7
Probability

> Everything existing in the universe is the fruit of chance.
> —Democritus

In this second major part of the book, we begin our study of probability, for two main reasons.

(1) Knowing some probability sharpens our understanding of uncertainty quantification in the kind of statistical models we've been using to describe patterns in data.

(2) Probability theory provides a handy framework that can guide our thinking when faced with the important decisions that shape our careers, our lives, and our world.

Probability in statistical inference

In Chapter 5, we considered three important questions about uncertainty in statistical models.

(1) How confident are we in an estimate of an effect size?

(2) How confident are we in our prediction?

(3) Are we sure that a postulated effect is even real. Or, on the other hand, could it plausibly be due to chance?

The techniques we used to address these questions—sampling distributions, bootstrapping, permutation tests, and so forth—all implicitly invoked the idea of probability. In this part of the book, we'll take a "back to basics" approach that enriches our understanding of these major ideas.

Probability in decision making

Let's imagine that the owner of a million-dollar beachfront home in Galveston, Texas approaches you with a financial proposition. "Every month for the next year," he begins, "I'll mail you a check

for $1000. But in return," he continues, "you agree to repair my house, up to its full value, if it suffers any damage from a hurricane."

Would you agree to his terms? Before you decide, you'll probably want to consider at least three questions:

1. *How likely is it that a hurricane will strike Galveston over the next 365 days?* If none do, you'll be $12,000 richer. If one does, you might lose a bundle.

2. *If a hurricane strikes, how much damage should you expect?* You can probably find data on past insurance claims that were filed in the wake of hurricanes. In interpreting this data, you'd want to know some basic facts about the beachfront home in question. Is it at ground level, or high up on stilts? How is the roof anchored to the walls? What are the home's physical geometry and aerodynamic properties? Do the windows have shatter-proof glass? And so on. (You'd also want to run a regression if you hope to make use of these facts, just like you've already learned to do.)

3. *How comfortable are you in accepting sure, steady gains, against the small probability of a catastrophic loss?* Even if you decide that the trade is mathematically in your favor, you recognize that the two sides pose hugely asymmetric risks and rewards. This prospect might not sit well with you. One is reminded of the (possibly apocryphal) conversation between two famous cotton traders of the pre-Depression era, recounted by author Edwin Lefèvre:[1]

   A: I can't sleep.
   B: Why not?
   A: I am carrying so much cotton that I can't sleep thinking about it. It is wearing me out. What can I do?
   B: Sell down to the sleeping point.

[1] *Reminiscinces of a Stock Operator*, Edwin Lefèvre, 1923. The book is out of copyright, and many copies are floating around on the web.

Not coincidentally, these three questions correspond to the three broad themes of this book—probability, statistical inference, and decision analysis.

Of course, most of us don't have a million dollars lying around, meaning that we're unlikely to be offered such a trade by any Galvestonians anytime soon. But insurance companies enter into these kinds of deals every day. So, too, do most of the 75 million homeowners in America, who take the opposite end of the trade.

And even if you're not on the market for insurance, you probably still face plenty of other decisions where versions of these three questions are relevant:

- State school or Ivy?
- Job or graduate school?
- Stocks, bonds, or savings account?
- Buy a house, or rent?
- Hire a plumber, or do it yourself?
- Comprehensive car insurance, or liability only?
- High deductible or high premium?

For decisions like these, knowing a bit about probability—and a bit about yourself—is essential. After all, in speaking about the stock market, Lefèvre's fictional character Larry Livingston might just as easily been describing any major life decision:[2]

> All stock-market mistakes wound you in two tender spots—your pocketbook and your vanity. . . . Of course, if a man is both wise and lucky, he will not make the same mistake twice. But he will make any one of ten thousand brothers or cousins of the original. The Mistake family is so large that there is always one of them around when you want to see what you can do in the fool-play line.

[2] *ibid.*

What is probability?

All attempts to quantify risk begin with probability. We'll explore three different definitions of this surprisingly subtle concept: the axiomatic, the frequentist, and the subjectivist definitions. These are the three different answers you'd get, respectively, if you posed the question "What is probability?" to a mathematician, a casino owner, and a stock-market investor.

Figure 7.1: What makes a die "fair"? Try formulating answers under both frequentist and subjectivist interpretations of probability.

First, there is the *axiomatic* approach to defining probability. This is the mathematician's definition, and the least satisfying of the three, so let's get it out of the way first. According to the axiomatic approach, the probability of an event A, denoted $P(A)$, is just a property that satisfies three *axioms*, or assumptions whose truth is taken for granted:

1. $0 \leq P(A) \leq 1$ for any event A. In words: all probabilities are numbers between zero and one.

2. If Ω is a certainty, then $P(\Omega) = 1$. In words: events that are certain to occur have probability one.

The symbol Ω is the capital Greek letter "Omega," and is mathematical shorthand for the union of all possible events that could occur in a given context. The notation "$A \cup B$" means that either A or B occurs.

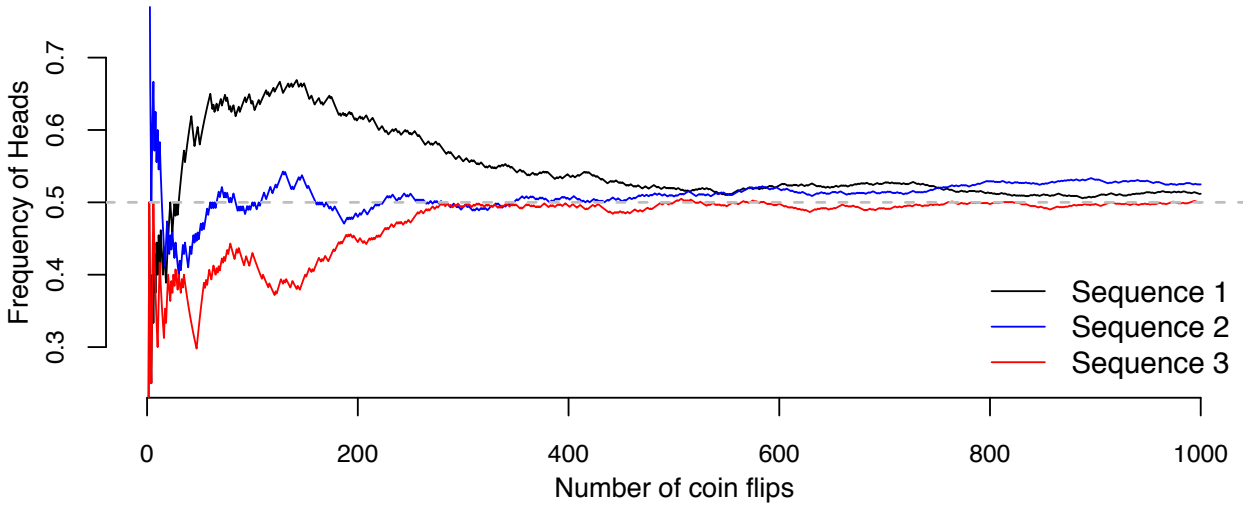


Figure 7.2: Three simulated sequences of 1000 coin flips each. The observed frequency of "heads" varies quite a bit in the beginning, but will eventually settle down near 0.5 if you keep flipping the coin.

3. If A and B cannot both happen, then $P(A \cup B) = P(A) + P(B)$. In words: the probabilities for mutually exclusive events add together.

These are called Kolmogorov's axioms, after the Russian mathematician of the same name.

Why do we call this approach unsatisfying? There is, of course, the perfectly reasonable question of why we should adopt these particular axioms. (One is always free to ask this question about axioms!) But the real issue is that, as a practical matter, the axioms are not very useful. If you are interested in actually assigning probabilities to real-world events—like the probability of rain tomorrow or a hurricane this year—then Kolmogorov is silent, and you're on your own.

Next, there's the *frequentist* interpretation of probability—or as we prefer to think of it, the casino owner's interpretation. Under the frequentist view, $P(A)$ is the long-run frequency with which the event A happens under repeated sampling. Think Las Vegas: dice, cards, roulette wheels, and so forth. A "fair coin" is the archetypal example in probability, and makes the frequency interpretation easy to understand: the coin is fair if, when we keep flipping it over and over again, the limiting frequency of heads in our sample is 50%, like in the figure above. If instead we get heads 40% of the time, or perhaps 70% of the time, then the probability of heads is 0.40 or 0.70, respectively.

