The Rules for Reasoning with Probabilities

Contents

Critical Thinker Academy

Introduction

In the tutorial course on different interpretations of the probability concept I often referred to the “probability calculus”, or the “mathematical theory of probability”. This is what most people are talking about when they refer to “probability theory”.

Probability theory provides basic rules for reasoning with probabilities. Given probabilities for events A and B, we can calculate the probability of “A and B”, “A or B”, “A given B”, and so on. These rules are part of our modern understanding of how we ought to reason about uncertainty. In this respect they function in a way similar to the rules of deductive propositional logic, which tell us how we ought to reason with truth-preserving inferences that involve conjunctions, disjunctions, conditionals, etc.

Critical thinking literacy requires that we be at least somewhat familiar with probabilistic relationships. Fallacies of probabilistic reasoning can only be understood in relation to these basic rules. For example, we should all know that the probability of two events both occurring is always smaller than the probability of either event occurring separately — to think otherwise is to commit the "conjunction fallacy".

Working through simple problems is a fun way to learn the basic concepts, but critical thinking doesn’t require that we become probability theory masters. **What's important is to begin to develop some intuitions about how probabilities combine, and some appreciation (and respect) for the power of statistical and probabilistic methods to guide us in our reasoning about risk and uncertainty**.

Part 1: Preliminary Concepts

1. What Has a Probability? Propositions vs Events

In this lecture I want to talk about what sorts of things can have probabilities. I know we did a whole course on philosophical interpretations of the probability concept, but in applications you generally see one of two different languages used to talk about probabilities, what I'll call a **"proposition language"** and an "**event language"**. These really are just different ways of saying the same thing, but it helps to know how to translate back and forth between them.

Let's say I toss a coin, and I ask: **What is the probability THAT the coin will land heads?**

Grammatically, what I'm attributing the probability to is a statement, "The coin will land heads." And the question we're asking is, what is the probability that this *proposition*, this *statement*, is *true*?

So, when we say P(H) = 0.5, that the probability of heads is 0.5, and we're using the language of propositions, we're interpreting H as the proposition that the coin will land heads, and we read the answer as saying that i**t's 50% likely that this proposition is true**.

Now, we can also ask the question this way: **What is the probability OF the coin landing heads?**

The grammar is subtly different. Here, the probability is being associated with an **event**, the event of the coin landing heads.

What's the difference? A proposition is a linguistic entity that asserts a claim that can be either true or false. An event is not a linguistic entity, and events don't assert anything; they aren't the sort of thing that can be true or false. An event is a state of affairs in the world that either happens or doesn't happen.

So in the event language, we interpret “the probability of heads is 0.5” as asserting that the probability of an event occurring, the event of the coin landing heads, is 0.5.

For the most part you can think of these as just two different ways of saying the same thing.

So why do we have these two different languages?

i. When the proposition language is more natural

Well, the proposition language is most natural when we’re talking about **what beliefs can be inferred on the basis of what evidence**, or **how likely it is that the conclusion of an argument is true, given the premises**. This is the domain of **“inductive logic”**, so you’re more likely to see this language in a logic text.

Also, the proposition language might be more natural under certain interpretations of the probability concept than others. For example, statements about **subjective probability**, where probabilities are associated with degrees of belief, are more natural in the proposition language than in the event language.

ii. When the event language is more natural

On the other hand, the event language is more commonly encountered in **statistics and probability theory textbooks**, and it’s often the more natural language when talking about **statistical analysis of data**.

It’s also more natural when we’re using, say, the **frequency interpretation of probability**, since frequencies are usually defined in terms of ratios of events.

The main point here is just you’ll encounter both ways of talking about probabilities, and in most cases you can convert back and forth between them. The mathematical rules are applied the same either way, it’s just a matter of how you interpret the language.

So the value of knowing both languages is similar to the value of knowing both languages in any bilingual community. You want to be able to understand both languages so you can understand the conversations that people are having.

What you want to avoid is thinking that there's only one right way to talk about probabilities, because a) it's just false, and b) it'll get in the way of a productive dialogue between people in each camp.

2. Probabilities Range Between 0 and 1

In this lecture we're going to talk about the range of mathematical values that a probability can take, and how we interpret probabilities at the extremes and between the extremes. We also talk about how the language of "necessity" and "contingency" relates to all this.

We commonly talk about probabilities in terms of percentages, but it's easy to forget that a percentage is just a fraction. 50% probability is 50/100, or 1/2. 25% is 25/100, or 1/4.

The first convention we adopt when talking about probabilities is that **probabilities are real numbers that can take on any value between 0 and 1**.

P(A) = 1

Now, the extremes are interesting. What do we mean when we say that the probability of an event or proposition is 0, or 1?

Well, if the probability is 1, in the event language that means **the event is certain to occur**, **there is no chance that it won't occur**. In the proposition language, it means that **the proposition must be true, it cannot be false**.

Now, this is just a convention. It doesn't say anything about when we can judge an event or proposition to have probability 1. But there are some obvious examples.

If we grant that that the only two options in a coin toss are to land heads or to land tails, then it's safe to say that the probability of the coin landing heads OR tails is 1, since these exhaust all the possibilities.

And this is the common way that this concept is used in probability theory and statistics. **Given a set of mutually exclusive and exhaustive possibilities, the probability that one of these possibilities will be realized is equal to 1.**

P(A) = 0

Now, a probability equal to 0 means just the opposite. This refers to **an event that can't possibly occur**, or **a proposition that can't possibly be true**.

In mathematics and logic there's a convention that contradictory statements can't be true, so **contradictions are automatically assigned a probability value of 0**. If a coin landing heads means that it didn't land tails, and vice versa, then it's a contradiction to say that the coin landed both heads and tails, at the same time — that's a contradiction, and can't possibly be true, so we assign it probability 0.

Necessity and Contingency

In philosophy we have a pair of concepts that we often use to distinguish events or propositions at the extremes from those that aren't at the extremes.

If a proposition or event has probability 1, that means it must be true, or it must occur, and we say that the proposition is NECESSARILY TRUE, or that it's a NECESSARY EVENT.

Similarly, if it has probability 0, that means it must be false, it can't possibly occur, and we say that it refers to a NECESSARILY FALSE proposition, or that the event is an IMPOSSIBLE EVENT.

So, if a proposition or an event has probability that is not 0 or 1, but lies between 0 and 1, that means that it's possible that the proposition is true, or that the event will occur. When this is the case, we commonly say that the proposition is a CONTINGENT proposition, rather than a necessary proposition. Similarly, we'd say that it's a contingent event, rather than a necessary event. In this context, contingent just means "possible" — it's possible that the event will occur. If it occurs, it occurs, but we understand that it could have happened otherwise, and that's what we mean when we say it's contingent rather than necessary.

Now, having just said this, I think it's important to point out that these concepts of necessary and contingent propositions and events are not part of formal probability theory, they’re really a part of a broader philosophical framework for interpreting the world, and they’re separable from probability theory.

A philosopher might argue, for example, that there is no principled way of distinguishing necessary propositions from contingent propositions, or might dispute the existence of necessary truths in a given domain. But these philosophical debates are largely independent of the conventions we use in probability theory and statistics.

I just wanted to make this point clear. The distinction between necessarily and contingently true propositions or events is an important one in philosophy, and probability theory can help us articulate what we mean by these concepts. But within probability theory, probability 1 or probability 0 has an independent formal meaning and formal justification, it’s part of the definition of what a probability is as a mathematical concept, and this definition is independent of any philosophical uses we might want to make of these concepts.

3. Mutually Exclusive Events

This is a short one. I used the expression **“mutually exclusive”** in the previous lecture. Here I want to clarify what this means.

Simply put, two propositions, or two events, are called **“mutually exclusive” if they can’t both be true, or occur, at the same time**.

So, a coin can’t land *both* heads *and* tails on the *same* toss.

A spin of a roulette wheel can’t land on *both* red *and* black on the *same* spin.

These are mutually exclusive events. Mutually exclusive events are sometimes also called “disjoint” events.

Obviously, then, if A and B can both be true at the same time, or occur at the same time, then they’re NOT mutually exclusive.

If A is “I draw a heart from a deck of cards”, and B is “I draw a King from a deck of cards”, those aren’t mutually exclusive — if you draw the King of Hearts then both A and B are true at the same time.

So, that’s it. I just wanted to make sure the concept was clear, because it’s important for understanding some of the rules for reasoning with probabilities.

4. Independent Events

Independence is a very important concept in probability theory.

The basic idea is straightforward. **Two events are said to be independent if the occurrence of one doesn't influence the probability of the occurrence of the other.**

- In other words, given two events A and B, if A occurring or not occurring has **no effect** on the probability of B occurring, then we say that A and B are probabilistically **independent**.
- On the other hand, if the occurrence or non-occurrence of A d**oes have an effect** on the probability of B, then we say that A and B are probabilistically **dependent**.

We can say the same thing in proposition language: Two propositions are independent if the truth of one doesn't make the truth of the other any more or less likely.

A schematic way of representing independence is like this. If A and B are probabilistically INDEPENDENT, then the following relations hold:

- P(A) = P(A given B)
- P(B) = P(B given A)

If A and B are independent, then the probability of A given B, is just the same as the probability of A all by itself.

And if they're independent, then it works the other way too. The probability of B given A is the same as the probability of B all by itself.

If these are NOT equal, then A and B are probabilistically DEPENDENT on one another. We're saying that the occurrence of B changes the probability of A, and vice versa:

- P(A) ≠ P(A given B)
- P(B) ≠ P(B given A)

Some examples:

Toss a coin. Assume this is a fair coin, so the probability of it landing heads is 50 percent, or 0.5.

Let's assume that it in fact landed heads. Given this, what is the probability that it will land heads again on the second toss?

A = the coin lands heads on toss 1

B = the coin lands heads on toss 2

In other words, what is P(B given A)?

Well, if you think it's higher or lower than 0.5, then you'd be making a mistake. These are probabilistically independent events:

P(B given A) = P(A) = 0.5

The probability of landing heads on a second toss is still 0.5. This would be the case even if you'd previously landed ten heads in a row — the probability of the next toss landing heads is still just 0.5. To believe otherwise is to commit what's known as the **"gambler's fallacy"** (we'll talk more about the gambler's fallacy in the course in fallacies of probabilistic reasoning).

Now, consider this example:

A = the dice roll is an **even number**

B = the dice roll is a **2**

Are these two events independent or dependent? If A is true, does it affect the probability of B being true, and vice versa?

We need to compare P(B) and P(B given A).

We know that the probability of rolling a 2 on a six-sided dice is 1 in 6. So the probability of event B by itself is 1/6.

But if A is true, then the set of possible values is restricted to just the even numbers. The probability of rolling a 2, given that it's even, is 1/3.

Thus, $P(B) \neq P(B|A)$. A and B are probabilistically **dependent** events. The occurrence of one affects the probability of the other.

Dependence is Symmetric, But Numerical Values May Differ

I want to note that dependence is a symmetrical relationship, in that it works both ways: **if the occurrence of A affects the probability of B, then the occurrence of B will also affect the probability of A**.

