An Introduction to Bayesian Reasoning and Methods

Kevin Ross

2021-01-12

Contents

Preface

This textbook presents an introduction to Bayesian reasoning and methods.

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

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 basic 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 (which this course is about). This chapter introduces the two interpretations.

1.1 Instances of randomness

Probability comes up in a wide variety of situations. Consider just a few examples.

1. The probability that a single flip of a fair coin lands on heads.
2. The probability you win the next Powerball lottery if you purchase a single ticket, 6-7-16-23-26, plus the Powerball number, 4.
3. The probability that a "randomly selected" Cal Poly student is a California resident.
4. The probability that it rains tomorrow in San Luis Obispo.
5. The probability that there are more Atlantic hurricanes in 2021 than in 2020.
6. The probability that the Green Bay Packers win the next Superbowl.
7. The probability that Democrats win both Senate seats in the Georgia runoff 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 1.1. 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 1.1

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! 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 flipping a coin 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 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 Packers are more likely than the Las Vegas (?!?!) Raiders to be the winner.

While all of the situations involve uncertainty, it seems that there are different "types" of uncertainty. Even though we don't know if a coin flip will land on heads or tails, the notion of "fairness" implies that these two outcomes 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 with certainty which teams will win individual games.

It also seems that there are different interpretations of probability. Given that a coin is fair, we might all agree that the probability that it lands on heads is 1/2. 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 the probability that it rains tomorrow in San Luis Obispo. Different weather forecasters, news stations, or websites might provide different values for this probability. There seems to be some subjectivity to probability in situations like tomorrow's weather or the next Superbowl.

Finally, some of these phenomenon are repeatedable. We could (in principle) flip a coin many times and how often the flips land on heads, 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 it 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** 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[1], like flipping a coin or drawing powerballs at random from a bin. 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, there is a *regular distribution of outcomes over a large number of (hypothetical) repetitions.*

- 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 2020-2021 seasons; how often would we expect the Eagles to win? The Raiders? (Hopefully a lot for the Eagles; probably not much for the Raiders).

[1] 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?

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

- 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.

- *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.

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%[2]. In particular, a probability requires much more interpretation than "is the probability greater than, less than, or equal to 50%?" As Example 1.1 suggests, there can be different interpretations of "probability", which we'll start to explore in the next section.

1.2 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, namely, *long run relative frequency* and *subjective probability*.

1.2.1 Long run relative frequency

We can 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%? One

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

Table 1.1: Results and running proportion of H for 10 flips of a fair coin.

Flip	Result	Running proportion of H
1	T	0.000
2	T	0.000
3	T	0.000
4	H	0.250
5	H	0.400
6	H	0.500
7	T	0.429
8	T	0.375
9	H	0.444
10	H	0.500

interpretation involves considering *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 1.1 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, T) so the running proportion of H after 2 flips is 0/2; and so on. Figure 1.1 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 1.2 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, 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 1.3 summarizes the results for our original set, while the plot on the right also displays the results for the three additional sets from Figure 1.3. 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. There is less variability in the proportion of H after 1000 flips than after 100. Now, even after 1000 flips the proportion of flips that result in H isn't guaranteed to be exactly 0.5, but we

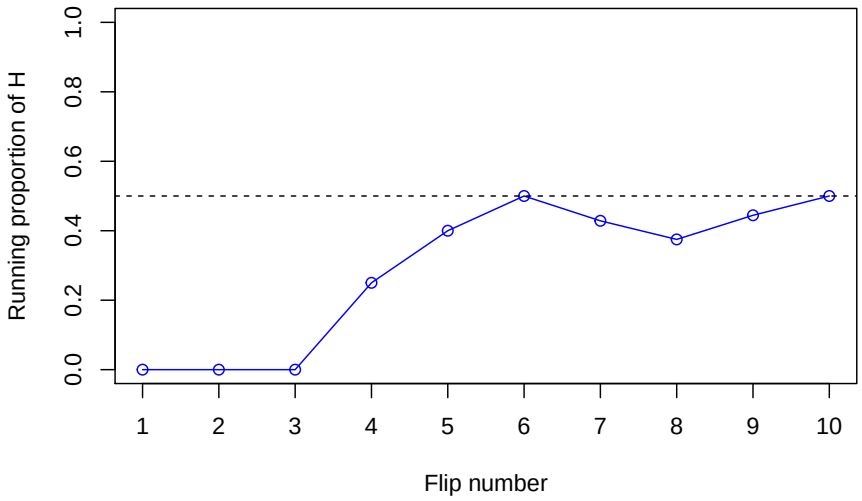


Figure 1.1: Running proportion of H versus number of flips for the 10 coin flips in Table 1.1.

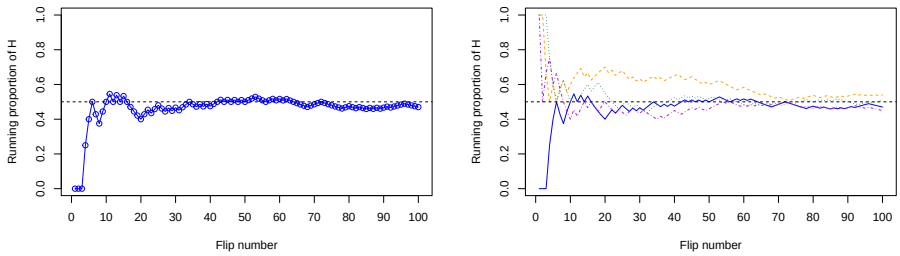


Figure 1.2: Running proportion of H versus number of flips for four sets of 100 coin flips.

see a tendency for the proportion to get closer to 0.5 as the number of flips increases.

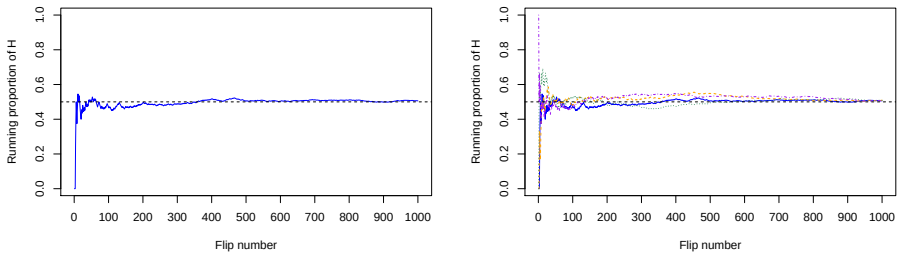


Figure 1.3: Running proportion of H versus number of flips for four sets of 1000 coin flips.

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 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 the outcome can be observed each time. 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 that event.

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.

Example 1.2. In each of the following, write a clearly worded sentence interpreting the numerical value of the probability as a long run relative frequency in context. (Just take the numerical values— 0.1, 0.25, and 0.73 — as given. We'll see how to compute probabilities later.)

1. The probability that a roll of a fair ten-sided die lands on 1 is 0.1.
2. The probability that two flips of a fair coin both land on H is 0.25.
3. The probability that in 100 flips of a fair coin the proportion of flips that land on H is between 0.45 and 0.55 is 0.73.

Solution. to Example 1.2

Show/hide solution

1. About 10% of rolls of a fair ten-sided result in a roll of 1. The phenomenon is a roll of a far ten-sided die and the event of interest is whether or not the die lans on 1.
2. In about 25% of sets of two fair coin flips, both flips in the set land on H. The phenomenon involves *two* flips of a coin, so we consider what would happen over many *sets* of two flips each.
3. In about 73% of sets of 100 fair coin flips, the proportion of H for the set is between 0.45 and 0.55. The phenomenon involves 100 coin flips, so we consider many sets of 100 coin flips each, each set resulting in a proportion of H that is either between 0.45 and 0.55 or not. Imagine adding many more paths to the plot on the right in Figure 1.2, each corresponding to a set of 100 flips, and seeing how many of the paths result in a value between 0.45 and 0.55 at flip 100.

1.2.2 Subjective probability

The long run relative frequency interpretation is natural in repeatable situations like flipping a coin, drawing Powerballs from a bin, or selecting a Cal Poly student at random.

On the other hand, it is difficult to conceptualize some scenarios in the long run. Superbowl 2021 will only be played once, the 2021 Georgia run off election will only be conducted once, 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, the Kansas City Chiefs are more likely than the Jacksonville Jaguars to win Superbowl 2021, each of the candidates in the Georgia Senate races are about equally likely to win, and if you've always been an apple-a-day person, there's a good chance you ate one on April 17, 2009. So it still makes sense to talk about probability in uncertain, but not necessarily repeated, situations.

However, the *meaning* of probability does seem different in a physically repeatable situations like coin flips than in single occurrences like the 2021 Superbowl. Let's switch sports and consider the 2020 World Series of Major League Baseball[3]. As of Sept 3,

[3]which already happened, but I'm too lazy to update the examples. Imagine that when you're reading this it's Sept 3, 2020.

- According to FiveThirtyEight, the Los Angeles Dodgers have a 31% chance of winning the 2020 World Series, the highest of any team, while the New York Yankees have a 10% chance.
- According to FanGraphs, the Dodgers have a 18% chance of winning the 2020 World Series, while the Yankees have an 6% chance.
- According to gambling site Odds Shark, the Dodgers have a 22% chance of winning the 2020 World Series, while the Yankees have an 18% chance.

Each website, as well as many others, assigns different probabilities to the Dodgers or Yankees winning. Which website, 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 2020 World Series will only happen once, we could still perform a simulation involving hypothetical repetitions. However, simulating the World Series involves first simulating the 2020 season to determine the playoff matchups, then simulating the playoffs to see which teams make the World Series, then simulating the World Series matchup itself. And simulating the 2020 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 Sept 3, according to FiveThirtyEight the Dodgers had a 68% chance of beating the Colorado Rockies in their game on Sept 4, but according to FanGraphs it was 67% and according to Odds Shark it was 72%. (The Dodgers won.) Even though these differences might seem small, many small differences over the course of the season could result in large differences in predictions for the World Series 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 baseball games or the World Series champion. Therefore, there is no single 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 or uncertainty* rather than long run relative frequencies. For example, in the FiveThirtyEight forecast, the Dodgers are *3.1 times more likely* to win the World Series than the Yankees (3.1 = 0.31 / 0.10). Relative likelihoods can also be compared across different forecasts or scenarios. For example, FiveThirtyEight believes that the Dodgers are about 1.4 times more likely to win the World Series than Odds Shark does. Also, FiveThirtyEight believes that the

likelihood that a fair coin lands on H is about 5 times larger than the likelihood that the Yankees win the 2020 World Series.

The FiveThirtyEight MLB 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 Los Angeles Dodgers will win the 2020 World Series"), 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 Sept 3), then in about 31% of repetitions the Dodgers would win the World Series, and in about 10% of repetitions the Yankees 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 1.3. What is your subjective probability that Professor Ross has a TikTok account? Consider the following two bets, and suppse you can choose only one.

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 1.3

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%.
   - 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. But the example gives a very rough idea of how you might discern a subjective probability of an event.

Disclaimer: we do not advocate gambling. We merely use gambling contexts to motivate probability concepts.

1.3 Working with probabilities

In the previous section we saw two different interpretations of probability: long run relative frequency and subjective. Fortunately, the mathematics of probability work the same way regardless of the interpretation. Furthermore, even with subjective probabilities it is helpful to consider what might happen in a simulation.

1.3.1 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. Put loosely, probabilities cannot be negative and the sum of probabilities over all possible outcomes must be 100%.

Example 1.4. As of Sept 3, FiveThirtyEight listed the following probabilities for who will win the 2020 World Series.

Team	Probability
Los Angeles Dodgers	31%
New York Yankees	10%
Houston Astros	9%
Tampa Bay Rays	9%
Other	

According to FiveThirtyEight (as of Sept 3):

1. What would you expect the results of 10000 repetitions of a simulation of the World Series 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 Dodgers do *not* win the 2020 World Series?
3. What must be the probability that one of the above four teams is the World Series champion?
4. What must be the probability that a team other than the above four teams is the World Series champion? That is, what value goes in the "Other" row in the table?

Solution. to Example 1.4

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 31%, or 3100, of the 10000 repetitions.

Team	Winner
Los Angeles Dodgers	3100
New York Yankees	1000
Houston Astros	900
Tampa Bay Rays	900
Other	4100
Total	10000

   Of course, there would be some variability from simulation to simulation, just like in the sets of 1000 coin flips in Figure 1.3. But the above counts represent about what we would expect.

2. 69%. Either the Dodgers win or they don't; if there's a 31% chance that the Dodgers win, there must be a 69% chance that they do not win. If we think of this as a simulation with 10000 repetitions, each repetition results in either the Dodgers winning or not, so if they win in 3100 of repetitions then they must not win in the other 6900.

3. 59%. There is only one World Series champion, so if say the Dodgers win then no other team can win. Thinking again of the simulation, the repetitions in which the Dodgers win are distinct from those in which the Yankees win. So if the Dodgers win in 3100 repetitions and the Yankees win in 1000 repetitions, then on a total of 4100 repetitions either the Dodgers or Yankees win. Adding the four probabilities, we see that the probability that one of the four teams above wins must be 59%.

