2
Probability and conditional probability

Probability is a rich language for communicating about uncertainty. And while most of us have an intuitive notion of what it means, it pays to be a bit more specific.

A probability is just a number that measures how likely it is that some event, like rain, will occur. If A is an event, $P(A)$ is its probability: $P(\text{coin lands heads}) = 0.5$, $P(\text{rainy day in Ireland}) = 0.85$, $P(\text{cold day in Hell}) = 0.0000001$, and so forth.

Probability from frequencies. Some probabilities are derived from data, like the knowledge that a coin comes up heads about 50% of the time in the long run, or that 11 people out of 100,000 die in a car accident. But it's also perfectly normal for a probability to reflect your subjective assessment or belief about something. Here, you should imagine a stock-market investor who has to decide whether to buy a stock or sell it. The performance of a stock over the coming months and years involves a bunch of one-off events that have never happened before, and will never be repeated. But that's OK. We can still talk about a probability like $P(\text{Apple stock goes up next month})$. We just have to recognize that this probability reflects someone's subjective judgment, rather than a long-run frequency from some hypothetical coin-flipping experiment.

Probability from judgment and/or betting markets. If you don't have any data, a great way to estimate the probability of some event is to get people to make bets on it. Let's take the example of the 2014 mens' final at Wimbledon, between Novak Djokovic and Roger Federer. This 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! Even at 32 years old and a bit past his prime, he was ranked #3 in the world, and had been in vintage form leading up

to the final.

How could you synthesize all this information to estimate a probability like $P(\text{Federer wins})$? Well, 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.[1] To interpret odds in sports betting, think "losses over wins." That is, if Federer and Djokovic played 33 matches, Federer would be expected to win 13 of them and lose 20, meaning that

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

The markets had synthesized all the available information for you, and concluded that the pre-match probability of a Federer victory was just shy of 40%. (Djokovic ended up winning in five sets.)

[1] 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.

Conditional probability

Another very important concept is that of a *conditional probability.* A conditional probability is the chance that some event A happens, given that another event B happens. We write this as $P(A \mid B)$ for short, where the bar $(\mid)$ means "given" or "conditional upon."

We're all accustomed to thinking about conditional probabilities in our everyday lives, even if we don't do so quantitatively. For example:

- $P(\text{rainy afternoon} \mid \text{cloudy morning})$,
- $P(\text{rough morning} \mid \text{out late last night})$,
- $P(\text{rough morning} \mid \text{out late last night, drank extra water})$,

and so forth. As the last example illustrates, it's perfectly valid to condition on more than one event.

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:

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

We'll see a few examples later where people get this wrong, and act as if $P(A \mid B)$ and $P(B \mid A)$ are the same. Don't do this.

Conditional probabilities are used to make statements about uncertain events in a way that reflects our assumptions and our partial knowledge of a situation. They satisfy all the same rules as ordinary probabilities, and we can compare them as such. For example, we all know that

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

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

The rules of probability

Probability is an immensely useful language, and there are only a few basic rules (called "axioms" in mathematical language):

(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

$$P(A \text{ or } B) = P(A) + P(B).$$

There's also a fourth, slightly more advanced rule for conditional probabilities:

(4) Let $P(A, B)$ be the *joint probability* that both A and B happen. Then the conditional probability $P(A \mid B)$ is:

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

An equivalent way of expressing Rule 4 is to multiply both sides of the equation by $P(B)$, to yield

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

We can use these two versions interchangeably.

