[SQUEAKING]

[RUSTLING]

[CLICKING]

ANKUR MOITRA: All right, everybody. Welcome back to class. So today we're going to pick up where we left off. So remember, last time we talked about the pigeonhole principle. I showed you a whole bunch of applications of it. And then we started this longer topic that's going to take us a couple of lectures, which is probability.

So just a quick recap of where we left off last time. So last time I started to introduce the basic terminology that we need for probability. We had these key definitions, like we had the notion of a sample space, which is really just telling you what are all the possible outcomes of your experiment.

You could think about something very small and finite, like I roll a die and I get a number between 1 and 6. Or you could even think about a sample space as being an infinite space. For purposes of our class, we're only going to consider discrete sample spaces, but you can even consider sample spaces like what happens if I take a coin and I flip it as many times until I get my first heads? That's something which has an infinite sample space. And what's important is that to the sample space we associate what's called a probability measure P.

So all this does is it takes in some particular outcome from the range of all possible outcomes. It better assign a non-negative probability to that. And additionally, your probability measure had better be normalized so that the sum of p of x over your entire sample space had better be normalized so that it's equal to 1, so that these probabilities make sense.

In fact, there are a lot of distributions that we're going to see over and over again. One of the simplest and most ubiquitous distributions is called the uniform distribution. So here, we have some finite sample space S. And then we just look at the probability measure that assigns the same fixed probability to every possible outcome. So in particular, in order for it to be normalized, each of those probabilities had better be 1 over the size of S.

So this is where we left off last time. We also defined this notion of what an event was. Sometimes we're not interested in one particular outcome, but really a collection of outcomes. Something like when I take a die and I roll it and I get an even number, that's not just any one atomic outcome, but it's really three outcomes for 2, 4, and 6. And we talk about the probability of an event as really just being the sum of the probability measure over all of the outcomes that land in the subset T.

So this is where we left off last time. And today, what we're really working towards is we're going to talk about something that's very important, but a little bit counterintuitive. We're going to talk about independence and conditioning.

So what's important to build up to this is really the idea that we cannot just consider the probabilities of individual events, but we can combine them in interesting ways. So for example, if I give you two events, T1 and T2, I can think about the union of T1 and T2. What this means is I'm really just treating this as the set of outcomes I'm interested in. This is another set of possible outcomes I'm interested in. And I'm forming a new event that takes everything that happens in either one of these or maybe both.

So in particular, in words, what does this new event that I've constructed mean? It's the event that at least one of these events, T1 or T2, happened. So that's one way to combine events. I could also think about combining events by, for example, taking T1 intersect T2. And then what am I going to get? I'm going to get a new event that's asking me whether both T1 and T2 happened.

And so this now leads to a very interesting concept, which is the notion of conditional probability. And I promise that this will sound very simple. But in the examples we'll see, some of them will become pretty counterintuitive. In fact, when you look, conditional probability is really in the world all around you in very important ways. We'll talk about some real world examples.

So first, let me tell you what this means. Let me just give you the definition. We can also combine two events, A and B, in this interesting way, where we want to assert for a fact that event B really did happen. And we want to look at if I tell you that event B happened, what's the new probability that this other event A happened.

So you can think about this in a few different ways. So in words, it's the probability of event A given or conditioned on that event B happens. See, one way to think about this is really in terms of pictures.

So we could start off with our full sample space S. That's all of the different possibilities. We could take one event A right here. And what's going on is that our probability measure really describes a way of throwing a dart at this board. Now, this dart always lands inside this box S, because that's the range of all possible outcomes. And then the probability of the event A is just the probability that that dart lands inside this oval that I've drawn right here.

So now when I have a pair of events and I have A and B right here, well, what I can do is imagine I do this experiment where I throw my dart at the board. And we already know it has to land inside the set S. But then what if I tell you that it's landed in this set B? And I promise you that event B happens. What I want to know is, given that promise, what's the new probability that it actually lands inside the set A?

So the only place where it could then land and event A could happen is right here in the intersection. And that's exactly how we define the conditional probability. It's literally just the probability of the intersection of the two events divided by the probability that I'm conditioning on. And this part is intuitive. Imagine that I have the same chance of hitting every single point inside this rectangle.

Then really, what's going on is that when I tell you that event B has happened, the probability of B is how I'm renormalizing my space of possible outcomes, because I know that now my dart is uniform over the set B. And what's important is not the amount of space that a takes up inside S, but the amount of space that A intersect B takes up inside B. So does this make sense? Are there any questions about the definition of conditional probability? Good? All right.

So let's develop a few different ways to think about conditional probability. Let me actually give one other simple example, which is related to it, which will have its own quirks. The other crucial notion that I want to introduce today is the notion of independence. So we talked about two events, A and B. And there's this very important concept, which we're going to use all the time, which is called independence.