4. 41%. Either one of the four teams above wins, or some other team wins. If one of the four teams above wins in 5900 repetitions, then in 4100 repetitions the winner is not one of these four teams.

Example 1.5. Suppose your subjective probabilities for the 2020 World Series champion satisfy the following conditions.

- The Astros and Rays are equally likely to win
- The Yankees are 1.5 times more likely than the Astros to win
- The Dodgers are 2 times more likely than the Yankees to win

- The winner is as likely to be among these four teams — Dodgers, Yankees, Astros, Rays — as not

Construct a table like in Example 1.4 of your subjective probabilities.

Solution. to Example 1.5

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 Astros.

- Suppose the Astros 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 Astros 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 Astros.
- The Astro and Rays are equally like to win, so the Rays also account for 1 unit.
- The Yankees are 1.5 times more likely than the Astros to win, so the Yankees account for 1.5 units. If 1 unit is 1000 repetitions, then the Yankees win in 1500 repetitions, 1.5 times more often than the Astros.
- The Dodgers are 2 times more likely than the Yankees to win, so the Dodgers account for $2 \times 1.5 = 3$ units. If 1 unit is 1000 repetitions, then the Dodgers 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 Astros account for 1 unit, so their probability of winning is 1/13 or about 7.7%. Likewise, the probability that the Dodgers win is 3/13 or about 23.1%.

Team	Units	Repetitions	Probability
Los Angeles Dodgers	3.0	3000	23.1%
New York Yankees	1.5	1500	11.5%
Houston Astros	1.0	1000	7.7%
Tampa Bay Rays	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 Dodgers are 2 times more likely ($2 = 23.1/11.5$) than the Yankees to win, and the Yankees are 1.5 times more likely ($1.5 \approx 11.5/7.7$) than the Astros to win.

We could have also solved this problem using algebra. Let x be the probability, as a decimal, that the Astros are the winner. Then x is also the probability that the Rays are the winner, $1.5x$ for the Yankees, and $3x$ for the Dodgers. 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 1.5 illustrates one way of formulating subjective probabilities. We start by specifying probabilities in relative terms, and then "normalize" these probabilities so that they add up to 100%. As in the problem, it helps to consider one outcome as a "baseline" and to specify all likelihoods relative to the baseline.

1.3.2 Odds

The words "probability", "chance", "likelihood", and "odds" are colloquially treated as synonyms. However, in the mathematical language of probability, *odds* provide a different way of reporting a probability. Rather than reporting probability on a 0% to 100% scale, odds report probabilities in terms of ratios.

Example 1.6. In Example 1.4 the odds that the Yankees win the World Series are 9 to 1 against.

1. What do you think that "9 to 1 against" means?
2. What are the odds of the Yankees *not* winning?
3. What are the odds of the Dodgers winning?
4. What are the odds of one of the Other teams winning?
5. The Philadelphia Phillies have 50 to 1 odds against winning. What is the probability that the Phillies win the World Series?

Solution. to Example 1.6

Show/hide solution

1. The probability that the Yankees win is 0.1, so the probability that they do not win is 0.9. These numbers are in a 9 to 1 ratio: the probability of not winning (0.9) is 9 times greater than the probability of winning (0.1). So the odds *against* the Yankees winning the World Series are 9 to 1; "against" because the Yankees are less likely to win than to not win.

2. The probabilities are still in the 9 to 1 ratio, but we can say that the odds are 9 to 1 *in favor* of the Yankees *not* winning. We could also say the odds are 1 to 9 in favor of the Yankees winning, but odds are typically reported with the larger value first — 9 to 1 instead of 1 to 9.

3. The probability that the Dodgers win is 0.31 and that they don't win is 0.69, and $0.69/0.31 \approx 2.2$. So the odds are 2.2 to 1 *against* the Yankees winning; "against" because the Yankees are less likely to win than to not win. Odds are often reported as whole numbers, so we could say the odds are 11 to 5 against the Dodgers winning.

4. The probability that an Other team wins is 0.41 and that an Other team doesn't win is 0.59, and $0.59/0.41 \approx 1.44$. So the odds are 1.44 to 1 against an Other team winning.

5. The probability that the Phillies do not win is 50 times greater than the probability that they do win. Let the event that the Phillies win account for 1 "unit" so that the event that they do not win accounts for 50 units, for a total of 51 units. So the probability that the Phillies win is $1/51 \approx 0.02$. Note that the probability of not winning, 50/51, is 50 times greater than the probability of winning.

   You could also solve this with algebra. Let x be the probability that the Phillies win, so $50x$ is the probability that they don't win. The probabilities must sum to 1, so set $x + 50x = 1$ and solve for x.

The **odds** of an event is a ratio involving the probability that the event occurs and the probability that the event does not occur. Odds can be expressed as either "in favor" of or "against" the event occurring.

$$\text{odds in favor} = \frac{\text{probability that the event occurs}}{\text{probability that the event does not occur}}$$

$$\text{odds against} = \frac{\text{probability that the event does not occur}}{\text{probability that the event occurs}}$$

In some situations odds are typically reported as odds against. While the odds of an event is a just a single number, odds are often reported as a ratio of whole numbers, e.g., 11 to 1, 7 to 2. In Example 1.6 the odds against the Dodgers winning the World Series are 2.2, which can be reported as 11 to 5 odds (against).

As discussed at the end of Section 1.2.2 bets can be used to discern probabilities or odds.

Example 1.7. Ron and Leslie agree to the following bet. They'll ask Professor Ross if he has a TikTok account. If he does, Leslie will pay Ron $200; if not,

Ron will pay Leslie $100. (Neither has any prior information about whether or not Professor Ross has a TikTok account.)

1. Given this setup, which of the following is being judged as more likely: that Professor Ross has a TikTok account, or that he does not? Why?
2. What are this bet's odds?
3. Ron and Leslie agree that this is a fair bet, and neither would accept worse odds. What is the subjective probability that Professor Ross has a TikTok account?
4. Suppose they were to hypothetically repeat this bet many times, say 3000 times. Given the probability from the previous part, how many times would you expect Leslie to win? To lose? What would you expect Leslie's net dollar winnings to be? In what sense is this bet "fair"? (Remember: Leslie's winnings are Ron's losses and vice versa.)

Solution. to Example 1.7

Show/hide solution

1. The larger potential payout corresponds to the *less* likely event. So Professor Ross is more likely to *not* have a TikTok account than to have one.
2. The payouts are in a 2 to 1 ratio, so the odds that Professor Ross has a TikTok account are 2 to 1 against.
3. The odds that Professor Ross has a TikTok account are 2 to 1 against, so Professor Ross is twice as likely to not have a TikTok account than to have one. This corresponds to a subjective probability[4] that Professor Ross has a TikTok account of 1/3 (and a probability that he does not have one of 2/3).
4. The probability that Leslie wins is 1/3, so you would expect her to win in 1000 of the 3000 repetitions. She wins $200 each time she wins, so you would expect her to win a total of $200,000 on games she wins. The probability that she loses is 2/3, so you would expect her to lose in 2000 of the 3000 repetions. She loses $100 each time, so you would expect her to lose a total of $200,000 on the games she wins. So you would expect Leslie's net winnings to be 0, and likewise for Ron. The bet is fair in the sense that neither party is expected to profit or lose in the long run.

The previous example illustrates that the odds of a fair bet on whether or not an event will occur imply a probability for the event.

[4]Technically, Ron and Leslie could still have different subjective probabilities. Leslie would not agree to worse odds, but she would accept better if Ron offered them. For example, given a potential loss of $200, Leslie would also agree to a potential payout from Ron of $125 rather than $100. That is, Leslie would accept odds of 1.6 to 1 against ($200/125 = 1.6$), corresponding to a subjective probability of 0.385 ($1/(1 + 1.6)$) . So Leslie's subjective probability that Professor Ross has a TikTok account is *at least* 1/3. Similarly, Ron's subjective probability that Professor Ross has a TikTok account is *at most* 1/3.

$$\text{probability that event occurs} = \frac{\text{odds in favor of the event}}{1 + \text{odds in favor of the event}}$$

$$= \frac{1}{1 + \text{odds against the event}}$$

Regardless of the interpretation — long run relative frequency or subjective — probabilities must follow basic logical consistency requirements. If these requirements are mistakenly not satisfied, bad things can happen.

Example 1.8. Donny Don't thinks the Dodgers have a pretty good chance to win the World Series. He thinks their only real competition is the Yankees. The following are Donny's subjective probabilities for which team will win the World Series.

Team	Probability
Los Angeles Dodgers	50%
New York Yankees	25%
Other	10%

1. What is wrong with Donny's probabilities?
2. What are Donny's odds that the Dodgers win? (Consider only Donny's probability that the Dodgers win.)
3. Would Donny agree to a bet where he pays you \$100 if the Dodgers win but you pay him \$100 if the Dodgers do not win?
4. What are Donny's odds that the Yankees win? Would Donny agree to a bet where he pays you \$150 if the Yankees win but you pay him \$50 if the Yankees do not win?
5. What are Donny's odds that a team other than the Dodgers or Yankees wins? Would Donny agree to a bet where he pays you \$180 if an other team wins but you pay him \$20 if the winner is either the Yankees or Dodgers?
6. Suppose you and Donny agree to make all of the bets in the three previous parts. Consider your net profit for each of the outcomes (Dodgers win, Yankees win, other wins). What do you notice?

Solution. to Example 1.8

Show/hide solution

1. Donny's probabilities do not add up to 100%.
2. Donny's odds that the Dodgers win are $\frac{0.5}{0.5} = 1$, or even odds

3. Donny believes that the Dodgers are equally likely to win as to not win so, yes, he would agree to this bet.

4. Donny's odds that the Yankees do not win are $\frac{0.75}{0.25} = 3$, or 3 to 1 odds against the Yankees winning. Donny believes that the Yankees are 3 times more likely to not win than to win. Since the payouts are in a 3 to 1 ratio with the larger payout corresponding to the Yankees winning (the less likely event), then Donny would agree to this bet.

5. Donny's odds that an other team does not win are $\frac{0.9}{0.1} = 9$, or 9 to 1 odds against an other team winning. Donny believes that an other team is 9 times more likely to not win than to win. Since the payouts are in a 9 to 1 ratio with the larger payout corresponding to an other team winning (the less likely event), then Donny would agree to this bet.

6. Given Donny's odds for each outcome, he would agree to each of these bets.

   - If the Dodgers win, you win the first bet but lose the other two, so your net profit is 100 - 50 - 20 = 30.
   - If the Yankees win, you win the second bet but lose the other two, so your net profit is 150 - 100 - 20 = 30
   - If an other team wins, you win the third bet but lose the other two, so your net profit is 180 - 100 - 50 = 30.

   Regardless of the outcome, you are guaranteed to earn a net profit of \$30. That's free money for you with no risk, and pretty bad business on Donny's part.

The previous problem contained an example of a "Dutch book". A **Dutch book**[5] is a set of probabilities and bets which guarantees a profit, regardless of the outcome of the gamble. Probabilities that fail to satisfy logical consistency requirements allow for the possibility of Dutch books. The fact that no one should ever want to get caught in a Dutch book, like Donny was in the previous problem, is used to justify why even subjective probabilities should satisfy logical consistency requirements.

1.4 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

[5] "Book" in the sense of a bookie taking bets.

examples in this section only motivate ideas. We will fill in lots more details throughout the course.

Example 1.9. How old do you think your instructor (Professor Ross) currently is[6]? Consider age on a continuous scale, e.g., you might be 20.73 or 21.36 or 19.50.

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 1.3. (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 midnight on Jan 4, 2021. 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 1.9

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

[6] 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.

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*.

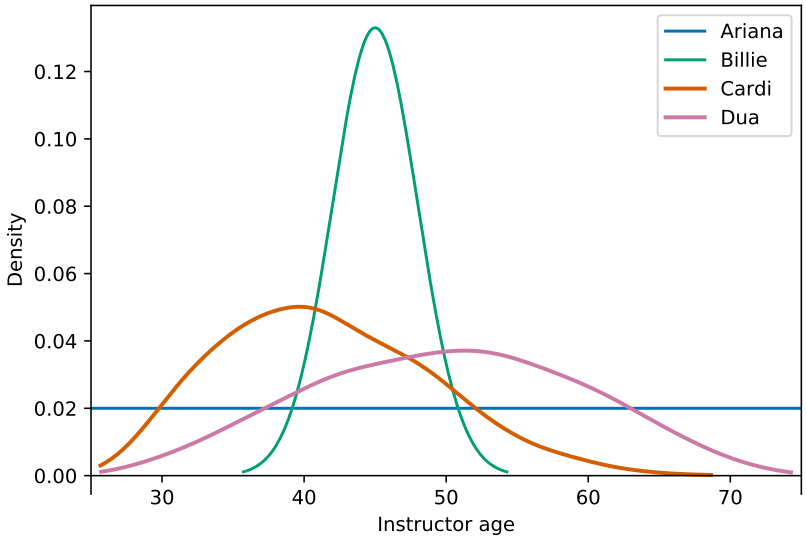
Recall that a **random variable** is a numerical quantity whose value is determined by the outcome of a random or uncertain phenomenon. The random phenomenon might involve physically repeatable randomness, as in "flip a coin 10 times and count the number of heads." But remember that "random" just means "uncertain" and there are lots of different kinds of uncertainty. For example, the total number of points scored in the 2021 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 50% chance that fewer than 45 points will be scored in Superbowl 2021".

