http://jgscott.github.com/

1
Probability foundations

All of probability in a single picture

Every year, about 245,000 women in the U.S. are diagnosed with breast cancer, and about 45,000 women will die from it.

But there are also over 2.1 million survivors of breast cancer living in the U.S., and as with many forms of cancer, early detection is a key predictor of a good outcome. In fact, by 2015, death rates from breast cancer had been steadily dropping since the 1970s. Researchers partially attribute this to the fact that mammography screening rates also increased over this period, so that more cases of breast cancer were detected at earlier, less threatening stages.[1]

[1] Myers et. al. "Benefits and Harms of Breast Cancer Screening: A Systematic Review." *Journal of the American Medical Association* 2015; 314(15):1615-34.

The revision. It therefore caught many people by surprise when, in late 2015, the American Cancer Society (ACS) revised its long-standing guidelines to recommend that most women have fewer mammograms. The *New York Times* devoted a front-page story to the announcement:[2]

> The American Cancer Society, which has for years taken the most aggressive approach to screening, issued new guidelines on Tuesday, recommending that women with an average risk of breast cancer start having mammograms at 45 and continue once a year until 54, then every other year for as long as they are healthy and likely to live another 10 years. The organization also said it no longer recommended clinical breast exams, in which doctors or nurses feel for lumps, for women of any age who have had no symptoms of abnormality in the breasts. Previously, the society recommended mammograms and clinical breast exams every year, starting at 40.

[2] "American Cancer Society, in a Shift, Recommends Fewer Mammograms." Denise Grady, *New York Times* front page, 20 October 2015. Available at http://www.nytimes.com/2015/10/21/health/breast-cancer-screening-guidelines.html.

The key changes here were for women under 45 or over 54, for whom biennial scans were now recommended; for women aged 45-54, annual scans remained the recommendation. Moreover, it's important to point out here (as the *New York Times* did, responsibly) that the revised guidelines applied only to women with an

"average" risk of breast cancer. Women with a personal or family history of breast cancer, or any other major risk factor, were still encouraged to get annual screenings from an early age.

Nonethess, the revised guidelines represented a big shift in thinking. Why the changes for the under-45s and over-54s? They arose from a careful attempt to balance the benefits of frequent early screening against the possible risks. The ACS cited three main points based on an exhaustive review of the clinical research.

1. For under 45s: younger women are less likely than older women to develop or die from breast cancer. No available evidence suggested that getting a scan every year, as opposed to every two years, decreases breast cancer mortality for younger (premenopausal) women. However, a woman with average risk of breast cancer who begins annual screening at age 40, as opposed to 45, does have a higher risk of experiencing a false-positive scan.[3] This can lead to unnecessary tests and procedures, which most women find stressful, onerous, expensive, or even painful.
2. Again for under 45s: early detection of a small tumor destined to become a big tumor is clearly a good thing. But not all small tumors are destined to become big tumors. The research showed that, in some cases, early detection could scare people into pursuing an overly aggressive course of treatment for small tumors that, if left alone, might never have caused the patient any harm.
3. For the over-54s: breast tumors in older women tend to be slower growing and less aggressive than in younger women. In the words of the researchers: "For women older than 55 years, biennial mammography is likely to provide the best balance of benefits to harms."[4]

[3] Here, a false positive means that a radiologist examines the mammogram of a healthy woman and wrongly concludes that she is likely to have breast cancer.

[4] Myers et. al., *ibid.*

The reaction. Plenty of clinicians were convinced by these arguments and agreed with the change in guidelines. But not everyone. In fact, the revision provoked a major backlash among some doctors and researchers. For example, three of them—Susan Drossman, Elisa Port, and Emily Sonnenblick—used the editorial page of the *New York Times* to very publicly renounce their lifelong support of the American Cancer Society:

> We profoundly disagree with these changes. All three of us, two breast radiologists and one breast surgeon, have been

> named "Mothers of the Year" by the American Cancer Society in recognition of our roles as mothers and physicians who have devoted our professional lives to the fight against breast cancer. One of us, Dr. Drossman, received her award the day before the new guidelines were issued. Because of our shared goals—early detection of breast cancer, improved treatments and saving lives—we were happy to support the cancer society. Now, we no longer wish to be involved. . . .
>
> Today, the overall survival rate for breast cancer in the United States is very close to 90 percent, the highest it has ever been. . . . Early detection with mammography and better treatment options are both directly responsible for that.[5]

[5] "Why the Annual Mammogram Matters," by Susan Drossman, Elisa Port, and Emily Sonnenblick. *New York Times*, 29 October, 2015, page A31. Available at http://www.nytimes.com/2015/10/29/opinion/why-the-annual-mammogram-matters.html.

Drs. Drossman, Port, and Sonnenblick went on to argue:

> Let's stop overemphasizing the 'harms' related to mammogram callbacks and biopsies. In breast cancer screening, most false positives—when a mammogram suggests something that requires further investigation—are easily resolved with additional images that put the concern to rest. Only 2 percent of screening cases cannot be resolved by further imaging and require a biopsy. . . . In our collective experience of 60 years, none of us have ever seen a potentially life-threatening complication related to a breast biopsy. We and our many colleagues who are intimately involved in patient care will continue to recommend to women annual screening mammograms starting at age 40.

To make matters more confusing, other clinical organizations—including the National Comprehensive Cancer Network, the American College of Obstetricians and Gynecologists, and the United States Preventive Services Task Force—kept their own guidelines in place. Some of these guidelines recommended earlier and more frequent mammograms than the ACS recommended; others, later or less frequent.

How to decide? One thing that everyone agrees on is that annual mammograms should still be offered to anyone who wants them. But the conflicting advice from all these experts leaves normal people in a position of uncertainty. If you're a woman in her 40s with no family history of breast cancer, what would you do? If you're not a woman, how would you counsel a loved one?

These questions lead to a cascade of follow-ups. Some of these are about knowing your own mind. For example, how stressful and difficult do you find medical procedures? Can you live with the higher likelihood of being called back to a doctor's office based

on a false positive? Or can you more easily live with the knowledge that, if you do develop cancer, the tumor might be larger than if you'd caught it earlier? The answers to these questions will differ from woman to woman, which is why most experts also agree that there is no single right answer to the question, "Should I get a mammogram every year?"

But other follow-up questions are about probability. For these, we need evidence rather than introspection. Just how likely is breast cancer? If you do get cancer, what are the chances of survival? How likely is a false positive from a routine mammogram?

The probability tree

We are statisticians, not medical doctors. We are not qualified to give clinical advice about breast-cancer screening, nor to weigh in on the controversy surrounding the American Cancer Society's new guidelines—which, both sides of the debate acknowledge, were the product of years of careful thought and discussion by dedicated experts genuinely trying to do the greatest good for the greatest number of women.

However, as statisticians, we can offer the following. If you want to make the best decision *for you*—about mammograms, or just about anything else—then it helps to know something about probability. In this spirit, allow us to explain why we're such big fans of Figure 1.1:

1. It can help someone reach a decision about screening mammograms that is right for them.

2. It can teach all of probability in a single picture.

Figure 1.1 is the brainchild of David Spiegelhalter and Jenny Gage of the University of Cambridge. These researchers asked themselves the question: how can we present the evidence on the benefits and risks of screening in a way that doesn't make an explicit recommendation, but that helps people reach their own conclusion? The result of their efforts was a series of *probability trees* like Figure 1.1, each one depicting the likely experiences of women with and without screening.

This particular figure tracks what we'd expect to happen to two hypothetical cohorts of 200 women, aged 50 to 70. In the cohort of 200 on the left, all women are screened; while in the cohort of 200 on the right, none are screened. The expected results for each