The frequentist interpretation of probability goes hand-in-hand with a commonly used (and misused) phrase, the "law of large numbers." The law of large numbers states that, if you were to perform an experiment (like the spinning of a roulette wheel) a very large number of times, the average frequency of a result (like the fraction of times the spin comes up red) will be very close to expected frequency calculated from probability theory. Casino owners depend on the law of large numbers to pay the electricity bill!

The frequentist interpretation of probability sounds appealingly objective, in that it merely involves counting and dividing to compute a frequency. In many practical situations, however, this apparent objectivity is elusive. For example, suppose you want to interpret the following statement:

> Someone diagnosed with colorectal cancer has a 63% probability of surviving at least five years from the initial diagnosis.[3]

[3] www.lungcanceralliance.org/facing/risks.html

We could certainly imagine a thought experiment in which an enormous cohort of patients are all diagnosed with colon cancer on the same day, and are all tracked for the next five years. We could then interpret the survival frequency of 63% as the five-year probability of survival.

Yet what are we to make of the fact that all patients are different, and in very important ways: age, sex, history of smoking or drug use, lifestyle, and genes, to name but a handful? For most real situations, these differences make the notion of "repeated events" somewhat philosophically problematic. At the very least, a frequentist probability statement requires some kind of subjective judgment along these lines in order to be meaningful. In this case, we must judge whether all the patients in the cohort are similar enough for their individual case histories to count as repeated measurements of the same event.

Finally, there's the *subjectivist* interpretation of probability.[4] Under the subjectivist interpretation, $P(A)$ is your degree of belief, expressed on a zero-to-one scale, that A will happen. Put another way, it is your subjective assessment of the "fair value" of a $1 bet that event A will happen. Here, imagine a stock-market investor who has to decide whether to take a "long" or a "short" position—that is, whether to buy stocks, or to sell them short. Neither Kolmogorov nor the casino analogy can help here, for the performance of the stock market over the subsequent days, months,

[4] This is often called the Bayesian interpretation of probability, though not everyone who uses Bayesian statistics actually subscribes to it.

and years is nothing like a series of roullette games. Rather, it involves a unique sequence of one-off events, never to be repeated. Such a situation requires two things: that the investor make a subjective assessment of the probability that stocks will go up, and that he or she backs this assessment with real money.

To those used to thinking of probability as a long-run frequency—as in, "If I flip the coin a million times, half of the flips will come up heads, and so the probability of getting heads is 50%"—this notion can seem crazy. But it's worth remembering that the frequency interpretation and the subjectivist interpretation of probability are not mutually exclusive. Remember, they are only different interpretations of the same underlying mathematical construction. Think of them as different pairs of glasses, like sunglasses and reading glasses: sometimes it's easier to see something using one pair, and sometimes the other.

We'll return to the frequency interpretation of probability later in the course. In fact, we'll depend heavily upon it to understand the alphabet soup of summaries returned by most statistical software packages. But for now, it is the subjectivist interpretation that can help us understand the answer to a very basic question: where do the "axioms" of probability come from?

Probability as degree-of-belief

"Probability does not exist."

Or so quipped Bruno de Finetti, a famous Italian mathematician of the mid-twentieth century. What he meant was roughly the following: probability is not a physical property of an object, like mass or electric charge, sitting out there in the world waiting to be measured by some intrepid scientist. Rather, it is a psychological property of you as a decision maker, like a dream or a memory. It is, in a word, subjective.

Let's see if we can make further sense of this idea by starting somewhere simple: your own preferences. Consider two options, which we'll label arbitrarily as A and B. We'll let A represent a "sure money" option, wherein someone gives you cash (say, $10); and we'll let B represent some other option (say, lunch at your favorite cafe). The foundation of the subjectivist theory of probability is the simple idea that you can articulate a choice between A and B. If you prefer A, then we write $A \succ B$; the $\succ$ symbol is

called your *preference relation.* Similarly, if you prefer B to A, we write $B \succ A$. If you are indifferent between the two options, we write $A \equiv B$.

Examine the following list of choices. Do you prefer A to B ($A \succ B$), B to A ($B \succ A$), or neither ($A \equiv B$)?

A: sure money	or	B: something else of value
\$1	or	\$10
\$10	or	UT wins a national championship in football
\$100	or	one month of unlimited talk, data, text, and MMS on an iPhone
\$30,000	or	a round-trip first-class ticket from Austin to Sydney over spring break
\$600,000	or	one standard 12.4kg bar of 99.99% gold bullion
\$3,000,000,000	or	a 1/100 ownership stake in Facebook

Easy enough. But things get a little more interesting when we contemplate making choices involving uncertain outcomes. First, two short working definitions: for our purposes here, a *contract* on some event X is an agreement stipulating that you will be paid \$100 if X happens, and \$0 otherwise. A *bet* is an exchange of money for a contract. (Obviously these are more general notions.)

With that in mind, for each of the contracts in the table on the following page, which side of the bet do you like better? In contemplating your preferences, think like a gambler and try to maximize your return. In other words, suppress any personal, nonfinancial interest you might have in one side versus the other.

A: sure money	or	B: a contract that pays you \$100 if . . .
\$30	or	It rains tomorrow.
\$8	or	The European Union has fewer members in 2 years than it does now.
\$23	or	The Republican nominee wins the next U.S. presidental election.
p	or	Your cell phone stops working at any time over the next 6 months.

This last entry is the interesting one: notice that the value of the sure money option, p, is left unspecified. How would changing the value of p affect your choice? Presumably if $p = \$0.01$ you'd take the contract ($B \succ A$), while if $p = \$99$ you'd take the cash ($A \succ B$).

The crucial question is: *for what value of p are you precisely indifferent between the cash and the contract?* This number quantifies your degree of belief in the proposition that your cell phone will stop working over the next six months. According to the subjectivist interpretation—and here's the philosophical leap—this "degree of

belief" is actually a probability. (Make that *your* probability.) Once you've set this probability, then the bet is deemed *fair* to you, and you ought to be indifferent as to which side you find yourself on.

Can people disagree about probabilities? Of course they can. In fact, at the heart of any interesting gamble is a disagreement about probability. One person thinks the probability of the Cowboys' winning the Super Bowl is p, and another person thinks it's q. It is then possible for the two of them to enter into a bet where both sides think they're getting a good deal.

Let's try a few more. For what values of p_1, p_2, p_3, and p_4 are you indifferent between the cash and contract?

A: sure money		B: a contract that pays you if . . .
p_1	or	1) The Democrats have control of the Senate after the next election.
p_2	or	2) The Houston Astros win the next World Series.
p_3	or	3) Scotland is still a part of Great Britain in 10 years.
p_4	or	4) A private company sends paying customers into space before 12/31/2012.

If you find it too hard to pin down the actual numbers that quantify your preferences, try at least putting them in order from smallest to largest.

One small point worth mentioning here: in considering these bets, we're assuming fairly modest sums of money. This way, we don't have to worry about something called *risk aversion* entering the picture. (More on that to come.) One way of operationalizing this assumption more formally is to stipulate the following: in setting your probabilities, you must do so in such a way that you are indifferent between the two sides of the bet, *and* indifferent among the number of contracts that change hands. This ensures that it is only the probability of an event, and not the magnitude of the possible win or loss, that influences your choice.

Figure 7.3: Odds posted at Cheltenham, a famed British horseracing ground.

Odds and probabilities

Sometimes it is useful to express $P(A)$ as in the form of odds—such as, 9:2, 3:1, or more generally x:y. These numbers have the interpretation that, if you wager y dollars that the specified event will occur, you will win x of profit if it actually does (in addition to getting your original stake back). Odds are usually quoted as the "odds against A." For example, if the odds against a Cowboys' Super Bowl victory are quoted as "9 to 2 against," then the odds

are 4.5, which means a probability of about 0.18. This manner of expressing probabilities is quite common in, for example, horse racing.

How do we know that 9/2 corresponds to a probability of 0.18? The following simple conversion formulas are useful:

$$\text{Odds Against A} = \frac{1 - P(A)}{P(A)} \tag{7.1}$$

$$P(A) = \frac{1}{(\text{Odds Against A}) + 1}, \tag{7.2}$$

where "Odds Against A" is interpreted as a decimal number (e.g. odds of 9:2 are $9/2 = 4.5$).

Odds can also be quoted as "odds in favor," which are simply the reciprocal of "odds against." If the odds against A are bigger than 1, then $P(A) < 0.5$. If the odds are less than 1, then $P(A) > 0.5$. You should verify this for yourself using the above formulas.