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 1.10. Continuing Example 1.9, the plot below displays the subjective probability distribution of the instructor's age of 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 1.10

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 possibe values, while Billie stretches this over the shortest. Ariana is basically saying the instructor 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.

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 1.11. 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. Are you more certain about the value of θ_b or θ_m; how is this reflected in your distributions?
5. 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 1.11

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 the figure on the left in 1.4 for what my subjective probability distributions might look like. Since I am more uncertain about θ_b, its density is "spread out" over a wider range of values.

5. 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 the figure on the right in 1.4 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.
- **Statistical inference** involves using data collected on a *sample* to make conclusions about a *population*.

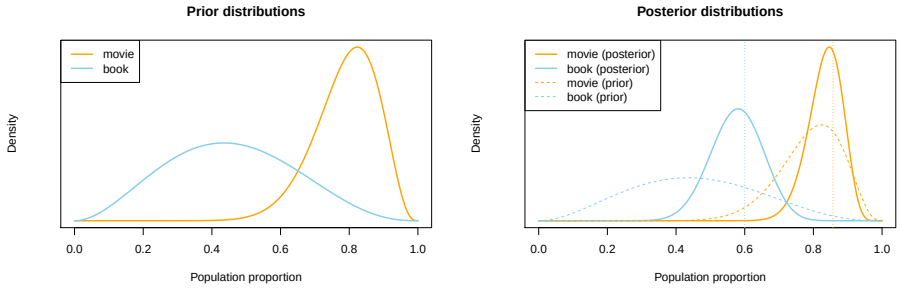


Figure 1.4: Example subjective distributions in Example 1.11. Left: prior to observing sample data. Right: after observing sample data.

- 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 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 2

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 2.1. 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" ("iterative").
- 14% selected "The scientific method identifies unchanging core principles and truths" (unchanging).
- 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. Use the information to construct an appropriate two-way table.

[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.

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 2.1

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 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 2.2. 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?

Solution. to Example 2.2

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, simply rescale the probabilities for Americans who have a postgraduate degree to add up to 1 (by dividing by 0.1511.)

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 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.

[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 — $H_i \cap H_j = \emptyset, i \neq j$.

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 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

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

"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 2.3. 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 a Bachelor's degree (instead of a postgraduate degree).

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 a 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.

Like the scientific method, Bayesian analysis is often an iterative process.

Example 2.4. Suppose that you are presented with six boxes, labeled 0, 1, 2, ..., 5, each containing five marbles. Box 0 contains 0 green and 5 gold marbles, box 1 contains 1 green and 4 gold, and so on with box i containing i green

and $5 - i$ gold. One of the boxes is chosen uniformly at random (perhaps by rolling a fair six-sided die), and then you will randomly select marbles from that box, without replacement. Based on the colors of the marbles selected, you will update the probabilities of which box had been chosen.

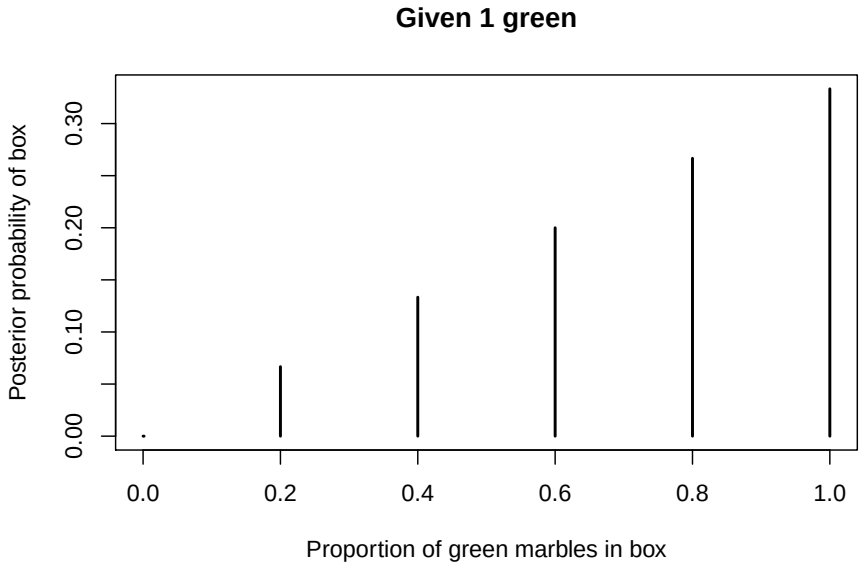
1. Suppose that a single marble is selected and it is green. Which box do you think is the most likely to have been chosen? Make a guess for the posterior probabilities for each box. Then construct a Bayes table to compute the posterior probabilities.

2. Now suppose a second marble is selected from the same box, without replacement, and its color is gold. Which box do you think is the most likely to have been chosen given these two marbles? Make a guess for the posterior probabilities for each box. Then construct a Bayes table to compute the posterior probabilities, *using the posterior probabilities after the selection of the green marble as the new prior probabilities before seeing the gold marble.*
3. Now construct a Bayes table corresponding to the original prior probabilities (1/6 each) and the evidence that the first ball selected was green and the second was gold. How do the posterior probabilities compare to the previous part?
4. In the previous part, the first ball selected was green and the second was gold. Suppose you only knew that in a sample of two marbles, 1 was green and 1 was gold. That is, you didn't know which was first or second. How would the previous part change? Should knowing the order matter? Does it?

Solution. to Example 2.4

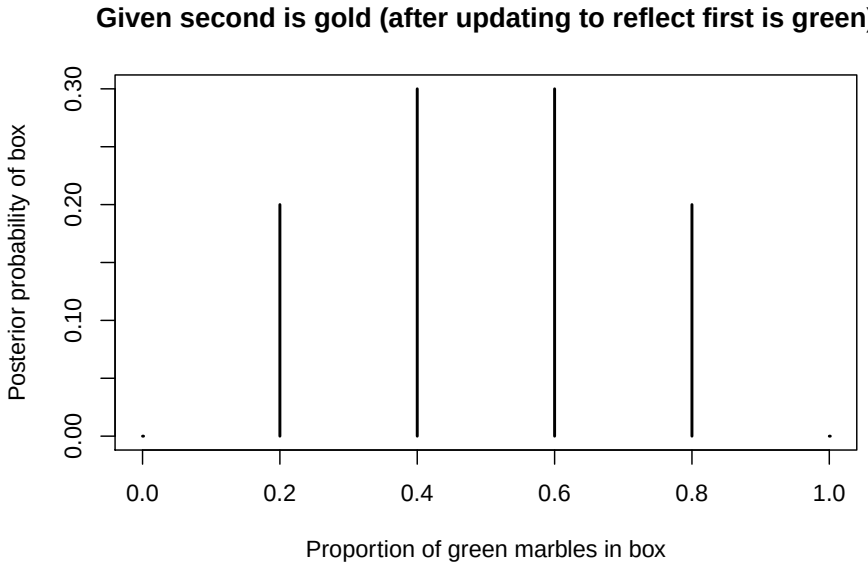
Since the prior probability is the same for each box, the posterior probability will be greatest for the box for which the likelihood of selecting a green marble (the evidence) is greatest, i.e., box 5 which has a likelihood of drawing a green marble of 1. The likelihood of drawing a green marble is 0 for box 0, so box 0 will have a posterior probability of 0. The Bayes table is below, along with a plot of the posterior probabilities. The likelihood column provides the probability of drawing a green marble from each of the boxes, which is $i/5$ for box i. Since the prior is "flat" the posterior probabilities are proportional to the likelihoods.

Green	prior	likelihood	product	posterior
0	0.1667	0.0	0.0000	0.0000
1	0.1667	0.2	0.0333	0.0667
2	0.1667	0.4	0.0667	0.1333
3	0.1667	0.6	0.1000	0.2000
4	0.1667	0.8	0.1333	0.2667
5	0.1667	1.0	0.1667	0.3333
sum	1.0000	NA	0.5000	1.0000



The posterior probabilities above quantify our uncertainty about the box after observing a single randomly selected marble is green. These probabilities serve as the prior probabilities before drawing any additional marbles. After drawing a green marble without replacement, each box has 4 marbles and 1 less green marble than before, and the likelihood of observing a second marble which is gold is computed for each of the 4-marble boxes. For example, after drawing a green marble, box 2 now contains 1 green marble and 3 gold marbles, so the likelihood of drawing a gold marble from box 2 is 3/4. (The likelihood for box 0 is technically undefined because the probability of drawing a green marble first from box 0 is 0. But since the prior probability for box 0 is 0, the posterior probability for box 0 will be 0 regardless of the likelihood.) The Bayes table is below. Since we have observed green and gold in equal proportion in our sample, the posterior probabilities are highest for the boxes with closest to equal proportions of green and gold (box 2 and box 3).

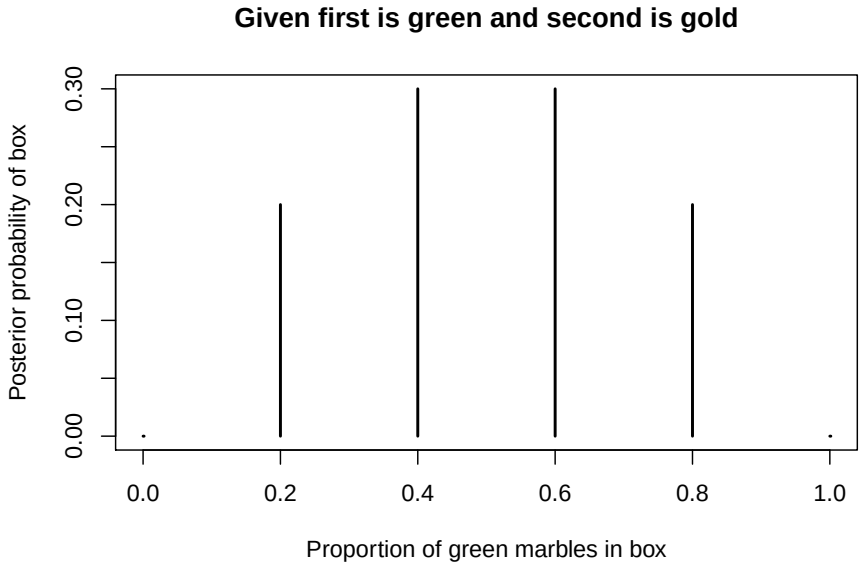
Green	prior	likelihood	product	posterior
0	0.0000	1.00	0.0000	0.0
1	0.0667	1.00	0.0667	0.2
2	0.1333	0.75	0.1000	0.3
3	0.2000	0.50	0.1000	0.3
4	0.2667	0.25	0.0667	0.2
5	0.3333	0.00	0.0000	0.0
sum	1.0000	NA	0.3333	1.0



Above we updated the posterior probabilities after the first marble and again after selecting the second. What if we start with equally likely prior probabilities and only update the posterior probabilities after selecting both marbles? The likelihood now represents the probability of drawing a green and then a gold marble, without replacement, from each of the boxes. For example, for box 2, the probability of drawing a green marble first is 2/5 and the conditional probability of then drawing a gold marble is 3/4, so the probability of drawing green and then gold is $(2/5)(3/4) = 0.3$.

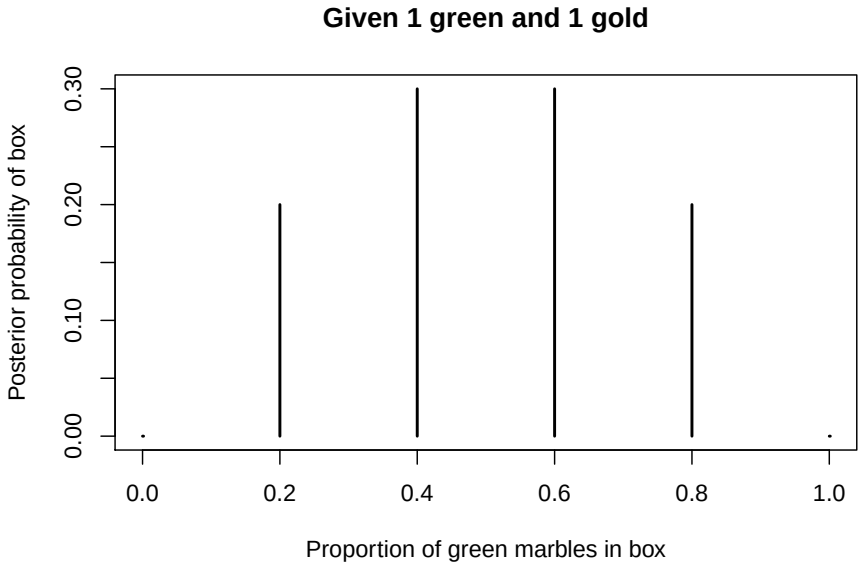
The Bayes table is below. Notice that the posterior probabilities are the same as in the previous part! It doesn't matter if we sequentially update our probabilities after each draw as in the previous part, or only once after the entire sample is drawn. The posterior probabilities are the same either way.

