesting problems involve multiple unknown parameters.

For example, we have considered the problem of estimating the population mean of a numerical variable assuming the population standard deviation was known. However, in practice both the population mean and population standard deviation are unknown. Even if we are only interested in estimating the population mean, we still need to account for the uncertainty in the population standard deviation.

When there are two (or more) unknown parameters the prior and posterior distribution will each be a *joint* probability distribution over *pairs* (or tuples/vectors) of possible values of the parameters.

Example 14.1. Assume body temperatures (degrees Fahrenheit) of healthy adults follow a Normal distribution with unknown mean μ and unknown standard deviation σ. Suppose we wish to estimate both μ, the population mean healthy human body temperature, and σ, the population standard deviation of body temperatures.

1. Assume a discrete prior distribution according to which

    - μ takes values 97.6, 98.1, 98.6 with prior probability 0.2, 0.3, 0.5, respectively.
    - σ takes values 0.5, 1 with prior probability 0.25, 0.75, respectively.
    - μ and σ are independent.

    Start to construct the Bayes table. What are the possible values of the parameter? What are the prior probabilities? (Hint: the parameter θ is a *pair* (μ, σ).)

2. Suppose two temperatures of 97.5 and 97.9 are observed, independently. Identify the likelihood.

3. Complete the Bayes table and find the posterior distribution after observing these two measurements. Compare to the prior distribution.

4. Suppose that we only observe that in a sample of size 2 the mean is 97.7. Is this information enough to evaluate the likelihood function and determine the posterior distribution?

5. The prior assumes that μ and σ are independent. Are they independent according to the posterior distribution?

Solution. to Example 14.1

1. See the table below. There are 3 possible values for μ and 2 possible values for σ so there are $(3)(2) = 6$ possible (μ, σ) pairs. Each row in the Bayes table represents a (μ, σ) pair. Since the prior assumes independence, the prior probability of any pair is the product of the marginal prior probabilities of μ and σ. For example, the probability probability that $\mu = 97.6$ and $\sigma = 0.5$ is $(0.2)(0.25) = 0.05$

2. The likelihood is similar to what we have seen in other examples concerning body temperature, but it is now a function of both μ and σ. That is, the likelihood is a function of two variables. The likelihood is determined by evaluating, for each (μ, σ) pair, the Normal(μ, σ) density at each of $y = 97.9$ and $y = 97.5$ and then finding the product:

$$f(y=(97.9,97.5)|\mu,\sigma) \propto \left[\sigma^{-1}\exp\left(-\frac{1}{2}\left(\frac{97.9-\mu}{\sigma}\right)^2\right)\right]\left[\sigma^{-1}\exp\left(-\frac{1}{2}\left(\frac{97.5-\mu}{\sigma}\right)^2\right)\right]$$

3. See the table below. As always, posterior is proportional to likelihood times prior. For the sample (97.9, 97.5), the sample mean is 97.7 and the sample standard deviation is 0.283. The posterior distribution pushes probability away from $\mu = 98.6$, and pushes more probability towards $\sigma = 0.5$.

```r
mu = c(97.6, 98.1, 98.6)
sigma = c(0.5, 1)
theta = expand.grid(mu, sigma) # all possible (mu, sigma) pairs
names(theta) = c("mu", "sigma")

# prior

prior_mu = c(0.20, 0.30, 0.50)
prior_sigma = c(0.25, 0.75)
prior = apply(expand.grid(prior_mu, prior_sigma), 1, prod)
prior = prior / sum(prior)
```

```
# data
y = c(97.9, 97.5) # single observed value

# likelihood
likelihood = dnorm(97.9, mean = theta$mu, sd = theta$sigma) *
  dnorm(97.5, mean = theta$mu, sd = theta$sigma)

# posterior
product = likelihood * prior
posterior = product / sum(product)

# bayes table
bayes_table = data.frame(theta,
                         prior,
                         likelihood,
                         product,
                         posterior)

kable(bayes_table, digits = 4, align = 'r')
```

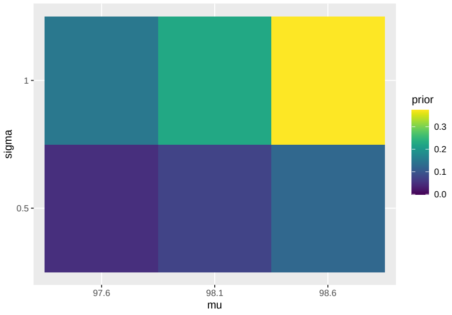
mu	sigma	prior	likelihood	product	posterior
97.6	0.5	0.050	0.5212	0.0261	0.2041
98.1	0.5	0.075	0.2861	0.0215	0.1680
98.6	0.5	0.125	0.0212	0.0027	0.0208
97.6	1.0	0.150	0.1514	0.0227	0.1778
98.1	1.0	0.225	0.1303	0.0293	0.2296
98.6	1.0	0.375	0.0680	0.0255	0.1997

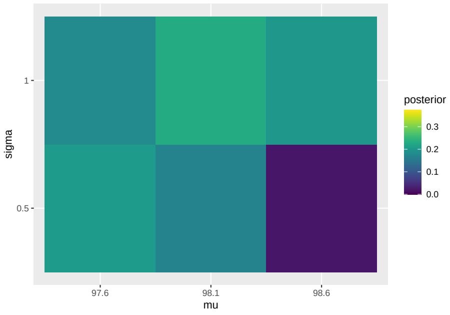
4. Intuitively, knowing only the posterior mean would not be sufficient, since it would not give us enough information to estimate the standard deviation σ. In order to evaluate the likelihood we need to compute $\frac{y-\mu}{\sigma}$ for each individual y value, so if we only had the sample mean we would not be able to fill in the likelihood column.

5. The posterior distribution represents some dependence between μ and σ. For example, consider the pair $\mu = 97.6$ and $\sigma = 0.5$. The marginal posterior probability that $\mu = 97.6$ is 0.3819. The marginal posterior probability that $\sigma = 0.5$ is 0.3929. But the joint posterior probability that $\mu = 97.6$ and $\sigma = 0.5$ is 0.2041, which is not the product of the marginal probabilities.

The plots below compare the prior and posterior distributions from the previous problem.

```
ggplot(bayes_table %>%
         mutate(mu = factor(mu),
                sigma = factor(sigma)),
       aes(mu, sigma)) +
    geom_tile(aes(fill = prior))  +
  scale_fill_viridis(limits = c(0, max(c(prior, posterior))))

ggplot(bayes_table %>%
         mutate(mu = factor(mu),
                sigma = factor(sigma)),
       aes(mu, sigma)) +
    geom_tile(aes(fill = posterior)) +
  scale_fill_viridis(limits = c(0, max(c(prior, posterior))))
```





Example 14.2. Continuing Example 14.1, let's assume a more reasonable, continuous prior for (μ, σ). We have seen that we often work with the precision $\tau = 1/\sigma^2$ rather than the SD. Assume a continuous prior distribution which assumes

- μ has a Normal distribution with mean 98.6 and standard deviation 0.3.
- τ has a Gamma distribution with shape parameter 5 and rate parameter 2.
- μ and τ are independent.

(This problem just concerns the prior distribution. We'll look at this posterior distribution in the next example.)

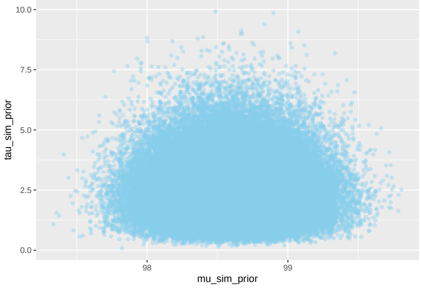
1. Simulate (μ, τ) pairs from the prior distribution and plot them.
2. Simulate (μ, σ) pairs from the prior distribution and plot them. Describe the prior distribution of σ.
3. Find and interpret a central 98% prior credible interval for μ.
4. Find a central 98% prior credible interval for the precision $\tau = 1/\sigma^2$.
5. Find and interpret a central 98% prior credible interval for σ.

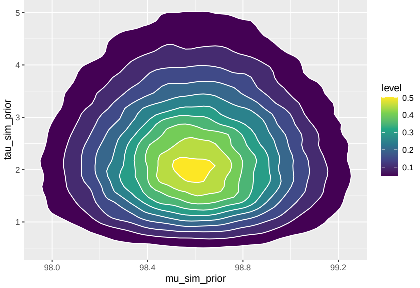
6. What is the prior credibility that both μ and σ lie within their credible intervals?

Solution. to Example 14.2

1. We could plot the prior distribution directly. However, distributions are usually only approximated via simulation, so we'll just simulate. The prior distribution is a distribution on (μ, τ) pairs.

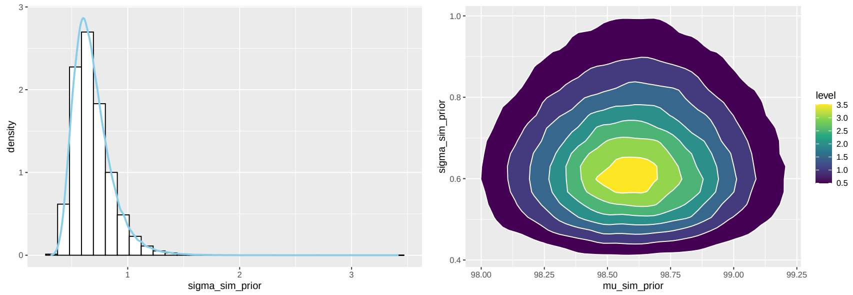
```
Nrep = 100000

mu_sim_prior = rnorm(Nrep, 98.6, 0.3)
tau_sim_prior = rgamma(Nrep, shape = 5, rate = 2)
sigma_sim_prior = 1 / sqrt(tau_sim_prior)
sim_prior = data.frame(mu_sim_prior, tau_sim_prior, sigma_sim_prior)

ggplot(sim_prior, aes(mu_sim_prior, tau_sim_prior)) +
  geom_point(color = "skyblue", alpha = 0.4)

ggplot(sim_prior, aes(mu_sim_prior, tau_sim_prior)) +
  stat_density_2d(aes(fill = ..level..),
              geom = "polygon", color = "white") +
 scale_fill_viridis_c()
```





2. See plots below. The prior distribution on τ induces a prior distribution on $\sigma = 1/\sqrt{\tau}$.

```
ggplot(sim_prior, aes(x = sigma_sim_prior)) +
 geom_histogram(aes(y=..density..), color = "black", fill = "white") +
 geom_density(size = 1, color = "skyblue")

ggplot(sim_prior, aes(mu_sim_prior, sigma_sim_prior)) +
  stat_density_2d(aes(fill = ..level..),
              geom = "polygon", color = "white") +
 scale_fill_viridis_c()
```



3. There is a prior probability of 98% that population mean human body temperature is between 97.9 and 99.3 degrees F.

```
quantile(mu_sim_prior, c(0.01, 0.99))
```

```
##   1%  99%
## 97.9 99.3
```

4. We can compute a credible interval like usual. Precision just doesn't have as practical an interpretation as standard deviation.

```
quantile(tau_sim_prior, c(0.01, 0.99))
```

```
##     1%    99%
## 0.6374 5.8224
```

5. There is a prior probability of 98% that population standard deviation of human body temperatures is between 0.41 and 1.25 degrees F.

```
quantile(sigma_sim_prior, c(0.01, 0.99))
```

```
##     1%    99%
## 0.4144 1.2526
```

6. Since μ and σ are independent according to the prior distribution, the probability that both parameters lie in their respective intervals is (0.98)(0.98)=0.9604. If we want 98% joint prior credibility, we need a different region.

Example 14.3. Continuing the previous example, we'll now compute the posterior distribution given a sample of two measurements of 97.9 and 97.5.

1. Assume a grid of μ values from 96.0 to 100.0 in increments of 0.01, and a grid of τ values from 0.1 to 25.0 in increments of 0.01. How many possible values of the pair (μ, τ) are there; that is, how many rows are there in the Bayes table?

2. Use grid approximation to approximate the joint posterior distribution of (μ, τ) Simulate values from the joint posterior distribution and plot them. Compute the posterior correlation between μ and τ; are they independent according to the posterior distribution?
3. Plot the simulated joint posterior distribution of μ and σ. Compare to the prior.
4. Suppose we wanted to approximate the posterior distribution without first using grid approximation. Describe how, in principle, you would use a naive (non-MCMC) simulation to approximate the posterior distribution. In practice, what is the problem with such a simulation?

Solution. to Example 14.3

1. There are (100-96)/0.01 = 400 values of μ in the grid (actually 401 including both endpoints) and (25-0.1)/0.01 = 2490 values of μ in the grid (actually 2491). There are almost 1 million possible values of the pair (μ, τ) in the grid.

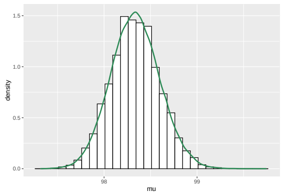
2. See below. Even though μ and τ are independent according to the prior distribution, there is a negative posterior correlation. (Below the posterior is computed via grid approximation. After the posterior distribution was computed, values were simulated from it for plotting.)

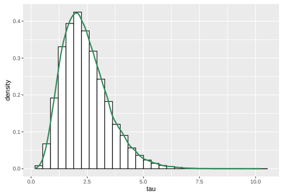
```
# parameters
mu = seq(96.0, 100.0, 0.01)
tau = seq(0.1, 25, 0.01)

theta = expand.grid(mu, tau)
names(theta) = c("mu", "tau")
theta$sigma = 1 / sqrt(theta$tau)

# prior
prior_mu_mean = 98.6
prior_mu_sd = 0.3

prior_precision_shape = 5
prior_precision_rate = 2

prior = dnorm(theta$mu, prior_mu_mean, sd = prior_mu_sd) *
  dgamma(theta$tau, shape = prior_precision_shape,
                    rate = prior_precision_rate)
prior = prior / sum(prior)

# data
y = c(97.9, 97.5)
```

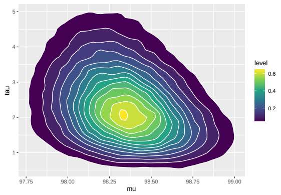
```
# likelihood
likelihood = dnorm(97.9, mean = theta$mu, sd = theta$sigma) *
  dnorm(97.5, mean = theta$mu, sd = theta$sigma)

# posterior
product = likelihood * prior
posterior = product / sum(product)

# posterior simulation
sim_posterior = theta[sample(1:nrow(theta), 100000, replace = TRUE, prob = posteri

cor(sim_posterior$mu, sim_posterior$tau)
```

```
## [1] -0.2888
```

```
#plots

ggplot(sim_posterior, aes(mu)) +
 geom_histogram(aes(y=..density..), color = "black", fill = "white") +
 geom_density(size = 1, color = "seagreen")

ggplot(sim_posterior, aes(tau)) +
 geom_histogram(aes(y=..density..), color = "black", fill = "white") +
 geom_density(size = 1, color = "seagreen")

ggplot(sim_posterior, aes(mu, tau)) +
  stat_density_2d(aes(fill = ..level..),
                geom = "polygon", color = "white") +
 scale_fill_viridis_c()
```





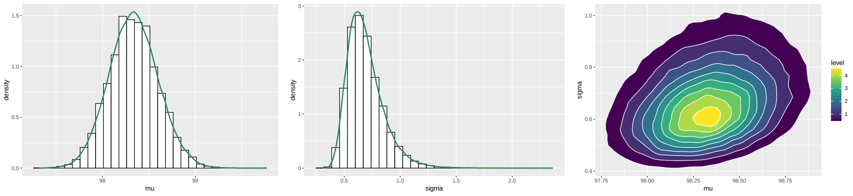


3. See below. We see that the posterior shifts the density towards smaller values of μ and σ. There is also a slight positive posterior correlation between μ and σ.

```
cor(sim_posterior$mu, sim_posterior$sigma)
```

```
## [1] 0.2804
```

```
#plots

ggplot(sim_posterior, aes(mu)) +
 geom_histogram(aes(y=..density..), color = "black", fill = "white") +
 geom_density(size = 1, color = "seagreen")

ggplot(sim_posterior, aes(sigma)) +
 geom_histogram(aes(y=..density..), color = "black", fill = "white") +
 geom_density(size = 1, color = "seagreen")

ggplot(sim_posterior, aes(mu, sigma)) +
  stat_density_2d(aes(fill = ..level..),
              geom = "polygon", color = "white") +
 scale_fill_viridis_c()
```



4. Simulate a value of (μ, σ) from their joint prior distribution, by simulating a value of μ from a Normal(98.6, 0.3) distribution and simulating, independently, a value of τ from a Gamma(5, 2) distribution and setting $\sigma = 1/\sqrt{\tau}$. Given μ and σ simulate two independent y values from a Normal(μ, σ) distribution. Repeat many times. Condition on the observed data by discarding any repetitions for which the y values are not (97.9, 97.5), to some reasonable degree of precision, say rounded to 1 decimal place. Approximate the posterior distribution using the remaining simulated values of (μ, σ).

   In practice, the probability of seeing a sample with y values of 97.9 and 97.5 is extremely small, so almost all repetitions of the simulation would be discarded and such a simulation would be extremely computationally inefficient. (For example, the values of μ and σ which maximize the likelihood of (97.9, 97.5) are 97.7 and 0.2, respectively, and even for those values and rounding to 1 decimal place the probability of seeing such a sample is only 0.015.)

The previous problem illustrates that grid approximation can quickly become computationally infeasible when there are multiple parameters (to obtain sufficient precision). Naively conditioning a simulation on the observed sample is also computationally infeasible, since except in the simplest situations the probability of recreating the observed sample in a simulation is essentially 0.

Therefore, we need more efficient computational methods, and MCMC will do the trick.

Example 14.4. The `temperature` data file contains 208 measurements of human body temperature (degrees F). The sample mean is 97.71 degrees F and the sample SD is 0.75 degrees F. Assuming the same prior distribution as in the previous problem, use JAGS to approximate the joint posterior distribution of μ and σ. Summarize the posterior distribution in context.

Solution. to Example 14.4

The JAGS code is below. A few comments on the code

- The data is read as individual values, so the likelihood of each `y[i]` is computed via a for loop.
- We have called the parameters by their names `mu`, `tau`, `sigma`, rather than just a single `theta`.
- We specify a prior distribution on `tau` and then define `sigma <- 1 / sqrt(tau)`.
- In JAGS `dnorm` is of the form `dnorm(mean, precision)`
- We are interested in the posterior distribution of μ and σ, so we include both parameters in the `model.names` argument of the `coda.samples` function.
- The output of `coda.samples` is a special object called an `mcmc.list`. Calling `plot` on this object produces a trace plot and a density plot for each parameter included in `variable.names`. But it does not automatically produce any joint distribution plots.
- We use `mcmc_scatter` from the `bayesplot` package to create a scatter plot of the joint posterior distribution, to which we can add contours.
- We can also extract the JAGS output as a matrix, put it in a data frame, and then use R or ggplot commands to create plots.
- The simulated values from an `mcmc.list` can be extracted as a matrix with `as.matrix` and then manipulated as usual, e.g., to compute the correlation.

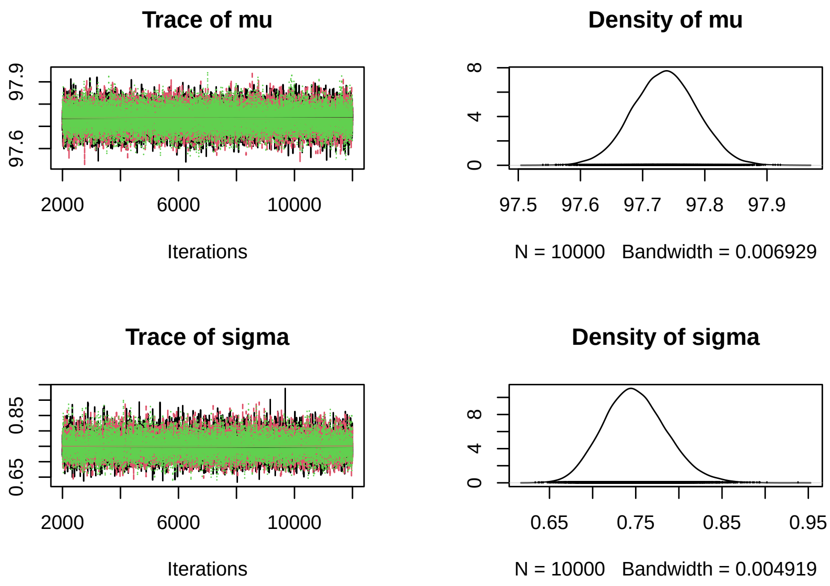
```
Nrep = 10000
Nchains = 3

# data
data = read.csv("_data/temperature.csv")
y = data$temperature
n = length(y)

# model
model_string <- "model{
  # Likelihood
  for (i in 1:n){
    y[i] ~ dnorm(mu, 1 / sigma ^ 2)
  }

  # Prior
  mu ~ dnorm(98.6, 1 / 0.3 ^ 2)

  sigma <- 1 / sqrt(tau)
  tau ~ dgamma(5, 2)

}"

# Compile the model
dataList = list(y=y, n=n)

model <- jags.model(textConnection(model_string),
                    data=dataList,
                    n.chains=Nchains)
```

```
## Compiling model graph
##    Resolving undeclared variables
##    Allocating nodes
## Graph information:
##    Observed stochastic nodes: 208
##    Unobserved stochastic nodes: 2
##    Total graph size: 222
##
## Initializing model
```

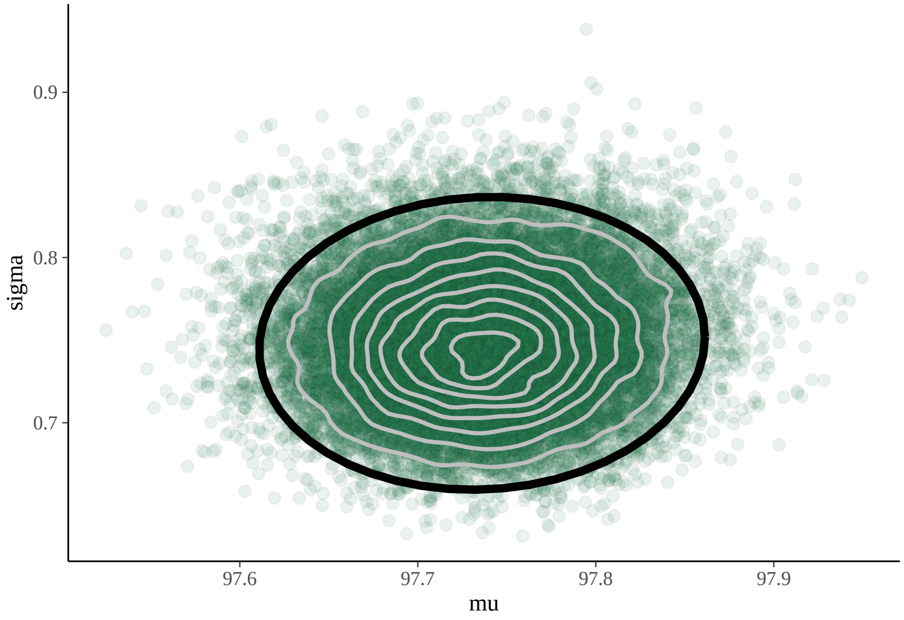
```
update(model, 1000, progress.bar="none")

posterior_sample <- coda.samples(model,
                                 variable.names=c("mu", "sigma"),
                                 n.iter=Nrep,
                                 progress.bar="none")


# Summarize and check diagnostics
summary(posterior_sample)
```

```
##
## Iterations = 2001:12000
## Thinning interval = 1
## Number of chains = 3
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##          Mean     SD Naive SE Time-series SE
## mu    97.736 0.0514 0.000297       0.000297
## sigma  0.749 0.0365 0.000211       0.000269
##
## 2. Quantiles for each variable:
##
##          2.5%    25%    50%    75%  97.5%
## mu    97.635 97.701 97.736 97.771 97.836
## sigma  0.682  0.724  0.748  0.773  0.825
```

```
plot(posterior_sample)
```



```
# Scatterplot from bayesplot package
color_scheme_set("green")
mcmc_scatter(posterior_sample, pars = c("mu", "sigma"), alpha = 0.1) +
  stat_ellipse(level = 0.98, color = "black", size = 2) +
  stat_density_2d(color = "grey", size = 1)
```



```
# posterior summary
posterior_sim = data.frame(as.matrix(posterior_sample))

head(posterior_sim)
```

```
##      mu  sigma
## 1 97.74 0.7210
## 2 97.68 0.7885
## 3 97.77 0.7331
## 4 97.72 0.7154
## 5 97.71 0.7449
## 6 97.74 0.7251
```

```
apply(posterior_sim, 2, mean)
```

```
##      mu   sigma
## 97.7362  0.7493
```

```
apply(posterior_sim, 2, sd)
```

```
##      mu   sigma
## 0.05138 0.03647
```

```
quantile(posterior_sim$mu, c(0.01, 0.99))
```

```
##    1%   99%
## 97.62 97.86
```

