MITOCW | 18.200 Lecture 3 Version 2.mp4

[SQUEAKING]

[RUSTLING]

[CLICKING]

PROFESSOR: Welcome back OpenCourseWare. So we're talking about random variables and expectation. And in fact, one theme that I want to mention is that a bunch of our probability lectures will be about how to get tools that give us a finer grained control of how random variables behave. At the end of today's lecture, we're going to talk about variance. And later on, we're going to talk about things like tail bounds.

All right, so let me start off with the main motivating problem that we're going to spend the first part of class talking about. This is a challenging problem that usually, the way I would get students' attention was I would say that if you're going to do tech or finance interviews, this is the kind of thing that they ask you during the interviews. But the truth is that everyone knows that they ask this, so they no longer ask this anymore. But it's still a really fun problem to think about.

So imagine we have n students in class, and they each turn in their psets. We grade them. And then, instead of figuring out who's who, we just randomly give out psets to different students. Well, one of the probability questions we can ask, which will be an excuse to learn a very important tool, as we'll be interested in what's the probability that we get everything wrong. So no one gets their correct pset.

OK, so we're going to see a very important tool, called inclusion-exclusion, in order to answer this question. But let's build up towards it and get some intuition. So the first thing I want to do is I want to understand this probability of this very complicated thing by breaking it up into simpler events. So what I'm going to do is I'm going to define a bunch of helper events that will help me compute this probability. So I'm going to let A i denote the event that the i-th student gets their correct pset.

So you can think about an event a few different ways. An event is literally a subset of the sample space. But that's not how I'm defining it here, because the sample space is gigantic to write down. So instead, I'm really defining it in words as a predicate that's applied to the different outcomes. So this is one event I care about.

And then, the ultimate thing we care about is we want to understand the following probability. We want to look at the probability that A 1 does not happen, the first student does not get their correct pset back, the second student does not get their correct pset back, all the way down to the n-th student also does not get their correct pset back. Now, we can write this a different way, which will be more convenient for the way that we're going to think about this. This is the same thing as 1 minus the probability that the complement of this event happens.

And what is the complement of the event that no student gets their pset correct is that at least one student gets their correct pset back? So in particular, we take the union of all of these events, either the first student gets the correct pset, or the second student, or the n-th student. So this is really the thing we want to compute today. So let's think about this in some special cases, which I claim we can always go back to our friend, the Venn diagram, to think about what's happening in this space.

So let's write out a simple Venn diagram just in the case where n equals 2. So I have my sample space. I have my event A 1. I have my event A 2. And what I care about is the probability that I throw a dart and that it lands in the union of A 1 and A 2, which is just this shaded region right here.

And we can just use some basic set theory to write this, because the probability of A 1 union A 2, well, we can break it up into the area by each of these different pieces in the diagram. So we can write it out as the probability of A 1. We can add in the probability of A 2. Now, this is not a correct expression. So what's wrong with this expression? Why is the probability of A 1 union A 2 not equal to the sum of their probabilities? Yes.

AUDIENCE: The intersection is counted twice.

PROFESSOR: Perfect. The intersection is counted twice. So here, what I'm going to do is I'm going to do a trick, which is, I'm just going to subtract off what I've overcounted. So this is the baby version of what inclusion-exclusion is. I want to get a handle on the union of these events. And I'm going to first overcount and then I'm going to correct my count.

All right, so nothing too complicated happened here. And we can try the same type of strategy, but maybe in the n equals 3 case. So we can write out the probability that A 1 or A 2 or A 3 happen. And we can do the same type of trick, where we look at the sum of the probabilities.

We add up the probability of A 1, the probability of A 2, the probability of A 3. But now we have the same problem, which is anything that happens in the intersection of these events is overcounted. So what I could try and do is subtract off pairwise differences. I could look at the probability. I could subtract off the probability A 1 and A 2 both happen.

I could subtract off the probability that A 1 and A 3 both happen. And last but not least, I'll subtract off the probability that A 2 and A 3 happen. So this is still not correct. There's something wrong with this expression.

Let's think about it again in terms of the Venn diagram. I can write out these three sets, A 1, A 2, and A 3. And if you think about what's happening in the Venn diagram, I've still miscounted.

So when I look at the union of-- if I look at probability of A 1, the probability of A 2, the probability of A 3, all this region in their intersection is overcounted. And I'm trying to subtract off these pairwise intersections. So next, I subtract off this pairwise intersection, this pairwise intersection, and this pairwise intersection. But what's wrong with my formula now? Any ideas?

So let's keep track pictorially of how the algebra connects. So I care about the union of A 1, A 2, and A 3. And I looked at the probability of A 1, that's everything in this disk, I looked at the probability of A 2, that's everything in this disk, and A 3, that's everything in this disk. That definitely overcounted the region in between. So what I did was I subtracted off this petal right here. And then I subtracted off this petal right here. And then I subtracted off this petal right here. So there's a problem. Anyone see it? Yeah.

AUDIENCE: You subtracted the center.

PROFESSOR: Yes, I've completely subtracted the center. That's exactly right. So, in fact, what I do is I add back in the center, and I'll get an actual correct expression.

So I add in the probability of A 1 and A 2 and A 3. So that is a correct expression. These are two correct expressions for n equals 2 and n equals 3. And we can just use these expressions to figure out what the answer is.

So let's go back to our definition of what these events are. So what is the probability of A 1? What's the probability that, if I have two students and I shuffle them randomly of who gets whose pset, that the first student gets the correct pset back. It's 1/2. So this is 1/2. This is 1/2, too. Now, just to make sure that everyone's awake, what's the probability that A 1 and A 2 happen? You have to be a little bit careful. Yeah?

AUDIENCE: 1/4.

PROFESSOR: It's not 1/4. Do you want to take a shot at why it's not 1/4? So let's think about what this event means. The first student gets their pset back, and the second student gets their pset back. So if you think about this, we could also write this in terms of conditional probability, the same way we talked about Bayes' rule last time.