Green	prior	likelihood	product	posterior
0	0.1667	0.0	0.0000	0.0
1	0.1667	0.2	0.0333	0.2
2	0.1667	0.3	0.0500	0.3
3	0.1667	0.3	0.0500	0.3
4	0.1667	0.2	0.0333	0.2
5	0.1667	0.0	0.0000	0.0
sum	1.0000	NA	0.1667	1.0



What if we know the sample contains 1 green and 1 gold marble, but we don't know which was drawn first? It seems that knowing the order shouldn't matter in terms of our posterior probabilities. Technically, the likelihood does change since there are two ways to get a sample with 1 green and 1 gold: green followed by gold or gold followed by green. Therefore, each likelihood will be two times larger than in the previous part. For example, for box 2, the probability of green then gold is $(2/5)(3/4)$ and the probability of gold then green is $(3/5)(2/4)$, so the probability of 1 green and 1 gold is $(2/5)(3/4) + (3/4)(2/5) = 2(0.3)$. However, the *ratios* of the likelihoods have not changed; since each likelihood is twice as large as it was in the previous part, the likelihood from this part is proportional to the likelihood from the previous part. Therefore, since the prior probabilities are the same as in the previous part and the likelihoods are *proportionally* the same as in the previous part, the posterior probabilities will also be the same as in the previous part.

Green	prior	likelihood	product	posterior
0	0.1667	0.0	0.0000	0.0
1	0.1667	0.4	0.0667	0.2
2	0.1667	0.6	0.1000	0.3
3	0.1667	0.6	0.1000	0.3
4	0.1667	0.4	0.0667	0.2
5	0.1667	0.0	0.0000	0.0
sum	1.0000	NA	0.3333	1.0



Bayesian analyses are often performed sequentially. 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 3

Odds and Bayes Factors

Example 3.1. 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 3.1

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 3.2. 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 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 3.3. Continuing Example 3.1

1. In symbols and words, what does one minus the answer to the probability in question in Example 3.1 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 3.3

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 $P(H^c|E) = 1 - 0.062 = 0.938$.

   $$\frac{P(H|E)}{P(H^c\text{E})} = \frac{0.062}{0.938} \approx 0.066$$

   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

$$\text{posterior odds} = \text{prior odds} \times \text{Bayes factor}$$
$$\frac{P(H|E)}{P(H^c|E)} = \frac{P(H)}{P(H^c)} \times BF$$

- 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**

$$\frac{P(H|E)}{P(H^c|E)} = \frac{P(H)}{P(H^c)} \times \frac{P(E|H)}{P(E|H^c)}$$
$$\text{posterior odds} = \text{prior odds} \times \text{Bayes factor}$$

Example 3.4. Continuing Example 3.1. 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 3.1.

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 3.4

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 ration 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	0.9770	0.0489	0.4100
Does not carry HIV	0.9500	0.0740	0.0703	0.5900
sum	1.0000	NA	0.1191	1.0000
ratio	0.0526	13.2027	0.6949	0.6949

Example 3.5. 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 representative selected 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. For now we'll only consider two potential values for θ: 1/2 or 2/3. We could write this as a pair of competing hypotheses.

$$H_1 = \{\theta = 1/2\}$$
$$H_2 = \{\theta = 2/3\}$$

1. Let Y be the number of couples in a random sample of n kissing couples that lean their heads to the right. What is the distribution of Y? Identify it by name and its relevant parameters.
2. Suppose that the researcher observed 12 kissing couples, 8 of whom leaned their heads to the right (a proportion of 8/12=0.667). Compute the relevant likelihoods and the corresponding Bayes factor.
3. Suppose that our prior belief is that the two hypotheses are equally likely. Determine the posterior probabilities for the two hypotheses.
4. Repeat the previous part but with a prior probability of 0.9 for H_1.
5. The full study actually used a sample of 124 kissing couples, of which 80 leaned their heads to the right (a proportion of $80/124 = 0.645$). Compute the relevant likelihoods and the corresponding Bayes factor.
6. Suppose that our prior belief is that the two hypotheses are equally likely. Determine the posterior probabilities for the two hypotheses given the data from the sample of 124 couples.
7. Repeat the previous part but with a prior probability of 0.9 for H_1.
8. Compare the results of the two samples ($n = 12$ versus $n = 124$). What do you observe about the influence of the prior?

Solution. to Example 3.5

1. Y, the number of couples in a random sample of n kissing couples that lean their heads to the right, has a Binomial distribution with parameters n and θ. The probability that y couples in the sample lean right is

$$P(Y = y) = \binom{n}{y}\theta^y(1-\theta)^{n-y}, \quad y = 0, 1, 2, \ldots, n$$

   which can be computed with `dbinom(y, n, theta)` in R.

2. The evidence is the event of observing 8 couples leaning to the right in a sample of 12, that is, $E = \{Y = 8\}$ where Y has a Binomial(12, θ)

distribution. If H_1 is true, Y has a Binomial(12, 1/2), so the likelihood is

$$P(E|H_1) = P(Y = 8|\theta = 1/2) = \binom{12}{8}(1/2)^8(1-1/2)^{12-8} = 0.121,$$

which is `dbinom(8, 12, 1/2)` in R. If H_2 is true, Y has a Binomial(12, 2/3) distribution, so the likelihood is

$$P(E|H_2) = P(Y = 8|\theta = 2/3) = \binom{12}{8}(2/3)^8(1-2/3)^{12-8} = 0.238,$$

which is `dbinom(8, 12, 2/3)` in R. The Bayes factor is

$$BF = \frac{P(E|H_1)}{P(E|H_2)} = \frac{0.121}{0.238} = 0.506$$

Observing 8 couples leaning right in a sample of 12 kissing couples is about 2 times more likely if $\theta = 2/3$ (H_2) than if $\theta = 1/2$ (H_1).

3. If the prior probabilities are equal, then the posterior probabilities will be in proportion to the likelihoods. So the posterior probability of H_2 will be about 2 times greater than the posterior probability of H_1. In terms of odds: the prior odds of H_1 are $0.5/0.5 = 1$, so the posterior odds of H_1 given E are 1×0.506. The Bayes table with the posterior probabilities is below.

theta	prior	likelihood	product	posterior
0.5	0.5	0.1208	0.0604	0.3364
0.667	0.5	0.2384	0.1192	0.6636
sum	1.0	NA	0.1796	1.0000
ratio	1.0	0.5068	0.5068	0.5068

4. Now the prior odds of H_1 are $0.9/0.1 = 9$; the prior probability of H_1 is 9 times greater than the prior probability of H_2. The posterior odds given E are $9 \times 0.506 = 4.56$; the posterior probability of H_1 is 4.56 times greater than the posterior probability of H_2. Even though observing 8 out of 12 couples leaning right is more likely if $\theta = 2/3$ (H_1) than if $\theta = 1/2$ (H_2), the posterior probability of H_1 is greater than the posterior probability of H_2 because of the large discrepancy in the prior probabilities.

theta	prior	likelihood	product	posterior
0.5	0.9	0.1208	0.1088	0.8202
0.667	0.1	0.2384	0.0238	0.1798
sum	1.0	NA	0.1326	1.0000
ratio	9.0	0.5068	4.5614	4.5614

5. Now the evidence is the event of observing 80 couples leaning to the right in a sample of 124, that is, $E = \{Y = 80\}$ where Y has a Binomial(124,

θ) distribution. If H_1 is true, Y has a Binomial(124, 1/2) distribution, so the likelihood is

$$P(E|H_1) = P(Y = 80|\theta = 1/2) = \binom{124}{80}(1/2)^{80}(1-1/2)^{124-80} = 0.00037,$$

which is `dbinom(80, 124, 1/2)` in R. If H_2 is true, Y has a Binomial(124, 2/3) distribution, so the likelihood is

$$P(E|H_2) = P(Y = 80|\theta = 2/3) = \binom{124}{80}(2/3)^{80}(1-2/3)^{124-80} = 0.0658,$$

which is `dbinom(8, 12, 2/3)` in R. The Bayes factor is

$$BF = \frac{P(E|H_1)}{P(E|H_2)} = \frac{0.00037}{0.0657} \approx 0.00566 \approx \frac{1}{176.64}$$

Observing 80 couples leaning right in a sample of 124 kissing couples is about 177 times more likely if $\theta = 2/3$ (H_2) than if $\theta = 1/2$ (H_1).

6. If the prior probabilities are equal, then the posterior probabilities will be in proportion to the likelihoods. So the posterior probability of H_2 will be about 177 times greater than the posterior probability of H_1. In terms of odds: the prior odds of H_1 are $0.5/0.5 = 1$, so the posterior odds of H_1 given E are 1×176.64. The Bayes table with the posterior probabilities is below.

theta	prior	likelihood	product	posterior
0.5	0.5	0.0004	0.0002	0.0056
0.667	0.5	0.0658	0.0329	0.9944
sum	1.0	NA	0.0331	1.0000
ratio	1.0	0.0057	0.0057	0.0057

7. Now the prior odds of H_1 are $09/0.1 = 9$; the prior probability of H_1 is 9 times greater than the prior probability of H_2. The posterior odds given E are $9 \times (1/176.64) = 1/19.63$; the posterior probability of H_2 is 19.63 times greater than the posterior probability of H_1. Even though our prior probability for H_1 was very large, the likelihood of the data is so small under H_1 compared with H_2 that the posterior probability for H_1 is small.

theta	prior	likelihood	product	posterior
0.5	0.9	0.0004	0.0003	0.0485
0.667	0.1	0.0658	0.0066	0.9515
sum	1.0	NA	0.0069	1.0000
ratio	9.0	0.0057	0.0510	0.0510

8. The prior had much more influence with the smaller sample size. When the sample size was large, the data, represented by the likelihoods, had much more weight in determining the posterior probabilities.

Chapter 4

Introduction to Estimation

Example 4.1. 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 representative selected 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.

Let Y be the number of couples in a random sample of n kissing couples that lean to right. Then Y has a Binomial(n, θ) distribution. Suppose that in a sample of $n = 12$ couples $y = 8$ leaned to the right.

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 the next few parts suppose $n = 12$. For now 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 $\theta = 0.1$
3. If $\theta = 0.3$ what is the distribution of Y? Compute and interpret the probability that $Y = 8$ if $\theta = 0.3$
4. If $\theta = 0.5$ what is the distribution of Y? Compute and interpret the probability that $Y = 8$ if $\theta = 0.5$.
5. If $\theta = 0.7$ what is the distribution of Y? Compute and interpret the probability that $Y = 8$ if $\theta = 0.7$.
6. If $\theta = 0.9$ what is the distribution of Y? Compute and interpret the probability that $Y = 8$ if $\theta = 0.9$.

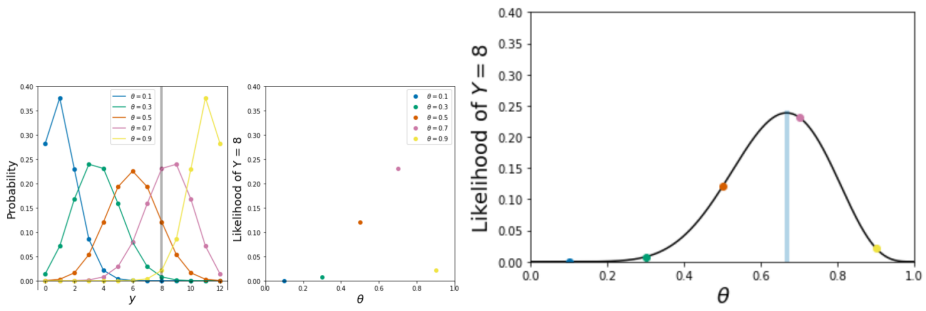
7. Now remember that θ is unknown. If you had to choose your estimate of θ from the values $0.1, 0.3, 0.5, 0.7, 0.9$, which one of these values would you choose based on of observing $y = 8$ couples leaning to the right in a sample of 12 kissing couples? Why?
8. Obviously our choice is not restricted 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.
9. 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?
10. What is our 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 4.1

Show/hide solution

1. Seems reasonable to use the sample proportion $8/12 = 0.667$.
2. 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)`.
3. 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)`.
4. 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)`.
5. 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)`.
6. 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)`.
7. Comparing the above, the probability of observing $y = 8$ is greatest when $\theta = 0.7$, so in some sense, the data seems most "consistent" with $\theta = 0.7$.
8. 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.