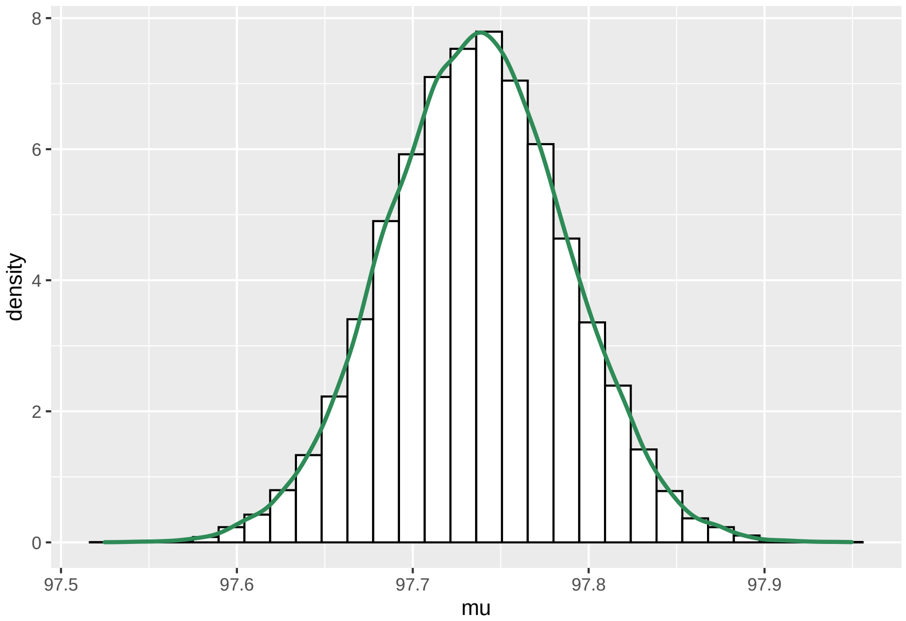
```
quantile(posterior_sim$sigma, c(0.01, 0.99))
```

```
##     1%    99%
## 0.6717 0.8423
```

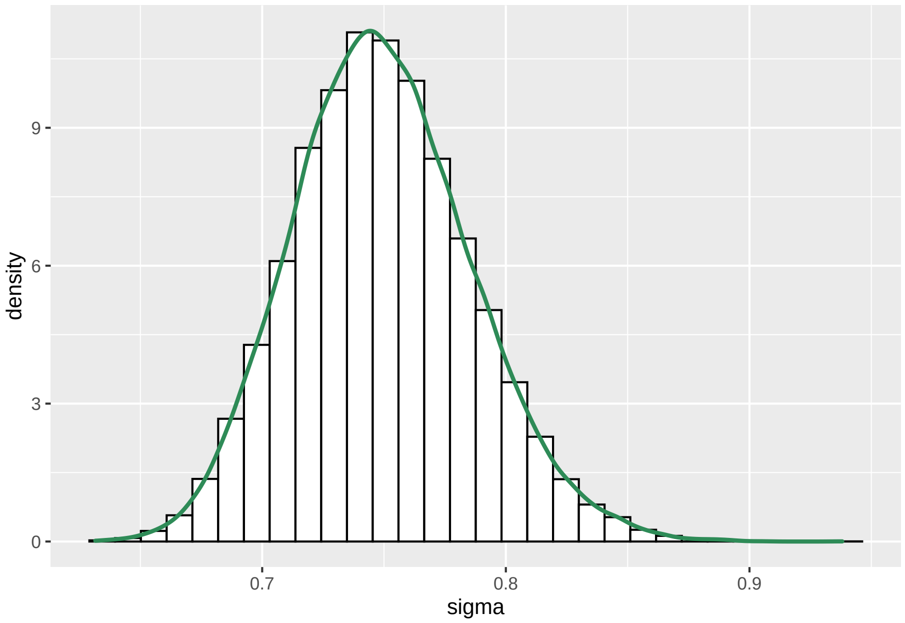
```
cor(posterior_sim)
```

```
##            mu   sigma
## mu    1.00000 0.04211
## sigma 0.04211 1.00000
```

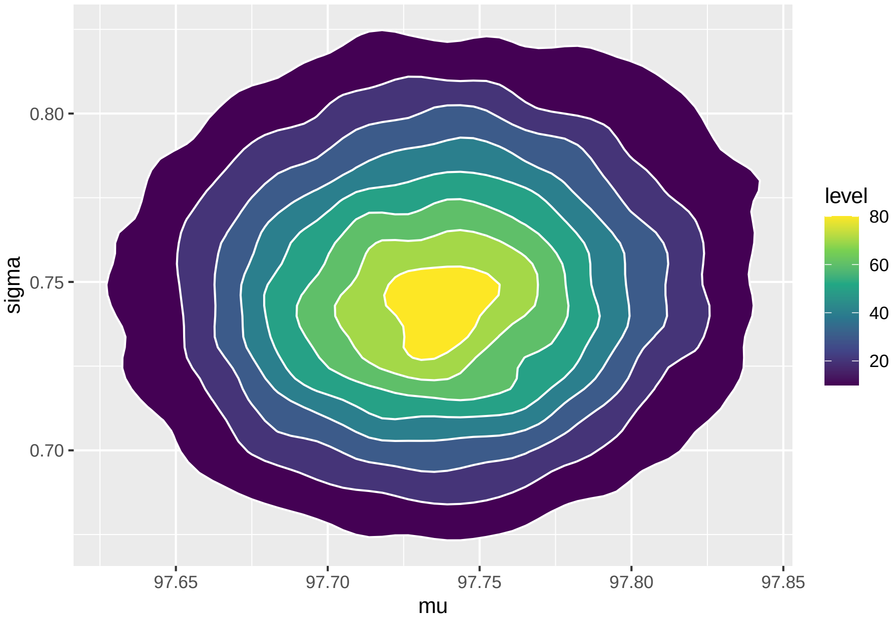
```
ggplot(posterior_sim, aes(mu)) +
 geom_histogram(aes(y=..density..), color = "black", fill = "white") +
 geom_density(size = 1, color = "seagreen")
```



```
ggplot(posterior_sim, aes(sigma)) +
 geom_histogram(aes(y=..density..), color = "black", fill = "white") +
 geom_density(size = 1, color = "seagreen")
```



```
ggplot(posterior_sim, aes(mu, sigma)) +
  stat_density_2d(aes(fill = ..level..),
                  geom = "polygon", color = "white") +
     scale_fill_viridis_c()
```



A few comments about the posterior distribution

- The joint posterior distribution appears to be roughly Bivariate Normal. The correlation is close to 0, indicating independence[1] between μ and σ in the posterior.
- The posterior distribution of μ is approximately Normal with posterior mean 97.7 (basically the sample mean) and posterior SD 0.05. There is a 98% posterior probability that the population mean human body temperature is between 97.6 and 97.9 degrees F.
- The posterior distribution of σ is approximately Normal with posterior mean 0.75 (basically the sample SD) and posterior SD 0.036. There is a 98% posterior probability that the population SD of human body temperatures is between 0.67 and 0.84 degrees F.
- Since μ and σ are roughly independent in the posterior, there is a posterior probability of 96% that both of the above statements are true, that is, that both parameters lie in their respective credible intervals. To have joint posterior credibility of 98%, we could lengthen each interval (to 99% for two independent intervals) to obtain a rectangular credibility region. The scatterplot also shows a 98% posterior credible ellipse (in black) for both μ and σ.

Example 14.5. Continuing the previous example, how could you use simulation to approximate the posterior predictive distribution of a single body tem-

[1] Remember, if X and Y are independent then the correlation is 0, but the converse is not true in general. However, if X and Y have a *Bivariate Normal* distribution and their correlation is 0, then X and Y are independent.

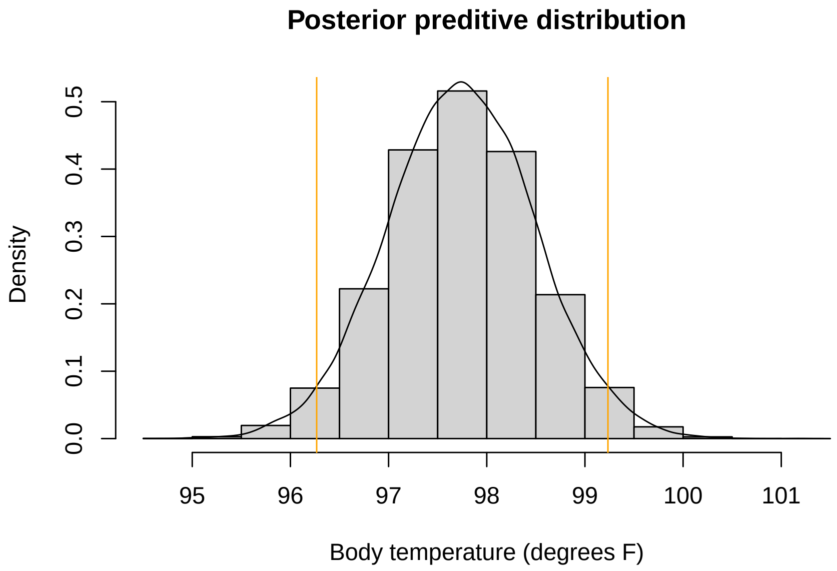
perature? Conduct the simulation and compute and interpret a 95% prediction interval.

Solution. to Example 14.5

- Simulate a (μ, σ) pair from the joint posterior distribution.
- Given μ and σ, simulate a value of y from a $N(\mu, \sigma)$ distribution.
- Repeat many times and summarize the simulated y values to approximate the posterior predictive distribution.

See the code below. JAGS has already returned a simulation from the joint posterior distribution of (μ, σ) For each of these simulated values, simulate a corresponding y value like usual.

```
theta_sim = as.matrix(posterior_sample)

y_sim = rnorm(nrow(theta_sim), theta_sim[, "mu"], theta_sim[, "sigma"])

hist(y_sim, freq = FALSE, xlab = "Body temperature (degrees F)",
     main = "Posterior preditive distribution")
lines(density(y_sim))
abline(v = quantile(y_sim, c(0.025, 0.975)), col = "orange")
```



```
quantile(y_sim, c(0.025, 0.975))
```

```
##  2.5% 97.5%
## 96.27 99.23
```

There is a posterior predictive probability of 95% that a body temperature is between 96.25 and 99.20 degrees F. Roughly, 95% of healthy human body temperatures are between 96.25 and 99.20 degrees F.

Chapter 15

Bayesian Analysis of a Numerical Variable

In this chapter we'll continue our study of a single numerical variable. When the distribution of the measured variable is symmetric and unimodal, the population mean is often the main parameter of interest. However, the population SD also plays an important role.

In the previous section we assumed that the measured numerical variable followed a Normal distribution. That is, we assumed a Normal likelihood function. However, the assumption of Normality is not always justified, even when the distribution of the measured variable is symmetric and unimodal. In this chapter we'll investigate an alternative to a Normal likelihood that is more flexible and robust to outliers and extreme values.

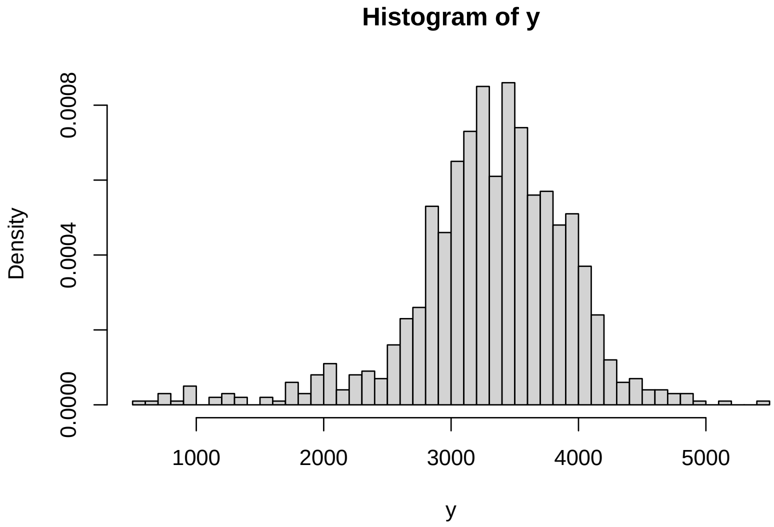
Example 15.1. In a previous assignment, we assumed that birthweights (grams) of human babies follow a Normal distribution with unknown mean μ and known SD $\sigma = 600$. (1 pound 454 grams.) We assumed a Normal(3400, 100) (in grams, or Normal(7.5, 0.22) in pounds) prior distribution for μ; this prior distribution places most of its probability on mean birthweight being between 7 and 8 pounds.

Now we'll assume, more realistically, that σ is unknown.

1. What does the parameter σ represent? What is a reasonable prior mean for σ? What range of values of σ will account for most of the prior probability?
2. Assume a Gamma prior distribution for σ with mean 600 and SD 200; this is a Gamma distribution with shape parameter $\alpha = 600^2/200^2$ and rate parameter $600/200^2$. Also assume that μ and σ are independent according to the prior distribution. Explain how you could use simulation to approximate the prior predictive distribution of birthweights. Run the simulation and summarize the results. Does the choice of prior seem reasonable?

3. The following summarizes data on a random sample[1] of 1000 live births in the U.S. in 2001.

```
data = read.csv("_data/birthweight.csv")
y = data$birthweight

hist(y, freq = FALSE, breaks = 50)
```



```
summary(y)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     595    3005    3350    3315    3714    5500
```

```
sd(y)
```

```
## [1] 631.3
```

```
n = length(y)
ybar = mean(y)
```

Does it seem reasonable to assume birthweights follow a Normal distribution?

[1] There are about 4 million live births in the U.S. per year. The data is available at the CDC website. We're only using a random sample to cut down on computation time.

4. Regardless of your answer to the previous question, continue to assume the model above. Use JAGS to find the posterior distribution.

5. How could you use simulation to approximate the posterior predictive distribution of birthweights? Run the simulation and find a 99% posterior prediction interval.

6. What percent of values in the observed *sample* fall outside the prediction interval? What does that tell you?

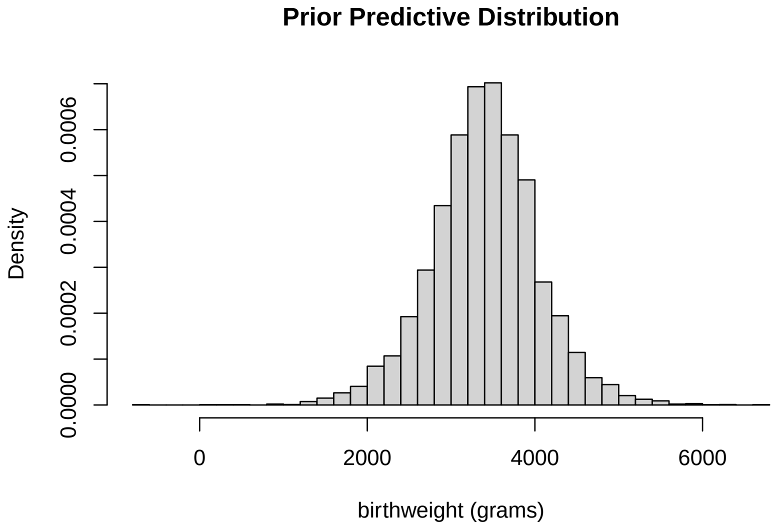
Solution. to Example 15.1

1. The parameter σ represents the population standard of individual birthweights: how much do birthweights vary from baby to baby? Our prior for μ says that a *mean* birthweight in the 7-8 or so pounds range seems reasonable. The parameter σ represents how much individual birthweights vary about this mean. Let's say that we think most babies weigh between 5 and 10 pounds; then we might want the interval [5, 10] to account for 95% of birthweights, or the values with 2 SDs of the mean, so we might want 5 pounds (the length of [5, 10]) to represent 4 SDs. So one reasonable prior mean of σ might be around 1.25 pounds, or around 600 grams or so. Let's choose a prior SD of 200 grams for σ to cover a reasonably wide range of values of σ: values like 100 grams which represent little variability in birthweights, to values like 1000 grams which represent a great deal of variability in birthweights. That is, we'll choose a Gamma prior distribution for σ with mean 600 and SD 200; this is a Gamma distribution with shape parameter $\alpha = 600^2/200^2$ and rate parameter $600/200^2$. Remember, this is only one choice for prior. There are many other reasonable choices.

2. Simulate a value μ from a Normal(3400, 100) distribution, and independently simulate a value of σ from a Gamma($600^2/200^2$, $600/200^2$) distribution. Given μ and σ, simulate a value y from a Normal(μ, σ) distribution. Repeat many times and summarize the simulated y values to approximate the prior predictive distribution. The results are below. According to this prior model, we predict that most babies weigh between about 4.5 and 10.5 pounds, which seems reasonable based on what we know about birthweights.

```
n_rep = 10000

mu_sim = rnorm(n_rep, 3400, sd = 100)

sigma_sim = rgamma(n_rep, 600 ^ 2 / 200 ^ 2, 600 / 200 ^2)

y_sim = rnorm(n_rep, mu_sim, sigma_sim)
```

```
hist(y_sim,
 breaks = 50,
 freq = FALSE,
 main = "Prior Predictive Distribution",
 xlab = "birthweight (grams)")
```



```
quantile(y_sim, c(0.025, 0.975))
```

```
##   2.5% 97.5%
##   2086  4685
```

3. Assuming a Normal distribution doesn't seem terrible, but it does appear that maybe the tails are a little heavier than we would expect for a Normal distribution, especially at the low birthweights. That is, there might be some evidence in the data that extremely low (and maybe high) birthweights don't quite follow what would be expected if birthweights followed a Normal distribution.

4. See JAGS code below. A few notes:

   - The JAGS code takes the full sample of size 1000 as an input.
   - The data consists of 1000 values assumed to each be from a Normal(μ, σ) distribution.
   - Remember that in JAGS, it's `dnorm(mean, precision)`.

- To find the likelihood of observing the entire sample, JAGS finds the likelihood of each of the individual values and then multiplies the values together for us to find the likelihood of the sample.
- This is accomplished in JAGS by specifying the likelihood via a for loop which evaluates the likelihood `y[i] ~ dnorm(mu, 1 / sigma ^ 2)` for each `y[i]` in the sample.

```
Nrep = 10000
Nchains = 3

# data
# data has already been loaded in previous code
# y is the full sample
# n is the sample size

# model
model_string <- "model{

  # Likelihood
  for (i in 1:n){
y[i] ~ dnorm(mu, 1 / sigma ^ 2)
  }


  # Prior
  mu ~ dnorm(3400, 1 / 100 ^ 2)

  sigma ~ dgamma(600 ^ 2 / 200 ^ 2, 600 / 200 ^2)

}"

# Compile the model
dataList = list(y=y, n=n)

model <- jags.model(textConnection(model_string),
                    data=dataList,
                    n.chains=Nchains)
```

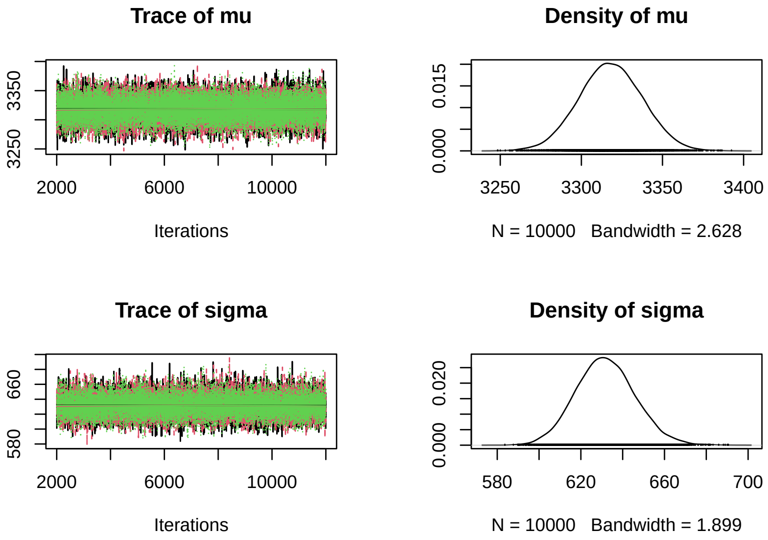
```
## Compiling model graph
##    Resolving undeclared variables
##    Allocating nodes
## Graph information:
##    Observed stochastic nodes: 1000
##    Unobserved stochastic nodes: 2
##    Total graph size: 1017
```

```
##
## Initializing model
```

```
update(model, 1000, progress.bar="none")

posterior_sample <- coda.samples(model,
                                variable.names=c("mu", "sigma"),
                                n.iter=Nrep,
                                progress.bar="none")

# Summarize and check diagnostics
summary(posterior_sample)
```

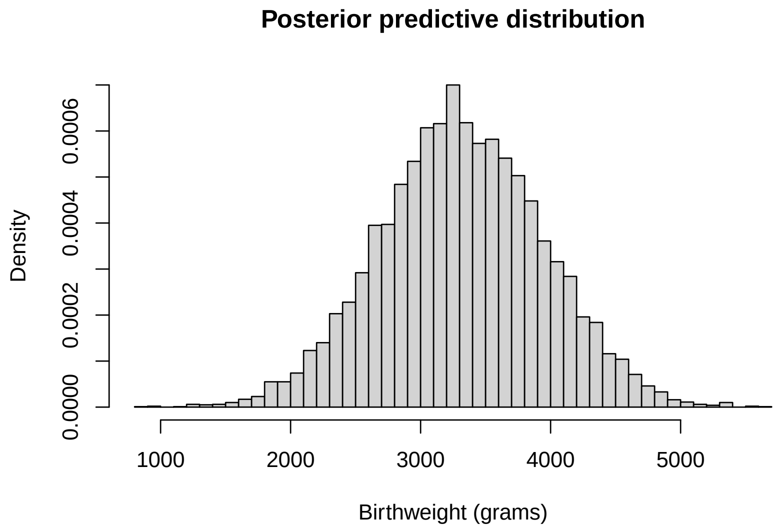
```
##
## Iterations = 2001:12000
## Thinning interval = 1
## Number of chains = 3
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##       Mean   SD Naive SE Time-series SE
## mu    3318 19.5   0.1125          0.115
## sigma  632 14.1   0.0815          0.105
##
## 2. Quantiles for each variable:
##
##       2.5%  25%  50%  75% 97.5%
## mu    3280 3305 3318 3331  3356
## sigma  605  622  631  641   661
```

```
plot(posterior_sample)
```



5. Simulate a (μ, σ) pair from the posterior distribution; JAGS has already done this for you. Then given μ and σ simulate a value of y from a Normal(μ, σ) distribution. Repeat many times and summarize the simulated values of y to approximate the posterior predictive distribution of birthweights.

```r
theta_sim = data.frame(as.matrix(posterior_sample))

y_sim = rnorm(Nrep, theta_sim$mu, theta_sim$sigma)

hist(y_sim, freq = FALSE, breaks = 50,
 xlab = "Birthweight (grams)",
 main = "Posterior predictive distribution")
```



```
quantile(y_sim, c(0.005, 0.995))
```

```
##  0.5% 99.5%
##  1714  4892
```

There is a posterior predictive probability of 99% that a randomly selected birthweight is between 1714 and 4892 grams. Roughly, this model and data say that 99% of birthweights are between 1714 and 4892 grams (3.8 and 10.8 pounds.).

6. Find the proportion of values in the observed sample that lie outside of the prediction interval

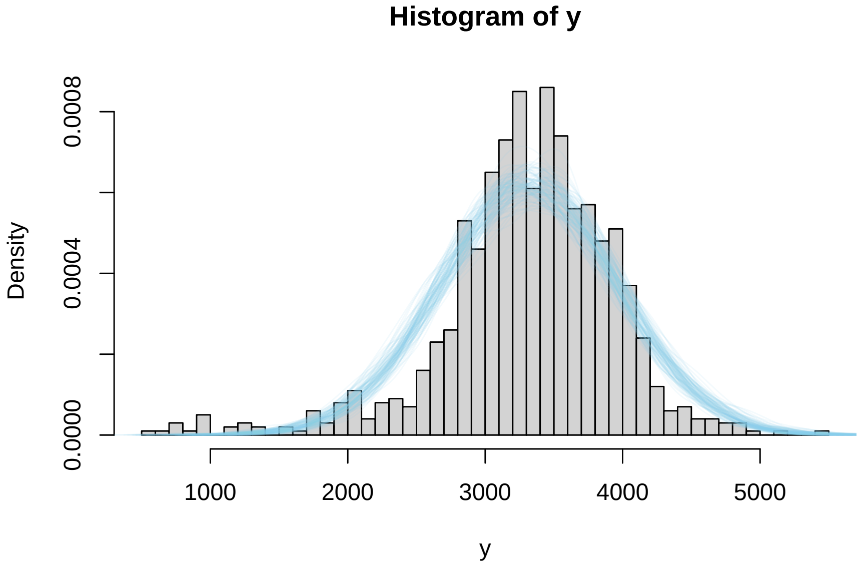
```
(sum(y < quantile(y_sim, 0.005)) + sum(y > quantile(y_sim, 0.995))) / n
```

```
## [1] 0.024
```

About 2.4 percent of birthweights in the sample fall outside of the 99% prediction interval, when we would only expect 1%. While not a large difference in magnitude, we are observing a higher percentage of birthweights in the tails than we would expect if birthweights followed a Normal distribution. So we have some evidence that a Normal model — that is, a Normal likelihood — might not be the best model for birthweights of all live births as it doesn't properly account for extreme birthweights.

The code below performs a posterior predictive check by simulating hypothetical samples of size 1000 from the posterior model, and comparing with the observed sample of size 1000. The simulation is similar to the posterior predictive simulation in the previous example, but now every time we simulate a (μ, σ) pair, we simulate a random sample of 1000 y values. Again, while not a terrible fit, there do seem to be more values in the tail — the lower tail especially — than would be expected under this model.

