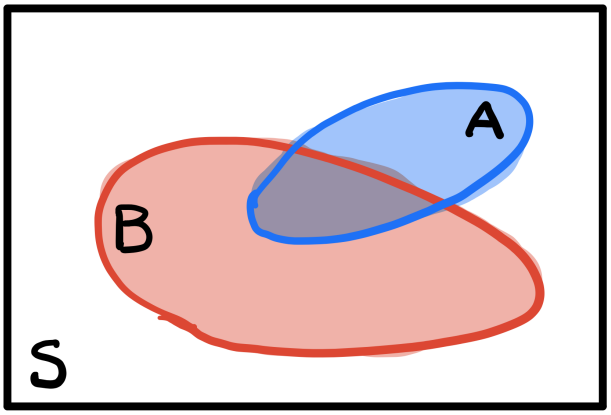

Figure 2.7

2.5 Conditional probability

Conditional probability answers the question "if I acquire knowledge that an event B occurred, how should this news change my assessment of the probability of another event A?" If I know B has occurred, then I am interested in computing a new probability $P(A \mid B)$, which is "the probability of A given B." Contrast this with regular ol' $P(A)$, which is the *marginal* or *unconditional* probability of A.

Definition 2.2. If $A, B \subseteq S$, then the **conditional probability** of A given B is

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \quad P(B) > 0. \tag{2.5}$$

To see what this definition is capturing, consider Figure 2.7. Because we *know* that B occurred, we can ignore everything outside of B. In particular, the only part of A that matters now is the part that intersects with B itself, hence the numerator. But if we are ignoring everything outside of B, then B has effectively become our new sample space, and the map $A \mapsto P(A \mid B)$ is a new probability measure on this revised space. This is why we divide by $P(B)$ in the definition, so that this new probability measure on the sample space B satisfies the axiom of total measure 1:

$$P(B \mid B) = \frac{P(B \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1.$$

Because the map $A \mapsto P(A \mid B)$ is a probability measure in its own right, it satisfies all the axioms and obeys all the rules (check this!).

2.5.1 The marginal/conditional decomposition

Because set intersection is a commutative operation, we have

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

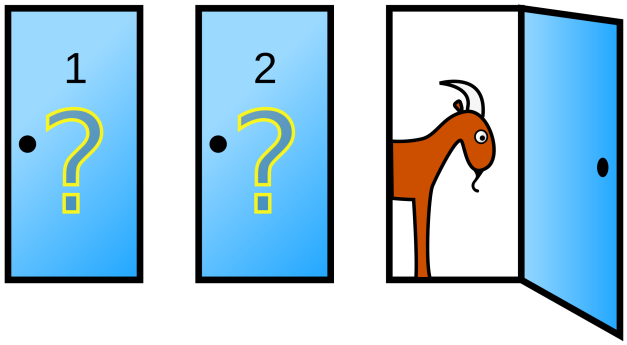

Figure 2.8

This, together with our original definition, provides us with "two identities for the price of one":

$$P(A \cap B) = P(A \mid B)P(B) = P(B \mid A)P(A).$$

I call this the **marginal/conditional decomposition**. It also goes by the *chain rule of probability* or the *multiplication rule*. In general, the probability of an intersection (a joint probability) can always be rewritten as the product of a marginal probability and a conditional probability. This is one of the most useful ways to compute the probability of an intersection, and one application of this is to rewriting the law of total probability:

$$P(A) = \sum_{i=1}^{\infty} P(A \cap B_i) = \sum_{i=1}^{\infty} P(A \mid B_i)P(B_i), \quad (B_i) \text{ is a partition of } S. \tag{2.6}$$

When you apply the law of total probability, you will probably use it in this second form.

2.5.2 Bayes' theorem

If we take our original definition of conditional probability and rewrite the numerator using a marginal/conditional decomposition, we get Bayes' theorem:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}. \tag{2.7}$$

This is useful because it is hyper-explicit about the exact relationship between the marginal probability $P(A)$ and the conditional probability $P(A \mid B)$. This formula tells you exactly how your beliefs about the event A ought to be updated or revised in light of the news that B occurred.