Another useful interpretation is that the odds express the ratio of two probabilities: the probability that A won't happen, divided by the probability that A will happen. So if the odds against victory are 3:1, then it's three times more likely that you'll lose than win.

Kolmogorov's axioms, coherence, and the Dutch book theorem

Here's one question you might have for de Finetti: since my probabilities are subjective, am I free to set them however I choose?

The surprising answer is: no! Your probabilities may be subjective, but they still have to obey some rules. The remarkable thing is that these rules happen to be identical to Kolmogorov's "axioms." But they are not axioms at all; they are direct logical consequences of the deeper principle of *coherence*. Coherence can be given a precise mathematical definition, but the essence of the concept is: *you can't choose probabilities that are guaranteed to lose you money.*

To sharpen your intuition about the principle of coherence, it's helps to imagine a situation where a bookmaker offers you a series of bets on the outcome of an uncertain event. Your goal is to make a fool of the bookmaker by placing bets that will win you money, regardless of the outcome.

"Probabilities must sum to 1."

As an example, let's imagine a hypothetical best-of-three wrestling match between two famous mathematical thinkers: Èvariste Galois and Tycho Brahe. Owing to the personalities involved, this would surely be a feisty battle. Galois died at age 20 from wounds suffered in a duel related to a broken love affair. Brahe also dueled at age 20—allegedly over a mathematical theory—and lost the tip of his nose in the process. He died at age 56; scholars remain unsure whether he suffered kidney failure after a night of overtoxication at a royal banquet, or was murdered by someone who spiked his drinks with mercury.

In advance of the big match, your local betting house, Brokelads, posts the following bets. The shop is willing to buy or sell at the posted price.

A: sure money	or	B: a contract that pays you $100 if . . .
$35	or	Galois wins.
$55	or	Brahe wins.

Notice that if you buy both contracts, you'll have paid $90, but will receive $100 no matter who wins. This is an example of *arbitrage*, or riskless profit.

Another word for arbitrage is *Dutch book*—a somewhat archaic term that means having a portfolio (or "book") of bets with a bookmaker that guarantees one a profit.

What if, on the other hand, the posted bets were these?

A: sure money	or	B: a contract that pays you $100 if . . .
$45	or	Galois wins.
$60	or	Brahe wins.

Here's another arbitrage opportunity—if you sell both contracts, you'll get $105, and will only pay $100 back to the bookmaker, no matter who wins.

On the other hand, suppose these are the prices:

A: sure money	or	B: a contract that pays you $100 if . . .
$47	or	Galois wins.
$53	or	Brahe wins.

Try either of the tricks above, and you're guaranteed a profit of

exactly $0. You could, of course, buy just one contract or the other, in the hopes of making a nonzero profit. But then you're just gambling.

What's the difference between the first two cases, where arbitrage was easy, and the third case, where it was impossible? In the first two cases, the implied probabilities of the two events summed to something other than 100%. Not so in the third case, where $47 + 53 = 100$.

From this, it's not hard to see that Kolmogorov's second axiom—that the probabilities for mutually exclusive, collectively exhaustive events must sum to 1—need not be postulated as an axiom at all. Rather, it is logical requirement of principal of coherence: your subjective probabilities must avoid certain loss.

"Probabilities for disjoint events add together."

Now let's imagine the house posts a slightly more complicated menu of bets:

A: sure money	or	B: a contract that pays you $100 if . . .
$50	or	Galois wins.
$35	or	Galois wins in two rounds.
$20	or	Galois wins in three rounds.

Remember, this is a three-round wrestling match, and so the first to two wins (like in tennis). You are allowed to buy, or sell, as many of the contracts as you like. Can you assemble a portfolio of bets that will guarantee a profit?

Try this: buy the "Galois wins" contract at $50. Then sell the contracts for "Galois wins in two rounds" ($35) and "Galois wins in three rounds" ($20). These transactions should net you $5. If Brahe wins, all three of these contracts become worthless, and you walk away with your sure profit. And if Galois wins—regardless of whether it's a two or three-round victory—then you'll owe $100 on one contract, but receive $100 on another. Use the winnings from one to pay what you owe on the other, and you still walk away with a sure profit.

Where the did bookmaker go wrong here? The answer is: he violated Kolmogorov's third axiom, thereby opening the door to a wily arbitrageur like yourself. Notice that "Galois wins in two rounds" and "Galois wins in three rounds" are mutually

exclusive, and together are equivalent to the event "Galois wins" (since any Galois victory must be in either two rounds or three). But according to the bookie,

$$P(\text{Galois wins}) = 0.50\,,$$

while

$$P(\text{Galois wins in two}) + P(\text{Galois wins in three}) = 0.35 + 0.20 = 0.55\,.$$

Yet clearly,

$$P(\text{Galois wins}) = P(\text{Galois wins in two}) + P(\text{Galois wins in three})\,,$$

meaning that the house has assigned two different probabilities to the same event.

If the bookie were assigning coherent probabilities, then he'd have been obeying Kolmogorov's third axiom, which says that this kind of inconsistency shouldn't happen. The probabilities for mutually exclusive events A and B must add together: $P(A \cup B) = P(A) + P(B)$.

"Probabilities are numbers between 0 and 1."

We've now derived Kolmogorov's second and third axioms—but what of the first, that probabilities must be numbers between 0 and 1? This one is the easiest of all:

A: sure money	or	B: a contract that pays you \$100 if . . .
−\$10	or	Galois wins.
\$110	or	Brahe wins.

This bookie probably wouldn't stay in business very long!

Beyond those three simple rules

WHERE DO WE stand now? We've provided a convincing elaboration of de Finetti's original point: that subjective probabilities must follow Kolmogorov's three axioms if they are to be coherent. Probabilities that don't satisfy these rules are at war with themselves. They lead to certain monetary loss in a hypothetical betting

We haven't given a formal mathematical derivation of this, merely an intuitive argument. But all that is needed is a little matrix algebra, and these claims can be proven fairly straightforwardly, giving us a result that has come to be known as the *Dutch Book theorem*.

situation where someone is allowed to buy or sell contracts at the prices implied by the stated probabilities.

But within these broad limits, individuals are free to make their own subjective choices. De Finetti makes this point nicely, if a bit stiffly:[5]

> [E]ach of these evaluations corresponds to a coherent opinion, to an opinion legitimate in itself, and every individual is free to adopt that one of these opinions which he prefers, or, to put it more plainly, that which he feels. The best example is that of a championship where the spectator attributes to each team a greater or smaller probability of winning according to his own judgment; the theory cannot reject *a priori* any of these judgments unless the sum of the probabilities attributed to each team is not equal to unity."

[5] "Foresight: its logical laws, its subjective sources" (1937). Quoted on page 22 of *Decision Theory*, by Parmigiani and Inoue (Wiley, 2009).

Subject to the constraint of coherence, *vive le différence.*

The rules of addition and multiplication

In the discussion of coherence and Kolmogorov's axioms, you may have noticed that we've left out two commonly used rules of probability, the addition and multiplication rules.

Addition rule: The probability that either A or B will happen is

$$P(A \cup B) = P(A) + P(B) - P(A, B)\,, \tag{7.3}$$

where $P(A, B)$ is the probability that both A and B happen at once.

Multiplication rule: The joint probability that A and B will both happen is

$$P(A, B) = P(A) \cdot P(B \mid A)\,, \tag{7.4}$$

where $P(B \mid A)$ is the conditional probability that B will happen, given that A happens.

Notice how the addition rule, which works for any events A and B, differs from Kolmogorov's third axiom, which makes the restrictive assumption that A and B are mutually exclusive. If A and B cannot both occur, then $P(A, B) = 0$ and we're back to the original rule.

These are perfectly valid rules for manipulating probabilities, and it is important that you know them. But where do they come from? Remember, we started this whole discussion of subjective

probability by trying to de-axiomatize Kolmogorov's three rules—that is, to show that they weren't axioms at all, but rather consequences of a deeper principle. Have we accomplished this, only to smuggle a new pair of axioms in through the back door?

A derivation of the addition rule

The answer is no. It turns out that the rule of addition can be easily derived from Kolmogorov's third rule, which it closely resembles. To see this, let's express the event $A \cup B$ in a slightly different, but equivalent way:

$$A \cup B = (A, \sim B) \text{ or } (\sim A, B) \text{ or } (A, B)\,.$$

The "$\sim A$" notation means that event A doesn't happen. This equation says that, in order for A or B to happen, one of the following must happen:

- A happens and B doesn't.
- B happens and A doesn't.
- Both A and B happen.