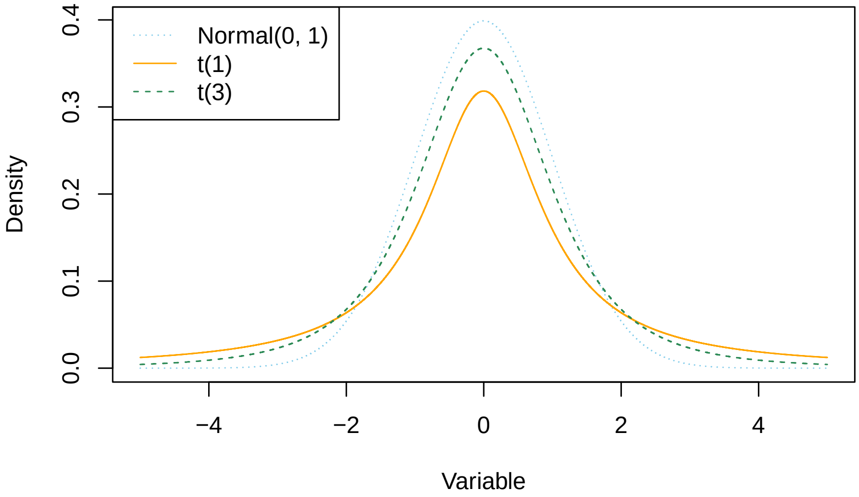
```r
# plot the observed data
hist(y, freq = FALSE, breaks = 50) # observed data

# number of samples to simulate
n_samples = 100

# simulate (mu, sigma) pairs from the posterior
# we just randomly select rows from theta_sim
index_sample = sample(Nrep, n_samples)

# simulate samples
for (r in 1:n_samples){

  i = index_sample[r]

  # simulate values from N(theta, sigma) distribution
  y_sim = rnorm(n, theta_sim[i, "mu"], theta_sim[i, "sigma"])

  # add plot of simulated sample to histogram
  lines(density(y_sim),
        col = rgb(135, 206, 235, max = 255, alpha = 25))
}
```



In the previous example we assumed a Normal likelihood. A Normal likelihood assumes that the population distribution of individual values of the measured numerical variable is Normal. Posterior predictive checking can be used to assess whether a Normal likelihood is appropriate for the observed data. If a Normal likelihood isn't an appropriate model for the data then other likelihood functions can be used. In particular, if the observed data is relatively unimodal and symmetric[2] but has more extreme values than can be accommodated by a Normal likelihood, a t-distribution or other distribution with heavy tails can be used to model the likelihood.

Normal distributions don't allow much room for extreme values. An alternative is to assume a distribution with heavier tails. For example, t-distributions have heavier tails than Normal. For t-distributions, the *degrees of freedom* parameter $d \geq 1$ controls how heavy the tails are. When d is small, the tails are much heavier than for a Normal distribution, leading to a higher frequency of extreme values. As d increases, the tails get lighter and a t-distribution gets closer to a Normal distribution. For d greater than 30 or so, there is very little different between a t-distribution and a Normal distribution except in the extreme tails. The degrees of freedom parameter d is sometimes referred to as the "Normality parameter", with larger values of d indicating a population distribution that is closer to Normal.

[2] If the observed data has multiple modes or is skewed, then other parameters like median or mode might be more appropriate measures of center than the population mean.



Example 15.2. Continuing the birthweight example, we'll now model the distribution of birthweights with a $t(\mu, \sigma, d)$ distribution.

1. How many parameters is the likelihood function based on? What are they?
2. What does assigning a prior distribution to the Normality parameter d represent?
3. The Normality parameter must satisfy $d \geq 1$, so we want a distribution for which only values above 1 are possible. One way to accomplish this is to let $d_0 = d - 1$, assign a Gamma distribution prior to $d_0 \geq 0$, and then let $d = 1 + d_0$. One common approach is to let the shape parameter of the Gamma distribution to be 1, and to let the scale parameter be 1/29, so that the prior mean of d is 30. Assume the same priors for μ and σ as in the previous example, and a Gamma(1, 1/29) prior for $(d - 1)$. Use JAGS to fit the model to the birthweight data and approximate and summarize the posterior distribution.
4. Consider the posterior distribution for d. Based on this posterior distribution, is it plausible that birthweights follow a Normal distribution?
5. Consider the posterior distribution for σ. What seems strange about this distribution? (Hint: consider the sample SD.)
6. The standard deviation of a Normal(μ, σ) distribution is σ. However, the standard deviation of a $t(\mu, \sigma, d)$ distribution is not σ; rather it is

$\sigma\sqrt{\frac{d}{d-2}} > \sigma$. When d is large, $\sqrt{\frac{d}{d-2}} \approx 1$, and so the standard deviation is approximately σ. However, it can make a difference when d is small.

Using the JAGS output, create a plot of the posterior distribution of $\sigma\sqrt{\frac{d}{d-2}}$. Does this posterior distribution of the population standard deviation seem more reasonable in light of the sample data?

7. How could you use simulation to approximate the posterior predictive distribution of birthweights? Run the simulation and find a 99% posterior prediction interval. How does it compare to the predictive interval from the model with the Normal likelihood?

Solution. to Example 15.2

1. The $t(\mu, \sigma, d)$ is based on 3 parameters: the population mean μ, the variability parameter σ (which is NOT the population SD; see below), and the Normality parameter d.

2. Assigning a prior distribution to d allows for a posterior distribution of d, which quantifies uncertainty about the degree of Normality versus “heavy-tailed-ness” of the distribution of birthweights. Assigning a prior distribution on d allows the model to explore different values of d to see what values seem most plausible given the observed data.

3. See the JAGS code below and output below. Note that the posterior distribution for μ is similar to the posterior distribution for μ from the model with the Normal likelihood.

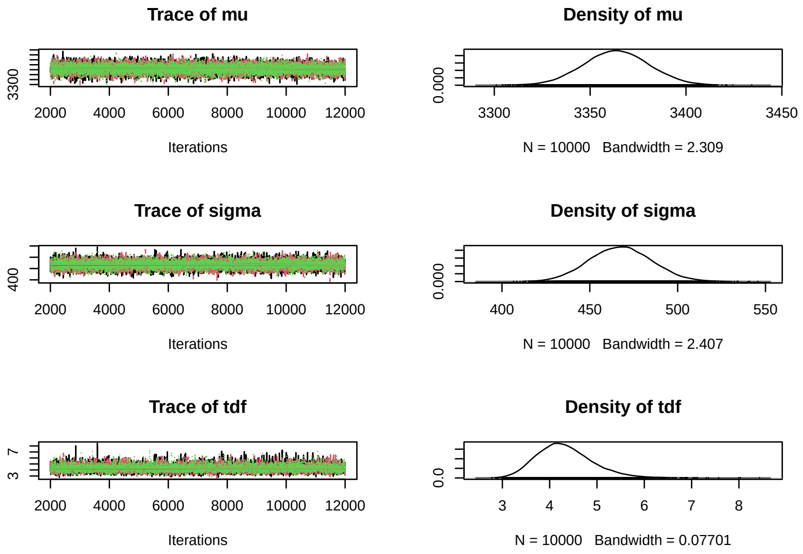
```
Nrep = 10000
Nchains = 3

# data
# data has already been loaded in previous code
# y is the full sample
# n is the sample size

# model
model_string <- "model{

  # Likelihood
  for (i in 1:n){
y[i] ~ dt(mu, 1 / sigma ^ 2, tdf)
  }


  # Prior
  mu ~ dnorm(3400, 1 / 100 ^ 2)

  sigma ~ dgamma(600 ^ 2 / 200 ^ 2, 600 / 200 ^ 2)

  tdf <- 1 + tdf0

  tdf0 ~ dexp(1 / 29)

}"

# Compile the model
dataList = list(y=y, n=n)

model <- jags.model(textConnection(model_string),
                data=dataList,
                n.chains=Nchains)
```

```
## Compiling model graph
##    Resolving undeclared variables
##    Allocating nodes
## Graph information:
##    Observed stochastic nodes: 1000
##    Unobserved stochastic nodes: 3
##    Total graph size: 1021
##
## Initializing model
```

```
update(model, 1000, progress.bar="none")

posterior_sample <- coda.samples(model,
                             variable.names=c("mu", "sigma", "tdf"),
                             n.iter=Nrep,
                             progress.bar="none")

# Summarize and check diagnostics
summary(posterior_sample)
```

```
##
## Iterations = 2001:12000
## Thinning interval = 1
## Number of chains = 3
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##           Mean     SD Naive SE Time-series SE
## mu    3363.69 17.124  0.09886        0.12760
## sigma  467.64 17.850  0.10306        0.20079
## tdf      4.32  0.593  0.00342        0.00703
##
## 2. Quantiles for each variable:
##
##           2.5%     25%     50%     75%   97.5%
## mu    3330.05 3352.21 3363.67 3375.15 3397.26
## sigma  433.34  455.48  467.48  479.61  503.24
## tdf      3.33    3.91    4.27    4.68    5.65
```

```
plot(posterior_sample)
```



```
# Extract the JAGS output as a data frame
theta_sim = data.frame(as.matrix(posterior_sample))
```

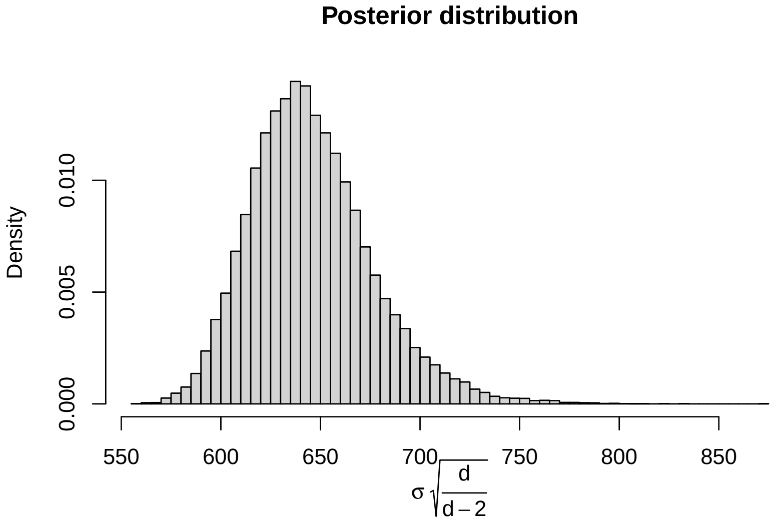
4. The posterior distribution for d places most of its probability on relatively small values of d. A 98% posterior credible interval for d is 3.2 to 6. Values in this interval indicate relatively heavy tails in comparison to a Normal distribution. Therefore, the posterior distribution indicates that it is not plausible that birthweights follow a Normal disttribution.

5. The posterior distribution for σ might seem initially strange. The sample standard is 631, but this value is not included the range of plausible values

of σ according to the posterior. This is an artifact of the interplay between σ and d in t-distributions, especially when d is small. See the next part for more details.

6. See below. This posterior distribution seems more reasonable given the sample SD.

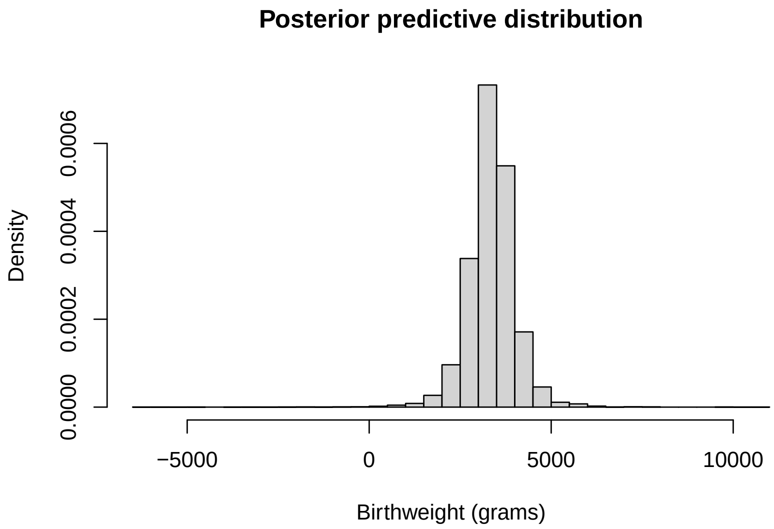
```
hist(theta_sim$sigma * sqrt(theta_sim$tdf / (theta_sim$tdf - 2)),
 freq = FALSE,
 breaks = 50,
 main = "Posterior distribution",
 xlab = expression(paste(sigma, " ", sqrt(frac(d, d-2)))))
```



7. Simulate a triple (μ, σ, d) from the joint posterior distribution; JAGS has already done this for you. Given (μ, σ, d), simulate a value y from a $t(\mu, \sigma, d)$ distribution. Repeat many times and summarize the simulated y values to approximate the posterior distribution.

   See the code and output below. The posterior predictive distribution now spans a more disperse range of birthweights that with the Normal model.

```
y_sim = theta_sim$mu + theta_sim$sigma * rt(Nrep, theta_sim$tdf)

hist(y_sim, freq = FALSE, breaks = 50,
 xlab = "Birthweight (grams)",
 main = "Posterior predictive distribution")
```



```
quantile(y_sim, c(0.005, 0.995))
```

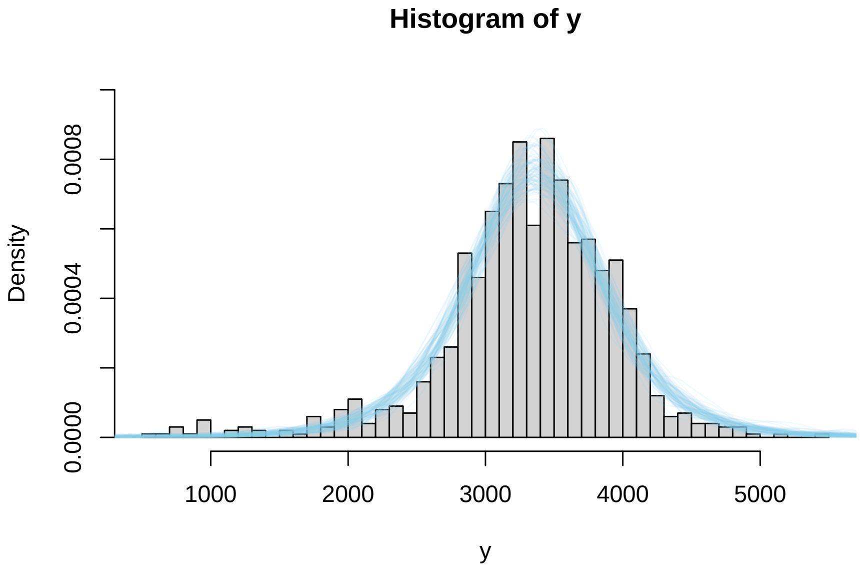
```
##  0.5% 99.5%
##  1104  5605
```

The code below performs a posterior predictive check by simulating hypothetical samples of size 1000 from the posterior model, and comparing with the observed sample of size 1000. The simulation is similar to the posterior predictive simulation in the previous example, but now every time we simulate a (μ, σ) pair, we simulate a random sample of 1000 y values.

We'll now try posterior predictive checks. We use the simulated values of θ from JAGS. For each value of θ, we generate a sample of size 1000 from a t-distribution centered at θ and with a standard deviation of 600. (`rt` in R returns values on a standardized scale which we then rescale.)

```
# plot the observed data
hist(y, freq = FALSE, breaks = 50,
     ylim = c(0, 0.001)) # observed data

# number of samples to simulate
n_samples = 100

# simulate (mu, sigma, d) triples from the posterior
# we just randomly select rows from theta_sim
index_sample = sample(Nrep, n_samples)
```

```
# simulate samples
for (r in 1:n_samples){

  # simulate values from a t-distribution
  y_sim = theta_sim[r, "mu"] + theta_sim[r, "sigma"] * rt(n, theta_sim[r, "tdf"])

  # add plot of simulated sample to histogram
  lines(density(y_sim),
        col = rgb(135, 206, 235, max = 255, alpha = 25))
}
```



We see that the actual sample resembles simulated samples more closely for the model based on the t-distribution likelihood than for the one based on the Normal likelihood. While we do want a model that fits the data well, we also do not want to risk overfitting the data. In this case, we do not want a few extreme outliers to unduly influence the model. However, it does appear that a model that allows for heavier tails than a Normal distribution could be useful here. Moreover, accommodating the tails improves the fit in the center of the distribution too.

When the degrees of freedom are very small, *t*-distributions can give rise to super extreme values. We see this in the posterior predictive distribution for birthweights, where there are some negative birthweights and birthweights over 10000 grams. The model based on the the *t*-likelihood seems to fit well over the range of observed values of birthweight. As usual, we should refrain from making predictions outside the ranges of the observed data.

In models with multiple parameters, there can be dependencies between parameters, so interpreting the marginal posterior distribution of any single parameter can be difficult. It is often more helpful to consider predictive distributions, which accounts for the joint distribution of all parameters. Interpreting predictive distributions is often more intuitive since predictive distributions live on the scale of the measured variable.

The observational units in the examples in this section are live births. Many of the extremely low birthweights are attributed to babies who were born live but did not survive. Rather than model birthweights with one single likelihood, it might be more appropriate to first categorize the births and use the "group" variable in the model. For example, me might include a variable to indicate whether a baby is full term or not (or even just the length of the pregnancy). We will see how to compare a numerical variable across groups in upcoming sections.

Chapter 16

Comparing Two Samples

Most interesting statistical problems involve multiple unknown parameters. For example, many problems involve comparing two (or more) populations or groups based on two (or more) samples. In such situations, each population or group will have its own parameters, and there will often be dependence between parameters. We are usually interested in difference or ratios of parameters between groups.

The example below concerns the familiar context of comparing two means. However, the way independence is treated in the example is not the most common. Soon we will see *hierarchical models* for comparing groups.

Example 16.1. Do newborns born to mothers who smoke tend to weigh less at birth than newborns from mothers who don't smoke? We'll investigate this question using birthweight (pounds) data on a sample of births in North Carolina over a one year period.

Assume birthweights follow a Normal distribution with mean μ_1 for nonsmokers and mean μ_2 for smokers, and standard deviation σ.

Note: our primary goal will be to compare the means μ_1 and μ_2. We're assuming a common standard deviation σ to simplify a little, but we could (and probably should) also let standard deviation vary by smoking status.

1. The prior distribution will be a joint distribution on (μ_1, μ_2, σ) triples. We could assume a prior under which μ_1 and μ_2 are independent. But why might we want to assume some prior *dependence* between μ_1 and μ_2? (For some motivation it might help to consider what the frequentist null hypothesis would be.)

2. One way to incorporate prior dependence is to assume a Multivariate Normal prior distribution. For the prior assume:

- σ is independent of (μ_1, μ_2)
- σ has a Gamma(1, 1) distribution
- (μ_1, μ_2) follow a Bivariate Normal distribution with prior means (7.5, 7.5) pounds, prior standard deviations (0.5, 0.5) pounds, and prior correlation 0.9.

Simulate values of (μ_1, μ_2) from the prior distribution[1] and plot them. Briefly[2] describe the prior distribution.

3. How do you interpret the parameter $\mu_1 - \mu_2$? Plot the prior distribution of $\mu_1 - \mu_2$, and find prior central 50%, 80%, and 98% credible interval. Also compute the prior probability that $\mu_1 - \mu_2 > 0$.

4. The following code loads and summarizes the sample data. Briefly describe the data.

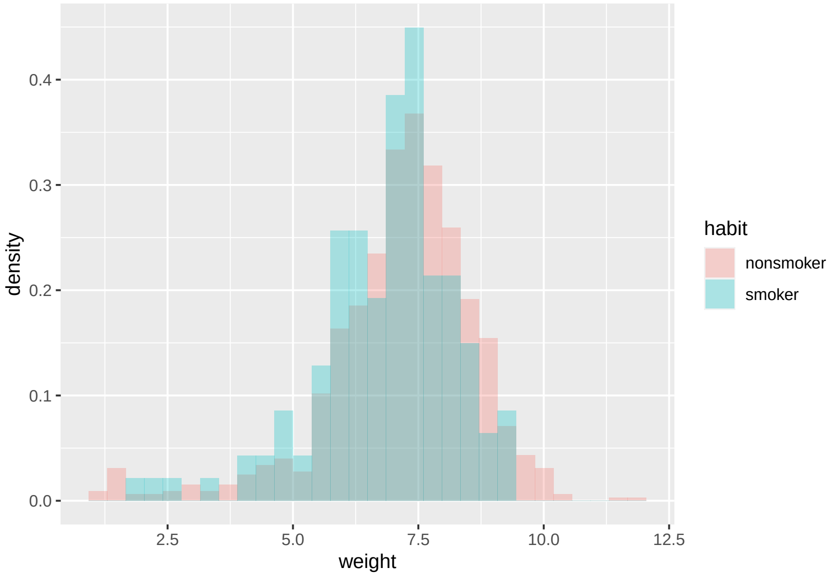
```r
data = read.csv("_data/baby_smoke.csv")

ggplot(data, aes(weight, fill = habit)) +
  geom_histogram(alpha = 0.3,
                 aes(y = ..density..),
                 position = 'identity')
```

[1] Values from a Multivariate Normal distribution can be simulated using `mvrnorm` from the `MASS` package. For Bivariate Normal, the inputs are the mean vector $[E(\mu_1), E(\mu_2)]$ and the covariance matrix

$$\begin{bmatrix} \text{Var}(\mu_1) & \text{Cov}(\mu_1, \mu_2) \\ \text{Cov}(\mu_1, \mu_2) & \text{Var}(\mu_2) \end{bmatrix}$$

where $\text{Cov}(\mu_1, \mu_2) = \text{Corr}(\mu_1, \mu_2)\text{SD}(\mu_1)\text{SD}(\mu_2)$.

[2] Why briefly? Because we want to focus on the *posterior* distribution.



```
data %>%
  group_by(habit) %>%
  summarize(n(), mean(weight), sd(weight)) %>%
  kable(digits = 2)
```

habit	n()	mean(weight)	sd(weight)
nonsmoker	873	7.14	1.52
smoker	126	6.83	1.39

5. Is it reasonable to assume that the two *samples* are independent? (In this case the y_1 and y_2 samples would be *conditionally independent* given (μ_1, μ_2, σ).)

6. Describe how you would compute the likelihood. For concreteness, how you would you compute the likelihood if there were only 4 babies in the sample: 2 non-smokers with birthweights of 8 pounds and 7 pounds, and 2 smokers with birthweights of 8.3 pounds and 7.1 pounds.

7. Use JAGS to approximate the posterior distribution. (The coding is a little tricky. See the code and some comments below.) Plot the posterior distribution. How strong is the dependence between μ_1 and μ_2 in the posterior? Why do you think that is?

8. Plot the posterior distribution of $\mu_1 - \mu_2$ and describe it. Compute and interpret posterior central 50%, 80%, and 98% credible intervals. Also compute and interpret the posterior probability that $\mu_1 - \mu_2 > 0$.

9. If we're interested in $\mu_1 - \mu_2$, why didn't we put a prior directly on $\mu_1 - \mu_2$ rather than on (μ_1, μ_2)?

10. Plot the posterior distribution of μ_1/μ_2, describe it, and find and interpret posterior central 50%, 80%, and 98% credible intervals.

11. Is there some evidence that babies whose mothers smoke tend to weigh less than those whose mothers don't smoke?

12. Can we say that smoking is the cause of the difference in mean weights?

13. Is there some evidence that babies whose mothers smoke tend to weigh *much* less than those whose mothers don't smoke? Explain.

14. One quantity of interest is the **effect size**, which is a way of measuring the magnitude of the difference between groups. When comparing two means, a simple measure of effect size (*Cohen's d*) is

$$\frac{\mu_1 - \mu_2}{\sigma}$$

Plot the posterior distribution of this effect size and describe it. Compute and interpret posterior central 50%, 80%, and 98% credible intervals.

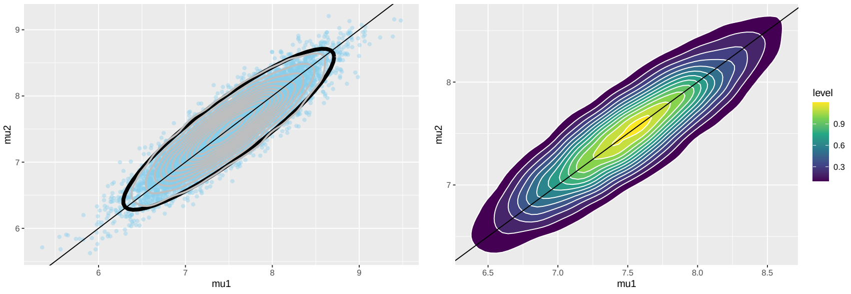
Solution. to Example 16.1

1. If our prior belief is that there is no difference in mean birthweights between babies of smokers and non-smokers, then our prior should place high probability on μ_1 being close to μ_2. Even if we want our prior to allow for different distributions for μ_1 and μ_2, there might still be some dependence. For example, we would assign a different prior conditional probability to the event that $\mu_1 > 8.5$ given $\mu_2 > 8.5$ than we would given $\mu_2 < 7.5$. Our prior uncertainty about mean birthweights of babies of nonsmokers informs our prior uncertainty about mean birthweights of babies in general, and hence also of babies of smokers.