It's the same thing as the probability of A 1 times the probability of A 2 conditioned on A 1. See, once I tell you that A 1 has happened, which happens with 50/50 chance, then I now change my sample space. And I have to look at the conditional probability. So any thoughts on why it's not a 1/4? Not one? Yes.

AUDIENCE: Even 1/2.

PROFESSOR: I heard a 1/2 over here. So you want to-- yeah, perfect. OK, so that's exactly right. It's 1/2, because if we think about these in terms of their sample space, there are only four possibilities in this-- there are only two possibilities in the sample space. And once we've chosen that the first student gets their correct pset back, there's no choice but the second student had better get their correct pset back as well.

So one of the things that makes computing this particular probability so nasty is that if these things were independent, life would be very easy, but they're not. See, when they're independent, I can multiply all of their individual probabilities. And we're seeing right here in this example that that's 100% false. But it turns out that it'll be morally true, at least when n becomes larger. So let's do this computation right here just to make sure we're on the same page.

What's the probability of A 1? What's the probability of A 2? What's the probability of A 3? They're all 1/3. Now, let's be careful.

What's the probability that A 1 and A 2 happen? So how large is my sample space? There are six possibilities in terms of who gets whose pset. And if A 1 happens, the first student gets the correct pset, the second student gets their correct pset, what can we say? What's the probability of this?

AUDIENCE: So there's three options for the first choice if they want [INAUDIBLE].

PROFESSOR: Yes.

AUDIENCE: So the second one, A 2, will have two choices?

PROFESSOR: Yes.

AUDIENCE: [INAUDIBLE]

PROFESSOR: OK, but what's the probability here?

AUDIENCE: [INAUDIBLE]?

PROFESSOR: Yeah.

AUDIENCE: Would it just be 1/6?

PROFESSOR: 1/6, perfect. So you gave a totally correct way to think about it. Another way to think about it is that there are six possibilities in the sample space. And when A 1 and A 2 happen, there's only one outcome that satisfies that condition, that's where everyone gets the correct pset, because if the first student does and the second student does, third student has no choice.

They get it correctly too. So all of these other probabilities here are all 1/6. That's absolutely right. And that will give us that the total probability of the event is 2/3.

OK, so one thing to emphasize is that these expressions I've written down in the n equals 2 and the n equals 3 case, these are valid not just in this particular problem, but they're true for any collection of events. I've made no assumptions whatsoever. It's just that if I had different events, they would have different probabilities, and I would have to compute all of the terms in this expression. All right, any questions so far?

So I think you'll agree that we probably should stop at n equals 3, in terms of writing down one expression, and reasoning about the Venn diagram. These things just get bigger and more complicated. But it turns out that there's a closed form expression, without looking at the Venn diagram and reasoning about the individual slices and pieces of area, that's going to give us the right answer in general.

So what I'm going to do is I'm going to guess what that formula is just by trying to extend the pattern that I've seen for the n equals 2 and n equals 3 case. That guess will be correct. And we'll check. We'll use that formula to compute what the answer to our challenge problem is. But then we'll go back and prove that formula. So that formula is not that hard to prove once you know what it ought to look like.

So what I claim, and this is called the inclusion-exclusion formula, is that the probability of any union of events A 1, A 2, doesn't matter. It's not specific to our particular problem. We can write this out by over and then under and then overcounting. So we're going to go through a series of corrections the same way we did in the n equals 2 and the n equals 3 case. Our naive approximation is just going to be the sum over all of the events of the probability of the event.

We know that that's not correct in general because we've over counted. So what we're going to do is we're going to subtract off the pairwise differences. We're going to take every pair of events, the same way we did in the n equals 3 case, and look at the probability of any pair of events happening, A i and A j.

The next term in the expression will fix the fact that we've undercounted. We'll look over triples the sum over all i less than j less than k of the probability of their threewise intersection, and so on. And the last term in this expression is just minus 1 to the n plus 1 times the probability of all of the events.

So that's the inclusion-exclusion formula. It takes a little bit to wrap your head around it, but it's very natural. It's really just this approach of successively approximating what the probability of this is by over and then under and then over and undercounting until you hone in on what the correct probability is. And all of this holds in general for no assumptions.

All right, so now it's not so obvious why this is actually helpful for our problem, because I took a simple-to-state problem, and I wrote down this behemoth of an expression right here. But what I claim is that all of the terms that happen in this expression, in our particular problem, turn out to be very easy to compute using the intuition we developed in the n equals 2 and n equals 3 case. So I'm going to ask you guys for help. And you guys are going to help me fill in what this looks like for our particular problem.

So returning to our challenge problem, the one we started off with, well, in the general n case, what is the probability that event A i happens? Who can help me out?

AUDIENCE: 1 over n

PROFESSOR: 1 over n, perfect. 1 over n, because I can just imagine that I'm handing back that student's pset first and I selected randomly. So I have a 1 over n chance of getting them their correct pset. Now, we know, as we go to higher intersections, we have to be more careful, because that's where the dependence between these events kicks in so that we can't just naively multiply these probabilities.

So the probability that A i and A j happen is not 1 over n squared. It's very close to that. So who can help me out and tell me what the answer is for this? And then we should see the pattern, hopefully. Yeah?

AUDIENCE: 1 over n times n minus 1.

PROFESSOR: Perfect, that's right. 1 over n times 1 over n minus 1. That's exactly right. The way to think about it is exactly in terms of the way I wrote it down with the conditional probabilities. If both of these events happen, well, I need student i to get the correct pset, student j to get the correct pset.

So let's imagine the thought experiment that we hand back student i's pset first. They have 1 over n chance of getting the correct one. And now, when they get the correct one, we only have n minus 1 possibilities for the remaining pset.

We know that student j's pset is still in that set because i did not get his pset. And so we have a 1 over n minus 1 chance of continuing so that we actually get both of their psets back. And just continuing the pattern, so the probability that we have a triple, A i and A j and A k is just going to be 1 over n times, 1 over n minus 1 times 1 over n minus 2.