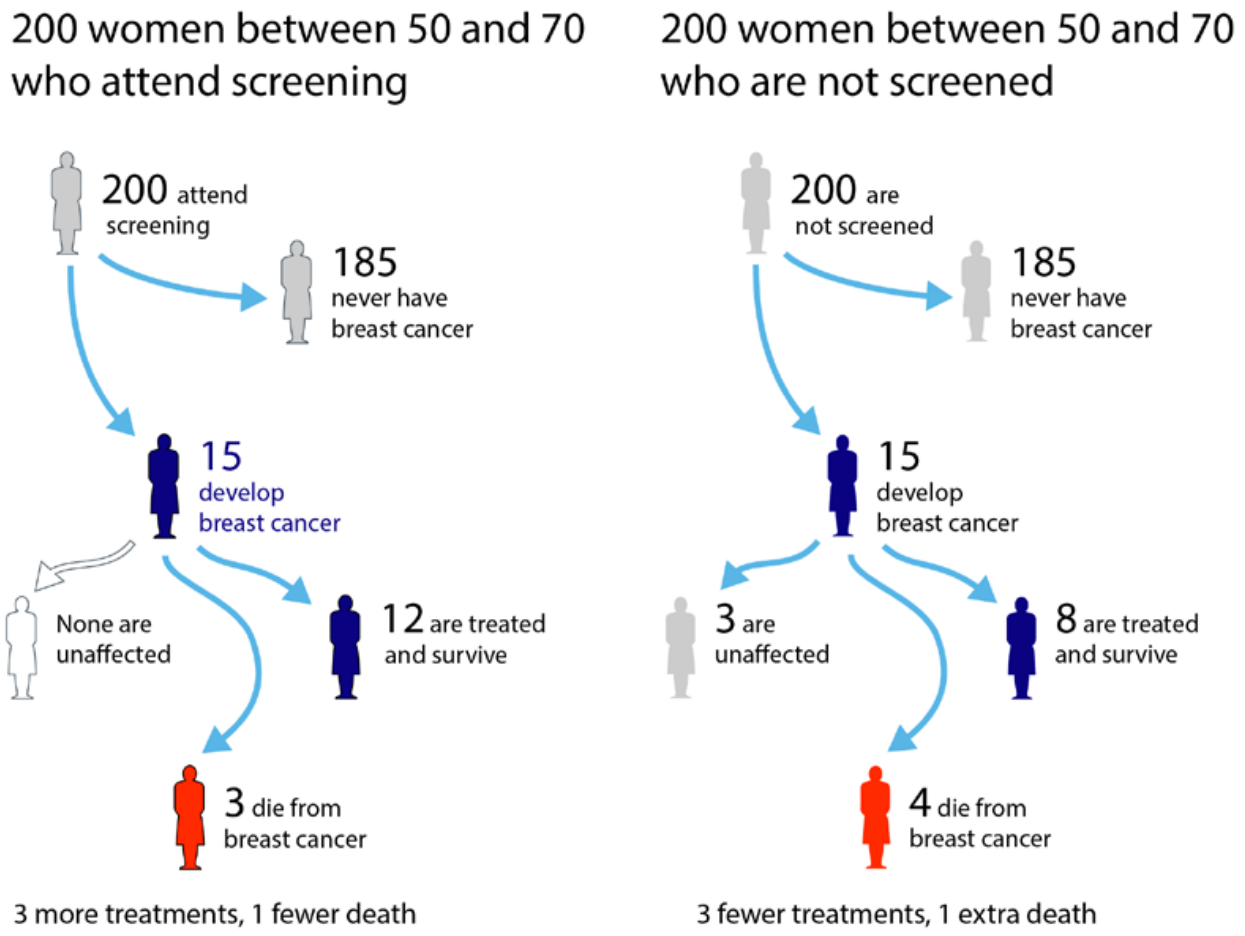

Figure 1.1: Two hypothetical cohorts of 200 women, ages 50-70. The 200 women on the left all go in for mammograms; the 200 on the right do not. The branches of the tree show how many women we would expect to experience various different outcomes. Figure from: "What can education learn from real-world communication of risk and uncertainty?" David Spiegelhalter and Jenny Gage, University of Cambridge. *Proceedings of the Ninth International Conference on Teaching Statistics* (ICOTS9, July, 2014). We're not the only fans of the picture: it won an award for excellence in scientific communication in 2014 from the UK Association of Medical Research Charities.

cohort are slightly different: on the right, we expect 1 fewer death, and 3 extra unnecessary screenings, versus the left. Which cohort would you prefer to be in?

Moreover, while this tree depicts the typical experience of 50-70 year olds, you can imagine that a similar probability tree with the appropriate numbers for, say, 40-50 year olds, or those over 70, would help an even wider cohort of women navigate this choice.

Finally, almost as a bonus, just about every major concept in probability is represented in this picture.

Expected value. In a group of 200 women, how many would we expect to get breast cancer?

Our best guess, or expected value, is about 15, regardless of whether they get screened or not.

Probability. How likely is breast cancer for a typical woman?

Fifteen cases of cancer in a cohort of 200 women means that an average woman aged 50-70 has a 7.5% chance of getting breast cancer ($15/200 = 0.075$).

Joint probability. Suppose that a typical woman does not go for a screening mammogram. How likely is she to get breast cancer and to die from it?

In the cohort of 200 unscreened women on the right, 4 are expected to get breast cancer and die from it. Thus the risk for a typical woman is about $4/200 = 0.02$, or 2%.

Rule of addition. Suppose a woman decides to go in for screening. What are her chances of not dying from breast cancer?

In the screened cohort, there are two kinds of survival outcomes: never getting breast cancer (185 women); or getting breast cancer, receiving treatment, and surviving (12 women). Adding these together, we find that 185+12=197 out of 200 women ultimately survive their experience, for a probability of 197/200, or 98.5%.

Conditional probability. Suppose that a woman decides to forego screening. If she then goes on to develop breast cancer, how likely is she to die from that cancer?

In the unscreened cohort, 15 women are expected to get breast cancer. Of these 15 women, 4 are expected to die from their cancer. Thus for an unscreened 50-70 year-old woman, the risk of dying from breast cancer, given that she develops breast cancer in the first place, is about 4/15, or about 27%. (Among screened women, this figure is 3/15, or 20%.)

This list of concepts isn't quite all of probability, but it's close!

Over the next several chapters, we will unpack these ideas in much more detail. Along the way, we'll meet several other examples of how probability can help you reason about uncertainty, in situations from the grave to the frivolous:

- How did the U.S. Army almost go drastically wrong in World War II when deciding where to beef up the armor on its bombers?
- Should a motorist be sent to jail on the sole basis of failing a roadside test for crack cocaine?

- Where should the search begin when a nuclear submarine is lost at sea?
- How does Netflix use a person's movie viewing history to recommend new shows?

And many more.

Reasoning about risk

Jared Diamond's constructive paranoia

For most of us, life is full of worry. Some people worry about tornados or earthquakes; other people won't get on an airplane. Some people worry more about lightning; others, about terrorists. And then there are the everyday worries: about love, money, career, status, conflict, young kids, old dogs, or the point of it all.

Jared Diamond worries a lot, too—about slipping in the shower.

Dr. Diamond is one of the most respected scientists in the world. Though he originally trained in physiology, Diamond left his most lasting mark on the popular imagination as the author of *Guns, Germs, and Steel: The Fates of Human Societies*. This Pulitzer-prize-winning book draws on ecology, anthropology, and geography to explain the major trends of human migration, conquest, and displacement over the last few thousand years.

Strangely enough, Diamond began to worry about slipping in the shower while conducting anthropological field research in the forests of New Guinea, 7,000 miles away from home, and a long day's walk from any shower. The seed of this worry was planted one day while he was out hiking in the wilds with some New Guineans. As night fell, Diamond suggested that they all make camp under the broad canopy of a nearby tree. But his companions reacted in horror, and refused. As Diamond tells it,

> They explained that the tree was dead and might fall on us. Yes, I had to agree, it was indeed dead. But I objected that it was so solid that it would be standing for many years. The New Guineans were unswayed, opting instead to sleep in the open without a tent.[6]

[6] Jared Diamond, "That Daily Shower Can Be a Killer." *New York Times*, January 29. 2013, page D1.

The New Guineans' fear initially struck Diamond as overblown. How likely could it possibly be that the tree would fall on them in the night? Surely they were being paranoid. For a famous professor like Diamond to get crushed by a tree while sleeping in the

forest would be the kind of freakish thing that made the newspaper, like getting struck by lightning at your own wedding or being killed by a falling vending machine.

But in the months and years after this incident, it began to dawn on Diamond that the New Guineans' "paranoia" was well founded. A dead tree might stay standing for somewhere between 3 and 30 years, so that the daily risk of a toppling was somewhere between 1 in 1,000 and 1 in 10,000. This is small, but far from negligible. Here's Diamond again:

> I came to realize that every night that I camped in a New Guinea forest, I heard a tree falling. And when I did a frequency/risk calculation, I understood their point of view. Consider: If you're a New Guinean living in the forest, and if you adopt the bad habit of sleeping under dead trees whose odds of falling on you that particular night are only 1 in 1,000, you'll be dead within a few years. In fact, my wife was nearly killed by a falling tree last year, and I've survived numerous nearly fatal situations in New Guinea.[7]

[7] *ibid.*

Having absorbed this attitude about the importance of everyday habits, Diamond began to apply it to his own life. He refers to it as a "hypervigilant attitude towards repeated low risks," or more memorably, "constructive paranoia."

Take the simple act of showering. If you're 75 years old, as Diamond was when he recounted this story, you can expect to live another 15 years. That's $15 \times 365 = 5{,}475$ more daily showers. So if your risk of a bad slip is "only" one in a thousand, you should expect to break your hip, or worse, about five times over that period. The implication is that, if you want a good chance of being around to blow out 90 candles, you must ensure that, by your own careful behavior, you reduce the risk of slipping in the shower to something much, much lower than one in a thousand.

And the same goes for all those other small risks we face day in, day out. Think about crossing a busy street, driving at night, touching the handle of a public toilet, or venturing out with the mad dogs and Englishmen into the mid-day sun. Each time the chance of a disaster is low. But most of us perform these actions again and again—and if we're slapdash about it, the expected number of disasters over several years can be alarmingly high. Diamond's conclusion? He needed to ensure that, for each repeated exposure to one of these risks, the chance of a disaster wasn't just low, but extremely low.

Cause	Expected deaths
Heart disease	203
Cancer	195
Respiratory disease	50
Stroke	42
Alzheimer's	28
Diabetes	25
Accidental poisoning	12
Car accident	11
Slips/falls	10
Homicide	5
Eating raw meat	2
Choking	1.5
Pregnancy	0.2
Dog bite	0.01
Falling vending machine	0.001
Hurricane	0.03
Tornado	0.02
Mass shooting	0.01
Lightning strike	0.01
Shark attack	0.0003
Plane crash	0.0001

(per 100,000)

Table 1.1: Expected deaths due to various causes over one year in an imaginary cohort of 100,000 Americans, of whom 99,200 are expected to survive.

Expected value

Jared Diamond's philosophy of constructive paranoia arises from an understanding of *expected value*. The expected value of a random event has a formal mathematical definition, which is important, and which we'll come to shortly. But the basic idea is simple: an expected value is just an average. And for repeated, independent events like taking a shower, or sleeping under a dead tree, this average is simple to calculate: it's the risk of the event in a single encounter, times the number of encounters.

For example, let's say that the risk of a dead tree falling down in the night is one chance in a thousand, and that you and 99 friends each sleep under your own dead tree every night for a year (so $365 \times 100 = 36{,}500$ person-nights of exposure). In your cohort of 100, how many would you expect to get crushed by a tree? The

math doesn't look good:

$$\begin{aligned}\text{Expected crushings} &= (\text{Risk of dead tree falling}) \times (\text{Number of exposures}) \\ &= \frac{1}{1000} \times (365 \times 100) \\ &= 36.5\,.\end{aligned}$$

You can expect about 36 of you to be crushed.

Let's look at some similar expected values for an imaginary cohort of 100,000 Americans—about the size of a small city, like Boulder or Green Bay. Table 1.1 shows how many of these 100,000 people we would expect to die annually due to various kinds of random events.[8] As you can see, the expected number of deaths due to the headline-grabbing causes in the bottom half of the table—from tornadoes to shark attacks to mass shootings—is tiny. And while 99,200 are expected to survive, most of the unlucky 800 succumb to chronic disease, cancer, the vagaries of age, or simple accidents. In fact, we'd expect 3 times as many people to die from a falling vending machine as from a shark, and 20 times as many to die from choking as from all six sensational causes put together.