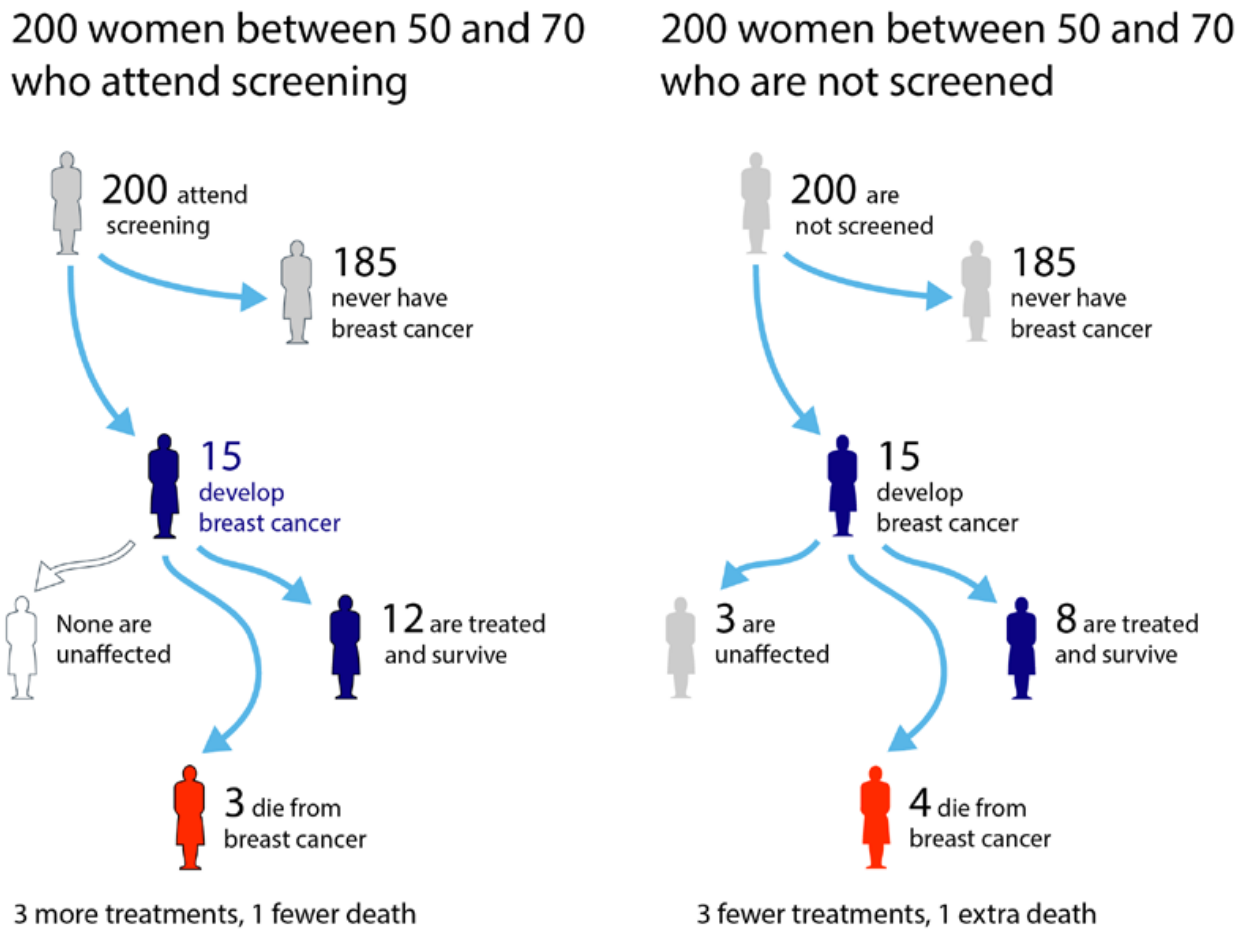

Figure 2.1: Two hypothetical cohorts of 200 women, ages 50-70. The 200 women on the left all go in for mammograms; the 200 on the right do not. The branches of the tree show how many women we would expect to experience various different outcomes. Figure from: "What can education learn from real-world 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.

To illustrate these rules, we'll turn to Figure 2.1, which is 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 2.1, 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

cohort are slightly different: on the right, we expect 1 fewer death, and 3 extra unnecessary screenings, versus the left.

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 ($15/200 = 0.075$). This is like the NP rule in reverse: if E is the expected value (here 15), then the probability is $P = E/N$.

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%.

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%.)