But I also want to point out that this **doesn't** mean that the numerical values will be the same. In our example, if we know the dice roll is even, then the probability of rolling a 2 is 1/3, rather than 1/6.

But lets do it the other way around. We roll the dice but we don't know anything about the outcome. What are the odds that the dice roll is even? Well, that's just 1/2, since the even numbers are 2, 4 and 6, and that's half of the possible outcomes.

But now let's say we know that we rolled a 2. This obviously affects the probability that the dice roll is even. In fact, it's certain that it's even — given that it's a 2, the probability that it's even is 1.

And this illustrates the point.

P(even) = 1/2, but P(even, given it's a 2) = 1

P(2) = 1/6, but P(2, given that it's even) = 1/3

They're probabilistically dependent whichever way you go, but the numerical values may differ depending on which way you go.

Part 2: The Basic Rules

1. The Negation Rule: P(not-A)

Okay, the first of the basic rules we're going to cover is the **negation rule**.

We're given an event A, and we know the probability of A occurring; call this P(A).

Question: What is the probability that A will NOT occur?

The rule is simple: Given P(A), the probability that A will occur, or that A is true, the probability of A NOT occurring, or NOT being true, is equal to 1 minus the probability of A.

P(not-A) = 1 - P(A)

Here's a simple example:

We know that the probability of rolling a 2 on a six-sided die is just 1 in 6. Our space of possible outcomes has six elements in it, and the outcome that we care about is just one of those elements, so the ratio is 1 in 6. Thus, P(2) = 1/6.

Now, what's the probability of NOT rolling a 2? Well, this is just the probability of rolling a 1 or a 3 or a 4 or a 5 or a 6 — anything that's not a 2. This is 5 of the 6 possible outcomes, so the probability is 5 in 6.

But note that 5/6 is just 1 − 1/6, or 1 - P(2), the probability of rolling a 2. If that's not clear just remember that 1 is equal to 6/6, so 6/6 - 1/6 equals 5/6. Thus,

P(not-A) = 1 - P(A)

P(not-2) = 1 - P(2)

P(not-2) = 1 − 1/6

P(not-2) = 5/6

$$P(\text{not-A}) = 1 - P(A)$$

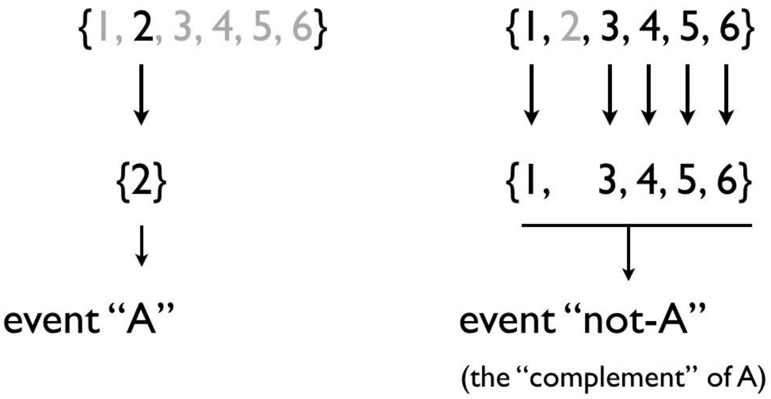
Example: dice rolls

$P(2) = 1/6$ $P(\text{not-2}) = 5/6 = 1 - P(2)$

$\{1, 2, 3, 4, 5, 6\}$ $\{1, 2, 3, 4, 5, 6\}$

$= 6/6 - 1/6$
$= 5/6$

Now, when you look at the sets of numbers in the curly brackets here you notice something interesting. These sets represent what is sometimes called the "sample space" of this experimental setup, the set of all possible elementary outcomes of a random trial. Notice that the events A and B can be represented by subsets of this sample space. Let's make this clearer.



{1, 2, 3, 4, 5, 6}

{2}

event "A"

{1, 2, 3, 4, 5, 6}

{1, 3, 4, 5, 6}

event "not-A"

(the "complement" of A)

The event of rolling a 2 is represented by this single element of the set: {2}

The event of NOT rolling a two is also represented by a subset of the sample space, in this case the remaining five elements: {1, 3, 4, 5, 6}

In set theory language, we'd call this set — {1, 3, 4, 5, 6} — the "complement" of the set of the singleton set {2}.

The general point is that given a space of possible outcomes, you can represent an event A as a subset of that space, and the negation of A, "not-A" is represented by the complement of that subset — all the members of the space that are NOT in A.

And when you conjoin these sets together, you recover the whole sample space, because A and not-A partition the sample space into two parts with nothing left over, so when you put the parts back together, you get the whole space back.

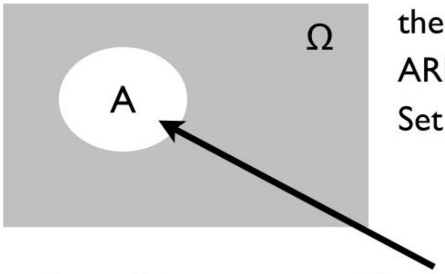
We can generalize this point graphically. First let's set up a convention that is often used to graphically represent probability relations. I find it really helps to develop your intuition about these relationships. This only works for certain kinds of sample spaces, but it's still very helpful.

Let's represent the total sample space, the complete set of possible elementary outcomes, by a square. Call it "omega". We'll set the area of this sample space equal to 1.

the sample space, Ω
AREA is proportional to PROBABILITY
Set the area of Ω = 1

Now, when we do this, different events, different possible outcomes, can be represented by subsets of this area. And the probability of those events will be proportional to the area of the subsets.

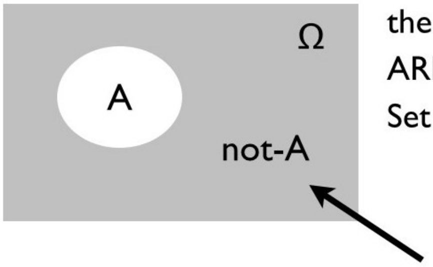
So, let event A be represented by a given area. The probability of A is proportional to the area of this subset — the larger the area, the more probable the event. If the area of A includes the whole sample space, omega, then the probability would be 1 — that event would happen with certainty. In general this will be a number less than 1.



the sample space, Ω
AREA is proportional to PROBABILITY
Set the area of Ω = 1

- Event A is represented by this subset
- the probability of A, P(A), is represented by the AREA of this subset (which will be a number less than 1)

So, how do we represent not-A on this diagram? Well, not-A is just the complement of A, it's the fraction of omega that is NOT in A.



the sample space, Ω
AREA is proportional to PROBABILITY
Set the area of Ω = 1

- Event not-A is represented by this subset, the **complement** of A
- the probability of not-A, **P(not-A)**, is represented by the AREA of this subset
- this area is equal to **P(not-A) = 1 - P(A)**

Now, from this we can see a couple of useful relations.

First, as we've seen, when we conjoin or take the union of A and not-A, we recover the whole sample space, which has probability 1. This simply involves adding up the areas, and they fit together like a the pieces of a jigsaw puzzle.

This schema shows us how to think about the probability of any event, or proposition, and it's negation, even ones that are hard to assign a numerical measure to.

For example, if A is the proposition that it's going to rain tomorrow, and we think this has a 75% chance of being true, then according to this rule we're compelled to assign a probability of 25% to the proposition that it won't rain tomorrow. This is a consistency constraint on how we're supposed to reason with probabilities as defined in standard probability theory. **If we don't do this then we're not following the rules of standard probability theory**. If you break this rule where standard probability is known to apply, like when you're at the roulette wheel or playing cards, then you'll just be in error, you'll misjudge the probabilities of complementary events.

The second thing I wanted to point out is that this rule builds in a rule of standard formal logic, which is known as the **law of the excluded middle**. For every proposition A, either A is true, or its negation is true, there's no third truth-value that A could have.

It's an interesting fact that there are logical systems, and mathematical systems, that reject this law. There's no room to go into those here, but I just wanted to point out that there are interesting relationships between probability theory and logic. But we're doing standard probability theory here, and it won't pay to get too distracted, so for the most part I'll be ignoring these possible digressions.

2. Restricted Disjunction Rule: P(A or B) = P(A) + P(B)

In this lecture we'll look at the **disjunction rule**, the rule for calculating the probability of a disjunction of events, or, in more familiar terms, given probabilities for events A and B, **what is the probability that EITHER A OR B will occur**? Statements of the form "A is true OR B is true" are known as "disjunctions" in math and logic, so that's where the rule gets its name.

There's a more general formulation for this rule, and there's a more restricted special case. **In this lecture we'll just deal with the special case, which occurs when the two events in question are mutually exclusive, meaning that they cannot both occur at the same time.**

Example 1

Let's consider dice rolls this time. The probability of any particular number, say, a 2, coming up on a single dice roll is 1/6, right?

So what would be the probability of rolling EITHER a 2 OR a 6?

P(2 or 6) = ?

Well, getting a 2 or a 6 is more likely than getting just one or the other by itself, so we know the probability is going to be higher. In this case we can actually count the elementary outcomes to get the answer.

There are six possible elementary outcomes, and the event in question picks out two of these outcomes. So the probability of getting a 2 or a 6 is just this ratio, which is equal to 2/6, or 1/3.

The algebraic translation of this reasoning is straightforward. What you're doing is ADDING the probabilities of the individual outcomes.

P(2 or 6)

= P(2) + P(6)

= (1/6) + (1/6)

= 2/6

= 1/3

The Rule

And this is our rule:

P(A or B) = P(A) + P(B)

if events A and B are “mutually exclusive”.

The probability of A or B is just the probability of A, plus the probability of B.

Now, note the disclaimer: the rule only works in this simple form if A and B are mutually exclusive, meaning that they can’t both occur at the same time — either A occurs, or B occurs, or neither occur, but they can’t both occur at the same time.

This is the case with our example. You can’t roll both a 2 and a 6 at the same time on a single dice roll, these are mutually exclusive outcomes. Similarly, you can’t toss both a head and a tail on a single coin toss.

Example 2

Let’s look at an another example.

What is the probability of drawing either a face card, or a 10, from a well shuffled deck of playing cards?

First we’ll need to make sure we’ve identified all the cards.

The face cards include Jacks, Queens and Kings. The ten is just a 10, so this is a total of four types of cards.

So we’d write this as the probability of getting a face card or a 10, which is just the probability of getting a jack or queen or a king or a 10.

Now, if these events are mutually exclusive then we can apply the restricted disjunction rule.

Are they? Sure they are. If you draw any one of these cards you can’t simultaneously draw any of the others, so they’re all mutually exclusive.

So, our expression would look like this:

P(face card OR 10)

= P(face card) + P(10)

= P(J or Q or K) + P(10)

= P(J) + P(Q) + P(K) + P(10)

The probability of drawing a jack or a queen or a king or a 10 is equal to the probability of drawing a jack, plus the probability of drawing a queen, plus the probability of drawing a king, plus the probability of drawing a 10.