2.5.3 Example: the Monty Hall problem

This is a famous problem in probability, based on the game show *Let's Make a Deal* hosted by Monty Hall. In the game, a contestant is presented with three doors. Behind one of the doors is the grand prize of a new car, and behind each of the other two doors is a goat (visualized in Figure 2.8). The goal is to ultimately select the door that has the prize behind it. The game proceeds as follows:

(i) You select one of the doors to open;

(ii) Instead of opening the door you chose, the host will open one of the two remaining doors that has a goat behind it;

- If the door you first chose has the prize behind it, then the host will select between the two remaining doors with equal probability;
- If the door you first chose has a goat behind it, then of course the host will select the only other door with a goat behind it;

(iii) The host will then give you a choice: stick with the original door you selected, or switch to the other remaining one. What should you do?

When most people encounter this problem for the first time, they guess that there is a 50/50 chance that the prize is behind either of the two remaining, unopened doors, but this is false. Let's say without loss of generality that the contestant initially selects Door 1, and then the host opens Door 2 to reveal a goat. Once it is revealed that Door 2 has a goat behind it, Door 3 becomes *conditionally* more likely to conceal the prize. This is because there are circumstances under which the host would be actively avoiding Door 3 because it conceals the prize. If the original choice of Door 1 has a goat behind it, then the host is constrained not to pick the door with the prize inside. The fact that he has not opened Door 3 tips you off to the possibility that he is hiding something, so his revealing Door 2 makes Door 3 more likely, and your best bet is to switch.

In order to see this beyond a shadow of a doubt, we will use Bayes' theorem to compute and compare the following:

$$P(D1 \text{ prize} \mid D2 \text{ open}) = \frac{P(D2 \text{ open} \mid D1 \text{ prize})P(D1 \text{ prize})}{P(D2 \text{ open})}$$
$$P(D3 \text{ prize} \mid D2 \text{ open}) = \frac{P(D2 \text{ open} \mid D3 \text{ prize})P(D3 \text{ prize})}{P(D2 \text{ open})}.$$

To begin, we know that

$$P(D1 \text{ prize}) = P(D2 \text{ prize}) = P(D3 \text{ prize}) = \frac{1}{3}.$$

Next note that

$$\begin{aligned} P(D2 \text{ open} \mid D1 \text{ prize}) &= 1/2 \\ P(D2 \text{ open} \mid D2 \text{ prize}) &= 0 \\ P(D2 \text{ open} \mid D3 \text{ prize}) &= 1. \end{aligned}$$

The law of total probability says

$$\begin{aligned} P(D2 \text{ open}) &= P(D2 \text{ open} \mid D1 \text{ prize})P(D1 \text{ prize}) + \\ &\quad P(D2 \text{ open} \mid D2 \text{ prize})P(D2 \text{ prize}) + \\ &\quad P(D2 \text{ open} \mid D3 \text{ prize})P(D3 \text{ prize}) \\ &= \frac{1}{2} \cdot \frac{1}{3} + 0 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} \\ &= 1/2, \end{aligned}$$

	$D = +$	$D = -$
$T = +$	True Positive (TP)	False Positive (FP)
$T = -$	False Negative (FN)	True Negative (TN)

Table 2.4: Contingency table of test results vs. disease status

so we're ready to go:

$$\begin{aligned} P(D1 \text{ prize} \mid D2 \text{ open}) &= \frac{P(D2 \text{ open} \mid D1 \text{ prize})P(D1 \text{ prize})}{P(D2 \text{ open})} \\ &= \frac{(1/2)(1/3)}{1/2} \\ &= 1/3 \end{aligned}$$

$$\begin{aligned} P(D3 \text{ prize} \mid D2 \text{ open}) &= \frac{P(D2 \text{ open} \mid D3 \text{ prize})P(D3 \text{ prize})}{P(D2 \text{ open})} \\ &= \frac{1 \cdot (1/3)}{1/2} \\ &= 2/3. \end{aligned}$$

You are more likely to win if you switch.

2.5.4 Example: disease testing

We are all too familiar with disease testing nowadays. Tests work pretty well in a lot of cases, but they are not infallible. A test could say that you have a disease when you don't (false positive), or it could fail to detect a disease that you in fact have (false negative). Let's be more concrete. Let D denote someone's true disease status; they either have the disease (+) or they don't (-). Let T denote the result of someone's test; the test either says they have the disease (+), or it says they don't (-). We will assume that these are the only two attributes a person possesses, so our little model is ignoring important things like symptoms, genetics, medical history, lifestyle, etc. Table 2.4 displays the sample space of this phenomenon.