2. Our main focus is on (μ_1, μ_2). We see that the prior places high density on (μ_1, μ_2) pairs with μ_1 close to μ_2.

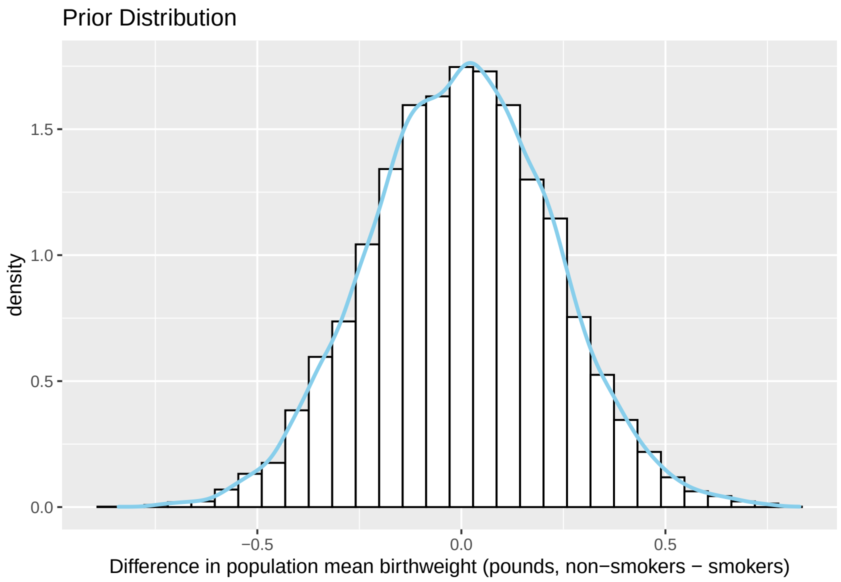
```
mu_prior_mean <- c(7.5, 7.5)
mu_prior_sd <- c(0.5, 0.5)
mu_prior_corr <- 0.9

mu_prior_cov <- matrix(c(mu_prior_sd[1] ^ 2,
              mu_prior_corr * mu_prior_sd[1] * mu_prior_sd[2],
              mu_prior_corr * mu_prior_sd[1] * mu_prior_sd[2],
              mu_prior_sd[2] ^ 2), nrow = 2)
```

```
library(MASS)

sim_prior = data.frame(mvrnorm(10000, mu_prior_mean, mu_prior_cov),
                           rgamma(10000, 1, 1))

names(sim_prior) = c("mu1", "mu2", "sigma")

ggplot(sim_prior, aes(mu1, mu2)) +
  geom_point(color = "skyblue", alpha = 0.4) +
  stat_ellipse(level = 0.98, color = "black", size = 2) +
  stat_density_2d(color = "grey", size = 1) +
  geom_abline(intercept = 0, slope = 1)

ggplot(sim_prior, aes(mu1, mu2)) +
  stat_density_2d(aes(fill = ..level..),
          geom = "polygon", color = "white") +
 scale_fill_viridis_c() +
  geom_abline(intercept = 0, slope = 1)
```



3. The parameter $\mu_1 - \mu_2$ is the difference in mean birthweights, smokers minus non-smokers. The prior mean of $\mu_1 - \mu_2$ is 0, reflecting a prior belief towards no difference in mean birthweight between smokers and non-smokers. Furthermore, there is a fairly high prior probability that the mean birthweight for smokers is close to the mean birthweight for non-smokers, with a difference of at most about 0.5 pounds. Under this prior, nonsmokers and smokers are equally likely to have the higher mean birthweight.

```
sim_prior <- sim_prior %>%
  mutate(mu_diff = mu1 - mu2)

ggplot(sim_prior,
   aes(mu_diff)) +
  geom_histogram(aes(y=..density..), color = "black", fill = "white") +
  geom_density(size = 1, color = "skyblue") +
  labs(x = "Difference in population mean birthweight (pounds, non-smokers - smoke
   title = "Prior Distribution")
```



```
quantile(sim_prior$mu_diff, c(0.01, 0.10, 0.25, 0.75, 0.90, 0.99))
```

```
##       1%      10%      25%      75%      90%      99% 
## -0.5173 -0.2863 -0.1505  0.1540  0.2825  0.5297
```

```
sum(sim_prior$mu_diff > 0 ) / 10000
```

```
## [1] 0.5059
```

4. The distributions of birthweights are fairly similar for smokers and non-smokers. The sample mean birthweight for smokers is about 0.3 pounds less than the sample mean birthweight for smokers. The sample SDs of birthweights are similar for both groups, around 1.4-1.5 pounds.

5. Yes, it is reasonable to assume that the two *samples* are independent. The data for smokers was collected separately from the data for non-smokers. That is, it is reasonable to assume *independence in the data*.

6. For each observed value of birthweight for non-smokers, evaluate the likelihood based on a $N(\mu_1, \sigma)$ distribution. For example, if birthweight of a non-smoker is 8 pounds, the likelihood is `dnorm(8, mu1, sigma)`; if birthweight of a non-smoker is 7 pounds, the likelihood is `dnorm(7, mu1,`

`sigma)`. The likelihood for the sample of non-smokers would be the products — assuming independence *within* sample — of the likelihoods of the individual values, as a function of μ_1 and σ: `dnorm(8, mu1, sigma) * dnorm(7, mu1, sigma) * ...`

The likelihood for the sample of smokers would be the products of the likelihoods of the individual values, as a function of μ_2 and σ: `dnorm(8.3, mu2, sigma) * dnorm(7.1, mu2, sigma) * ...`

The likelihood function for the full sample would be the product — assuming independence *between* samples — of the likelihoods for the two samples, a function of μ_1, μ_2 and σ.

7. Here is the code; there are some comments about syntax at the end of this chapter.

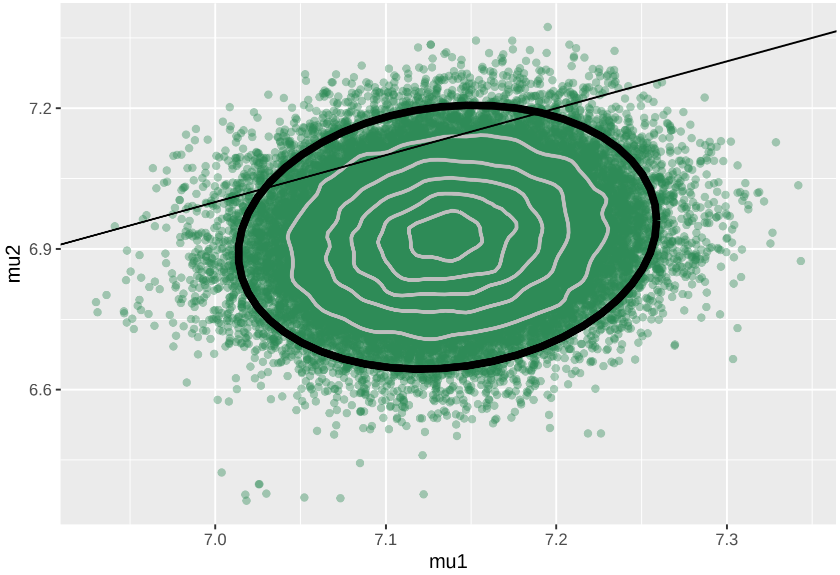
```
# data
y = data$weight

x = (data$habit == "smoker") + 1

n = length(y)

n_groups = 2

# Prior parameters
mu_prior_mean <- c(7.5, 7.5)
mu_prior_sd <- c(0.5, 0.5)
mu_prior_corr <- 0.9

mu_prior_cov <- matrix(c(mu_prior_sd[1] ^ 2,
              mu_prior_corr * mu_prior_sd[1] * mu_prior_sd[2],
              mu_prior_corr * mu_prior_sd[1] * mu_prior_sd[2],
              mu_prior_sd[2] ^ 2), nrow = 2)

# Model
model_string <- "model{

  # Likelihood
  for (i in 1:n){
  y[i] ~ dnorm(mu[x[i]], 1 / sigma ^ 2)
  }

  # Prior
  mu[1:n_groups] ~ dmnorm.vcov(mu_prior_mean[1:n_groups],
                          mu_prior_cov[1:n_groups, 1:n_groups])
  sigma ~ dgamma(1, 1)

}"

dataList = list(y = y, x = x, n = n, n_groups = n_groups,
            mu_prior_mean = mu_prior_mean, mu_prior_cov = mu_prior_cov)

# Compile
Nrep = 10000

n.chains = 5

model <- jags.model(textConnection(model_string),
                data = dataList,
                n.chains = n.chains)
```

```
## Compiling model graph
##    Resolving undeclared variables
##    Allocating nodes
## Graph information:
##    Observed stochastic nodes: 999
##    Unobserved stochastic nodes: 2
##    Total graph size: 2016
##
## Initializing model
```

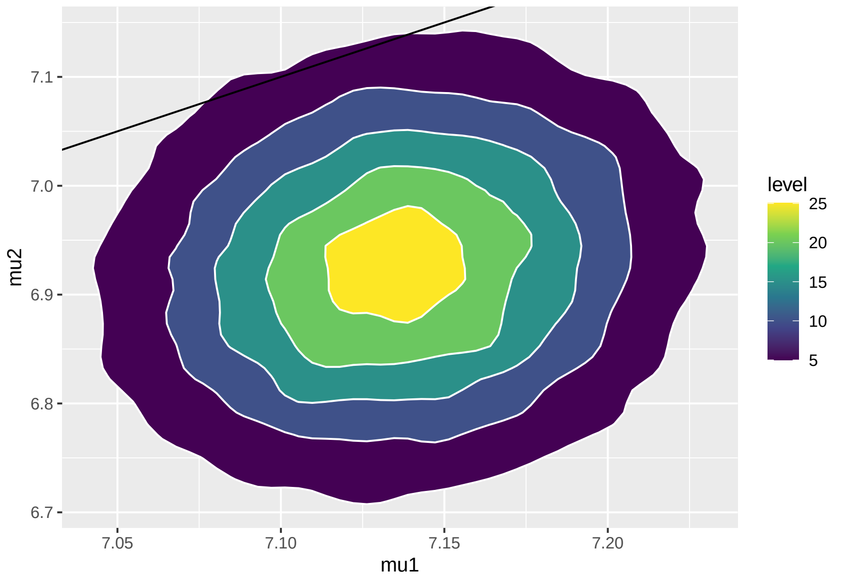
```
# Simulate
update(model, 1000, progress.bar = "none")

posterior_sample <- coda.samples(model,
                             variable.names = c("mu", "sigma"),
                             n.iter = Nrep,
                             progress.bar = "none")

sim_posterior = as.data.frame(as.matrix(posterior_sample))
names(sim_posterior) = c("mu1", "mu2", "sigma")
head(sim_posterior)
```

```
##     mu1   mu2 sigma
## 1 7.050 7.035 1.528
## 2 7.069 7.069 1.457
## 3 7.082 7.091 1.464
## 4 7.090 7.089 1.485
## 5 7.055 7.000 1.476
## 6 7.075 6.931 1.475
```

```
ggplot(sim_posterior, aes(mu1, mu2)) +
  geom_point(color = "seagreen", alpha = 0.4) +
  stat_ellipse(level = 0.98, color = "black", size = 2) +
  stat_density_2d(color = "grey", size = 1) +
  geom_abline(intercept = 0, slope = 1)
```



```
ggplot(sim_posterior, aes(mu1, mu2)) +
  stat_density_2d(aes(fill = ..level..),
          geom = "polygon", color = "white") +
 scale_fill_viridis_c() +
  geom_abline(intercept = 0, slope = 1)
```



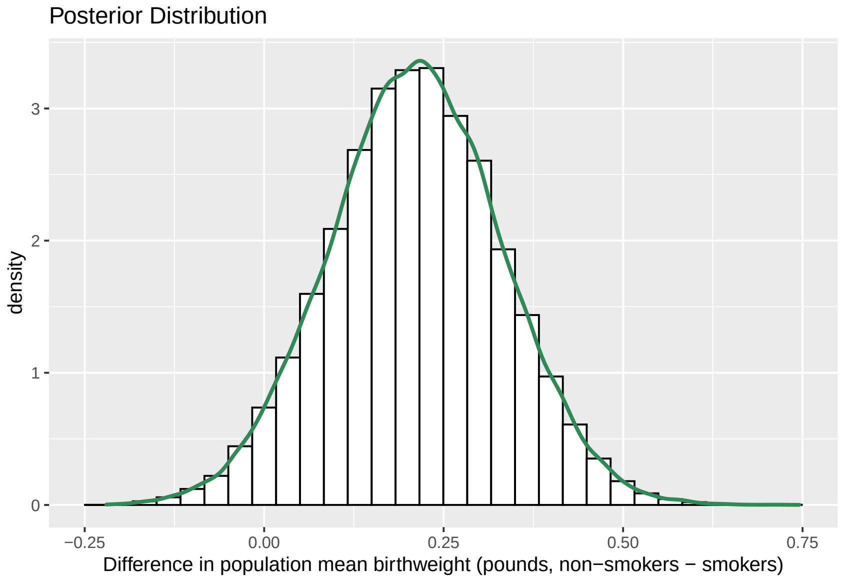
```
cor(sim_posterior$mu1, sim_posterior$mu2)
```

```
## [1] 0.127
```

The posterior mean of μ_1 is close to the sample mean birthweight for non-smokers, and the posterior mean of μ_2 is close to the sample mean birthweight for smokers. The posterior SD of μ_1 is smaller than that of μ_2, reflecting the larger sample size for non-smokers than smokers. The posterior distribution places most of its probability on (μ_1, μ_2) pairs with $\mu_1 > \mu_2$, representing a much stronger belief (than prior) that mean birthweight for smokers is less than for non-smokers. The posterior correlation between μ_1 and μ_2 is about 0.13, which is much smaller than the prior correlation. Even though there was fairly strong dependence between μ_1 and μ_2 in the prior, there was *independence between the samples in the data*, represented in the likelihood. With the large sample sizes (especially for non-smokers) the data has more influence on the posterior than the prior does.

8. The code below summarizes the posterior distribution of the difference in means $\mu_1 - \mu_2$. JAGS has already simulated (μ_1, μ_2) pairs from the posterior distribution, so we just need to compute $\mu_1 - \mu_2$ for each pair.

```
sim_posterior = sim_posterior %>%
  mutate(mu_diff = mu1 - mu2)

ggplot(sim_posterior,
   aes(mu_diff)) +
  geom_histogram(aes(y=..density..), color = "black", fill = "white") +
  geom_density(size = 1, color = "seagreen") +
  labs(x = "Difference in population mean birthweight (pounds, non-smokers - smokers)",
   title = "Posterior Distribution")
```



```
quantile(sim_posterior$mu_diff, c(0.01, 0.10, 0.25, 0.75, 0.90, 0.99))
```

```
##        1%       10%       25%       75%       90%       99% 
## -0.06925  0.05636  0.13074  0.29190  0.36488  0.48873
```

```
sum(sim_posterior$mu_diff > 0 ) / length(sim_posterior$mu_diff)
```

```
## [1] 0.9604
```

The posterior distribution of $\mu_1 - \mu_2$ is approximately Normal. The posterior mean of $\mu_1 - \mu_2$ is about 0.21 pounds, which is a compromise between the prior mean of $\mu_1 - \mu_2$ of 0 (no difference) and the difference in sample means of about 0.31 pounds.

There is a posterior probability of 50% that mean birthweight for non-smokers is between 0.13 and 0.29 pounds greater than mean birthweight for non-smokers.

There is a posterior probability of 80% that mean birthweight for non-smokers is between 0.06 and 0.36 pounds greater than mean birthweight for non-smokers.

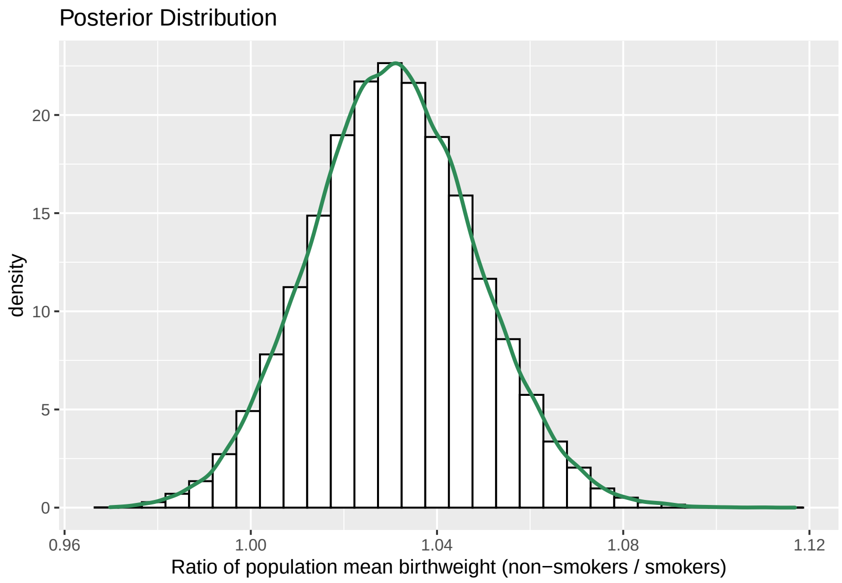
There is a posterior probability of 98% that mean birthweight for non-smokers is between 0.07 pounds less and 0.49 pounds greater than mean birthweight for non-smokers.

There is a posterior probability of about 96 percent that the mean birthweight for non-smokers is greater than the mean birthweight for smokers.

9. Putting a prior directly on $\mu_1 - \mu_2$ does allow us to make inference about the difference in mean birthweights. But what if we also want to estimate the mean birthweight for each group? Having a posterior distribution just on the difference between the two groups does not allow us to estimate the mean for either group. Also, $\mu_1 - \mu_2$ is the absolute difference in means, but what if we want to measure the difference in relative terms? Putting a prior distribution on (μ_1, μ_2) enables us to make posterior inference about mean birthweight for both non-smokers and smokers and any parameter (difference, ratio) that depends on the means.

10. See code and output below. JAGS has already simulated (μ_1, μ_2) pairs from the posterior distribution, so we just need to compute μ_1/μ_2 for each pair.

```
sim_posterior = sim_posterior %>%
  mutate(mu_ratio = mu1 / mu2)

ggplot(sim_posterior,
   aes(mu_ratio)) +
  geom_histogram(aes(y=..density..), color = "black", fill = "white") +
  geom_density(size = 1, color = "seagreen") +
  labs(x = "Ratio of population mean birthweight (non-smokers / smokers)",
   title = "Posterior Distribution")
```



```
quantile(sim_posterior$mu_ratio, c(0.01, 0.10, 0.25, 0.75, 0.90, 0.99))
```

```
##      1%    10%    25%    75%    90%    99%
## 0.9903 1.0080 1.0187 1.0426 1.0537 1.0732
```

The posterior distribution of μ_1/μ_2 is approximately Normal. The posterior mean of μ_1/μ_2 is about 1.03, which is a compromise between the prior mean of $\mu_1/\mu_2 = 1$ (no difference) and the ratio of sample means of 1.046.

There is a posterior probability of 50% that the mean birthweight of non-smokers is between 1.02 and 1.04 *times* greater than the mean birthweight of smokers.

There is a posterior probability of 80% that the mean birthweight of non-smokers is between 1.01 and 1.05 *times* greater than the mean birthweight of smokers.

There is a posterior probability of 98% that the mean birthweight of non-smokers is between 0.99 and 1.07 *times* greater than the mean birthweight of smokers.

11. Yes, there is some evidence. Even though we started with fairly strong prior credibility of no difference, with the relatively large sample sizes, the difference in sample means observed in the data was enough to overturn the prior beliefs. Now, the 98% credible interval for $\mu_1 - \mu_2$ does contain 0, indicating some plausibility of no difference. But there's nothing special

about 98% credibility, and we should look at the whole posterior distribution. According to our posterior distribution, we place a high degree of plausibility on the mean birthweight for smokers being less than the mean birthweight of non-smokers.

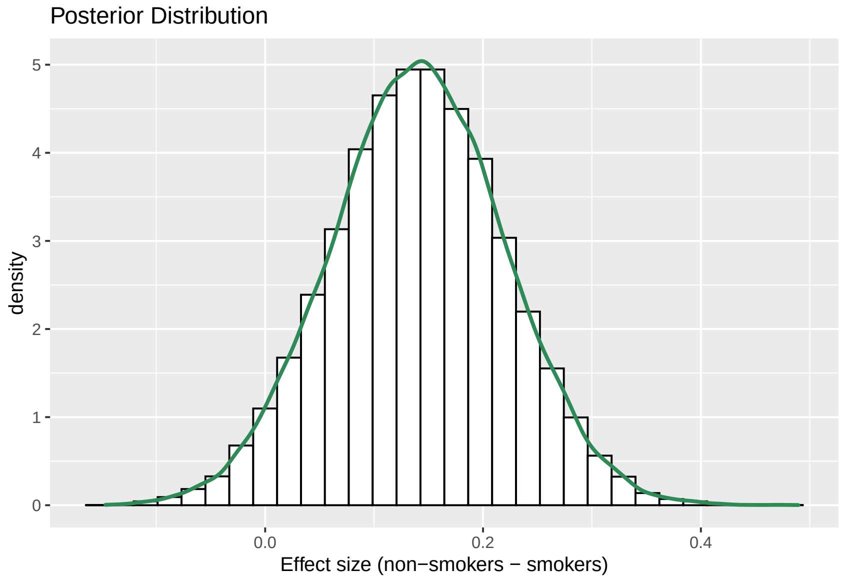
12. The question of causation has nothing to do with whether we are doing a Bayesian or frequentist analysis. Rather, the question of causation concerns: *how were the data collected?* In particular, was this an experiment with random assignment of the explanatory variable? It wasn't; it was an observational study (you can't randomly assign some mothers to smoke). Therefore, there is potential for confounding variables. Maybe mothers who smoke tend to be less healthy in general than mothers who don't smoke, and maybe some other aspect of health is more closely associated with lower birthweight than smoking is.

13. The posterior distribution of $\mu_1 - \mu_2$ does not give much plausibility to large differences in mean birthweight. Almost all of the posterior probability is placed on the absolute difference being less than 0.5 pounds, and the relative difference being no more than 1.07 times. Just because we have evidence that there is a difference, doesn't necessarily mean that the difference is large in practical terms.

14. The observed effect size is about $0.3/1.5 = 0.2$. Birthweights vary naturally from baby to baby by about 1.5 pounds, so a difference of 0.3 pounds seems relatively small. The sample mean birthweight for non-smokers is *0.2 standard deviations* greater than the sample mean birthweight for smokers.

    The following simulates and summarizes the posterior distribution of the population effect size. JAGS has already simulated (μ_1, μ_2, σ) triples for us; we just need to compute $(\mu_1 - \mu_2)/\sigma$ for each triple.

```
sim_posterior = sim_posterior %>%
  mutate(effect_size = (mu1 - mu2) / sigma)

ggplot(sim_posterior,
   aes(effect_size)) +
  geom_histogram(aes(y=..density..), color = "black", fill = "white") +
  geom_density(size = 1, color = "seagreen") +
  labs(x = "Effect size (non-smokers - smokers)",
   title = "Posterior Distribution")
```



```
quantile(sim_posterior$effect_size, c(0.01, 0.10, 0.25, 0.75, 0.90, 0.99))
```

```
##       1%      10%      25%      75%      90%      99% 
## -0.04583  0.03741  0.08694  0.19396  0.24273  0.32664
```

The posterior mean of $(\mu_1 - \mu_2)/\sigma$ is about 0.14, which is a compromise between the prior mean of $(\mu_1 - \mu_2)/\sigma$ of 0 (no difference) and the sample effect size of 0.2.

There is a posterior probability of 50% that the mean birthweight of non-smokers is between 0.09 and 0.19 *standard deviations* greater than the mean birthweight of smokers.

There is a posterior probability of 80% that the mean birthweight of non-smokers is between 0.04 and 0.24 *standard deviations* greater than the mean birthweight of smokers.

There is a posterior probability of 98% that the mean birthweight of non-smokers is between 0.05 standard deviations less than and 0.33 *standard deviations* greater than the mean birthweight of smokers.

The posterior distribution indicates that the effect size is pretty small. The difference between mean birthweight of smokers and non-smokers is small relative to the variability in birthweights.

(Of course, smoking has many other adverse health effects. But looking at birthweight alone, based on this data set we cannot conclude that there is a large difference in mean birthweight between smokers and non-smokers.)

The previous example introduced many important ideas.

Independence in the data versus in the prior/posterior

It is typical to assume *independence in the data*, e.g., independence of values of the measured variables within and between samples (conditional on the parameters). Whether independence in the data is a reasonable assumption depends on how the data is collected.

But whether it is reasonable to assume prior *independence of parameters* is a completely separate question and is dependent upon our subjective beliefs about any relationships between parameters.

Transformations of parameters

The primary output of a Bayesian data analysis is the full joint posterior distribution on all parameters. Given the joint distribution, the distribution of transformations of the primary parameters is readily obtained.

Effect size for comparing means

When comparing groups, a more important question than "is there a difference?" is "*how large* is the difference?" An **effect size** is a measure of the magnitude of a difference between groups. A difference in parameters can be used to measure the absolute size of the difference in the measurement units of the variable, but effect size can also be measured as a relative difference.