But these three events are mutually exclusive. Therefore, applying Kolmogorov's third rule, we can add the probabilities of the three events on the righthand side to get the probability of the event of the lefthand side:

$$P(A \cup B) = P(A, \sim B) + P(\sim A, B) + P(A, B)\,.$$

Now we do something that seems a bit silly at first: we add and subtract the same term from the righthand side. This doesn't change matters, so we still preserve the equation:

$$P(A \cup B) = P(A, \sim B) + P(\sim A, B) + P(A, B) + P(A, B) - P(A, B)\,.$$

Why would we just add 0 to the righthand side? Because now we can group terms in a different way and notice something interesting.

$$P(A \cup B) = \Big\{P(A, \sim B) + P(A, B)\Big\} + \Big\{P(\sim A, B) + P(A, B)\Big\} - P(A, B)\,. \tag{7.5}$$

The first term in braces is the sum of the probabilities for two mutually exclusive events:

- A occurs and B doesn't.
- A occurs and B occurs.

But together, these events are equivalent to the event "A occurs." Applying Kolmogorov's third rule again, we see that

$$\Big\{P(A, \sim B) + P(A, B)\Big\} = P(A)\,.$$

By a similar argument, it must also be the case that

$$\Big\{P(B, \sim A) + P(B, A)\Big\} = P(B)\,.$$

Put these two results together and substitute into the original equation (7.5), and the law of addition pops right out:

$$P(A \cup B) = P(A) + P(B) - P(A, B)\,. \tag{7.6}$$

The rule of addition is therefore a mathematical consequence of Kolmogorov's three rules. Since all coherent decision makers will choose probabilities that obey these three rules, then their probabilities must obey the rule of addition as well.

Bayes' rule and conditional probability

Unlike the addition rule, the multiplication rule cannot be derived directly from Kolmogorov's axioms. Instead, we will demonstrate its validity using the same kind of Dutch-book argument that gave us the original three rules proposed as axioms by Kolmogorov.

Before we get there, however, we will explore a profound and important result in probability that arises directly from the multiplication rule: Bayes' rule, the granddaddy of them all. Once we see why this consequence of the multiplication rule is so important, we'll return to the question of why the rule itself must be true.

Bayes' rule describes conditional probabilities. Suppose you examine the online book-purchase histories of two friends, and you see the following:

Bertrand	Pablo
Proof and Consequences	*A History of Non-representational Art*
Never at Rest: A Biography of Isaac Newton	*The Art and Architecture of Mesopotamia*
The Great Theorems of the 20th Century	*Michelangelo's David*

What sorts of books are you likely to buy these friends for their birthdays? In particular, who do you think is more likely to enjoy the newest best-seller, *Social Grace: A Mathematician Tries Really Hard*?

If you used the purchase histories to inform your judgment, then you'll feel right at home with conditional probability statements. That's because you understand that any probability assessment is conditional upon the data at hand, and can change when you learn new data.

Conditional probability statements are what doctors, judges, weather forecasters try to make every day of their lives, as they assimilate information to reach an informed assessment.

The probability that a patient complaining of chest pains has suffered a heart attack: Does the patient feel the pain radiating down his left side? What does his ECG look like? Does his blood test reveal elevated levels of myoglobin?

The probability of rain this afternoon in Austin: What are the current temperature and barometric pressure? What does the radar show? Was it raining this morning in Dallas?

The probability that a person on trial is actually guilty: Did the accused have a motive? Means? Opportunity? Was any biological evidence left at the scene—maybe a bloody glove—that reveals a likely DNA match?

Probabilities simply cannot be isolated from the world. They are always contingent upon what we know. When our knowledge changes, they too must change.

But how? Suppose we start with a subjective probability assessment, such as $P(A) = 0.99$ for the event A: "the next engine off the assembly line will pass inspection." We might have arrived at this judgment, for example, by calculating the rate at which engines off the same assembly line passed inspection over the previous month. (In the absence of any other information, this is surely as good a guess as any.)

Now suppose we're given some new information, such as

B_1: the assembly-line crew has been working the last 7 hours straight, and keeps eyeing the clock; or

B_2: The lab that performs stress tests has been complaining about the quality of the recent shipment of steel from Shenzhen; or

B_3: The last 10 engines off the assembly line all failed inspection.

What is $P(A \mid B_1)$? What about $P(A \mid B_2, B_3)$? How should we incorporate this new information into our subjective probability assessment to arrive at a new, updated probability?

Figure 7.4: Bayes' rule is named after Thomas Bayes (above), an English reverend of the 18th century who first derived the result. It was published posthumously in 1763 in "An Essay towards solving a Problem in the Doctrine of Chances."

This process of learning requires Bayes' rule: an equation that transforms a *prior* probability $P(A)$ into a *posterior* probability $P(A \mid B)$. Bayes' rule can be derived straightforwardly from the multiplication rule:

$$P(A, B) = P(A) \cdot P(B \mid A)\,,$$

where $P(B \mid A)$ is the conditional probability that B will happen, given that A happens. Notice that this could equally well be written

$$P(A, B) = P(B) \cdot P(A \mid B)\,.$$

If we equate the two righthand sides, we see that

$$P(B) \cdot P(A \mid B) = P(A) \cdot P(B \mid A)\,,$$

Simply divide through by $P(B)$, and we arrive at Bayes' rule:

$$P(A \mid B) = \frac{P(A) \cdot P(B \mid A)}{P(B)}\,. \tag{7.7}$$

Each piece of this equation has a name:

- $P(A \mid B)$ is the posterior probability: how probable is A, now that we've seen data B?
- $P(A)$ is the prior probability: how probable is A, before ever having seen data B?
- $P(B \mid A)$ is the likelihood: if A were true, how likely is it that we'd see data B?
- $P(B)$ is the marginal probability of B: how likely is it that we'd see data B anyway, regardless of whether A is true or not? To calculate this quantity, it helps to remember the addition rule:

$$P(B) = P(B \mid A) \cdot P(A) + P(B \mid \sim A) \cdot P(\sim A)\,.$$

The base-rate fallacy: an example of Bayes' rule in action

The rules of probability theory are not simply mathematical curiosities to be manipulated on the page. They are very useful in explaining properties of the real world, and Bayes' rule is the most useful of all.

Suppose you're serving on a jury in the city of New York, with a population of roughly 10 million people. A man stands before

you accused of murder, and you are asked to judge whether he is guilty (G) or not guilty ($\sim G$). In his opening remarks, the prosecutor tells you that the defendant has been arrested on the strength of a single, overwhelming piece of evidence: that his DNA matched a sample of DNA taken from the scene of the crime. Let's call denote this evidence by the letter D. To convince you of the strength of this evidence, the prosecutor calls a forensic scientist to the stand, who testifies that the probability that an innocent person's DNA would match the sample found at the crime scene is only one in a million. The prosecution then rests its case.

Would you vote to convict this man?

If you answered "yes," you might want to reconsider! You are charged with assessing $P(G \mid D)$—that is, the probability that the defendant is guilty, given the information that his DNA matched the sample taken from the scene. Bayes' rule tells us that

$$P(G \mid D) = \frac{P(G) \cdot P(D \mid G)}{P(D)} = \frac{P(G) \cdot P(D \mid G)}{P(D \mid G) \cdot P(G) + P(D \mid \sim G) P(\sim G)} .$$

We know the following quantities:

- The prior probability of guilt, $P(G)$, is about one in 10 million. New York City has 10 million people, and one of them committed the crime.
- The probability of a false match, $P(D \mid \sim G)$, is one in a million, because the forensic scientist testied to this fact.

To use Bayes' rule, let's make one additional assumption: that the likelihood, $P(D \mid G)$, is equal to 1. This means we're assuming that, if the accused were guilty, there is a 100% chance of seeing a positive result from the DNA test.

Let's plug these numbers into Bayes' rule and see what we get:

$$\begin{aligned} P(G \mid D) &= \frac{\frac{1}{10{,}000{,}000} \cdot 1}{1 \cdot \frac{1}{10{,}000{,}000} + \frac{1}{1{,}000{,}000} \cdot \frac{9{,}999{,}999}{10{,}000{,}000}} \\ &\approx 0.09 . \end{aligned}$$

The probability of guilt looks to be only 9%! This result seems shocking in light of the forensic scientist's claim that $P(D \mid \sim G)$ is so small: a "one in a million chance" of a positive match for an innocent person. Yet the prior probability of guilt is very low—$P(G)$ is a mere one in 10 million—and so even very strong evidence still only gets us up to $P(G \mid D) = 0.09$.

