CHAPTER 2

Introduction: Credibility, Models, and Parameters

Contents

2.1. Bayesian Inference Is Reallocation of Credibility Across Possibilities 16
2.1.1 Data are noisy and inferences are probabilistic 19
2.2. Possibilities Are Parameter Values in Descriptive Models 22
2.3. The Steps of Bayesian Data Analysis 25
2.3.1 Data analysis without parametric models? 30
2.4. Exercises 31

I just want someone who I can believe in,
Someone at home who will not leave me grievin'.
Show me a sign that you'll always be true,
and I'll be your model of faith and virtue.[1]

The goal of this chapter is to introduce the conceptual framework of Bayesian data analysis. Bayesian data analysis has two foundational ideas. The first idea is that Bayesian inference is reallocation of credibility across possibilities. The second foundational idea is that the possibilities, over which we allocate credibility, are parameter values in meaningful mathematical models. These two fundamental ideas form the conceptual foundation for every analysis in this book. Simple examples of these ideas are presented in this chapter. The rest of the book *merely* fills in the mathematical and computational details for specific applications of these two ideas. This chapter also explains the basic procedural steps shared by every Bayesian analysis.

[1] This chapter introduces ideas of mathematical models, credibility of parameter values, and the semantics of models. The poem plays with the words "model," "believe," and "true" in an everyday context, and hints that Bayesian methods (personified) may be someone to believe in. (And yes, grammatically, the first line should be "in whom I can believe," but the poem is supposed to be colloquial speech. Besides, the grammatically correct version is iambic not dactylic!)

Doing Bayesian Data Analysis, 2nd Edition
http://dx.doi.org/10.1016/B978-0-12-405888-0.00002-7

2.1. BAYESIAN INFERENCE IS REALLOCATION OF CREDIBILITY ACROSS POSSIBILITIES

Suppose we step outside one morning and notice that the sidewalk is wet, and wonder why. We consider all possible causes of the wetness, including possibilities such as recent rain, recent garden irrigation, a newly erupted underground spring, a broken sewage pipe, a passerby who spilled a drink, and so on. If all we know until this point is that some part of the sidewalk is wet, then all those possibilities will have some prior credibility based on previous knowledge. For example, recent rain may have greater prior probability than a spilled drink from a passerby. Continuing on our outside journey, we look around and collect new observations. If we observe that the sidewalk is wet for as far as we can see, as are the trees and parked cars, then we re-allocate credibility to the hypothetical cause of recent rain. The other possible causes, such as a passerby spilling a drink, would not account for the new observations. On the other hand, if instead we observed that the wetness was localized to a small area, and there was an empty drink cup a few feet away, then we would re-allocate credibility to the spilled-drink hypothesis, even though it had relatively low prior probability. This sort of reallocation of credibility across possibilities is the essence of Bayesian inference.

Another example of Bayesian inference has been immortalized in the words of the fictional detective Sherlock Holmes, who often said to his sidekick, Doctor Watson: "How often have I said to you that when you have eliminated the impossible, whatever remains, however improbable, must be the truth?" (Doyle, 1890, chap. 6) Although this reasoning was not described by Holmes or Watson or Doyle as Bayesian inference, it is. Holmes conceived of a set of possible causes for a crime. Some of the possibilities may have seemed very improbable, *a priori*. Holmes systematically gathered evidence that ruled out a number of the possible causes. If all possible causes but one were eliminated, then (Bayesian) reasoning forced him to conclude that the remaining possible cause was fully credible, even if it seemed improbable at the start.

Figure 2.1 illustrates Holmes' reasoning. For the purposes of illustration, we suppose that there are just four possible causes of the outcome to be explained. We label the causes A, B, C, and D. The heights of the bars in the graphs indicate the credibility of the candidate causes. ("Credibility" is synonymous with "probability"; here I use the everyday term "credibility" but later in the book, when mathematical formalisms are introduced, I will also use the term "probability.") Credibility can range from zero to one. If the credibility of a candidate cause is zero, then the cause is definitely not responsible. If the credibility of a candidate cause is one, then the cause definitely *is* responsible. Because we assume that the candidate causes are mutually exclusive and exhaust all possible causes, the total credibility across causes sums to one.

The upper-left panel of Figure 2.1 shows that the prior credibilities of the four candidate causes are equal, all at 0.25. Unlike the case of the wet sidewalk, in which

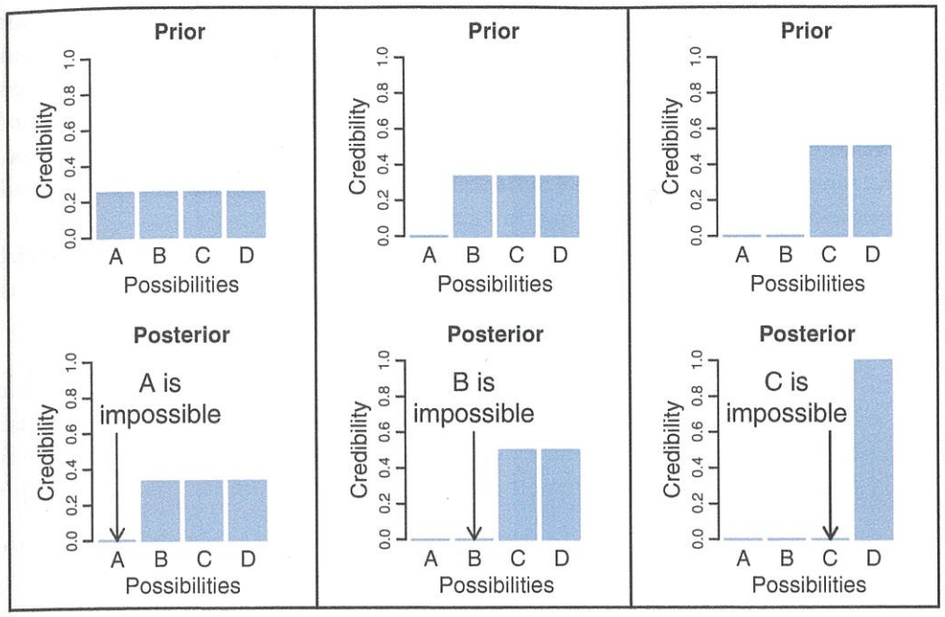

Figure 2.1 The upper-left graph shows the credibilities of the four possible causes for an outcome. The causes, labeled A, B, C, and D, are mutually exclusive and exhaust all possibilities. The causes happen to be equally credible at the outset; hence all have prior credibility of 0.25. The lower-left graph shows the credibilities when one cause is learned to be impossible. The resulting posterior distribution is used as the prior distribution in the middle column, where another cause is learned to be impossible. The posterior distribution from the middle column is used as the prior distribution for the right column. The remaining possible cause is fully implicated by Bayesian reallocation of credibility.

prior knowledge suggested that rain may be a more likely cause than a newly erupted underground spring, the present illustration assumes equal prior credibilities of the candidate causes. Suppose we make new observations that rule out candidate cause A. For example, if A is a suspect in a crime, we may learn that A was far from the crime scene at the time. Therefore, we must re-allocate credibility to the remaining candidate causes, B through D, as shown in the lower-left panel of Figure 2.1. The re-allocated distribution of credibility is called the *posterior distribution* because it is what we believe after taking into account the new observations. The posterior distribution gives zero credibility to cause A, and allocates credibilities of 0.33 (i.e., 1/3) to candidate causes B, C, and D.

The posterior distribution then becomes the prior beliefs for subsequent observations. Thus, the prior distribution in the upper-middle of Figure 2.1 is the posterior distribution from the lower left. Suppose now that additional new evidence rules out candidate cause B. We now must re-allocate credibility to the remaining candidate

causes, C and D, as shown in the lower-middle panel of Figure 2.1. This posterior distribution becomes the prior distribution for subsequent data collection, as shown in the upper-right panel of Figure 2.1. Finally, if new data rule out candidate cause C, then all credibility must fall on the remaining cause, D, as shown in the lower-right panel of Figure 2.1, just as Holmes declared. This reallocation of credibility is not only intuitive, it is also what the exact mathematics of Bayesian inference prescribe, as will be explained later in the book.

The complementary form of reasoning is also Bayesian, and can be called judicial *exoneration*. Suppose there are several possible culprits for a crime, and that these suspects are mutually unaffiliated and exhaust all possibilities. If evidence accrues that one suspect is definitely culpable, then the other suspects are exonerated.

This form of exoneration is illustrated in Figure 2.2. The upper panel assumes that there are four possible causes for an outcome, labeled A, B, C, and D. We assume that the causes are mutually exclusive and exhaust all possibilities. In the context of suspects for a crime, the credibility of the hypothesis that suspect A committed the crime is the

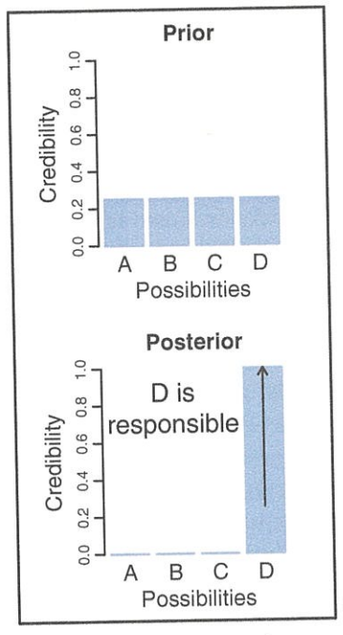

Figure 2.2 The upper graph shows the credibilities of the four possible causes for an outcome. The causes, labeled A, B, C and D, are mutually exclusive and exhaust all possibilities. The causes happen to be equally credible at the outset, hence all have prior credibility of 0.25. The lower graph shows the credibilities when one cause is learned to be responsible. The nonresponsible causes are "exonerated" (i.e., have zero credibility as causes) by Bayesian reallocation of credibility.

culpability of the suspect. So it might be easier in this context to think of culpability instead of credibility. The prior culpabilities of the four suspects are, for this illustration, set to be equal, so the four bars in the upper panel of Figure 2.2 are all of height 0.25. Suppose that new evidence firmly implicates suspect D as the culprit. Because the other suspects are known to be unaffiliated, they are exonerated, as shown in the lower panel of Figure 2.2. As in the situation of Holmesian deduction, this exoneration is not only intuitive, it is also what the exact mathematics of Bayesian inference prescribe, as will be explained later in the book.

2.1.1. Data are noisy and inferences are probabilistic

The cases of Figures 2.1 and 2.2 assumed that observed data had definitive, deterministic relations to the candidate causes. For example, the fictional Sherlock Holmes may have found a footprint at the scene of the crime and identified the size and type of shoe with complete certainty, thereby completely ruling out or implicating a particular candidate suspect. In reality, of course, data have only probabilistic relations to their underlying causes. A real detective might carefully measure the footprint and the details of its tread, but these measurements would only probabilistically narrow down the range of possible shoes that might have produced the print. The measurements are not perfect, and the footprint is only an imperfect representation of the shoe that produced it. The relation between the cause (i.e., the shoe) and the measured effect (i.e., the footprint) is full of random variation.

In scientific research, measurements are replete with randomness. Extraneous influences contaminate the measurements despite tremendous efforts to limit their intrusion. For example, suppose we are interested in testing whether a new drug reduces blood pressure in humans. We randomly assign some people to a test group that takes the drug, and we randomly assign some other people to a control group that takes a placebo. The procedure is "double blind" so that neither the participants nor the administrators know which person received the drug or the placebo (because that information is indicated by a randomly assigned code that is decrypted after the data are collected). We measure the participants' blood pressures at set times each day for several days. As you can imagine, blood pressures for any single person can vary wildly depending on many influences, such as exercise, stress, recently eaten foods, etc. The measurement of blood pressure is itself an uncertain process, as it depends on detecting the sound of blood flow under a pressurized sleeve. Blood pressures are also very different from one person to the next. The resulting data, therefore, are extremely messy, with tremendous variability within each group, and tremendous overlap across groups. Thus, there will be many measured blood pressures in the drug group that are higher than blood pressures in the placebo group, and vice versa. From these two dispersed and overlapping heaps of numbers, we want to infer how big a difference there is between the groups, and how certain we can

be about that difference. The problem is that for any particular real difference between the drug and the placebo, its measurable effect is only a random impression.

All scientific data have some degree of "noise" in their values. The techniques of data analysis are designed to infer underlying trends from noisy data. Unlike Sherlock Holmes, who could make an observation and completely rule out some possible causes, we can collect data and only incrementally adjust the credibility of some possible trends. We will see many realistic examples later in the book. The beauty of Bayesian analysis is that the mathematics reveal exactly how much to re-allocate credibility in realistic probabilistic situations.