9. $f(y|\theta) = P(Y = 8|theta) = \binom{12}{8}\theta^8(1\theta)^{12-8}$. 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 plot below.
10. 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 probability 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 4.2. We'll now take a Bayesian approach to estimating θ in Example 4.1. 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?
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. What does this prior distribution say about θ? 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. What

does this prior distribution say about θ? Redo the previous parts. How does the posterior distribution change?

Solution. to Example 4.2

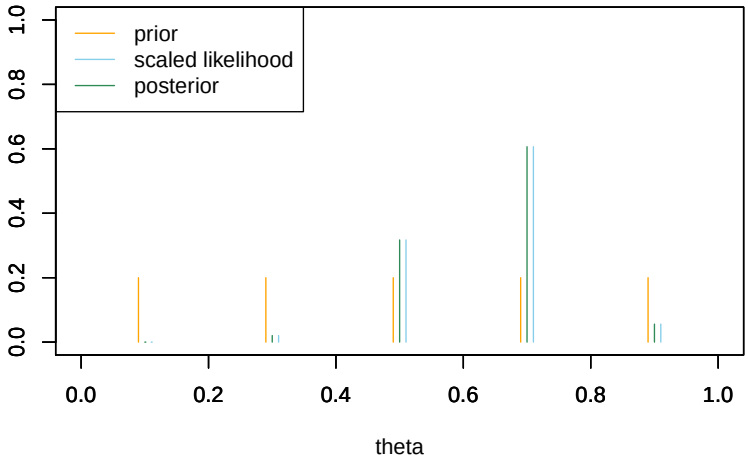
1. See plot below; the prior is "flat".
2. The likelihood is computed as in Example 4.1.
3. See the Bayes table below. Since the prior is flat, the posterior is proportional to the likelihood.

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

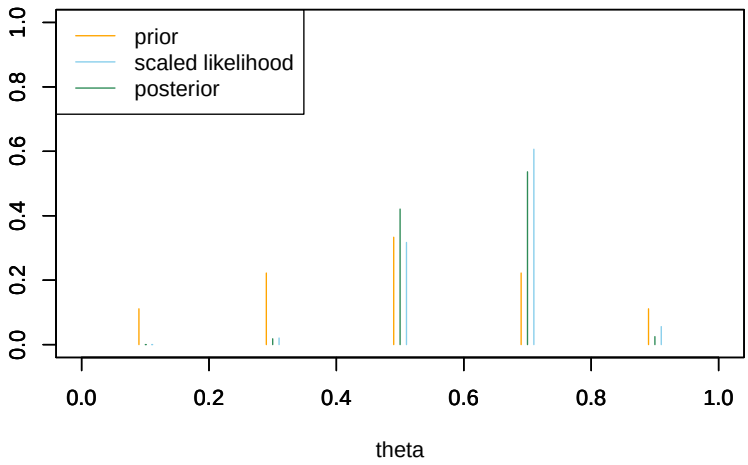
```
# plots
plot(theta-0.01, prior, type='h', xlim=c(0, 1), ylim=c(0, 1), col="orange", xlab='
```

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



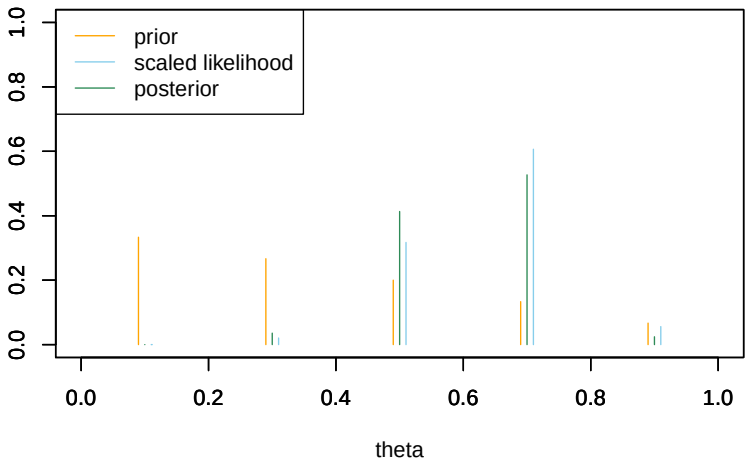
4. See table and plot below. Because the posterior probability is greater for 0.5 than for 0.7, the posterior probability of $\theta = 0.5$ is greater than in the previous part, and the posterior probability of $\theta = 0.7$ is less.

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



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



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 4.3. Continuing Example 4.2. 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: $0, 0.0001, 0.0002, \ldots, 0.9999, 1$. 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?
3. Now assume a prior distribution which is proportional to $1 - 2|\theta - 0.5|$ for $\theta = 0, 0.0001, 0.0002, \ldots, 0.9999, 1$. 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?
4. Now assume a prior distribution which is proportional to $1 - \theta$ for $\theta = 0, 0.0001, 0.0002, \ldots, 0.9999, 1$. 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?
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?

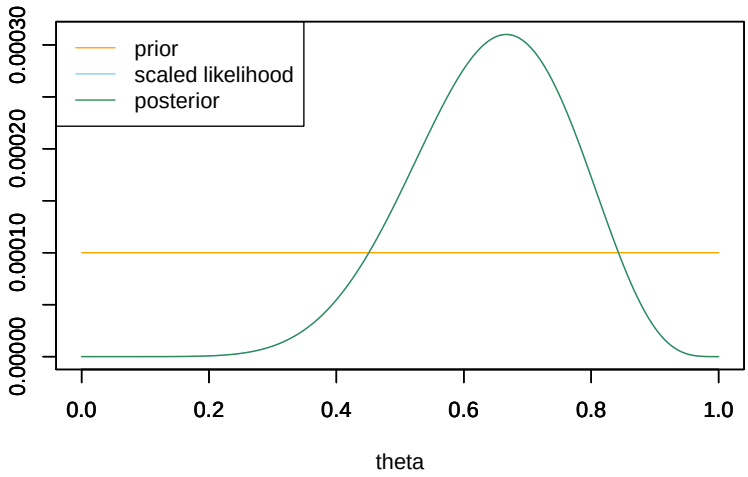
Solution. to Example 4.3

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

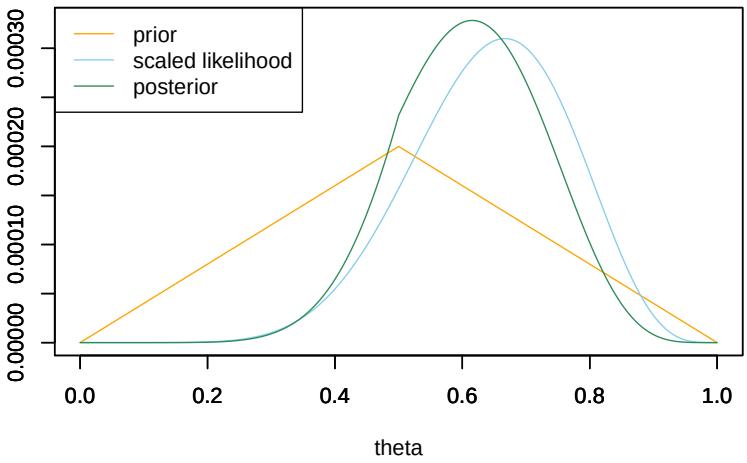
```
# prior
theta = seq(0, 1, 0.0001)
prior = rep(1, length(theta))
```

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

plot_posterior(theta, prior, likelihood)
```



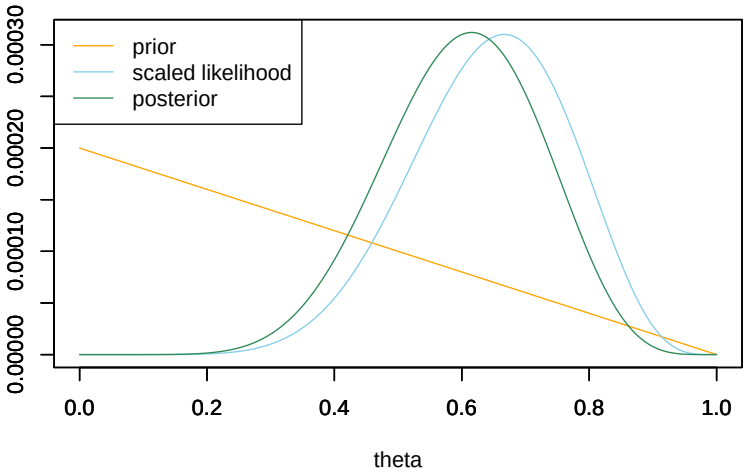
3. See plot below. The posterior is a compromise between the "triangular" prior which places highest prior probability near 0.5, and the likelihood. For this posterior, the posterior probability is greater near 0.5 than for the one in the previous part.

```
# 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 part 0.5, as opposed to what happens with the uniform prior.

```
# prior
theta = seq(0, 1, 0.0001)
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 4.4. Continuing Example 4.1. 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.
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?

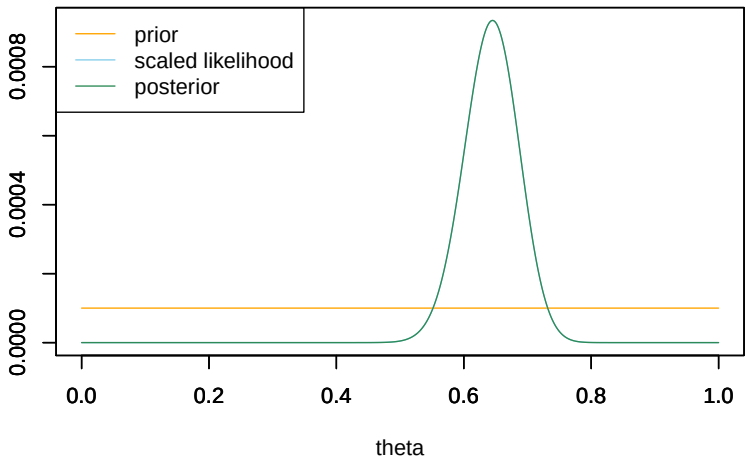
3. Now assume a prior distribution which is proportional to $1 - 2|\theta - 0.5|$ for $\theta = 0, 0.0001, 0.0002, \ldots, 0.9999, 1$. Then suppose again that $y = 80$ couples in a sample of $n = 124$ kissing couples lean right. Use software to plot the prior distribution, the likelihood function, and 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?

Solution. to Example 4.4

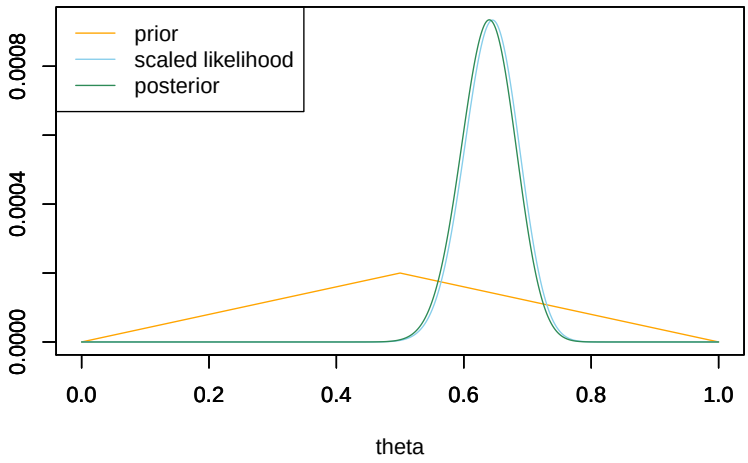
1. See plot below. The likelihood function is $f(y = 80|\theta) = \binom{124}{80}\theta^{80}(1-\theta)^{124-80}, 0 \le \theta \le 1$, the likelihood of observing a value of $y = 80$ from a Binomial(124, θ) distribution (`dbinom(80, 124, theta)`). The maximum likelihood estimate of θ is the sample proportion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

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.

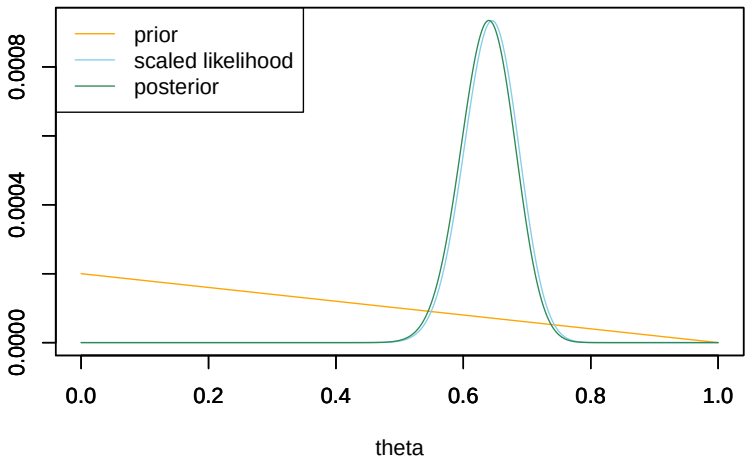
```
# prior
theta = seq(0, 1, 0.0001)
prior = rep(1, length(theta))
prior = prior / sum(prior)

# 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, they 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.

Recall that the likelihood function is the probability (or density for continuous data) of observing the sample data y viewed as a *function of the parameter* θ. When the data y takes values on a continuous scale, the likelihood is determined by the *probability density function* of Y given θ, $f(y|\theta)$. In the likelihood function, the observed value of the data y is treated as a fixed constant, and the likelihood of observing that y is evaluated for all potential values of θ.