[8] Centers for Disease Control, http://www.cdc.gov/nchs/fastats/.

In light of these numbers, it might be wise to heed Jared Diamond's advice. He cites research showing that most people's attitude toward danger is confused: we overestimate the risk of dramatic events like those in the bottom half of Table 1.1, while simultaneously underestimating the risk of the familiar causes in the top half. To be fair, this has a lot to do with living in a world of mass media and near-instant communication. Thousands of people anonymously choke to death every year. But if someone gets attacked by a shark or blows himself up in a train station anywhere in the world, you will hear about it, no matter how unlikely the actual risk. As people in the statistics business put it: newspapers love numerators. While this makes for good copy, it short-circuits our natural cues for reasoning intuitively about risk.

But as Diamond explains, dwelling on the spectacular numerators isn't a smart way to stay alive. Many of life's mundane risks, from heart disease to car accidents to falling in the shower, do not strike out of the blue. Rather, they are direct results of our own day-to-day behavior:

> Having learned both from those studies and from my New Guinea friends, I've become as constructively paranoid about showers, stepladders, staircases and wet or uneven sidewalks as my New Guinea friends are about dead trees. As I drive, I

> remain alert to my own possible mistakes (especially at night), and to what incautious other drivers might do.
>
> My hypervigilance doesn't paralyze me or limit my life: I don't skip my daily shower, I keep driving, and I keep going back to New Guinea. I enjoy all those dangerous things. But I try to think constantly like a New Guinean, and to keep the risks of accidents far below 1 in 1,000 each time.[9]

[9] *ibid.*

It turns out that your mother and the New Guineans agree about the importance of good habits in successfully navigating all the sidewalks, roads, mid-day suns, and public toilets of a full and busy life. Look both ways, don't drive while tired, wear sunscreen, wash your hands—and don't sleep under dead trees.

Probability: a language for uncertainty

All of these examples illustrate the concept of a *random variable*, which is a generic term for any uncertain outcome. For example:

- the number of trees that fall over tonight in a particular patch of New Guinean forest.
- the number of women aged 50-70, out of a group of 200, who will get breast cancer.
- how many users will click on a particular Google ad in the next hour.

For these random variables, we've seen how to form expected values by mutiplying the risk times the exposure.

But here's where we run up against the limitations of thinking about risk purely in terms of a simple frequency/exposure calculation. One problem is a lack of generality. For example, it's not at all clear how we could use this approach to calculate an expected value for *these* uncertain outcomes:

- the rate of U.S. unemployment in 18 months.
- the value of your retirement portfolio in 30 years.
- your extra lifetime earnings from going to graduate school.

What does a risk/exposure calculation even look like here? And what about these uncertain outcomes?

- the winner of next year's Tour de France.

- whether a defendant is guilty or innocent.
- whether you'll like the next person you're matched up with through a dating app.

Here the possible outcomes aren't even numbers.

But the bigger problem is that an expected value conveys nothing about *uncertainty*. We may expect that 11 people in Green Bay, Wisconsin (pop. 100,000) will die this year in a car accident. But it could be 5, or 20. It's a random variable; no one knows for sure.

To really understand risk deeply, we need a better language for helping us to communicate clearly about uncertainty. That language is probability.

The basic rules of probability

A probability is a number that measures how likely some event is. We use the notation $P(A)$ to denote the probability of the event A: $P(\text{coin lands heads}) = 0.5$, $P(\text{rainy day in Ireland}) = 0.85$, $P(\text{cold day in Hell}) = 0.0000001$, and so forth.

There are only four basic rules of probability. The first three are:

1. All probabilities are numbers between 0 and 1, with 0 meaning impossible and 1 meaning certain.
2. Either an event occurs (A), or it doesn't (not A):
$$P(\text{not } A) = 1 - P(A)\,.$$
3. If two events are mutually exclusive (i.e. they cannot both occur), then
$$P(A \text{ or } B) = P(A) + P(B)\,.$$

These are sometimes called Kolmogorov's rules, after a Russian mathematician. (Like chess and gymnastics, probability is a very Russian pursuit.)

There's also a fourth, slightly more advanced rule for conditional probabilities. A conditional probability, denoted $P(A \mid B)$, is the probability that A happens, given that B has already happened, or is assumed to happen.[10] Here's the rule:

4. Let $P(A, B)$ be the joint probability that both A and B happen. Then the conditional probability $P(A \mid B)$ is: [11]
$$P(A \mid B) = \frac{P(A, B)}{P(B)}\,. \tag{1.1}$$

[10] The vertical bar ($\mid$) means "given" or "conditional upon."

[11] A minor technical point: this only works if $P(B) > 0$. You can't condition upon an impossible event, i.e. where $P(B) = 0$.

Conditional probabilities reflect our assumptions and our partial knowledge. They satisfy all the same rules as ordinary probabilities, and we can compare them as such. For example, we all know that

$$P(\text{rainy afternoon} \mid \text{cloudy morning}) > P(\text{rainy afternoon} \mid \text{sunny morning}),$$
$$P(\text{shark attack} \mid \text{swimming in ocean}) > P(\text{shark attack} \mid \text{watching TV}),$$
$$P(\text{heart disease} \mid \text{swimmer}) < P(\text{heart disease} \mid \text{couch potato}),$$

and so forth, even if we don't know the exact numbers.

A key fact about conditional probabilities is that they are not symmetric: $P(A \mid B) \neq P(B \mid A)$. In fact, these two numbers are sometimes very different. For example, just about everybody who plays professional basketball in the NBA practices very hard:

$$P(\text{practices hard} \mid \text{plays in NBA}) \approx 1.$$

But sadly, most people who practice hard with a dream of playing in the NBA will fall short:[12]

$$P(\text{plays in NBA} \mid \text{practices hard}) \approx 0.$$

[12] The average NBA player is 6′ 7″, or over 200 cm.

Finally, we say that two events A and B are *independent* of each other if one event conveys no information about the other. To explain this concept mathematically, we invoke conditional probability. Two events A and B are independent if

$$P(A) = P(A \mid B) = P(A \mid \text{not } B).$$

That is, knowing whether B occurred doesn't change your assessment of how likely A is to occur, and vice versa. If two events are independent, then their joint probability is the product[13] of their individual probabilities: $P(A, B) = P(A) \cdot P(B)$. Thus, for example, the probability of rolling two dice and getting 1–1 ("snake eyes") is $1/6 \cdot 1/6 = 1/36$.

[13] You can derive this fact by applying a bit of algebra to Rule 4, the rule for conditional probabilities, under the assumption that A and B are independent. For non-independent events, this simplification doesn't work and you must go back to Rule 4.

Example: Pete Rose's hitting streak. As far as the rules of probability go: surprisingly, that's it! Every other concept and rule can be worked out from these four. But as we'll see, despite their simplicity, these rules are stunningly powerful for helping us reason about random outcomes. Here's a short, teaser example.

Before he got banned from baseball for life for betting on his own games with the Cincinatti Reds, Pete Rose was an extraordinary baseball player. Rose holds the Major League record for

career hits, at 4,256; most baseball fans consider it unlikely that his record will ever be broken. He is also famous for his performance over the summer of 1978, when he got a hit in every game from June 14 to August 1st—a hitting streak of 44 games. For a couple of weeks over that summer, fans and journalists believed that Rose had a real chance at breaking Joe DiMaggio's 56-game hitting streak, the all-time record standing since 1941.

What probability might reasonably be associated with a hitting streak of that length? First, let's make some basic assumptions.

- Rose bats 4 times per game, and each at-bat is independent (i.e. the current at-bat doesn't affect the next one).
- In 1978 Rose was a .300 hitter, so we'll say that his average chance of getting a hit in any single at-bat was 30%. (By comparison, DiMaggio was a .328 hitter in 1941, an era of baseball that was friendlier to batters than Rose's was.)

Next, we'll appeal to the second of the four rules of probability listed earlier: either Rose gets a hit in a game, or he doesn't (i.e. gets 4 outs). Therefore, for a single game, $P(\text{Rose gets a hit}) = 1 - P(\text{Rose makes 4 outs})$. Why write it this way? Because the probability that Rose makes four outs is the easier of the two numbers to calculate directly. Rose has a 70% chance of getting out in a single at-bat, and the at-bats are assumed independent. So for a single game,

$$P(\text{Rose makes 4 outs}) = 0.7 \cdot 0.7 \cdot 0.7 \cdot 0.7 \approx 0.24\,,$$

meaning that

$$P(\text{Rose gets a hit}) = 1 - 0.24 = 0.76\,.$$

Finally, we're ready to calculate the probability of Rose's 44-game hitting streak. Let H_i denote the event "Rose hits safely in game i." Since each game is assumed to be independent,

$$\begin{aligned} P(H_1, H_2, \ldots, H_{44}) &= P(H_1) \times P(H_2) \times \cdots \times P(H_{44}) \\ &= (0.76)^{44} \\ &\approx \frac{1}{175440} \approx 0.0000057\,. \end{aligned}$$

From this, we can make three inferences:

1. The odds[14] against any particular player as good as Pete Rose starting such a hitting streak today are 175,000 to 1.

[14] Especially in a gambling context, you will sometimes see probabilities expressed as *odds*, such as, 9 to 2 or 3 to 1. Odds are usually quoted as the "odds on A" or "odds against A." The odds against some event A are just a different way of expressing $P(A)$. For example, suppose that the odds against a Patriots' Super Bowl victory are quoted as 9 to 2. This means that if we played the season over again 11 times ($11 = 9 + 2$), we would expect the Patriots to win the Super Bowl 2 times, and to not win it the other 9 times. The conversion formula is

$$\text{Odds against A} = \frac{1 - P(A)}{P(A)},$$

where the odds are interpreted as a decimal fraction like 9/2.

2. This doesn't mean that the 44-game hitting streak will never be beaten. The chance might be small, but remember: there are an awful lot of baseball games, and expected value = risk times exposure.
3. Joe DiMaggio's 56-game hitting streak in 1941 was statistically more impressive than Rose's in 1978. If you do a similar calculation using DiMaggio's average, you get $(0.792)^{56} = 1$ in 2.13 million. That one's going to be hard to beat.