So we say that two events A and B are independent if the probability of A intersect B is equal to the probability of A times the probability of B. So this is, in some sense, the simplest way that the probability of intersecting two events could work is that whether A happens, you think of it as it contributes this P of A. And whether B happens contributes this P of B.

But in fact, just as an exercise, using our definition for conditional probability, we can see that independence is really equivalent to a much simpler thing, which is that probability of A given B is equal to the probability of A. So this comes just from dividing through by the probability of B And using my definition for conditional independence. So these are two completely equivalent conditions. You can check either one. And that's what it means for two things to be independent.

So what's going on is that in my dartboard example in pictures, I might have my entire rectangle represents S. I might have this left half represents A. And maybe this bottom half represents B. And so what's going on is if I look at the probability of A, that's how much area this rectangle for A takes up out of the entire rectangle, but then when I condition on B, it doesn't change what that fraction is. So this is one way to think about what independence is.

And I want to caution you that even though it's a very simple definition, it's actually very counterintuitive in some ways. There are easy pairs of events that are natural to think about why they're independent. And then there are some events that are independent that are very counterintuitive.

So let's do the simple one first. Let's do some simple examples to get some intuition for independence. So imagine that we throw a dice twice. And this is a fair die, so it's the uniform distribution over all six possibilities. And when I look at the pair of throws, my sample space now has 36 possibilities.

For every possibility for the first throw x1, every possibility for the second throw x2. So I look at all possibilities of x1 x2 that are in this set 1, 2, up to 6. So this is my sample space, and everything is uniform over the sample space. All of the 36 possibilities for the choice of some number between 1 and 6, and some number between 1 and 6, are all equally likely.

And now I can define some events. Let's look at events A, which is that x1 is odd. And let's look at the event that B, that x2 is at most 3. So probably you all have very good intuition for this. So who thinks these two events A and B are independent? OK. Who thinks they're not? Who is still tired? OK. All right. Good honesty.

All right. So these two are independent. Now, one way to check independence is literally just to follow the definition. So when in doubt, just compute the thing. But we're going to see part of the intuition for why they're independents we'll see in the computation itself.

So what does it mean to be independent? Let's check. Let's look at the probability of A and B. So in our sample space, there are 36 possibilities. How many possibilities meet the condition that both event A and event B hold? Not a hard question. Yeah? A quarter of them. Yeah.

So let me just say 9 out of 36. Why is it 9? Because I have three possibilities for x1 is odd. 1, 3, 5. I have three possibilities for x2 is at most 3, 1, 2, 3. And every choice of one of the possibilities for A and every choice for one of the possibilities of B defines a possible outcome that satisfies both A and B. So they're independent because they have this kind of product structure that whatever happens for A, I can choose something there, and then I can define something that happens for B. So literally, if you think about it, it really represents this picture, if you think about it the right way.

So in any case, we can check that this really does equal the probability of A times the probability of B. Because what's the probability of A? It's just 1/2. 1/2 of the roles are odd. What's the probability of B? It's 1/2 because 1/2 of them or 3 or less. So

The way to think about it is that independence is sometimes intuitive. Because what's going on here is I could literally think about rolling two separate die. I roll one die here, and I care about what the roll is. And then you go to another room in building two and you roll another die.

And they're independent, because we literally are not talking to each other. These events are just happening in different spaces. But independence is a bit more counterintuitive than that, because there are other ways that events can be independent that satisfy our definition, that aren't so easy to check. Let me also mention that a lot of times we use the shorthand, which just means that A and B are independent as events.

All right. So let's go a little further with this example so that I can convince you that independence is subtle. So let's define a new event C, which is going to be the event x1 minus x2 is equal to 3. Now, this is a bit stranger of an event, because it involves both x1 and x2.

Now, what are the possible outcomes in the sample space where event C happens? Can anyone list them for me?

STUDENT: [INAUDIBLE]

ANKUR MOITRA: Perfect. And those are the only possibilities. So another way to define this event is really as the subset of the sample space, which we're interested in, which is the possibility 4, 1. That's the possibility 5, 2. And lastly, we have the possibility 6, 3. That's the only subset of the sample space that I'm interested in. Those are the only things that can happen where this event C is on, is true.

And we can look at the same type of thing. Let's look at our definition for independence. Let's look at the probability that A and C see both happen. So what was A? x1 is odd. And there's only one possibility on my list where x1 is odd and C happens. So this means the probability of my intersection is 1 over 36.

Unfortunately, this is not equal to the probability of A times the probability of C. We know the probability of A is 1/2, and we know the probability of C is 3 over 36, just based on our description of the event. So these two events A and C are not independent. A is not independent from C.