Conflating $P(\sim G \mid D)$ with $P(D \mid \sim G)$ is so common that it has an informal name: the prosecutor's fallacy,[6] or more generally the base-rate fallacy.[7] Words may mislead, but Bayes' rule never does!

An alternate way of thinking about this result is the following. Of the 10 million innocent people in New York, ten would have DNA matches merely by chance. The one guilty person would also have a DNA match. Hence there are 11 people with a DNA match, only one of whom is guilty, and so $P(G \mid D) \approx 1/11$.

[6] en.wikipedia.org/wiki/Prosecutor's_fallacy

[7] en.wikipedia.org/wiki/Base_rate_fallacy

Understanding Bayes' rule using trees

Let's try a second example. You may have encountered the following brain-teaser, which is the frequent subject of "first caller with the right answer wins a prize"-style contests on the radio. Here's how one San Francisco radio station described the puzzle:

> There are two tribes in the jungle. The truth tellers always tell the truth and the liars always lie. A scientist comes to a fork in the road. He knows that the truth tellers' tribe is in one direction and the liars' tribe is in the other direction. But he does not know which direction is the truth tellers' tribe. There is a native sitting at the intersection of the roads. The scientist does not know whether this native is a truth teller or a liar. The scientist may ask the native one and only one question to determine the direction to the truth tellers' tribe. What question should he ask?

The scientist should ask "Which way to your village?" (Ponder for a minute why this is guaranteed to be informative.)

To illustrate Bayes' theorem, let's amend this example to include probabilities, and to adjust for the fact that it's not really "PC" to talk about lying tribes of jungle natives anymore. Suppose you face the following situation:

> You are driving through unfamiliar territory in East Texas in your burnt-orange car sporting a bumper sticker from the University of Texas. You reach a fork in the road. In one direction lies College Station; in another direction, Austin. The road sign pointing to Austin has been stolen, but you see a man selling watermelons out of his pickup truck. You pull over and ask him for directions.
>
> You know that there are two kinds of people in this part of Texas, Longhorns and Aggies, with Aggies outnumbering Longhorns by a 60/40 margin. But you don't know which one this man is. If he's a Longhorn, he is sure to help you out to the best of his ability, and you judge that there is only a 5% chance that he will get confused and point you in the wrong direction. But you believe that, if he is an Aggie and you ask him for directions, there is a 70% chance that he will see the bumper sticker on your car and send you the opposite way.
>
> Having once won a pair of concert tickets by solving a similar brain teaser posed by a radio station, you decide to ask him "Which way is your university?" He stares for a moment at your bumper sticker, then smiles and points to the left. You go in the direction indicated, and two hours later you arrive in Austin.

Given that you ended up in Austin, what is the probability that the man you encountered was a Longhorn?

One way of solving this is to use Bayes' rule directly:

$$\begin{aligned} P(\text{Longhorn} \mid \text{pointed to Austin}) &= \frac{P(\text{Longhorn}) \cdot P(\text{pointed to Austin} \mid \text{Longhorn})}{P(\text{pointed to Austin})} \\ &= \frac{0.4 \cdot 0.95}{0.6 \cdot 0.7 + 0.4 \cdot 0.95} \\ &= 0.475\,. \end{aligned}$$

There is slightly better than an even chance that you were talking to an Aggie.

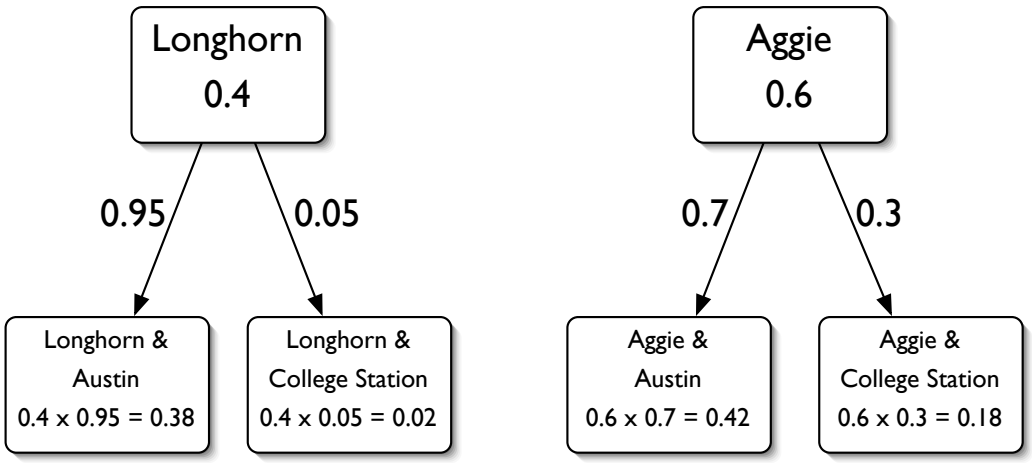
So Bayes' rule gets use the answer with little fuss, assuming you're happy with the "plug and chug" approach. But an alternative, very intuitive way of solving this problem—and of understanding Bayes' theorem more generally—is to use a *tree*. First, start by listing the possible states of the world, along with their probabilities:

You might even make the size of the boxes correspond to the probabilities, though we've avoided this embellishment here.

Next, draw arrows from each state of the world to the possible observational consequences. Along the arrows, put the conditional probabilities that you will observe each data point, given the corresponding state of the world:

At the terminal leaves of the tree, multiply out the probabilities according to the multiplication rule: $P(A, B) = P(A) \cdot P(B \mid A)$. So, for example, the probability that the man is an Aggie and that he points you to Austin is $0.7 \times 0.6 = 0.42$. The sum of all the probabilities in the leaves of the tree must be 1; this is a mutually exclusive, exhaustive partition of all possible states of the world.

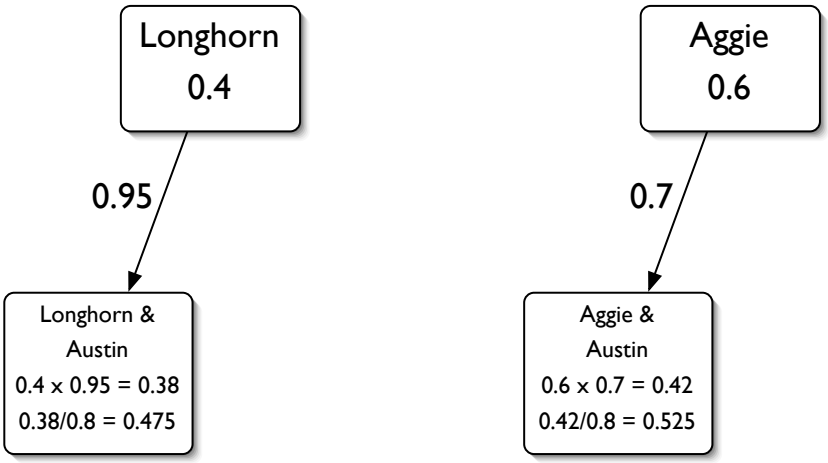
But now that you've arrived back home, you know that the man was pointing to Austin. In light of this data, how you can use the tree to compute the probability that he was a Longhorn?



The answer: simply cut off all the branches corresponding to data that wasn't actually observed, leaving only the actual data:

The remaining leaf probabilities are proportional to the posterior probabilities of the corresponding states of the world. These probabilities, of course, do not sum to 1 anymore. But this is easily fixed: simply divide each probability by the sum of all remaining probabilities (in this case, $0.38 + 0.42 = 0.8$):

As before, we find that there is a probability of 0.475 that the man at the watermelon stand was a Longhorn. Just so you can



keep things straight, the sum of 0.8, by which you divided the terminal-leaf probabilities in the final step to ensure that they summed to 1, corresponds exactly to the denominator, P(points to Austin), in Bayes' rule.

Coherent conditional probabilities

So the importance of the multiplication rule is that it leads us directly to Bayes' rule. Now, at last, we return to the question of why you should accept the multiplication rule in the first place. This question can also be answered with a Dutch-book style argument—that is, by imagining a hypothetical bookmaker whose probabilities fail to respect the multiplication rule, and by showing that it is possible to own a portfolio of bets that make him a sure loser.