Now, what's the probability of drawing a jack? Well, there are four jacks in a deck of 52 cards, one for each suit. So the probability of drawing a jack is 4/52. And this will be the same for all the cards.

This is our answer, the rest is just algebra:

P(face card OR 10)

= P(J) + P(Q) + P(K) + P(10

= (4/52) + (4/52) + (4/52) + (4/52)

= (1/13) + (1/13) + 1/13) + (1/13)

= 4/13

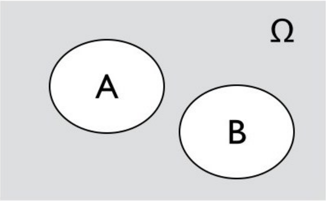
Note however that you can simplify the algebra by recognizing that 4/52 is equal to 1/13. So the answer is 4/13, which is roughly 0.31, or 31 percent.

The Sample Space View

Before we leave we should look at this expression from the sample space perspective to get some additional insight into what it means.

If you recall from the tutorial on negations, we showed that you can represent the total sample space by a unit area, call it "omega". Subsets on this sample space represent events, and the probability of the event is just the area of the subspace as a fraction of the total sample space. The total sample space has area equal to 1.

The restricted disjunction rule looks like this, graphically:



Restricted Disjunction Rule

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

**If A and B are MUTUALLY EXCLUSIVE
it means that the areas of A and B
DO NOT OVERLAP**

Events A and B are represented by areas on the sample space. The larger the area, the larger the probability associated with the event. The probability of A or B occurring is represented by the sum of these areas — you just add up the areas of A and B, and this sum will obviously be a larger fraction of the total sample space.

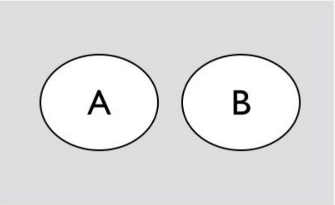
What it means to say that A and B are mutually exclusive is that these areas don't overlap, there are no regions of intersection. So if A occurs, B doesn't occur, and vice versa.

As we'll see in the next lecture, if A and B are not mutually exclusive, this means that their areas DO overlap, and this would correspond to cases where A and B both occur at the same time. As a result it's a bit trickier to calculate the sum of the areas, it's not a simple algebraic sum of the two areas taken separately.

3. General Disjunction Rule: P(A or B) = P(A) + P(B) - P(A and B)

In the last lecture we looked at the disjunction rule applied to the special case where the events in question are mutually exclusive. Now lets look at **the general rule, which also applies to cases where the events are NOT mutually exclusive**.

Here's our Venn diagram depiction of mutually exclusive events.



What is the probability that the card is a King or a Jack?

P(A or B) = P(A) + P(B)

The grey square is the whole sample space, the areas of A and B represent the two events in question, and the size of these areas is proportional to the probability of each event.

In this example of drawing either a King or a Jack from a deck of cards, these are mutually exclusive events, so we represent these as having no overlap in the sample space, and the probability of drawing either a King or a Jack is just the algebraic some of their individual probabilities.

Now let's consider a different case. What is the probability of drawing a card that is either a King or a Spade?

P(K or S) = ?

Let's start with the first part: what is the probability of drawing a King?

Well, there are four kings in a deck of cards, one for each suit, so that's just 4 out of 52. Thus,

P(K) = 4/52

Now what's the probability of drawing a spade?

Well, there are four suits, so one in every four cards is a spade. So the probability is 1 in 4, but we'll write this as 13 out of 52 so that our sums have a common denominator.

P(S) = 13/52

Now if we were to just add up these probabilities to get the probability of drawing a King or a Spade, the answer would be this:

P(K or S)

= P(K) + P(S)

= (4/52) + (13/52)

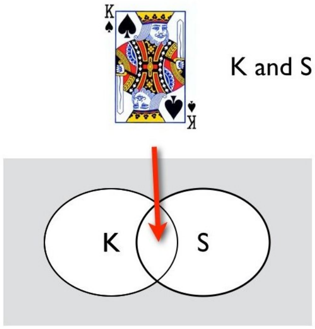
= 17/52

But there's a problem with this answer. The problem is that in calculating the probabilities for drawing a king and a spade, **we've double-counted one card — namely, the KING OF SPADES.**

The King of Spades is BOTH a KING AND a SPADE, so this card is included in the calculation of both probabilities — he's included in the probability of drawing a King, and he's included in the probability of drawing a spade.

But there's only ONE King of Spades in the deck, so by counting him twice **we're overestimating the probability of drawing either a King or a Spade**, and that's an error.

Graphically our situation looks like this:



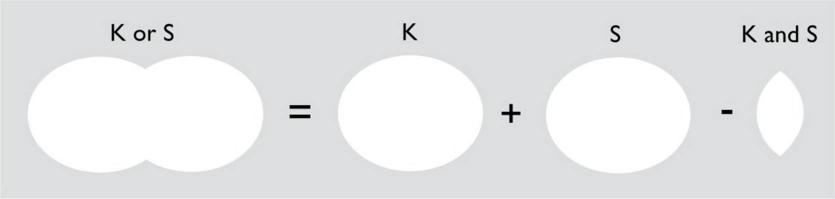
In this case our events are NOT mutually exclusive, there are cases where they overlap, and in this particular case, the overlap represents the event of drawing the King of Spades. (By the way, I know that these areas shouldn't be the same size, but for this introduction here it'll be helpful to keep them the same size.)

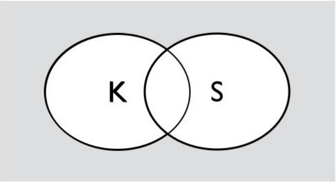
In a Venn diagram you define the overlap region as "K and S", the cards that are both Kings and Spades.

And you can see how if we're just adding up the areas of K and S then we'd be counting the overlap region twice. What we're going for is the area of the white space, that peanut-shaped area defined by the external boundaries of K and S.

To get THAT area, all you need to do is SUBTRACT the area of the overlap region from the sum of the two separate areas.

This picture let's you visualize what's going on.





What is the probability that the card is a King or a Spade?

The probability of drawing a King OR a Spade is represented by that peanut shaped area on the left, and you get it by subtracting the overlap region from the sum of the two areas.

This gives us the algebraic expression that we need to fix the error caused by double-counting the overlap region.

P(K or S) = P(K) + P(S) - P(K and S)

You just add up the probabilities of the two events taken separately, and then subtract the probability associated with the conjunction of the two events. In this case there's only one card that is both a king and a spade, the probability of drawing that card is just 1 in 52.

P(K or S)

= (4/52) + (13/52) - (1/52)

= 16/52

So we subtract 1 in 52 from our sum, and we get the correct answer, which is 16/52, rather than 17/52.

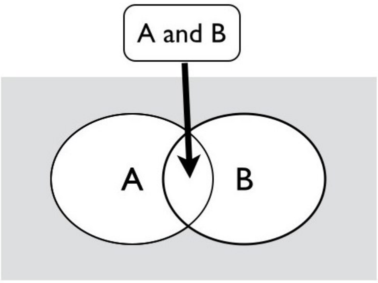
General Rule

Here's the general disjunction rule in terms of arbitrary events A and B:

P(A or B) = P(A) + P(B) - P(A and B)

Notice that this general rule includes the restricted rule as a special case, since if A and B are mutually exclusive, then A and B don't overlap and the conjunction term on the right goes to zero, and we recover the restricted rule.

P(A or B) = P(A) + P(B) - P(A and B)



The General Disjunction Rule

Another Example

Let's look at another example. What is the probability of drawing either a Face Card (F) or a Spade (S)? A face card, remember, is a Jack, Queen or King, any card with a face on it.

Here's our rule:

P(F or S) = P(F) + P(S) - P(F and S)

It's equal to the probability of drawing a face card, plus the probability of drawing a spade, minus the probability of drawing a card that is BOTH a face card and a spade.

With examples like these you can just count the cards to get the probabilities.

Let's start with P(F). There are 3 face cards per suit, and four suits, so that gives us 12 out of 52 cards that are face cards. So P(F) = 12/52.

P(S) is easy, there are 13 cards in a suit, so P(S) = 13/52.

Now, how many card are there that are *both* face cards *and* spades? Well, just the Jack, Queen and King of spades, so that's 3 out of 52. So P(F and S) = 3/52.

And the rest is simple algebra:

P(F or S) = P(F) + P(S) - P(F and S)

= (12/52) + (13/52) - (3/52)

= 22/52

and our answer is 22/52, or 11/26, which is roughly 0.26, or 26 percent.

This is how you use the general rule for calculating probabilities of disjunctions. You just have to remember to check whether the events are mutually exclusive, and if not, you need to subtract the probability of the conjunction of the two events, the cases where both events occur, or where both of the corresponding propositions are true.

4. Restricted Conjunction Rule: P(A and B) = P(A) x P(B)

In this lecture we'll look at the **conjunction rule**, the rule for calculating the probability of a conjunction of events. But we'll deal with a **special case**, where the events in question are **independent**, and the rule takes on a very simple form.

Coin Tosses

Let's consider coin tosses again. The probability of a single coin landing heads is 1/2, right?

Now what if we toss two coins at the same time? What is the probability that both coins will land heads?

Most people will see the answer right away, because we're familiar with these sorts of cases. We know that the probability of **both** landing heads is going to be **less** than the probability of just one landing heads. When we're dealing fractions, we know that **multiplying fractions gives us a smaller number.**

In this case, if we **multiply the probabilities of the two independent events, we get the right answer**: 1/2 times 1/2 equal to 1/4. There's a 25% chance of both coins landing heads.

If we go for three coins, it's the same idea. We just multiply the probabilities of each individual event, and the answer is half as small again, 1/8.

The Rule

This is the **restricted conjunction rule**:

P(A and B) = P(A) x P(B)

If A and B are independent events, the probability of the conjunction of two events, which is just the probability of the two events both occurring, or of the corresponding propositions both being true, is just the product of the probabilities taken separately.

If the events are not independent, then we need to use a modified version of this rule, but we'll take that up in the next lecture.

Dice

Let's just look at one more example.

Using a six-sided dice, what is the probability of rolling **three sixes in a row**? That is, what is

P(6 and 6 and 6) = ?

Dice rolls, like coin tosses, are independent events, so we can use the restricted conjunction rule. The probability of rolling a six is just one in six. So the probability of rolling three sixes in a row is just 1/6 x 1/6 x 1/6, which is one in 216:

P(6 and 6 and 6)

= P(6) x P(6) x P(6)

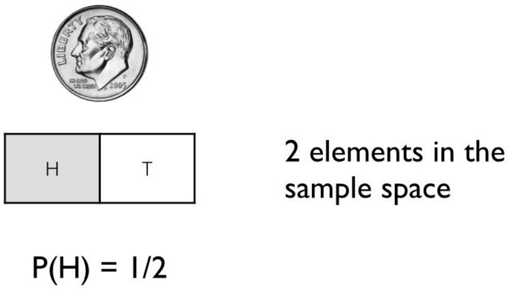
= (1/6) x (1/6) x (1/6)

= 1/216

Sample Space View

This is a pretty simple rule, but before we leave I want you to think about what this rule means from the **sample space perspective**. In this framework, events are associated with subsets of the sample space, and probabilities are associated with the ratios of the corresponding subsets to the total space of possible outcomes.

