CHAPTER 4

What Is This Stuff Called Probability?

Contents

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).

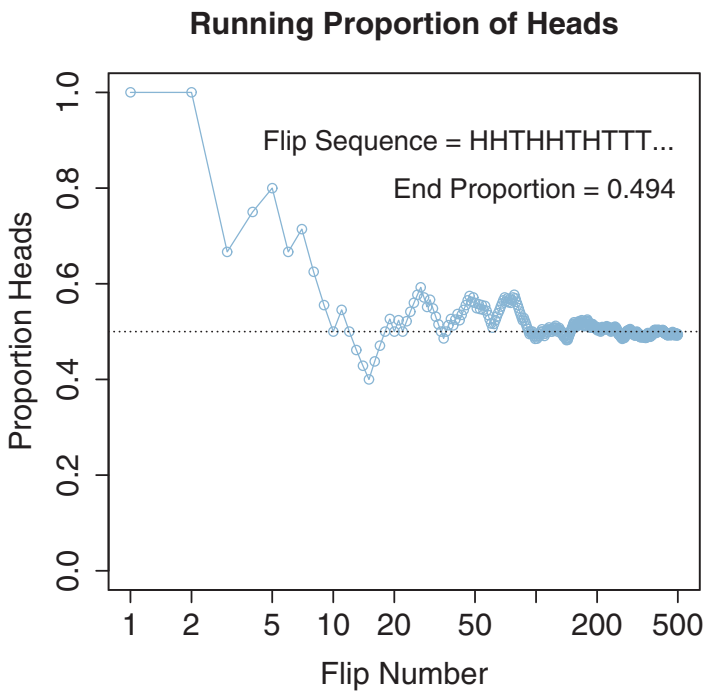

Figure 4.1 Running proportion of heads when flipping a coin. The x-axis is plotted on a logarithmic scale so that you can see the details of the first few flips but also the long-run trend after many flips. R code for producing this figure is discussed in Section 4.5.

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 $1,473/10,000 = 0.1473$.