Here is a simplified illustration of Bayesian inference when data are noisy. Suppose there is a manufacturer of inflated bouncy balls, and the balls are produced in four discrete sizes, namely diameters of 1.0, 2.0, 3.0, and 4.0 (on some scale of distance such as decimeters). The manufacturing process is quite variable, however, because of randomness in degrees of inflation even for a single size ball. Thus, balls of manufactured size 3 might have diameters of 1.8 or 4.2, even though their average diameter is 3.0. Suppose we submit an order to the factory for three balls of size 2. We receive three balls and measure their diameters as best we can, and find that the three balls have diameters of 1.77, 2.23, and 2.70. From those measurements, can we conclude that the factory correctly sent us three balls of size 2, or did the factory send size 3 or size 1 by mistake, or even size 4?

Figure 2.3 shows a Bayesian answer to this question. The upper graph shows the four possible sizes, with blue bars at positions 1, 2, 3, and 4. The prior credibilities of the four sizes are set equal, at heights of 0.25, representing the idea that the factory received the order for three balls, but may have completely lost track of which size was ordered, hence any size is equally possible to have been sent.

At this point, we must specify the form of random variability in ball diameters. For purposes of illustration, we will suppose that ball diameters are centered on their manufactured size, but could be bigger or smaller depending on the amount of inflation. The bell-shaped curves in Figure 2.3 indicate the probability of diameters produced by each size. Thus, the bell-shaped curve centered on size 2 indicates that size-2 balls are usually about 2.0 units in diameter, but could be much bigger or smaller because of randomness in inflation. The horizontal axis in Figure 2.3 is playing double duty as a scale for the ball sizes (i.e., blue bars) and for the measured diameters (suggested by the bell-shaped distributions).

The lower panel of Figure 2.3 shows the three measured diameters plotted as circles on the horizontal axis. You can see that the measured diameters are closest to sizes 2 or 3, but the bell-shaped distributions reveal that even size 1 could sometimes produce balls of those diameters. Intuitively, therefore, we would say that size 2 is most credible, given the data, but size 3 is also somewhat possible, and size 1 is remotely possible, but size 4 is rather unlikely. These intuitions are precisely reflected by Bayesian analysis,

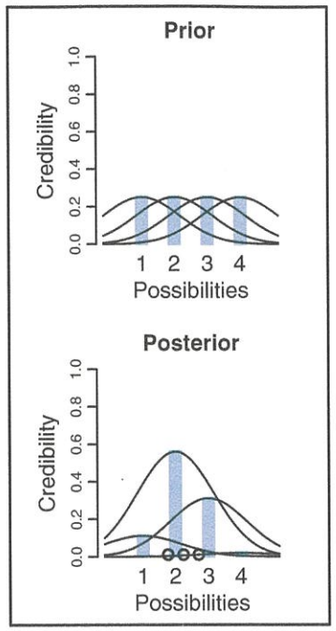

Figure 2.3 The upper graph shows the prior credibilities of the four candidate means in normal distributions, located at values of 1, 2, 3, and 4. Superimposed on the means are the corresponding normal distributions. The horizontal axis is playing double duty as a scale for the means (marked by the blue bars) and for the data (suggested by the normal distributions). The three observed data values are plotted as circles on the floor of the lower panel. Bayesian reallocation of credibility across the four candidate means indicates that the mean at 2 is most credible given the data, the mean at 3 is somewhat credible, and so on.

which is shown in the lower panel of Figure 2.3. The heights of the blue bars show the exact reallocation of credibility across the four candidate sizes. Given the data, there is 56% probability that the balls are size 2, 31% probability that the balls are size 3, 11% probability that the balls are size 1, and only 2% probability that the balls are size 4.

Inferring the underlying manufactured size of the balls from their "noisy" individual diameters is analogous to data analysis in real-world scientific research and applications. The data are noisy indicators of the underlying generator. We hypothesize a range of possible underlying generators, and from the data we infer their relative credibilities.

As another example, consider testing people for illicit drug use. A person is taken at random from a population and given a blood test for an illegal drug. From the result of the test, we infer whether or not the person has used the drug. But, crucially, the test is not perfect, it is noisy. The test has a non-trivial probability of producing false positives and false negatives. And we must also take into account our prior knowledge

that the drug is used by only a small proportion of the population. Thus, the set of possibilities has two values: The person uses the drug or does not. The two possibilities have prior credibilities based on previous knowledge of how prevalent drug use is in the population. The noisy datum is the result of the drug test. We then use Bayesian inference to re-allocate credibility across the possibilities. As we will see quantitatively later in the book, the posterior probability of drug use is often surprisingly small even when the test result is positive, because the prior probability of drug use is small and the test is noisy. This is true not only for tests of drug use, but also for tests of diseases such as cancer. A related real-world application of Bayesian inference is detection of spam in email. Automated spam filters often use Bayesian inference to compute a posterior probability that an incoming message is spam.

In summary, the essence of Bayesian inference is reallocation of credibility across possibilities. The distribution of credibility initially reflects prior knowledge about the possibilities, which can be quite vague. Then new data are observed, and the credibility is re-allocated. Possibilities that are consistent with the data garner more credibility, while possibilities that are not consistent with the data lose credibility. Bayesian analysis is the mathematics of re-allocating credibility in a logically coherent and precise way.

2.2. POSSIBILITIES ARE PARAMETER VALUES IN DESCRIPTIVE MODELS

A key step in Bayesian analysis is defining the set of possibilities over which credibility is allocated. This is not a trivial step, because there might always be possibilities beyond the ones we include in the initial set. (For example, the wetness on the sidewalk might have been caused by space aliens who were crying big tears.) But we get the process going by choosing a set of possibilities that covers a range in which we are interested. After the analysis, we can examine whether the data are well described by the most credible possibilities in the considered set. If the data seem not to be well described, then we can consider expanding the set of possibilities. This process is called a posterior predictive check and will be explained later in the book.

Consider again the example of the blood-pressure drug, in which blood pressures are measured in one group that took the drug and in another group that took a placebo. We want to know how much difference there is in the tendencies of the two groups: How big is the difference between the typical blood pressure in one group versus the typical blood pressure in the other group, and how certain can we be of the difference? The magnitude of difference *describes* the data, and *our goal is to assess which possible descriptions are more or less credible.*

In general, data analysis begins with a family of candidate descriptions for the data. The descriptions are mathematical formulas that characterize the trends and spreads in the data. The formulas themselves have numbers, called parameter values, that determine the exact shape of mathematical forms. You can think of *parameters as control knobs on*

mathematical devices that simulate data generation. If you change the value of a parameter, it changes a trend in the simulated data, just like if you change the volume control on a music player, it changes the intensity of the sound coming out of the player.

In previous studies of statistics or mathematics, you may have encountered the so-called normal distribution, which is a bell-shaped distribution often used to describe data. It was alluded to above in the example of the inflated bouncy balls (see Figure 2.3). The normal distribution has two parameters, called the mean and standard deviation. The mean is a control knob in the mathematical formula for the normal distribution that controls the location of the distribution's central tendency. The mean is sometimes called a *location parameter.* The standard deviation is another control knob in the mathematical formula for the normal distribution that controls the width or dispersion of the distribution. The standard deviation is sometimes called a *scale parameter.* The mathematical formula for the normal distribution converts the parameter values to a particular bell-like shape for the probabilities of data values.

Figure 2.4 shows some data with candidate normal distributions superimposed. The data are shown as a histogram, which plots vertical bars that have heights indicating

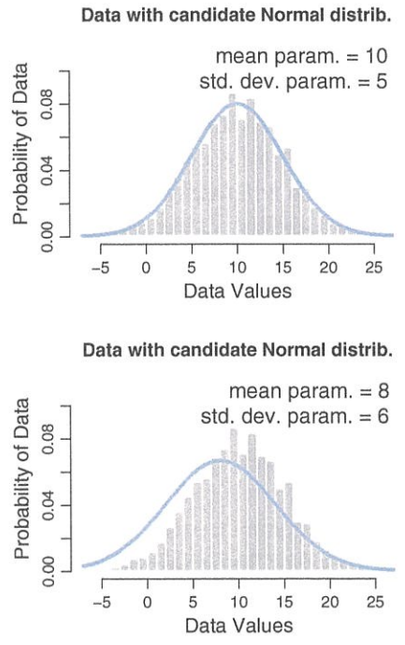

Figure 2.4 The two graphs show the same data histogram but with two different candidate descriptions by normal distributions. Bayesian analysis computes the relative credibilities of candidate parameter values.

how much of the data falls within the small range spanned by the bar. The histogram appears to be roughly unimodal and left-right symmetric. The upper panel superimposes a candidate description of the data in the form of a normal distribution that has a mean of 10 and a standard deviation of 5. This choice of parameter values appears to describe the data well. The lower panel shows another choice of parameter values, with a mean of 8 and a standard deviation of 6. While this candidate description appears to be plausible, it is not as good as the upper panel. The role of Bayesian inference is to compute the exact relative credibilities of candidate parameter values, while also taking into account their prior probabilities.

In realistic applications, the candidate parameter values can form an infinite continuum, not only a few discrete options. The location parameter of the normal distribution can take on any value from negative to positive infinity. Bayesian inference operates without trouble on infinite continuums.

There are two main desiderata for a mathematical description of data. First, the mathematical form should be comprehensible with meaningful parameters. Just as it would be fruitless to describe the data in a language that we do not know, it would be fruitless to describe the data with a mathematical form that we do not understand, with parameters that we cannot interpret. In the case of a normal distribution, for example, the mean parameter and standard-deviation parameter are directly meaningful as the location and scale of the distribution. Throughout this book, we will use mathematical descriptions that have meaningful parameters. Thus, Bayesian analysis reallocates credibility among parameter values within a meaningful space of possibilities defined by the chosen model.

The second desideratum for a mathematical description is that it should be descriptively adequate, which means, loosely, that the mathematical form should "look like" the data. There should not be any important systematic discrepancies between trends in the data and the form of the model. Deciding whether or not an apparent discrepancy is important or systematic is not a definite process. In early stages of research, we might be satisfied with a rough, "good enough" description of data, because it captures meaningful trends that are interesting and novel relative to previous knowledge. As the field of research matures, we might demand more and more accurate descriptions of data. Bayesian analysis is very useful for assessing the relative credibility of different candidate descriptions of data.

It is also important to understand that mathematical descriptions of data are not necessarily causal explanations of data. To say that the data in Figure 2.4 are well described by a normal distribution with mean of 10 and standard deviation of 5 does not explain what process caused the data to have that form. The parameters are "meaningful" only in the context of the familiar mathematical form defined by the normal distribution; the parameter values have no necessary meaning with respect to causes in the world. In some applications, the mathematical model might be motivated as a description of a natural

process that generated the data, and thereby the parameters and mathematical form can refer to posited states and processes in the world. For example, in the case of the inflated bouncy balls (Figure 2.3), the candidate parameter values were interpreted as "sizes" at the manufacturer, and the underlying size, combined with random inflation, caused the observed data value. But reference to physical states or processes is not necessary for merely describing the trends in a sample of data. In this book, we will be focusing on generic data description using intuitively accessible model forms that are broadly applicable across many domains.

2.3. THE STEPS OF BAYESIAN DATA ANALYSIS

In general, Bayesian analysis of data follows these steps:

1. Identify the data relevant to the research questions. What are the measurement scales of the data? Which data variables are to be predicted, and which data variables are supposed to act as predictors?
2. Define a descriptive model for the relevant data. The mathematical form and its parameters should be meaningful and appropriate to the theoretical purposes of the analysis.
3. Specify a prior distribution on the parameters. The prior must pass muster with the audience of the analysis, such as skeptical scientists.
4. Use Bayesian inference to re-allocate credibility across parameter values. Interpret the posterior distribution with respect to theoretically meaningful issues (assuming that the model is a reasonable description of the data; see next step).
5. Check that the posterior predictions mimic the data with reasonable accuracy (i.e., conduct a "posterior predictive check"). If not, then consider a different descriptive model.

Perhaps the best way to explain these steps is with a realistic example of Bayesian data analysis. The discussion that follows is abbreviated for purposes of this introductory chapter, with many technical details suppressed. For this example, suppose we are interested in the relationship between weight and height of people. We suspect from everyday experience that taller people tend to weigh more than shorter people, but we would like to know by how much people's weights tend to increase when height increases, and how certain we can be about the magnitude of the increase. In particular, we might be interested in predicting a person's weight based on their height.

The first step is identifying the relevant data. Suppose we have been able to collect heights and weights from 57 mature adults sampled at random from a population of interest. Heights are measured on the continuous scale of inches, and weights are measured on the continuous scale of pounds. We wish to predict weight from height. A scatter plot of the data is shown in Figure 2.5.