And maybe this is intuitive too, because these two events depend on the same things. They both depend on what happens for roll one. But what I claim is that there are slight changes to event C that actually make these things be independent.

Let's define a fourth and final event. Let's look at this event D. That's x1 minus x2 is equal to 4 instead of 3. Now we have only two possibilities. We have the possibility 5, 1, and we have the possibility 6, 2. And we can just check our definition for independence. So the probability of A and D. Well, there's only one thing that survives. So my probability of the intersection is still 1 over 36. And you can check that this really is equal to the probability of A, which is 1/2, times the probability of D, which is 2 over 36.

So this is an example which would be a little bit hard to guess unless you literally did the calculation, because it doesn't look like they're independent, but they still are. So are there any questions about independence? Does this make sense? OK.

All right. So let's go a little further. I want to tell you about some more powerful tools in probability that play very nicely with conditioning and independence. Let me tell you about this basic fact that we're going to use all over the place. And then I'll give you a way to visualize it. This fact is called the law of total probability.

Even though we already defined whatever we need, just in terms of the sample space and the probability measure, there are a lot of tricks that are going to happen where there are clever ways to compute probabilities that don't involve listing the entire sample space and just adding things up in a boring way.

So a lot of these tools are really a way to organize and think about the sample space to try and make probability more intuitive. And the basic version of the law of total probability is that if I have an event A and I care about its probability, we already have a definition for how to compute that probability just by going over the entire sample space. But what I claim is that we can break it up into two different terms, one where A and B both happen. And then I can add in the probability that A happens and B does not happen.

So this is just if I'm interested in the probability that A happens, well either A happens and B happens or A happens and B doesn't happen. These together cover the space of all things that could happen so that A still occurs, and they don't overlap at all.

So in particular, the proof of this is there really is no proof. It's literally just we take the definition of what the probability of A is. It's the sum over all x in my set A of p of x. And all I'm doing is I'm breaking the sum up into two pieces, it's the sum over all x in A and B of p of x. That's literally this right here plus the sum over all x in A and not B of probability of x.

And so pictorially, we can go back to our favorite diagram. We have this event A that we care about. We have this other event B. And then we can break up this event A into two pieces, one where A happens and B does not, and one where A happens and B does. And together, they cover the space.

So, in fact, there's an even fancier version of the law of total probability, because no one says you have to do this with one other event. In fact, you can do this with a ton of other events. So let's talk about the most general version of the law of total probability.

We can be interested in the probability of some event A, and we can write this as the sum over a collection of other events of the probability of A and BI. But now we need some conditions on the BIs. We need that all of the BIs are disjoint. So BI intersect BJ is the empty set for I not equal to J. They're pairwise disjoint. And we also need that together, they cover the entire space. The union of all of these BIs is my sample space.

So in particular for this most general definition for the law of total probability, really, the way that I recover that special case we already talked about is I consider two events, B1 and B2, and B2 is the complement of B1. And that way, I just get two terms and the sum. They're definitely disjoint. B1 and its complement don't intersect. And together, they cover the space.

But instead of breaking up my entire sample space into only two possibilities, I can think about breaking it up into three, four, five, how many ever is convenient. So any questions about the law of total probability? Good? All right.

So let's go back to visualizing these things. I want to try to make these things as intuitive as possible. So in fact, there's a clever way to think about them using a diagram, which I find very helpful. So we can think about our random process as really taking some sort of path in a tree. Maybe we start off our random process, and then we have a whole bunch of possibilities, like B1, B2, B3.

So what I'm deciding here is really the answer to the question, which of the BIs happens? Because in the setup of the law of total probability, all these BIs are pairwise disjoint, and they cover the space. So when I start this process, before I actually tell you what the outcome is, where the dart lands, I'm first going to tell you which region the dart lands in.

Does it land in B1? Does it land in B2? Does it land in B3? The assumption that these things are pairwise disjoint just means that one and only one of these things happens. But I have a different probability of traversing these edges depending on what the probability of those events are.

And then finally, I can look at conditioned on B1 happening whether the event A actually happens. So I can continue this process. And then what I can ask is, how does, knowing that BI happens, how does that change the probability of A?

So this is a nice way to visualize it. Because what's going on in here is you can write it out instead using conditional probability. Another equivalent way to do this is it's the sum over I of the probability of A conditioned on BI times the probability of BI. Literally nothing is happening here. I'm just using the definition of what conditional probability is.

And then you can see that algebraically, what's happening is first you're telling me which BI happened. That contributes this term right here. And then once you tell me that BI happens, I have to also take into account the remaining conditional probability. And when I add up over all the possibilities, that gives me the total probability of the event. So these are very basic algebraic manipulations, but they're powerful. So any questions? All makes perfect sense?

STUDENT: So what are the nodes, the bottom?