We give special names to the probabilities of the various outcomes:

- **prevalence**: $P(D = +)$ is the marginal probability of contracting the disease. If you had no other information, this is the baseline likelihood that a randomly selected member of the population is sick;
- **false negative rate (FNR)**: $P(T = - \mid D = +)$ is the probability of the test saying you don't have the disease when you d;
- **false positive rate (FPR)**: $P(T = + \mid D = -)$ is the probability of the test saying you have the disease when you don't
- **sensitivity (true positive rate)**: $P(T = + \mid D = +) = 1 - P(T = - \mid D = +)$ is the probability of the test correctly detecting that you have the disease;

- **specificity (true negative rate)**: $P(T = - \mid D = -) = 1 - P(T = + \mid D = -)$ is the probability of the test correctly detecting that you *don't* have the disease.

We would love for the sensitivity and the specificity of a test to both be high, but there is a trade-off between them. Up to a certain point, you cannot improve sensitivity without making specificity worse, and vice versa. You can always make either the sensitivity or specificity perfect, but it will be at the expense of the other:

- Imagine a test that automatically returns (+) for everyone that takes it. In that case, everyone with the disease will be correctly diagnosed, and so the sensitivity $P(T = + \mid D = +) = 1$ is perfect. But there are a lot of healthy people being diagnosed with a disease they don't have, and so the specificity $P(T = - \mid D = -) = 0$ is perfectly wretched;
- Conversely, imagine a test that automatically returns (-) for everyone that takes it. So everyone with the disease is being told they don't have it, and the sensitivity $P(T = + \mid D = +) = 0$ sucks. But everyone without the disease is correctly being told they're healthy, and so the specificity $P(T = - \mid D = -) = 1$ is now perfect.

Here's another analogy to drive it home. Think of D as someone's guilt (+) or innocence (-) of committing a crime, and think of T as the verdict of guilt (+) or innocence (-) rendered by a judge.

- If the judge wants to trivially max out sensitivity, she should just declare everyone before her guilty and send everyone to jail. In this way, no criminals are walking the street, but now you have a lot of innocent people locked up;
- If the judge wants to trivially max out specificity, she could let everyone go free. Now no innocent people are jailed, but you have criminals on the loose.

We seek the "Goldilocks sweet spot" between these two absurd extremes, but there's a real tradeoff. A "perfect test" would have sensitivity and specificity both equal to 1, but if the outcomes are truly random (meaning not perfectly predictable), then a perfect test is not possible. You'll always make some mistakes some of the time.

Example 2.8. Imagine we had a test with sensitivity $P(T = + \mid D = +) = 0.88$ and false positive rate $P(T = + \mid D = -) = 0.14$, and a disease with prevalence $P(D = +) = 0.01$. You test positive for this disease, and you want to know what the chance is that the test is wrong. In particular, you want to calculate $P(D = - \mid T = +)$. We know from Bayes' theorem that this is equal to:

$$P(D = - \mid T = +) = \frac{P(T = + \mid D = -)P(D = -)}{P(T = +)}.$$

We are given $P(T = + \mid D = -) = 0.14$ and $P(D = -) = 1 - P(D = +) = 1 - 0.01 = 0.99$, so we are done if we can calculate $P(T = +)$. The way to do this is with the law of total probability. Table 2.4 is our sample space, and we will partition it according to the value of D. So the law of