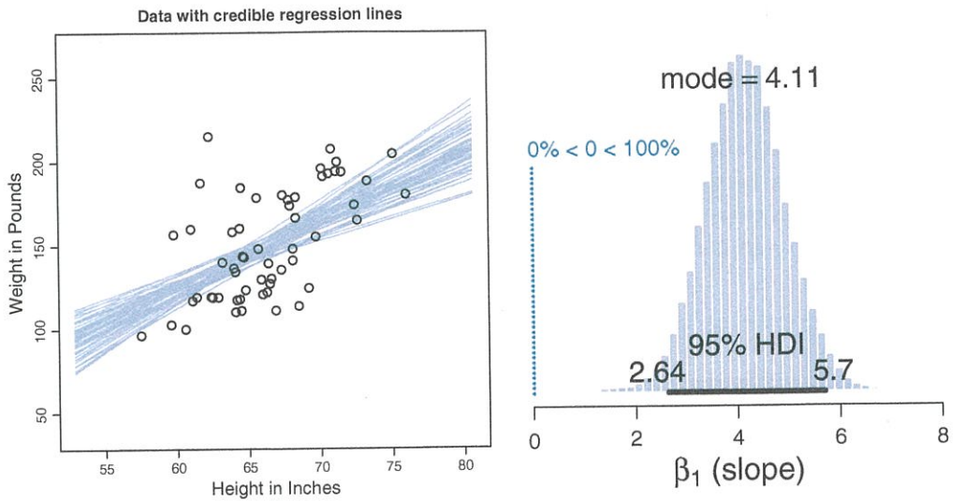

Figure 2.5 Data are plotted as circles in the scatter plot of the left panel. The left panel also shows a smattering of credible regression lines from the posterior distribution superimposed on the data. The right panel shows the posterior distribution of the slope parameter (i.e., β_1 in Equation 2.1).

The second step is to define a descriptive model of the data that is meaningful for our research interest. At this point, we are interested merely in identifying a basic trend between weight and height, and it is not absurd to think that weight might be proportional to height, at least as an approximation over the range of adult weights and heights. Therefore, we will describe predicted weight as a multiplier times height plus a baseline. We will denote the predicted weight as $\hat{y}$ (spoken "y hat"), and we will denote the height as x. Then the idea that predicted weight is a multiple of height plus a baseline can be denoted mathematically as follows:

$$\hat{y} = \beta_1 x + \beta_0 \tag{2.1}$$

The coefficient, β_1 (Greek letter "beta"), indicates how much the predicted weight increases when the height goes up by one inch.[2] The baseline is denoted β_0 in Equation 2.1, and its value represents the weight of a person who is zero inches tall. You might suppose that the baseline value should be zero, *a priori*, but this need not be the case for describing the relation between weight and height of mature adults, who have a limited range of height values far above zero. Equation 2.1 is the form of a line,

[2] Here is a proof that β_1 indicates how much that $\hat{y}$ goes up when x increases by 1 unit. First, at height x, the predicted weight is $\hat{y}_x = \beta_1 x + \beta_0$. Second, at height $x + 1$, the predicted weight is $\hat{y}_{x+1} = \beta_1(x + 1) + \beta_0 = \beta_1 x + \beta_1 + \beta_0$. Therefore, the change in predicted weight is $\hat{y}_{x+1} - \hat{y}_x = \beta_1$.

in which β_1 is the slope and β_0 is the intercept, and this model of trend is often called linear regression.

The model is not complete yet, because we have to describe the random variation of actual weights around the predicted weight. For simplicity, we will use the conventional normal distribution (explained in detail in Section 4.3.2.2), and assume that actual weights y are distributed randomly according to a normal distribution around the predicted value $\hat{y}$ and with standard deviation denoted σ (Greek letter "sigma"). This relation is denoted symbolically as

$$y \sim \text{normal}(\hat{y}, \sigma) \tag{2.2}$$

where the symbol "$\sim$" means "is distributed as." Equation 2.2 is saying that y values near $\hat{y}$ are most probable, and y values higher or lower than $\hat{y}$ are less probable. The decrease in probability around $\hat{y}$ is governed by the shape of the normal distribution with width specified by the standard deviation σ.

The full model, combining Equations 2.1 and 2.2, has three parameters altogether: the slope, β_1, the intercept, β_0, and the standard deviation of the "noise," σ. Note that the three parameters are meaningful. In particular, the slope parameter tells us how much the weight tends to increase when height increases by an inch, and the standard deviation parameter tells us how much variability in weight there is around the predicted value. This sort of model, called linear regression, is explained at length in Chapters 15, 17, and 18.

The third step in the analysis is specifying a prior distribution on the parameters. We might be able to inform the prior with previously conducted, and publicly verifiable, research on weights and heights of the target population. Or we might be able to argue for a modestly informed prior based on consensual experience of social interactions. But for purposes of this example, I will use a noncommittal and vague prior that places virtually equal prior credibility across a vast range of possible values for the slope and intercept, both centered at zero. I will also place a vague and noncommittal prior on the noise (standard deviation) parameter, specifically a uniform distribution that extends from zero to a huge value. This choice of prior distribution implies that it has virtually no biasing influence on the resulting posterior distribution.

The fourth step is interpreting the posterior distribution. Bayesian inference has re-allocated credibility across parameter values, from the vague prior distribution, to values that are consistent with the data. The posterior distribution indicates combinations of β_0, β_1, and σ that together are credible, given the data. The right panel of Figure 2.5 shows the posterior distribution on the slope parameter, β_1 (collapsing across the other two parameters). It is important to understand that Figure 2.5 shows a distribution of parameter values, not a distribution of data. The blue bars of Figure 2.5 indicate the credibility across the *continuum* of candidate slope values, analogous to the blue

bars in the examples of Sherlock Holmes, exoneration, and discrete candidate means (in Figures 2.1–2.3). The posterior distribution in Figure 2.5 indicates that the most credible value of the slope is about 4.1, which means that weight increases about 4.1 pounds for every 1-inch increase in height. The posterior distribution also shows the uncertainty in that estimated slope, because the distribution shows the relative credibility of values across the continuum. One way to summarize the uncertainty is by marking the span of values that are most credible and cover 95% of the distribution. This is called the *highest density interval* (HDI) and is marked by the black bar on the floor of the distribution in Figure 2.5. Values within the 95% HDI are more credible (i.e., have higher probability "density") than values outside the HDI, and the values inside the HDI have a total probability of 95%. Given the 57 data points, the 95% HDI goes from a slope of about 2.6 pounds per inch to a slope of about 5.7 pounds per inch. With more data, the estimate of the slope would be more precise, meaning that the HDI would be narrower.

Figure 2.5 also shows where a slope of zero falls relative to the posterior distribution. In this case, zero falls far outside any credible value for the slope, and therefore we could decide to "reject" zero slope as a plausible description of the relation between height and weight. But this discrete decision about the status of zero is separate from the Bayesian analysis *per se*, which provides the complete posterior distribution.

Many readers may have previously learned about null hypothesis significance testing (NHST) which involves *sampling distributions* of summary statistics such as t, from which are computed p values. (If you do not know these terms, do not worry. NHST will be discussed in Chapter 11.) It is important to understand that the posterior distribution in Figure 2.5 is *not* a sampling distribution and has nothing to do with p values.

Another useful way of understanding the posterior distribution is by plotting examples of credible regression lines through the scatter plot of the data. The left panel of Figure 2.5 shows a random smattering of credible regression lines from the posterior distribution. Each line plots $\hat{y} = \beta_1 x + \beta_0$ for credible combinations of β_1 and β_0. The bundle of lines shows a range of credible possibilities, given the data, instead of plotting only a single "best" line.

The fifth step is to check that the model, with its most credible parameter values, actually mimics the data reasonably well. This is called a "posterior predictive check." There is no single, unique way to ascertain whether the model predictions systematically and meaningfully deviate from the data, because there are innumerable ways in which to define systematic deviation. One approach is to plot a summary of predicted data from the model against the actual data. We take credible values of the parameters, β_1, β_0, and σ, plug them into the model Equations 2.1 and 2.2, and randomly generate simulated y values (weights) at selected x values (heights). We do that for many, many

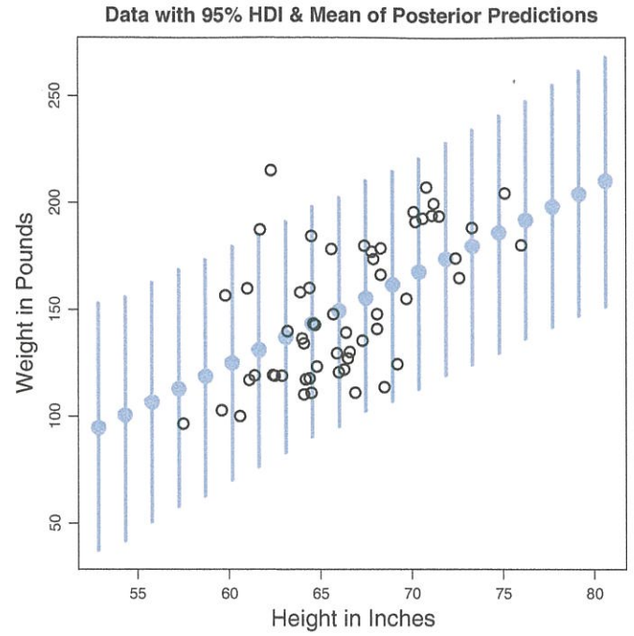

Figure 2.6 The data of Figure 2.5 are shown with posterior predicted weight values superimposed at selected height values. Each vertical bar shows the range of the 95% most credible predicted weight values, and the dot at the middle of each bar shows the mean predicted weight value.

credible parameter values to create representative distributions of what data would look like according to the model. The results of this simulation are shown in Figure 2.6. The predicted weight values are summarized by vertical bars that show the range of the 95% most credible predicted weight values. The dot at the middle of each bar shows the mean of the predicted weight values. By visual inspection of the graph, we can see that the actual data appear to be well described by the predicted data. The actual data do not appear to deviate systematically from the trend or band predicted from the model.

If the actual data did appear to deviate systematically from the predicted form, then we could contemplate alternative descriptive models. For example, the actual data might appear to have a nonlinear trend. In that case, we could expand the model to include nonlinear trends. It is straightforward to do this in Bayesian software, and easy to estimate the parameters that describe nonlinear trends. We could also examine the distributional properties of the data. For example, if the data appear to have outliers relative to what is predicted by a normal distribution, we could change the model to use a heavy-tailed distribution, which again is straightforward in Bayesian software.

We have seen the five steps of Bayesian analysis in a fairly realistic example. This book explains how to do this sort of analysis for many different applications and types of descriptive models. For a shorter but detailed introduction to Bayesian analysis for comparing two groups, with explanation of the perils of the classical t test, see the article by Kruschke (2013a). For an introduction to Bayesian analysis applied to multiple linear regression, see the article by Kruschke, Aguinis, and Joo (2012). For a perspective on posterior predictive checks, see the article by Kruschke (2013b) and Section 17.5.1 (among others) of this book.

2.3.1. Data analysis without parametric models?

As outlined above, Bayesian data analysis is based on meaningfully parameterized descriptive models. Are there ever situations in which such models cannot be used or are not wanted?

One situation in which it might appear that parameterized models are not used is with so-called *non*parametric models. But these models are confusingly named because they actually do have parameters; in fact they have a potentially infinite number of parameters. As a simple example, suppose we want to describe the weights of dogs. We measure the weights of many different dogs sampled at random from the entire spectrum of dog breeds. The weights are probably not distributed unimodally, instead there are probably subclusters of weights for different breeds of dogs. But some different breeds might have nearly identical distributions of weights, and there are many dogs that cannot be identified as a particular breed, and, as we gather data from more and more dogs, we might encounter members of new subclusters that had not yet been included in the previously collected data. Thus, it is not clear how many clusters we should include in the descriptive model. Instead, we infer, from the data, the relative credibilities of different clusterings. Because each cluster has its own parameters (such as location and scale parameters), the number of parameters in the model is inferred, and can grow to infinity with infinite data. There are many other kinds of infinitely parameterized models. For a tutorial on Bayesian nonparametric models, see Gershman and Blei (2012); for a recent review, see Müller and Mitra (2013); and for textbook applications, see Gelman et al. (2013). We will not be considering Bayesian nonparametric models in this book.

There are a variety of situations in which it might seem at first that no parameterized model would apply, such as figuring out the probability that a person has some rare disease if a diagnostic test for the disease is positive. But Bayesian analysis does apply even here, although the parameters refer to discrete states instead of continuous distributions. In the case of disease diagnosis, the parameter is the underlying health status of the individual, and the parameter can have one of two values, either "has disease" or "does not have disease." Bayesian analysis re-allocates credibility over those two parameter values based on the observed test result. This is exactly analogous to the discrete possibilities considered by Sherlock Holmes in Figure 2.1, except that the test results yield probabilistic information instead of perfectly conclusive information. We will do exact Bayesian computations for this sort of situation in Chapter 5 (see specifically Table 5.4).