So now we're almost home free, because we can plug in all of these values that we've computed into our expression. But we're going to need a little bit of counting in order to actually simplify this expression, because one thing we need to know is, OK, all of these probability of A i intersect A j. We know what those values are. They're 1 over n times 1 over n minus 1. But how many of them are there?

So that leads us to a very important concept that we're going to talk a ton about, which is called a binomial coefficient. So in particular, how many events of the form A i intersect A j are there? Each one of these possible pairwise intersections shows up in our inclusion-exclusion formula.

And I'm going to tell you what the answer is. In fact, it's such an important quantity that it's going to come up all over the place. It has a special name. It's called the binomial coefficient. And it's defined as n times n minus 1 over 2. That's what we mean when we write this expression of n choose two.

What this expression means, so when we look at it, we say this in words, n choose 2, is because I have n possibilities for what i is because it can be anything from 1 to n. I have n possibilities for what j is because it can be anything from 1 to n. And what I'm doing when I look at all of these events is I'm choosing two things out of them, a value for i and a value for j.

But the key is making sure that you count this the right way, because a priori, I have n possibilities for the choice of i, and then I have n minus 1 possibilities for the choice of j, because it can't be the same as i. But the key is that I have this extra factor of 2, because I've overcounted. Because I only care about the unordered pair, i, j, I don't care about what order they have. So each pair of events happens two different ways, both for i being the first thing, j being the second and vice versa.

All right, so now we're in good shape. In fact, more generally, if we look at how many terms there are for the r way intersection, what we can do is we can define the n choose r binomial coefficient, which is just defined as n times n minus 1 all the way down to n minus r plus 1 over r factorial. And it's the same type of intuition. If I'm trying to choose r values out of n possibilities, I have n possibilities for the first thing, n minus 1 all the way down to n minus r plus 1. But I have to divide by something to account for the fact that I've overcounted in this expression.

So we'll talk much more about counting, actually, immediately after this lecture. That's what we'll be starting next week. But let's get down to business, and let's actually write out what our expression is. Then something beautiful is going to happen.

So we care about this probability in the union of all the A i's. And if we work out what this expression is, well, I have n times 1 over n. That just comes from adding up all these probabilities of the A i's. There are n of them, and they're each 1 over n.

And then I have my next term. My next term is minus-- I'll write it out as n times n minus 1 over 2. So there are n choose 2 terms there. We know that their probabilities are 1 over n times n minus 1.

And let me write out just the next term and we'll stop there. We also get n choose 3, so n times n minus 1 times n minus 2 over 6, namely 3 factorial. And this is all times 1 over n times n minus 1 times n minus 2. And this expression keeps going on the same way. I get a higher n choose r. I get more n's here. And the signs are alternating the same way they are in the inclusion-exclusion formula.

And now you can see, I get a huge amount of cancellation. And things simplify. And I'm going to get something very beautiful at the end of the day. So I'm going to get that this overall probability, which we computed through inclusion-exclusion, is 1 minus-- I'll write it as 1 over 2 factorial suggestively-- plus 1 over 3 factorial minus 1 over 4 factorial, and so on. And now magic happens. So does this look familiar from calculus? Hopefully. Anyone brave enough to guess what this infinite sum is? Yeah?

AUDIENCE: 1 over e.

PROFESSOR: Perfect, 1 over e. That's exactly right. It's 1 over e. So this very simple-to-state question, the e pops out. It's a very beautiful answer. It involves all kinds of things that are going to show up in more detail in class. We have this basic building block and probability, the inclusion-exclusion formula. We have the starting point of counting, which is going to be a whole unit. And then we've connected it back to infinite series. And we're going to do a lot with generating functions later.

I think it's also good to try and keep a mental model for what's going on here, because we did do a bunch of calculations. It turns out that there's actually a very simple heuristic. See, in the n equals 2 and n equals 3 case, remember, we had the subtlety when we were computing the intersections of events, that events were not independent. That's why that 1/2 was 1/2 and was not 1/4.

But it turns out there's actually a simple heuristic that is not correct, but it works, so caution here. This is not a correct argument, but it's more of a back-of-the-envelope calculation. So the A i's, I claim they're almost independent. So they're not independent.

We already saw that in the n equals 2 case. But what I claim is that as n gets larger and larger, then they do become independent. You can think about this intuitively. If I have an infinitely large class, the first person, if they get their pset back correctly or not, it really ought to not affect very much whether the second student does too, because I haven't really changed the denominator very much.

So it turns out that, really, the dominant terms in this inclusion-exclusion formula all happen for smaller intersections of events, because these probabilities are becoming astronomically tiny. So wherever the action is really happening in this formula, things almost act like they're independent. So if you just pretend that they are, then you actually do get the right answer.

So let's look at the probability that, not A 1 and not all the way up to A n, well, we can pretend-- and this is not literally true-- that this is the product of all of these probabilities. And each one of these things behaves like 1 minus 1 over n, because I'm very likely to not get my correct pset back. And I'm raising them to the n-th power. And this, in the limit, as n goes to infinity, really is 1 over E.

OK, so inclusion-exclusion is actually a way to make some of these intuitions precise when you don't literally have independence, but you have something very close to it. So any questions? So far, we've just used the inclusion-exclusion formula. And what I still owe you is I owe you a proof of the inclusion-exclusion formula. Everyone following? Yeah? OK.

All right, so let's prove this formula. It turns out that it's really not too hard to prove once you know what you're trying to prove. So let's take our inclusion-exclusion formula that we want to prove. And there are actually tons of different proofs of this fact. I'll give you a particularly simple one, which is we're just going to prove it by induction.

So what do I have to do when I prove it by induction? I have to prove the base case. Actually, we already proved the base case because we already proved it pictorially, using the Venn diagram. So we're good. We proved it, for example, in the n equals 2 case.

You can check it's true for n equals 1, 2. That's just a technicality. But now, what I want to do in the inductive step, I'm going to assume it's true for some value n. And what I want to prove-- I want to show that then it must be true for the next largest value, so for n plus 1.