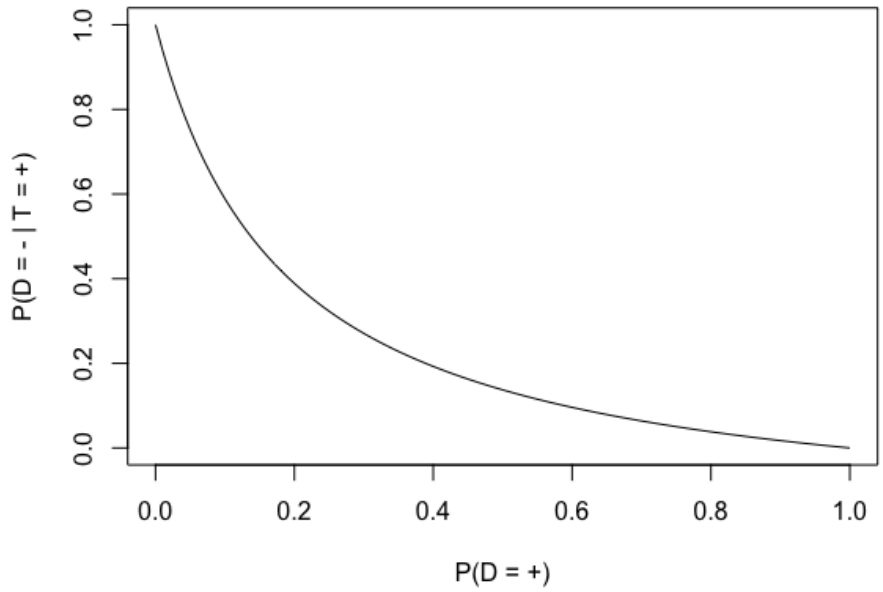

Figure 2.9: As the disease becomes less rare, the test is less likely to be wrong.

total probability[1] allows us to write:

$$\begin{aligned} P(T=+) &= P(T=+ \cap D=-) + P(T=+ \cap D=+) \\ &= P(T=+ \mid D=-)P(D=-) + P(T=+ \mid D=+)P(D=+) \\ &= 0.14 \cdot 0.99 + 0.88 \cdot 0.01 \\ &= 0.1474. \end{aligned}$$

So we have

$$P(D=- \mid T=+) = \frac{P(T=+ \mid D=-)P(D=-)}{P(T=+)} = \frac{0.14 \cdot 0.99}{0.1474} \approx 0.94.$$

So despite the fact that the sensitivity (0.88) and the specificity (1 - 0.14 = 0.86) of the test are fairly high, your test result has a high probability of being wrong. The main reason for this is that the disease is so rare. Before you got tested, the probability that you had the disease was $P(D=-) = 0.01$. Now that you have been tested, the revised probability is $P(D=+ \mid T=+) = 1 - P(D=- \mid T=+) \approx 0.06$. So your chances have definitely gone up in light of the new information about your test result, which is bad news, but it's hardly a death sentence. Figure 2.9 displays the value of $P(D=- \mid T=+)$ for different values of the prevalence.

2.6 Independent events

[1] In the notation of Section 2.3.4, the event of interest is $A = \{T=+\}$ and our partition of the sample space consists of $B_1 = \{D=-\}$ and $B_2 = \{D=+\}$.

Definition 2.3. Two events $A,\ B \subseteq S$ are **independent** if $P(A \cap B) = P(A)P(B)$.

This just sort of comes out of nowhere, but it has a nice interpretation. Consider two independent events $A,\ B \subseteq S$. If you observe that B occurs, then the conditional probability of A is

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A). \tag{2.8}$$

So your probability assessment of the event A was not changed by the news that B occurred. These events "live separate lives," and knowledge of one does not teach you anything about the likelihood of the other.

Example 2.9. Recall the example of rolling two fair, six-sided die. The outcomes are visualized in Table 2.2. Next consider the following events:

A = "Die 1 is less than or equal to 3"
B = "Die 2 is equal to 1"
C = "The sum of the rolls is 10"
D = "The sum of the rolls is 7"
E = "The rolls are the same (doubles)"
F = "The sum is even".

Since the dice are fair, outcomes are equally likely, so by inspecting Table 2.2 and counting, we know that $P(A) = 18/36 = 1/2$, $P(B) = 6/36$, $P(C) = 3/36$, $P(D) = 6/36$, $P(E) = 6/36$, and $P(F) = 18/36 = 1/2$. Furthermore, we see that $P(A \cap B) = 3/36$, $P(A \cap C) = 0$, $P(A \cap D) = 3/36$, and $P(E \cap F) = 6/36$. As a consequence, we see that

$$\begin{aligned}
P(A \mid B) &= \frac{P(A \cap B)}{P(B)} = \frac{3/36}{6/36} = 1/2 = P(A) \\
P(A \mid C) &= \frac{P(A \cap C)}{P(C)} = \frac{0}{3/36} = 0 \neq P(A) \\
P(A \mid D) &= \frac{P(A \cap D)}{P(D)} = \frac{3/36}{6/36} = 1/2 = P(A) \\
P(E \mid F) &= \frac{P(E \cap F)}{P(F)} = \frac{6/36}{18/36} = 1/3 \neq P(E) \\
P(F \mid E) &= \frac{P(E \cap F)}{P(E)} = \frac{6/36}{6/36} = 1 \neq P(F).
\end{aligned}$$