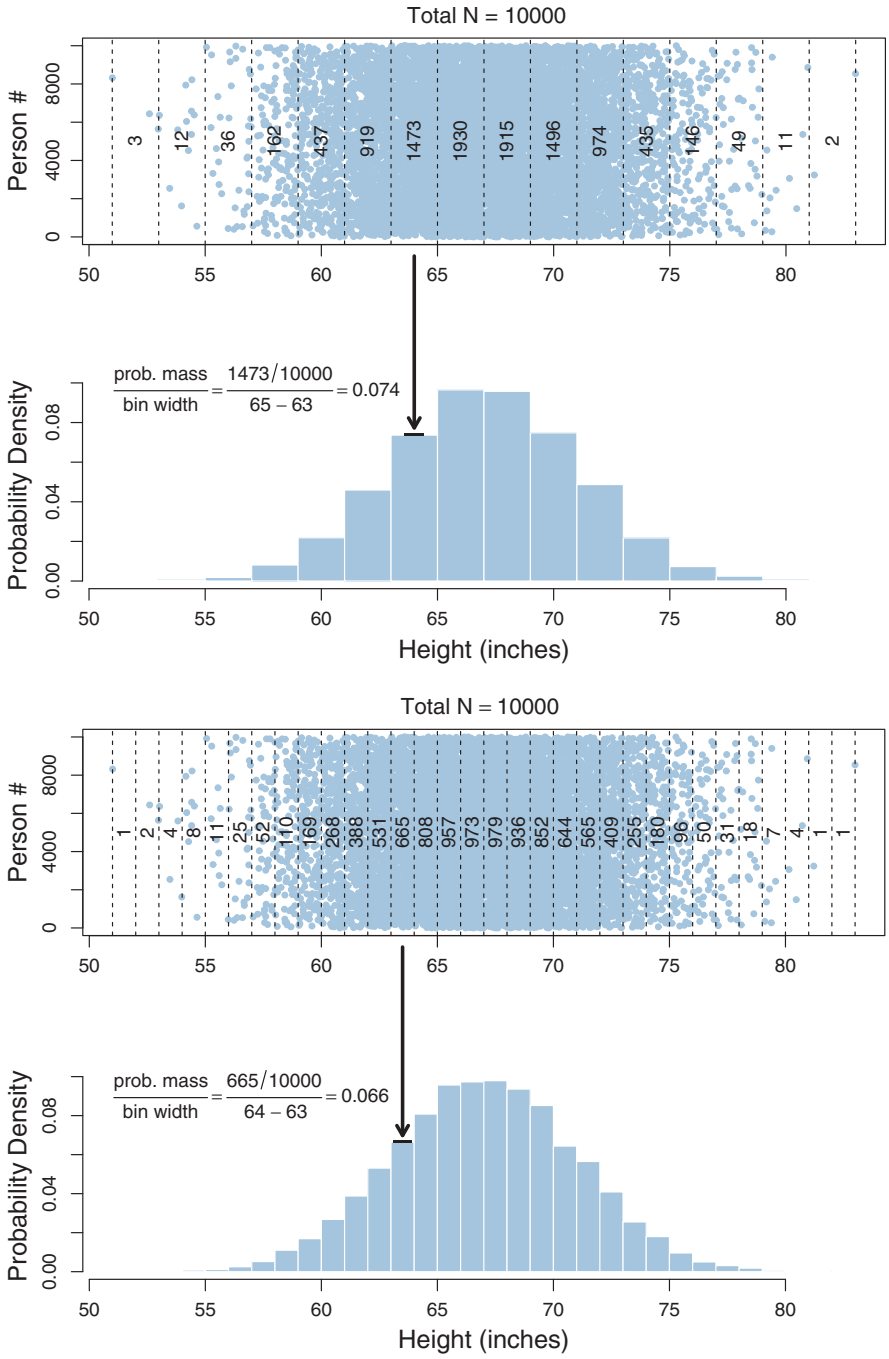

Figure 4.2 Examples of computing probability density. Within each main panel, the upper plot shows a scatter of 10,000 heights of randomly selected people, and the lower plot converts into probability density for the particular selection of bins depicted.

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.