Here's the sample space for single coin toss:



It's just a listing of the set of possible elementary outcomes, and in this case there are just two, heads and tails. The probability of the coin landing heads is equal to the ratio of outcomes where it lands heads, divided by the total number of elements in the sample space, which in this case is obviously just 1 out of 2 or 1/2.

Here's the sample space for the event where two coins are tossed:

H, H	H, T
T, H	T, T

4 elements in the sample space

P(H and H) = 1/4

You've got twice as many possible combinations of outcomes — heads heads, heads tails, tails heads, and tails tails. This sample space has four elementary outcomes, and the probability of two heads is just the ratio of the number of elements where the outcome is two heads, divided by the total number of elementary outcomes, which is 4. So the probability for this event is just 1 in 4.

Here's the sample space for three coin tosses:

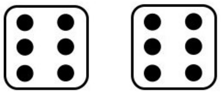
H, H, H	H, H, T	H, T, H	T, H, H
H, T, T	T, H, T	T, T, H	T, T, T

8 elements in the sample space

P(H and H and H) = 1/8

There are eight possible combinations of heads and tails, and the probability of landing three heads is just 1/8.

Here's the sample space for rolling two dice:



(1)(1)	(1)(2)	(1)(3)	(1)(4)	(1)(5)	(1)(6)
(2)(1)	(2)(2)	(2)(3)	(2)(4)	(2)(5)	(2)(6)
(3)(1)	(3)(2)	(3)(3)	(3)(4)	(3)(5)	(3)(6)
(4)(1)	(4)(2)	(4)(3)	(4)(4)	(4)(5)	(4)(6)
(5)(1)	(5)(2)	(5)(3)	(5)(4)	(5)(5)	(5)(6)
(6)(1)	(6)(2)	(6)(3)	(6)(4)	(6)(5)	(6)(6)

36 elements in the sample space

P(6 and 6)

= P(6) x P(6)
= 1/6 x 1/6
= 1/36

There are 36 possible outcomes. The probability of rolling two sixes is 1/6 times 1/6, or 1 in 36.

It's rare that you'll have to write out the sample space like this to solve a problem, but it's a helpful reminder of what the mathematical rule means, and why it gives us the right answers.

5. General Conjunction Rule: P(A and B) = P(A) x P(B|A)

In the last lecture we looked at the **restricted conjunction rule**, which is the rule for calculating the probability of a conjunction of events, when those events are **independent** of one another. What independence means in this context is that if one of the events occurs, this has no effect on the probability of the other event occurring. In that case you simply multiply the probabilities for each event.

Now we need to look at the more **general case**, where the events are **not independent**.

An Example Where the Restricted Rule Fails

Let's consider dice rolls once again. Let's call E the event that the dice roll is an **even number**.

And let's call P the event that the dice roll is a **prime number**.

Here are the event spaces that correspond to these dice rolls:

E = {2, 4, 6}

P = {1, 2, 3, 5}

We're interested in the probability that the dice roll is BOTH even AND prime:

P(E and P) = ?

Now, if we were to use the restricted conjunction rule, we'd just multiply these probabilities together. Let's see how that would work.

The probability of a dice roll being even is just 3/6, or 1/2, since we've only got six possibilities and three of those are even. So P(E) = 3/6.

The probability of a dice roll being prime is 4 in 6, since the primes make up 4 of the six possible rolls. So P(P) = 4/6.

Now if we use the **restricted conjunction rule**, the calculation looks like this:

P(E and P)

= P(E) x P(P)

= (3/6) x (4/6)

= 12/36

= 1/3

We just multiply these numbers together, and we get a final answer of 12 out of 36, which is equal to 1/3. So according to this calculation, on any given dice roll there's a one in three chance of getting a roll that is both even and prime.

But we know this answer can't be right.

How do we know this? Because by inspection we know that there's only ONE possible dice roll that is both even and prime — it's the 2.

But if there's only one possible dice roll that is both even and prime, then we know the answer. The answer has to be 1/6. **But the restricted rule gives us 1/3, or 2/6 — it overestimates the probability.**

Why Does The Restricted Rule Fail?

So, this example shows us that the restricted conjunction rule doesn't apply to this case. Why doesn't it apply?

It doesn't apply because E and P are not independent events. We're interested in the probability that E and P are both true of a given dice roll, but if P is true (if we know the dice roll is a prime) then that affects the probability that E is true (that it's even). If it's prime then just look at the options — {1, 2, 3, 5} — there's only one even number in that list of four prime numbers (the 2), so the probability of the roll being even, *given that it's prime*, is 1/4, not 1/2.

And similarly if we know that the dice roll is even, that affects the probability that it's also prime. In this case, if we know it's even, then our options are {2, 4 ,6}, and only one of those is prime (the 2), so the probability of it being prime, *given that it's even*, is 1/3.

So, we know that the restricted conjunction rule doesn't work, and we know it doesn't work because the rule doesn't take into account the probabilistic dependence of the two events on one another. This gives us an idea for how we might modify the conjunction rule to fix this problem.

The General Conjunction Rule

Instead of just multiplying the probabilities of A and B, we can try multiplying the probability of A with the probability not of B, but of B GIVEN A, or P(B|A):

General Conjunction Rule

P(A and B) = P(A) x P(B|A)

An expression like this, P(B|A), is called a **"conditional probability"**. We've got a whole other lecture on conditional probabilities, but for now it's enough to just read it as "the probability of B given A".

Our Example Using the General Rule

Let's try out this new rule with our example.

First of all, let's remember that we know what the answer is supposed to be, just by inspection. There's only one dice roll that is both even and prime, so the probability has to be 1/6. Let's see if our general formula actually gives us this answer.

Here's our formula:

P(E and P) = P(E) x P(P|E)

The probability of a dice roll being even and prime equals the probability of it being even times the probability of it being prime, given that it's even.

And here are the numbers:

P(E and P)

= P(E) x P(P|E)

= 1/2 x 1/3

= 1/6

The probability that a roll is even is 1/2. The probability that a roll is prime, given that it's even, is 1/3, since among the even numbers, 2, 4 and 6, only one of these, the 2, is prime, so the probability is 1/3.

1/2 times 1/3 is 1/6, and lo and behold, we get the right answer!

Now, you might be wondering whether it also works the other way, if we consider the probability of a roll being even, given that it's prime.

And the answer is, yes it does. Let's do it that way. If we switch around the As and Bs in our general rule, the result still holds.

P(E and P) = P(P) x P(E|P)

When you plug in the numbers you get this:

P(E and P)

= P(P) x P(E|P)

= (4/6) x (1/4)

= 4/24

= 1/6

The probability of a roll being a prime number is 4/6, since 1, 2, 3 and 5 are primes.

The probability of a roll being even, given that it's prime, is 1/4, since of those four primes, only one, the 2, is even.

4/6 times 1/4 is 4/24, which is 1/6.

Same result. It works both ways.

Summing Up

So, to sum up, this is the general conjunction rule:

P(A and B) = P(A) x P(B|A)

P(A and B) = P(B) x P(A|B)

It’s written here in two forms, depending on which you choose as the conditional probability term, but they’re equivalent formulations and they’ll give you the same answers.

Note also how this rule reduces to the restricted conjunction rule when A and B are independent. In that case, following our definition of probabilistic independence, **the conditional probabilities just reduce to the unconditional probabilities and you recover the simple restricted rule.**

In the next lecture we’re going to focus our attention on those conditional probabilities.

6. General Conditional Probability Rule

In the last lecture we looked at the general conjunction rule, which involves the use of conditional probabilities. Now if you look at this expression for conditional probability:

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}$$

which I'm calling the **"general conditional probability" rule**, you might notice that it's **exactly the same formula as the general conjunction rule, it's just rearranged**. This is, in fact, how conditional probability is **defined** in standard probability theory. In this lecture I want to explore what this formula means from the **sample space perspective**, with the hope that it'll help to develop some intuitions about why it works.

Let me say up front that I'm not going to be talking about the Bayes' Rule formulation of conditional probability here, I'm going to save that for another lecture.

An Example

Let's start off with a simple example to refresh our memories. Consider these two events: the first event is rolling a 2 on a dice roll. We'll label this event with the number 2. The second event is rolling an even number on a dice roll. We'll label this event "E". Thus,

2 = the dice roll is 2

E = the dice roll is even

We'll call a probability a **"categorical probability"** if we're just talking about the probability of an event that is not conditional on other events, we're just asking about the probability of A occurring, we're not asking about the probability of A given some other event occurs. We call this P(A).

We contrast "categorical" probabilities with **"conditional" probabilities** — here we're asking about the probability of A, *given that* *some other event, B, has occurred*. Or in other words, we're asking about the probability of A, on the condition that B occurs. We call this P(A|B)**.**

So, with these two events, what are the categorical probabilities?

That's easy: the probability of rolling a 2 is 1/6, and the probability of rolling an even number is 3/6, or 1/2.

P(2) = 1/6

P(E) = 1/2

But of course we're interested in conditional probabilities, so a natural question to ask is, what is the probability of rolling a 2 , *given that it's even?*

P(2|E) = ?

We know the answer to this just by inspection: if the dice roll is either a 2, 4 or 6, then the probability that it's a 2 is just one third. Thus,

P(2|E) = 1/3

Now let's see how this answer squares with our definition of conditional probability.

Here's our definition in terms of general events A and B:

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

The probability of A given B is equal to the probability of the conjunction of A AND B, divided by the unconditional probability of B all by itself.

If we substitute our events for this example it looks like this:

$$P(2|E) = \frac{P(2\text{and } E)}{P(E)}$$

The probability of rolling a 2, given that it's even, equals the probability of rolling both a 2 and an even number, divided by the probability of just rolling an even number.

We know the value of the denominator term, it's just one half: P(E) = 1/2. The numerator is the only tricky part. It's a conjunction, and in the last lecture we covered the general conjunction rule. In fact this is just another way of writing the general conjunction rule. The question we want to ask is this: How many possible dice rolls are there, where the dice roll is **both a 2 and even**?

Answer: Just one.

Rolling a 2 is the only dice roll that is both a 2 and even.