We'll start, as before, with a definition: a *a conditional bet* on event A, given event B, is an agreement with the following terms:

- If event B happens, then you agree to pay \$$p$ for an ordinary contract on event A (which, as before, pays \$100 if event A happens).
- If event B fails to happen, then the bet is called off and no money changes hands.

The price p is agreed upon before the parties know whether B happens. If you pay \$$p$ and receive the corresponding guarantee, you are buying the contract; if you're receive cash and issue the guarantee, you are selling the contract. How should such a

contract be priced?

Let's go back to our imaginary three-round wrestling match between Galois and Brahe. Suppose a bookie posts the following prices:

A:	or	B:
\$30	or	A: Brahe wins the match.
\$40	or	B: Brahe wins the first round.
\$25	or	A, B: Brahe wins the first round and the match.
\$50	or	$A\|B$: a conditional contract that Brahe wins the match, given that he wins the first round.

He will buy or sell any number of contracts at the stated prices. Can you find a portfolio of bets that will give you a sure profit?

Try this: buy 1 contract on B, and sell 2 contracts of A, B. Then enter into 2 conditional contracts on $A|B$. You'll net \$10 from your trading, and:

- If Brahe loses the first round, all the contracts are worthless, no money changes hands on the conditional contract, and you keep your \$10.

- If Brahe wins the first round and loses the match, you'll win \$100 on your B contract, and lose two lots of \$50 on the conditional contracts you bought. Your net profit is, again, \$10.

- If Brahe wins the first round and wins the match, you'll win \$100 on your B contract, and win two lots of \$50 on the conditional contracts you bought. You can use this money to pay off the \$200 you owe on the A, B contracts you sold. Your net profit is, again, \$10.

The problem is that the bookie violated the product rule. His prices implied that $P(A, B) = 0.25$, but that $P(B) \cdot P(A \mid B) = 0.4 \cdot 0.5 = 0.2$. These two quantities must be equal to avoid arbitrage.

Random variables and probability distributions

The past several pages have given you a new appreciation for the rules of probability: where they come from, what they mean, and how they can be used. With these tools in hand, we are now ready to define a few core concepts that will anchor the rest of our discussion of statistical inference and decision analysis.

A *random variable* X is any uncertain quantity—the result of a coin flip, for example, or the price of a particular stock next month. A *probability distribution* is a list of all possible outcomes for that random variable, along with their probabilities. For example, suppose you own a special six-sided die that comes up with an even number twice as often as it comes up odd. If you were to roll the die—say, in an effort to win a rigged bet with someone who though the die was fair—then the random variable would be the number showing when it stopped rolling, and the probability distribution would be:

Value, x	Probability, $P(X = x)$
1	1/9
2	2/9
3	1/9
4	2/9
5	1/9
6	2/9

Check for yourself that the probabilities sum to 1, as they must. Note how we use big X to denote the random variable, and little x to denote the possible outcomes of the random variable. We call a table like the one above a *probability mass function*, or PMF: for any possible outcome x, the probability mass function $P(X = x)$ tells you how likely the random variable X is to take the value x.[8]

[8] Technically, we can only specify a PMF for a discrete random variable, whose possible outcomes can be counted on your fingers and toes. To specify the probability distribution of a continuous random variable, which can take on any of a continuous range of possible outcomes, we need to use something called a probability density function, or PDF. More on that later.

Moments: summarizing joint variation for random variables

You have likely already encountered two ways of a summarizing a single random variable: by quoting its expected value and its variance (or its standard deviation). These quantities describe two important features—the center and dispersion, respectively—of a probability distribution. The *expected value* of a probability distribution is a probability-weighted average of the possible values

of the random variable. Formally, if the random variable has N possible outcomes $\{x_1, \ldots, x_N\}$ having corresponding probabilities $\{p_1, \ldots, p_N\}$, then the expected value is

$$E(X) = \sum_{i=1}^{N} p_i x_i \,.$$

The expected value is a measure of the "center" of a probability distribution. The expected value can also be referred to as the *mean*, as long as this is carefully distinguished from the sample mean.

Next, the *variance* of a probability distribution is a measure of dispersion. It is "the expected squared deviation from the expected value", or

$$\text{var}(X) = \text{E}\big(\{X - \text{E}(X)\}^2\big) \,.$$

The standard deviation of a probability distribution is $\sigma = \text{sd}(X) = \sqrt{\text{var}(X)}$. The standard deviation is more interpretable than the variance, because it has the same units (dollars, miles, etc.) as the random variable itself.

The mean and the variance are both examples of *moments*, which is a term borrowed from physics. Moments summarize the shape of a probability distribution—its center, its spread, and so forth. *Statistics*, on the other hand, summarize the shape of a sample. This conceptual distinction is easily forgotten when doing data analysis: for every moment, there is a corresponding statistic with a very similar definition, and this can create confusion. It's therefore worth repeating: moments describe *probability distributions*, while statistics describe *samples*. This chapter is about moments of random variables, not statistics.

The familiar concepts of mean and variance will still be useful, but no longer sufficient, when two or more variables are in play. In this sense, a quantitative relationship is much like a human relationship: you can't describe one by simply listing off facts about the characters involved. You may know that Homer likes donuts, works at the Springfield Nuclear Power Plant, and is fundamentally decent despite being crude, obese, and incompetent. Likewise, you may know that Marge wears her hair in a beehive, despises the *Itchy and Scratchy Show*, and takes an active interest in the local schools. Yet these facts alone tell you little about their marriage. A quantitative relationship is the same way: if you ignore the interactions of the "characters," or individual variables involved, then you will miss the best part of the story.

Moments of one variable

The expected value and variance are just two examples of a general concept called a *moment*. Informally, a moment is a quantitative description of the geometry or shape of a probability distribution. The term is borrowed from physics, where it comes up in many contexts: magnetic moment, electric-dipole moment, moment of inertia, and so forth.

The idea of a moment is best understood with the help of an ice-skating bat spinning in place. By changing the placement of his wings, the bat changes his *moment of inertia*, a one-number summary of how his body mass is distributed in space relative to his axis of rotation. Because angular momentum is conserved, the bat's moment of inertia is inversely related to his rotational velocity. This physical law is exploited to dramatic visual effect in competition, where ice skaters often end a routine by drawing their arms in closer and closer, thereby spinning faster and faster.

Other examples of a moment in physics follow the same conceptual pattern as the moment of inertia: each is a concise geometric description of a body or physical system. The term itself is etymologically related to "momentum;" according to the *Oxford English Dictionary*, it seems to have first been used in this sense in Kater and Lardner's 1830 *Treatise on Mechanics*.[9]

By metaphorical extension, a moment in probability theory is a geometric description of your uncertainty about a random variable. The mean of a random variable X describes where the probability distribution for X is centered; metaphorically, it is like the skater's center of gravity. The variance describes how dispersed the distribution is around its center; metaphorically, it is like the spread of the skater's arms.

Moments in physics have precise mathematical definitions.

[9] Although Milton came close in 1641: "All the moments and turnings of humane occasions are mov'd to and fro as upon the axle of discipline." (*The reason of church-governement urg'd against prelaty*, X.135).

For example, if a system comprises n different masses $m_1, \ldots, m_n$ placed at radiuses $r_1, \ldots, r_n$ around a common rotational axis, then the system's moment of inertia is

$$I = \sum_{i=1}^{n} m_i r_i^2 .$$

In staring at any formula like this, the important question to ask yourself is: *how does the math formalize the intuition?* Here, it does so straightforwardly. The further the objects are from the center, the larger the values of r_i, and therefore the larger the moment of inertia. The quantity I is a summary of the geometric distribution of mass in the system. It doesn't tell you everything about the system, but it does tell you something useful: does the mass tend to concentrate near the axis of rotation, or far away from it?

Moments in probability theory also have precise mathematical definitions. Suppose that X is a discrete random variable that takes on values $x_1, \ldots, x_n$ with probabilities $p_1, \ldots, p_n$, respectively. The kth moment of X is defined as the expected value of the kth power of X, or

$$E(X^k) = \sum_{i=1}^{n} p_i x_i^k .$$

For continuous random variables, there is calculus-based version of the formula:

$$E(X^k) = \int_{\Omega} x^k p(x)\, dx ,$$

where $p(x)$ is the *probability density function* (or p.d.f.) of the random variable X, and Ω is the space of all possible values that X might take on. If you compute the area between two points under the curve of the probability density function, you will get the probability that the random variable will take on a value between those two points.