ANKUR MOITRA: Right. So this would be these are all of the possible outcomes in the sample space. So what I'm doing is instead of just telling you first where the dart lands, like which x happens, first I'm going to tell you whether B1 happens, B2 happens, or B3 happens. Once I tell you that B1 happens, I know my dart lands in that space. I look at all the x's that belong to B1, and maybe I just shade in the x's that belong to that subset that are also in my set A.

So for example, the fraction of nodes here that are colored is exactly the probability of A conditioned on B1. So you can see that I have this probability of taking B1, and then I have a conditional probability. Or I have the probability of taking this branch and then another conditional probability. This is pictorially what the law of probability is telling you. You can think about it many different ways.

This is tricky, so don't be embarrassed to ask questions. That's why I'm here. Any other questions? Everyone thinks this is way too easy? Way too hard? Just right? OK, I'll take it.

All right. Let's see. All right, let's check our understanding, because some of these things are a little bit counterintuitive. How many people have heard of the Monty Hall problem? OK, how many people feel totally comfortable with Monty Hall problem and think I should skip it or how many people want to hear it? Who wants me to skip it? OK, that's a very small number, so going to do it.

So you've all seen the Monty Hall problem before. What I claim is that a lot of these ways of thinking about things in terms of the law of total probability, in terms of this visual I gave you here, they give you a nice intuitive way to think about what's going on. So I'll give you maybe an explanation you haven't seen before.

So remember in the Monty Hall problem, there's a game show and there's three doors. And behind one of the doors is a nice prize. Maybe there's a car. But you don't know which door has the prize. So the way that the game works is that first, you pick a door.

Now, whatever door you pick, if I pick this door, there's definitely another door that doesn't have the prize. If I pick this door that has the prize, there's still another door that doesn't have the prize. So either way, the host can always, no matter what you pick for your first door, reveal to you another door that you didn't select which does not have the price, and show you that you shouldn't pick that door. And the basic question is, do you switch?

So this is a very famous example. I think the first time you see it, it is a bit hard to wrap your head around. But it's been done in virtually every class that covers probability so often that maybe it's become a little bit more intuitive. As you all know, the answer is that you switch, that then you'll win this game with 2/3 probability.

Now, you can figure that out by computing conditional probabilities. You have to be very careful about the definitions of your events. But I think a very intuitive way to think about it is in terms of this tree that I drew before, because we can think about all the things that can happen in the game.

So in the beginning, when you choose a door, remember that the prize is put behind a door at random. So whatever door you choose, you have a 2/3 probability of selecting a door that has no prize, and you have a 1/3 chance of selecting the door that has the prize. Of course, the game isn't over then, and we can just think about what happens depending on your choice of strategy. You could, at that juncture, decide to switch. You could decide to not switch. And the same thing happens here. You could decide to switch or not switch.

And the way to think about it is that once I select a door that doesn't have the prize, then the host has no choice but to reveal to me the only other door that doesn't have the prize. And then when I switch, I definitely win. So I win as soon as I select a door in the beginning with no prize if I'm committed to the strategy switch. And if I don't switch, then of course, I lose.

And the same thing happens here, but in reverse. If I got unlucky and I selected the door with the prize and I committed to the strategy switch, then the host would show me one of these two doors without the prize, and I would switch, and I would lose my prize. But this only happens with probability 1/3, because that's the probability that this tree associates with it, and otherwise I'd win. So you can see that once I've committed to the fact that I'm going to switch, all I need to do is I need to get not super lucky, just I need to select the door which doesn't have the prize, and then automatically you can see that I will win.

So we'll do more examples of these kinds of things later, especially in recitation, like drawing out these diagrams to organize the state space. Let me tell you one more really cool tool, and then we'll do some real world applications of it. This involves, again, conditional probability. That's called Bayes' rule. And it's a very basic formula.

So imagine that we compute something the probability of B given A. It turns out there's a way to express this in terms of the conditional probability the other way around. You just have to correct by the right factors of P of B and P of A. So this is a very powerful rule. The proof is quite straightforward. It's just algebraic. Because all we have to do is just invoke the definitions.

So by definition, how did we define the probability of A given B? It was just the probability of A and B divided by the probability of B. That's literally what our definition was. I can take this definition, and I can just multiply by this correction term, which is probability of B over probability of A.

And if I do this on both sides of my equation, well, one side I'm going to get this right hand side. So I'll get probability of A given B times probability of B over the probability of A. And on my other side, the probability of B's are going to cancel. I'm going to get probability of A and B over probability of A. And I claim that this is the same thing as the probability of B given A. And that's because it doesn't matter what order I intersect these two things. Probability of A and B is the same thing as the probability of B and A. And then this is, again, literally the definition of conditional probability.