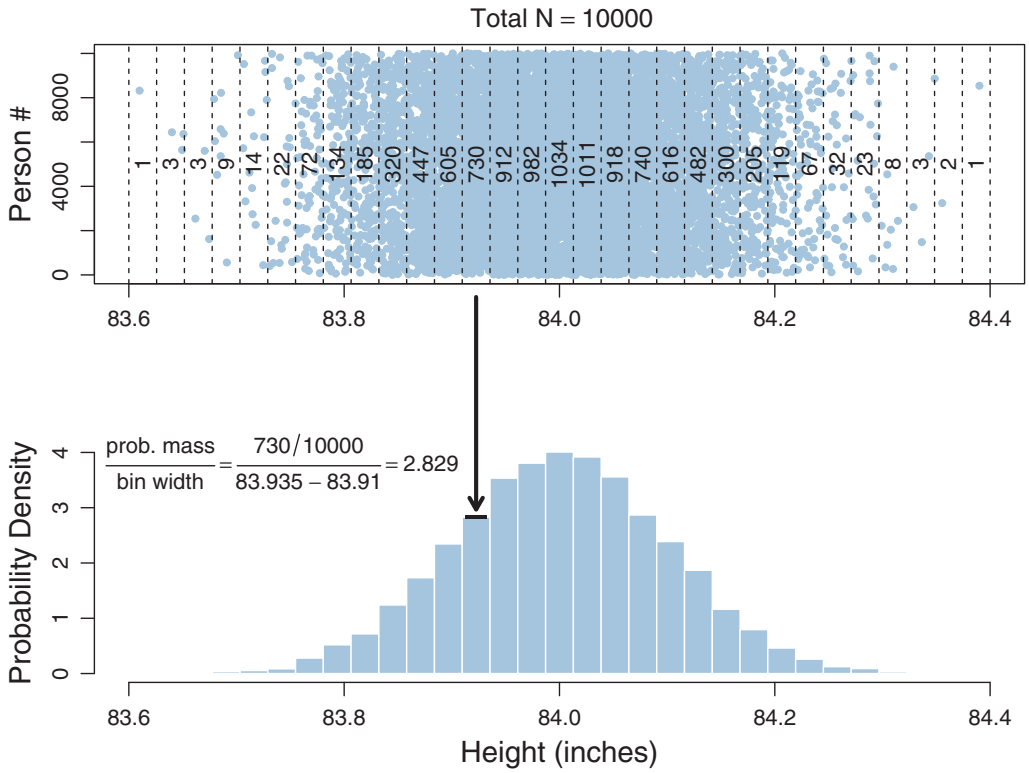

Figure 4.3 Example of probability density greater than 1.0. Here, all the probability mass is concentrated into a small region of the scale, and therefore the density can be high at some values of the scale. The annotated calculation of density uses rounded interval limits for display. (For this example, 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:

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

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:

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

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 $\int \mathrm{d}x$ 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.

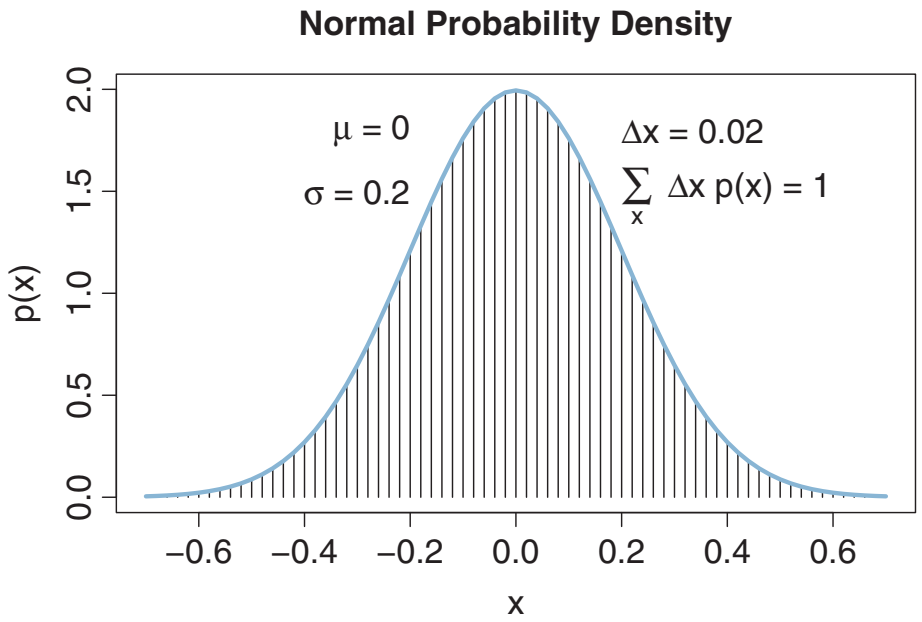

Figure 4.4 A normal probability density function, shown with a comb of narrow intervals. The integral is approximated by summing the width times height of each interval.

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 $\int \mathrm{d}x\, p(x)\, (x - M)^2$. 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 $\int \mathrm{d}x\, p(x)\, (x - M)^2$ 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.