Finally, there might be some situations in which the analyst is loathe to commit to any parameterized model of the data, even tremendously flexible infinitely parameterized models. If this is the case, then Bayesian methods cannot apply. These situations are rare, however, because mathematical models are enormously useful tools. One case of trying to make inferences from data without using a model is a method from NHST called *resampling* or *bootstrapping*. These methods compute p values to make decisions, and p values have fundamental logical problems that will be discussed in Chapter 11. These methods also have very limited ability to express degrees of certainty about characteristics of the data, whereas Bayesian methods put expression of uncertainty front and center.

2.4. EXERCISES

Look for more exercises at https://sites.google.com/site/doingbayesiandataanalysis/

Exercise 2.1. [Purpose: To get you actively manipulating mathematical models of probabilities.] Suppose we have a four-sided die from a board game. On a tetrahedral die, each face is an equilateral triangle. When you roll the die, it lands with one face down and the other three faces visible as a three-sided pyramid. The faces are numbered 1-4, with the value of the bottom face printed (as clustered dots) at the bottom edges of all three visible faces. Denote the value of the bottom face as x. Consider the following three mathematical descriptions of the probabilities of x. Model A: $p(x) = 1/4$. Model B: $p(x) = x/10$. Model C: $p(x) = 12/(25x)$. For each model, determine the value of $p(x)$ for each value of x. Describe in words what kind of bias (or lack of bias) is expressed by each model.

Exercise 2.2. [Purpose: To get you actively thinking about how data cause credibilities to shift.] Suppose we have the tetrahedral die introduced in the previous exercise, along with the three candidate models of the die's probabilities. Suppose that initially, we are not sure what to believe about the die. On the one hand, the die might be fair, with each face landing with the same probability. On the other hand, the die might be biased, with the faces that have more dots landing down more often (because the dots are created by embedding heavy jewels in the die, so that the sides with more dots are more likely to land on the bottom). On yet another hand, the die might be

been developed for different applications. I would list various web resources here, but aside from the main sites for R and RStudio, other sites are continuously evolving and changing. Therefore, search the web for the latest packages and documentation. There are also numerous books about R and specific applications of R.

3.10. EXERCISES

Look for more exercises at https://sites.google.com/site/doingbayesiandataanalysis/

Exercise 3.1. [Purpose: Actually doing Bayesian statistics, eventually, and the next exercises, immediately.] Install R on your computer. (And if that's not exercise, I don't know what is.)

Exercise 3.2. [Purpose: Being able to record and communicate the results of your analyses.] Open the program `ExampleOfR.R`. At the end, notice the section that produces a simple graph. Your job is to save the graph so you can incorporate it into documents in the future, as you would for reporting actual data analyses. Save the graph in a format that is compatible with your word processing software. Import the saved file into your document and explain, in text, what you did. (Notice that for some word processing systems you could merely copy and paste directly from R's graphic window to the document window. But the problem with this approach is that you have no other record of the graph produced by the analysis. We want the graph to be saved separately so that it can be incorporated into various reports at a future time.)

Exercise 3.3. [Purpose: Getting experience with the details of the command syntax within R.] Adapt the program `SimpleGraph.R` so that it plots a cubic function ($y = x^3$) over the interval $x \in [-3, +3]$. Save the graph in a file format of your choice. Include a properly commented listing of your code, along with the resulting graph.

CHAPTER 4

What Is This Stuff Called Probability?

Contents

4.1. The Set of All Possible Events . . . 72
4.1.1 Coin flips: Why you should care . . . 73
4.2. Probability: Outside or Inside the Head . . . 73
4.2.1 Outside the head: Long-run relative frequency . . . 74
4.2.1.1 Simulating a long-run relative frequency . . . 74
4.2.1.2 Deriving a long-run relative frequency . . . 76
4.2.2 Inside the head: Subjective belief . . . 76
4.2.2.1 Calibrating a subjective belief by preferences . . . 76
4.2.2.2 Describing a subjective belief mathematically . . . 77
4.2.3 Probabilities assign numbers to possibilities . . . 77
4.3. Probability Distributions . . . 78
4.3.1 Discrete distributions: Probability mass . . . 78
4.3.2 Continuous distributions: Rendezvous with density . . . 80
4.3.2.1 Properties of probability density functions . . . 82
4.3.2.2 The normal probability density function . . . 83
4.3.3 Mean and variance of a distribution . . . 84
4.3.3.1 Mean as minimized variance . . . 86
4.3.4 Highest density interval (HDI) . . . 87
4.4. Two-Way Distributions . . . 89
4.4.1 Conditional probability . . . 91
4.4.2 Independence of attributes . . . 92
4.5. Appendix: R Code for Figure 4.1 . . . 93
4.6. Exercises . . . 95

Oh darlin' you change from one day to the next,
I'm feelin' deranged and just plain ol' perplexed.
I've learned to put up with your raves and your rants:
The mean I can handle but not variance.[1]

Inferential statistical techniques assign precise measures to our uncertainty about possibilities. Uncertainty is measured in terms of *probability*, and therefore we must establish the properties of probability before we can make inferences about it. This chapter introduces the basic ideas of probability. If this chapter seems too abbreviated for you, an excellent beginner's introduction to the topics of this chapter has been written by Albert and Rossman (2001, pp. 227-320).

[1] This chapter discusses ideas of probability distributions. Among those ideas are the technical definitions of the *mean* and *variance* of a distribution. The poem plays with colloquial meanings of those words.

Doing Bayesian Data Analysis, 2nd Edition
http://dx.doi.org/10.1016/B978-0-12-405888-0.00004-0

4.1. THE SET OF ALL POSSIBLE EVENTS

Suppose I have a coin that I am going to flip. How likely is it to come up a head? How likely is it to come up a tail?[2] How likely is it to come up a torso? Notice that when we contemplate the likelihood of each outcome, we have in mind a set of all possible outcomes. Torso is not one of the possible outcomes. Notice also that a single flip of a coin can result in only one outcome; it cannot be both heads and tails in a single flip. The outcomes are mutually exclusive.

Whenever we ask about how likely an outcome is, we always ask with a set of possible outcomes in mind. This set exhausts all possible outcomes, and the outcomes are all mutually exclusive. This set is called the *sample space*. The sample space is determined by the measurement operation we use to make an observation of the world. In all of our applications throughout the book, we take it for granted that there is a well-defined operation for making a measurement. For example, in flipping a coin, we take it for granted that there is a well-defined way to launch the coin and catch it, so that we can decide exactly when the coin has stopped its motion and is stable enough to be declared one outcome or the other.[3] As another example, in measuring the height of a person, we take it for granted that there is a well-defined way to pose a person against a ruler and decide exactly when we have a steady enough reading of the scale to declare a particular value for the person's height. The mechanical operationalization, mathematical formalization, and philosophical investigation of measurement could each have entire books devoted to them. We will have to settle for this single paragraph.

Consider the probability that a coin comes up heads when it is flipped. If the coin is fair, it should come up heads in about 50% of the flips. If the coin (or its flipping mechanism) is biased, then it will tend to come up heads more than or less than 50% of the flips. The probability of coming up heads can be denoted with parameter label θ (Greek letter theta); for example, a coin is fair when $\theta = 0.5$ (spoken "theta equals point five").

We can also consider our degree of belief that the coin is fair. We might know that the coin was manufactured by a government mint, and therefore we have a high degree of belief that the coin is fair. Alternatively, we might know that the coin was manufactured by Acme Magic and Novelty Company, and therefore we have a high degree of belief that the coin is biased. The degree of belief about a parameter can be denoted $p(\theta)$. If the coin was minted by the federal government, we might have a strong belief that the coin

[2] Many coins minted by governments have the picture of an important person's head on one side. This side is called "heads" or, technically, the "obverse." The reverse side is colloquially called "tails" as the natural opposite of "heads" even though there is rarely if ever a picture of a tail on the other side!

[3] Actually, it has been argued that *flipped* coins always have a 50% probability of coming up heads, and only *spun* coins can exhibit unequal head-tail probabilities (Gelman & Nolan, 2002). If this flip-spin distinction is important to you, please mentally substitute "spin" for "flip" whenever the text mentions flipping a coin. For empirical and theoretical studies of coin-flip probabilities, see, e.g., Diaconis, Holmes, and Montgomery (2007).

is fair; for example we might believe that $p(\theta = 0.5) = 0.99$, spoken "the probability that theta equals 0.5 is 99 percent." If the coin was minted by the novelty company, we might have a strong belief that the coin is biased; for example we might believe that $p(\theta = 0.5) = 0.01$ and that $p(\theta = 0.9) = 0.99$.

Both "probability" of head or tail outcome and "degree of belief" in biases refer to sample spaces. The sample space for flips of a coin consists of two possible outcomes: head and tail. The sample space for coin bias consists of a continuum of possible values: $\theta = 0.0$, $\theta = 0.01$, $\theta = 0.02$, $\theta = 0.03$, and all values in between, up to $\theta = 1.0$. When we flip a given coin, we are sampling from the space of head or tail. When we grab a coin at random from a sack of coins, in which each coin may have a different bias, we are sampling from the space of possible biases.

4.1.1. Coin flips: Why you should care

The fairness of a coin might be hugely consequential for high stakes games, but it isn't often in life that we flip coins and care about the outcome. So why bother studying the statistics of coin flips?

Because coin flips are a surrogate for myriad other real-life events that we do care about. For a given type of heart surgery, we may classify the patient outcome as survived more than a year or not, and we may want to know what is the probability that patients survive more than one year. For a given type of drug, we may classify the outcome as having a headache or not, and we may want to know the probability of headache. For a survey question, the outcome might be agree or disagree, and we want to know the probability of each response. In a two-candidate election, the two outcomes are candidate A and candidate B, and before the election itself we want to estimate, from a poll, the probability that candidate A will win. Or perhaps you are studying arithmetic ability by measuring accuracy on a multi-item exam, for which the item outcomes are correct or wrong. Or perhaps you are researching brain lateralization of a particular cognitive process in different subpopulations, in which case the outcomes are right-lateralized or left-lateralized, and you are estimating the probability of being left-lateralized in the subpopulation.

Whenever we are discussing coin flips, which might not be inherently fascinating to you, keep in mind that we could be talking about some domain in which you are actually interested! The coins are merely a generic representative of a universe of analogous applications.

4.2. PROBABILITY: OUTSIDE OR INSIDE THE HEAD

Sometimes we talk about probabilities of outcomes that are "out there" in the world. The face of a flipped coin is such an outcome: We can observe the flip, and the probability of coming up heads can be estimated by observing several flips.

But sometimes we talk about probabilities of things that are not so clearly "out there," and instead are just possible beliefs "inside the head." Our belief about the fairness of a coin is an example of something inside the head. The coin may have an intrinsic physical bias, but now I am referring to our *belief* about the bias. Our beliefs refer to a space of mutually exclusive and exhaustive possibilities. It might be strange to say that we randomly sample from our beliefs, like we randomly sample from a sack of coins. Nevertheless, the mathematical properties of probabilities outside the head and beliefs inside the head are the same in their essentials, as we will see.

4.2.1. Outside the head: Long-run relative frequency

For events outside the head, it's intuitive to think of probability as being the long-run relative frequency of each possible outcome. For example, if I say that for a fair coin the probability of heads is 0.5, what I mean is that if we flipped the coin many times, about 50% of the flips would come up heads. In the long run, after flipping the coin many, many times, the relative frequency of heads would be very nearly 0.5.

We can determine the long-run relative frequency by two different ways. One way is to approximate it by actually sampling from the space many times and tallying the number of times each event happens. A second way is by deriving it mathematically. These two methods are now explored in turn.

4.2.1.1 Simulating a long-run relative frequency

Suppose we want to know the long-run relative frequency of getting heads from a fair coin. It might seem blatantly obvious that we should get about 50% heads in any long sequence of flips. But let's pretend that it's not so obvious: All we know is that there's some underlying process that generates an "H" or a "T" when we sample from it. The process has a parameter called θ, whose value is $\theta = 0.5$. If that's all we know, then we can approximate the long-run probability of getting an "H" by simply repeatedly sampling from the process. We sample from the process N times, tally the number of times an "H" appeared, and estimate the probability of H by the relative frequency.

It gets tedious and time-consuming to sample a process manually, such as flipping a coin. Instead, we can let the computer do the repeated sampling much faster (and hopefully the computer feels less tedium than we would). Figure 4.1 shows the results of a computer simulating many flips of a fair coin. The R programming language has pseudo-random number generators built into it, which we will use often.[4] On the first flip, the computer randomly generates a head or a tail. It then computes the proportion

[4] Pseudo-random number generators (PRNGs) are not actually random; they are in fact deterministic. But the properties of the sequences they generate mimic the properties of random processes. The methods used in this book rely heavily on the quality of PRNGs, which is an active area of intensive research (e.g., Deng & Lin, 2000; Gentle, 2003).