And the probability of rolling a 2 is just 1/6, right? So now we have our numbers:

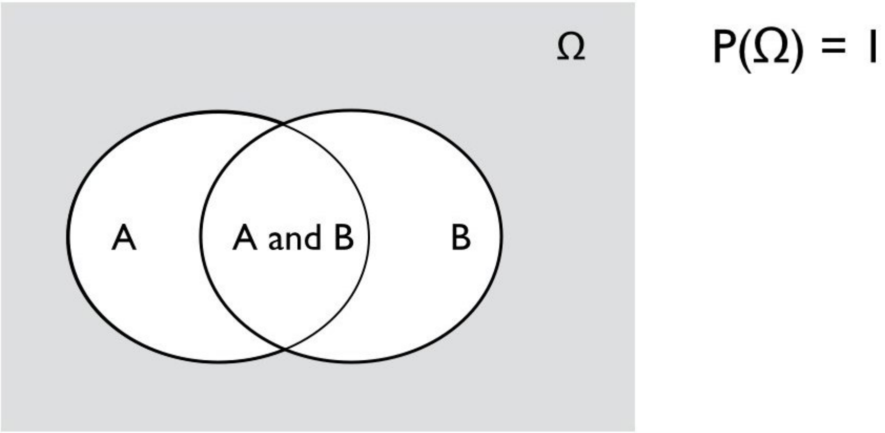
$$P(2 | E) = \frac{\left(\frac{1}{6}\right)}{\left(\frac{1}{2}\right)} = \frac{2}{6} = \frac{1}{3}$$

And this is, indeed, the answer that we figured out just by inspection.

So the formula works. Now, I know for a fact that a lot of students don't have a good intuitive sense of why it works. They're not sure exactly why the conjunction is relevant, and they're not sure why we're dividing by the probability of the conditioning event. To help see why the formula makes sense, it helps to look at the situation from the sample space perspective.

The Sample Space View

Here's our generic sample space that we're familiar with from previous tutorials.

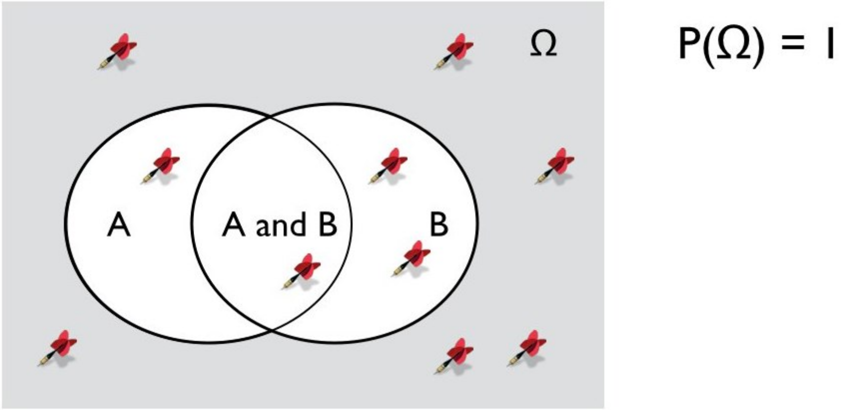


The grey square, labeled "omega", represents the set of all possible outcomes of a probabilistic trial, or what we've been calling the "sample space". Events are represented by subsets of this sample space, and probabilities of events are represented by the area of the subset associated with a given event.

So in this example the ovals labeled A and B are events, and the areas of A and B are proportional to the probability of A and B occurring.

If we consider the area of the whole sample space, omega, then we assign this probability 1, which means that the outcome has to land somewhere inside this area.

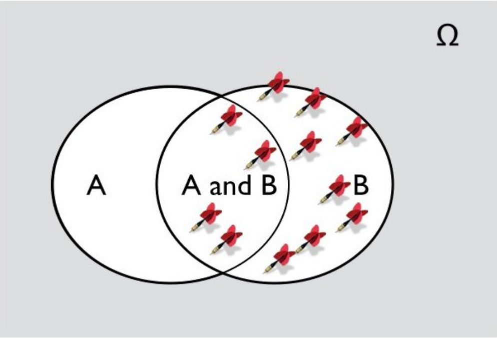
If we think of a probabilistic trial by analogy with throwing a dart at a board, then we're saying that the dart is guaranteed to land somewhere inside the grey square.



Now, in this diagram A and B overlap. This represents events where A and B both occur at the same time, indicating that these aren't mutually exclusive events. In previous lectures we used the example of drawing a playing card that is both a face card and a spade. In this video we used the example of getting a dice roll that is both an even number and a 2. The area of this overlap region represents the probability that both A and B will occur.

Now let’s ask the question, how is conditional probability represented on this diagram? Let’s think about what’s going on when we ask “what is the probability of A, given B?”.

What we’re saying is that in this case, we know some additional information that we didn’t know before. We know that event B occurred. This is like saying that we know that our dart landed somewhere inside B.



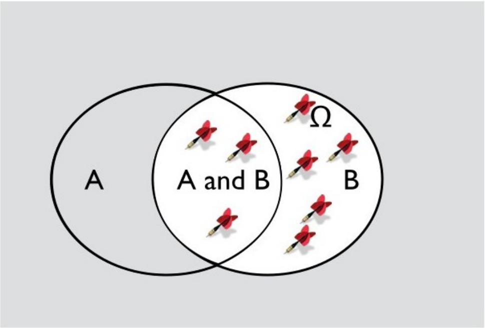
$$P(\Omega) = 1$$

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

So we’re saying, given that we know the dart landed in B, *what are the odds that it also landed in A*? In other words, given that we know the dart landed inside B ,what are the odds that the dart landed *in the overlap region between B and A*?

The overlap region makes up a fraction of B, and that’s precisely the fraction that we’re trying to estimate with the conditional probability rule. We’re asking for the ratio of the overlap region to the area of B.

We can say the same thing with a slightly different emphasis. When we know that B is true, or that B occurred, what we’re saying is that we’re no longer dealing with the whole sample space. We’re dealing with a reduced sample space, and treating this as our new “omega”.



P(A|B)

- this is what "given B" means

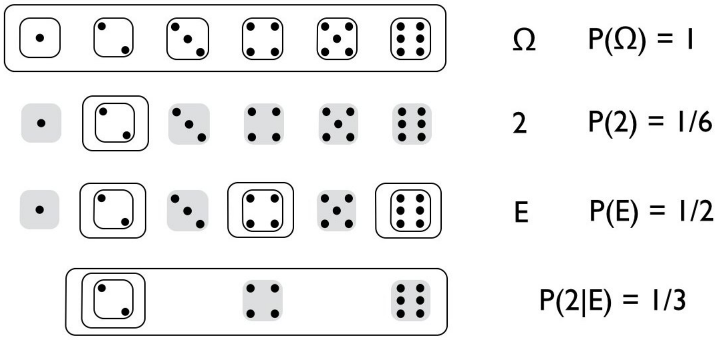
$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

The conditional probability of A given B is the area of the overlap region, the events where A and B both occur, divided not by the area of the original sample space, but by the area of the reduced sample space, B.

Now I think it's much easier to visualize what the numerator and the denominator represent in the general rule for conditional probability.

The Sample Space View and the Dice Problem

This discussion is consistent with what we did when we solved the dice problem.



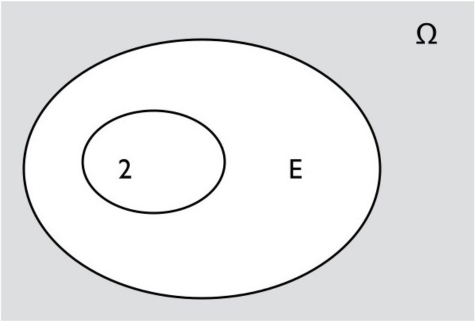
Omega is equal to the set of equally probable outcomes 1 through 6. We imagine assigning an area to this set equal to 1.

The event of rolling a 2 is a subset of this sample space. In this case the 2 takes up exactly one sixth of the total area, so the probability is 1/6.

The event of rolling an even number is a different, larger subset. It takes up exactly one half of the total area, so the probability is 1/2.

Now, if we we're considering the probability of rolling a 2, given that it's even, we're dealing with a reduced sample space. We're treating the evens as the new sample space, and looking at the proportion of events corresponding to the number 2, as a fraction of this new sample space. And we get the answer, 1/3.

Note that in this case the overlap is complete. The 2s are entirely within the evens. This corresponds to a sample space that looks like this:

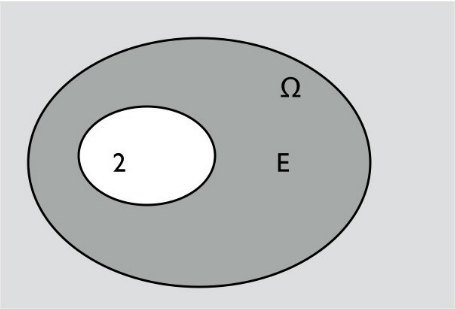


$$P(2|E) = \frac{P(2 \text{ and } E)}{P(E)}$$

The overlap is complete, with the 2 a proper subset of the evens. But the formula works all the same.

If A is included inside B, then the intersection of A and B is just equal to the area of A. In this case, the intersection of 2 and E is just equal to the area of 2. This simplifies the calculation.

The numerator is just the area of 2, and the denominator is just the area of E. Plugging in the probabilities we get the answer, 1/3.



$$P(2|E) = \frac{P(2 \text{ and } E)}{P(E)}$$

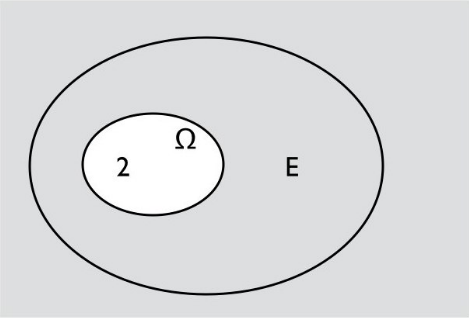
In this case, P(2 and E) = P(2)

$$P(2|E) = \frac{P(2)}{P(E)}$$

$$= \frac{1/6}{1/2} = 1/3$$

Now, for the sake of completeness, let's work it the other way. What's the probability of the dice landing even, given that it's a 2?

We know the answer to this already, it's got to be equal to 1, since 2 is an even number. The calculation gets it right too.



$$P(E|2) = \frac{P(E \text{ and } 2)}{P(2)}$$

once again, P(E and 2) = P(2)

$$P(E|2) = \frac{P(2)}{P(2)}$$

$$= \frac{1/2}{1/2} = 1$$

We use the fact, once again, that the area of overlap is just equal to the area of 2, so P(E and 2) is just equal to P(2). And now the conditional probability is just 1/2 over 1/2, which is equal to 1. It's like asking, what's the probability that the dart landed on the 2, given that it landed on the 2? That's a sure bet!

Okay, that wraps up this introduction to the general conditional probability rule in probability theory. I hope this lectrue gives you a better sense of why the rule has the form it does and why it works.