There is an striking correspondence between this and the formula for the moment of inertia: the probabilities p_i are like the masses m_i, while the values x_i are like the radiuses r_i. In fact, the analogy with physical mass is so instructive that, when we describe a probability distribution by listing the possible values x_i together with their probabilities p_i, we are said to be specifying the *probability mass function* of the distribution.

Likewise, the kth *central moment* of X is

$$E\big[\{X - E(X)\}^k\big] = \sum_{i=1}^{n} p_i \{x_i - E(X)\}^k .$$

The mean of a probability distribution is its first moment; the variance is its second central moment. Higher-order moments also have geometric interpretations. For example, a probability distribution's skewness (or lopsidedness) is measured by the third moment, while its tail weight (or propensity to produce extreme events) is measured by the fourth moment.

The following two points about moments are worth remembering.

1. Moments are merely summaries. Two probability distributions can have the same mean and the same variance, and yet be very different. See Figure 7.5. With few exceptions, the only way to perfectly characterize an entire distribution is to quote the probability mass function—or, for a continuous random variable, the probability density function.

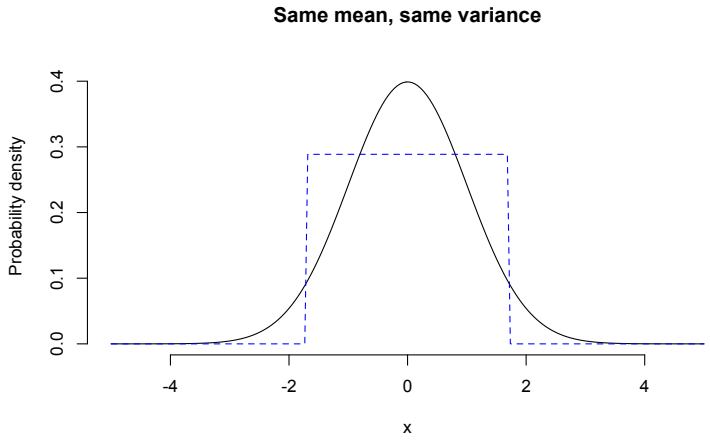


Figure 7.5: Two probability distributions with the same mean and same variance.

2. Like everything in mathematics, the definition of a moment is just a human convention, agreed upon by a body of working scientists and statisticians. There is nothing holy about this definition; it just happens to be one that conveys information which people find useful.

Joint distributions and covariance

A moment summarizes the probability distribution for one variable. To summarize relationships between more than one variable, we will appeal to the concept of a *mixed moment*, which summarizes a joint distribution.

A *joint distribution* is an exhaustive list of joint outcomes for two or more variables at once, together with the probabilities for each of these outcomes. For example, the table below depicts a simple, stylized joint distribution for the rain and average wind speed on a random day in February.

Outcome	Wind (mph)	Rain (inches)	Probability
1	5	1	0.4
2	5	3	0.1
3	15	1	0.1
4	15	3	0.4

In this simple case, each variable can take one of only two values, and so there are only four possible joint outcomes, whose probabilities must sum to 1. Notice, too, that the joint distribution depicts a positive relationship between wind and rain: when one is high, the other tends to be high as well.

To quantify this relationship, define the *covariance* of two random variables X and Y as

$$\text{cov}(X,Y) = E\Big\{[X - E(X)][Y - E(Y)]\Big\} = \sum_{i=1}^{n} p_i \big[x_i - E(X)\big]\big[y_i - E(Y)\big] .$$

This sum is over all possible joint outcomes for X and Y. In the wind/rain example, the expected values for wind speed (X) and rainfall (Y) are

$$\begin{aligned} E(X) &= \sum_{i=1}^{n} p_i x_i = 0.4 \cdot 5 + 0.1 \cdot 5 + 0.1 \cdot 15 + 0.4 \cdot 15 = 10 \\ E(Y) &= \sum_{i=1}^{n} p_i y_i = 0.4 \cdot 1 + 0.1 \cdot 1 + 0.1 \cdot 3 + 0.4 \cdot 3 = 2 \end{aligned}$$

Plugging these numbers into the formula for covariance, we get

$$\begin{aligned} \text{cov}(X,Y) &= E\Big\{[X - E(X)][Y - E(Y)]\Big\} \\ &= 0.4 \cdot (5-10)(1-2) + 0.1 \cdot (5-10)(3-2) + 0.1 \cdot (15-10)(1-2) + 0.4 \cdot (15-10)(3-2) \\ &= 0.4 \cdot (5) + 0.1 \cdot (-5) + 0.1 \cdot (-5) + 0.4 \cdot (5) \\ &= 3 . \end{aligned}$$

Again, ask yourself: *how does the mathematical definition of covariance formalize the intuition behind the concept of dependence?* Try reasoning through the formula, and its application to this example, on your own.

You may notice the following: in the third line of the above computation, the positive terms correspond to joint outcomes when wind speed and rainfall are on the *same side* of their respective means—that is, both above the mean, or both below it. The negative terms, on the other hand, correspond to outcomes where the two quantities are on *opposite sides* of their respective means. In this case, the "same side" outcomes are more likely than the "opposite side" outcomes, and therefore the covariance is positive.

Correlation as standardized covariance

The covariance is our first example of a mixed moment. It provides one way of quantifying the direction and magnitude of association between two random variables X and Y.

One difficulty that arises in interpreting covariance, however, is that it depends upon the scale of measurement for the two sets of observations. For example, suppose we measured rain in millimeters, rather than inches, as in the following table.

Outcome	Wind (mph)	Rain (mm)	Probability
1	5	25.4	0.4
2	5	76.2	0.1
3	15	25.4	0.1
4	15	76.2	0.4

Now $E(Y) = 50.8$, and the wind and rain variables have covariance

$$\begin{aligned} \text{cov}(X,Y) &= 0.4 \cdot (5-10)(-25.4) + 0.1 \cdot (5-10)(25.4) + 0.1 \cdot (15-10)(-25.4) + 0.4 \cdot (15-10)(25.4) \\ &= 76.2\,. \end{aligned}$$

This is 25.4 times as big as 3, the answer from before. And yet we wouldn't say that wind and rain are 25.4 times as "dependent" as they were before; the new numbers describe exactly the same probability distribution, just in different units. Clearly we need a measure of dependence that is invariant to changes in scale. (Interestingly, 25.4 is precisely the number of millimeters in a single inch, a fact which might suggest to you how covariances behave when you multiply one of the variables by a constant. More on that later.)

One such scale-invariant measure is Pearson's *product-moment correlation coefficient*, often called simply the correlation coefficient. (There are other kinds of correlation coefficients as well, and so sometimes we must distinguish them from one another.) The Pearson coefficient, named after English statistician Karl Pearson, is on a standardized scale running from -1 (perfect negative correlation) to $+1$ (perfect positive correlation).

The Pearson correlation coefficient for two random variables X and Y is just their covariance, rescaled by their respective variances:

$$\text{cor}(X,Y) = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)} \cdot \sqrt{\text{var}(Y)}}\,.$$

Let's apply this definition to joint distribution for wind speed and rain fall measured in inches:

$$\text{cor}(X, Y) = \frac{3}{\sqrt{25} \cdot \sqrt{1}} = 0.6\,.$$

And with rain measured in millimeters,

$$\text{cor}(X, Y) = \frac{76.2}{\sqrt{25} \cdot \sqrt{645.16}} = 0.6\,.$$

There is a common factor of 25.4 that appears in both the numerator and denominator. It cancels, leaving us with a scale-invariant quantity. If the correlation between two variables is 0, then they are said to be uncorrelated.[10]

[10] Although not necessarily independent!

Functions of random variables

A VERY IMPORTANT set of equations in probability theory describes what happens when you construct a new random variable as a linear combination of other random variables—that is, when

$$W = aX + bY + c$$

for some random variables X and Y and some constants a, b, and c.

The fundamental question here is: how does *joint* variation in X and Y (that is, correlation) influence the behavior a random variable formed by adding X and Y together? To jump straight to the point, it turns out that

$$E(W) = aE(X) + bE(Y) + c \tag{7.8}$$

$$\text{var}(W) = a^2\,\text{var}(X) + b^2\,\text{var}(Y) + 2ab\,\text{cov}(X, Y)\,. \tag{7.9}$$

Why would you care about a linear combination of random variables? Consider a few examples:

- You know the distribution for X, the number of points a basketball team will score in one quarter of play. Then the random variable describing the points the team will score in four quarters of play is $W = 4x$.
- A weather forecaster specifies a probability distribution for tomorrow's temperature in Celsius (a random variable, C). You can compute the moments of C, but you want to convert

to Fahrenheit (another random variable, F). Then F is also a random variable, and is a linear combination of the one you already know: $F = (9/5)C + 32$.