So the proof of this Bayes rule is very simple. In fact, anyone know anything historically about why people were interested in this. It's kind of an interesting historical fact. Yeah?

STUDENT: They wanted to prove God.

ANKUR MOITRA: Yeah, they wanted to prove God, the existence of God using Bayes' rule. I don't think they got there, and we're not going to get there today either, but we'll still do some fun stuff with it. It's kind of interesting. A lot of the foundations of probability have some theological component to them, but I'll leave that as a discussion for another time.

So we've got Bayes' rule. Let's do a powerful real world example. The numbers in this example are all completely real. So let me tell you some facts. So approximately 1% of women aged 40 to 50 get breast cancer. That is a shockingly high percentage, but it's a true fact.

And one of the ways that you can try and screen for breast cancer is a mammogram. So a mammogram is not a perfect test. So it has a 10% false positive rate. I'll tell you what I mean by that in a second. And it has similarly for its false negative rate.

So let's look at two events which we really care about, because our test is imperfect. We can look at A is the event that we test positive. And we can look at the event B that we really have the condition that's being tested for. In this case, breast cancer. These are the two events. Yeah?

STUDENT: [INAUDIBLE]

ANKUR MOITRA: Yeah. It is. Yeah. So developed in that time frame. So here, just to be clear, so what do I mean by the false positive rate? So this would be the probability of A given not B. So this is the chance the test says positive even though you don't have it. And we can also look at the false negative rate, which is the probability of not A given B. So you actually do have the condition, but the test doesn't pick it up.

So both of these things, a lot of these tests, we have much higher accuracy tests. But the idea is that you usually deploy some lower accuracy test as a screening procedure to try and get a rough idea. But then the question is, how much stock should you actually put in the test results?

So one of the conditional probabilities that's very important is what's the probability of B given A? So what does this mean? You test positive, and then you're very worried, but you want to know whether you actually have the condition or not.

So let me take a poll from the audience. Let's just say approximately, what do you think the probability of B given A is? After all, the mammogram has 10% false positive and false negative rate. So maybe you think you've got a 90% chance of having the condition. Maybe it's 75%. Maybe it's 25%. Maybe it's around 10%. Maybe it's 1%.

So you all said you understand conditional probability. You all said you're ready for a test on it. So here's my test. So what do you think the answer is? And this is a very medically important question, because when you test positive for something, you have many treatment options. Some of them are more or less invasive than others. So how sure are you that you have that condition? Is it worth it to do some very invasive and risky procedure?

So who thinks the answer is one, 90%? Who thinks the answer is two, 75%? Three, 25%. You guys have to vote. You can't abstain. You're a doctor. Someone's relying on you. Let's try this again. Who thinks the answer is one, 90%? Just take a vote. Take a chance. I'm not going to grade you on this. Who thinks the answer is two, 75%? Who thinks the answer is three, 25%? Who thinks the answer is four, 10%? Who thinks the answer is five, 1%?

OK, so if you answered two, then you're in good company, because that's the average physician's answer. There was actually a survey of hundreds of physicians. And when you average their scores, you get 75%. Unfortunately, you're in a group that completely got the wrong answer, because the answer is not at all 75%. It's actually about 8%. So 10% is very close. For anyone who answered four, could you give me some kind of intuition? Yeah.

STUDENT: I guess 1% is really low to begin with. So I don't know. [INAUDIBLE]

ANKUR MOITRA: Yeah, that's right. That's great Intuition Because if you think about it, already, I mean, this is a high percentage, but it's still in an absolute sense low. So you can think that the vast majority, when you think about a typical person who doesn't have it, 99% chance, then there's still a reasonable chance that they would still test positive if they were tested. And so among the people who are tested positive, that's the region in the dartboard I care about. Still, the disproportionate share of them are people who aren't going to have it.

So this is actually from a very famous book that I encourage you to read if you want to be depressed. It's called *Thinking Fast and Slow* by Kahneman. And it goes through all of these examples about how peoples, and especially doctors, it kind of picks on doctors, make very counterintuitive decisions.

So when I used to teach 042 more regularly, I would still do this example. I think it's a beautiful example. And then in 2016, we had our first kid, a baby girl. And I remember that when we were doing all these tests, not breast cancer, but other things like testing for the risk of Down syndrome, one of the crazy things is that at the time, what you do is you do an amniocentesis to do that. That's where you put a needle into the belly, and then you extract fluid to try and figure out whether various genetic conditions are happening.

The crazy thing is that a lot of the things they were testing for had very low percentages of occurring. They're terrible things when they happen, like Down syndrome. But the risk of causing a miscarriage with an amniocentesis, when I looked it up, was something like 1 inch 700. So I have to say that real world examples from having kids makes these things feel very differently. So if you're curious, I can tell you all kinds of conditional probabilities that doctors get wrong that relate to child rearing, but that's another topic for another time.