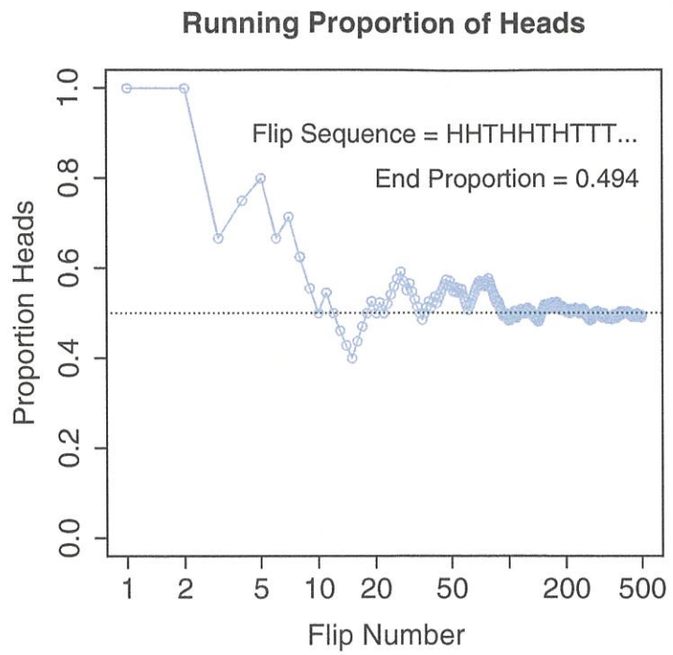

Figure 4.1 Running proportion of heads when flipping a coin. The x-axis is plotted on a logarithmic scale so that you can see the details of the first few flips but also the long-run trend after many flips. R code for producing this figure is discussed in Section 4.5.

of heads obtained so far. If the first flip was a head, then the proportion of heads is $1/1 = 1.0$. If the first flip was a tail, then the proportion of heads is $0/1 = 0.0$. Then the computer randomly generates a second head or tail, and computes the proportion of heads obtained so far. If the sequence so far is HH, then the proportion of heads is $2/2 = 1.0$. If the sequence so far is HT or TH, then the proportion of heads is $1/2 = 0.5$. If the sequence so far is TT, then the proportion of heads is $0/2 = 0.0$. Then the computer generates a third head or tail, and computes the proportion of heads so far, and so on for many flips. Figure 4.1 shows the running proportion of heads as the sequence continues.

Notice in Figure 4.1 that at the end of the long sequence, the proportion of heads is *near* 0.5 but not necessarily exactly equal to 0.5. This discrepancy reminds us that even this long run is still just a finite random sample, and there is no guarantee that the relative frequency of an event will match the true underlying probability of the event. That's why we say we are *approximating* the probability by the long-run relative frequency.

4.2.1.2 Deriving a long-run relative frequency

Sometimes, when the situation is simple enough mathematically, we can derive the exact long-run relative frequency. The case of the fair coin is one such simple situation. The sample space of the coin consists of two possible outcomes, head and tail. By the assumption of fairness, we know that each outcome is equally likely. Therefore, the long-run relative frequency of heads should be exactly one out of two, i.e., 1/2, and the long-run relative frequency of tails should also be exactly 1/2.

This technique is easily extended to other simple situations. Consider, for example, a standard six-sided die. It has six possible outcomes, namely 1 dot, 2 dots, ..., 6 dots. If we assume that the die is fair, then the long-run relative frequency of each outcome should be exactly 1/6.

Suppose that we put different dots on the faces of the six-side die. In particular, suppose that we put 1 dot on one face, 2 dots on two faces, and 3 dots on the remaining three faces. We still assume that each of the six faces is equally likely. Then the long-run relative frequency of 1 dot is exactly 1/6, and the long-run relative frequency of 2 dots is exactly 2/6, and the long-run relative frequency of 3 dots is exactly 3/6.

4.2.2. Inside the head: Subjective belief

How strongly do you believe that a coin minted by the US government is fair? If you believe that the coin could be slightly different than exactly fair, then how strongly do you believe that the probability of heads is $\theta = 0.51$? Or $\theta = 0.49$? If instead you are considering a coin that is ancient, asymmetric, and lopsided, do you believe that it inherently has $\theta = 0.50$? How about a coin purchased at a magic shop? We are not talking here about the true, inherent probability that the coin will come up heads. We are talking about our degree of belief in each possible probability.

To specify our subjective beliefs, we have to specify how likely we think each possible outcome is. It can be hard to pin down mushy intuitive beliefs. In the next section, we explore one way to "calibrate" subjective beliefs, and in the subsequent section we discuss ways to mathematically describe degrees of belief.

4.2.2.1 Calibrating a subjective belief by preferences

Consider a simple question that might affect travelers: How strongly do you believe that there will be a snowstorm that closes the interstate highways near Indianapolis next New Year's Day? Your job in answering that question is to provide a number between 0 and 1 that accurately reflects your belief probability. One way to come up with such a number is to calibrate your beliefs relative to other events with clear probabilities.

As a comparison event, consider a marbles-in-sack experiment. In a sack we put 10 marbles: 5 red, and 5 white. We shake the sack and then draw a marble at random. The probability of getting a red marble is, of course, $5/10 = 0.5$. We will use this sack of marbles as a comparison for considering snow in Indianapolis on New Year's Day.

Consider the following two gambles that you can choose from:

- Gamble A: You get $100 if there is a traffic stopping snowstorm in Indianapolis next New Year's Day.
- Gamble B: You get $100 if you draw a red marble from a sack of marbles with 5 red and 5 white marbles.

Which gamble would you prefer? If you prefer Gamble B, that means you think there is less than a 50-50 chance of a traffic-stopping snowstorm in Indy. So at least you now know that your subjective belief about the probability of traffic-stopping snowstorm is less than 0.5.

We can narrow down the degree of belief by considering other comparison gambles. Consider these two gambles:

- Gamble A: You get $100 if there is a traffic stopping snowstorm in Indianapolis next New Year's Day.
- Gamble C: You get $100 if you draw a red marble from a sack of marbles with 1 red and 9 white marbles.

Which gamble would you prefer? If you now prefer Gamble A, that means you think there is more than a 10% chance of traffic-stopping snowstorm in Indy on New Year's Day. Taken together, the two comparison gambles have told you that your subjective probability lies somewhere between 0.1 and 0.5. We could continue to consider preferences against other candidate gambles to calibrate your subjective belief more accurately.

4.2.2.2 Describing a subjective belief mathematically

When there are several possible outcomes in a sample space, it might be too much effort to try to calibrate your subjective belief about every possible outcome. Instead, you can use a mathematical function to summarize your beliefs.

For example, you might believe that the average American woman is 5′4″ tall, but be open to the possibility that the average might be somewhat above or below that value. It is too tedious and may be impossible to specify your degree of belief that the average height is 4′1″, or 4′2″, or 4′3″, and so on up through 6′1″, 6′2″, and 6′3″ etc. So you might instead describe your degree of belief by a bell-shaped curve that is highest at 5′4″ and drops off symmetrically above and below that most-likely height. You can change the width and center of the curve until it seems to best capture your subjective belief. Later in the book, we will talk about exact mathematical formulas for functions like these, but the point now is merely to understand the idea that mathematical functions can define curves that can be used to describe degrees of belief.

4.2.3. Probabilities assign numbers to possibilities

In general, a probability, whether it's outside the head or inside the head, is just a way of assigning numbers to a set of mutually exclusive possibilities. The numbers, called "probabilities," merely need to satisfy three properties (Kolmogorov, 1956):

1. A probability value must be nonnegative (i.e., zero or positive).
2. The sum of the probabilities across all events in the entire sample space must be 1.0 (i.e., one of the events in the space must happen, otherwise the space does not exhaust all possibilities).
3. For any two mutually exclusive events, the probability that one *or* the other occurs is the *sum* of their individual probabilities. For example, the probability that a fair six-sided die comes up 3-dots *or* 4-dots is $1/6 + 1/6 = 2/6$.

Any assignment of numbers to events that respects those three properties will also have all the properties of probabilities that we will discuss below. So whether a probability is thought of as a long-run relative frequency of outcomes in the world, or as a magnitude of a subjective belief, it behaves the same way mathematically.

4.3. PROBABILITY DISTRIBUTIONS

A probability *distribution* is simply a list of all possible outcomes and their corresponding probabilities. For a coin, the probability distribution is trivial: We list two outcomes (head and tail) and their two corresponding probabilities (θ and $1 - \theta$). For other sets of outcomes, however, the distribution can be more complex. For example, consider the height of a randomly selected person. There is some probability that the height will be 60.2″, some probability that the height will be 68.9″, and so forth, for every possible exact height. When the outcomes are continuous, like heights, then the notion of probability takes on some subtleties, as we will see.

4.3.1. Discrete distributions: Probability mass

When the sample space consists of discrete outcomes, then we can talk about the probability of each distinct outcome. For example, the sample space of a flipped coin has two discrete outcomes, and we talk about the probability of head or tail. The sample space of a six-sided die has six discrete outcomes, and we talk about the probability of 1 dot, 2 dots, and so forth.

For continuous outcome spaces, we can *discretize* the space into a finite set of mutually exclusive and exhaustive "bins." For example, although heights of people are a continuous scale, we can divide the scale into a finite number of intervals, such as $< 51''$, $51''$ to $53''$, $53''$ to $55''$, $55''$ to $57''$, …, $> 83''$. Then we can talk about the probability that a randomly selected person falls into any of those intervals. Suppose that we randomly sample 10,000 people and measure the heights very accurately. The top panel of Figure 4.2 shows a scatter plot of the 10,000 measurements, with vertical dashed lines marking the intervals. In particular, the number of measurements that fall within the interval $63''$ to $65''$ is 1,473, which means that the (estimated) probability of falling in that interval is $1,473/10,000 = 0.1473$.

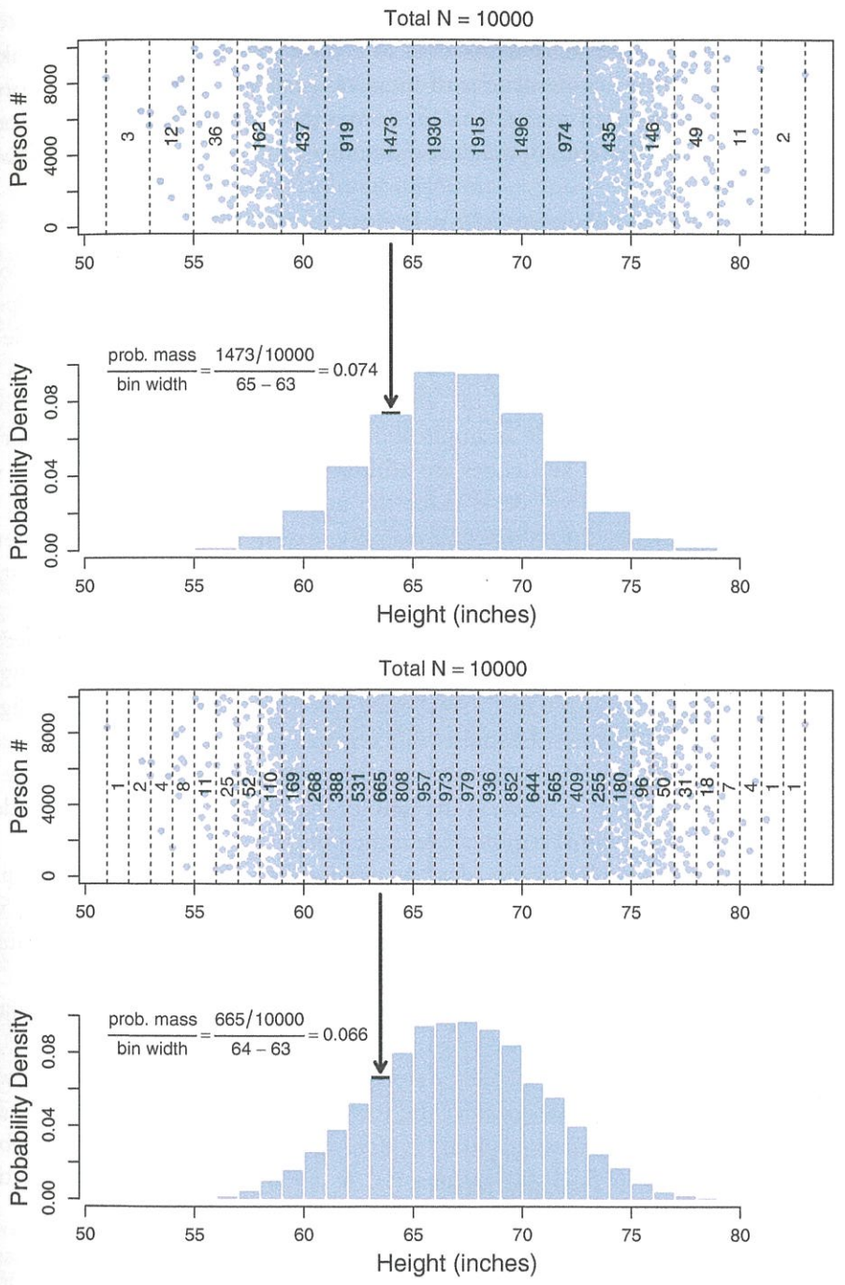

Figure 4.2 Examples of computing probability density. Within each main panel, the upper plot shows a scatter of 10,000 heights of randomly selected people, and the lower plot converts into probability density for the particular selection of bins depicted.