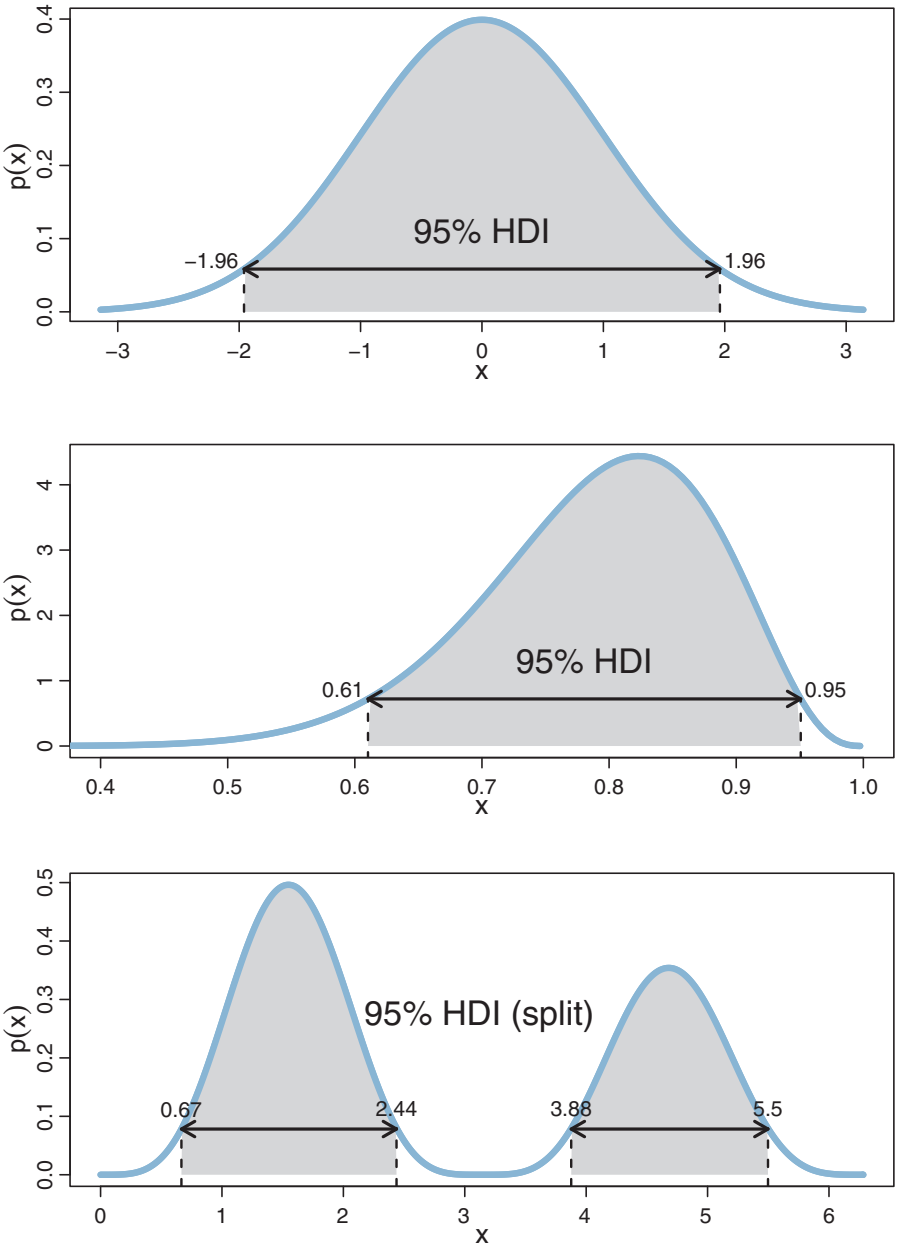

Figure 4.5 Examples of 95% highest density intervals (HDIs). For each example, all the *x* values inside the interval have higher density than any *x* value outside the interval, and the total mass of the points inside the interval is 95%. The 95% area is shaded, and it includes the zone below the horizontal 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 $\int_{x:\, p(x)>W} \mathrm{d}x\; p(x) = 0.95$.

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: $p(\text{blue}) \cdot p(\text{blond}) = 0.36 \cdot 0.21 = 0.08 \neq 0.16 = p(\text{blue}, \text{blond})$.

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`

command. The example in Figure 4.1 was created by typing `set.seed(47405)` and then `source("RunningProportion.R")`.