Recall that a continuous random variable[1] U follows a **Normal (a.k.a., Gaussian) distribution** with mean μ and standard deviation $\sigma > 0$ if its probability

[1]Why U and not X or Y? In a Bayesian analysis, we might assume the data follows a Normal distribution, but we might also use a Normal distribution to quantify the uncertainty about a parameter. We generally associate X and Y with data. U is supposed to represent a more general variable, which could be either data or a parameter.

density function is[2]

$$f_U(u) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{u-\mu}{\sigma}\right)^2\right), \quad -\infty < u < \infty.$$
$$\propto \frac{1}{\sigma} \exp\left(-\frac{1}{2}\left(\frac{u-\mu}{\sigma}\right)^2\right), \quad -\infty < u < \infty.$$

The constant $1/\sqrt{2\pi}$ ensures that the total area under the density is 1, but it doesn't effect the shape of the density.

In R, `dnorm(u, mu, sigma)`.

Example 4.5. Assume body temperatures (degrees Fahrenheit) of healthy adults follow a Normal distribution with unknown mean μ and known[3] standard deviation $\sigma = 1$. Suppose we wish to estimate μ, the population mean healthy human body temperature.

1. Assume first the following discrete prior distribution for μ which places probability 0.10, 0.25, 0.30, 0.25, 0.10 on the values 97.6, 98.1, 98.6, 99.1, 99.6, respectively. Suppose a single temperature value of 97.9 is observed. Construct a Bayes table and find the posterior distribution of μ. In particular, how do you determine the likelihood?
2. Now suppose a second temperature value, 97.5, is observed, independently of the first. Construct a Bayes table and find the posterior distribution of μ after observing these two measurements, using the posterior distribution from the previous part as the prior distribution in this part.
3. Now consider the original prior again. Determine the likelihood of observing temperatures of 97.9 and 97.5 in a sample of size 2. Then construct a Bayes table and find the posterior distribution of μ after observing these two measurements. Compare to the previous part.
4. Consider the original prior again. Suppose that we take a random sample of two temperature measurements, but instead of observing the two individual values, we only observe that the sample mean is 97.7. Determine the likelihood of observing a sample mean of 97.7 in a sample of size 2. (Hint: if $\bar{Y}$ is the sample mean of n values from a $N(\mu, \sigma)$ distribution, what is the distribution of $\bar{Y}$?) Then construct a Bayes table and find the posterior distribution of μ after observing this sample mean. Compare to the previous part.

Solution. to Example 4.5

1. The likelihood is determined by evaluating the Normal(μ, 1) density at $y = 97.9$ for different values of μ: `dnorm(97.9, mu, 1)` or

$$f(97.9|\mu) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{97.9-\mu}{1}\right)^2\right)$$

[2] exp is just another way of writing the exponential function, $\exp(u) = e^u$.

[3] It's unrealistic to assume the population standard deviation is known. We'll consider the case of unknown standard deviation later.

See the table below. Posterior probability is shifted towards the smaller values of μ since those give the higher likelihood of the observed value $y = 97.9$.

```
# prior
theta = seq(97.6, 99.6, 0.5)
prior = c(0.10, 0.25, 0.30, 0.25, 0.10)
prior = prior / sum(prior)

# data
y = 97.9 # single observed value
sigma = 1


# likelihood
likelihood = dnorm(y, theta, sigma) # 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, digits = 4, align = 'r')
```

theta	prior	likelihood	product	posterior
97.6	0.10	0.3814	0.0381	0.1326
98.1	0.25	0.3910	0.0978	0.3400
98.6	0.30	0.3123	0.0937	0.3258
99.1	0.25	0.1942	0.0485	0.1688
99.6	0.10	0.0940	0.0094	0.0327

2. See the table below. More posterior probability is shifted towards the smaller values of μ.

```
# prior
prior = posterior

# data
y = 97.5 # single observed value
sigma = 1
```

```r
# likelihood
likelihood = dnorm(y, theta, sigma) # 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, digits = 4, align = 'r')
```

theta	prior	likelihood	product	posterior
97.6	0.1326	0.3970	0.0527	0.2048
98.1	0.3400	0.3332	0.1133	0.4407
98.6	0.3258	0.2179	0.0710	0.2761
99.1	0.1688	0.1109	0.0187	0.0728
99.6	0.0327	0.0440	0.0014	0.0056

3. See the table below. Since the two measurements are independent, the likelihood is the product of the likelihoods for $y = 97.9$ and $y = 97.5$. The posterior is the same in the previous part. It doesn't matter if we update the posterior after each observations, or all at once.

```r
# prior
theta = seq(97.6, 99.6, 0.5)
prior = c(0.10, 0.25, 0.30, 0.25, 0.10)
prior = prior / sum(prior)

# data
y = c(97.9, 97.5) # two observed values
sigma = 1


# likelihood
likelihood = dnorm(y[1], theta, sigma) * dnorm(y[2], theta, sigma)  # function of

# 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')
```

theta	prior	likelihood	product	posterior
97.6	0.10	0.1514	0.0151	0.2048
98.1	0.25	0.1303	0.0326	0.4407
98.6	0.30	0.0680	0.0204	0.2761
99.1	0.25	0.0215	0.0054	0.0728
99.6	0.10	0.0041	0.0004	0.0056

4. For a sample of size n from a $N(\mu, \sigma)$ distribution, the sample mean follows a $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ distribution. The likelihood is determined by evaluating the Normal(μ, $\frac{1}{\sqrt{2}}$) density at $y = 97.7$ for different values of μ: `dnorm(97.7, mu, 1 / sqrt(2))` or

$$f_{\bar{Y}}(97.7|\mu) \propto \exp\left(-\frac{1}{2}\left(\frac{97.7-\mu}{1/\sqrt{2}}\right)^2\right)$$

See the table below. While the likelihood is not the same as in the previous part, it is *proportionally* the same; that is, the likelihood in this part has the same *shape* as the likelihood in the previous part. Therefore, the posterior distributions are the same.

```
# prior
theta = seq(97.6, 99.6, 0.5)
prior = c(0.10, 0.25, 0.30, 0.25, 0.10)
prior = prior / sum(prior)

# data
n = 2
y = 97.7 # sample mean
sigma = 1


# likelihood
likelihood = dnorm(y, theta, sigma / sqrt(n)) # 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, digits = 4, align = 'r')
```

theta	prior	likelihood	product	posterior
97.6	0.10	0.5586	0.0559	0.2048
98.1	0.25	0.4808	0.1202	0.4407
98.6	0.30	0.2510	0.0753	0.2761
99.1	0.25	0.0795	0.0199	0.0728
99.6	0.10	0.0153	0.0015	0.0056

It is often not necessary to know all the individual data values to evaluate the *shape* of the likelihood as a function of the parameter θ, but rather simply the values of a few summary statistics.
For example, when estimating the population mean of a Normal distribution with known standard deviation σ, it is sufficient to know the sample mean for the purposes of evaluating the shape of the likelihood of the observed data under different potential values of the population mean.

If $Y_1, \ldots, Y_n$ is a random sample from a $N(\mu, \sigma)$ distribution, then $\bar{Y}$ has a $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ distribution

- σ measures the unit-to-unit variability of individual values of the variable over all possible units in the population. For example, how much do body temperatures vary from person-to-person over many people?
- $\frac{\sigma}{\sqrt{n}}$ measures the sample-to-sample variability of sample means over all possible samples of size n from the population. For example, how much do sample mean body temperatures from sample to sample over many samples of n people?

Example 4.6. Continuing the previous example. We'll now use a grid approximation and assume that any multiple of 0.0001 between 96.0 and 100.0 is a possible value of μ: $96.0, 96.0001, 96.0002, \ldots, 99.9999, 100.0$.

1. Assume a discrete uniform prior distribution over μ values in the grid. Suppose that the sample mean temperature is $\bar{y} = 97.7$ in a sample of $n = 2$ temperature measurements. 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 μ?

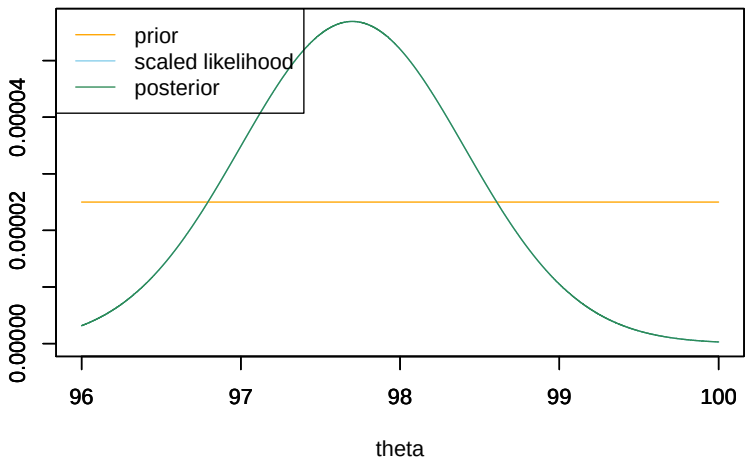
2. Now assume a prior distribution which is proportional to a Normal distribution with mean 98.6 and standard deviation 0.7 over μ values in the grid. Suppose that the sample mean temperature is $\bar{y} = 97.7$ in a sample of $n = 2$ temperature measurements. Use software to plot the prior distribution, the (scaled) likelihood function, and then find the posterior and plot it. Describe the posterior distribution. What does it say about μ?
3. Compare the posterior distributions corresponding to the two different priors. How does each posterior distribution compare to the prior and the likelihood? Comment on the influence that the prior distribution has.

Solution. to Example 4.6

1. Since the prior is flat, the posterior has the same shape as the likelihood. The highest posterior probability is near the observed sample mean of 97.7.

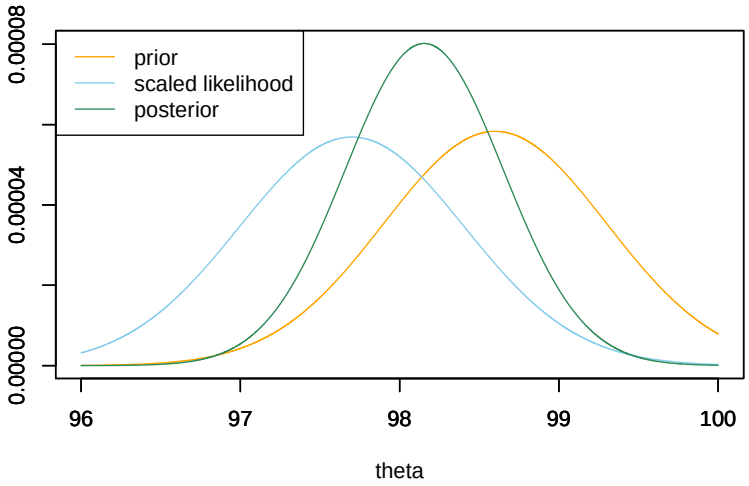
```
# prior
theta = seq(96, 100, 0.0001)
prior = rep(1, length(theta))
prior = prior / sum(prior)

# data
n = 2 # sample size
y = 97.7 # sample mean
sigma = 1

# likelihood
likelihood = dnorm(y, theta, sigma / sqrt(n)) # 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=range(theta), ylim=ylim, col="orange", xlab='theta', yla
  par(new=T)
  plot(theta, likelihood/sum(likelihood), type='l', xlim=range(theta), ylim=ylim, col="skybl
  par(new=T)
  plot(theta, posterior, type='l', xlim=range(theta), ylim=ylim, col="seagreen", xlab='', yl
  legend("topleft", c("prior", "scaled likelihood", "posterior"), lty=1, col=c("orange", "sk
}
```

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



2. Be sure to distinguish between the Normal distribution in the prior, which quantifies our prior uncertainty about μ, and the Normal distribution used to determine the likelihood which models variability of temperatures in the population. The poster is a compromise between likelihood and prior.

```
# prior
theta = seq(96, 100, 0.0001)
prior = dnorm(theta, 98.6, 0.7)
prior = prior / sum(prior)

# data
n = 2 # sample size
y = 97.7 # sample mean
sigma = 1

# likelihood
likelihood = dnorm(y, theta, sigma / sqrt(n)) # function of theta

plot_posterior(theta, prior, likelihood)
```



3. When the prior is flat, the posterior has the shape of the likelihood. Otherwise, the posterior is a compromise between prior and likelihood. When the sample size is so small, the prior will have a lot of influence on the posterior.