When comparing a numerical variable between two groups, one measure of the population effect size is **Cohens's** d

$$\frac{\mu_1 - \mu_2}{\sigma}$$

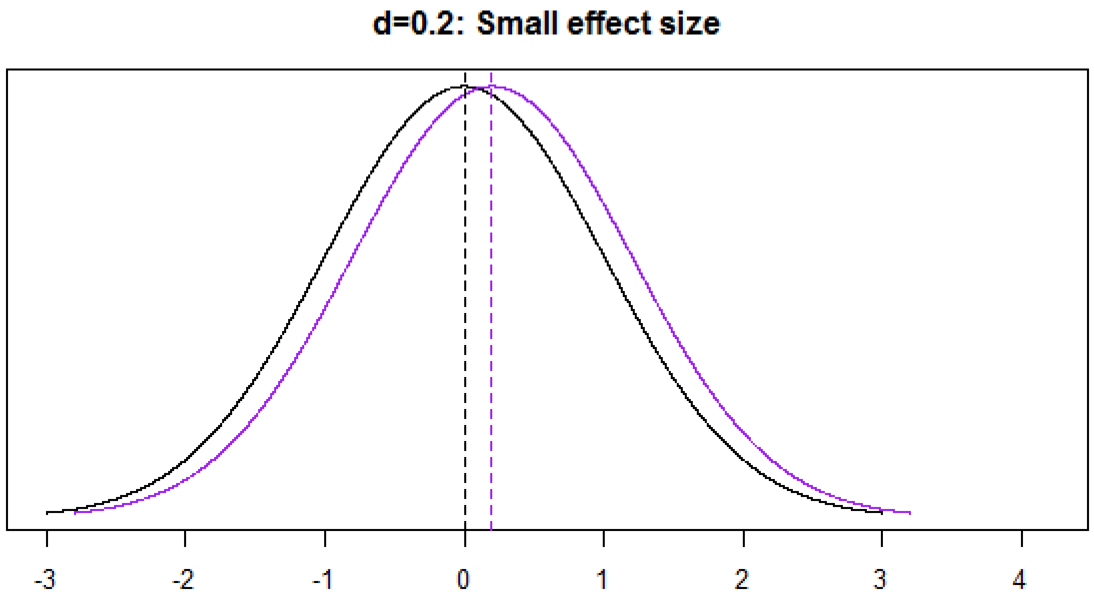
The values of any numerical variable vary naturally from unit to unit. The SD of the numerical variable measures the degree to which individual values of the variable vary naturally, so the SD provides a natural "scale" for the variable. Cohen's d compares the magnitude of the difference in means relative to the natural scale (SD) for the variable

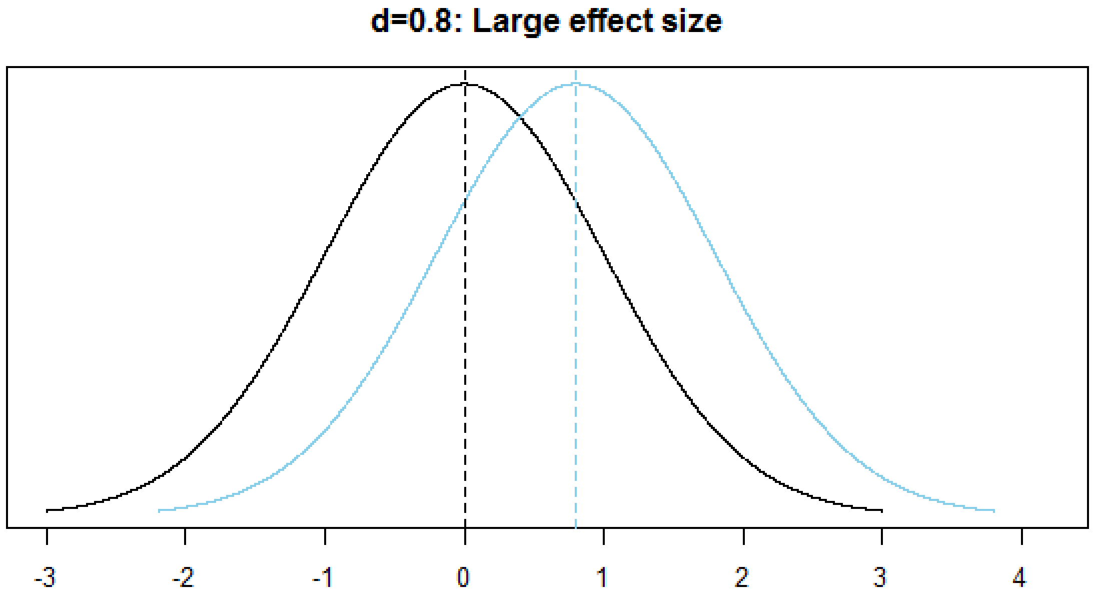
Some rough guidelines for interpreting $|d|$:

d	0.2	0.5	0.8	1.2	2.0
Effect size	Small	Medium	Large	Very Large	Huge

For example, assume the two population distributions are Normal and the two population standard deviations are equal. Then when the effect size is 1.0 the median of the distribution with the higher mean is the 84th percentile of the distribution with the lower mean, which is a very large difference.

d	0.2	0.5	0.8	1.0	1.2	2.0
Effect size	Small	Medium	Large		Very Large	Huge
Median of population 1 is (blank) percentile of population 2	58th	69th	79th	84th	89th	98th





Notes on the JAGS code.

- You should be able to define the prior parameters for the Multivariate Normal distribution within JAGS, but I keep getting an error. So I'm defining prior parameters outside of JAGS and then passing them in with the data. (I can never remember what you can do in JAGS and what you need to do in R and pass to JAGS.)

- The Bivariate Normal prior is coded in JAGS using `dmnorm.vcov` which has two parameters: a mean *vector* and a covariance *matrix*. (There is also `dmnorm` is which is parametrized by the precision matrix.)
- The prior `mu[1:2] ~ dmnorm(...)` creates a vector `mu` with two components `mu[1]` and `mu[2]`. When `"mu"` is called in the `variable.names = c("mu", "sigma")` argument of `coda.samples` JAGS will return the vector `mu` — that is, both components `mu[1]` and `mu[2]`. See the output of posterior_sample.
- Group variables (like non-smoker/smoker) need to be coded as numbers in JAGS, starting with 1. So `x` recodes smoking status as 1 for non-smokers and 2 for smokers.
- We have data on individual birthweights, so we evaluate the likelihood of each individual value `y[i]` using a Normal distribution and then use a for loop to find the likelihood for the sample.
- Notice that the mean used in the likelihood depends on the group: `mu[x[i]]`. For example, if element `i` has birthweight `y[i] = 8` and is a non-smoker `x[i] = 1`, then the likelihood is evaluated using a Normal distribution with mean μ_1; for this element `y[i] ~ dnorm(mu[x[i]], ...)` in JAGS is like calling `dnorm(y[i], mu[x[i]], ...) = dnorm(8, mu[1], ...)` in R. If element `i` has birthweight `y[i] = 7.3` and is a smoker `x[i] = 2`, then the likelihood is evaluated using a Normal distribution with mean μ_2; for this element `y[i] ~ dnorm(mu[x[i]], ...)` in JAGS is like calling `dnorm(y[i], mu[x[i]], ...) = dnorm(7.3, mu[2], ...)` in R.
- The `variable.names = c("mu", "sigma")` argument of `coda.samples` tells JAGS which simulation output to save. Given the joint posterior distributon of the primary parameters, it is relatively easy to obtain the posterior distribution of transformations of these parameters outside of JAGS in R.

Non-Normal Likelihood

The sample data exhibits a long left tail, similar to what we observed in the previous chapter. Therefore, a non-Normal model might be more appropriate for the distribution of birthweights. The code and output below uses a t-distribution for the likelihood, similar to what was done in the previous section. The posterior distribution of (μ_1, μ_2) is fairly similar to what was computed above for the model with the Normal likelihood. The model below based on the t-distribution likelihood shifts posterior credibility a little more towards mean birthweights for non-smokers being greater than birthweights for smokers. But the difference is still small in absolute terms; at most 0.5 pounds or so. In terms of comparing population means, the choice of likelihood (Normal versus t) does not make much of a difference in this example. That is, the *inference* regarding $\mu_1 - \mu_2$ appears not to be too sensitive to the choice of likelihood. However, if we were using the model to *predict* birthweights, then the t-distribution based model might be more appropriate, as we observed in the previous chapter.

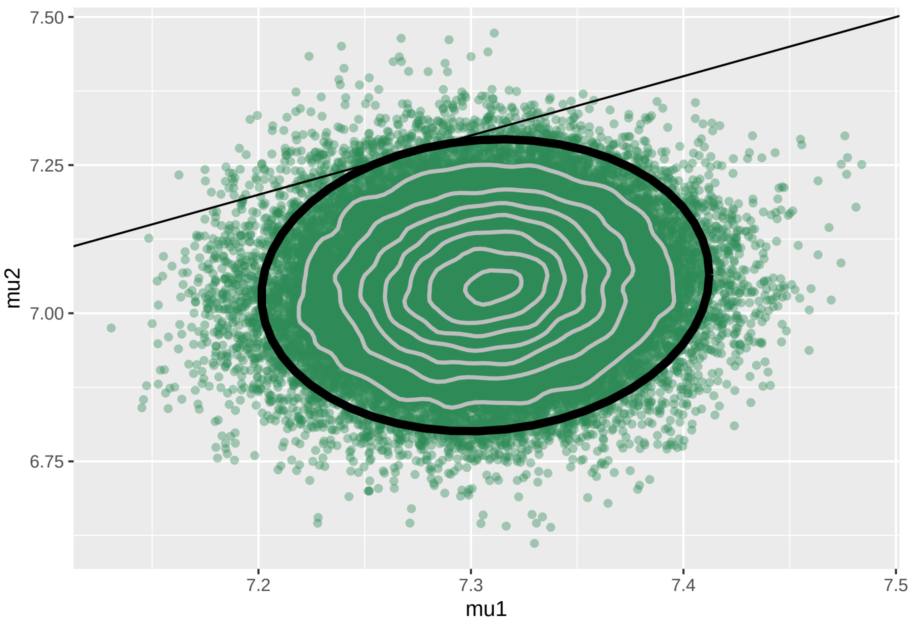
```
# Model
model_string <- "model{

  # Likelihood
  for (i in 1:n){
      y[i] ~ dt(mu[x[i]], 1 / sigma ^ 2, tdf)
  }

  # Prior
  mu[1:n_groups] ~ dmnorm.vcov(mu_prior_mean[1:n_groups],
                              mu_prior_cov[1:n_groups, 1:n_groups])

  sigma ~ dgamma(1, 1)

  tdf <- 1 + tdf0

  tdf0 ~ dexp(1 / 29)

}"

dataList = list(y = y, x = x, n = n, n_groups = n_groups,
                mu_prior_mean = mu_prior_mean, mu_prior_cov = mu_prior_cov)

# Compile
Nrep = 10000

n.chains = 5

model <- jags.model(textConnection(model_string),
                    data = dataList,
                    n.chains = n.chains)
```

```
## Compiling model graph
##    Resolving undeclared variables
##    Allocating nodes
## Graph information:
##    Observed stochastic nodes: 999
##    Unobserved stochastic nodes: 3
##    Total graph size: 2020
##
## Initializing model
```

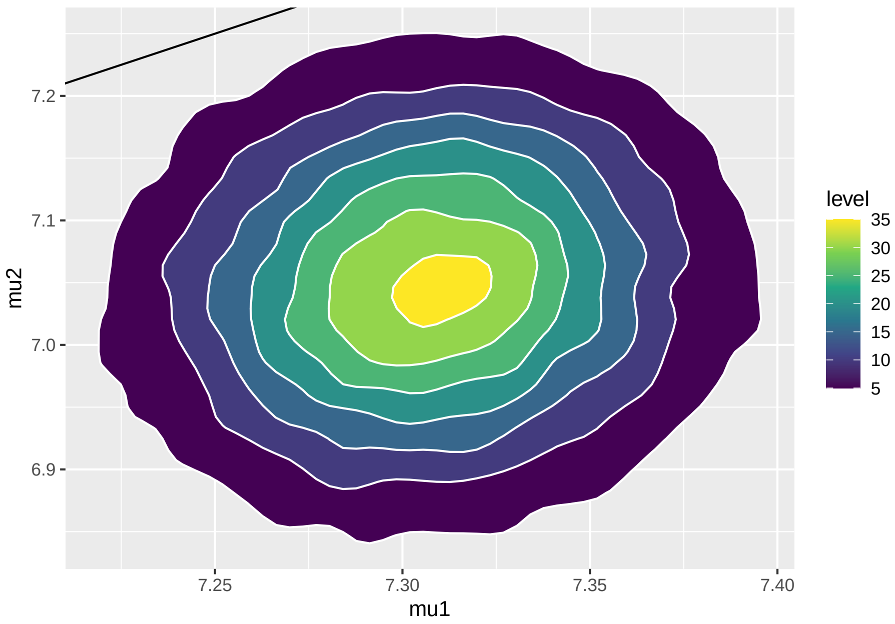
```
# Simulate
update(model, 1000, progress.bar = "none")

posterior_sample <- coda.samples(model,
                                 variable.names = c("mu", "sigma", "tdf"),
                                 n.iter = Nrep,
                                 progress.bar = "none")


sim_posterior = as.data.frame(as.matrix(posterior_sample))
names(sim_posterior) = c("mu1", "mu2", "sigma", "tdf")
head(sim_posterior)
```

```
##     mu1   mu2 sigma   tdf
## 1 7.225 6.850 1.080 3.953
## 2 7.219 6.982 1.061 4.057
## 3 7.217 6.982 1.094 4.248
## 4 7.232 7.042 1.034 2.978
## 5 7.242 6.958 1.000 2.934
## 6 7.275 6.970 1.012 3.954
```

```
ggplot(sim_posterior, aes(mu1, mu2)) +
  geom_point(color = "seagreen", alpha = 0.4) +
  stat_ellipse(level = 0.98, color = "black", size = 2) +
  stat_density_2d(color = "grey", size = 1) +
  geom_abline(intercept = 0, slope = 1)
```



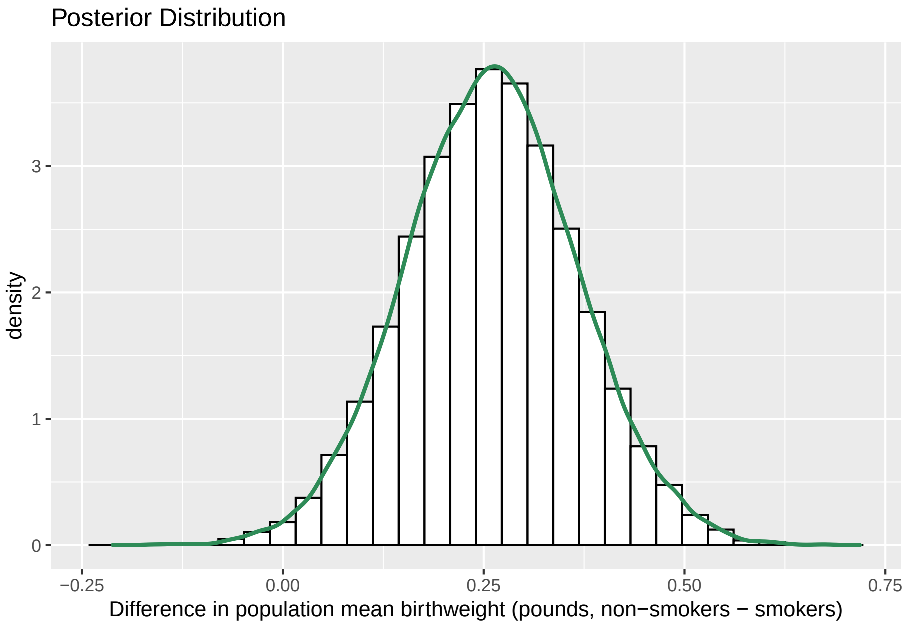
```
ggplot(sim_posterior, aes(mu1, mu2)) +
  stat_density_2d(aes(fill = ..level..),
              geom = "polygon", color = "white") +
 scale_fill_viridis_c() +
  geom_abline(intercept = 0, slope = 1)
```



```
cor(sim_posterior$mu1, sim_posterior$mu2)
```

```
## [1] 0.06929
```

```
sim_posterior = sim_posterior %>%
      mutate(mu_diff = mu1 - mu2)

ggplot(sim_posterior,
       aes(mu_diff)) +
  geom_histogram(aes(y=..density..), color = "black", fill = "white") +
  geom_density(size = 1, color = "seagreen") +
  labs(x = "Difference in population mean birthweight (pounds, non-smokers - smokers)",
       title = "Posterior Distribution")
```



```
quantile(sim_posterior$mu_diff, c(0.01, 0.10, 0.25, 0.75, 0.90, 0.99))
```

```
##      1%     10%     25%     75%     90%     99% 
## 0.01004 0.12352 0.18782 0.33075 0.39720 0.51219
```

```
sum(sim_posterior$mu_diff > 0 ) / length(sim_posterior$mu_diff)
```

```
## [1] 0.992
```

Chapter 17

Introduction to Markov Chain Monte Carlo (MCMC) Simulation

Bayesian data analysis is based on the posterior distribution of relevant parameters given the data. However, in many situations the posterior distribution cannot be determined analytically or via grid approximation. Therefore we use simulation to approximate the posterior distribution and its characteristics. But in many situations it is difficult or impossible to simulate directly from a distribution, so we turn to indirect methods.

Markov chain Monte Carlo (MCMC)[1] methods provide powerful and widely applicable algorithms for simulating from probability distributions, including complex and high-dimensional distributions.

Example 17.1. A politician campaigns on a long east-west chain of islands[2]. At the end of each day she decides to stay on her current island, move to the island to the east, or move to the island to the west. Her goal is to visit all the islands proportional to their population, so that she spends the most days on the most populated island, and proportionally fewer days on less populated islands. But, (1) she doesn't know how many islands there are, and (2) she doesn't know the population of each island. However, when she visits an island she can determine its population. And she can send a scout to the east/west adjacent islands to determine their population before visiting. How can the politician achieve her goal in the long run?

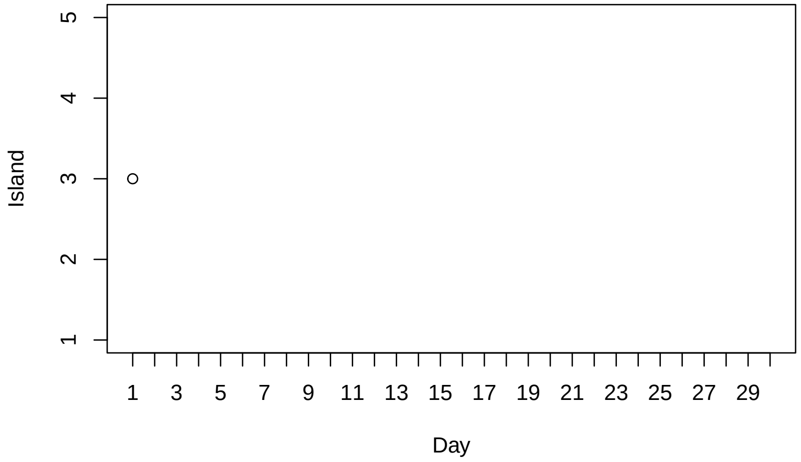
[1] For some history, and an origin of the use of the the term "Monte Carlo", see Wikipedia. Monte Carlo methods consist of a broad class of algorithms for obtaining numerical results based on random numbers, even in problems that don't explicitly involve probability (e.g., Monte Carlo integration).

[2] This island hopping example is inspired by Kruschke, *Doing Bayesian Data Analysis*.

Suppose that every day, the politician makes her travel plans according to the following algorithm.

- She flips a fair coin to *propose* travel to the island to the east (heads) or west (tails). (If there is no island to the east/west, treat it as an island with population zero below.)
- If the proposed island has a population greater than that of the current island, then she travels to the proposed island.
- If the proposed island has a population less than that of the current island, then:
  - She computes a, the ratio of the population of the proposed island to the current island.
  - She travels to the proposed island with probability a,
  - And with probability $1 - a$ she spends another day on the current island.

1. Suppose there are 5 islands, labeled $1, \dots, 5$ from west to east, and that island θ has population θ (thousand), and that she starts at island 3 on day 1. How could you use a coin and a spinner to simulate *by hand* the politician's movements over a number of days? Conduct the simulation and plot the path of her movements.



2. Construct *by hand* a plot displaying the simulated relative frequencies of days in each of the 5 islands.
3. Now write code to simulate the politician's movements for many days. Plot the island path.

4. Plot the simulated relative frequencies of days in each of the 5 islands.

5. Recall the politician's goal of visiting each island in proportion to its population. Given her goal, what should the plot from the previous part look like? Does the algorithm result in a reasonable approximation?

6. Suppose again that the number of islands and their populations is unknown, but that the population for island θ is proportional to $\theta^2 e^{-0.5\theta}$, where the islands are labeled from west to east 1, 2, ... Based on this information, can she implement her algorithm? That is, is it sufficient to know that the populations are *in proportion to* $\theta^2 e^{-0.5\theta}$ without knowing the actual populations? In particular, is there enough information to compute the "acceptance probability" a?

7. Write code to run the algorithm for the previous situation and compare the simulated relative frequencies to the target distribution. Does the algorithm result in a reasonable approximation?

8. Why doesn't the politician just always travel to or stay on the island with the larger population?

9. Is the next island visited dependent on the current island?

10. Is the next island visited dependent on *how she got to* the current island? That is, given the current island is her next island independent of her past history of previous movements?

11. What would happen if there were an island among the east-west chain with population 0? (For example, suppose there are 10 islands, she starts on island 1, and island 5 has population 0.) How could she modify her algorithm to address this issue?

Solution. to Example 17.1

1. Starting at island θ, flip a coin to propose a move to either $\theta - 1$ or $\theta + 1$. If the proposed move is to $\theta + 1$, it will be accepted because the population is larger. If the proposed move is to $\theta - 1$ it will be accepted with probability $(\theta - 1)/\theta$; otherwise the move will be rejected and she will stay in the current island.

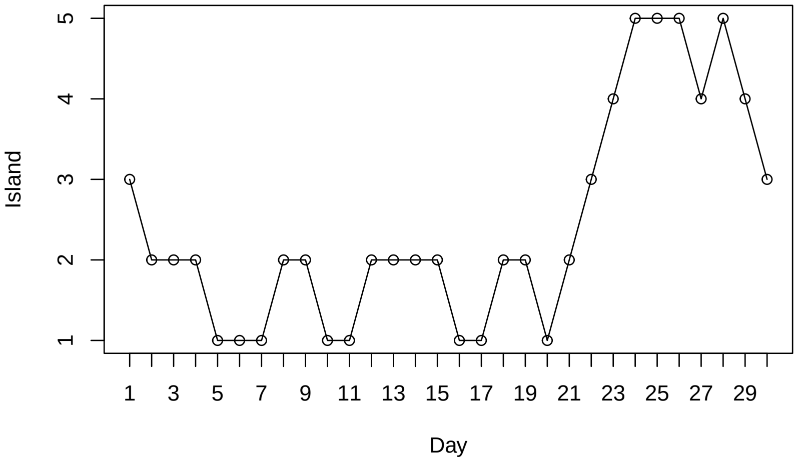
   For example, starting in island 3 she proposes a move to either island 2 or island 4. If the proposed move is to island 4, it is accepted and she moves to island 4. If the proposed move is to island 2, it is accepted with probability 2/3. If it is accepted she moves to island 2; otherwise she stays at island 3 for another day.

   If she is on island 1 and she proposes a move to the west, the proposal is rejected (because the population of island "0" is 0) and she spends another day on island 1. Likewise if she proposes a move to the east from island 5.

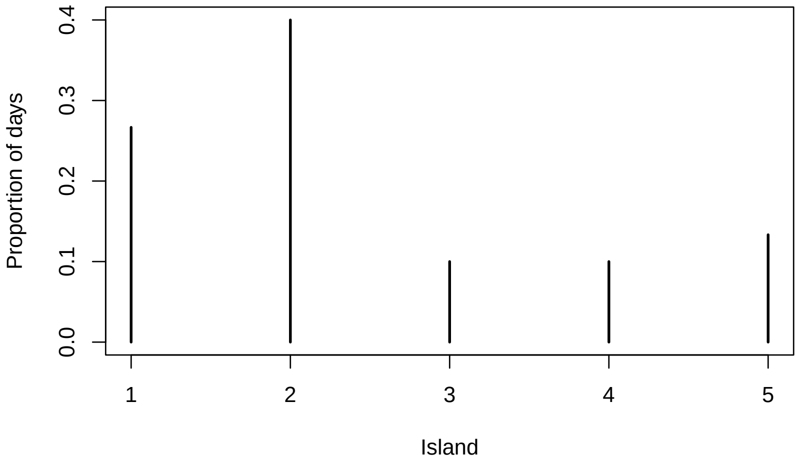
The acceptance probabilities are

$$
\begin{aligned}
a(1 \to 0) &= 0 & \qquad a(1 \to 2) &= 1 \\
a(2 \to 1) &= 1/2 & \qquad a(2 \to 3) &= 1 \\
a(3 \to 2) &= 2/3 & \qquad a(3 \to 4) &= 1 \\
a(4 \to 3) &= 3/4 & \qquad a(4 \to 5) &= 1 \\
a(5 \to 4) &= 4/5 & \qquad a(5 \to 6) &= 0
\end{aligned}
$$

Here is an example plot for 30 days.