If you want to control the window size created for the graph and subsequently save the figure, you can load the graphics functions defined in the utility programs that accompany this book. Here is a sequence of commands that open and save a separate graphics window:

```
source("DBDA2E-utilities.R") # Definitions of openGraph, saveGraph, etc.
set.seed(47405)              # Optional, for a specific pseudo-random sequence.
openGraph(width=6,height=6)
source("RunningProportion.R")
saveGraph(file="RunningProportionExample",type="jpg")
```

The previous paragraphs explain how to use the script `RunningProportion.R`, but if you want to explore its internal mechanics, you can open it in RStudio's editing window. You will see a script of commands like this:

```
N = 500       # Specify the total number of flips, denoted N.
pHeads = 0.5  # Specify underlying probability of heads.
# Generate a random sample of N flips (heads=1, tails=0):
flipSequence = sample( x=c(0,1), prob=c(1-pHeads,pHeads), size=N, replace=TRUE)
# Compute the running proportion of heads:
r = cumsum( flipSequence ) # Cumulative sum: Number of heads at each step.
n = 1:N                    # Number of flips at each step.
runProp = r / n            # Component by component division.
# Graph the running proportion:
plot( n , runProp , type="o" , log="x" , col="skyblue" ,
      xlim=c(1,N) , ylim=c(0.0,1.0) , cex.axis=1.5 ,
      xlab="Flip Number" , ylab="Proportion Heads" , cex.lab=1.5 ,
      main="Running Proportion of Heads" , cex.main=1.5 )
# Plot a dotted horizontal reference line:
abline( h=pHeads , lty="dotted" )
# Display the beginning of the flip sequence:
flipLetters = paste( c("T","H")[flipSequence[1:10]+1] , collapse="" )
displayString = paste0( "Flip Sequence = " , flipLetters , "..." )
text( N , .9 , displayString , adj=c(1,0.5) , cex=1.3 )
# Display the relative frequency at the end of the sequence.
text( N , .8 , paste("End Proportion =",runProp[N]) , adj=c(1,0.5) , cex=1.3 )
```

The first two commands, above, merely specify the number of flips and the underlying probability of heads for the simulated coin. The fourth line introduces a new function that is predefined in standard distributions of R, namely the `sample` function. It generates pseudo-random samples from the set of elements defined by the user-supplied argument `x`, which in this case is a vector containing a 0 (zero) and a 1 (one). The argument `prob` specifies the probability with which each component of `x`

should be sampled. You can read more about the `sample` function by typing `?sample` at R's command line.

The sixth line, above, uses the cumulative sum function, `cumsum`, which is predefined in R. You can read about the `cumsum` function by typing `?cumsum` at R's command line. The function computes, at each component of a vector, the cumulative sum up to that component. For example, `cumsum ( c(1,1,0,0,1) )` produces the vector `1 2 2 2 3`.

As you read through the remaining lines of R code, above, it helps to run each line individually in R to see what it does. For example, to see the contents of the variable `runProp`, just type `runProp` and R's command line. To learn about the various arguments of the `plot` function, type `?plot` at R's command line.

The line that defines the variable `flipLetters` may seem a bit mysterious. Whenever you encounter a complex command like this in R, it can be a good strategy to unpack the commands from the inside out (as was mentioned in Section 3.7.6 in relation to diagnosing programming errors). Try entering these commands in R (or just select the text in RStudio and click Run):

```
                flipSequence[1:10]
                flipSequence[1:10]+1
      c("T","H")[flipSequence[1:10]+1]
paste( c("T","H")[flipSequence[1:10]+1] , collapse="" )
```

4.6. EXERCISES

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

Exercise 4.1. [Purpose: To gain experience with the `apply` function in R, while dealing with a concrete example of computing conditional probabilities.] The eye-color hair-color data from Table 4.1 are built into R as the array named HairEyeColor. The array is frequencies of eye and hair color for males and females. Run the following code in R:

```
show( HairEyeColor ) # Show data
EyeHairFreq = apply( HairEyeColor, c("Eye","Hair"), sum ) # Sum across sex
EyeHairProp = EyeHairFreq / sum( EyeHairFreq )  # joint proportions, Table 4.1
show( round( EyeHairProp , 2 ) )
HairFreq = apply( HairEyeColor , c("Hair") , sum ) # Sum across sex and eye
HairProp = HairFreq / sum( HairFreq ) # marginal proportions, Table 4.1
show( round( HairProp , 2 ) )
EyeFreq = apply( HairEyeColor , c("Eye") , sum ) # Sum across sex and eye
EyeProp = EyeFreq / sum( EyeFreq ) # marginal proportions, Table 4.1
show( round( EyeProp , 2 ) )
EyeHairProp["Blue",] / EyeProp["Blue"] # conditional prob, Table 4.2
```