So that'll be my strategy. And all we're going to do is we're going to have to write down this mess of an expression and simplify it. So let me make my life easier. I'm just going to fix my expression right here.

So let's write down the expression we want to show. We can change all these n's to n plus 1. So this is the sum from i equals 1 to n plus 1. These are all n plus 1 values, and so on.

And this now becomes n plus 2. And this becomes n plus 1. So this is the formula that we want to show. This is what the expression would be at a larger value of n.

And all I'm going to do is I'm going to collect terms in this expression so that I can appeal to what I know that the expression already holds for some smaller value of n. And then I'm going to see whatever is left over, what are the terms that are left over, and I'm going to have to interpret those. And we're going to connect it back to the n equals 2 case.

So let's collect terms in this expression and see what's left over OK, so first, I'm going to let star denote this expression right here. This is the expression for n plus 1. And I'm going to group this term square into different terms.

So let me start off with the terms, which I'm going to call triangle, that I'm just going to pull out my old inclusion-exclusion formula. So what if I look at the sum from i equals 1 to n of the probabilities of all these A i's? You can see that in my expression for n plus 1, I'm just removing the probability of A sub n plus 1. I'm just removing that term.

And I'll only keep the terms that don't involve the n plus 1, OK? And then I have some terms here for this pairs. But my pairs only involve things up to n. So I'm going to look at the sum over all i less than j. But I'm going to put an upper bound that the n plus 1 term does not show up in this expression. And I'll have, again, the probability of the intersection, the pairwise intersection. And I can just continue this process here.

So the terms that I've defined in triangle, they're not all the terms that are in square, but there are many of them. And now I just want to deal with what's left over. So I'll define one more thing. And then I can at least write down an expression. So let's look at almost all the terms that we've left off the table.

Let's start with the pairwise intersections. So we can look from i equals 1 to n of the probability that A i happens and A n plus 1 happens, because all I'm doing is I'm looking at this expression right here. By putting an upper bound that I'm only considering things up to n, I've left off a whole bunch of terms.

Let me look at all the pairs where i is something and j is n plus 1. Those are definitely terms I've left out. So I'm going to group them together in circle. And the same way I can look at the next term in circle, which is the sum of i less than j all between 1 and n of the probability of the threewise intersection, A i and A j and A n plus 1.

All right, so this is just a whole bunch of bookkeeping. And let me write down the expression that I claim connects all of these different groupings of terms. At the end of the day, what I care about is this expression I have from inclusion-exclusion for n plus 1. That's square.

And I claim that this is equal to all the terms that don't involve n plus 1 plus the probability of A n plus 1, that's not something that showed up in my circle expression, and then, minus circle. So you can see that I've actually flipped the signs here, because originally, when I had pairwise intersections, I had a minus sign. And here, I've put it as plus. But it'll be clear why I did this in a minute.

So is everyone OK with this expression? I claim that we are basically home once we understand this grouping of terms. So now's a good time to ask questions? Does this make sense? Give me a nod yes. OK.

All right, so let's put everything together now that we've done the painful work, which is writing down all these indices, and let's check that everything works out. So I need one more slightly clever thing. So what I'm going to do is I'm going to interpret each of these expressions. So first of all, triangle, I claim is equal to the probability of A 1 union A 2 all the way up to A n. Not A n plus 1, but A n.

Why is that true? Just to make sure we're paying attention. Why is that true? Because we assumed it by induction. This is literally what we assume by induction. Remember, our strategy is to assume it's true up to n. And what I did was I pulled out all the terms in my n plus 1 expression. And actually, we get the same structure. And so we already know this by induction, perfect, OK?

And now we need to be slightly clever, which is what I'm going to do is you can see that all of these events show up all over the place right here, these intersections between A i and A n plus 1. I'm going to call these new events. I'm going to define some helper events B i. So B i is defined as the event that A i happens and A n plus 1 happens.

Let's check our definition. I claim that this expression right here is b 1 intersect-- sorry, B i intersect B j. Why is that? Literally, the definition of B i is A i intersect A n plus 1.

So A i intersect A n plus 1, intersect A j intersect A n plus 1. What happens when I intersect with A n plus 1 twice? It's the same thing as doing it once. So this is literally the pairwise intersection.

And now, all of a sudden, we're in good shape, right? Because what I claim now is that circle has a very nice definition, which is that circle is the probability of B 1 union B 2 all the way up to union B n. That's literally the same structure. I have here. This is the same inclusion-exclusion formula had before. I just had to squint and figure out what are new events which this is expressing inclusion-exclusion on. So that's where the cleverness happened.

And now I can remember what these B i's are. So this is the same thing I claim as the probability that A 1 or A 2, all the way up to or A n happens, and A n plus 1, because each one of these things just takes one of the A i's and intersects it with A n plus 1. So this is really just De Morgan's laws in disguise.

So last piece of-- I won't even call this cleverness, right, which is, I'm going to call this event right here, C. The real event I care about is the union of all of the A i's from 1 to n plus 1. But let me look at C as just the union of from 1 to n. And let's write out what we know, and then we'll see that we've figured out our answer.

So writing out, putting together all of this expression, we have that square-- well, we broke it up into different terms. This right here, this triangle is just the probability of the event, C. That's how we defined what C was. It was just the first n events.

And then what do we have? We have the probability of A n plus 1. And then we subtract off the probability of C and A n plus 1. I just put everything we know together. So does this make sense?

Why am I done? Yeah? This is the base case, perfect. This is the n equals 2 case. So by the n equals 2 case, we know that this is exactly equal to the probability of C union A n plus 1. And we just have to remember what C is, because C is the union of all of these A 1 up to A n. And then, we tack on A n plus 1. And we've proven inclusion-exclusion. So that's the proof.

So it turns out that once you know what the expression is supposed to be, the rest of it is very natural, because this expression shows up in disguise in subparts of the expression we have. All right, so this was one of the longest proofs we've done. Are we good? Yeah?