So we have a little bit of time left. What I'm going to do is I'm going to start the next topic. We're going to take these ideas even further. So far, what we've talked about is mostly events. We talked about ways to reason about events. We talked about what events were, how to combine them in different ways.

And we talked about these fancy ways about reasoning about their probability, either through the law of total probability or using conditional probabilities. We talked about how to reason about it through this tree diagram I gave you. That's a way to organize the sample space. We talked about the Monty Hall problem. We talked about some real world examples.

Well, it turns out that there's an even more powerful abstraction that we can do beyond events where things get even more interesting. Some of the theory goes over, but it changes in cool ways. So let me explain this in terms of how we're going to generalize what we've seen. So last topic for today I want to tell you about is what are called random variables.

So where are we right now? We talked about events, which are a subset of the sample space. We talked about a lot of properties of events. We talked about the property of independence and how, in some cases, it can be intuitive, like when the experiments are literally done separated from each other. Other times, it can be counterintuitive and you just have to check the definition.

We talked about conditioning. We talked about unions and intersections. And then random variables, instead of being a subset of the sample space, what they are is they're functions defined on the sample space. So what they do is they take an outcome and they ascribe to that outcome some real value. So this is the intuitive thing. Let's do the definition.

So a random variable is a function f that goes from my sample space and ascribes to each outcome some real value. And in fact, we'll be interested in a lot of properties of random variables. So one of the most useful and ubiquitous properties is called its expectation. So let's just define that now too. So the associated expectation of the random variable as just literally the average over your probability space of the values the function takes.

So I write it here. This is a little block E. And it's defined to be equal to the sum over all outcomes in my space of the probability of that outcome times the real value of the corresponding random variable. So this is just a weighted average of the value of the random variable, weighted by the probability of each one of these outcomes.

So random variables come up all the time. We don't just care about events. One of the simplest ways it might come up is even in understanding something like betting, where there are a lot of different outcomes that you could be betting on. And then something like the random variable would describe the payoffs to different parties. So let's do some simple examples to get some intuition for what the random variable is.

Let's do one where the sample space is infinite, just for fun. So let's toss a coin until you get a heads. We'll start off with a fair coin, and then we'll vary this experiment.

So first, what is the sample space? Well, we could have the first coin as heads already, and we stop there. We could have tails, heads. Could have tails, tails, heads, and so on. So any sequence of tails, any length from 0 to infinity with a heads appended at the end, that's our sample space. And we also care about what's the associated probability measure.

So furthermore, the probability of having I T's and then one H is equal to 1/2 to the I plus 1. Because I better get my first thing is a tails. That's half probability. My second thing is tails and so on. And there are I plus 1 things here. So the chance of getting exactly the string is just 1/2 raised to the I plus 1. And now we have enough information that we can compute the expected value.

So let's look at the random variable we'll be interested in will be just the number of coin tosses. So of course, this game in principle could go on forever, but the expected value gives us some baseline of what we expect it to be. We're going to talk about a lot fancier tools later in a few lectures' time. Because when we get back to random variables, we're going to talk about tools for arguing why random variables in some cases are very close to their expectation. That's very important in things like understanding the error of polling. But for right now, we just want to work with what a random variable is and get some crude idea of what we expect it to be approximately.

So the natural way to do this, to compute the expectation of f, is literally just to use the expression we gave you. So there's a 1/2 probability that the game is over instantly and that we get heads to begin with. There's 1/4 probability that this game goes on for two total flips. There is 1/8 probability that it goes on for three total flips and so on. And so we can write this as a series. It's the sum from k equals 1 to infinity of k times 1 over 2 to the k.

Already on the homework, we've shown you some examples of how to evaluate sums like this. We're going to talk about this in a lot more detail when we get to the counting unit, which is right after the probability unit. So we'll give you tools not only for evaluating the sums, but for thinking about what kinds of things they count.

But in any case, we can evaluate this very quickly just by arranging things in the right way. So from this expression, it's not that easy to guess what the answer is. But if we break up this quarter, this 2 times 1/4 into the sum of 2/4, it becomes a lot easier. So let's write it out pictorially. We have 1/2 from the first term, and then we have 2/4. And then we have 3/8. And this thing goes on.

And what we can do is we can evaluate the sum in a different way, line by line, row by row first. So if we have 1/2 plus 1/4 plus 1/8 plus 1/16 and so on. Well, this, you probably all know, the answer is 1 for what that part of the sum contributes. And if instead we have 1/4 plus 1/8 plus 1/16, well, that's going to give us 1/2. And the next thing will give us 1/4. And now when we sum up all of these guys along the column, we're going to get 2.