⋆ *Probability distributions*

A few definitions, and a bit of math, will help us understand probability more deeply. Recall that a random variable is just a term for any uncertain outcome. We use the term *sample space* to mean the set of possible outcomes for a random variable. There are three common types of random variables, corresponding to three different types of sample spaces.

Categorical: the outcome will be one of many categories. For example, which party will win the next U.S. presidential election: Democrats, Republicans, or Other? Will your next interaction with customer service be Good, Fair, or Unrepeatable?

Discrete: the possible outcomes are whole numbers (1, 2, 3, etc.). Most of the examples we saw in our discussion of everyday risks—numbers of shark attacks, falls in the shower, and so forth—were discrete random variables.

Continuous: the random variable could be anything within a continuous range of numbers, like the price of Apple stock tomorrow, or the size of subsurface oil reservoir.

Discrete and continuous random variables are sometimes grouped together and called *numerical* random variables, since the possible outcomes are all numbers.

A *probability distribution* is just a list of probabilities for all the outcomes in the sample space of a random variable. For example, imagine that you've just pulled up to your new house after a long cross-country drive, only to discover that the movers have buggered off and left all your furniture and boxes sitting in the front yard.[15] What a mess! You decide to ask your new neighbors for some help getting your stuff indoors. Assuming your neighbors are the kindly type, how many pairs of hands might come

[15] This actually happened to a friend of mine. -JS

to your aid? Let's use the letter X to denote the (unknown) size of the household next door. The table at right shows a probability distribution for X, taken from U.S. census data in 2015; you might find this easier to visualize using the barplot in Figure 1.2.

This probability distribution provides a complete representation of your uncertainty in this situation. It has all the key features of any probability distribution:

1. There is a random variable, or uncertain quantity—here, the size of the household next door (X).
2. There is a *sample space*, or set of possible outcomes for the random variable—here, the numbers 1 through 8.
3. Finally, there are probabilities for each outcome in the sample space—here provided via a simple look-up table. Notice that the table uses big X to denote the random variable itself, and little x to denote the elements of the sample space.

Size of household, x	Probability, $P(X = x)$
1	0.280
2	0.336
3	0.155
4	0.132
5	0.060
6	0.023
7	0.011
8	0.003

Table 1.2: Probability distribution for household size in the U.S. in 2015. There is a vanishingly small probability for a household of size 9 or higher, which is just rounded off to zero here.

Most probability distributions won't be this simple, but they will all require specifying these three features.

⋆ *Expected value: the mathematical definition*

When you knock on your neighbors' door in the hopes of getting some help with your moving fiasco, how many people should you "expect" to be living there?

The *expected value* of a probability distribution for a numerical random variable is just an average of the items in the sample space—but a weighted average, rather than an ordinary average. If you take the 8 numbers in the sample space of Figure 1.2 and form their ordinary average, you get

$$\text{Ordinary average} = \frac{1}{8} \cdot 1 + \frac{1}{8} \cdot 2 + \cdots + \frac{1}{8} \cdot 7 + \frac{1}{8} \cdot 8 = 4.5\,.$$

Here, the weight on each number in the sample space is $1/8 = 0.125$, since there are 8 numbers. This is *not* the expected value; it give each number in the sample space an equal weight, ignoring the fact that these numbers have different probabilities.

To calculate an expected value, we instead form an average using *unequal* weights, given by the probabilities of each item in the sample space:

$$\text{Expected value} = (0.280) \cdot 1 + (0.336) \cdot 2 + \cdots + (0.011) \cdot 7 + (0.003) \cdot 8 \approx 2.5\,.$$

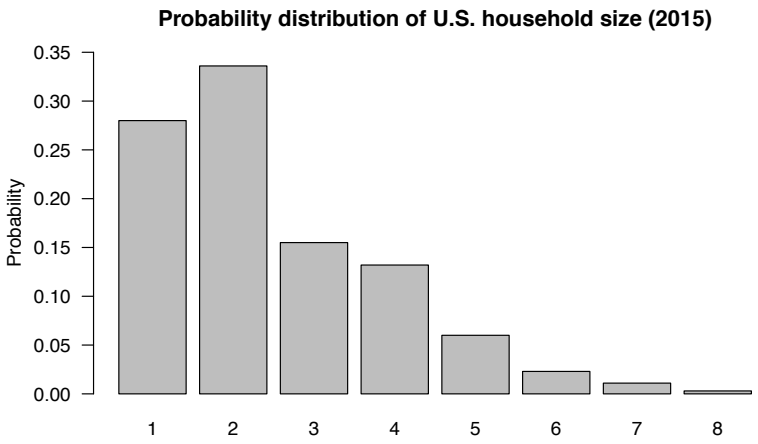

Figure 1.2: Probability distribution for the size of a random U.S. household in 2015. The elements of the sample space (the numbers $x = 1$ through $x = 8$) are shown along the horizontal axis. The probabilities $P(X = x)$ are shown on the vertical axis.

The more likely numbers (e.g. 1 and 2) get higher weights than 1/8, while the unlikely numbers (e.g. 7 and 8) get lower weights.

This example conveys something important about expected values. Even if the world is black and white, an expected value is often grey. For example, the expected American household size is 2.5; a baseball player expects to get 0.25 hits per at bat; the typical person "expects" to be born with 1 testicle; and so forth.

As a general rule, suppose that the possible outcomes for a random variable X are the numbers $x_1, \ldots, x_N$. The formal definition for the expected value[16] of X is

$$E(X) = \sum_{i=1}^{N} P(X = x_i) \cdot x_i \,. \tag{1.2}$$

[16] Later, in Chapter 3, we'll learn how something called the binomial distribution allows us to connect this definition with our earlier understanding of expected value as a risk/frequency calculation.

Joint distributions

During World War II, the size of the Allied air campaign over Europe was truly staggering. Every morning, huge squadrons of B-17 Flying Fortress bombers, each with a crew of 10 men, would take off from their air bases in the south of England, to make their way across the Channel and onwards to their targets in Germany. By 1943, they were dropping nearly 1 million pounds of bombs per week. At its peak strength, in 1944, the U.S. Army Air Forces (AAF) had 80,000 aircraft and 2.6 million people—4% of the U.S. male population—in service.

As the air campaign escalated, so too did the losses. In 1942,

the AAF lost 1,727 planes; in 1943, 6,619; and in 1944, 20,394. In a single mission over Germany in August of 1943, 376 B-17 bombers were dispatched from 16 different air bases in the south of England, in a joint bombing raid on factories in Schweinfurt and Regensburg. Only 316 planes came back—a daily loss rate of 16%. Some units were devastated: the 381st Bomb Group, flying out of RAF Ridgewell, lost 9 of its 20 bombers that day.[17]

Individual airmen had to complete 25 combat missions to be sent home. When the chances of returning safely from a *single* mission were that dicey, World War II airmen could be forgiven for thinking that they'd been sent to England to die.

But in the face of these bleak odds, the crews of the B-17s had at least three major defenses.

1. Their own tail and turret gunners, to defend the plane below and from the rear.
2. Their fighter escorts: the squadrons of P-47 Thunderbolts, RAF Spitfires, and P-51 Mustangs sent along to protect the bombers from the Luftwaffe.
3. A Hungarian-American statistician named Abraham Wald.

Abraham Wald never shot down a Messerschmitt or even saw the inside of a combat aircraft. Nonetheless, he made an outsized contribution to the Allied war effort, and no doubt saved the lives of many American bomber crews, using an equally potent weapon: conditional probability.

[17] Figures from *Statistical Abstract of the United States*, U.S. Census Bureau, (1944, 1947, 1950); and the Army Air Forces Statistical Digest (World War II), available at archive.org.

Figure 1.3: Abraham Wald as a young man.

Wald and the missing bombers

Abraham Wald was born in 1902 in Austria-Hungary, where he went on to earn a Ph.D. in mathematics from the University of Vienna. Wald was Jewish, and when the Nazis invaded in 1938, he—like so many brilliant European mathematicians and scientists of that era—fled to America.

Wald soon went to work as part of the Applied Mathematics Panel, which had been convened by order of President Roosevelt to function as something of a mathematical tech-support hotline for the U.S. military. It was during these years of service to his adopted country that Wald prevented the military brass from making a major blunder, thereby saving many lives.

Here's the problem Wald analyzed.[18] While some airplanes came back from bombing missions in Germany unscathed, many

[18] Distilled from: Mangel and Samaniego, "Abraham Wald's work on aircraft survivability." *Journal of the American Statistical Association* 79 (386): 259-67.

others had visibly taken hits from enemy fire. In fact, someone examining the planes just after they landed would likely have found bullet holes and flak damage everywhere: on the fuselage, across the wings, on the engine block, and sometimes even near the cockpit.

At some point, a clever person, whose identity is lost to history, had the idea of analyzing the distribution of these hits over the surface of the returning planes. The thinking was that, if you could find patterns in where the B-17s were taking enemy fire, you could figure out where to reinforce them with extra armor, to improve survivability. (You couldn't reinforce them everywhere, or they would be too heavy to fly.)

Researchers at the Center for Naval Analyses took this idea and ran with it. They examined data on hundreds of damaged airplanes that had returned from bombing runs in Germany. They found a very striking pattern[19] in where the planes had taken enemy fire. It looked something like this:

Location	Number of planes
Engine	53
Cockpit area	65
Fuel system	96
Wings, fuselage, etc.	434

[19] Alas, the actual data used in the original analyses cannot be located. But Wald wrote a report for the Navy on his methods, and we have attempted to simulate a data set that hews as closely as possible to the assumptions and (patchy) information that he provides in that report ("A Method of Estimating Plane Vulnerability Based on Damage of Survivors", from 1943). These and subsequent numbers are for hypothetical cohort of 800 airplanes, all taking damage.

If you turn that into a probability distribution, so that the numbers sum to 1, you get the following:

Location	Probability of hit
Engine	0.08
Cockpit area	0.10
Fuel system	0.15
Wings, fuselage, etc.	0.67

Thus of all the planes that took hits and made it back to base, 67% of them had taken those hits on the wings and fuselage. In the aggregate, no other part of the plane had taken nearly as much damage. The conclusion of the Navy researchers was simple: put more armor on the wings and fuselage.