All right, in fact, I think-- yeah one of the other proofs we're going to do, maybe, I will be doing in three or four lectures time is the Chernoff bound, which will be so complicated that we're going to have you diagram it in recitation. So that'll be fun. But you should already start to think about not just verifying these things line by line, but why did I present it the way that I did?

So I didn't literally have to show you the n equals 3 case. But we can see that that built up a lot of intuition for what the terms are doing, why they need to alternate in signs. And you can see some elements of at least how I thought about how to put these things together.

OK, so now I want to tell you one other brief aside, and then we'll move on to our second topic for today. There is one thing that I told you about last time, which was when we talked about random variables, I talked about their expectation. And there was one key property that I told you about expectations, which is you add up two random variables, f and g, and then, the expectation of their sum is the sum of their expectations.

So this was called linearity of expectation. I proved it for you. The proof was completely trivial. It was just rearranging things algebraically. But what I claim is that it's actually deeper than it sounds.

So right now, what we went through, all of this work computing was the probability that no student gets their correct pset back. But another thing I could do through random variables is I could also ask, what about the expected number of students who get their correct pset back? So I had to do all of this work. I had to tell you about binomial coefficients, prove the inclusion-exclusion formula to answer a very similar sounding problem, but I claim that I have enough space left on the board right here to compute what this expression is, because the way that we can think about this random variable, the number right, really is the sum of a bunch of indicator variables for different events. I claim that this random variable is just the sum from i equals 1 to n of this indicator function of A i, whether the i-th student gets their correct pset back.

Now, by linearity of expectation, this is the same thing as the sum from i equals 1 to n times the of the expectation of this indicator A i. So remember from last time that this indicator of an event is just the simple random variable that's 1 when the event happens and is 0 otherwise. So what is the expectation of the indicator of an event is just the probability that that event happens?

So each of these things we already computed as 1 over n. So I'm going to get the sum from i equals 1 to n of 1 over n. And that's the same thing as 1. So the expected number of students you-- the number of students you expect to get the correct pset back is 1.

So this is something important to keep track of is that it turns out that when you're dealing with expectations, you don't have to worry about all of the stuff about the dependency between the events, whether they're independent or not. And this is a really unique property of expectations. It won't be true when we get to other things like variance, but they'll still be helpful expressions we can do that can make our life easier for computing. OK, good?

Let's go on to the second part of today's lecture, where what I want to do is I want to dig a bit more into random variables. We've talked about the expectation of random variables. And what I want to do is I want to introduce concepts like the variance of random variables. We're going to use these things extensively.

So let me start off first with just a bit of a reminder about some of the cool things you can do with random variables. So last time, I told you this example of mean time to failure. That was one interesting illustration of the concept of expectation.

If something is failing independently, like a computer, every day, what's the expected number of days that we expect to go before we actually have to replace that part? So remember, just as a reminder, we had this concept of mean time to failure. And if we had the CPU fails independently with probability p each day, then one of the quantities we cared about was the expectation of T, which is the failure time for the computer, so how often do we have to replace it. And we wrote out this expression for it that we simplified.

And we showed that the answer was 1 over p, which makes sense, because if I have a coin that's fair, the expected number of flips until I get my first heads is two. And the smaller p is, the larger this expectation will be. So this was just an algebraic illustration of what we can do with the definition of expectation.

I want to try a harder problem that builds off of mean time to failure. So let's consider the following experiment. What I'm going to do is I'm going to ask students in a huge class, OK, I'm going to ask them their birthday. And I want to know how many until we get all 365 days as answers.

This is a leap year. So I guess we could ask a slightly different question this year. But we're going to assume that everyone's birthday is uniformly at random. And this is another very basic probability question. It sounds a bit like mean time to failure, but there's a twist on it, which is we don't care about one thing happening, like our first heads or the failure of a computer, we care about a coverage condition that we cover all of the different days.

So a lot of times, you'll hear this colloquially referred to as the "coupon collector problem." We could ask this not with 365 as the number of things we want to cover, but some other parameter n. And you could imagine that every weekend you get a coupon in the paper, and it's uniformly at random for which coupon it is. And what you'd like to know is how many papers do you have to look at until you get all of the possible coupons.

But already, we have some powerful tools for how to address this question. It's really not a straightforward question until you see the trick. You see, if you think about what I did in this example of computing the expected number of students who get their correct pset back, there was already something very powerful we had, which was that when I have a random variable that I care about, this random variable being the number of students who get the right answer, well, what I can do is I can decompose it into simpler building blocks that are easier to understand. So that's what I did there, because I broke up this random variable, which is the number right, into counting over all of these events just as a sum of indicator variables.

Here, we're going to have to do the same type of trick, but we're going to do it in a more clever way. So same type of trick in action. So we're going to write T, which is our time where we collect all of these coupons or get all of these birthdays as a sum of simpler random variables. So what I'm going to do is I'm going to write T is T 0 plus T 1 all the way up to T 364. And T i, just to be clear, is the number of steps between the i-th distinct and the i plus first.

So what's going on is we have a progress measure of how many distinct birthdays we've seen. It doesn't matter to me what particular days they were, because everyone's birthday is uniformly at random. But I'm going to care about, in this progress measure, how much time do I have in between taking a step?

So at the beginning, this might happen very quickly, because in the beginning, I've seen no birthdays. I ask a student, and I'm definitely going to cover a new day I haven't seen before. But towards the end, when I've covered all 364 out of 365 days, it's potentially going to take me a while in order to get that one remaining distinct birthday I haven't seen, to go from covering 364 to covering 365. So this is a big trick. Otherwise, it would not be at all obvious how to do this. Yes?

AUDIENCE: Are they like, the number of students [INAUDIBLE]?

PROFESSOR: Exactly. Sorry, number of steps meaning the number of students. So what this means is-- let's just break down what this expression means. So i is 0 here. So this would be the number of students I have to ask in order to get to the first distinct day. And that's clearly 1.