The probability of a discrete outcome, such as the probability of falling into an interval on a continuous scale, is referred to as a probability *mass*. Loosely speaking, the term "mass" refers the amount of stuff in an object. When the stuff is probability and the object is an interval of a scale, then the mass is the proportion of the outcomes in the interval. Notice that the sum of the probability masses across the intervals must be 1.

4.3.2. Continuous distributions: Rendezvous with density[5]

If you think carefully about a continuous outcome space, you realize that it becomes problematic to talk about the probability of a specific value on the continuum, as opposed to an interval on the continuum. For example, the probability that a randomly selected person has height (in inches) of exactly 67.21413908 ... is essentially nil, and that is true for *any* exact value you care to think of. We can, however, talk about the probability mass of intervals, as we did in the example above. The problem with using intervals, however, is that their widths and edges are arbitrary, and wide intervals are not very precise. Therefore, what we will do is make the intervals infinitesimally narrow, and instead of talking about the infinitesimal probability mass of each infinitesimal interval, we will talk about the ratio of the probability mass to the interval width. That ratio is called the probability *density*.

Loosely speaking, density is the amount of stuff per unit of space it takes up. Because we are measuring amount of stuff by its mass, then density is the mass divided by the amount space it occupies. Notice that a small mass can have a high density: A milligram of the metal lead has a density of more than 11 grams per cubic centimeter, because the milligram takes up only 0.000088 cubic centimeters of space. Importantly, we can conceive of density *at a point* in space, as the ratio of mass to space when the considered space shrinks to an infinitesimal region around the point.

Figure 4.2 shows examples of this idea. As previously mentioned, the upper panel shows a scatter plot of heights (in inches) of 10,000 randomly selected people, with intervals of width 2.0. To compute the average probability density in the interval 63″ to 65″, we divide the interval's probability mass by the interval's width. The probability mass is (estimated as) $1,473/10,000 = 0.1473$, and the interval width is 2.0 units (i.e., $65 - 63$), hence the average probability density in the interval is 0.074 (rounded). This is the average probability density over the interval. For a more precise density over a narrower interval, consider the lower panel of Figure 4.2. The interval 63″ to 64″ has (estimated) mass of $665/10,000$, and hence the average probability density in the interval is $(665/10,000)/(64 - 63) = 0.066$ (rounded). We can continue narrowing the intervals and computing density.

[5] "There is a mysterious cycle in human events. To some generations much is given. Of other generations much is expected. This generation of Americans has a rendezvous with destiny." Franklin Delano Roosevelt, 1936.

The example in Figure 4.2 illustrates the estimation of density from a finite sample across noninfinitesimal intervals. But to compute density for an infinitesimal interval, we must conceive of an infinite population continuously spread across the scale. Then, even an infinitesimal interval may contain some nonzero (though infinitesimal) amount of probability mass, and we can refer to probability density at a point on the scale. We will soon see mathematical examples of this idea.

Figure 4.3 shows another example, to emphasize that probability densities can be larger than 1, even though probability mass cannot exceed 1. The upper panel of Figure 4.3 shows heights in inches of 10,000 randomly selected *doors* that are manufactured to be 7 feet (84 inches) tall. Because of the regularity of the manufacturing process, there is only a little random variation among the heights of the doors, as can be seen in the figure by the fact that the range of the scale is small, going only from 83.6″ to 84.4″. Thus, all the probability mass is concentrated over a small range of the scale.

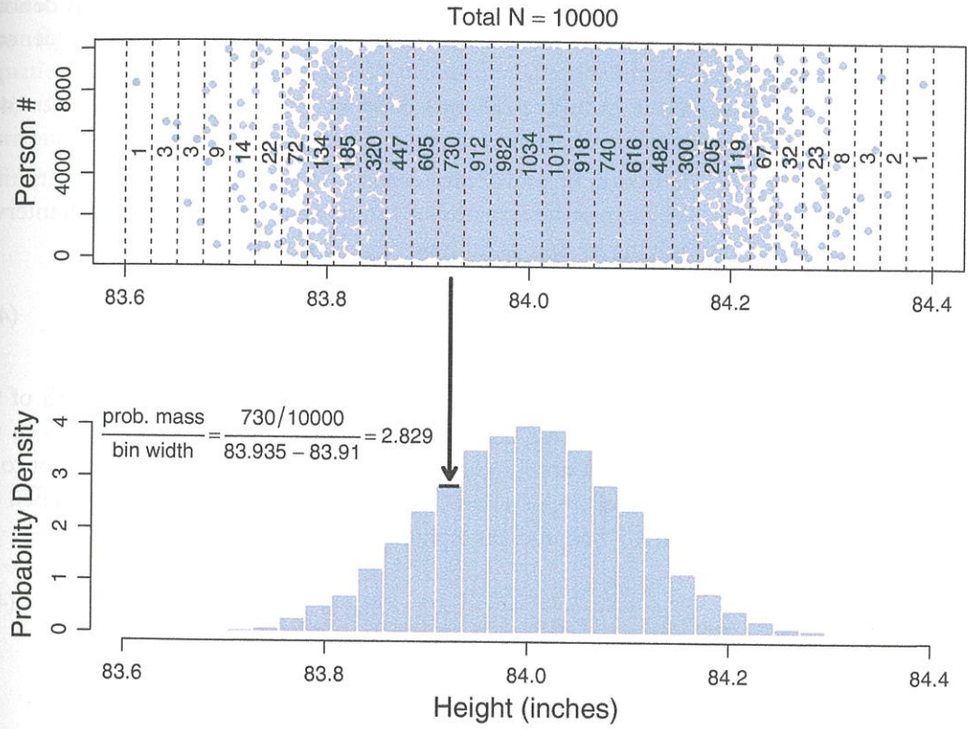

Figure 4.3 Example of probability density greater than 1.0. Here, all the probability mass is concentrated into a small region of the scale, and therefore the density can be high at some values of the scale. The annotated calculation of density uses rounded interval limits for display. (For this example, we can imagine that the points refer to manufactured doors instead of people, and therefore the *y* axis of the top panel should be labeled "Door" instead of "Person.")

Consequently, the probability density near values of 84 inches exceeds 1.0. For example, in the interval 83.9097″ to 83.9355″, there is a probability mass of $730/10,000 = 0.073$. But this mass is concentrated over a bin width of only $83.9355 - 83.9097 = 0.0258$, hence the average density within the interval is $0.073/0.0258 = 2.829$. There is nothing mysterious about probability densities larger than 1.0; it means merely that there is a high concentration of probability mass relative to the scale.

4.3.2.1 Properties of probability density functions

In general, for any continuous value that is split up into intervals, the sum of the probability masses of the intervals must be 1, because, by definition of making a measurement, some value of the measurement scale must occur. We can write that fact as an equation, but we need to define some notation first. Let the continuous variable be denoted x. The width of an interval on x is denoted Δx (the symbol "Δ" is the Greek letter, capital delta). Let i be an index for the intervals, and let $[x_i, x_i+\Delta x]$ denote the interval between x_i and $x_i+\Delta x$. The probability *mass* of the ith interval is denoted $p([x_i, x_i+\Delta x])$. Then the sum of those probability masses must be 1, which is denoted as follows:

$$\sum_i p([x_i, x_i+\Delta x]) = 1 \tag{4.1}$$

Recall now the definition of probability density: It is the ratio of probability mass over interval width. We can rewrite Equation 4.1 in terms of the density of each interval, by dividing and multiplying by Δx, as follows:

$$\sum_i \Delta x \frac{p([x_i, x_i+\Delta x])}{\Delta x} = 1 \tag{4.2}$$

In the limit, as the interval width becomes infinitesimal, we denote the width of the interval around x as dx instead of Δx, and we denote the probability *density* in the infinitesimal interval around x simply as $p(x)$. The probability density $p(x)$ is not to be confused with $p([x_i, x_i+\Delta x])$, which was the probability mass in an interval. Then the summation in Equation 4.2 becomes an integral:

$$\underbrace{\sum_i}_{\int} \underbrace{\Delta x}_{dx} \underbrace{\frac{p([x_i, x_i+\Delta x])}{\Delta x}}_{p(x)} = 1 \quad \text{that is,} \quad \int dx\, p(x) = 1 \tag{4.3}$$

In this book, integrals will be written with the dx term next to the integral sign, as in Equation 4.3, instead of at the far right end of the expression. Although this placement is not the most conventional notation, it is neither wrong nor unique to this book. The placement of dx next to the integral sign makes it easy to see what variable is being integrated over, without have to put subscripts on the integral sign. This usage can be especially helpful if we encounter integrals of functions that involve multiple variables. The placement of dx next to the integral sign also maintains grouping of terms when rewriting discrete sums and integrals, such that $\sum_x$ becomes $\int dx$ without having to move the dx to the end of the expression.

To reiterate, in Equation 4.3, $p(x)$ is the probability density in the infinitesimal interval around x. Typically, we let context tell us whether we are referring to a probability mass or a probability density, and use the same notation, $p(x)$, for both. For example, if x is the value of the face of a six-sided die, then $p(x)$ is a probability mass. If x is the exact point-value of height, then $p(x)$ is a probability density. There can be "slippage" in the usage, however. For example, if x refers to height, but the scale is discretized into intervals, then $p(x)$ is really referring to the probability mass of the interval in which x falls. Ultimately, you'll have to be attentive to context and tolerant of ambiguity.

4.3.2.2 The normal probability density function

Any function that has only nonnegative values and integrates to 1 (i.e., satisfies Equation 4.3) can be construed as a probability density function. Perhaps the most famous probability density function is the *normal* distribution, also known as the Gaussian distribution. A graph of the normal curve is a well-known bell shape; an example is shown in Figure 4.4.

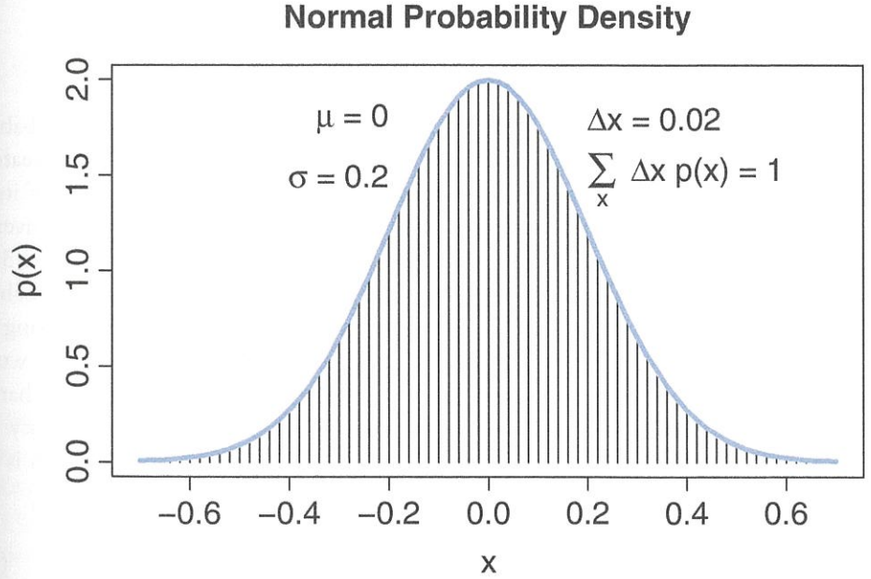

Figure 4.4 A normal probability density function, shown with a comb of narrow intervals. The integral is approximated by summing the width times height of each interval.

The mathematical formula for the normal probability density has two parameters: μ (Greek mu) is called the *mean* of the distribution and σ (Greek sigma) is called the *standard deviation*. The value of μ governs where the middle of the bell shape falls on the x-axis, so it is called a location parameter, and the value of σ governs how wide the bell is, so it is called a scale parameter. As discussed in Section 2.2, you can think of the parameters as control knobs with which to manipulate the location and scale of the distribution. The mathematical formula for the normal probability density is

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left[\frac{x-\mu}{\sigma}\right]^2\right). \tag{4.4}$$

Figure 4.4 shows an example of the normal distribution for specific values of μ and σ as indicated. Notice that the peak probability density can be greater than 1.0 when the standard deviation, σ, is small. In other words, when the standard deviation is small, a lot of probability mass is squeezed into a small interval, and consequently the probability density in that interval is high.