But Wald pointed out that this recommendation suffered from a crucial flaw: *the Navy researchers only had data on the survivors.* The

planes that had been shot down were not available for analysis—and only the pattern of bullet holes on those missing planes could definitively tell the story of a B-17's vulnerabilities. In fact, Wald's thinking was very nearly the opposite of the Navy researchers. If all those planes had safely made it home with damage to the wings and fuselage, then these areas probably weren't all that vulnerable!

The importance of getting the right conditional probability

Wald's argument was, essentially, that the Navy researchers were using the wrong conditional probability. They had looked at the data and concluded that the wings and fuselage were vulnerable, on the basis of the fact that

$$P(\text{hit on wings or fuselage} \mid \text{returns safely}) \approx 0.67\,.$$

But that's the right answer to the wrong question. Instead, the Navy researchers needed to calculate

$$P(\text{returns safely} \mid \text{hit on wings or fuselage}) = ?$$

This might be a very different number. Remember: $P(\text{practices hard} \mid \text{plays in NBA}) \approx 1$, while $P(\text{plays in NBA} \mid \text{practices hard}) \approx 0$.

But to actually calculate a probability like $P(\text{returns safely} \mid \text{hit on wings or fuselage})$ required that Wald approach the data set like a forensic scientist. Essentially, he had to reconstruct the typical encounter of a B-17 with an enemy fighter, using only the mute testimony of the bullet holes on the planes that had made it back, coupled with some educated guesswork. So he went to work. He analyzed the likely attack angle of enemy fighters. He chatted with engineers. He studied the properties of a shrapnel cloud from a flak gun. He suggested to the army that they fire thousands of dummy bullets at a plane sitting on the tarmac. And yes, he did a lot of math.[20]

Remarkably, when all was said and done, Wald was able to reconstruct an estimate for the *joint distribution* for the two distinct types of events that each airplane experienced: where it took a hit, and whether it returned home safely. In other words, although Wald couldn't bring the missing planes back into the air, he could bring their statistical signature back into the data set. For a cohort of 800 bombers that took damage, Wald's estimate of this joint distribution would have looked something like the following table. You'll notice that the numbers in the left column correspond

[20] We don't go into detail on Wald's methods here, which were very complex. But later statisticians have taken a second look at those methods, with the hindsight provided by subsequent advances in the field. They have concluded, very simply: "Wald's treatment of these problems was definitive." (Mangel and Samaniego, *ibid.*)

exactly to the table given earlier: the pattern of hits to airplanes that made it back home. What's new is the right column: Wald's forensic reconstruction of the pattern of hits to planes that had been shot down.

	Returned	Shot down
Engine	53	57
Cockpit area	65	46
Fuel system	96	16
Wings, fuselage, etc.	434	33

Table 1.3: An example of how Abraham Wald could have reconstructed the joint distribution over hit type and outcome for our hypothetical cohort of 800 planes taking enemy fire.

For example, Wald's method would have estimated that 53 of the 800 planes, or 6.6% overall, experienced the joint event (hit type = engine, outcome = returned home safely).

This estimate for the joint distribution over two random variables, hit type and outcome, now allowed Wald to answer the right question. Of the 467 planes that had taken hits to wings and fuselage, 434 of them had returned home, while 33 of them had not. Thus Wald estimated that the conditional probability of survival, given a hit to the wings and fuselage, was

$$P(\text{returns safely} \mid \text{hit on wings or fuselage}) = \frac{434}{434+33} \approx 0.93\,.$$

It turns out that B-17s were pretty robust to taking hits on the wings or fuselage.

On the other hand, of the 110 planes that had taken damage to the engine, only 53 only returned safely. Therefore

$$P(\text{returns safely} \mid \text{hit on engine}) = \frac{53}{53+57} \approx 0.48\,.$$

Similarly,

$$P(\text{returns safely} \mid \text{hit on cockpit area}) = \frac{65}{65+46} \approx 0.59\,.$$

The bombers were much more likely to get shot down if they took a hit to the engine or cockpit area. Thus Wald's final, life-saving advice ran exactly counter to that of the Center for Naval Analyses: *put the armor where you don't see the bullet holes*.

Postscript. In the story of Abraham Wald and the missing airplanes, the path of counterintuitive facts eventually turns a full 360

degrees. Imagine asking any random person off the street: "Where should we put extra armor on airplanes to help them survive enemy fire?" We haven't done this survey, but we strongly suspect that most thoughtful people would answer: where the pilot and the engines are! But the data initially seem to contradict this intuition, and most people are happy to run with the data: if the planes are taking damage on the wings and the fuselage, then by God, let's put the armor there instead. It turns out that only a very careful analysis like Wald's can rehabilitate the initial, intuitive guess over the spurious attempt at data-based reasoning.

The moral of the story is that data alone isn't enough. You have to know enough about conditional probability to be able to pose the right question in the first place.

How Netflix knows your taste in movies so well

The same math that Abraham Wald used to analyze bullet holes on B-17s also underpins the modern digital economy of films, television, music, and social media. To give one example: Netflix, Hulu, and other video-streaming services all use this same math to examine what shows their users are watching, and apply the results of their number-crunching to recommend new shows.

To see how this works, suppose that you're designing the movie-recommendation algorithm for Netflix, and you have access to the entire Netflix database, showing which customers have liked which films. Your goal is to leverage this vast data resource to make automated, personalized movie recommendations. The better these recommendations are, the more likely your customers are to keep their accounts on auto-pay.

You decide to start with an easy case: assessing how probable it is that a user will like the film *Saving Private Ryan* (event A), given that the same user has liked the HBO series *Band of Brothers* (event B). This is almost certainly a good bet: both are epic dramas about the Normandy invasion and its aftermath. Therefore, you might think: job done! Recommend away.

For this particular pair of shows, fine. But keep in mind that you want to be able to do this kind of thing automatically. It would not be cost effective to put a human in the loop here, laboriously tagging all possible pairs of movies for similar themes or content—to say nothing of all of the other stuff that might make two different films appeal to the same person.

The key insight here is to frame the problem in terms of con-

ditional probability: if $P(A \mid B)$ is high, then B watchers are also likely to be A watchers. This is where your database, coupled with the rule for conditional probability, comes in handy. Suppose the data on your 50 million users looks like this:

	Liked *Band of Brothers*	Didn't watch or didn't like
Liked *Saving Private Ryan*	2.8 million	6.5 million
Didn't watch or didn't like	0.7 million	40 million

Just as with the example of Wald and the missing airplines, here we have a joint distribution for two random variables: A = whether a user liked *Saving Private Ryan*, and B = whether the user liked *Band of Brothers*. From this joint distribution, we can easily work out the conditional probability that we need. Of the 50 million users in the database, 2.8 + 0.7 = 3.5 million of them watched and liked *Band of Brothers*. Of these 3.5 million people, 2.8 million (or 80%) also liked *Saving Private Ryan*. Therefore,

$$P(\text{likes } \textit{Saving Private Ryan} \mid \text{likes } \textit{Band of Brothers}) = \frac{2.8 \text{ million}}{3.5 \text{ million}} = 0.8\,.$$

Note that you could also jump straight to the math and use Equation 1.1, the rule for conditional probabilities, like this:

$$P(A \mid B) = \frac{P(A,B)}{P(B)} = \frac{2.8/50}{(2.8+0.7)/50} = 0.8\,.$$

You'd get the same answer in the end.

The key thing that makes this approach work so well is that it's automatic. Computers aren't yet very good at automatically scanning films for thematic content. But they're brilliant at calculating conditional probabilities from a vast database of users' movie-watching histories.

The same trick works for books, too. Suppose you examine the online book-purchase histories of two friends Albert and Pablo, and discovered the following items:

Albert	Pablo
Proof and Consequences	*A Short History of Non-representational Art*
A Body in Motion: Newton's Guide to Productivity	*Achtung, Maybe? Dali and the Surreal*
Obscure Theorems of the 14th Century	*Your Face is Offside: Cubism and Soccer*

What sorts of books are you likely to recommend to these friends for their birthdays? Amazon learned to automate this process long ago, to the chagrin of independent bookstores everywhere. Similar math also underpins recommender systems for music (Spotify), ads (Google), and even friends (Facebook).

The digital economy truly is ruled by conditional probability.

⋆ *Joint, conditional, and marginal probabilities*

Joint probabilities. To understand the basic math behind joint, conditional, and marginal probabilities, we'll return to the story of Abraham Wald and the B-17s. We start by turning Table 1.3, which contains counts of different joint event types for a cohort of 800 airplanes, into a table of probabilities:

	Returned	Shot down
Engine	0.066	0.071
Cockpit area	0.081	0.058
Fuel system	0.120	0.020
Wings, fuselage, etc.	0.542	0.042

This is the joint distribution for two random variables: X = hit type, along the rows; and Y = outcome, along the columns. The entries in the table give the joint probabilities $P(X = x, Y = y)$. For example, 2% of all planes both took a hit in the fuel system and got shot down: $P(X = \text{fuel system}, Y = \text{shot down}) = 0.02$. Up to round-off error, these 8 probabilities all sum to 1.[21]

[21] In general, a joint distribution is a list of all possible joint events for multiple random variables, together with their joint probabilities.