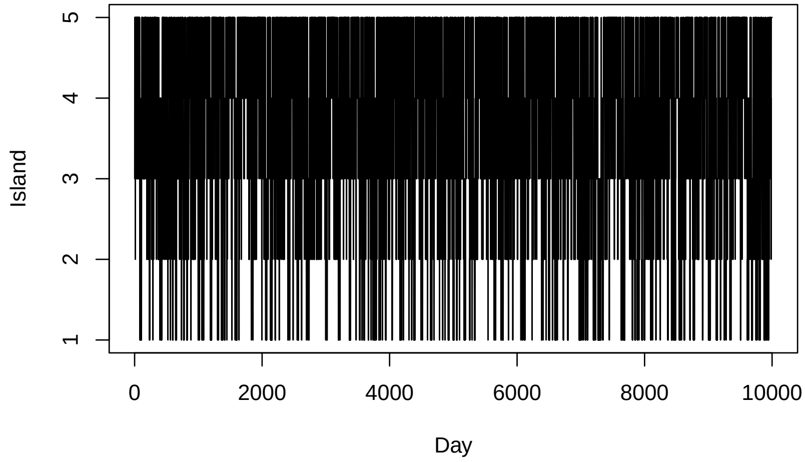


2. The plot below corresponds to the path from the previous plot.



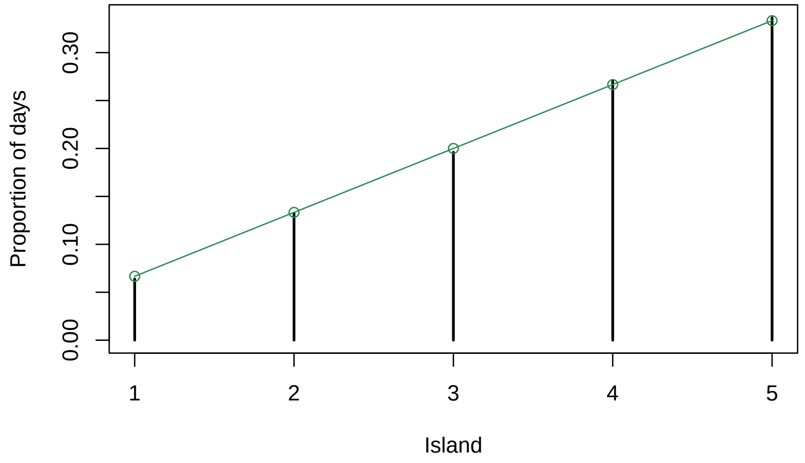
3. Some code is below. There are different ways to implement this algorithm, but note the proposal and acceptance steps below.

```r
n_states = 5
theta = 1:n_states
pi_theta = theta

n_steps = 10000
theta_sim = rep(NA, n_steps)
theta_sim[1] = 3 # initialize

for (i in 2:n_steps){
  current = theta_sim[i - 1]
  proposed = sample(c(current + 1, current - 1), size = 1, prob = c(0.5, 0.5))
  if (!(proposed %in% theta)){ # to correct for proposing moves outside of boundaries
proposed = current
  }
  a = min(1, pi_theta[proposed] / pi_theta[current])
  theta_sim[i] = sample(c(proposed, current), size = 1, prob = c(a, 1-a))
}

# trace plot
plot(1:n_steps, theta_sim, type = "l", ylim = range(theta), xlab = "Day", ylab = "Island")
```



4. The plot below corresponds to the path from the previous plot.

```
plot(table(theta_sim) / n_steps, xlab = "Island", ylab = "Proportion of days")

points(theta, theta / 15, type = "o", col = "seagreen")
```



5. Since the population of island θ is proportional to θ, the target probability distribution satisfies $\pi(\theta) \propto \theta, \theta = 1, \ldots, 5$. That is, $\pi(\theta) = \theta/15, \theta =$

$1, \ldots, 5$. This distribution is depicted in green in the previous plot. We see that the algorithm does produce a reasonable approximation.

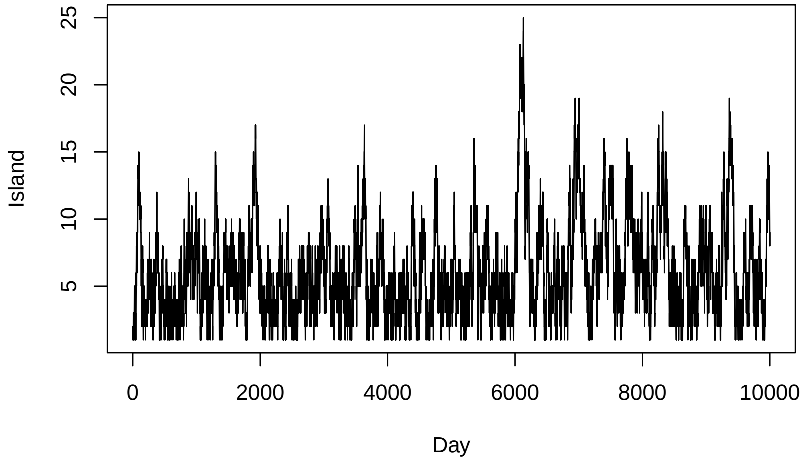
6. She can still make east-west proposals with a coin flip. And she can still decide whether to accept the proportion based on relative population. If she is currently on island θ_{current} and she proposes a move to island θ_{proposed}, she will accept the proposal with probability

$$a(\theta_{\text{current}} \to \theta_{\text{proposed}}) = \frac{\theta^2_{\text{proposed}} e^{-0.5\theta_{\text{proposed}}}}{\theta^2_{\text{current}} e^{-0.5_{\text{current}}}}$$

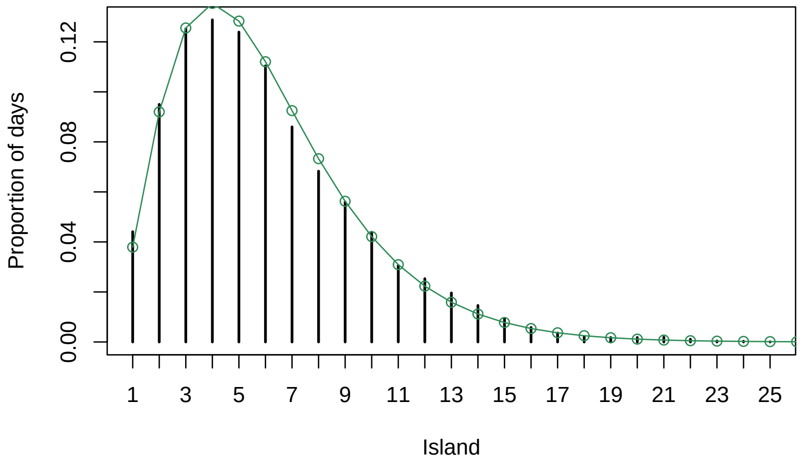
In terms of running the algorithm it is sufficient to know that the populations are *in proportion to* $\theta^2 e^{-0.5\theta}$ without knowing the actual populations.

7. See below; it seems like a reasonable approximation.

```
n_states = 30
theta = 1:n_states
pi_theta = theta ^ 2 * exp(-0.5 * theta) # notice: not probabilities


n_steps = 10000
theta_sim = rep(NA, n_steps)
theta_sim[1] = 1 # initialize

for (i in 2:n_steps){
  current = theta_sim[i - 1]
  proposed = sample(c(current + 1, current - 1), size = 1, prob = c(0.5, 0.5))
  if (!(proposed %in% theta)){ # to correct for proposing moves outside of boundaries
proposed = current
  }
  a = min(1, pi_theta[proposed] / pi_theta[current])
  theta_sim[i] = sample(c(proposed, current), size = 1, prob = c(a, 1-a))
}

# trace plot
plot(1:n_steps, theta_sim, type = "l", ylim = range(theta_sim), xlab = "Day", ylab = "Island
```



```r
plot(table(theta_sim) / n_steps, xlab = "Island", ylab = "Proportion of days")
points(theta, pi_theta / sum(pi_theta), type = "o", col = "seagreen")
```



8. She wants to visit the islands in proportion to their population. So she still wants to visit the smaller islands, just not as often as the larger ones.

But if she always move towards islands with larger populations, she would not visit the smaller ones at all.

9. Yes, the next island visited dependent on the current island. For example, if she is on island 3 today, tomorrow she can only be on island 2, 3, or 4.
10. No the next island visited is not dependent on *how she got to* the current island. The proposals and acceptance probability only depend on the current state, and not past states (given the current state).
11. With only east-west proposals, since she would never visit an island with population 0 (because the acceptance probability would be 0), she could never get to islands on the other side. She could modify her algorithm to cast a wider net in her proposals, instead of just proposing moves to adjacent islands.

The goal of a Markov chain Monte Carlo method is to simulate from a probability distribution of interest. In Bayesian contexts, the distribution of interest will usually be the posterior distribution of parameters given data.

A **Markov chain** is a random process that exhibits a special "one-step" dependence structure. Namely, conditional on the most recent value, any future value is conditionally independent of any past values. In a Markov chain: "Given the present, the future is conditionally independent of the past." Roughly, in terms of simulating the next value of a Markov chain, all that matters is where you are now, not how you got there.

The idea of MCMC is to build a Markov chain whose long run distribution — that is, the distribution of state visits after a large number of "steps" — is the probability distribution of interest. Then we can *indirectly* simulate a representative sample from the probability distribution of interest, and use the simulated values to approximate the distribution and its characteristics, by running an appropriate Markov chain for a sufficiently large number of steps. The Markov chain does not need to be fully specified in advance, and is often constructed "as you go" via an algorithm like a "modified random walk". Each step of the Markov chain typically involves

- A **proposal** for the next state, which is generated according to some known probability distribution or mechanism,
- A **decision of whether or to accept the proposal**. The decision usually involves probability
  - With probability a, accept the proposal and step to the next state
  - With probability $1-a$, reject the proposal and remain in the current state for the next step.

In principle, proposals can be fairly naive and not related to the target distribution (though in practice choice of proposal is very important since it affects computational efficiency). Furthermore, the target distribution of interest only needs to be specified up to a constant of proportionality, and the state space of possible values does not need to be fully specified in advance.

The island hopping example illustrated an MCMC algorithm for a discrete parameter θ. Recall that most parameters in statistical models take values on a continuous scale, so most posterior distributions are continuous distributions. MCMC simulation can also be used to approximate the posterior distribution and related characteristics for continuous distributions.

The following examples illustrates how MCMC can be used to approximate the posterior distribution in a Beta-Binomial setting. Of course, this scenario can be handled analytically and so MCMC is not necessary. However, it will help to see how the ideas work in a familiar context where an analytical solution is available.

Example 17.2. Suppose we wish to estimate θ, the proportion of Cal Poly students who have read a non-school related book in 2022. Assume a Beta(1, 3) prior distribution for θ. In a sample of 25 Cal Poly students, 4 have read a book in 2021. We'll use MCMC to approximate the posterior distribution of θ.

1. Without actually computing the posterior distribution, what can we say about it based on the assumptions of the model?
2. What are the possible "states" that we want our Markov chain to visit?
3. Given a current state θ_{current} how could we *propose* a new value θ_{proposed}, using a continuous analog of "random walk to neighboring states"?
4. How would we decide whether to accept the proposed move? How would we compute the probability of accepting the proposed move?
5. Suppose the current state is $\theta_{\text{current}} = 0.20$ and the proposed state is $\theta_{\text{proposed}} = 0.15$. Compute the probability of accepting this proposal.
6. Suppose the current state is $\theta_{\text{current}} = 0.20$ and the proposed state is $\theta_{\text{proposed}} = 0.25$. Compute the probability of accepting this proposal.
7. Write code to run the algorithm and plot the simulated values of θ.
8. What is the posterior distribution? Does the distribution of simulated values of θ provide a reasonable approximation to the posterior distribution?

Solution. to Example 17.2

1. Since posterior is proportional to likelihood times prior, we know that

$$\begin{aligned}\pi(\theta|y=4) &\propto f(y=4|\theta)\pi(\theta)\\ &\propto \left(\theta^4(1-\theta)^{25-4}\right)\left(\theta^{1-1}(1-\theta)^{3-1}\right)\end{aligned}$$

2. θ takes values in (0, 1) so each possible value in (0, 1) is a state.

3. There are different approaches, but here's a common one. We want to propose a new state in the neighborhood of the current state. Given θ_{current}, propose θ_{proposed} from a $N(\theta_{\text{current}}, \delta)$ distribution where the standard deviation δ represents the size of the "neighborhood". For example, if $\theta_{\text{current}} = 0.5$ and $\delta = 0.05$ then we would draw the proposal from the $N(0.5, 0.05)$ distribution, so there's a 68% chance the proposal is between 0.45 and 0.55 and a 95% chance that it's between 0.40 and 0.60.

4. Remember, the goal is to approximate the posterior distribution, so we want to visit the θ states in proportion to their posterior density. If the proposed state has higher posterior density than the current state, $\pi(\theta_{\text{proposed}}|y=4) > \pi(\theta_{\text{current}}|y=4)$, then we accept the proposal. Otherwise, accept the proposal with a probability based on the relative posterior densities of the proposed and current states. That is, we accept the proposed move with probability

$$a(\theta_{\text{current}} \to \theta_{\text{proposed}}) = \min\left(1, \frac{\pi(\theta_{\text{proposed}}|y=4)}{\pi(\theta_{\text{current}}|y=4)}\right) = \min\left(1, \frac{\left(\theta_{\text{proposed}}^4 (1-\theta_{\text{proposed}})^{25-4}\right)\left(\theta_{\text{proposed}}^{1-1} (1-\theta_{\text{proposed}})^{3-1}\right)}{\left(\theta_{\text{current}}^4 (1-\theta_{\text{current}})^{25-4}\right)\left(\theta_{\text{current}}^{1-1} (1-\theta_{\text{current}})^{3-1}\right)}\right)$$

5. The posterior density is larger for the proposed state, so the proposed move is accepted with probability 1.

$$a(0.20 \to 0.15) = \min\left(1, \frac{\pi(0.15|y=4)}{\pi(0.20|y=4)}\right) = \min\left(1, \frac{\left(0.15^4 (1-0.15)^{25-4}\right)\left(0.15^{1-1} (1-0.15)^{3-1}\right)}{\left(0.20^4 (1-0.20)^{25-4}\right)\left(0.20^{1-1} (1-0.20)^{3-1}\right)}\right) = 1$$

```
(dbeta(0.15, 1, 3) * dbinom(4, 25, 0.15)) / (dbeta(0.2, 1, 3) * dbinom(4, 25, 0.2))
```

```
## [1] 1.276
```

6. The posterior density is smaller for the proposed state, so based on the ratio of the posterior densities, the proposed move is accepted with probability 0.553.

$$a(0.20 \to 0.25) = \min\left(1, \frac{\pi(0.25|y=4)}{\pi(0.20|y=4)}\right) = \min\left(1, \frac{\left(0.25^4 (1-0.25)^{25-4}\right)\left(0.25^{1-1} (1-0.25)^{3-1}\right)}{\left(0.20^4 (1-0.20)^{25-4}\right)\left(0.20^{1-1} (1-0.20)^{3-1}\right)}\right) = 0.553$$

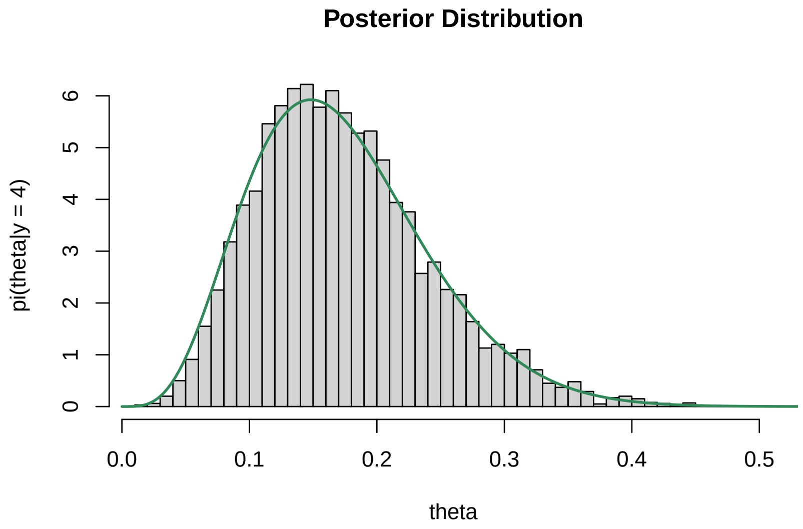
```
(dbeta(0.25, 1, 3) * dbinom(4, 25, 0.25)) / (dbeta(0.2, 1, 3) * dbinom(4, 25, 0.2))
```

```
## [1] 0.5533
```

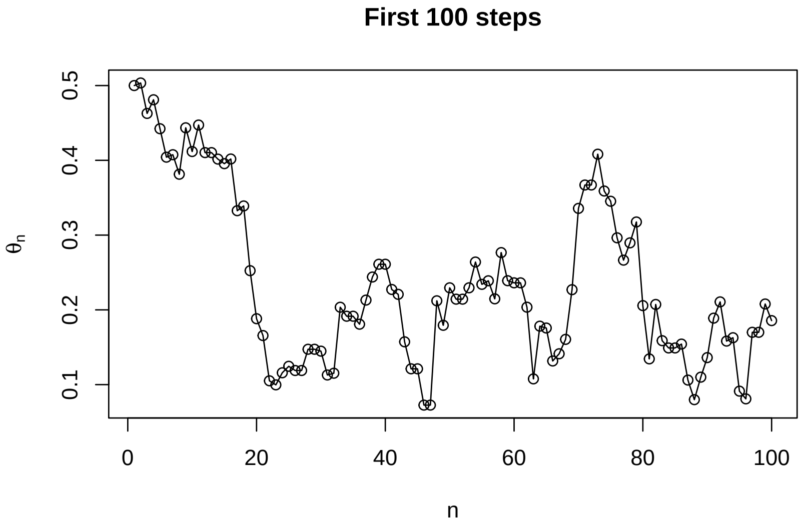
7. See below. The Normal distribution proposal can propose values outside of (0, 1), so we set $\pi(\theta|y=4)$ equal to 0 for $\theta \notin (0,1)$. This way, proposals to states outside (0, 1) will never be accepted.

```
n_steps = 10000
delta = 0.05
```

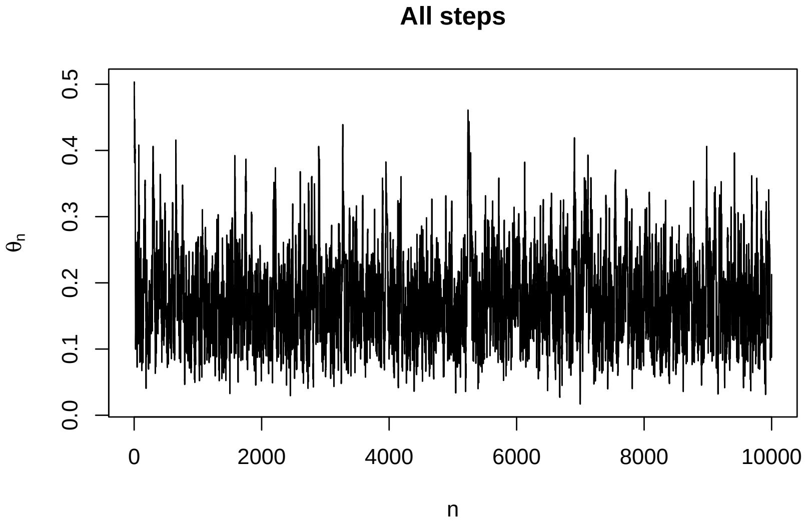
```
theta = rep(NA, n_steps)
theta[1] = 0.5 # initialize

# Posterior is proportional to prior * likelihood
pi_theta <- function(theta) {
  if (theta > 0 & theta < 1) dbeta(theta, 1, 3) * dbinom(4, 25, theta) else 0
}

for (n in 2:n_steps){
  current = theta[n - 1]
  proposed = current + rnorm(1, mean = 0, sd = delta)
  accept = min(1, pi_theta(proposed) / pi_theta(current))
  theta[n] = sample(c(current, proposed), 1, prob = c(1 - accept, accept))
}

# simulated values of theta
hist(theta, breaks = 50, freq = FALSE,
 xlab = "theta", ylab = "pi(theta|y = 4)", main = "Posterior Distribution")

# plot of theoretical posterior density of theta
x_plot= seq(0, 1, 0.0001)
lines(x_plot, dbeta(x_plot, 1 + 4, 3 + 25 - 4), col = "seagreen", lwd = 2)
```



```
plot(1:100, theta[1:100], type = "o", xlab = "n",
 ylab = expression(theta[n]), main = "First 100 steps")
```



```
plot(1:n_steps, theta, type = "l", xlab = "n",
 ylab = expression(theta[n]), main = "All steps")
```



8. The theoretical posterior distribution is the Beta(5, 24) distribution, depicted in green above. The distribution of simulated values of θ provides a reasonable approximation to the posterior distribution.

The goal of an MCMC method is to simulate θ values from a probability distribution $\pi(\theta)$. One commonly used MCMC method is the **Metropolis algorithm**[3] To generate θ_{new} given θ_{current}:

1. Given the current value θ_{current} propose a new value θ_{proposed} according to the *proposal (or "jumping") distribution j*.

   $$j(\theta_{\text{current}} \to \theta_{\text{proposed}})$$

   is the conditional density that θ_{proposed} is proposed as the next state given that θ_{current} is the current state
2. Compute the *acceptance probability* as the ratio of target density at the current and proposed states

   $$a(\theta_{\text{current}} \to \theta_{\text{proposed}}) = \min\left(1, \frac{\pi(\theta_{\text{proposed}})}{\pi(\theta_{\text{current}})}\right)$$

3. Accept the proposal with probability $a(\theta_{\text{current}} \to \theta_{\text{proposed}})$ and set $\theta_{\text{new}} = \theta_{\text{proposed}}$. With probability $1 - a(\theta_{\text{current}} \to \theta_{\text{proposed}})$ reject the proposal and set $\theta_{\text{new}} = \theta_{\text{current}}$.

   - If $\pi(\theta_{\text{proposed}}) \geq \pi(\theta_{\text{current}})$ then the proposal will be accepted with probability 1.
   - Otherwise, there is a positive probability or rejecting the proposal and remaining in the current state. But this still counts as a "step" of the MC.

The Metropolis algorithm assumes the proposal distribution is symmetric. That is, the algorithm assumes that the proposal density of moving in the direction $\theta \to \tilde{\theta}$ is equal to the proposal density of moving the direction $\tilde{\theta} \to \theta$.

$$j(\theta \to \tilde{\theta}) = j(\tilde{\theta} \to \theta)$$

A generalization, the **Metropolis-Hastings algorithm**, allows for asymmetric proposal distributions, with the acceptance probabilities adjusted to accommodate the asymmetry.

$$a(\theta_{\text{current}} \to \theta_{\text{proposed}}) = \min\left(1, \frac{\pi(\theta_{\text{proposed}})j(\theta_{\text{proposed}} \to \theta_{\text{current}})}{\pi(\theta_{\text{current}})j(\theta_{\text{current}} \to \theta_{\text{proposed}})}\right)$$

The Metropolis algorithm only uses the target distribution π through ratios of the form $\frac{\pi(\theta_{\text{proposed}})}{\pi(\theta_{\text{current}})}$. Therefore, π only needs to be specified up to a constant of proportionality, since even if the normalizing constant were known it would

[3] The algorithm is named after Nicholas Metropolis, a physicist who led the research group which first proposed the method in the early 1950s, consisting of Arianna Rosenbluth, Marshall Rosenbluth, Augusta Teller, and Edward Teller. It is disputed whether Metropolis himself had anything to do with the actual invention of the algorithm.

cancel out anyway. This is especially useful in Bayesian contexts where the target posterior distribution is only specified up to a constant of proportionality via

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

We will most often use MCMC methods to simulate values from a posterior distribution $\pi(\theta|y)$ of parameters θ given data y. The Metropolis (or Metropolis-Hastings algorithm) allows us to simulate from a posterior distribution *without computing the posterior distribution.* Recall that the inputs of a Bayesian model are (1) the data y, (2) the likelihood $f(y|\theta)$, and (3) the prior distribution $\pi(\theta)$. The target posterior distribution satisfies

$$\pi(\theta|y) \propto f(y|\theta)\pi(\theta)$$

Therefore, the acceptance probability of the Metropolis algorithm can be computed based on the form of the prior and likelihood alone

$$a(\theta_{\text{current}} \to \theta_{\text{proposed}}) = \min\left(1,\ \frac{\pi(\theta_{\text{proposed}}|y)}{\pi(\theta_{\text{current}}|y)}\right) = \min\left(1,\ \frac{f(y|\theta_{\text{proposed}})\pi(\theta_{\text{proposed}})}{f(y|\theta_{\text{current}})\pi(\theta_{\text{current}})}\right)$$

To reiterate

- Proposed values are *simulated* according to the *proposal* distribution j.
- Proposals are *accepted* based on probabilities determined by the *target* distribution π.

The Metropolis-Hastings algorithm works for *any* proposal distribution which allows for eventual access to all possible values of θ. That is, if we run the algorithm long enough then the distribution of the simulated values of θ will approximate the target distribution $\pi(\theta)$. Thus we can choose a proposal distribution that is easy to simulate from. However, in practice the choice of proposal distribution is extremely important — especially when simulating from high dimensional distributions — because it determines how *fast* the MC converges to its long run distribution. There are a wide variety of MCMC methods and their extensions that strive for computational efficiency by making “smarter” proposals. These algorithms include Gibbs sampling, Hamiltonian Monte Carlo (HMC), and No U-Turn Sampling (NUTS). We won’t go into the details of the many different methods available. Regardless of the details, all MCMC methods are based on the two core principles of proposal and acceptance.

Example 17.3. We have seen how to estimate a process probability p in a Binomial situation, but what if the number of trials is also random? For example, suppose we want to estimate both the average number of three point shots Steph Curry attempts per game (μ) and the probability of success on a single attempt (p). Assume that

- Conditional on μ, the number of attempts in a game N has a Poisson(μ) distribution
- Conditional on n and p, the number of successful attempts in a game Y has a Binomial(n, p) distribution.
- Conditional on (μ, p), the values of (N, Y) are independent from game to game

For the prior distribution, assume

- μ has a Gamma(10, 2) distribution
- p has a Beta(4, 6) distribution
- μ and p are independent

In his two most recent games, Steph Curry made 4 out of 10 and 6 out of 11 attempts.

1. What is the likelihood function?
2. Without actually computing the posterior distribution, what can we say about it based on the assumptions of the model?
3. What are the possible "states" that we want our Markov chain to visit?
4. Given a current state θ_{current} how could we *propose* a new value θ_{proposed}, using a continuous analog of "random walk to neighboring states"?
5. Suppose the current state is $\theta_{\text{current}} = (8, 0.5)$ and the proposed state is $\theta_{\text{proposed}} = (7.5, 0.55)$. Compute the probability of accepting this proposal.
6. Write code to run the algorithm and plot the simulated values of θ.
7. Write and run JAGS code to approximate the posterior distribution, and compare with the previous part.