But then, as I go later, it's the number of students I-- I just got my i-th new birthday. I've covered i days. I ask another student, ask another student, I ask another student. How long do I go until I get a new birthday out of that set? Any other questions? This is a big trick. It's hard to know that this is what you're supposed to do without seeing this trick before, but it's a trick we're going to use all the time.

So now, by linearity of expectation, we know that the expectation of T is the sum from i equals 0 to 364 of the expectation of all of these T i's. This is literally just using linearity of expectation. We know T is the sum of all of these T i's. That's how many students I have to ask until I cover all of the birthdays. And what I claim is that this expectation of T i is very easy to compute.

So let's think about what happens with this expectation of T i. Let's draw a picture in case it's helpful. So we can think about all of the different birthdays. We have 365 of them. And we've seen i different birthdays so far. And we're asking how long until we see the next distinct birthday.

So we've seen all of these different birthdays in this region. Because each dart I'm throwing is uniformly at random, I can just pretend that all the birthdays I've seen are all at the beginning. Nothing's going to change. And what I'm asking is how long, when I throw a dart that randomly hits one of these segments, until it actually is outside the set I've seen so far.

So we know that this probability that any particular dart lands outside this set is just 365 minus i over 365. This makes intuitive sense. When i equals 0 and I haven't seen anything, I'm definitely going to get a new birthday. When I've seen all but one of the birthdays, I have a 1 over 365 chance.

And so, now we're in good shape, because from the mean time to failure argument, we know exactly what this expectation of T is. It's equal to the sum of 1 over these P i's So in the beginning, we just have 365 over 365. Then we have 365 over 364 all the way down to adding up 365 over 1, which is 365. So it turns out that the answer for this, you can just compute it numerically, it's about 6.47% times 365.

So I have to ask about 6 and 1/2 times as large a class as I would need in principle to cover all the distinct days. In fact, the same argument works even with 365 replaced by n. So more generally, for the coupon collector problem, you'll get n over n plus n over n minus 1 all the way down to adding n. That's the sum of i equals 1 to it's n times the sum of i equals 1 to n of 1 over n.

And it turns out that this is equal to n times the n-th harmonic number. It behaves like n log n. So this gives you a very good back-of-the-envelope calculation for a lot of randomized things that this coupon collector problem shows up all the time. So basically, the 6.47 just comes from taking the natural log of the appropriate domain size. OK, any questions? Good?

All right, so we've done a whole bunch of examples today. So we did the pset problem. We did this coupon collector problem. These are really important examples to build your probabilistic intuition. I want to tell you a little bit more about some powerful general tools. And then we'll call it a day.

So we're going to do a little bit more theory, the same way that last time I told you about random variables and their expectation, I want to tell you about the variance and what happens when I take the expectation of products of random variables. And in fact, the first thing I'm going to do is, the same way that I defined the notion of two events being independent, I'm going to tell you what it means for two random variables to be independent. And then we'll see , when two random variables are independent, what kinds of computations become easier using that.

So here's our key definition that we're going to explore for the rest of what's left in lecture. So remember, we have this notion of a random variable, which is just a function from the outcome space to a real value. And now, imagine I give you two random variables, f and g. Well, we're going to say that they're independent if the following events are independent. So let's look at all the x's where f of x equals alpha. And let's look at all of the x's where g of x equals some beta.

All right, so a bunch of things are happening here. So I have a random variable, which remember, takes an outcome and maps it to a real value. So f and g are my two functions that map outcomes to real values. What I can do is I can use a random variable to define an event.

So now I'm going the other direction. Before, what I told you about, was how to take an event and define a random variable. That took us to the notion of an indicator variable. And we used that in that computation. Now I'm going the other direction, which is I'm saying, imagine I give you a random variable and it takes on a whole bunch of discrete values.

And what I'll do is I'll chop up the domain, the sample space, into the parts that correspond to different values of that random variable. So I'm going the other direction. I'm taking a random variable and defining an event, because at the end of the day, this is just a subset of the outcome space. So this is an event that's defined based on the choice of alpha. I get a different event for different choices of alpha.

This is an event that's defined based on the choice of beta and g. And what I want is that I want these two events are independent, which is just the same definition of independence we had before. I want the probability of the intersection of those two events to be equal to the product of their probabilities.

That's literally how we define independence of events. And what it means for random variables to be independent is this more stringent thing, which is I want that for any choice of alpha and beta that induces some choice of an event, one for one random variable, one for the other, I want every choice of alpha and beta to define two events that are independent. Does this make sense?

So this is a more complicated definition than we had before because we have one quantifier here. So checking independence for events is very easy. You just compute a bunch of probabilities and check that the equation holds. Now, computing independence for random variables is trickier, because in principle, you have to worry about every possible value of alpha and beta and compute all of those probabilities, OK? Nevertheless, it's a very powerful notion.

One thing you can do is let's just do a simple consistency check. Sorry for going back and forth and back and forth. But we can look at the indicator of two different events, A and B. Those are random variables.

What does it mean for those functions to be independent? I claim that those functions, those random variables, are independent if and only if their events are. So this just means that our definition makes sense, because we're thinking of random variables as a generalization of events because I can have not just 0 1 values, but I can have any real values.

But my definition better at least recover what my definition was in the case of events. And that's true. You can check that if you create two random variables based on the indicator functions of two different events, that they will meet this condition if and only if the corresponding events meet that condition too, because you really only have to check alpha and beta being 0 and 1. And you can check that everything works out. So that's good. So we have a reasonable notion of independence.

What I want to work up today is some idea of why this is powerful. And I want to compare and contrast this to linearity of expectation that we've already used to great effect. So here's the key lemma, which we'll prove. The proof is pretty boring, but it's a very powerful lemma.

The proof is really just about massaging the definitions the same way we had for linearity of expectation. What I claim is that if we have two random variables, f and g, so let's say f and g are independent random variables, then, what I claim is that the expectation of f of x times g of x is equal to the product of their expectation. So that's the lemma I want to prove. And I promise you the proof is very easy. But this is still a very powerful lemma.