7. Total Probability Rule

We're working up to a derivation of Bayes' Rule from the general rule for conditional probability. An important step in that proof involves an application of what's called the "**law of total probability**". This is what it looks like:

$$\mathbf{P(A) = P(B)\ P(A \mid B) + P(\text{not-}B)\ P(A \mid \text{not-}B)}$$

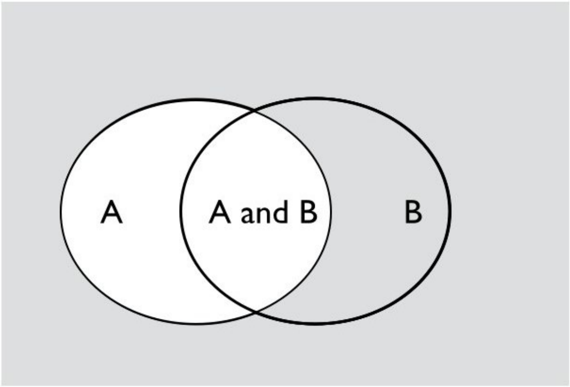
This is actually a very handy rule, very useful for working out certain kinds of problems, and it's just interesting enough to warrant its own discussion, so here it is.

We'll start off with an algebraic derivation using Venn diagrams that will help us understand that scary expression above, and then look at an example calculation.

The basic idea behind the law of total probability is that **you can think of the unconditional probability of an event as a sum of conditional probabilities, where you're conditioning over all the various alternative ways that the event could come about.**

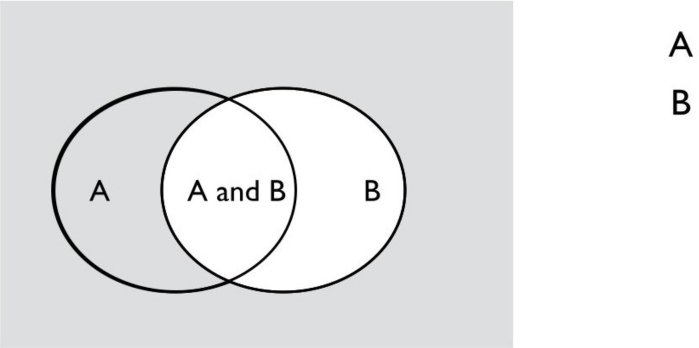
Below is the standard Venn diagram we've been using. We're interested in how different kinds of events are represented here. How many are there?

At first glance it looks like three. There's event A, represented by the white oval:

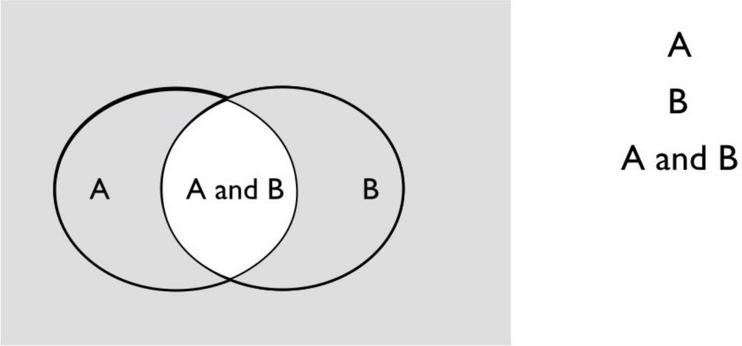


A

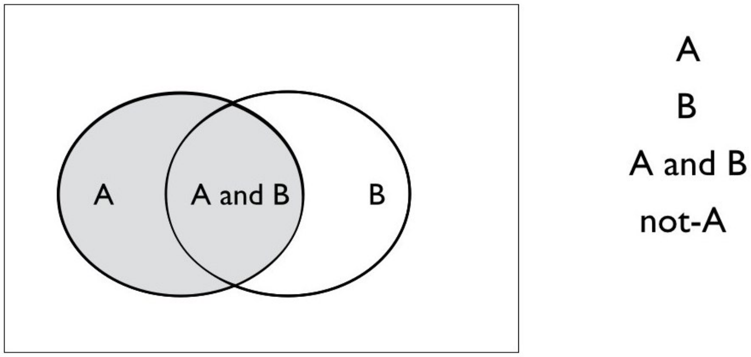
There's event B, represented by this other white oval:



And there's the event where A and B both occur, represented by the overlap region:

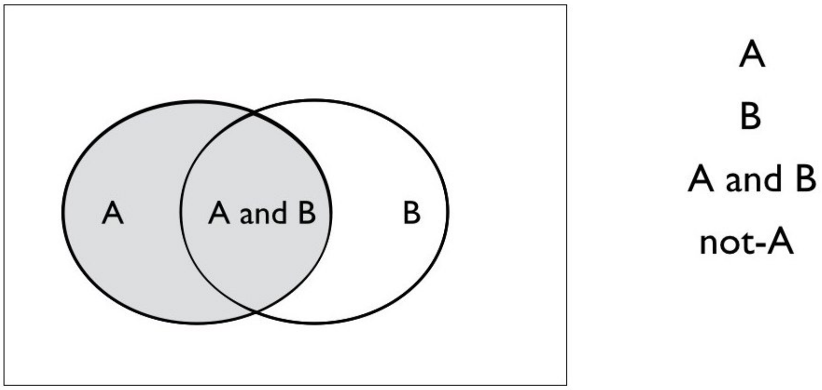


But there are more events on this diagram. There's the event associated with A NOT occurring, which looks like this:



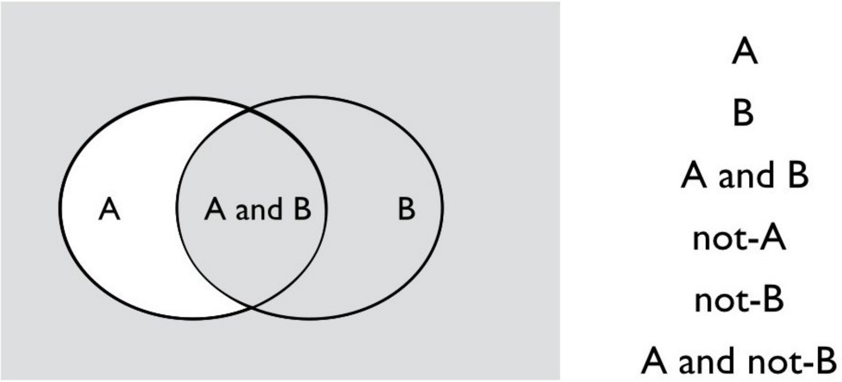
We call this the logical or set-theoretic **complement of A** — everything in the event space which is **outside of A**, covering all the cases where A does **not** occur.

Similarly, there's the complement of B:



which represents "**not-B**", the event in which B does **not** occur.

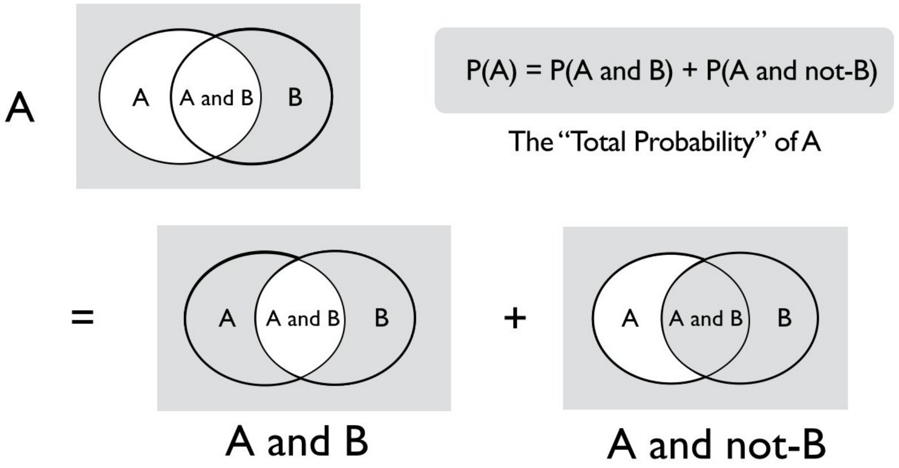
And you could have the area outside of the overlap region as well, representing all the cases where A and B don't occur at the same time, but I want to look at a specific overlap region. I want to look at the **overlap between A and not-B**:



These are the events that are inside A, that are also OUTSIDE of B, the events where A occurs but B does NOT occur.

So, what's the point of this exercise? The point is that we have more than one way of representing an event, and we can use these alternate constructions to give us information about an event.

Consider for example event A again. Here it's just a simple event and we're not paying attention to what events its overlapping with. But when do pay attention, we see the following:


A
A
A and B
B
P(A) = P(A and B) + P(A and not-B)
The "Total Probability" of A
=
A
A and B
B
+
A
A and B
B
A and B
A and not-B

Event A can be viewed as the sum of two overlap regions. The first is the overlap region between A and B. The second is the overlap region between A and NOT-B. This gives us a new logical representation of event A:

A is just the sum of (A and B) and (A and not-B)

But then we immediately have a representation for the **probability** of A, since these areas correspond to probabilities.

P(A) = P(A and B) + P(A and not-B)

The probability of A is equal to the probability that A and B will both occur, plus the probability that A will occur and B not occur.

This is a version of the law of total probability.

This doesn't quite correspond to the big long formula we saw at the start of the video with all the conditional probabilities in it, but to get there we just do a little rearranging and apply the general conjunction rule, which if you recall, expresses conjunctions in terms of conditional probabilities.

So we start with this expression:

P(A) = P(A and B) + P(A and not-B)

And we can switch the order of terms since conjunction is a commutative operation, you get the same answer either way:

P(A) = P(B and A) + P(not-B and A)

And now we just apply the **general conjunction rule** to each of these conjunctions:

P(B and A) = P(B) x P(A|B)

P(not-B and A) = P(not-B) x (P(A|not-B)

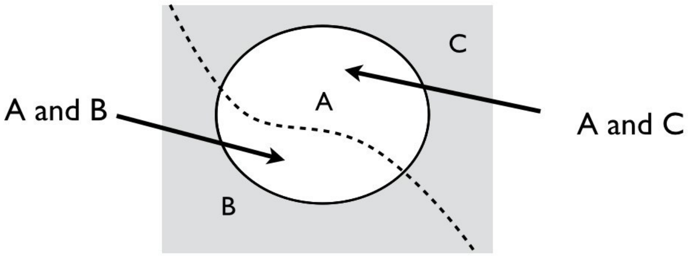
… and substitute them in our total probability expression to give this:

P(A) = P(B) x P(A|B) + P(not-B) x P(A|not-B)

… which is just the **law of total probability** that we gave at the top of this lecture.

We need to pay attention to the order when applying this rule so we don't get confused about what's playing the role of the As and the Bs, but it's just a simple substitution.

One thing I want to note here is that there's nothing special about having only two events, A and B. In general you can imagine A overlapping with any number of events.



A = (A and B) + (A and C)

P(A) = P(A and B) + P(A and C)

Here are two events, B and C, that overlap with event A. Event B is everything to the left of the dotted line in the event space, event C is everything to the right of the dotted line in the event space.

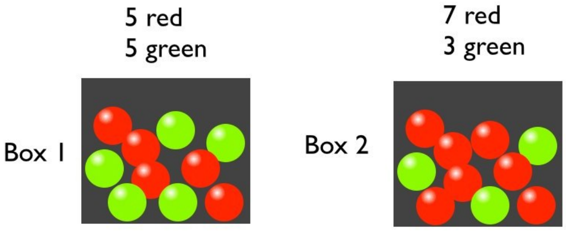
The law of total probability says that we can represent the probability of A as the sum of the intersections of these two events with the event A.