Figure 4.4 also illustrates that the area under the normal curve is, in fact, 1. The x axis is divided into a dense comb of small intervals, with width denoted Δx. The integral of the normal density is approximated by summing the masses of all the tiny intervals as in Equation 4.2. As can be seen in the text within the graph, the sum of the interval areas is essentially 1.0. Only rounding error, and the fact that the extreme tails of the distribution are not included in the sum, prevent the sum from being exactly 1.

4.3.3. Mean and variance of a distribution

When we have a numerical (not just categorical) value x that is generated with probability $p(x)$, we can wonder what would be its average value in the long run, if we repeatedly sampled values of x. For example, if we have a fair six-sided die, then each of its six values should come up 1/6th of the time in the long run, and so the long-run average value of the die is $(1/6)1 + (1/6)2 + (1/6)3 + (1/6)4 + (1/6)5 + (1/6)6 = 3.5$. As another example, if we play a slot machine for which we win \$100 with probability 0.001, we win \$5 with probability 0.14, and otherwise we lose \$1, then in the long run our payoff is $(0.001)(\$100) + (0.14)(\$5) + (0.859)(-\$1) = -\$0.059$. In other words, in the long run we lose about 6 cents per pull of the bandit's arm. Notice what we did in those calculations: We weighted each possible outcome by the probability that it happens. This procedure defines the *mean* of a probability distribution, which is also called the *expected value*, and which is denoted $E[x]$:

$$E[x] = \sum_x p(x)\, x \tag{4.5}$$

Equation 4.5 applies when the values of x are discrete, and so $p(x)$ denotes a probability mass. When the values of x are continuous, then $p(x)$ denotes a probability density and the sum becomes an integral over infinitesimal intervals:

$$E[x] = \int \mathrm{d}x\, p(x)\, x \tag{4.6}$$

The conceptual meaning is the same whether x is discrete or continuous: $E[x]$ is the long-run average of the values.

The mean value of a distribution typically lies near the distribution's middle, intuitively speaking. For example, the mean of a normal distribution turns out to be the value of its parameter μ. In other words, it turns out to be the case that $E[x] = \mu$. A specific example of that fact is illustrated in Figure 4.4, where it can be seen that the bulk of the distribution is centered over $x = \mu$; see the text in the figure for the exact value of μ.

Here's an example of computing the mean of a continuous distribution, using Equation 4.6. Consider the probability density function $p(x) = 6x(1-x)$ defined over the interval $x \in [0, 1]$. This really is a probability density function: It's an upside down parabola starting at $x = 0$, peaking over $x = 0.5$, and dropping down to baseline again at $x = 1$. Because it is a symmetric distribution, intuition tells us that the mean should be at its midpoint, $x = 0.5$. Let's check that it really is:

$$\begin{aligned}
E[x] &= \int \mathrm{d}x\, p(x)\, x \\
&= \int_0^1 \mathrm{d}x\; 6x(1-x)\, x \\
&= 6\int_0^1 \mathrm{d}x\; \left(x^2 - x^3\right) \\
&= 6\left[\frac{1}{3}x^3 - \frac{1}{4}x^4\right]_0^1 \\
&= 6\left[\left(\frac{1}{3}1^3 - \frac{1}{4}1^4\right) - \left(\frac{1}{3}0^3 - \frac{1}{4}0^4\right)\right] \\
&= 0.5
\end{aligned} \tag{4.7}$$

We will be doing relatively little calculus in this book, and Equation 4.7 is about as advanced as we'll get. If your knowledge of calculus is rusty, don't worry, just keep reading for conceptual understanding.

The *variance* of a probability distribution is a number that represents the dispersion of the distribution away from its mean. There are many conceivable definitions of how far the values of x are dispersed from their mean, but the definition used for the specific term

"variance" is based on the squared difference between x and the mean. The definition of variance is simply the mean squared deviation (MSD) of the x values from their mean:

$$\text{var}_x = \int \text{d}x\, p(x)\ (x - E[x])^2 \tag{4.8}$$

Notice that Equation 4.8 is just like the formula for the mean (Equation 4.6) except that instead of integrating x weighted by x's probability, we're integrating $(x - E[x])^2$ weighted by x's probability. In other words, the variance is just the average value of $(x - E[x])^2$. For a discrete distribution, the integral in Equation 4.8 becomes a sum, analogous to the relationship between Equations 4.5 and 4.6. The square root of the variance, sometimes referred to as root mean squared deviation (RMSD), is called the *standard deviation* of the distribution.

The variance of the normal distribution turns out to be the value of its parameter σ squared. Thus, for the normal distribution, $\text{var}_x = \sigma^2$. In other words, the standard deviation of the normal distribution is the value of the parameter σ. In a normal distribution, about 34% of the distribution lies between μ and $\mu + \sigma$ (see Exercise 4.5). Take a look at Figure 4.4 and visually identify where μ and $\mu + \sigma$ lie on the x axis (the values of μ and σ are indicated in the text within the figure) to get a visual impression of how far one standard deviation lies from the mean. Be careful, however, not to overgeneralize to distributions with other shapes: Non-normal distributions can have very different areas between their mean and first standard deviation.

A probability distribution can refer to probability of measurement values or of parameter values. The probability can be interpreted either as how much a value could be sampled from a generative process, or as how much credibility the value has relative to other values. When $p(\theta)$ represents credibility values of θ, instead of the probability of sampling θ, then the mean of $p(\theta)$ can be thought of as a value of θ that represents a typical credible value. The standard deviation of θ, which measures how wide the distribution is, can be thought of as a measure of uncertainty across candidate values. If the standard deviation is small, then we believe strongly in values of θ near the mean. If the standard deviation is large, then we are not very certain about what value of θ to believe in. This notion of standard deviation as representing uncertainty will reappear often. A related measure of the width of a distribution is the highest density interval, described below.

4.3.3.1 Mean as minimized variance

An alternative conceptual emphasis starts with the definition of variance and derives a definition of mean, instead of starting with the mean and working to a definition of variance. Under this alternative conception, the goal is to define a value for the *central tendency* of a probability distribution. A value represents the central tendency of the distribution if the value is close to the highly probable values of the distribution.

Therefore, we define the central tendency of a distribution as whatever value M minimizes the long-run expected distance between it and all the other values of x. But how should we define "distance" between values? One way to define distance is as squared difference: The distance between x and M is $(x - M)^2$. One virtue of this definition is that the distance from x to M is the same as the distance from M to x, because $(x - M)^2 = (M - x)^2$. But the primary virtue of this definition is that it makes a lot of subsequent algebra tractable (which will not be rehearsed here). The central tendency is, therefore, the value M that minimizes the expected value of $(x - M)^2$. Thus, we want the value M that minimizes $\int \text{d}x\, p(x)\ (x - M)^2$. Does that look familiar? It's essentially the formula for the variance of the distribution, in Equation 4.8, but here thought of as a function of M. Here's the punch line: It turns out that the value of M that minimizes $\int \text{d}x\, p(x)\ (x - M)^2$ is $E[x]$. In other words, the mean of the distribution is the value that minimizes the expected squared deviation. In this way, the mean is a central tendency of the distribution.

As an aside, if the distance between M and x is defined instead as $|x - M|$, then the value that minimizes the expected distance is called the *median* of the distribution. An analogous statement applies to the *modes* of a distribution, with distance defined as zero for any exact match, and one for any mismatch.

4.3.4. Highest density interval (HDI)

Another way of summarizing a distribution, which we will use often, is the highest density interval, abbreviated HDI.[6] The HDI indicates which points of a distribution are most credible, and which cover most of the distribution. Thus, the HDI summarizes the distribution by specifying an interval that spans most of the distribution, say 95% of it, such that every point inside the interval has higher credibility than any point outside the interval.

Figure 4.5 shows examples of HDIs. The upper panel shows a normal distribution with mean of zero and standard deviation of one. Because this normal distribution is symmetric around zero, the 95% HDI extends from -1.96 to $+1.96$. The area under the curve between these limits, and shaded in grey in Figure 4.5, has area of 0.95. Moreover, the probability density of any x within those limits has higher probability density than any x outside those limits.

[6] Some authors refer to the HDI as the HDR, which stands for highest density *region*, because a region can refer to multiple dimensions, but an interval refers to a single dimension. Because we will almost always consider the HDI of one parameter at a time, I will use "HDI" in an effort to reduce confusion. Some authors refer to the HDI as the HPD, to stand for highest probability density, but which I prefer not to use because it takes more space to write "HPD interval" than "HDI." Some authors refer to the HDI as the HPD, to stand for highest *posterior* density, but which I prefer not to use because *prior* distributions also have HDIs.

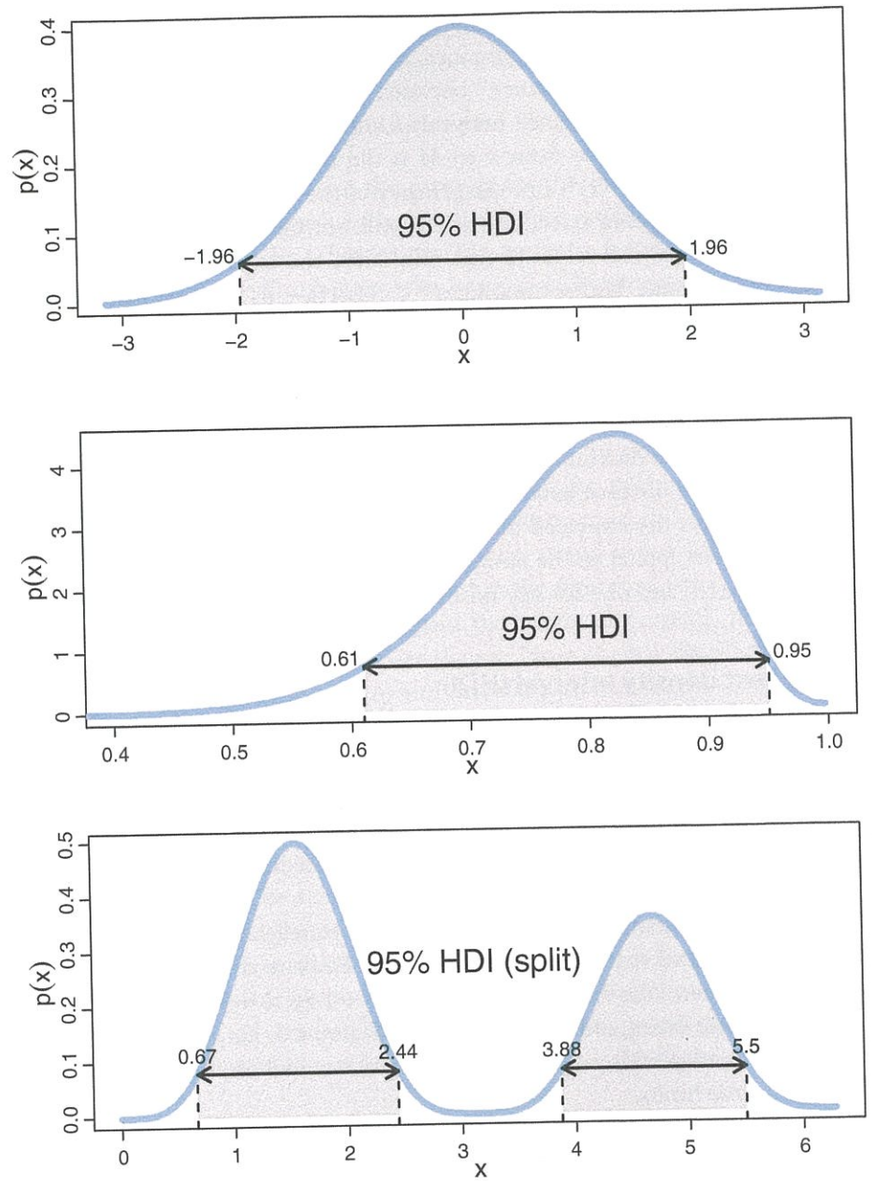

Figure 4.5 Examples of 95% highest density intervals (HDIs). For each example, all the x values inside the interval have higher density than any x value outside the interval, and the total mass of the points inside the interval is 95%. The 95% area is shaded, and it includes the zone below the horizontal arrow. The horizontal arrow indicates the width of the 95% HDI, with its ends annotated by (rounded) x values. The height of the horizontal arrow marks the minimal density exceeded by all x values inside the 95% HDI.