Example 4.7. Continuing the previous example. In a recent study[4], the sample mean body temperature in a sample of 208 healthy adults was 97.7 degrees F.

We'll again use a grid approximation and assume that any multiple of 0.0001 between 96.0 and 100.0 is a possible value of μ: $96.0, 96.0001, 96.0002, \ldots, 99.9999, 100.0$.

1. Before performing a Bayesian analysis, use software to plot the likelihood, and compute the maximum likelihood estimate of μ based on this data.
2. Assume a discrete uniform prior distribution over μ values in the grid. 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 μ?
3. Now assume a prior distribution which is proportional a Normal distribution with mean 98.6 and standard deviation 0.7 over μ values in the grid. 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 μ?
4. Compare the posterior distributions corresponding to the two different priors. How does each posterior distribution compare to the prior and the likelihood? Comment on the influence that the prior distribution has. How does this compare to the $n = 2$ situation?

[4]Source and a related article.

Solution. to Example 4.7

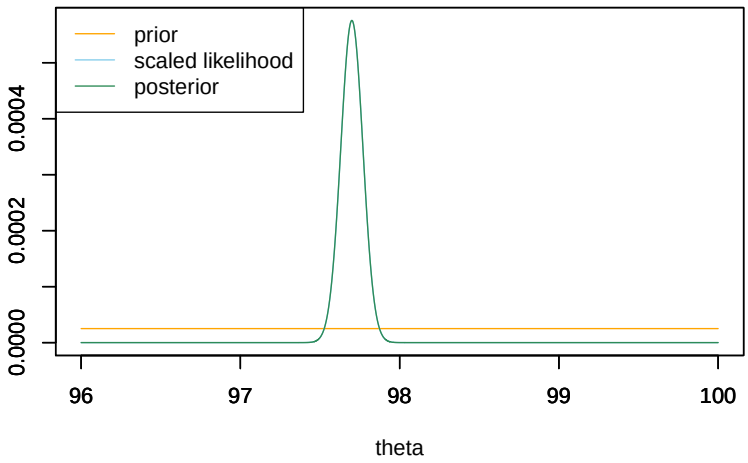
1. The likelihood is determined by evaluating the Normal(μ, $\frac{1}{\sqrt{208}}$) density at $y = 97.7$ for different values of μ: `dnorm(97.7, mu, 1 / sqrt(208))` or

$$f_{\bar{Y}}(97.7|\mu) \propto \exp\left(-\frac{1}{2}\left(\frac{97.7-\mu}{1/\sqrt{208}}\right)^2\right)$$

   See a plot of the likelihood below. The MLE of μ is the observed sample mean of 97.7.

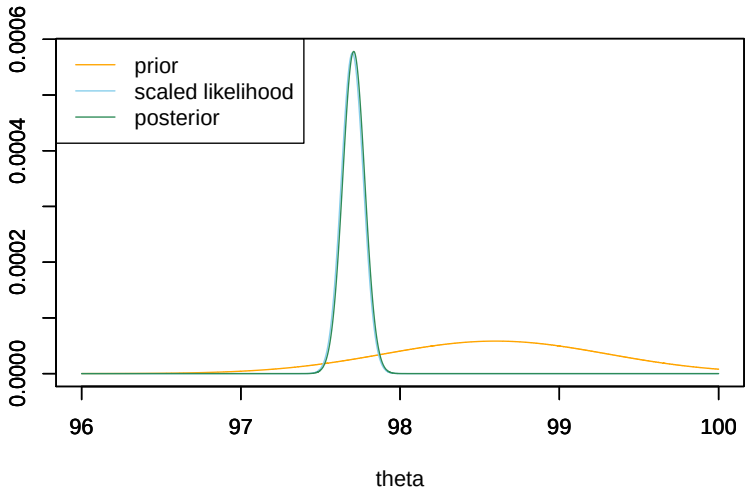
2. Since the prior is flat, the posterior has the same shape as the likelihood. With such a large sample size, the likelihood is pretty peaked. So the posterior probability is concentrated in a fairly narrow range of values around 97.7.

```
# prior
theta = seq(96, 100, 0.0001)
prior = rep(1, length(theta))
prior = prior / sum(prior)

# data
n = 208 # sample size
y = 97.7 # sample mean
sigma = 1

# likelihood
likelihood = dnorm(y, theta, sigma / sqrt(n)) # function of theta

plot_posterior(theta, prior, likelihood)
```



3. Even though the prior probability is highest near 98.6, the likelihood at these values is so small that they have small posterior probability. The posterior distribution is about the same as in the previous part.

```
# prior
theta = seq(96, 100, 0.0001)
prior = dnorm(theta, 98.6, 0.7)
prior = prior / sum(prior)

# data
n = 208 # sample size
y = 97.7 # sample mean
sigma = 1

# likelihood
likelihood = dnorm(y, theta, sigma / sqrt(n)) # function of theta

plot_posterior(theta, prior, likelihood)
```



4. The posterior distributions are about the same in each case. With a large sample size, the likelihood is fairly peaked, and so the likelihood is close to 0 outside of a narrow range of values around the observed sample mean of 97.7. Therefore, the posterior probability is concentrated in this range, regardless of the prior.

Chapter 5

Introduction to Inference

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 5.1. Continuing the kissing study in Example 4.2 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.

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".
4. Find the variance of the prior distribution θ, a.k.a, the "prior variance".
5. Find the standard deviation of the prior distribution of θ, a.k.a, the "prior standard deviation".

   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	0.1111	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

6. Find the mode of the posterior distribution of θ, a.k.a., the "posterior mode".
7. Find the median of the posterior distribution of θ, a.k.a., the "posterior median".
8. Find the expected value of the posterior distribution of θ, a.k.a., the "posterior mean".
9. Find the variance of the posterior distribution θ, a.k.a, the "posterior variance".
10. Find the standard deviation of the posterior distribution of θ, a.k.a, the "posterior standard deviation".
11. How have the posterior values changed from the respective prior values?

Solution. to Example 5.1

Show/hide solution

1. The prior mode is 0.5, the value of θ with the greatest prior probability.
2. The prior median is 0.5. (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 $\theta = 0.5$.)
3. The prior mean is 0.5. Remember that an expected value is a probability-weighted average value
$$0.1(1/9) + 0.3(2/9) + 0.5(3/9) + 0.7(2/9) + 0.9(1/9) = 0.5.$$
4. The prior variance is 0.0533. Remember that variance is the probability-weighted average squared deviation from the mean
$$(0.1-0.5)^2(1/9)+(0.3-0.5)^2(2/9)+(0.5-0.5)^2(3/9)+(0.7-0.5)^2(2/9)+(0.9-0.5)^2(1/9) = 0.0533$$
5. The prior standard deviation is 0.231. Remember that standard deviation is the square root of the variance: $\sqrt{0.0533} = 0.231$.
6. The posterior mode is 0.7, the value of θ with the greatest posterior probability.
7. The posterior median is 0.7. (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 $\theta = 0.5$.)
8. The posterior mean is 0.614. Now the posterior probabilities are used in the probability-weighted average value
$$0.1(0.000) + 0.3(0.018) + 0.5(0.421) + 0.7(0.536) + 0.9(0.025) = 0.614.$$

9. The posterior variance is 0.013. Now the posterior probabilities are used in the probability-weighted average squared deviation from the mean

   $$(0.1-0.5)^2(0.000)+(0.3-0.5)^2(0.018)+(0.5-0.5)^2(0.421)+(0.7-0.5)^2(0.536)+(0.9-0.5)^2(0.025) = 0.013$$

10. The posterior standard deviation is $\sqrt{0.013} = 0.115$.

11. The measures of center (mean, median, mode) shift from the prior value of 0.5 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 (0.614) lies between the prior mean (0.5) and the sample propotion (0.667).

    The measures of variability (SD, variance) are smaller for the posterior than for the prior. After observing some data, there is less uncertainty about θ. The prior probability is "spread" over the five possible values of θ, while almost all of the posterior probability is concentrated at 0.5 and 0.7.

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 point estimates of θ are the posterior mean, the posterior median, and 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, θ plays the role of the variable U and the posterior distribution provides the probability-weights.

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.

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 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 $\text{SD}(U) = \sqrt{\text{Var}(U)}$. Standard deviation is measured in the same measurement units as the variable itself.

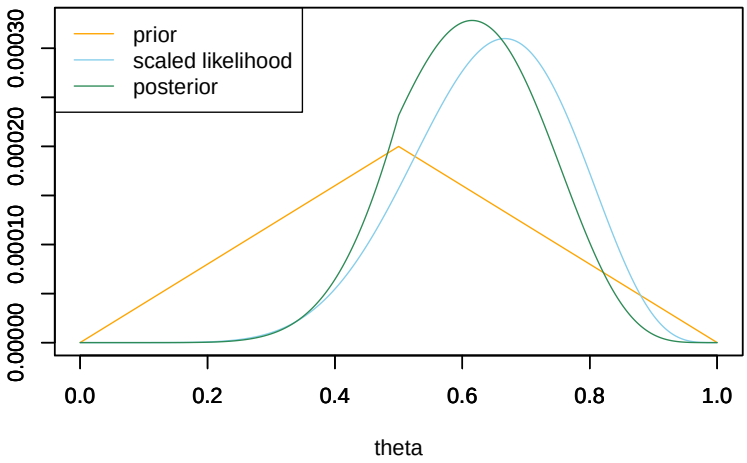
In the calculation of a posterior standard deviation, θ plays the role of the variable U and the posterior distribution provides the probability-weights.

Example 5.2. Continuing the kissing study in Example 4.3. Now assume a prior distribution which is proportional to $1 - 2|\theta - 0.5|$ for $\theta = 0, 0.0001, 0.0002, \ldots, 0.9999, 1$. 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".
4. Find the variance of the prior distribution θ, a.k.a, the "prior variance".
5. Find the standard deviation of the prior distribution of θ, a.k.a, the "prior standard deviation".
6. For what range of values is the prior probability that θ lies in that range equal to 95%?
7. Find the prior probability that θ is greater than 0.5.

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

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



8. Find the mode of the posterior distribution of θ, a.k.a., the "posterior mode".
9. Find the median of the posterior distribution of θ, a.k.a., the "posterior median".
10. Find the expected value of the posterior distribution of θ, a.k.a., the "posterior mean".
11. Find the variance of the posterior distribution θ, a.k.a, the "posterior variance".
12. Find the standard deviation of the posterior distribution of θ, a.k.a, the "posterior standard deviation".
13. For what range of values is the posterior probability that θ lies in that range equal to 95%?
14. Find the posterior probability that θ is greater than 0.5.
15. How have the posterior values changed from the respective prior values?

Solution. to Example 5.2

Show/hide solution

```
## prior

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

```
## [1] 0.5
```

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

```
## [1] 0.5
```

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

```
## [1] 0.5
```

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

```
## [1] 0.04166666
```

```
# prior sd
sqrt(prior_var)
```

```
## [1] 0.2041241
```

```
# prior 95% credible interval
prior_cdf = cumsum(prior)
c(theta[max(which(prior_cdf <= 0.025))], theta[min(which(prior_cdf >= 0.975))])
```

```
## [1] 0.1117 0.8882
```

```
# prior prob(theta > 0.5)
sum(prior[theta > 0.5])
```

```
## [1] 0.4999
```

```
## posterior

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

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

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

```
## [1] 0.6126
```

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

```
## [1] 0.6113453
```

```
# posterior variance
post_var = sum(theta ^ 2 * posterior) - post_ev ^ 2
post_var
```

```
## [1] 0.01302593
```

```
# posterior sd
sqrt(post_var)
```

```
## [1] 0.1141312
```

```
# posterior 95% credible interval
posterior_cdf = cumsum(posterior)
c(theta[max(which(posterior_cdf <= 0.025))], theta[min(which(posterior_cdf >= 0.975))])
```

```
## [1] 0.3857 0.8253
```

```
# prior prob(theta > 0.5)
sum(posterior[theta > 0.5])
```

```
## [1] 0.829704
```

In the previous problem, the center of the posterior distribution is closer to the sample proportion than the center of the prior distribution. There is less uncertainty about θ after observing some data, so the posterior standard deviation is less than the prior standard deviation. The 95% posterior interval is narrower than the prior interval, and its centered is shifted towards the posterior mean. The posterior concentrates more probability above 0.5 than the prior does.