Marginal probabilities. Next, we add an additional row and column of *marginal* (or overall) probabilities of the different event types and outcomes, like this:

	Returned	Shot down	Marginal
Engine	0.066	0.071	0.137
Cockpit area	0.081	0.058	0.139
Fuel system	0.120	0.020	0.140
Wings, fuselage, etc.	0.542	0.042	0.584
Marginal	0.809	0.191	1

These are called the marginal probabilities because we calculate them by summing across the relevant margin (i.e. row or column)

of the table. This just reflects the fact that the probability of some event (like returning safely) is the sum of the probabilities for all the distinct ways that event can happen. For example, an airplane that takes a hit to the engine can do so in two ways: it can take the hit and return, or it can take the hit and not return. Therefore,

$$\begin{aligned} P(\text{hit to engine}) &= P(\text{returned}, \text{hit to engine}) + P(\text{shot down}, \text{hit to engine}) \\ &= 0.066 + 0.071 = 0.137\,. \end{aligned}$$

The rest of the marginal probabilities are calculated similarly, e.g.

$$\begin{aligned} P(\text{returned}) &= 0.066 + 0.081 + 0.120 + 0.542 \\ &= 0.809\,. \end{aligned}$$

Conditional probabilities. Finally, we are ready to understand the rule for conditional probabilities. You'll recall that this was the fourth of the basic rules of probability quoted earlier. It goes like this:

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

Remember how we used Table 1.3 to calculate $P(\text{returns} \mid \text{hit to engine})$? We looked at the total number of planes that had taken a hit to the engine. We then asked: of these planes, how many also returned home safely? As an equation, this gives us

$$\begin{aligned} P(\text{returns} \mid \text{engine hit}) &= \frac{\text{Number taking engine hit and returned safely}}{\text{Number taking engine hit}} \\ &= \frac{53}{110} \approx 0.48\,. \end{aligned}$$

You'll notice we get the exact same answer if we use the rule for conditional probabilities: $P(A \mid B) = P(A, B)/P(B)$. These probabilities are estimated using the relevant fractions from the data set:

$$\begin{aligned} P(\text{returns} \mid \text{engine hit}) &= \frac{\text{Fraction taking engine hit and returning safely}}{\text{Fraction taking engine hit}} \\ &= \frac{53/800}{110/800} \\ &= \frac{0.066}{0.137} \approx 0.48\,. \end{aligned}$$

While the rule for conditional probabilities may look a bit intimidating, it just codifies exactly the same intuition we used to calculate $P(\text{returns} \mid \text{engine hit})$ from the table of counts.

Hot streaks and coincidences: understanding independence

Ken Cho, a tech entrepreneur in Austin, Texas, has been nicknamed the "Forrest Gump of financial disasters." You might recall that, in the film, Forrest Gump ends up witnessing some of the most important historical events of the 20th century. Similarly, Ken Cho had a ringside seat for two of the biggest and most startling bankruptcies in history. Before starting his own company, Mr. Cho worked in finance at both Enron and Lehman Brothers—two associations that, taken together, make for a strikingly unlucky CV.

If you started chatting with a random person on an airplane and discovered that he or she had worked at both Enron and Lehman Brothers, you'd probably be a bit surprised at such poor luck. But what kind of probability might we reasonably associate with this coincidence?

Understanding independence

Many common errors in reasoning about probability boil down to a misunderstanding of *independence*. You may recall this concept from several pages ago. In probability theory, we say that two events are independent if they convey no information about each other: $P(A) = P(A \mid B) = P(A \mid \text{not } B)$.

When we're doing probability calculations, we sometimes *assume* that two events are independent of each other, especially when the causes of the events are believed to be unrelated. If you're flipping coins or rolling dice, this seems reasonable.[22]

But in other cases, you can get tripped up by naïvely assuming independence. Life is full of lurking variables that can induce unexpected correlations, often in surprisingly subtle ways. In 2012, for example, data scientists at the predictive-analytics firm Kaggle claimed to discovered that, based on an analysis of the used-car market, orange used cars are more dependable than used colors in less flamboyant colors. The hypothesis was that owners of orange cars tend to be more devoted to their cars than the average person, and that this difference shows up in the reliability statistics.[23] As companies trawl ever larger data sets to look for patterns, more of these bizarre correlations are bound to come to our attention. Many will be spurious, but at least some will be real.

Let's take the example of Ken Cho's unlucky CV. How small is $P(E, L)$, the joint probability that a randomly chosen American

[22] Although the mathematician Persi Diaconis is somewhat famous (among math dorks and magicians) for being able to flip a coin very non-independently, so that it comes up heads every time. He is also the person who is usually credited for the mathematical result that it takes seven shuffles to adequately randomize a deck of cards.

[23] "Big Data Uncovers Some Weird Correlations." Deborah Gage, *Wall Street Journal* online edition, March 23, 2014.

adult worked at both Enron and Lehman Brothers?

You might naïvely calculate it as follows. There were about 200 million working-age Americans throughout the period in question (2001–7). At the time of their implosions, Enron and Lehman Brothers had about 20,000 and 26,000 employees, respectively. Therefore, if we assume that these events are independent, we might estimate $P(E, L)$ as

$$\begin{aligned} P(E, L) &= P(E) \cdot P(L) \\ &\approx \frac{20,000}{200,000,000} \cdot \frac{26,000}{200,000,000} \quad \approx 1.3 \times 10^{-8}, \end{aligned}$$

or about 1 in 100 million. This looks pretty unusual! If this is true, we would only expect that there are two such people (200 million adults, times a probability of 1 in 100 million) in the entire country.

However, the deal-breaking flaw in this calculation is the assumption of independence. For non-independent events, Rule 4 (for conditional probabilities) says that

$$P(E, L) = P(E) \cdot P(L \mid E).$$

That is, we should be using the conditional probability that someone worked for Lehman Brothers (L), given that they also worked in finance for Enron (E). These two events are far from independent. Once we condition on event E, we know that Mr. Cho worked in the finance industry, making it much more likely that he would also have held a job at Lehman Brothers. A rough estimate is that there were about 2 million professionals working in this sector of the finance industry at the time, so that the correct denominator in $P(L \mid E)$ is more like 2 million, not 200 million.[24] Therefore, a better estimate for $P(E, L)$ would be

$$\begin{aligned} P(E, L) &= P(E) \cdot P(L \mid E) \\ &\approx \frac{20,000}{200,000,000} \cdot \frac{26,000}{2,000,000} \quad \approx 1.3 \times 10^{-6}, \end{aligned}$$

or more like one in a million. In the context of a country as large as the U.S., this no longer looks unusual. In fact, since there are 200 million working-age adults, we would actually expect that there are about 200 such Forrest Gumps out there who held jobs at both Lehman and Enron.

[24] The real denominator is probably even smaller than 2 million, because these were considered *excellent* jobs in the finance industry, and candidates for them would have been drawn from a smaller pool. But we're just going for a ballpark figure here.

Another example: colorblindness. Colorblindness runs in families. In fact, there is at least one family out there[25] in which there are

[25] Mine.

7 male cousins from three different sets of parents—four Garvey brothers, two Wappler brothers, and one Scott—all of whom of are red-green colorblind. Christmas with this family involves some notably poor choices of sweaters, chromatically speaking.

How unlikely is it that all 7 male cousins in an extended family will be colorblind? Working this out exactly gets a bit tedious. So instead, we'll use the rule for joint probabilities to calculate a simpler probability: the chance that a randomly selected pair of brothers from the U.S. population will be be red-green colorblind. Let A indicate that the first brother is colorblind, and B that the second brother is colorblind. To calculate $P(A, B)$, we need to properly account for non-independence. That is, we need both $P(A)$ and $P(B \mid A)$.

The quantity $P(A)$ is the probability that any randomly selected U.S. male will be red-green colorblind. It's known that about 8% of men are red-green colorblind, so we'll take $P(A) = 0.08$.

What about $P(B \mid A)$? Remember, we are conditioning on the knowledge that the first brother is colorblind, and since colorblindness is genetic, $P(B \mid A)$ will be larger than 0.08. Specifically, colorblindness is an X-linked trait: a colorblind male must have inherited an X chromosome from his mother that contains the colorblindness gene.[26] To make things simple, let's assume that the brothers' mother has normal color vision (which is true of 99.5% of women). If this is the case, the only way the first brother could be colorblind is if their mother has one normal X chromosome, and one X chromosome with the colorblindness gene. The second brother inherits one of his mother's two X chromosomes; either X is equally likely. From this, we can deduce that $P(B \mid A) = 0.5$.

[26] This is why colorblindness is so much rarer in women than in men. Men have only one X chromosome, and so they need only one copy of the gene to end up colorblind. But females need two copies of the gene, one on each X chromosome, to end up colorblind. This is much less likely.

Putting these facts together, we find that

$$\begin{aligned} P(\text{both brothers colorblind}) &= P(\text{1st brother colorblind}) \cdot P(\text{2nd brother colorblind} \mid \text{1st brother colorblind}) \\ &= 0.08 \cdot 0.5 \\ &= 0.04\,. \end{aligned}$$

So about 4% of all pairs of brothers will be colorblind. For a family of four boys, the probability drops to $0.08 \cdot 0.5^3$, or 1%.

Independence and the "hot hand"

Less obviously, we can sometimes check whether events A and B are independent, even if we don't know this to be true beforehand. We can do this by carefully observing their frequencies

	Frequency of made shots after. . .						
Player	3 misses	2 misses	1 miss	overall	1 hit	2 hits	3 hits
Clint Richardson	0.5	0.47	0.56	0.5	0.5	0.49	0.48
Julius Erving	0.52	0.51	0.51	0.52	0.52	0.53	0.48
Lionel Hollins	0.5	0.49	0.46	0.46	0.46	0.46	0.32
Maurice Cheeks	0.77	0.6	0.6	0.54	0.56	0.55	0.59
Caldwell Jones	0.5	0.48	0.47	0.43	0.47	0.45	0.27
Andrew Toney	0.52	0.53	0.51	0.4	0.46	0.43	0.34
Bobby Jones	0.61	0.58	0.58	0.47	0.54	0.53	0.53
Steve Mix	0.7	0.56	0.52	0.48	0.52	0.51	0.36
Daryl Dawkins	0.88	0.73	0.71	0.58	0.62	0.57	0.51

Table 1.4: Data on the "hot hand" phenomenon for the 1980–81 Philadelphia 76ers. Keep in mind that some of the sample sizes used to calculate these frequencies are quite small, and thus potentially non-representative.

and verifying whether $P(A) = P(A \mid B)$. A good example here is related to the "hot-hand" or "winning streak" phenomenon in sports. Basketball fans in particular—and even coaches, players, and broadcasters—tend to believe in the hot hand: that if a player makes a shot, then he or she is more likely to make the next shot.[27] To express this idea in an equation, believers in the hot hand would assert that

$$P(\text{makes 2nd shot} \mid \text{makes 1st}) > P(\text{makes 2nd shot} \mid \text{misses 1st}).$$

You can imagine analogous formulations of this idea in walks of life other than sports, from picking stocks to playing poker to creating hit songs or viral videos.