Let's explicitly calculate this using the rule conditional probability (Equation 2.1, on page 11) instead. The rule says

$$P(\text{survives} \mid \text{gets cancer}) = \frac{P(\text{gets cancer and survives})}{P(\text{gets cancer})}\,.$$

We'll take this equation piece by piece.

- Out of 200 women, we expect that 15 will develop cancer. This is the denominator in our equation:

$$P(\text{gets cancer}) = \frac{15}{200}\,.$$

- Out of 200 women, we expect that 11 will develop cancer and survive. This is the numerator in our equation:

$$P(\text{gets cancer and survives}) = \frac{11}{200}\,.$$

- Therefore, using the rule for conditional probability,

$$P(\text{survives} \mid \text{cancer}) = \frac{11/200}{15/200} = 11/15\,.$$

Example: Abraham Wald and the WWII Bombers

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. And the bad days were very bad. 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.[2]

Like Yossarian in *Catch-22*, World War II airmen were painfully aware that each combat mission was a role of the dice. What's more, they had to complete 25 missions to be sent home. With such poor chances of returning from a *single* mission, they 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.

[2] Numbers taken 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.

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.

Where should the military reinforce its planes?

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.

Figure 2.2: Abraham Wald.

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. Here's the problem Wald analyzed.[3] While some airplanes came back from bombing missions in Germany unscathed, many others had visibly taken hits from enemy fire. 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[4] 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

If you turn those frequencies into probabilities, so that the numbers sum to 1, you get the numbers in the table below.

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

[4] 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.

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.

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

But that's the right answer to the wrong question. Wald recognized that this number suffered from a crucial flaw: *it only included data on the survivors.* The planes that had been shot down were missing from the analysis—and only the pattern of bullet holes on those missing planes could definitively tell the story of a B-17's vulnerabilities.

Instead, he recognized that it was essential to calculate the *inverse* probability, namely

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

This might be a very different number. Remember: $P(\text{practices hard} \mid \text{plays in NBA}) \approx 1$, while $P(\text{plays in NBA} \mid \text{practices hard}) \approx 0$. Conditional probabilities aren't symmetric.

Of course, Wald had no data on the planes that had been shot down. Therefore, to actually calculate the probability $P(\text{returns safely} \mid \text{hit on wings or fuselage})$ 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 guessing. So Wald 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.[5]

Remarkably, when all was said and done, Wald was able to reconstruct an estimate for the *joint frequencies* 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

[5] 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.*)

bring their statistical signature back into the data set. For our hypothetical cohort of 800 bombers that took damage, Wald's best guess would have looked something like this:

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

Table 2.1: An example of how Abraham Wald could have reconstructed the joint frequency 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). You'll notice that the numbers in the left column correspond 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.

This estimate for the joint frequencies for two random outcomes, 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.

Postscript. In the story of Abraham Wald and the missing B-17s, 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 suggest otherwise. This implies that we should turn 180 degrees away from our intuition: if the planes are taking damage on the wings and the fuselage, then let's put the armor there instead. But that's wrong, and 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.

Conditional probability and recommender systems

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—for example, by assigning a film a five-star rating. 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.

As with Abraham Wald and the missing bombers, it's all about asking the right question. The key insight here is to frame the problem in terms of conditional probability. Suppose that, for some pair of films A and B, the probability $P(\text{random user likes } A \mid \text{random user likes } B)$ is high—say, 80%. Now we learn that Linda liked film B, but hasn't yet seen film A. Wouldn't A be a good recommendation? Based on her liking of A, there's an 80% chance she'll like it.