Substituting our formulas for the conjunctions in terms of conditional probabilities, we get this expression, which is the law of total probability for this case:

P(A) = P(B) P(A|B) + P(C) x P(A|C)

An Example Calculation

Okay, let's look at an example total probability calculation.



What is the probability that a ball drawn from a randomly chosen box is green?

You're given two boxes, each with ten balls in them. You're told that box 1 has 5 red ball and 5 green balls in it, and box 2 has 7 red balls and 3 green balls. You can't see inside them, so you don't know which is which. You pick a box at random and draw out a ball.

Question: What is the probability that the ball you drew out is green?

This is a typical total probability question. What you're actually given is a lot of information about conditional probabilities and from this you're supposed to work out an unconditional probability, the probability of drawing a green ball.

Let's first define our events:

- Let G be the event that the drawn ball is green.
- Let B1 and B2 be the events of picking box 1 and box 2 respectively.

Now let's look at the equation for total probability. Here's the equation:

P(G) = P(B1) x P(G|B1) + P(not-B1) x P(G|not-B1)

P(G) = P(B1) x P(G|B1) + P(B2) x P(G|B2)

In the second line I've simplified the expression in an obvious way, since in this context, the act of not choosing box 1 is equivalent to choosing box 2, since they're mutually exclusive and exhaustive alternatives.

So this reads, the probability of drawing a green ball is equal to the probability of choosing box 1, times the probability of drawing a green ball out of box 1, plus the probability of choosing box 2, times the probability of drawing a green ball out of box 2.

Now we write down the information we have and see if we can fill in these terms.

The probabilities for choosing box 1 and box 2 are easy, they're both 1/2 or 0.5 since we're randomly choosing between them. So

P(B1) = 0.5

P(B2) = 0.5

And the probabilities for choosing green given that we're in a particular box are easy too, since they're just the fractions given: half the balls in box 1 are green, so that's 0.5, and 3 of the 10 balls in box 2 are green, so that 0.3. Thus,

$P(G|B1) = 0.5$

$P(G|B2) = 0.3$

And now we just substitute these values:

$P(G) = P(B1) \times P(G|B1) + P(B2) \times P(G|B2)$

$P(G) = (0.5)(0.5) + (0.5)(0.3)$

$P(G) = (0.25) + (0.15)$

$P(G) = 0.4$ or a 40% chance

Which makes sense. The probability of drawing a green ball should be less than 50%, since in box 2 only 30% of the balls are green. But it should be higher than 30%, since there's a 50% chance that you'll pick box 1, which has more green balls in it. **The answer is a weighted sum of the two conditional probabilities.**

We're in a good position now to introduce **Bayes' Rule**. Bayes' rule is intended to answer a different version of this problem.

Here's the problem:

> **Given that a green ball was in fact drawn, and not knowing which box it was drawn from, what is the probability that it came from box 1 (or alternatively, box 2)?**

This is a typical Bayes' Rule question.

Our intuition tells us that it's more likely to have come from box 1 than box 2, since box 1 has more green balls, but Bayes' Rule can tell us *exactly what this probability is*, given the information we have.

So let's move on now to Bayes' Rule.

8. Bayes' Rule

In this lecture we're going to take a look at **Bayes' Rule**, which is arguably the most famous rule of probability theory. It's famous because it is provides a model for an important process in human reasoning, namely, learning from experience. Bayes' rule can tell us how we should modify the strength of our belief in a particular hypothesis after we've learned some new bit of evidence.

Here we're just going to show how Bayes' Rule follows from the general rule for conditional probabilities, and look at two example calculations.

A Derivation of Bayes' Rule

Here's the general rule for conditional probability. It serves as the basic definition of conditional probability in probability theory:

$$P(A \mid B) = \frac{P(\text{A and B})}{P(B)}$$

We can rewrite this formula by swapping out the numerator with the general formula for conjunctions:

$$P(A \mid B) = \frac{P(A) \text{ x } P(B \mid A)}{P(B)}$$

This is just a rearranged form of the general conditional probability rule, so really we're just swapping around terms here.

What we have now is the simplest form of Bayes' Rule. But there's a more useful formulation that is more commonly used, which we get by rewriting the denominator in terms of the **total probability of B**:

P(B) = P(A) x P(B | A) + P(not-A) x P(B | not-A)

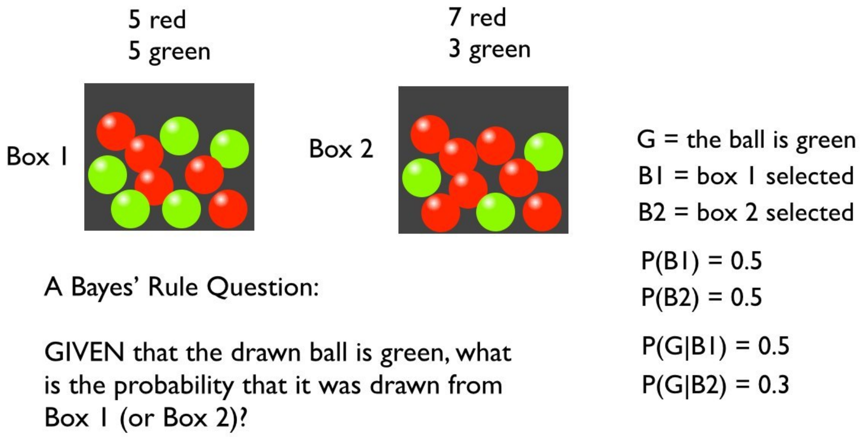
Substitute this expression into the denominator above gives the following:

$$P(A \mid B) = \frac{P(A) \text{ x } P(B \mid A)}{P(A) \text{ x } P(B \mid A) + P(\text{not - A}) \text{ x } P(B \mid \text{not - A})}$$

If you don't have a clue where the expression for total probability comes from then I'd recommend going back to the previous lecture where we derived it. But this is the "working version" of Bayes' Rule, it's the version we generally use in calculations.

Now let's look at the problem we left off with at the of the last lecture and see how Bayes' rule can be used to solve it.

A Bayes' Rule Calculation



A Bayes' Rule Question:

GIVEN that the drawn ball is green, what is the probability that it was drawn from Box 1 (or Box 2)?

G = the ball is green
B1 = box 1 selected
B2 = box 2 selected

$P(B1) = 0.5$
$P(B2) = 0.5$

$P(G|B1) = 0.5$
$P(G|B2) = 0.3$

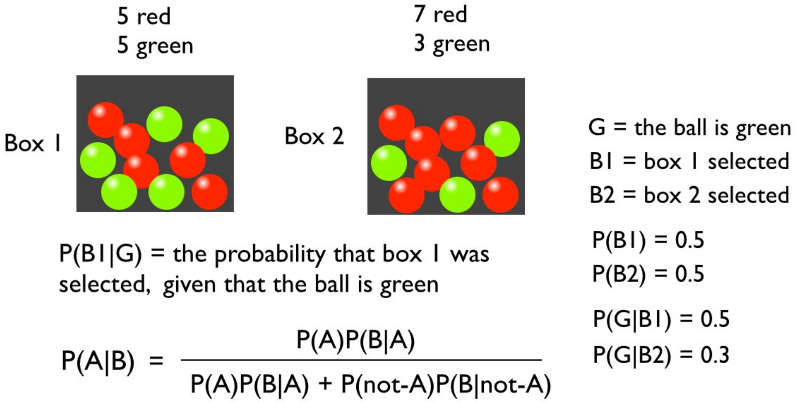
We're given two boxes, each with a different proportion of red and green balls in them. Box 1 has 5 red and 5 green balls, while box 2 has 7 red and 3 green. A box is randomly chosen, we don't know which one, and a ball is drawn. The ball is green.

Question: What is the probability that the ball came from box 1? Or alternately, from box 2?

The information on the right in the figure above summarizes what we know about the setup. The probability of picking any given box is 0.5 because it's random. And the probability of picking a green ball out of any particular box is given by the proportion of green balls in the box, which are 0.5 and 0.3 respectively.

Now, what is that we want to calculate? It's $P(B1|G)$, the probability that box 1 was selected, given that the ball is green.

Now let's look at Bayes' Rule.



5 red
5 green

7 red
3 green

Box 1

Box 2

G = the ball is green
B1 = box 1 selected
B2 = box 2 selected

P(B1) = 0.5
P(B2) = 0.5

P(G|B1) = 0.5
P(G|B2) = 0.3

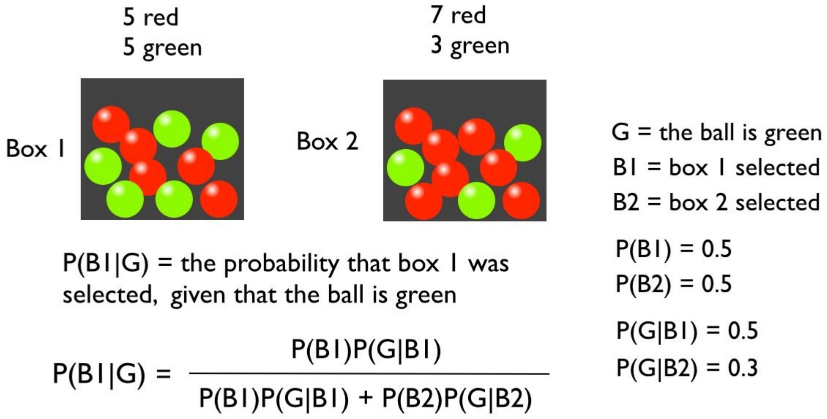
P(B1|G) = the probability that box 1 was selected, given that the ball is green

$$P(A|B) = \frac{P(A)P(B|A)}{P(A)P(B|A) + P(\text{not-}A)P(B|\text{not-}A)}$$

This is the general rule in terms of generic As and Bs. We need to rewrite this in terms of G, B1 and B2.

Note that A is going to become B1, the event of choosing box 1. What is not-A? Well, the setup says that we've only got two options, box 1 or box 2, so if it's not box 1 then it's got to be box 2. So not-A is going to be replaced with B2, the event of selecting box 2.

And here's what we get when me make those substitutions.