So is the hot hand actually real? This turns out to be a surprisingly tricky question to answer. In a famous study from 1985, three economists looked at data from both the NBA and college basketball and concluded as follows:

> Detailed analyses of the shooting records of the Philadelphia 76ers provided no evidence for a positive correlation between the outcomes of successive shots. The same conclusions emerged from free-throw records of the Boston Celtics, and from a controlled shooting experiment with the men and women of Cornell's varsity teams.[28]

For example, Table 1.4 shows the authors' data for the 9 players on the 1980–81 Philadelphia 76ers. The table shows how frequently players made shots after streaks of different lengths (e.g. after 2 hits in a row, or after 1 miss). For example, Julius Erving made 52% of his shots overall, 52% of his shots after 1 made

[27] A popular video game from the 1990s, NBA Jam, immortalized this idea for anyone of that era. If you made three shots in a row, a game announcer would bellow "He's on fire!!" Your basketball avatar would temporarily be granted otherwordly speed, hops, and accuracy—and yes, the ball would actually be on fire whenever you touched it.

[28] Gilovich, Thomas; Tversky, A.; Vallone, R. (1985). "The Hot Hand in Basketball: On the Misperception of Random Sequences." *Cognitive Psychology* 3 (17): 295–314.

basket, and 48% of his shots after 3 made baskets—no "hot hand" at all. If you examine the table closely, you'll find that there's not much evidence for the hot-hand hypothesis for any of the players.[29] If anything, the evidence seems to go the other way: that most players on the 76ers were less likely to make a shot after 2 or 3 made baskets. Ironically, this might reflect the fact that the players themselves believed in the hot-hand phenomenon: if a player who fancies himself "hot" starts to take riskier shots, his shooting percentage will predictably drop.

[29] The authors of the 1985 study verified this using formal statistical hypothesis tests.

However, some recent studies have questioned both the methods and the conclusions of the original 1985 study. For example, two researchers at Stanford analyzed data from Major League Baseball and claimed to have found robust evidence for a hot hand across many different statistical categories.[30] Their basic argument is that in most sports, it would be really hard to find evidence for the hot-hand phenomenon, even if it really existed. The reason: defenses adapt. For example, as they point out, "a hot shooter in basketball should be defended more intensely. . . which will lower his shooting percentage." This happens a lot less in baseball, where the scope for defensive adaptation is limited.

[30] The Hot-Hand Fallacy: Cognitive Mistakes or Equilibrium Adjustments? Evidence from Major League Baseball. Brett Green and Jeffrey Zwiebel, Stanford University 2013. As of this writing, the paper had not been peer reviewed.

So the jury is still out on the hot-hand phenomenon in sports. However, it seems fair to say that any effects that might be found in the future are likely to be small, given that nobody else has found them yet despite looking pretty intensely.

The law of total probability

THE law of total probability sounds impressive, but is actually quite simple. It says: the probability of any event is the sum of the probabilities for all the ways in which the event can happen.[31]

[31] The law of total probability is really just Kolmogorov's third rule in disguise. The distinct ways in which some event A can happen are mutually exclusive. Therefore we just sum all their probabilities together to get $P(A)$.

This is most easily explained by example. Suppose that we're looking at houses on the Austin real estate market. The matrix in Table 1.5 shows estimates for the joint probabilities for various bedroom/bathroom combinations, using 2015 real-estate data. You can check that the probabilities sum to 1 across the entire matrix, as they must, by Kolmogorov's first rule.

Now consider the event "a house has 3 bedrooms." In studying the matrix of joint probabilities, we see that a house can have three bedrooms in four distinct ways: with one, two, three, or four bathrooms. Each of these combinations (3 bedrooms, x bathrooms)

	Bathrooms				
Bedrooms	1	2	3	4	Marginal
1	0.003	0.001	0.000	0.000	0.004
2	0.068	0.113	0.020	0.000	0.201
3	0.098	0.249	0.126	0.004	0.477
4	0.015	0.068	0.185	0.015	0.283
5	0.002	0.005	0.017	0.006	0.030
6	0.001	0.001	0.002	0.001	0.005
Marginal	0.187	0.437	0.350	0.026	

Table 1.5: Joint probabilities for various combinations of bedrooms and bathrooms for homes on the Austin real-estate market in the fall of 2015.

has a specific joint probability. By the law of total probability, we add these joint probabilities together to get the total (or marginal) probability of 3 bedrooms:

$$\begin{aligned} P(\text{3 bed}) &= P(\text{3 bed, 1 ba}) + P(\text{3 bed, 2 ba}) + P(\text{3 bed, 3 ba}) + P(\text{3 bed, 4 ba}) \\ &= 0.098 + 0.249 + 0.126 + 0.004 \\ &= 0.477\,. \end{aligned}$$

A formal statement of the law. More generally, suppose that events $B_1, B_2, \ldots, B_N$ constitute an exhaustive partition of all possibilities in some situation. That is, the events themselves are mutually exclusive, but one of them must happen. This can be expressed mathematically as

$$P(B_i, B_j) = 0 \text{ for any } i \neq j, \quad \text{and} \quad \sum_{i=1}^{N} P(B_i) = 1\,. \tag{1.3}$$

Now consider any event A. If Equation 1.3 holds, then

$$\begin{aligned} P(A) &= \sum_{i=1}^{N} P(A, B_i) \\ &= \sum_{i=1}^{N} P(B_i) \cdot P(A \mid B_i)\,. \end{aligned} \tag{1.4}$$

Equation 1.4 is what is usually called the law of total probability.

An example: asking embarrassing questions under an invisibility cloak

Suppose that you want to learn about the prevalence of drug use among college students. You decide to conduct a survey at a large state university to find out how many of the students there have

smoked marijuana in the last year. But there's an obvious problem here, common to a lot of survey research: if you ask people questions about sex or drugs, they often just lie. This makes your survey difficult to interpret.

Here's a cute trick for getting around this problem by cleverly exploiting the law of total probability. Suppose that, instead of asking people point-blank about their marijuana habits, you give them the following instructions.

1. Flip a coin. Look at the result, but keep it private.
2. If the coin comes up heads, please use the space provided to write an answer to question Q1: "Is the last digit of your Social Security number odd?"
3. If the coin comes up tails, please use the space provided to write an answer to question Q2: "Have you smoked marijuana in the last year?"

The key fact here is that only the respondent knows which question he or she is answering. This gives people plausible deniability. Someone answering "yes" might have easily flipped heads and answered the first, innocuous question rather than the second, embarrassing one, and the designer of the survey would never know the difference.

Moreover, despite the partial invisibility cloak we've provided to the marijuana users in our sample, we can still use the results of the survey to answer the question we care about: what fraction of students have used marijuana in the past year? We'll use the following notation:

- Let Y be the event "a randomly chosen student answers yes."
- Let Q_1 be the event "the student provided an answer to question 1, about their Social Security number."
- Let Q_2 be the event "the student provided an answer to question 2, about their marijuana use."

From the survey, we have an estimate of $P(Y)$, which is the overall fraction of survey respondents providing a "yes" answer. We really want to know $P(Y \mid Q_2)$, the probability that a randomly chosen student will answer "yes", given that he or she was answering the marijuana question. The problem is that we don't know which students were answering the marijuana question!

However, we can use the law of total probability, as follows:

$$P(Y) = P(Q_1) \cdot P(Y \mid Q_1) + P(Q_2) \cdot P(Y \mid Q_2)\,.$$

We know that $P(Q_1) = P(Q_2) = 0.5$, since a coin flip was used to determine whether Q_1 or Q_2 was answered. Moreover, we also know that $P(Y \mid Q_1) = 0.5$, since it is equally likely that someone's Social Security number will end in an even or odd digit.[32]

We can use this information to simplify the equation above:

$$P(Y) = 0.5 \cdot 0.5 + 0.5 \cdot P(Y \mid Q_2)\,,$$

or equivalently,

$$P(Y \mid Q_2) = 2 \cdot \{P(Y) - 0.25\}\,.$$

Suppose, for example, that 35% of survey respondents answer yes, so that $P(Y) = 0.35$. This implies that

$$P(Y \mid Q_2) = 2 \cdot (0.35 - 0.25) = 0.2\,.$$

We therefore estimate that about 20% of students have smoked marijuana in the last year.

[32] This survey design relies upon the fact that the survey designer doesn't know anyone's Social Security number. If you were running this survey in a large company, where people's SSNs were actually on file, you'd need to come up with some other innocuous question whose answer was unknown to the employer, but for which $P(Y \mid Q_1)$ was known.

2
Updating probabilities

The 2014 mens' final at Wimbledon, between Novak Djokovic and Roger Federer, was one of the most anticipated tennis matches in years. Djokovic, at 27 years old, was the top-ranked player in the world and at the pinnacle of the sport. And Federer was—well, Federer! His effortless grace on court had carried him to 17 prior Grand Slam titles, seven of them at Wimbledon, and he was widely regarded as the greatest player in history.[1] Even at 32 years old and a bit past his prime, Federer was ranked #3 in the world, and had been enjoying a superlative run of form leading up to the final. In the minds of most tennis fans, this might be Federer's last best shot at an 18th Grand Slam title.

[1] As someone who grew up a big Pete Sampras fan, even I must agree!

But as world #1, Djokovic was still favored to win the match. To be specific, if you walked into any betting shop in Britain just before the match started, you would been quoted odds of 20/13 on a Federer victory.[2] Recall our discussion of odds: this means you would need to stake a $13 bet in order to win a profit of $20. Using the conversion formula for odds and probabilities given earlier, this implies that the bookies were assessing the pre-match probability of a Federer victory as just shy of 40%:

[2] There are approximately 9,000 betting shops in the United Kingdom. In fact, it is estimated that approximately 4% of all retail storefronts in England are betting shops.

$$P(\text{Federer wins match}) = \frac{1}{\frac{20}{13}+1} \approx 0.4\,.$$

Of course, that was before the match. If you waited until the end of the first set before walking into the betting shop, you would have been quoted odds of 2/3, meaning you would have need to post a $3 stake to win a $2 profit. From this we can deduce (using the conversion formula) that the probability of a Federer victory was now assessed to be 60%. What happened?