So this is very natural. The expected number of flips for a coin until you get your first heads is 2. Makes sense. And in fact, we can do this in a more clever way too. We'll use this in a minute, actually. So let me explain the even more general setting. Which leads me to something called mean time to failure. This is something that comes up in a lot of computer science applications.

Let me give you the story behind it, but it'll really be the same kind of exercise we just did. Imagine we've got some computer as a part of some giant computing system. And it's not a perfect computer, and it fails independently on each day with some probability P.

Well, one of the things you might be interested in, what's called the mean time to failure, is how many days you expect the computer to last before you replace it. This is a very basic probabilistic statement. It shows up all the time, especially in things like actuarial sciences or literally in computing.

Because when you're building giant computing systems like ChatGPT, you have to create this huge infrastructure of computers, and they often fail. And you have to have some idea for how often that's going to happen and how much the cost is to replace these parts, to figure out economically what you're sinking into the system.

But we can do the same exercise we did the same way. I'll do it in a slightly more clever way. So we have this random variable T. That is the time that this computer fails. So we have some probability P that it happens on the first day. And then we have some probability 1 minus P times P that it happens on the second day. Let's just do out one more, 1 minus P squared times P, then it happens on the third day.

But there's a clever trick for doing this, which is that this is equal to P from this first term times 1 minus P times 1 plus the expectation of T. So this looks a bit magical. Can anyone give me some intuition for what exactly I just did in this expression? Why is this true? Yeah.

That's right, that's right. So first I got my P from here. I pulled out this 1 minus P, because everything's got a 1 minus P in front of it. But then the important thing is that when I don't fail on the first day, the entire process starts over again. It has no memory of what happened in the past. I've gone one extra day. But then whatever happens to that same identical random variable that I care about is going to be what happens to the process thereon out.

So in fact, this makes it much easier. Instead of messing around with these infinite sums, I can just rearrange things algebraically. And what I'll get is that the expectation of T is equal to 1 over P. Any questions? Make sense? All right. So the mean time to failure just depends on this failure probability in this very simple way.

So let's get back to expectation. And then I'm going to tell you something unlike the Monty Hall problem that you're definitely going to find counterintuitive. So we talked about random variables. We talked about events. Actually, there's a way to go back and forth between the two.

So it turns out that we can actually treat events as random variables. This will be a very important trick that we're going to use all the time. So this is called the indicator function. So in particular, if you give me an event A, well, I can define this function that has this funny notation. Usually you write it as a bold faced 1 with an A subscript. But this is just a random variable. It takes in some outcome from the sample space, and it spits out a real value. It spits out the value 1 if your outcome belongs to the event, fine. And it outputs a 0 otherwise.

So in particular, if you give me an event, this describes an associated random variable. This is just a recipe for it. My random variable takes on real values. Actually, it takes on values just 1 or 0. It's 1 if the event is on, it's 0 otherwise.

And now, the good thing is that a lot of what we do with random variables, we can use to reason about events. So let's connect this back to events. So I can look at, for example, the expectation of my indicator function over a random outcome. So let's just be clear, because we've introduced a lot of notation already today for probability.

Remember that when I write this bold E, I'm looking at that experiment where I throw my dart and I get some outcome in my space. That tells me what x I have. And remember that an event is just whether or not x belongs to A. But what's going on here is that I have this function that just outputs 1 if x belongs to A and outputs 0 otherwise.

So in this case, we can write out what the definition of the expectation is. It's the sum over all outcomes in my set of the probability of the outcome times the value of the random variable. And this is the same thing as the sum over all x in A of the probability of x. Because the only way that this term survives is if x belongs to A. Otherwise, the term is 0, and it goes away. So this is really just bookkeeping for keeping track of which P of x is contribute to my sum. And this way, I change the subscript of my sum from being over all outcomes to only the outcomes in my set.

So in particular, the nice thing is that we get back the probability of the event. So you can see that we can use random variables as a placeholder for computing things like the probabilities of events just by looking at their expectation. So this is just a bunch of bookkeeping. Does this make sense? Any questions? All right.

So now one of the really powerful things about expectation is that it behaves really nicely in some settings. It behaves as nicely as you could want. So let me tell you a powerful fact. Some of you have probably already seen it before. But linearity of expectation.

So it turns out that when you add random variables, computing their expectation is easy. So in particular, I can look at a random variable f of x. I can look at a random variable g of x. I mean, it seems intuitive, but it's not true. Yeah?

STUDENT: If A and B both happen, then it gets double counted on the left side but not on the right side.

ANKUR MOITRA: That's right, that's right. So what's going on here is we can go back to our diagram. We have these two events A and B. And when I look at the expectation of 1A of x, well, I'm just looking at this dart and throwing it in here. And every time it lands inside A, I'm getting $1. Every time it lands outside, I'm getting no dollars. And I'm looking at the expected number of dollars I win. And that's the same thing as the probability that the dart actually lands in that set.