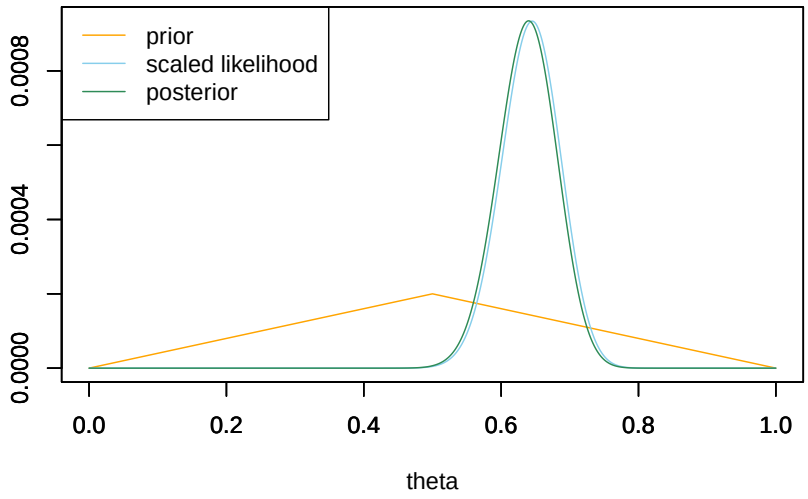
Bayesian inference for a parameter is based on its posterior distribution. Since a Bayesian analysis treats parameters as random variables, it is possible to make posterior probability statements about a parameter.

A Bayesian **credible interval** is an interval of values for the parameter that has at least the specified probability, e.g., 95%. 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. The endpoints of a 95% ***central*** **posterior credible interval** correspond to the 2.5th and the 97.5th percentiles of the posterior distribution.

Central credible intervals are easier to compute, but are not the only or most widely used credible intervals. A 95% **highest posterior density interval** is the interval of values that contains 95% of the 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 and unimodal, central posterior credible intervals and highest posterior density intervals are similar.

Example 5.3. Continuing the kissing study in Example 4.4, we'll now perform a Bayesian analysis on the actual study data in which 80 couples out of a sample of 124 leaned right. Assume a prior distribution which is proportional to $1 - 2|\theta - 0.5|$ for $\theta = 0, 0.0001, 0.0002, \ldots, 0.9999, 1$. Use software to answer the following questions. Recall the prior, likelihood, and posterior.

```
# prior
theta = seq(0, 1, 0.0001)
prior = 1 - 2 * abs(theta - 0.5) # shape of prior
prior = prior / sum(prior) # scales so that prior sums to 1

# 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 a 95% central posterior credible interval for θ. How does the credible interval compare to the one from the previous example (with $n = 12$)?
2. Write a clearly worded sentence reporting the credible interval from the previous part in context.
3. Given the shape of the posterior distribution, how could your approximate a posterior 95% central posterior credible interval?
4. Find the posterior probability that θ is greater than 0.5. How does this probability compare to the one from the previous example (with $n = 12$)?
5. Write a clearly worded sentence reporting the probability from the previous part in context.
6. Given the shape of the posterior distribution, how could your approximate the posterior probability that θ is greater than 0.5?
7. Now consider the other two prior distributions from Example 4.4. Would any of the conclusions from this problem change substantially if we had chosen one of the other priors?

Solution. to Example 5.3

Show/hide solution

```
## posterior

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

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

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

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

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

```
## [1] 0.6378017
```

```
# posterior variance
post_var = sum(theta ^ 2 * posterior) - post_ev ^ 2
post_var
```

```
## [1] 0.001803819
```

```
# posterior sd
sqrt(post_var)
```

```
## [1] 0.04247139
```

```
# posterior 95% credible interval
posterior_cdf = cumsum(posterior)
c(theta[max(which(posterior_cdf <= 0.025))], theta[min(which(posterior_cdf >= 0.975))])
```

```
## [1] 0.5526 0.7188
```

```
# prior prob(theta > 0.5)
sum(posterior[theta > 0.5])
```

```
## [1] 0.999182
```

Show/hide solution

1. A 95% central posterior credible interval for θ is [0.552, 0.719]. This interval is narrower (more precise) than the one for $n = 12$. With a larger sample size, the likelihood is more "peaked" and so the posterior probability is concentrated over a narrower range of values.

2. There is a posterior probability of 95% that the population proportion of kissing couples who lean heads to the right is between 0.552 and 0.719.
3. The posterior distribution is approximately Normal, with posterior mean 0.638 and posterior standard deviation 0.042. A Normal distribution places 95% of probability on values that fall within 2 standard deviations of the mean. So an approximate 95% posterior credible interval has endpoints $0.638 \pm 2 \times 0.042$, yielding an interval of $[0.552, 0.723]$.
4. The posterior probability that θ is greater than 0.5 is 0.9992. This probability compare to the one from the previous example since with the larger sample size, the posterior standard deviation is smaller and the posterior distribution is concentrated even more near the observed sample proportion of 0.645.
5. There is a posterior probability of 99.92% that the population proportion of kissing couples who lean heads to the right is greater than 0.5.
6. Use standardization (z-scores) and the empirical rule. A value of 0.5 is 3.24 standard deviations below the posterior mean: $(0.5 - 0.638)/0.042 = -3.24$. By the empirical rule for Normal distributions, 99.7% of the probability corresponds to values within 3 standard deviations of the mean. Therefore the posterior probability that θ is less than 0.5 is pretty small.
7. We saw in Example 4.4 that with the sample size of $n = 124$, the posterior distribution was basically the same for each of the three priors. Sothe conclusions from this problem would not change substantially if we had chosen one of the other priors.

In many situations, the posterior distribution of a single parameter is approximately Normal, so an approximate 95% credible interval has endpoints

$$\text{posterior mean} \pm 2 \times \text{posterior SD}$$

Also, posterior probabilities of hypotheses about a parameter can often be approximated with Normal distribution calculations — standardizing and using the empirical rule.

Example 5.4. We'll now compare to the Bayesian analysis in the previous example to a frequentist analysis. Recall the actual study data in which 80 couples out of a sample of 124 leaned right.

1. Compute a 95% confidence interval for θ.
2. Write a clearly worded sentence reporting the confidence interval in context.
3. Explain what "95% confidence" means?
4. Conduct a (null) hypothesis (significance) test of whether the sample data provide strong evidence that more than half of all kissing couples lean their heads to the right. Compute the corresponding p-value.
5. Write a clearly worded sentence reporting the hypothesis test in context.
6. Interpret the p-value.

7. Compare the *numerical results* of the Bayesian and frequentist analysis. How does the *interpretation* of these results differ between the two approaches?

Solution. to Example 5.4

Show/hide solution

1. The observed sample proportion is $\hat{p} = 80/124 = 0.645$ and its standard error is $\sqrt{\hat{p}(1-\hat{p})/n}$. The usual formula for a 95% confidence interval for a population prportion is

$$\hat{p} \pm 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

   Plugging in $n = 124$ and $\hat{p} = 80/124$ yields the interval $[0.559, 0.731]$.

2. We estimate with 95% confidence that the population proportion of kissing couples who lean heads to the right is between 0.559 and 0.731.

3. Confidence is in the estimation procedure. Over many samples, 95% 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 null hypothesis is $H_0 : \theta = 0.5$. The alternative hypothesis is $H_a : \theta > 0.5$. The standard deviation of the null distribution is $\sqrt{0.5(1-0.5)/124} = 0.045$. The standardized statistic is $(0.645 - 0.5)/0.045 = 3.23$. Assuming the null distribution is approximately Normal, the p-value is approximately 0.0006.

5. With a p-value of 0.0006 we have strong evidence to reject the null hypothesis and conclude that the population proportion of kissing couples who lean heads to the right is greater than 0.5

6. Interpreting the p-value

   - If the population proportion of kissing couples who lean heads to the right is equal to 0.5
   - Then we would observe a sample proportion of 0.645 or more in about 0.06% of random samples of size 124
   - Since we actually observed a sample proportion of 0.645, which would be unlikely if the population proportion were 0.5
   - The data provide evidence that the population proportion is not 0.5

7. The numerical results are similar: the 95% posterior credible interval is similar to the 95% confidence interval, and the p-value (0.0006) is similar to the posterior probability that θ is less than 0.5 (0.0008 = 1-0.9992).

> However, the *interpretation* of these results is very different between the two approaches. The Bayesian approach provides probability statements about the parameter; the frequentist approach develops procedures based on the probability of what might 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. 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 how our prior "beliefs" about the parameters have been updated to reflect the observed data.

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.

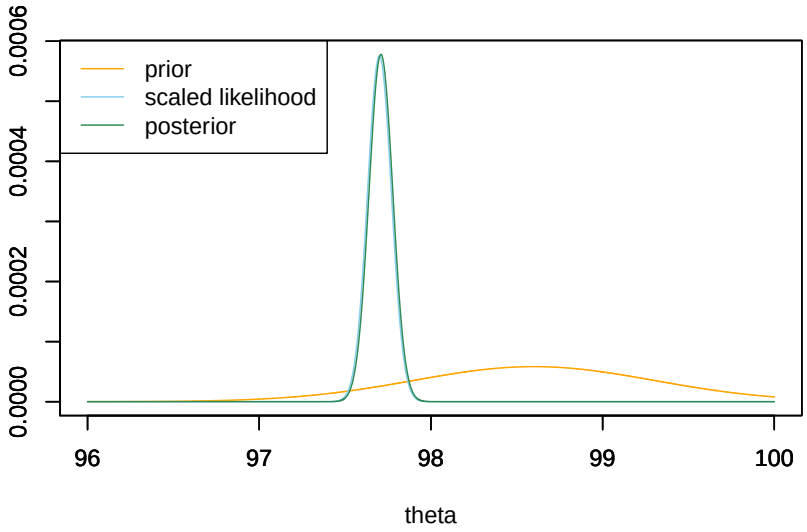
Example 5.5. Continuing Example 4.7. Assume body temperatures (degrees Fahrenheit) of healthy adults follow a Normal distribution with unknown mean μ and known standard deviation $\sigma = 1$. Suppose we wish to estimate μ, the population mean healthy human body temperature. In a recent study[1], the sample mean body temperature in a sample of 208 healthy adults was 97.7 degrees F.

We'll again use a grid approximation and assume that any multiple of 0.0001 between 96.0 and 100.0 is a possible value of μ: $96.0, 96.0001, 96.0002, \ldots, 99.9999, 100.0$. Assume a prior distribution which is proportional a Normal distribution with mean 98.6 and standard deviation 0.7 over μ values in the grid. Recall the prior, likelihood, and posterior.

```
# prior
theta = seq(96, 100, 0.0001)
prior = dnorm(theta, 98.6, 0.7)
```

[1] Source and a related article.

```
prior = prior / sum(prior)

# data
n = 208 # sample size
y = 97.7 # sample mean
sigma = 1

# likelihood
likelihood = dnorm(y, theta, sigma / sqrt(n)) # function of theta

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



1. What does the prior standard deviation of 0.7 represent?
2. What does the population standard deviation of 1 represent?
3. Compute the posterior standard deviation. What does it represent?
4. Compute the posterior mean.
5. Compute a 95% credible interval for μ.
6. Write a clearly worded sentence reporting the credible interval in context.
7. Compute the posterior probability that μ is less than 98.6.
8. Write a clearly worded sentence reporting the probability in the previous part in context.

Solution. to Example 5.5

Show/hide solution

1. The prior standard deviation of 0.7 quantifies, in a single number, our degree of prior uncertainty about the population mean human body temperature μ. We have a prior probability of 68% that μ is between 97.9 and 99.3, a prior probability of 95% that μ is between 97.2 and 100, etc (assuming a Normal prior).
2. The population standard deviation of 1 represents the person-to-person variability in body temperatures. If we were to measures body temperatures for many people, body temperatures would vary by about 1 degree F from person to person. About 68% of body temperatures would be within 1 degree of μ, about 95% would be within 2 degrees of μ, etc (assuming that individual body temperatures follows a Normal distributions.)
3. The posterior standard deviation of 0.069 (see code below) quantifies, in a single number, our degree of posterior uncertainty about the population mean human body temperature μ after observing the sample data.
4. The posterior mean is 97.71, which is pretty close to the observed sample mean.
5. Code gives [97.57, 97.85]. Since the posterior distribution is approximately Normal, we can approximate the endpoints of the confidence interval with $97.71 \pm 2 \times 0.069$.
6. There is a posterior probability of 95% that the population mean human body temperature is between 97.57 and 97.85 degrees Fahrenheit.
7. The posterior probability that μ is less than 98.6 is essentially 1. The value 98.6 is 12.9 standard deviations above the posterior mean: $(98.6 - 97.71)/0.069 = 12.9$.
8. There is a posterior probability of close to 100% that the population mean human body temperature is less than 98.6 degrees Fahrenheit.

Show/hide solution

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

```
## [1] 97.70874
```

```
# posterior variance
post_var = sum(theta ^ 2 * posterior) - post_ev ^ 2
post_var
```

```
## [1] 0.004760979
```

```r
# posterior sd
sqrt(post_var)
```

```
## [1] 0.06899985
```

```r
# posterior 95% credible interval
posterior_cdf = cumsum(posterior)
c(theta[max(which(posterior_cdf <= 0.025))], theta[min(which(posterior_cdf >= 0.975))])
```

```
## [1] 97.5734 97.8440
```

```r
# prior prob(theta < 98.6)
sum(posterior[theta < 98.6])
```

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