So usually, when I'm trying to learn something new, I try and connect it back to something that I already understand. So let's think about the special case, again, where these random variables are indicators of events. This is an indicator of an event A. This is an indicator of an event B.

So what would the expectation of the indicator of event A look like? It would be the probability of the event A. What would the expectation of the indicator function of the other event look like? It would be the probability of the other event.

So this is the product of their probabilities. What is that supposed to be equal to? What happens if I take two indicator functions of two different events, A and B? What do I get? What is that? That's a whole bunch of English to parse. I take the indicator of two different events, A and B, and I multiply the indicator random variables. What is that? Can anyone describe that to me in English? Yeah.

AUDIENCE: [INAUDIBLE]

PROFESSOR: Yes, exactly. So the product of them is literally the indicator of the intersection of the events. And then, the expectation of that, as you said, is the probability that both of them happen. So this recovers something we already know, which literally, this is the definition of what independence is for events. So this is just a generalization, building a more general theory of what's going on.

All right, so let's prove this thing. Ah, that's tedious. Let's just write out the definitions and see what happens. So I've got the expectation of f of x and g of x. Well, whatever this mess is inside, that's a random variable. It takes in an outcome x and it spits out a real value that's given by the product of the values of f and g.

So I can take its expectation just by literally using the definition of the expectation. It's the sum over all x's in the sample space of p of x, the probability of that event, times f of x times g of x. I haven't done anything here. I've just literally used the definition. Whatever this monstrosity is, it's a random variable. And this is the definition of what its expectation is.

And now what we're going to do is, here's the one clever part, is we're going to break up the sum into different pieces in a way that corresponds you know clearly, we have to use this definition of independence, otherwise, the thing is false. And independence is fundamentally about breaking up the sample space into different pieces. So I'm going to break up that sum into different pieces mimicking the way that I've broken it up in this definition.

So literally, the way that I'm going to do that, is it's going to be the sum over alpha and beta-- and that should sound familiar-- and then the sum over all x that are in the sample space, but moreover, where f of x is equal to alpha and g of x is equal to beta. So I'm just breaking this sum up into different pieces, different contributions. And then I have the same thing. I have the probability of that outcome.

I have f of x. I have g of x. Except, because I've broken this event up, I know what f of x is and I know what g of x is. They're alpha and beta, respectively. So this is the same thing as alpha times beta times the probability of x.

So now we're in good shape. And we're just going to have to rearrange what the sum is. So let's massage the sum until it stares us in the face, how to appeal to independents. So this sum is the same thing as sum over alpha and beta times alpha-- of alpha times beta times the probability that f of x is equal to alpha and g of x is equal to beta.

So what have I done here? This alpha and beta, I can actually just pull it outside the sum. And when I pull this alpha and beta outside the sum, I'm left with the sum of all these probabilities of these atomic outcomes. And this is really just an event. It's the event that f of x equals alpha and g of x equals beta. So I get the probability of that event.

And now I can appeal to independence, because this is the same thing as the sum over alpha and beta, again, alpha times beta. But now I can break up these probabilities, f of x equals alpha times the probability that g of x equals beta. That's what it means for the two events that I've defined here to be independent is literally that the probability of their intersection is the product of their probabilities.

And so now I'm in good shape, because what I can do is I can turn this into two different sums, the sum over alpha and the sum over beta. And I can pull out terms that don't depend on beta. And I claim that this is the same thing as the sum over alpha of alpha times the probability that f of x equals alpha all times the sum over beta of beta times the probability that g of x equals beta.

So I really haven't done much other than figure out this way to break up the sample space so that I can appeal to my definition. But the rest of it is just algebra. So what is this expression right here? Any ideas?

This wasn't quite my original definition, but hopefully, you should see that it's the same. What property of the random variable f is that expression computing? Yeah. Expectation, that's right.

So my original definition for expectation was the sum over x's in the sample space. Here, I'm doing the same thing, but I'm collecting all the x's together that have f of x equal alpha and arranging it that way. So this is exactly just another definition for the expectation of f of x. And likewise, this right here is the expectation of g of x. Awesome, so we're done. Any questions?

OK, let's do a diagnostic question. This is definitely the way I like to think about these things. I just proved something for you. Learning about proof-based math, especially learning about how to write it, isn't just about verifying line by line what I do on the board. It's about trying to understand why I organized my proof that way, and also, where I use the different assumptions.

So the critical thing here for this lemma, which is different than linearity of expectation-- see, linearity of expectation, I used nothing. For this lemma, I use something. I use something very strong, which was that f and g are independent. And we talked about how it's a bit of a pain in the neck to actually verify independence, because you have to check it over all alpha and beta.

Did I really need that property of independence? So obviously, your intuition is that I did use it crucially. So can you think of an example where we have two random variables, f and g, but f and g are not independent, and where this lemma fails, where the expectation of the product is not equal to the product of their expectations?

So if you're able to push back all of the symbols and think about this intuitively, this would be very easy. But it's not always that easy. Is anyone bold enough to think of an example where I have two f and g's, two random variables that violate my assumptions of the lemma? They are not independent. And the conclusion of my lemma is false, too? Yeah.

AUDIENCE: Maybe some indicator variables of just [INAUDIBLE].

PROFESSOR: Yes, perfect. Let me raise you one, which is, let's just define f to be the indicator function of some event A, and let's let G be the indicator function of the complement of A. That would do the trick, right? So maybe these both have some nonzero probability of happening. The product of their probabilities is nonzero.

But what's the product of them? Well, that's the indicator function of both events happening. And no event and its complement can both happen. So we're done. So these are not independent. And the statement is false. So especially as we go to more and more general abstraction, it's always good to keep in some mental model that helps you navigate what's going on.

So we've only got a few minutes left, which is perfect because I want to introduce you to one last concept which is also powerful. This is the thing I've been promising you, which is the notion of variance. So that's our last topic for today. And I'll tell you one basic but very powerful property of variance that we're going to see at length in the sequel, when we get back to probability, when we talk about tail bounds. And we can already see some of the intuition.