And now same thing is true for B. But when I add these two things up, what's going on is if I actually landed in the intersection of these two, I'd be getting $2. So really what's going on is that this is not equal, because this is no longer an indicator random variable. It doesn't just take the value 0 and 1. It also takes the value 2 sometimes. The only way this would be true is if A and B were actually disjoint. All right. Does that make sense? All right. Sounds good.

So I'll tell you one really interesting problem, and then we'll end there, and we'll go on to other topics next time. So you all have seen the Monty Hall problem before. That's no surprise. It turns out that conditional probability, conditional expectations can still be very counterintuitive in ways that even math professors can be totally wrong and have the wrong intuition.

So let me tell you about this amazing example that's due to my colleague upstairs. In fact, I saw him right before class. It's called Mossel's dice paradox. Elchanan Mossel is a very famous probabilist. And when he came up with this, his advisor told him, I know you've done a lot of great research and you've proved amazing things, but this will be what you're famous for. And he was very disappointed.

So here is the experiment. You can throw a dice until you get a 6. And we're going to be interested in the number of rolls that you have to take in expectation. Now, if that were it, there would be no paradox. We already talked about that for the mean time to failure. Because if my probability of rolling a 6 is 1/6, I can look at the expected number of rolls, and that would just be 6. Fine. So nothing is hard here.

But the experiment is you're going to throw a dice until you get a 6. But now we want to compute the conditional expectation. So conditioned on all the throws being even, what if I tell you that all the throws are even? What is the expected number of rolls?

So very simple problem. It takes a little bit to wrap your head around it. But the way that this experiment goes, I'm sitting here throwing the die until I get a 6. And you're not watching me do the experiment. But then before I ask you anything about the experiment, I tell you, actually, while I was doing it, I only got evens. It was the darndest thing. And now telling you that I only got even throws, I'm asking you how that changes your expected number of rolls.

So it turns out that there's a very natural way to try and answer this question. And almost everyone who sees this the first time gets it wrong the same way. It's amazing. So there's one argument that leads you to the wrong answer. So you'd claim, you'd think, that the answer is 3. Let me tell you why you might think that. So let's go through what we know.

So we know that all rolls are 2, 4, or 6, because you just promised me that all the roles were even. So those are the only possibilities. Nothing else shows up when I roll the die. And now, if you're rolling a die and it's only got three sides, 2, 4, and 6, then how long do you have to go until you end up with your first 6 and stop the experiment? Well, 6 is 1/3 of the possibilities. So you'd expect just by this argument from this mean time to failure, that maybe the expected number of rolls is 1 over 1/3. So namely 3.

And this is actually completely wrong. So in fact, Elchanan was teaching over at Berkeley when he came up with this example, and he was doing a review for some probability class. And he wanted to make this argument that probability is very counterintuitive, and you just have to compute the darn algebraic expression to do it.

And he came up with this random example. There is no special intuition behind it. He came up with it on the fly. And then he started doing the proof out, and he got the answer 3/2. And he was so convinced that somewhere in there was some algebraic mistake he'd made on the board. So the truth is, when you're speaking to a large audience, your IQ decreases by 20 points at least.

And he told them that, don't worry, I'll tell you what's wrong with my computation, and I'll email the class and I'll tell you what I got wrong so that you're ready for your exam. He went home and he checked it. And the darnedest thing. The answer is actually 3/2. It's a very counterintuitive problem. In fact, I remember, I'll tell you-- I guess this is videotaped. I'm not sure how much I should say.

But we have a mailing list among some of the three faculty, especially ones who teach some of the intro proof based math classes. And then when this thing came out, David Karger emailed me and told me, this would be a great example for 042. We should do this. And one faculty member who shall not be named was on the mailing list and did not believe the answer. It was the craziest thing. He was arguing with us back and forth on this thread I have. And he just was convinced that the answer was 3.

And actually, should I cut the video for this? So let me tell you the other part of the story. So then he went on this debate for maybe a month or two, not believing the answers we gave him, even about this very basic statement. And then he finally came into our theory faculty lunch, which I have to go to after this, and he said, aha I understand what was wrong. You were all explaining it wrong. Let me tell you my explanation. So he started explaining it in this very convoluted way.

And Peter was there with me at lunch. Peter was co-teaching. And Peter usually doesn't say that much in the faculty lunches. And he said, now I know why no one understands probability when you teach it, which is kind of amazing. So I'll scrub the rest of the names from the story. But we'll do some examples in class where we actually ask you to compute this, and you'll see some tricks for doing this. And it'll be a good exercise for some of your writing problems.

So we'll stop there a bit early, and next time we will continue with inclusion, exclusion and other applications in probability.