- You know the joint distribution describing your uncertainty as to the future prices of two stocks X and Y. A portfolio of stocks is a linear combination of the two; if you buy 100 shares of the first and 200 of the second, then

$$W = 100X + 200Y$$

is a random variable describing the value of your portfolio.

- Your future grade on the statistics midterm is X_1, and your future grade on the final is X_2. You describe your uncertainty for these two random variables with some joint distribution. If the midterm counts 40% and the final 60%, then your final course grade is the random variable

$$C = 0.4X_1 + 0.6X_2\,,$$

a linear combination of your midterm and final grades.

- The speed of Rafael Nadal's slice serve is a random variable S_1. The speed on his flat serve is S_2. If Rafa hits 70% slice serves, his opponent should anticipate a random service speed equal to $0.7S_1 + 0.3S_2$.

In all five cases, it is useful to express the moments of the new random variable in terms of the moments of the original ones. This saves you a lot of calculational headaches! We'll now go through the mathematics of deriving Equations (7.8) and (7.9).

Multiplying a random variable by a constant

Let's first examine what happens when you make a new random variable W by multiplying some other random variable X by a constant:

$$W = aX\,.$$

This expression means that, whenever $X = x$, we have $W = ax$. Therefore, if X takes on values $x_1, \ldots, x_n$ with probability $p_1, \ldots, p_n$, then we know that

$$E(X) = \sum_{i=1}^{n} x_i p_i\,,$$

and so

$$E(W) = \sum_{i=1}^{n} a x_i p_i = a \sum_{i=1}^{n} x_i p_i = aE(X)\,.$$

The constant a simply comes out in front of the original expected value. Mathematically speaking, this means that the expectation is a linear function of a random variable.

The variance of W can be calculated in the same way. By definition,

$$\text{var}(X) = \sum_{i=1}^{n} p_i \{x_i - E(X)\}^2\,.$$

Therefore,

$$\begin{aligned}
\text{var}(W) &= \sum_{i=1}^{n} p_i \{a x_i - E(W)\}^2 \\
&= \sum_{i=1}^{n} p_i \{a x_i - aE(X)\}^2 \\
&= \sum_{i=1}^{n} p_i a^2 \{x_i - E(X)\}^2 \\
&= a^2 \sum_{i=1}^{n} p_i \{x_i - E(X)\}^2 \\
&= a^2\, \text{var}(X)
\end{aligned}$$

Now we have a factor of a^2 out front.

What if, in addition to multiplying X by a constant a, we also add another constant c to the result? This would give us

$$W = aX + c\,.$$

To calculate the moments of this random variable, revisit the above derivations on your own, adding in a constant term of c where appropriate. You'll soon convince yourself that

$$\begin{aligned}
E(W) &= aE(X) + c \\
\text{var}(W) &= a^2 \text{var}(X)\,.
\end{aligned}$$

The constant simply gets added to the expected value, but doesn't change the variance at all.

A linear combination of two random variables

Suppose X and Y are two random variables, and we define a new random variable as $W = aX + bY$ for real numbers a and b. Then

$$\begin{aligned} E(W) &= \sum_{i=1}^{n} p_i\{ax_i + by_i\} \\ &= \sum_{i=1}^{n} p_i a x_i + \sum_{i=1}^{n} p_i b y_i \\ &= a\sum_{i=1}^{n} p_i x_i + b\sum_{i=1}^{n} p_i y_i \\ &= aE(X) + b(E(Y)\,. \end{aligned}$$

Again, the expectation operator is linear.

The variance of W, however, takes a bit more algebra:

$$\begin{aligned} \text{var}(W) &= \sum_{i=1}^{n} p_i\Big\{[ax_i + by_i] - [aE(X) + bE(Y)]\Big\}^2 \\ &= \sum_{i=1}^{n} p_i\Big\{[ax_i - aE(X)] + [by_i - bE(Y)]\Big\}^2 \\ &= \sum_{i=1}^{n} p_i\Big\{[ax_i - aE(X)]^2 + [by_i - bE(Y)]^2 + 2ab[x_i - E(X)][y_i - E(Y)]\Big\} \\ &= \sum_{i=1}^{n} p_i[ax_i - aE(X)]^2 + \sum_{i=1}^{n} p_i[by_i - bE(Y)]^2 + \sum_{i=1}^{n} p_i 2ab[x_i - E(X)][y_i - E(Y)] \\ &= \text{var}(aX) + \text{var}(bY) + 2ab\text{cov}(X,Y) \\ &= a^2\text{var}(X) + b^2\text{var}(Y) + 2ab\text{cov}(X,Y) \end{aligned}$$

The covariance of X and Y strongly influences the variance of their linear combination. If the covariance is positive, then the variance of the linear combination is *more than* the sum of the two individual variances. If the covariance is negative, then the variance of the linear combination is *less than* the sum of the two individual variances.

An example: portfolio choice under risk aversion

Let's revisit the portfolio-choice problem posed above. Say you plan to allocate half your money to one asset X, and the other half to some different asset Y. Look at Equations (7.8) and (7.9), which specify the expected value and variance of your portfolio in terms of the moments of the joint distribution for X and Y. If you are a risk-averse investor, would you prefer to hold two assets with a positive covariance or a negative covariance?

To make things concrete, let's imagine that the joint distribution for X and Y is given in the table at right. Each row is a possible joint outcome for X and Y: the first column lists the possible values of X; the second, the possible values of Y; and the third, the probabilities for each joint outcome. You should interpret the numbers in the X and Y columns as the value of \$1 at the end of the investment period—for example, after one year. If $X = 1.1$ after a year, then your holdings of that stock gained 10% in value.

x	y	$P(x,y)$
1.0	1.0	0.15
1.0	1.1	0.10
1.0	1.2	0.05
1.1	1.0	0.10
1.1	1.1	0.20
1.1	1.2	0.10
1.2	1.0	0.05
1.2	1.1	0.10
1.2	1.2	0.15

Table 7.1: Positive covariance.

Under this joint distribution, a single dollar invested in a portfolio with a 50/50 allocation between X and Y is a random variable W. This random variable has an expected value of 1.1 and variance

$$\begin{aligned}\text{var}(W) &= 0.5^2\text{var}(X) + 0.5^2\text{var}(Y) + 2 \cdot 0.5^2 \cdot \text{cov}(X,Y) \\ &= 0.5^2 \cdot 0.006 + 0.5^2 \cdot 0.006 + 2 \cdot 0.5^2 \cdot (0.002) \\ &= 0.004\,,\end{aligned}$$

for a standard deviation of $\sqrt{0.004}$, or about 6.3%.

What if, on the other hand, the asset returns were negatively correlated, as they are in the table at right? (Notice which entries have been switched, compared to the previous distribution.)

x	y	$P(x,y)$
1.0	1.0	0.05
1.0	1.1	0.10
1.0	1.2	0.15
1.1	1.0	0.10
1.1	1.1	0.20
1.1	1.2	0.10
1.2	1.0	0.15
1.2	1.1	0.10
1.2	1.2	0.05

Table 7.2: Negative covariance.

Under this new joint distribution, the expected value of \$1 invested in a 50/50 portfolio is still 1.1. But since the covariance between X and Y is now negative, the variance of the portfolio changes:

$$\begin{aligned}\text{var}(W) &= 0.5^2\text{var}(X) + 0.5^2\text{var}(Y) + 2 \cdot 0.5^2 \cdot \text{cov}(X,Y) \\ &= 0.5^2 \cdot 0.006 + 0.5^2 \cdot 0.006 + 2 \cdot 0.5^2 \cdot (-0.002) \\ &= 0.002\,,\end{aligned}$$

for a standard deviation of $\sqrt{0.002}$, or about 4.5%. Same expected return, but lower variance, and therefore more attractive to a risk-averse investor!

What's going on here? Intuitively, under the first portfolio, where X and Y are positively correlated, the bad days for X and Y tend to occur together. So do the good days. (When it rains, it pours; when it's sunny, it's 100 degrees.) But under the second portfolio, where X and Y are negatively correlated, the bad days and good days tend to cancel each other out. This results in a lower overall level of risk.

The morals of the story are:

1. Correlation creates extra variance.
2. Diversify! (Extra variance hurts your compounded rate of return.)