Solution. to Example 17.3

1. For an (n, y) pair for a single game, the likelihood satisfies

   $$f((n, y)|\mu, p) \propto (e^{-\mu}\mu^n)(\,p^y(1-p)^{n-y})$$

   Since the games are assumed to be independent, we evaluate the likelihood for each observed (n, y) pair and then find the product

   $$f(((10,4),(11,6))|\mu,p) \propto (e^{-\mu}\mu^{10}p^4(1-p)^{10-4})(e^{-\mu}\mu^{11}p^6(1-p)^{11-6}) = e^{-2\mu}\mu^{21}p^{10}(1-p)^{21-10}$$

   Notice that the likelihood can be evaluated based on (1) the total number of games, 2, (2) the total number of attempts, 21, and (3) the total number of successful attempts, 10.

2. Posterior is proportional to prior times likelihood. The priors for μ and p are independent.

   $$\pi(\mu,p|((10,4),(11,6))) \propto (\mu^{10-1}e^{-2\mu}p^4(1-p)^6)(e^{-2\mu}\mu^{21}p^{10}(1-p)^{21-10})$$

3. Each (μ, p) pair with $\mu > 0$ and $0 < p < 1$ is a possible state.

4. Given $\theta_{\text{current}} = (\mu_{\text{current}}, p_{\text{current}})$ we can propose a state using a *Bivariate Normal* distribution centered at the current state. The proposed values of μ and p could be chosen independently, but they could also reflect some dependence.

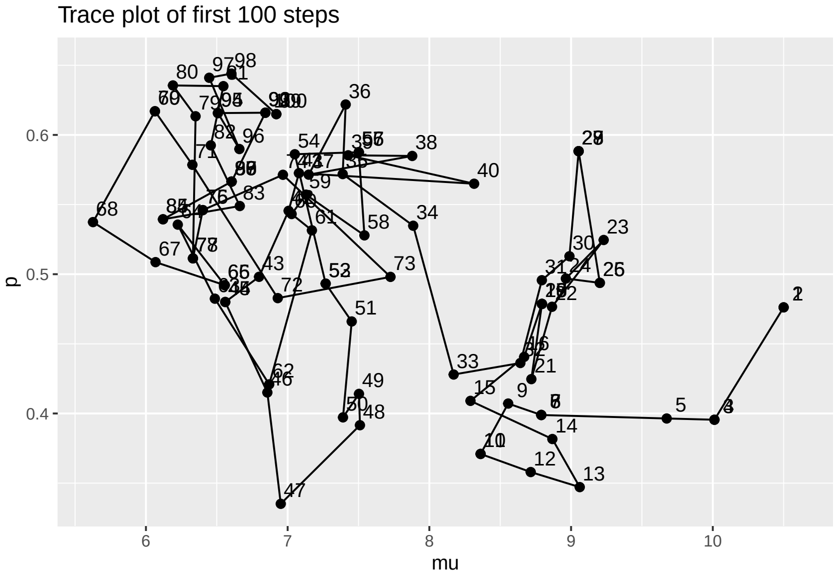
5. Suppose the current state is $\theta_{\text{current}} = (8, 0.5)$ and the proposed state is $\theta_{\text{proposed}} = (7.5, 0.55)$. Compute the probability of accepting this proposal.

```
pi_theta <- function(mu, p) {
  if (mu > 0 & p > 0 & p < 1) {
dgamma(mu, 10, 2) * dbeta(p, 4, 6) * dpois(21, 2 * mu) * dbinom(11, 21, p)
  } else {
0
  }
}

pi_theta(7.5, 0.55) / pi_theta(8, 0.5)
```

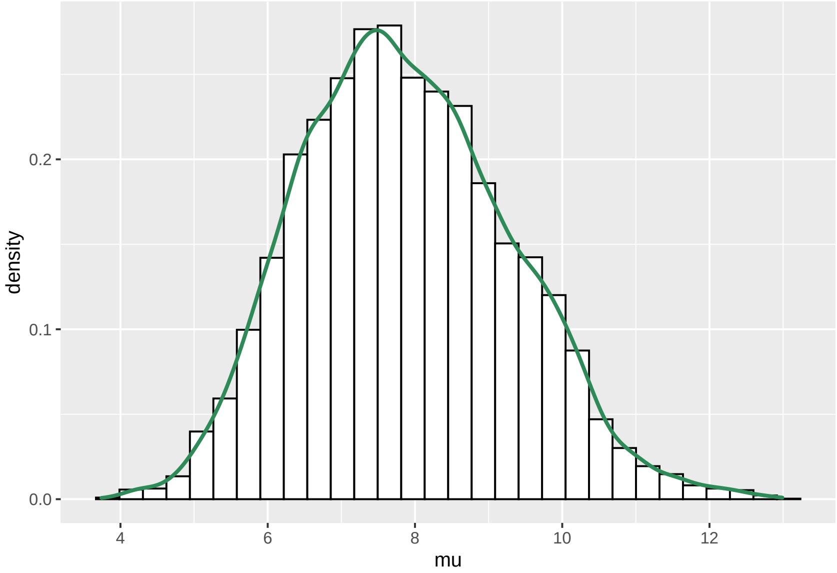
```
## [1] 0.8334
```

6. The code and output is below. Now the state is two-dimensional, (μ, p), and the trace plot shows how the Markov chain explores this two-dimensional space as it steps. You can't tell from this trace plot when the Markov chain rejects a proposal and stays in the same place since the plot points get overlaid in that case, but you can see where the step numbers coincide.

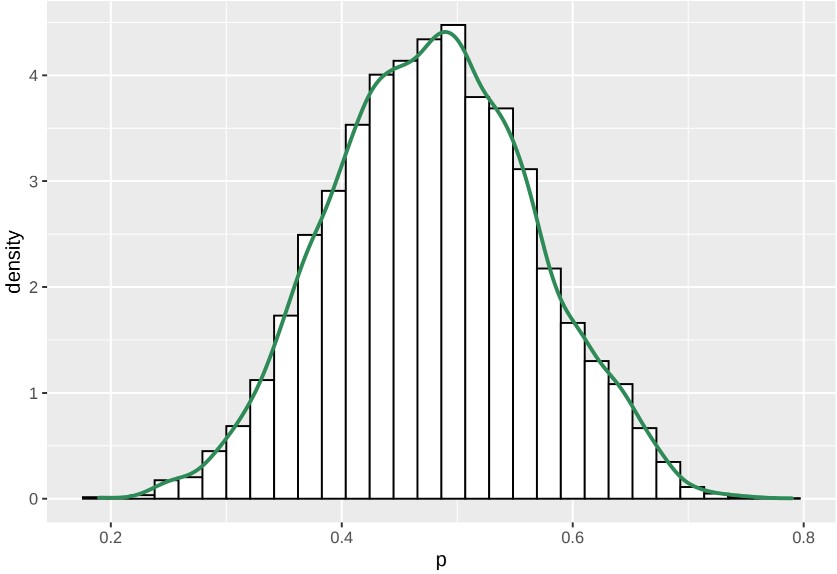
```
n_steps = 11000
delta = c(0.4, 0.05) # mu, p

theta = data.frame(mu = rep(NA, n_steps),
               p = rep(NA, n_steps))
theta[1, ] = c(10.5, 10 / 21) # initialize

for (n in 2:n_steps){
  current = theta[n - 1, ]
  proposed = current + rnorm(2, mean = 0, sd = delta)
  accept = min(1, pi_theta(proposed$mu, proposed$p) / pi_theta(current$mu, current$p))
  accept_ind = sample(0:1, 1, prob = c(1 - accept, accept))
  theta[n, ] = proposed * accept_ind + current * (1 - accept_ind)
}

# Trace plot of first 100 steps
ggplot(theta[1:100, ] %>%
    mutate(label = 1:100),
   aes(mu, p)) +
  geom_path() +
  geom_point(size = 2) +
  geom_text(aes(label = label, x = mu + 0.1, y = p + 0.01)) +
  labs(title = "Trace plot of first 100 steps")
```



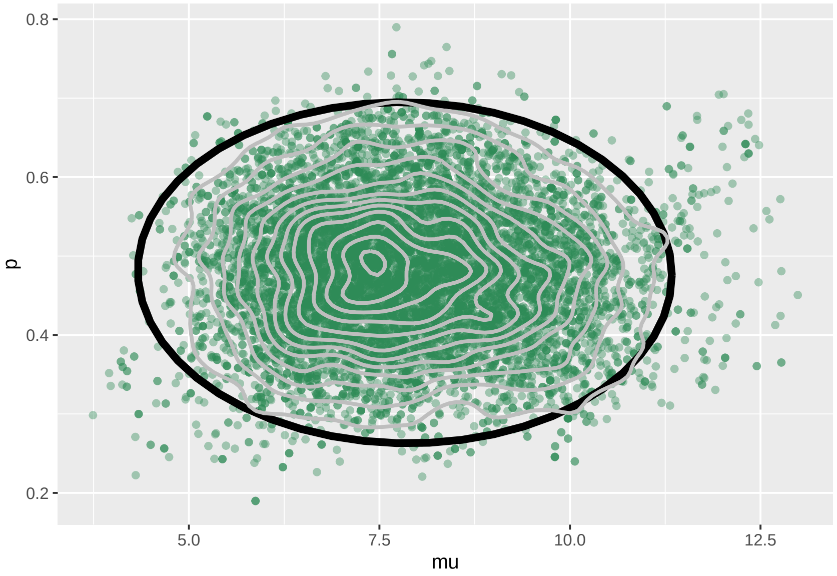
```
# Delete the first 1000 steps - we'll see why in the next chapter
theta = theta[-(1:1000), ]

ggplot(theta, aes(x = mu)) +
 geom_histogram(aes(y=..density..), color = "black", fill = "white") +
 geom_density(size = 1, color = "seagreen")
```



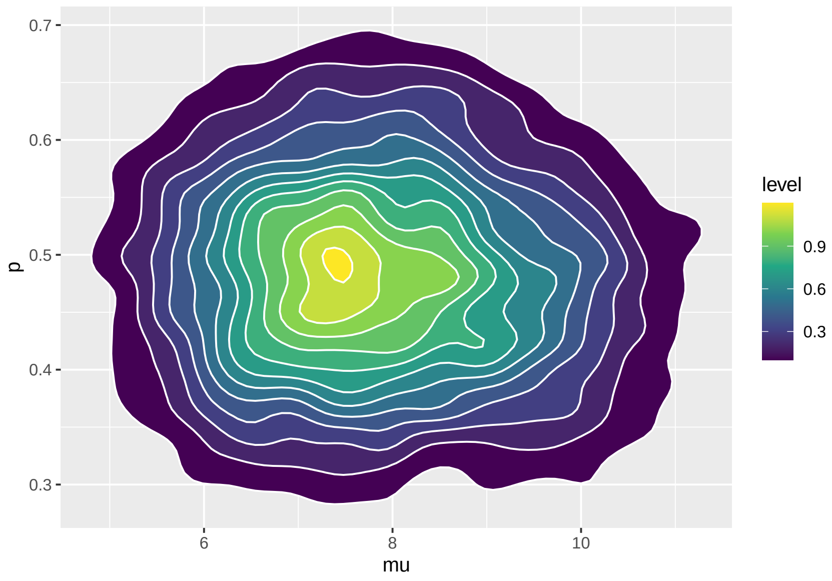
```
ggplot(theta, aes(x = p)) +
 geom_histogram(aes(y=..density..), color = "black", fill = "white") +
 geom_density(size = 1, color = "seagreen")
```



```
ggplot(theta, aes(mu, p)) +
  geom_point(color = "seagreen", alpha = 0.4) +
  stat_ellipse(level = 0.98, color = "black", size = 2) +
  stat_density_2d(color = "grey", size = 1) +
  geom_abline(intercept = 0, slope = 1)
```



```
ggplot(theta, aes(mu, p)) +
  stat_density_2d(aes(fill = ..level..),
                  geom = "polygon", color = "white") +
  scale_fill_viridis_c() +
  geom_abline(intercept = 0, slope = 1)
```



7. The JAGS code is below. The results are similar, but not quite the same as our code from scratch in the previous part. JAGS has a lot of built in features that improve efficiency. In particular, JAGS is making smarter proposals and is not rejecting as many proposals as our from-scratch algorithm. The scatterplots of simulated (μ, p) pairs kind of show this; the plot based on the from-scratch algorithm is "thinner" than the one based on JAGS because the from-scratch algorithm rejects proposals and sits in place more often.

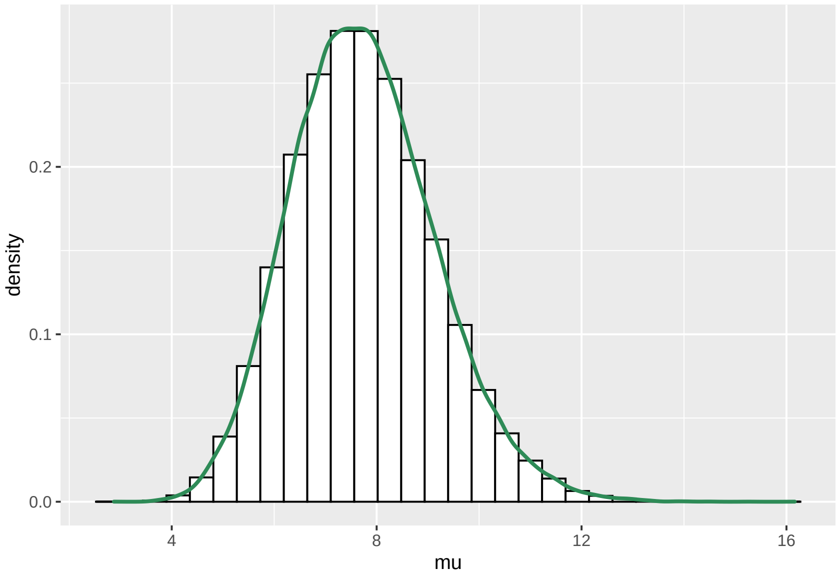
```
# data
n = c(10, 11)
y = c(4, 6)
n_sample = 2


# Model
model_string <- "model{

  # Likelihood
  for (i in 1:n_sample){

  y[i] ~ dbinom(p, n[i])

  n[i] ~ dpois(mu)
  }
  # Prior

  mu ~ dgamma(10, 2)

  p ~ dbeta(4, 6)

}"

dataList = list(y = y, n = n, n_sample = n_sample)

# Compile
Nrep = 10000

n.chains = 5

model <- jags.model(textConnection(model_string),
                data = dataList,
                n.chains = n.chains)
```

```
## Compiling model graph
##    Resolving undeclared variables
##    Allocating nodes
## Graph information:
##    Observed stochastic nodes: 4
##    Unobserved stochastic nodes: 2
##    Total graph size: 11
##
## Initializing model
```

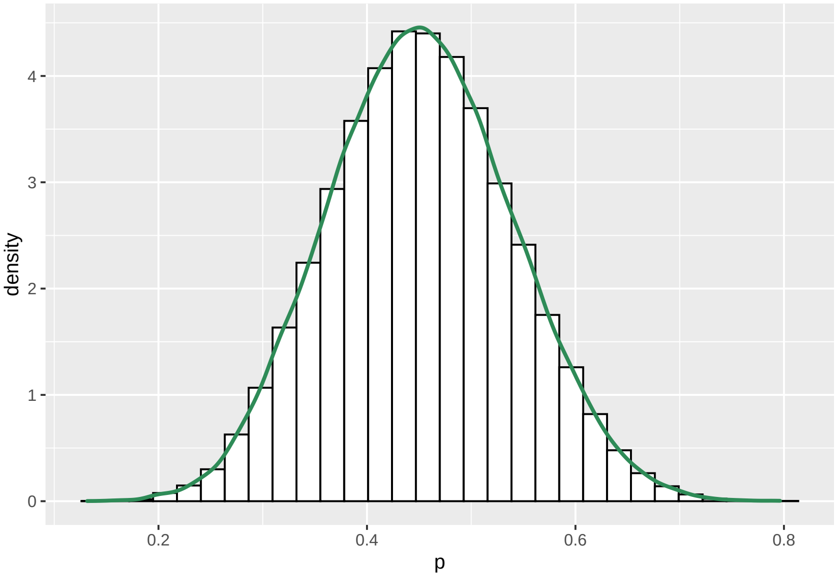
```
# Simulate
update(model, 1000, progress.bar = "none")

posterior_sample <- coda.samples(model,
                             variable.names = c("mu", "p"),
                             n.iter = Nrep,
                             progress.bar = "none")

sim_posterior = as.data.frame(as.matrix(posterior_sample))
head(sim_posterior)
```

```
##      mu      p
## 1 9.034 0.3177
## 2 8.791 0.5250
## 3 8.290 0.3150
## 4 7.839 0.3179
## 5 8.326 0.5670
## 6 8.983 0.5996
```

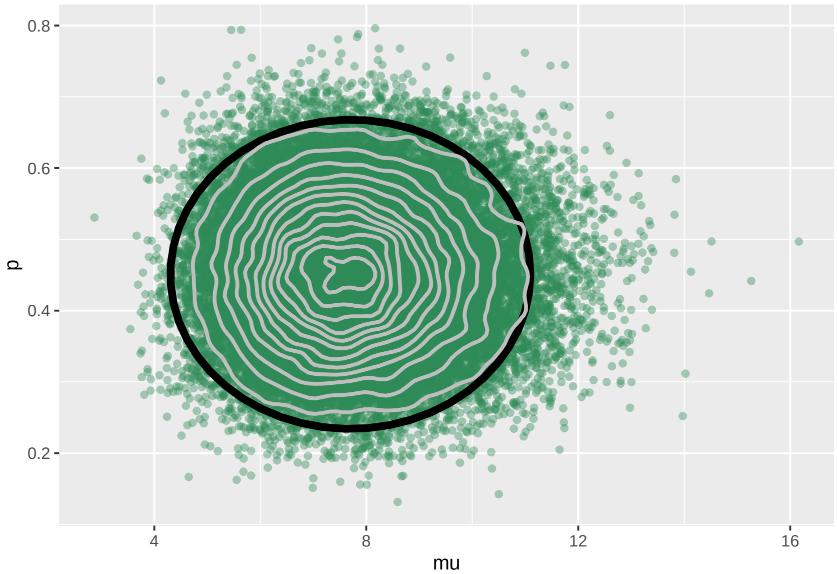
```
ggplot(sim_posterior, aes(x = mu)) +
 geom_histogram(aes(y=..density..), color = "black", fill = "white") +
 geom_density(size = 1, color = "seagreen")
```



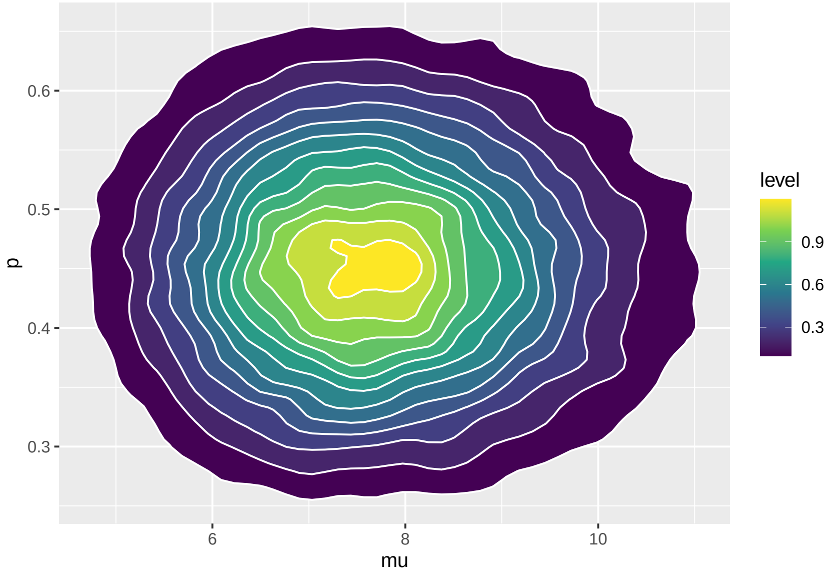
```
ggplot(sim_posterior, aes(x = p)) +
 geom_histogram(aes(y=..density..), color = "black", fill = "white") +
 geom_density(size = 1, color = "seagreen")
```



```
ggplot(sim_posterior, aes(mu, p)) +
  geom_point(color = "seagreen", alpha = 0.4) +
  stat_ellipse(level = 0.98, color = "black", size = 2) +
  stat_density_2d(color = "grey", size = 1) +
  geom_abline(intercept = 0, slope = 1)
```



```
ggplot(sim_posterior, aes(mu, p)) +
  stat_density_2d(aes(fill = ..level..),
          geom = "polygon", color = "white") +
 scale_fill_viridis_c() +
  geom_abline(intercept = 0, slope = 1)
```



Chapter 18

Some Diagnostics for MCMC Simulation

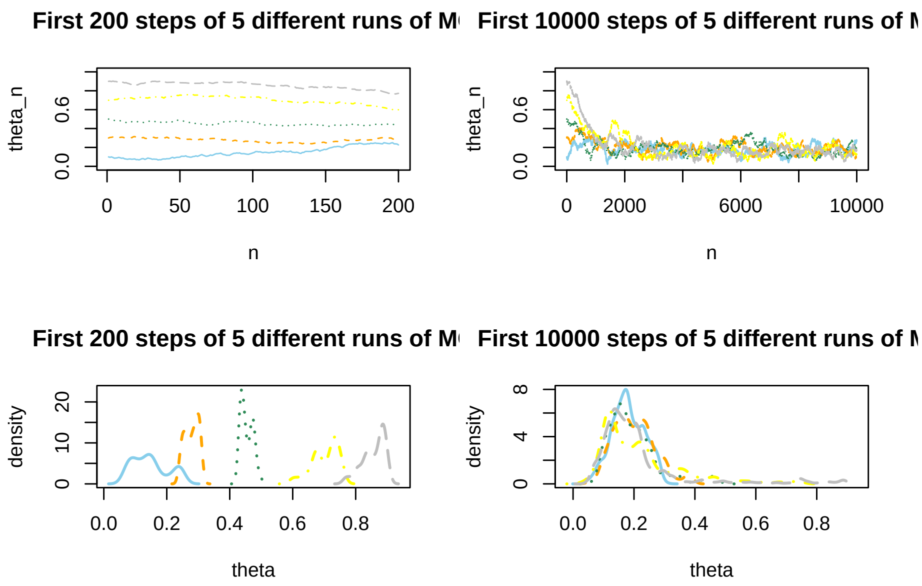
The goal we wish to achieve with MCMC is to simulate from a probability distribution of interest (e.g., the posterior distribution). The idea of MCMC is to build a Markov chain whose long run distribution is the probability distribution of interest. Then we can simulate a sample from the probability distribution, and use the simulated values to summarize and investigate characteristics of the probability distribution, by running the Markov chain for a sufficiently large number of steps. In practice, we stop the chain after some number of steps; how can we tell if the chain has sufficiently converged?

In this chapter we will introduce some issues to consider in determining if an MCMC algorithm "works".

- Does the algorithm produce samples that are *representative* of the target distribution of interest?
- Are estimates of characteristics of the distribution (e.g. posterior mean, posterior standard deviation, central 98% credible region) based on the simulated Markov chain *accurate* and stable?
- Is the algorithm *efficient*, in terms of time or computing power required to run?

Example 18.1. Recall Example 17.2 in which we used a Metropolis algorithm to simulate from a Beta(5, 24) distribution. Given current state θ_c, the proposal was generated from a $N(\theta_c, \delta)$ distribution, where δ was a specified value (determining what constitutes the "neighborhood" in the continuous random walk analog). The algorithm also needs some initial value of θ to start with; in Example 17.2 we used an initial value of 0.5. What is the impact of the initial value?

The following plots display the values of the first 200 steps and their density, and the values of 10,000 steps and their density, for 5 different runs of the Metropolis chain each starting from a different initial value: 0.1, 0.3, 0.5, 0.7, 0.9. The value of δ is 0.005. (We're setting this value to be small to illustrate a point.) What do you notice in the plots? How does the initial value influence the results?



Solution. to Example 18.1

Show/hide solution

With such a small δ value the chain tends to take a long time to move away from its current value. For example, the chain that starts at a value of 0.9 tends to stay near 0.9 for the first hundreds of steps. Values near 0.9 are rare in a Beta(5, 24) distribution, so this chain generates a lot of unrepresentative values before it warms up to the target distribution. After a thousand or so iterations all the chains start to overlap and become indistinguishable regardless of the initial condition. However, the density plots for each of the chains illustrate that the initial steps of the chain still carry some influence.

The goal of an MCMC simulation is to simulate a *representative* sample of values from the target distribution. While an MCMC algorithm should converge eventually to the target distribution, it might take some time to get there. In particular, it might take a while for the influence of the initial state to diminish. **Burn in** refers to the process of discarding the first several hundred or thousand steps of the chain to allow for a "warm up" period. Only values simulated after the burn in period are used to approximate the target distribution.

The update step in rjags runs the MCMC simulation for a burn in period, consisting of n.iter steps. (The n.iter in update is not the same as the n.iter in coda.samples.) The update function merely "warms-up" the simulation, and the values sampled during the update phase are not recorded.