5 red
5 green

7 red
3 green

Box 1

Box 2

G = the ball is green
B1 = box 1 selected
B2 = box 2 selected

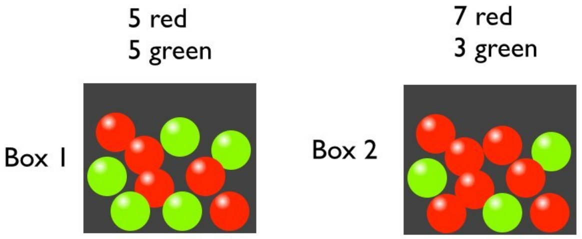
P(B1) = 0.5
P(B2) = 0.5

P(G|B1) = 0.5
P(G|B2) = 0.3

P(B1|G) = the probability that box 1 was selected, given that the ball is green

$$P(B1|G) = \frac{P(B1)P(G|B1)}{P(B1)P(G|B1) + P(B2)P(G|B2)}$$

This is Bayes' Rule for this particular question. And notice that every term in this expression is known, they're all on the side in the information given. So now we just substitute:



G = the ball is green
B1 = box 1 selected
B2 = box 2 selected

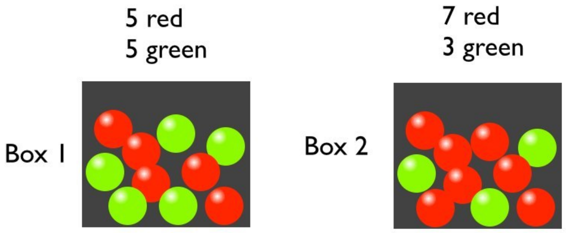
P(B1) = 0.5
P(B2) = 0.5

P(G|B1) = 0.5
P(G|B2) = 0.3

P(B1|G) = the probability that box 1 was selected, given that the ball is green

$$P(B1|G) = \frac{(0.5)(0.5)}{(0.5)(0.5) + (0.5)(0.3)}$$

You see now how the numbers are actually pretty easy to work with. When we evaluate the products and do the sums we get the answer:



G = the ball is green
B1 = box 1 selected
B2 = box 2 selected

P(B1) = 0.5
P(B2) = 0.5

P(G|B1) = 0.5
P(G|B2) = 0.3

P(B1|G) = the probability that box 1 was selected, given that the ball is green

$$P(B1|G) = \frac{0.25}{0.25 + 0.15} = \frac{0.25}{0.4} = 0.625$$

The probability that the ball came from Box 1 is 62.5%

This example illustrates how the Bayes calculation can be viewed as a model of **learning from experience**. Before we knew the color of the ball, we were completely ignorant about which box was chosen, and the probability of it coming from box 1 was just 50%, reflecting this ignorance. Now, after having come to know the color of the ball, we can revise the probability of the hypothesis, and we see that it's more likely that the ball came from box 1 than from box 2. Which is exactly what we would expect, given the proportions of green and red balls in the boxes, but Bayes' rule gives us a precise estimate of how much more likely it is. That's the power of the rule.

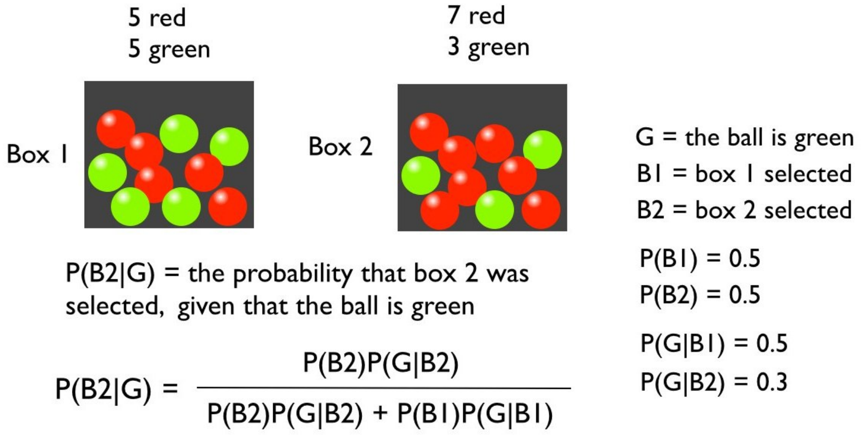
Now, can you guess what the probability is of the ball coming from box 2? There are two ways to do this, a short way and a longer way. The short way is just to realize that we've only got two possibilities here and their probabilities have to add up to 1, so the short way is like this:

$P(B2|G) = 1 - P(B1|G)$

$= 1 - 0.625$

$= 0.375$

The probability is 0.375, or roughly 37.5%. The long way to do is to solve Bayes' Rule for Box 2 instead of Box 1. That setup looks like this:



G = the ball is green
B1 = box 1 selected
B2 = box 2 selected

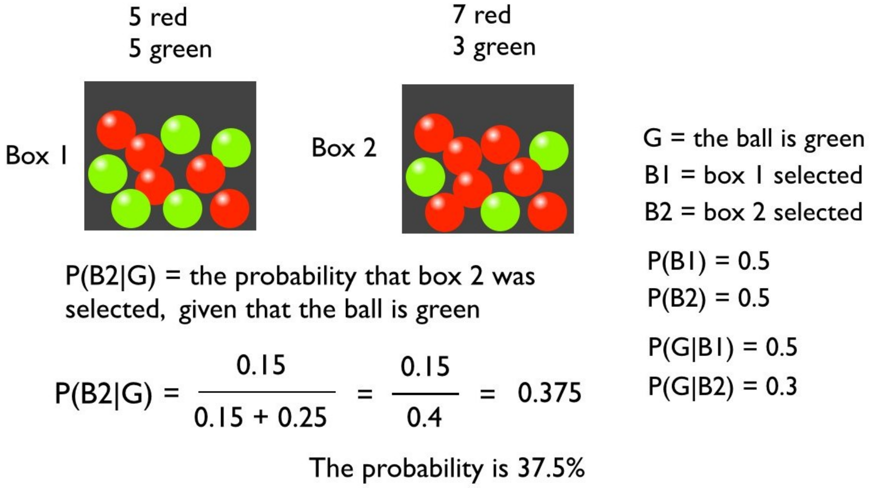
P(B1) = 0.5
P(B2) = 0.5

P(G|B1) = 0.5
P(G|B2) = 0.3

P(B2|G) = the probability that box 2 was selected, given that the ball is green

$$P(B2|G) = \frac{P(B2)P(G|B2)}{P(B2)P(G|B2) + P(B1)P(G|B1)}$$

And when you plug in the right values, you get this:



G = the ball is green
B1 = box 1 selected
B2 = box 2 selected

P(B1) = 0.5
P(B2) = 0.5

P(G|B1) = 0.5
P(G|B2) = 0.3

P(B2|G) = the probability that box 2 was selected, given that the ball is green

$$P(B2|G) = \frac{0.15}{0.15 + 0.25} = \frac{0.15}{0.4} = 0.375$$

The probability is 37.5%

which is the same answer. So it works either way.

Another Example

Let's look at another example.

Here's a typical sort of problem you often see on math tests.

> **"Two computer companies sell computer chips to a technology company. Company A sold 100 chips of which 5 were defective. Company B sold 300 chips of which 21 were defective."**
>
> **Question: What is the probability that a given defective chip came from Company B?**

When we approach a problem like this the first thing we should do is define our variables and write down what we know in terms of those variables.

So let

A = the chip came from Company A

B = the chip came from Company B

D = the chip is defective

The question we're being asked to solve is, what is the probability of B, that the chip came from company B, given D, that it was defective? That is, **what is P(B | D)**?

Now let's write down what we know:

$P(D|A) = 5/100 = 0.05$

$P(D|B) = 21/300 = 7/100 = 0.07$

These are the conditional probabilities of getting a defective chip from each of the companies, respectively. You can just read these off the question, 5 out of 100 and 21 out of 300, which give us 0.05 and 0.07 when we simplify them.

Now, we're also going to need the unconditional probabilities for A and B — these are the prior probabilities P(A) and P(B) before taking into account the new information that the chip was defective.

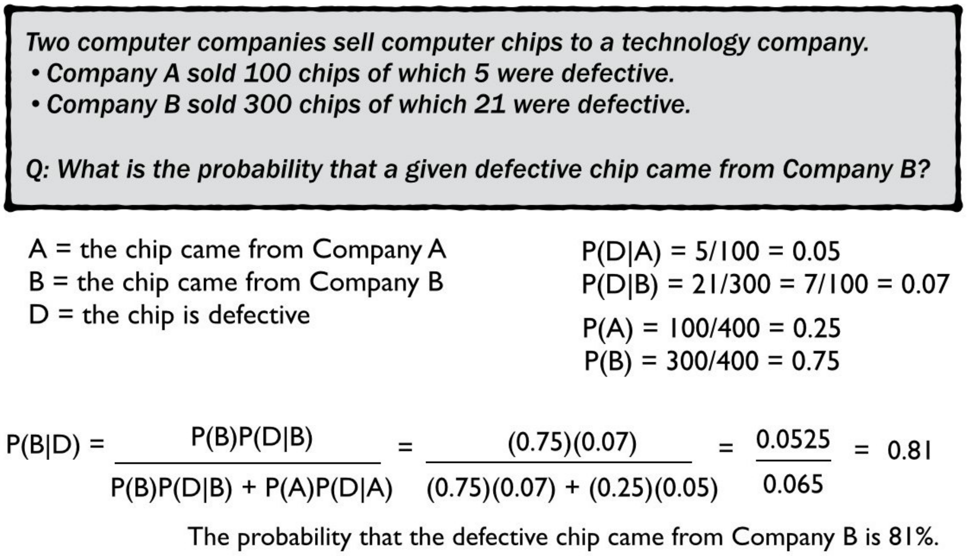
In this case the prior probabilities aren't distributed equally across the alternatives. The question says that 400 chips in total were bought, with 100 coming from company A and 300 coming from company B. So any random chip is more likely to have come from B than from A.

More precisely, the probability of a chip coming from Company A is 0.25, or 25%, and from Company B is 0.75, or 75%:

P(A) = 100/400 = 0.25

P(B) = 300/400 = 0.75

Now, to solve this problem, all we have to do is apply Bayes' Rule. When we substitute our values we get this:



Two computer companies sell computer chips to a technology company.

- *Company A sold 100 chips of which 5 were defective.*
- *Company B sold 300 chips of which 21 were defective.*

Q: What is the probability that a given defective chip came from Company B?

A = the chip came from Company A
B = the chip came from Company B
D = the chip is defective

P(D|A) = 5/100 = 0.05
P(D|B) = 21/300 = 7/100 = 0.07
P(A) = 100/400 = 0.25
P(B) = 300/400 = 0.75

$$P(B|D) = \frac{P(B)P(D|B)}{P(B)P(D|B) + P(A)P(D|A)} = \frac{(0.75)(0.07)}{(0.75)(0.07) + (0.25)(0.05)} = \frac{0.0525}{0.065} = 0.81$$

The probability that the defective chip came from Company B is 81%.

P(B|D) = .81, or 81%.

The prior probability of any given chip coming from company B was 75%. But once we learned that the chip was defective, and we knew the ratio of defective chips that come from company B, and that it actually has a worse track record, percentage-wise, than company A, this information raised the probability that the defective chip came from company B to 81%.

Okay, I think this will do for an introduction to Bayes' Rule. In the tutorial course on fallacies of probabilistic reasoning we'll come back to Bayes' rule and talk about how most people's intuitive judgments about conditional probabilities are flawed because they fail to consider base rates, the prior probabilities of events, in their estimates. This can have very serious consequences when we're dealing with, for example, a doctor's estimate of how likely it is that someone has the HIV virus, given that they've tested positive for it. But we'll save that discussion for another course.