But how can we learn $P(\text{likes } A \mid \text{likes } B)$? This is where your database, coupled with the rule for conditional probability, comes in handy. Suppose that there are 5 million people in your database who have seen both *Saving Private Ryan* and *Band of Brothers*, and that the ratings data on these 5 million users looks like this:

	Liked *Band of Brothers*	Didn't like
Liked *Saving Private Ryan*	2.8 million	0.3 million
Didn't like	0.7 million	1.2 million

Once again, we have information on two random outcomes: A = whether a user liked *Saving Private Ryan*, and B = whether the user liked *Band of Brothers*. From this information, we can easily work out the conditional probability that we need. Of the 5 million users in the database who have watched both programs, 2.8 + 0.7 = 3.5 million of them liked *Band of Brothers*. Of these 3.5 million people, 2.8 million (or 80%) also liked *Saving Private Ryan*. Therefore,

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

Note that you could also jump straight to the math, and use the rule for conditional probabilities (Equation 2.1, on page 11), like this:

$$P(A \mid B) = \frac{P(A,B)}{P(B)} = \frac{2.8/5}{(2.8+0.7)/5} = 0.8\,.$$

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 very good (yet) at automatically scanning films for thematic content. But they're brilliant at calculating conditional probabilities from a vast database of users' movie-watching histories.

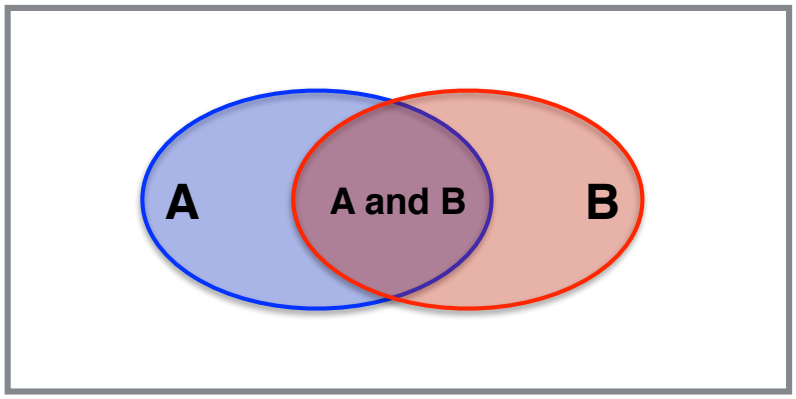

Figure 2.3: A Venn diagram depicting the addition rule for probabilities.

The addition rule

The addition rule tells us how to compute the probability of "either/or" type events. It's easy to do this for an either/or event like (Astros win World Series or Cubs win World Series). What makes this easy is that these two events are mutually exclusive. Therefore, to compute the probability of either one happening (i.e. the union of the two events), we sum the probabilities of the individual events.

But what about events that aren't mutually exclusive? We cannot, for example, directly use this rule to compute a probability such as $P([\text{snow on Christmas eve}] \text{ or } [\text{snow on Christmas day}])$, because it is possible for both events to occur in the same year.

For "either/or" events like these, we need a more general rule, usually called the *addition rule*. The addition rule says that, for any two events A and B, the probability that either A or B will happen is

$$P(A \text{ or } B) = P(A) + P(B) - P(A, B)\,, \tag{2.2}$$

where $P(A, B)$ is the joint probability of A and B.

It's easy to visualize the addition rule using a Venn diagram. Imagine throwing darts at the rectangular area in Figure 2.3. What is $P(A \text{ or } B)$—the probability that a randomly thrown dart will land either in the blue oval (A) or the red oval (B)?

If we assume that the outer rectangle has total area 1, then this probability is just the total area covered by the union of the two ovals ($A \cup B$). To calculate this area, we can start by adding the areas of the blue oval and red ovals together. But then we've

double-counted the purple area of overlap: once for the A oval, and once for the B oval. So we have to subtract this area back off, to correct for the double-counting:

$$\text{Area of } A \cup B = (\text{Area of } A) + (\text{Area of } B) - (\text{Area of } A \cap B)\,.$$

The overlap is written $A \cap B$ and read alound as "A intersect B."

You'll notice that this simple formula parallels the addition rule for probabilities,

$$P(A \text{ or } B) = P(A) + P(B) - P(A, B)\,,$$

which you can think of in exactly the same way: add the probabilities for the two events, then subtract off the bit that you've double-counted.