The big swing in odds reflected the fact that Federer had won the first set, 7-6, in a closely fought tie break. Thus the new 60% figure represented something fundamentally different: it was the *conditional probability* that Federer would win the match, given that

he had won the first set:

$$P(\text{Federer wins match} \mid \text{Federer wins 1st set}) = \frac{1}{\frac{2}{3}+1} = 0.6\,.$$

The change from 40% to 60% in the probability of a Federer victory reflects the new information that Federer had won the first set. This left Djokovic in a hole: he would need to win 3 of the remaining 4 sets, while Federer would only need to win 2 of them.[3]

[3] In the end, Djokovic won the match in 5 sets to capture his second Wimbledon title. He would also go on to win a rematch against Federer in the 2015 Wimbledon final.

Bayes' rule

Conditional probability measures how likely some event A is to occur, given that some other event B has occurred. It stands to reason, therefore, that conditional probabilities must change to reflect new information or events—like the information that Federer has won the first set in a five-set tennis match.

This leads us to the most interesting and useful rule in probability theory: Bayes' rule, which gives us a concrete recipe for updating conditional probabilities to explicitly incorporate the effect of new information. Bayes' rule sits at the heart of the information economy. It is used almost everywhere, from search engines to spam filters to self-driving cars. Bayes' rule has also played a major part in many fascinating episodes throughout history. Alan Turing used Bayes's rule to help break secret Nazi codes. The U.S. Navy used it to help find the sunken USS Scorpion, a nuclear submarine lost at sea in June of 1968. It is a very powerful rule.

Updating conditional probabilities

Conditional probability statements are what doctors, judges, weather forecasters try to make every day of their lives, as they take in new 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 are always contingent upon what we know. When our knowledge changes, they too must change—and Bayes' rule tells us how to change them. 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?

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:

$$P(A \mid B) = \frac{P(A) \cdot P(B \mid A)}{P(B)}\,. \tag{2.1}$$

Each piece of this equation has a name:

Figure 2.1: 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."

- $P(A)$ is the prior probability: how probable is A, before ever having seen data B?
- $P(A \mid B)$ is the posterior probability: how probable is A, now that we've 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? This calculation is usually the tedious part of applying Bayes' rule, and it is done using the law of total probability:

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

An example: have you found the two-headed coin?

Let's see an example of Bayes' rule in action. Imagine a jar with 1024 normal quarters. Into this jar, a friend places a single two-headed quarter (i.e. with heads on both sides). Your friend then gives the jar a good shake to mix up the coins. You draw a single coin at random from the jar, and without examining it closely, flip the coin ten times. The coin comes up heads all ten times. Are you holding the two-headed quarter, or an ordinary quarter?

Let's use D to denote the data that the coin came up heads ten times in a row. We will use Bayes' rule to compute $P(T \mid D)$, the probability that you are holding the two-headed quarter (T), given the data:

$$P(T \mid D) = \frac{P(T) \cdot P(D \mid T)}{P(D)}\,.$$

Let's take this equation one piece at a time. First, what is $P(T)$, the prior probability that you are holding the two-headed quarter? Well, there are 1025 quarters in the jar: 1024 ordinary ones, and one two-headed quarter. Assuming that your friend mixed the coins in the jar well enough, then you are just as likely to draw one coin as another, and so $P(T)$ must be $1/1025$.

Next, what about $P(D \mid T)$, the likelihood of flipping ten heads in a row, given that you chose the two-headed quarter? Clearly this is 1—there is no possibility of seeing anything else.

Finally, what about $P(D)$, the marginal probability of flipping ten heads in a row? As is almost always the case when using Bayes' rule, $P(D)$ is the hard part to calculate. We will use the

law of total probability to do so:

$$P(D) = P(T) \cdot P(D \mid T) + P(\sim T) \cdot P(D \mid \sim T)\,.$$

Taking the pieces on the right-hand one by one:

- As we saw above, the prior probability of the two-headed coin, $P(T)$, is $1/1025$.
- This means that the prior probability of an ordinary coin, $P(\sim T)$, must be $1024/1025$.
- Also from above, we know that $P(D \mid T) = 1$.
- Finally, we can calculate $P(D \mid \sim T)$ quite easily. If the coin is an ordinary quarter, then there is a 50% chance of getting heads on any one coin flip. Each flip is independent. Therefore,

$$\begin{aligned} P(D \mid \sim T) &= \frac{1}{2} \times \frac{1}{2} \times \cdots \times \frac{1}{2} \quad \text{(10 times)} \\ &= \left(\frac{1}{2}\right)^{10} = \frac{1}{1024}\,. \end{aligned}$$

We can now put all these pieces together:

$$\begin{aligned} P(T \mid D) &= \frac{P(T) \cdot P(D \mid T)}{P(T) \cdot P(D \mid T) + P(\sim T) \cdot P(D \mid \sim T)} \\ &= \frac{\frac{1}{1025} \cdot 1}{\frac{1}{1025} \cdot 1 + \frac{1024}{1025} \cdot \frac{1}{1024}} = \frac{1/1025}{2/1025} \\ &= \frac{1}{2}\,. \end{aligned}$$

Perhaps surprisingly, there is only a 50% chance that you are holding the two-headed coin. Yes, flipping ten heads in a row with a normal coin is very unlikely. But so is drawing the one two-headed coin from a jar of 1024 normal coins! In fact, as the math shows, both explanations for the data are equally unlikely, which is why we're left with a posterior probability of 0.5.

Two-headed coins in the wild. OK, you might be thinking that our example involving the two-headed coin is pretty artificial. But it's not. In fact, in the real world, an awful lot of time and energy is spent looking for metaphorical two-headed coins—for example, in any industry where companies compete strenuously for talented employees. To see why, let's change the story just a little bit.

Suppose you're in charge of a large trading desk at a major Wall Street bank. You have 1025 employees under you, and each one is responsible for managing a portfolio of stocks to make money for your firm and its clients.

One day, a young trader knocks on your door and confidently asks for a big raise. You ask her to make a case for why she deserves one. She replies:

> Look at my trading record. I've been with the company for ten months, and in each of those ten months, my portfolio returns have been in the top half of all the portfolios managed by my peers on the trading floor. If I were just an average trader, this would be very unlikely. In fact, the probability that an average trader would see above-average results for ten months in a row is only $(1/2)^{10}$, which is less than one chance in a thousand. Since it's unlikely I would be that lucky, the implication is that I am a talented trader, and I should therefore get a raise.

The math of this scenario is exactly the same as the one involving the big jar of quarters. Metaphorically, the trader is claiming to be a two-headed coin (T): she performs above average, every single month without fail.

But from your perspective, things are not so clear. Is the trader lucky, or good? There are 1025 people in your office (i.e. 1025 coins). Now you're confronted with the data that one of them has had an above-average monthly return for ten months in a row (i.e. D = "flipped heads ten times in a row"). This is admittedly unlikely, and this person might therefore be an excellent performer, worth paying a great deal to retain. But excellent performers are probably also rare, so that the prior probability $P(T)$ is pretty small to begin with.

To make a decision, you need to know the posterior probability, $P(T \mid D)$. In light of what you know about Bayes' rule, which of the following replies is the most sensible?

(A) "You're right. Here's a giant raise."

(B) "Thank you for letting me know. While I need more data to give you a raise, you've had a good ten months. I'll review your case again in 6 months and will look closely at the facts you've showed me."

The best answer depends very strongly on your beliefs about whether excellent stock traders are common or rare. For example,

suppose you believe that 10% of all stock traders are truly excellent, in the sense that they can reliably finish with above-average returns, month after month; and that the other 90% just muddle through and collect their thoroughly average bonus checks. Then $P(T) = 0.1$, and

$$P(T \mid D) = \frac{0.1 \cdot 1}{0.1 \cdot 1 + 0.9 \cdot \frac{1}{1024}} \approx 0.991\,,$$

so that there is better than a 99% chance that your employee is among those 10% of excellent performers. You should give her a raise, or risk letting some other bank save you the trouble.

What if, however, you believed that excellence were much rarer, say $P(T) = 1/10000$? In that case,

$$P(T \mid D) = \frac{0.0001 \cdot 1}{0.0001 \cdot 1 + 0.9999 \cdot \frac{1}{1024}} \approx 0.093\,.$$

In this case, even though the ten-month hot streak was unusual—$P(D \mid\sim T)$ is small, at $1/1024$—there is still more than a 90% chance that your employee got lucky. This is inconsistent with the evidence, meaning that $P(T)$ is probably much smaller than 10%.

How rare are two-headed coins? The moral of the story is that the prior probability in Bayes' rule—in this case, the baseline rate of excellent stock traders, i.e. "two-headed coins"—plays a very important role in correctly estimating conditional probabilities. If ignore this prior probability, or conflate the likelihood $P(D \mid T)$ with the posterior probability $P(T \mid D)$, you may be setting yourself up to make a big mistake.

So just how rare are two-headed coins? While it's very difficult to know the answer to this question in something like stock-trading, it is worth pointing out one fact: in the above example, a prior probability of 10% is almost surely too large. If this prior probability were right, then out of your office of 1025 traders, you would expect there to be $0.1 \times 1025 \approx 100$ of them with 10-month winning streaks all at your door clamoring for a raise. (Traders are not known for being shy about their winning streaks, or anything else.) This line of reasoning demonstrates that, while the prior often reflects your own knowledge about the world, it can just as often be informed by data.

The best, or at least the most famous, two-headed coin in all of stock picking must be Warren Buffett, known as the "Oracle of Omaha." Over the last 50 years, Warren Buffett has beaten the

market so badly that it almost defies belief: between 1964 and 2013, the share price of his holding company, Berkshire Hathaway, has risen by about 1 million percent, versus only 2300% for the S&P 500 stock index.

Bayes' rule and the law

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 $P(G \mid D) = 0.09$.

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 $P(G \mid D) \approx 1/11$.

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,[4] or more generally the base-rate fallacy.[5] Words may mislead, but Bayes' rule never does!

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

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

TBD

Bayes' rule and medicine

Bayes search

USS Scorpion, Air France.