So A and B are independent events. This makes sense because the dice are rolled separately and their outcomes should not affect one another. So knowing that Die 2 is a 1 should not teach you anything about Die 1. A and C are dependent events, because they are mutually exclusive. If you know that you rolls sum to ten, that guarantees that Die 1 is greater than a three. A and D are independent. If you roll a double, that necessarily means that your sum is also even, so in fact $E \subseteq F$, and these events are quite *dependent*.

Remark 2.4. Something that inevitably confuses folks is the difference between disjoint and independent events. The colloquial meanings of those words make it feel like the concepts would be similar, but they are actually quite different. Recall:

- **Disjoint events**: they cannot occur at the same time, so $A \cap B = \varnothing$;
- **Independent events**: knowledge of one does not teach you anything about the other, so $P(A \mid B) = P(A)$.

Right off the bat, one difference we see is that disjointedness is a property of the sets themselves. Independence is a property of the probability measure and how it factors. So disjoint events remain disjoint regardless what measure is being used, but two events that are independent with respect to one probability measure may not be independent under another.

But here is a more vivid illustration of the difference. Say you had two disjoint events A and B, and $P(A) > 0$. If you compute the conditional probability of one given the other, you get

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(\varnothing)}{P(B)} = \frac{0}{P(B)} = 0 \neq P(A).$$

So if A and B are disjoint, and I know B occurs, then this *guarantees* that A does not. Far from teaching us nothing about A, knowing B teaches us everything about it. In this way we see that disjointedness is in fact a rather severe form of *dependence*.

Remark 2.5. Can we visualize independent events with a Venn diagram? No. And the reason again is that independence is not fundamentally a property of the sets themselves, but a property of the measure P that is acting on the sets. Visualizing P is an inherently difficult task, and so independence cannot be easily drawn with our familiar blobs.

2.7 Summary of probability fundamentals

We are studying probability spaces:

- (**sample space**) a nonempty set S containing all possible outcomes of a random phenomenon;
- (**events**) subsets $A \subseteq S$;
- (**probability measure**) a function $A \mapsto P(A)$ that assigns probabilities to events.

These are the axioms we impose by fiat on the probability measure:

- (**total measure 1**) $P(S) = 1$;
- (**nonnegativity**) $P(A) \geq 0$ for any $A \subseteq S$;
- (**countable additivity**) if (A_i) is a sequence of pairwise disjoint events, then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

We can deduce from these axioms several basic rules:

- (**complement**) $P(A) = 1 - P(A^c)$;
- (**monotonicity**) if $A \subseteq B$, then $P(A) \leq P(B)$;
- (**inclusion/exclusion**) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$;
- (**total probability**) if (B_i) partitions S, then

$$P(A) = \sum_{i=1}^{\infty} P(A \cap B_i) = \sum_{i=1}^{\infty} P(A \mid B_i)P(B_i);$$

- (**counting**) if S is finite and all outcomes are equally likely, then $P(A) = \#(A)/\#(B)$;
- $P(\varnothing) = 0$;
- $0 \leq P(A) \leq 1$.

These are properties that probability shares with other phenomena like the measures of length, area, and volume in Euclidean space. Conditioning and independence are the special sauce that set probability apart as its own unique branch of mathematics:

- (**conditioning**) $P(A \mid B) = P(A \cap B)/P(B)$, assuming $P(B) > 0$;
- (**marginal/conditional decomposition**) $P(A \cap B) = P(A \mid B)P(B) = P(B \mid A)P(A)$;
- (**Bayes' theorem**) $P(A \mid B) = P(B \mid A)P(A)/P(B)$;
- (**independent events**) $P(A \cap B) = P(A)P(B)$, implying that $P(A \mid B) = P(A)$.

Master this page, and you'll be one step closer to Easy Street. See you there!