The middle panel of Figure 4.5 shows a 95% HDI for a skewed distribution. By definition, the area under the curve between the 95% HDI limits, shaded in grey in the figure, has area of 0.95, and the probability density of any x within those limits is higher than any x outside those limits. Importantly, notice that the area in the left tail, less than the left HDI limit, is larger than the area in right tail, greater than the right HDI limit. In other words, the HDI does not necessarily produce equal-area tails outside the HDI. (For those of you who have previously encountered the idea of equal-tailed credible intervals, you can look ahead to Figure 12.2, p. 342, for an explanation of how HDIs differ from equal-tailed intervals.)

The lower panel of Figure 4.5 shows a fanciful bimodal probability density function. In many realistic applications, multimodal distributions such as this do not arise, but this example is useful for clarifying the definition of an HDI. In this case, the HDI is split into two subintervals, one for each mode of the distribution. However, the defining characteristics are the same as before: The region under the curve within the 95% HDI limits, shaded in grey in the figure, has total area of 0.95, and any x within those limits has higher probability density than any x outside those limits.

The formal definition of an HDI is just a mathematical expression of the two essential characteristics. The 95% HDI includes all those values of x for which the density is at least as big as some value W, such that the integral over all those x values is 95%. Formally, the values of x in the 95% HDI are those such that $p(x) > W$ where W satisfies $\int_{x:p(x)>W} dx\ p(x) = 0.95$.

When the distribution refers to credibility of values, then the width of the HDI is another way of measuring uncertainty of beliefs. If the HDI is wide, then beliefs are uncertain. If the HDI is narrow, then beliefs are relatively certain. As will be discussed at length in Chapter 13, sometimes the goal of research is to obtain data that achieve a reasonably high degree of certainty about a particular parameter value. The desired degree of certainty can be measured as the width of the 95% HDI. For example, if μ is a measure of how much a drug decreases blood pressure, the researcher may want to have an estimate with a 95% HDI width no larger than 5 units on the blood pressure scale. As another example, if θ is a measure of a population's preference for candidate A over candidate B, the researcher may want to have an estimate with a 95% HDI width no larger than 10 percentage points.

4.4. TWO-WAY DISTRIBUTIONS

There are many situations in which we are interested in the conjunction of two outcomes. What is the probability of being dealt a card that is both a queen *and* a heart? What is the probability of meeting a person with both red hair *and* green eyes? When playing a board game involving a die and a spinner, we have degrees of belief about both the die *and* the spinner being fair.

Table 4.1 Proportions of combinations of hair color and eye color

	Hair color				
Eye color	**Black**	**Brunette**	**Red**	**Blond**	**Marginal (eye color)**
Brown	0.11	0.20	0.04	0.01	0.37
Blue	0.03	0.14	0.03	0.16	0.36
Hazel	0.03	0.09	0.02	0.02	0.16
Green	0.01	0.05	0.02	0.03	0.11
Marginal (hair color)	0.18	0.48	0.12	0.21	1.0

Some rows or columns may not sum exactly to their displayed marginals because of rounding error from the original data. Data adapted from Snee (1974).

As a specific example for developing these ideas, consider Table 4.1, which shows the probabilities of various combinations of people's eye color and hair color. The data come from a particular convenience sample (Snee, 1974), and are not meant to be representative of any larger population. Table 4.1 considers four possible eye colors, listed in its rows, and four possible hair colors, listed across its columns. In each of its main cells, the table indicates the *joint probability* of particular combinations of eye color and hair color. For example, the top-left cell indicates that the joint probability of brown eyes and black hair is 0.11 (i.e., 11%). Notice that not all combinations of eye color and hair color are equally likely. For example, the joint probability of blue eyes and black hair is only 0.03 (i.e., 3%). We denote the joint probability of eye color e and hair color h as $p(e, h)$. The notation for joint probabilities is symmetric: $p(e, h) = p(h, e)$.

We may be interested in the probabilities of the eye colors overall, collapsed across hair colors. These probabilities are indicated in the right margin of the table, and they are therefore called *marginal* probabilities. They are computed simply by summing the joint probabilities in each row, to produce the row sums. For example, the marginal probability of green eyes, irrespective of hair color, is 0.11. The joint values indicated in the table do not all sum exactly to the displayed marginal values because of rounding error from the original data. The marginal probability of eye color e is denoted $p(e)$, and it is computed by summing the joint probabilities across the hair colors: $p(e) = \sum_h p(e, h)$.

Of course, we can also consider the marginal probabilities of the various hair colors. The marginal probabilities of the hair colors are indicated on the lower margin of Table 4.1. For example, the probability of black hair, irrespective of eye color, is 0.18. The marginal probabilities are computed by summing the joint probabilities within each column. Thus, $p(h) = \sum_e p(e, h)$.

In general, consider a row variable r and a column variable c. When the row variables are continuous instead of discrete, then $p(r, c)$ is a probability density, and the summation for computing the marginal probability becomes an integral, $p(r) = \int \mathrm{d}c\, p(r, c)$, where the resulting marginal distribution, $p(r)$, is also a probability density. This summation

process is called *marginalizing over* c or *integrating out* the variable c. Of course, we can also determine the probability $p(c)$ by marginalizing over r: $p(c) = \int \mathrm{d}r\, p(r, c)$.

4.4.1. Conditional probability

We often want to know the probability of one outcome, given that we know another outcome is true. For example, suppose I sample a person at random from the population referred to in Table 4.1. Suppose I tell you that this person has blue eyes. Conditional on that information, what is the probability that the person has blond hair (or any other particular hair color)? It is intuitively clear how to compute the answer: We see from the blue-eye row of Table 4.1 that the total (i.e., marginal) amount of blue-eyed people is 0.36, and that 0.16 of the population has blue eyes and blond hair. Therefore, of the 0.36 with blue eyes, the fraction 0.16/0.36 has blond hair. In other words, of the blue-eyed people, 45% have blond hair. We also note that of the blue-eyed people, 0.03/0.36 = 8% have black hair. Table 4.2 shows this calculation for each of the hair colors.

The probabilities of the hair colors represent the credibilities of each possible hair color. For this group of people, the general probability of having blond hair is 0.21, as can be seen from the marginal distribution of Table 4.1. But when we learn that a person from this group has blue eyes, then the credibility of that person having blond hair increases to 0.45, as can be seen from Table 4.2. This reallocation of credibility across the possible hair colors *is* Bayesian inference! But we are getting ahead of ourselves; the next chapter will explain the basic mathematics of Bayesian inference in detail.

The intuitive computations for conditional probability can be denoted by simple formal expressions. We denote the conditional probability of hair color given eye color as $p(h|e)$, which is spoken "the probability of h *given* e." The intuitive calculations above are then written $p(h|e) = p(e, h)/p(e)$. This equation is taken as the *definition* of conditional probability. Recall that the marginal probability is merely the sum of the cell probabilities, and therefore the definition can be written $p(h|e) = p(e, h)/p(e) = p(e, h)/\sum_h p(e, h)$. That equation can be confusing because the h in the numerator is a specific value of hair color, but the h in the denominator is a variable that takes on all possible values of hair color. To disambiguate the two meanings of h, the equation can be written $p(h|e) = p(e, h)/p(e) = p(e, h)/\sum_{h^*} p(e, h^*)$, where h^* indicates possible values of hair color.

Table 4.2 Example of conditional probability

	Hair color				
Eye color	**Black**	**Brunette**	**Red**	**Blond**	**Marginal (eye color)**
Blue	0.03/0.36 = 0.08	0.14/0.36 = 0.39	0.03/0.36 = 0.08	0.16/0.36 = 0.45	0.36/0.36 = 1.0

Of the blue-eyed people in Table 4.1, what proportion have hair color h? Each cell shows $p(h|\text{blue}) = p(\text{blue}, h)/p(\text{blue})$ rounded to two decimal points.

The definition of conditional probability can be written using more general variable names, with r referring to an arbitrary row attribute and c referring to an arbitrary column attribute. Then, for attributes with discrete values, conditional probability is defined as

$$p(c|r) = \frac{p(r,c)}{\sum_{c^*} p(r,c^*)} = \frac{p(r,c)}{p(r)} \tag{4.9}$$

When the column attribute is continuous, the sum becomes an integral:

$$p(c|r) = \frac{p(r,c)}{\int \mathrm{d}c\, p(r,c)} = \frac{p(r,c)}{p(r)} \tag{4.10}$$

Of course, we can conditionalize on the other variable, instead. That is, we can consider $p(r|c)$ instead of $p(c|r)$. It is important to recognize that, in general, $p(r|c)$ is *not* equal to $p(c|r)$. For example, the probability that the ground is wet, given that it's raining, is different than the probability that it's raining, given that the ground is wet. The next chapter provides an extended discussion of the relationship between $p(r|c)$ and $p(c|r)$.

It is also important to recognize that there is no temporal order in conditional probabilities. When we say "the probability of x given y" we do *not* mean that y has already happened and x has yet to happen. All we mean is that we are restricting our calculations of probability to a particular subset of possible outcomes. A better gloss of $p(x|y)$ is, "among all joint outcomes with value y, this proportion of them also has value x." So, for example, we can talk about the conditional probability that it rained the previous night given that there are clouds the next morning. This is simply referring to the proportion of all cloudy mornings that had rain the night before.

4.4.2. Independence of attributes

Suppose I have a six-sided die and a coin. Suppose they are fair. I flip the coin and it comes up heads. Given this result from the coin, what is the probability that the rolled die will come up 3? In answering this question, you probably thought, "the coin has no influence on the die, so the probability of the die coming up 3 is 1/6 regardless of the result from the coin." If that's what you thought, you were assuming that the spinner and the coin are *independent*.

In general, when the value of y has no influence on the value of x, we know that the probability of x given y simply is the probability of x in general. Written formally, that idea is expressed as $p(x|y) = p(x)$ for all values of x and y. Let's think a moment about what that implies. We know from the definition of conditional probability, in Equations 4.9 or 4.10, that $p(x|y) = p(x,y)/p(y)$. Combining those equations implies that $p(x) = p(x,y)/p(y)$ for all values of x and y. After multiplying both sides by $p(y)$, we get the implication that $p(x,y) = p(x)p(y)$ for all values of x and y. The implication goes the other way, too: When $p(x,y) = p(x)p(y)$ for all values of x and y, then $p(x|y) = p(x)$ for all values of x and y. Therefore, either of these conditions is our mathematical definition of independence of attributes. To reiterate, to say that attributes x and y are independent means that $p(x|y) = p(x)$ for all values of x and y, which is mathematically equivalent to saying that $p(x,y) = p(x)p(y)$ for all values of x and y.

Consider the example of eye color and hair color back in Table 4.1 (page 90). Are the attributes independent? Intuitively from everyday experience, we know that the answer should be no, but we can show it mathematically. All we need, to disprove independence, is some eye color e and some hair color h for which $p(h|e) \neq p(h)$, or equivalently for which $p(e,h) \neq p(e)p(h)$. We already dealt with such a case, namely blue eyes and blond hair. Table 4.1 shows that the marginal probability of blond hair is $p(\text{blond}) = 0.21$, while Table 4.2 shows that the conditional probability of blond hair given blue eyes is $p(\text{blond}|\text{blue}) = 0.45$. Thus, $p(\text{blond}|\text{blue}) \neq p(\text{blond})$. We can instead disprove independence by cross-multiplying the marginal probabilities and showing that they do not equal the joint probability: $p(\text{blue}) \cdot p(\text{blond}) = 0.36 \cdot 0.21 = 0.08 \neq 0.16 = p(\text{blue, blond})$.

As a simple example of two attributes that *are* independent, consider the suit and value of playing cards in a standard deck. There are four suits (diamonds, hearts, clubs, and spades), and thirteen values (ace, 2, ..., 9, jack, queen, king) of each suit, making 52 cards altogether. Consider a randomly dealt card. What is the probability that it is a heart? (Answer: 13/52 = 1/4.) Suppose I look at the card without letting you see it, and I tell you that it is a queen. Now what is the probability that it is a heart? (Answer: 1/4.) In general, telling you the card's value does not change the probabilities of the suits, therefore value and suit are independent. We can verify this in terms of cross multiplying marginal probabilities, too: Each combination of value and suit has a 1/52 chance of being dealt (in a fairly shuffled deck). Notice that 1/52 is exactly the marginal probability of any one suit (1/4) times the marginal probability of any one value (1/13).

Among other contexts, independence will come up when we are constructing mathematical descriptions of our beliefs about more than one parameter. We will create a mathematical description of our beliefs about one parameter, and another mathematical description of our beliefs about the other parameter. Then, to describe what we believe about combinations of parameters, we will often assume independence, and simply multiply the separate credibilities to specify the joint credibilities.

4.5. APPENDIX: R CODE FOR FIGURE 4.1

Figure 4.1 was produced by the script `RunningProportion.R`. To run it, simply type `source("RunningProportion.R")` at R's command line (assuming that your working directory contains the file!). Each time you run it, you will get a different result because of the pseudo-random number generator. If you want to set the pseudo-random number generator to a specific starting state, use the `set.seed`