So first, let me just define what the variance is. It's another property of a random variable. So the variance of a random variable is the following quantity. We write it as var of f. It's the expectation of f of x minus the expectation of f of x squared.

All right, so a bunch of things are happening here. So this, whatever it is, it's a scalar. And it's whatever the average value is for the random variable. I'm creating a new random variable that shifts the average down to 0. So the average of this new random variable is 0.

And when I square this thing, I'm measuring how spread out it is. So the variance, as the name suggests, is how much you expect the random variable to vary. It's a very coarse grained way to define how much a random variable varies. And we'll develop a lot of finer grained tools that will give us a better understanding of what's going on. But already, we'll be able to use this to great effect.

So that was my informal claim is that it measures how spread out A random variable is. Let me give you a basic fact, which is easy to show. So first of all, I claim that the variance of f is non-negative. That's true for any random variable. Can anyone argue why this is true so that I don't have to prove it? There should just be a one sentence. Yeah.

AUDIENCE: [INAUDIBLE]

PROFESSOR: That's right, that's right. Let's just unpack what this is. It's the expectation of a random variable. So what we do is we take a weighted average, according to the probabilities in the sample space, of the values that this random variable takes on. But the random variable I care about is this thing that's really-- it's the outcome of squaring the thing.

So it's always non-negative. The weighted average of a bunch of non-negative things had better be non-negative. In fact, let me raise you one. I claim that not only is that statement true and your explanation is perfect, I claim that when the variance of a function is 0, there's something very strong we can say about the random variable.

So what happens when the variance of a random variable is 0? Yes. It's constant, that's right. So f is constant value. So it doesn't matter what point you give me, what outcome in the sample space, there's always one fixed value and it only takes on that value. That's exactly right.

Now, there's one other alternative formula for the variance that's sometimes helpful. In some cases, one is a better starting point than the other. They're both equivalent. But it pays to just keep both in mind and use whichever one is easier. So let me give you another alternative expression for the variance. And then let me tell you a powerful property of the variance that kicks in when we have independence.

So an alternative-- completely alternative formula is that the variance of f is equal to the expectation of f of x squared minus the expectation of f of x, but with the square on the outside. And this happens because, from our first expression, when we expand out the square, we'll get that the variance of f is equal to the expectation of f of x quantity squared minus 2 times the value of the function times the expectation of the function, and then, plus the expectation of f of x squared.

And now we can appeal to linearity of expectation. And life becomes good. So we have the expectation of f of x with the square on the inside. Well, this part right here is not a random variable. That's a scalar.

So it really just pulls outside of the expectation in a nice way. And we get another E of f of x. So we get minus 2 copies of the expectation of f of x squared with the square on the outside. We add one more copy back in. And in total, we're just subtracting 1 copy with the square on the outside. So this just uses the properties of the expectation and linearity of expectation. But this gives us our two different definitions.

And now we can say the last property that we're going to show for today. And I'll tell you why this property is so important. So I claim that if we have two functions, f and g, that are independent, then the variance of their sum is equal to the sum of their variances. So this is definitely not true in general. This looks a bit like linearity of expectation. But it's really very different, because linearity of expectation held always. This only holds when f and g are independent.

So you can prove this. Let's just do a quick proof, and then I'll tell you about some cool applications. So by definition-- well, the variance of f of x plus g of x-- well, we can write it out from our first definition. We add up f of x plus g of x. We subtract off their expectation. And then we're going to regroup terms and we'll see what happens.

So we subtract off the expectation of their sum, f of x plus g of x. And we take the square of this entire thing. That's literally the definition. And now what we can do is we can appeal to linearity of expectation. And we'll get that this entire thing is equal to-- well, let's group together terms.

It'll be the expectation of f of x minus the expectation of f of x plus g of x minus its expectation of g of x. And this entire thing is squared. So all I'm doing is, remember, the square is happening outside all this mess. I'm pushing the expectation through. So this is the expectation of f of x plus the expectation of g of x. I'm pulling one term with f and one term with g.

But now what I can do is I can expand the square. And if I expand the square, I'm going to get this quantity squared. That's literally the definition of the variance of f. I'm going to get this quantity squared. That's the variance of g. And I'm going to get cross terms.

But now, the key thing is that cross term is 0 because of independence. I have the expectation of this times this. Because these things are independent, it's the product of their expectations. And each individual thing is 0. So that's really the last part of the argument. I can write that out a bit more precisely.

So the key thing is really what happens with the cross term. So when we get this cross term that looks like 2 times the expectation of f of x minus E of f of x times g of x minus the expectation of g of x, then what's going to happen is by independence, we can write this out as the product of their expectations. And this is all times the same thing, the expectation of g of x minus its own expectation. And now we're in good shape, because literally, by linearity of expectation, this is 0.

It's important to not lose track of what these different things mean. This term right here is just the scalar shift. And what's going on is I'm just shifting so that the mean of my random variable is 0. And the fact that they're independent allows me to write it as a product. So that's exactly why I get 0. So we can put together all these pieces, and we have this beautiful statement that the sum of the variances for two independent random variables adds when we add up the two random variables.

So I'll leave you with a closing example, which we're going to talk about in much more detail. There's a super-important example from this that's becoming very relevant these days, which is about polling. So we might have some particular events that we care about, which are things like the probability any particular person votes one way or the other. And we care about something like what the sum of those events are going to be because we want to forecast elections.

So it turns out that the way that you can do this, at least in principle, is if you pick a random person and ask them who they're going to vote for, well, that'll give you some estimate of what the underlying baseline probability is. And then you can do this not for one person, but for many people. So when you look at that estimator, it ends up being the sum of independent random variables. And the fact that that converges very quickly to the true average for the random variables, as we'll see, really comes out of this fact that the variance adds instead of increasing by a larger quantity.

So we'll talk a lot more about that in later lectures when we talk about things called concentration bounds. But I just want to give you a hint of where we're going with these tools. And we'll practice a bunch of explicit computations of variance to get a handle first before we go to tail bounds. So we'll stop there.