The JAGS code below simulates 5 different chains, from 5 different initial conditions, each with a burn in period of 1000 steps, after which 10,000 steps of each chain are simulated. The output consists of 50,000 simulated values of θ.

```
# Data
n = 25
y = 4

# Model
model_string <- "model{

  # Likelihood
  y ~ dbinom(theta, n)

  # Prior
  theta ~ dbeta(1, 3)

}"

data_list = list(y = y, n = n)

# Compile
model <- jags.model(textConnection(model_string),
                    data = data_list,
                    n.chains = 5)
```

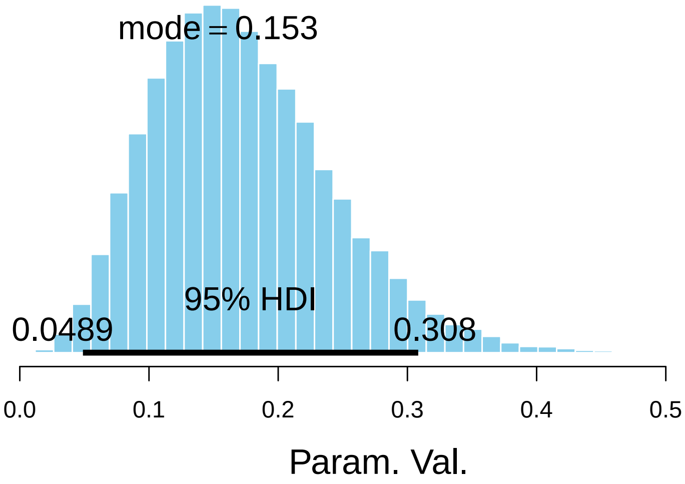
```
## Compiling model graph
##    Resolving undeclared variables
##    Allocating nodes
## Graph information:
##    Observed stochastic nodes: 1
##    Unobserved stochastic nodes: 1
##    Total graph size: 5
##
## Initializing model
```

```
# Simulate
update(model, n.iter = 1000, progress.bar = "none")

Nrep = 10000
```

```
posterior_sample <- coda.samples(model,
                                 variable.names = c("theta"),
                                 n.iter = Nrep,
                                 progress.bar = "none")
summary(posterior_sample)
```

```
##
## Iterations = 2001:12000
## Thinning interval = 1
## Number of chains = 5
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##           Mean             SD       Naive SE Time-series SE
##       0.172895       0.069324       0.000310       0.000435
##
## 2. Quantiles for each variable:
##
##   2.5%    25%    50%    75%  97.5%
## 0.0609 0.1223 0.1652 0.2153 0.3286
```

```
plotPost(posterior_sample)
```



```
##               ESS   mean median   mode hdiMass  hdiLow hdiHigh compVal
## Param. Val. 25109 0.1729 0.1652 0.1535    0.95 0.04886  0.3084      NA
##             pGtCompVal ROPElow ROPEhigh pLtROPE pInROPE pGtROPE
## Param. Val.         NA      NA       NA      NA      NA      NA
```

```
nrow(as.matrix(posterior_sample))
```

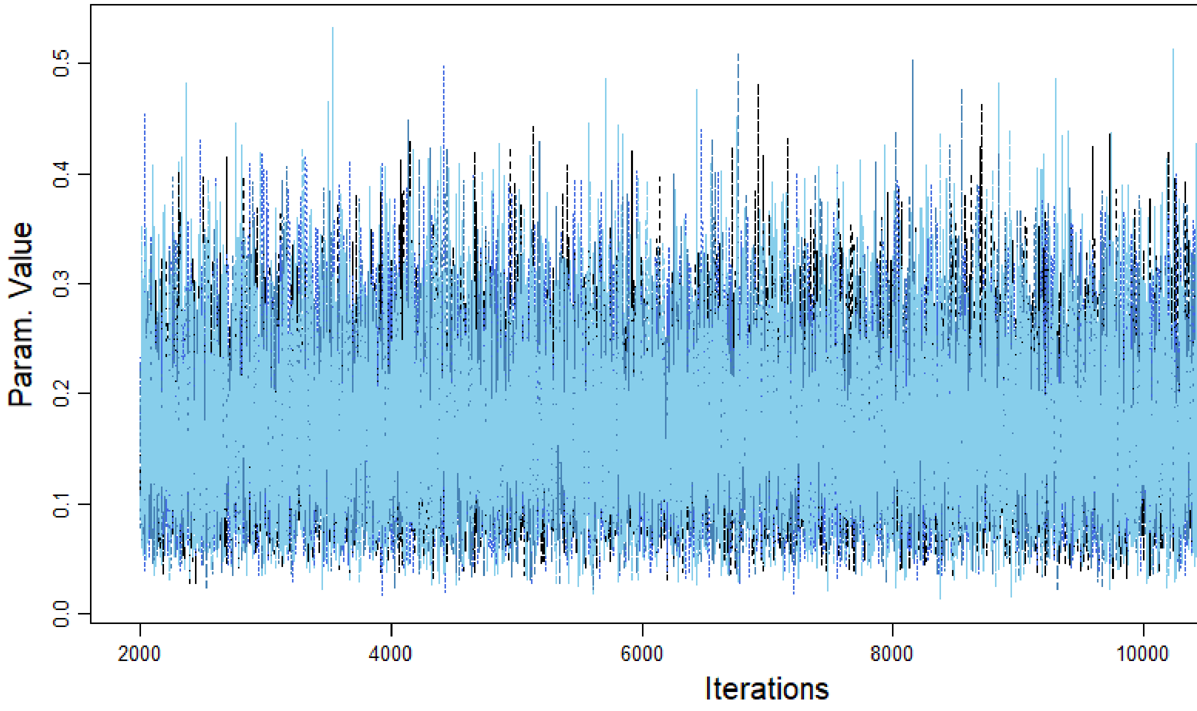
```
## [1] 50000
```

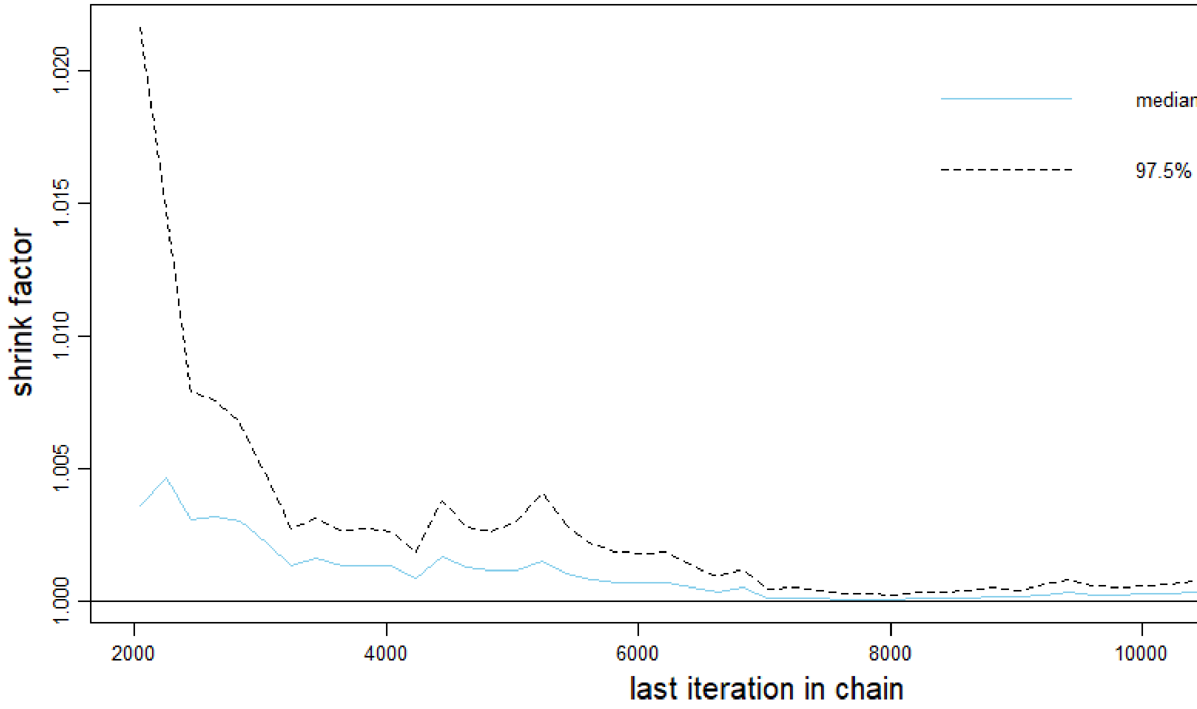
In practice, it is common to use a burn in period of several hundred or thousands of steps. To get a better idea of how long the burn in period should be, run the chain starting from several disperse initial conditions to see how long it takes for the paths to "overlap". JAGS will generate different initial values, but you can also specify them with the `inits` argument in `jags.model`. After the burn in period, examine trace plots or density plots for multiple chains; if the plots do not "overlap" then there is evidence that the chains have not converged, so they might not be producing representative samples from the target distribution, and therefore a longer burn in period is needed.

The **Gelman-Rubin statistic** (a.k.a., shrink factor) is a numerical check of convergence based on a measure of variance between chains relative to variability with chains. The idea is that if multiple chains have settled into representative sampling, then the average difference between chains should be equal to the average difference within chains (i.e., across steps). Roughly, if multiple chains are all producing representative samples from the target distribution, then given a current value, it shouldn't matter if you take the next value from the chain or if you hop to another chain. Thus, after the burn-in period, the shrink factor should be close to 1. As a rule of thumb, a shrink factor above 1.1 is evidence that the MCMC algorithm is not producing representative samples.

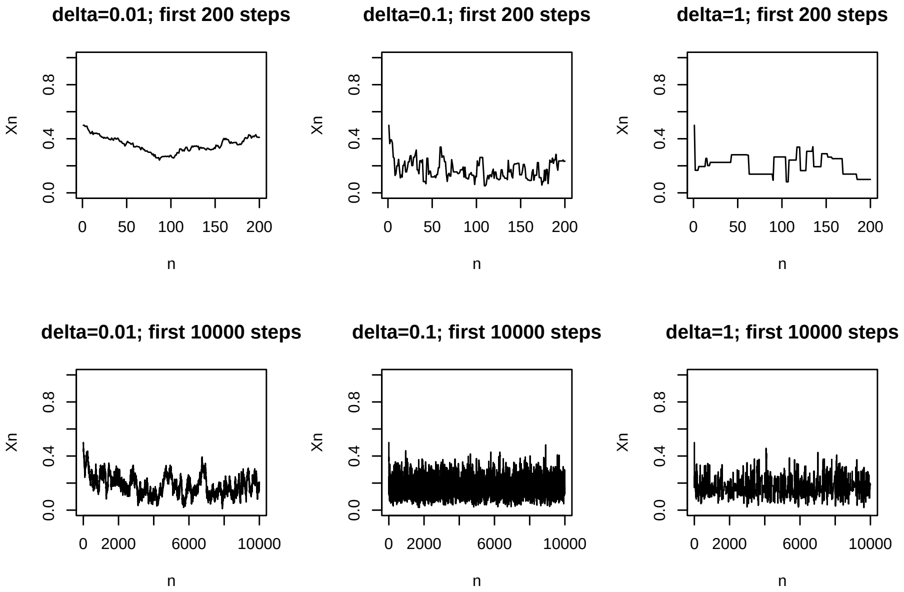
Recall that there are several packages available for summarizing MCMC output, and these packages contain various diagnostics. For example, the output of the `diagMCMC` function in the `DBDA2E-utilities.R` file includes a plot the shrink factor.

```
diagMCMC(posterior_sample)
```


Param. Value
0.0
0.1
0.2
0.3
0.4
0.5
2000
4000
6000
8000
10000
Iterations


shrink factor
1.000
1.005
1.010
1.015
1.020
median
97.5%
2000
4000
6000
8000
10000
last iteration in chain

Example 18.2. Continuing with Metropolis sampling from a Beta(5, 24) distribution, the following plots display the results of three different runs, one each for $\delta = 0.01$, $\delta = 0.1$, $\delta = 1$, all with an initial value of 0.5. Describe the differences in the behavior of the chains. Which chain seems "best"? Why?



Solution. to Example 18.2

Show/hide solution

When $\delta = 0.01$ only values that are close to the current value are proposed. A proposed value close to the current value will have a density that is close to, if not greater, than that of the current value. Therefore, most of the proposals will be accepted, but these proposals don't really go anywhere. With $\delta = 0.01$ the chain moves often, but it does not move far.

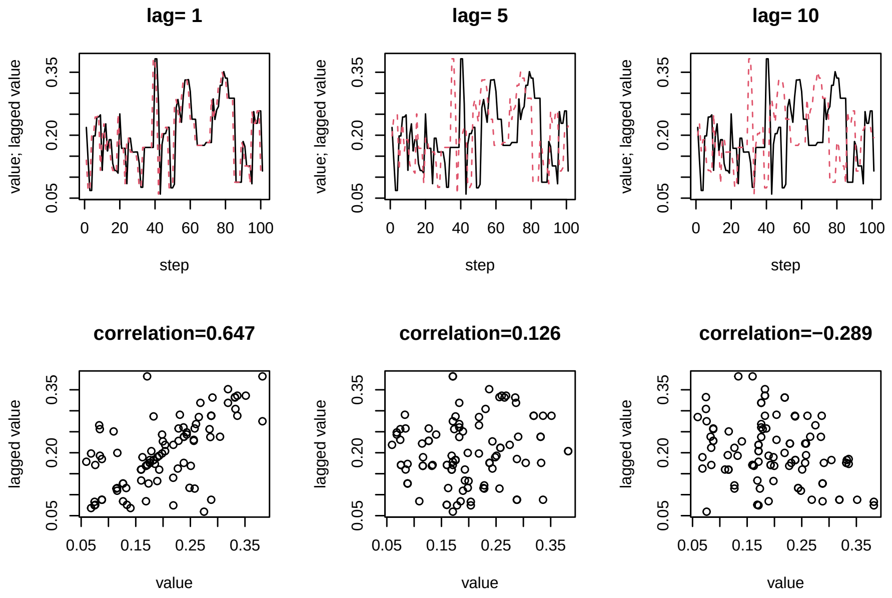
When $\delta = 1$ a wide range of values will be proposed, including values outside of (0, 1). Many proposed values will have density that is much less than that of the current value, if not 0. Therefore many proposals will be rejected. With $\delta = 1$ the chain tends to get stuck in a value for a large number of steps before moving (though when it does move, it can move far.)

Both of the above cases tend to get stuck in place and require a large number of steps to explore the target distribution. The case $\delta = 0.1$ is a more efficient. The proposals are neither so narrow that it takes a long time to move nor so wide that many proposals are rejected. The fast up and down pattern of the trace plot shows that the chain with $\delta = 0.2$ explores the target distribution much more efficiently than the other two cases.

The values of a Markov chain at different steps are *dependent*. If the degree of dependence is too high, the chain will tend to get "stuck", requiring a large number of steps to fully explore the target distribution of interest. Not only will the algorithm be inefficient, but it can also produce inaccurate and unstable estimates of chararcteristics of the target distribution.

If the MCMC algorithm is working, trace plots should look like a "fat, hairy catepillar."[1] Plots of the autocorrelation function (ACF) can also help determine how "clumpy" the chain is. An autocorrelation measures the correlation between values at different lags. For example, the lag 1 autocorrelation measures the correlation between the values and the values from the next step; the lag 2 autocorrelation measures the correlation between the values and the values from 2 steps later.

Example 18.3. Continuing with Metropolis sampling from a Beta(5, 24) distribution, the following plots display, for the case $\delta = 0.1$, the actual values of the chain (after burn in) and the values lagged by 1, 5, and 10 time steps. Are the values at different steps dependent? In what way are they not too dependent?



Solution. to Example 18.3

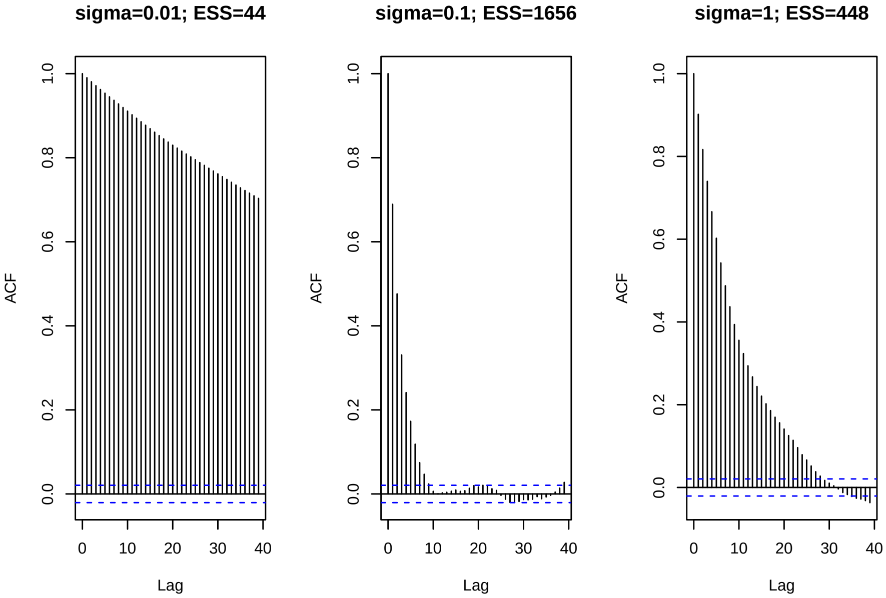
Show/hide solution

[1] I've seen this description in many references, but I don't know who first used this terminology.

Yes, the values are dependent. In particular, the lag 1 autocorrelation is about 0.8, and the lag 5 autocorrelation is about 0.4. However, the autocorrelation decays rather quickly as a function of lag. The lag 10 autocorrelation is already close to 0. In this way, the chain is "not too dependent"; each value is only correlated with the values in the next few steps.

An autocorrelation plot displays the autocorrelation within in a chain as a function of lag. If the ACF takes too long to decay to 0, the chain exhibits a high degree of dependence and will tend to get stuck in place.

The plot below displays the ACFs corresponding to each of the δ values in Example 18.2. Notice that with $\delta = 0.2$ the ACF decays fairly quickly to 0, while in the other cases there is still fairly high autocorrelation even after long lags.



Example 18.4. Continuing with Metropolis sampling from a Beta(5, 24) posterior distribution. We know that the posterior mean is $5/29 = 0.172$. But what if we want to approximate this via simulation?

1. Suppose you simulated 10000 *independent* values from a Beta(5, 24) distribution, e.g. using `rbeta`. How would you use the simulated values to estimate the posterior mean?
2. What is the *standard error* of your estimate from the previous part? What does the standard error measure? How could you use simulation to approximate the standard error?
3. Now suppose you simulated 10000 from a Metropolis chain (after burn in). How would you use the simulated values to estimate the posterior mean?

What does the standard error measure in this case? Could you use the formula from the previous part to compute the standard error? Why?

4. Consider the three chains in Example 18.2 corresponding to the three δ values 0.01, 0.1, and 1. Which chain provides the most reliable estimate of the posterior mean? Which chain yields the smallest standard error of this estimate?

Solution. to Example 18.4

Show/hide solution

1. Simulate 10000 values and compute the sample mean of the simulated values.
2. For the Beta(5, 24) distribution, the population SD is $\sqrt{(5/29)(1-5/29)/(29+1)} = 0.07$. The standard error of the sample mean of 10000 values is $0.07/\sqrt{10000} = 0.0007$. The standard error measures the sample-to-sample variability of sample means over many samples of size 10000. To approximate the standard error via simulation: sample 10000 values from a Beta(5, 24) distribution and compute the sample mean, then repeat many times and find the standard deviation of the simulated sample means.
3. You would still use the sample mean of the 10000 values to approximate the posterior mean. The standard error measures how much the sample mean varies from run-to-run of the Markov chain. To approximate the standard error via simulation: simulate 10000 steps of the Metropolis chain and compute the sample mean, then repeat many times and find the standard deviation of the simulated sample means. The standard error formula from the previous part assumes that the 10000 values are *independent*, but the values on the Markov chain are not, so we can't use the same formula.
4. Among these three, the chain with $\delta = 0.1$ provides the most reliable estimate of the posterior mean since it does the best job of sampling from the posterior distribution. While there is dependence in all three chains, the chain with $\delta = 0.1$ has the least dependence and so comes closest to independent sampling, so it would have the smallest standard error.

A Markov chain that exhibits a high degree of dependence will tend to get stuck in place. Even if you simulate 10000 steps of the chain, you don't really get 10000 "new" values. The **effective sample size (ESS)** is a measure of how much independent information there is in an autocorrelated chain. Roughly, the effective sample size answers the question: what is the equivalent sample size of a completely independent chain?

The effective sample size[2] of a chain with N steps (after burn in) is

$$\text{ESS} = \frac{N}{1 + 2\sum_{\ell=1}^{\infty} \text{ACF}(\ell)}$$

where the infinite sum is typically cut off at some upper lag (say $\ell = 20$). For a completely independent chain, the autocorrelation would be 0 for all lags and the ESS would just be the number of steps N. The more quickly the ACF decays to 0, the larger the ESS. The more slowly the ACF decays to 0, the smaller the ESS.

The larger the ESS of a Markov chain, the more accurate and stable are MCMC-based estimates of characteristics of the posterior distribution (e.g., posterior mean, posterior standard deviation, 98% credible region). That is, if the ESS is large and we run the chain multiple times, then estimates do not vary much from run to run.

The *standard error* of a statistic is a measure of its accuracy. The standard error of a statistic measures the sample-to-sample variability of values of the statistic over many samples of the same size. A standard error can be approximated via simulation.

- Simulate a sample and compute the value of the statistic.
- Repeat many times and find the standard deviation of simulated values of the statistic.

For many statistics (means, proportions) the standard error based on a sample of n *independent* values is on the order of $\frac{1}{\sqrt{n}}$.

For example, the standard error of a sample mean measures the sample-to-sample variability of sample means over many samples of the same size. The standard error of a sample mean based on an *independent* sample of size n is

$$\frac{\text{population SD}}{\sqrt{n}}$$

where the population SD measures the variability of individual values of the variable.

The usual $\frac{1}{\sqrt{n}}$ formulas for standard errors are based on samples of n *independent* values. However, in a Markov chain the values will be *dependent*. The **Monte Carlo standard error (MCSE)** is the standard error of a statistic generated based on an MCMC algorithm. The MCSE of a statistic measures the run-to-run variability of values of the statistic over many runs of the chain of the same number of steps. A MCSE can be approximated via simulation

[2] The `coda` library in R contains a lot of diagnostic tests for MCMC methods, including the function `effectiveSize`.

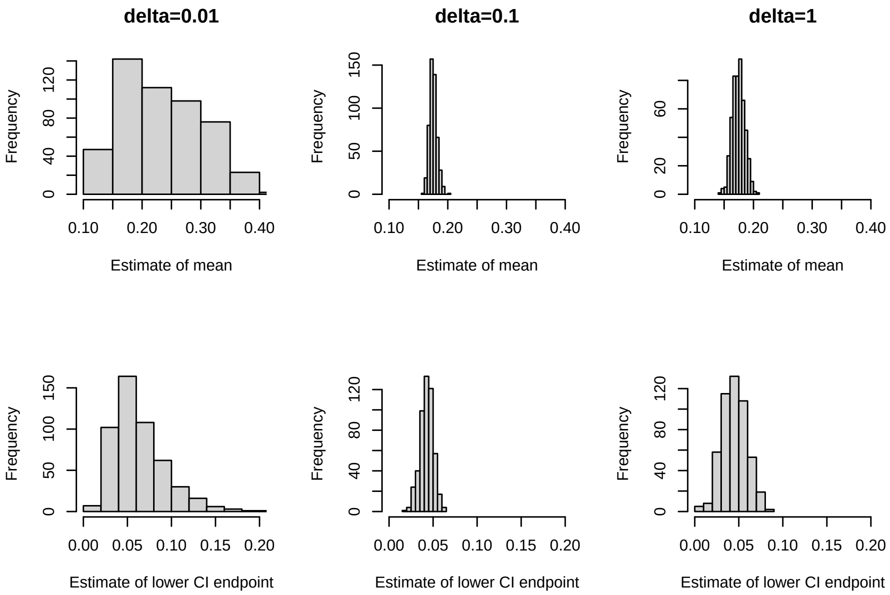
- Simulate many steps of the Markov chain and compute the value of the statistic for the simulated chain
- Repeat many times and find the standard deviation of simulated values of the statistic

For many statistics (means, proportions) the MCSE based on a chain with effective sample size ESS is on the order of $\frac{1}{\sqrt{\text{ESS}}}$.

The MCSE effects the accuracy of parameter estimates based on the MCMC method. If the chain is too dependent, the ESS will be small, the MCSE will be large, and resulting estimates will not be accurate. That is, two different runs of the chain could produce very different estimates of a particular characteristic of the target distribution.

How large of an ESS is appropriate depends on the particular characteristic of the posterior distribution being estimated. A larger ESS will be required for accurate estimates of characteristics that depend heavily on sparse regions of the posterior that are visited relatively rarely by the chain, like the endpoints of a 98% credible interval.

The plot belows correspond to each of the δ values in Example 18.2. Each plot represents 500 runs of the chain, each run with 1000 steps (after burn in). For each run we computed both the sample mean (our estimate of the posterior mean) and the 0.5th percentile (our estimate of the lower endpoint of a central 99% credible interval.) Therefore, each plot in the top row displays 500 simulated sample means, and each plot in the bottom row displays 500 simulated 0.5th percentiles. The MCSE is represented by the degree of variability in each plot. We see that for both statistics the MCSE is smallest when $\delta = 0.1$, corresponding to the smallest degree of autocorrelation and the largest ESS.



For most of the situations we'll see in this course, standard MCMC algorithms will run fairly efficiently, and checking diagnostics is simply a matter of due diligence. However, especially in more complex models, diagnostic checking is an important step in Bayesian data analysis. Poor diagnostics can indicate the need for better MCMC algorithms to obtain a more accurate picture of the posterior distribtuion. Algorithms that use "smarter" proposals will usually lead to better results.

Bibliography

Dogucu, M., Johnson, A., and Ott, M. (2022). *Bayes Rules! An Introduction to Applied Bayesian Modeling.* Chapman and Hall/CRC, Boca Raton, Florida, 1st edition. ISBN 978-0367255398.

Kruschke, J. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan.* Academic Press, 2nd edition.

McElreath, R. (2020). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan, 2nd Edition.* CRC Press, 2 edition.