In your write-up, include each line above and its results. *Explain* what each line does (in a bit more detail than the inline comments). Extend the above commands by also computing the probabilities of the hair colors given Brown eyes, and the probabilities of the eye colors given Brown hair.

Exercise 4.2. [Purpose: To give you some experience with random number generation in R.] Modify the coin flipping program in Section 4.5 `RunningProportion.R` to simulate a biased coin that has $p(H) = 0.8$. Change the height of the reference line in the plot to match $p(H)$. Comment your code. *Hint:* Read the help for the `sample` command.

Exercise 4.3. [Purpose: To have you work through an example of the logic presented in Section 4.2.1.2.] Determine the exact probability of drawing a 10 from a shuffled pinochle deck. (In a pinochle deck, there are 48 cards. There are six values: 9, 10, Jack, Queen, King, Ace. There are two copies of each value in each of the standard four suits: hearts, diamonds, clubs, spades.)

(A) What is the probability of getting a 10?

(B) What is the probability of getting a *10 or Jack*?

Exercise 4.4. [Purpose: To give you hands-on experience with a simple probability density function, in R and in calculus, and to reemphasize that density functions can have values larger than 1.] Consider a spinner with a [0,1] scale on its circumference. Suppose that the spinner is slanted or magnetized or bent in some way such that it is biased, and its probability density function is $p(x) = 6x(1 - x)$ over the interval $x \in [0, 1]$.

(A) Adapt the program `IntegralOfDensity.R` to plot this density function and approximate its integral. Comment your code. Be careful to consider values of x only in the interval $[0, 1]$. *Hint:* You can omit the first couple of lines regarding `meanval` and `sdval`, because those parameter values pertain only to the normal distribution. Then set `xlow=0` and `xhigh=1`, and set `dx` to some small value.

(B) Derive the exact integral using calculus. *Hint:* See the example, Equation 4.7.

(C) Does this function satisfy Equation 4.3?

(D) From inspecting the graph, what is the maximal value of $p(x)$?

Exercise 4.5. [Purpose: To have you use a normal curve to describe beliefs. It's also handy to know the area under the normal curve between μ and σ.]

(A) Adapt the code from `IntegralOfDensity.R` to determine (approximately) the probability mass under the normal curve from $x = \mu - \sigma$ to $x = \mu + \sigma$. Comment your code. *Hint:* Just change `xlow` and `xhigh` appropriately, and change the `text` location so that the `area` still appears within the plot.

(B) Now use the normal curve to describe the following belief. Suppose you believe that women's heights follow a bell-shaped distribution, centered at 162 cm with about two-thirds of all women having heights between 147 and 177 cm. What should be the μ and σ parameter values?

Exercise 4.6. [Purpose: Recognize and work with the fact that Equation 4.9 can be solved for the joint probability, which will be crucial for developing Bayes' theorem.] School children were surveyed regarding their favorite foods. Of the total sample, 20% were 1st graders, 20% were 6th graders, and 60% were 11th graders. For each grade, the following table shows the proportion of respondents that chose each of three foods as their favorite:

	Ice cream	Fruit	French fries
1st graders	0.3	0.6	0.1
6th graders	0.6	0.3	0.1
11th graders	0.3	0.1	0.6

From that information, construct a table of joint probabilities of grade and favorite food. Also, say whether grade and favorite food are independent or not, and how you ascertained the answer. *Hint:* You are given p(grade) and p(food|grade). You need to determine p(grade,food).