[SQUEAKING]

[RUSTLING]

[CLICKING]

ERIK DEMAINE: All right, welcome back to 6.1200. We continue probability. And not only that, we continue on expectation today. And we will also get to a second measure of random variables, which is variance. But first, most of today will be expectation.

Remember what we did last class? So for random variable x over a sample space s, so it's just a function from s to real numbers, we had three different formulas, at least, for expectation in different scenarios. The first one was the definition, sum over all outcomes, omega probability of the outcome times x evaluated at that outcome. Sorry, this is a omega.

Or we could rewrite that sum according to the distinct values of capital X. And so sum over the range of x, and take the probability that big X equals little x multiplied by little x. That's equivalent. And in the special case where x is a function to natural numbers, then we could write this as an infinite series.

Sum i equals 1 to infinity, probability of x greater than or equal to i. And that turned out to be equivalent to this thing by expanding it out. So if the CDF is nice, that's a nice formula, if it's in the special case over natural numbers.

For indicator random variables where x is 0 or 1, another special case, then the expectation is just equal to the probability equaling 1, which is like probability of the event that this is an indicator for. And the big thing was linearity of expectation, which told us, if we summed and also multiply it by a constant inside the expectation, we could push the expectation inside and/or pull the sum and the multiplication out. So we're going to use all of these things today and see how they contrast with other fun properties.

But I'm going to start out with some motivating examples to give us some interesting problems to think about. So let's talk about a bunch of events. E 1 up to E n, so these are subsets of our sample space. And we can think of those events being good things or bad things. Maybe these are different ways in which your computer could fail or your algorithm could fail. Or maybe they're good things. Maybe these are you'd like at least one of these to happen, or you'd like at least half of them to happen or whatever.

So we're going to think about, in particular, the number of events that happen. So let's define a random variable N to be the number of E i events that happen. And first problem is, what is the expected value of N. So what is the expected number of events in general?

There's a nice formula for this, which is sum i equals 1 to N of the probability of event i. OK, this is pretty much a warm-up. It's very similar to things that we did last time. In fact, we did a special case of this last time. But let me remind you of those techniques. And let's actually do it for this example.

So remember the technique was, if I have some random variable that is tricky to analyze and I want to compute its expectation, right, split it up into a sum of different random variables. So here, there's a pretty obvious way to split it up. N is supposed to count the number of E i's. And so each of these suggests we should count them individually.

Let's define an indicator, random variable I sub i for each event. And that is counting, among all the events that are known as E i-- there's only one of them-- how many of them occurred? 0 or 1. Did it occur or not? So then N is going to be just the sum, I equals 1 to n of I i. Aye aye, Captain.

Then we can compute the expectation using all these fun things. So we've got, the expected value of N is, by linearity, the sum of the expected values of the I i's. And because the I i's are indicator random variables, their expectation is exactly the probability that they're equal to 1. So this is sum I equals 1 to n probability.

I keep forgetting my x's next to my E's. I i equals 1. And that's just by, definition of the I i's, that is the probability of E i. OK, so very easy proof, but refreshing how we use all these fun facts.

We did a special case of this last time, which was we were counting, if I flip a coin n times and the probability of heads is p, so this is biased coins are heads probability p. Then the expected number of heads was equal to p times n. That's something we computed last time in exactly the same way. So this isn't terribly new. But this is a more general form. Here, all the events were the same probability. Now they could have different probabilities. Still, you sum them. Cool, we will use this again as well.

All right, so that's simple warm-up problem, just counting the number of events. Now we'd like to understand a little deeper how many events are going to occur in particular? A really interesting question is, when does at least one event occur.

So if you think of these, for example, as being different failure modes for your system, you want to know, what's the chance of anything failing out of all the different ways that things could fail. And those events may not be independent. So all of this works, even if the E i's are dependent on each other. Maybe they're all the same, who knows? But this summation is exactly the right answer for expectation by linearity.

So let's start with a very simple upper bound, called the Union bound, on the chance of something happening. So this is the probability of the union, hence the name Union bound, is at most the sum of the probabilities of the events. In other words, using the terminology from over here, the probability that at least one event occurs, that's the probability that N is at least 1, is at most the expectation of the number of events, because these two things are equal by this previous theorem here. So these two statements are equivalent because the number being at least 1, that's at least one event occurring, that's the same as the union of these events.

OK, cool. So this is like the Sum rule. But Sum rule had an equality and was only when the events were disjoined. This works all the time, but it can give a very weak upper bound. Depending on your events, this sum may be larger than 1. So that doesn't tell you very much. To say a probability is less than or equal to 1, it's like, yeah, I knew that. You didn't have to do any work.

Let me, in parallel, write some examples here. So let's do our same example of n independent coin flips, probability of heads is p. Then the Union bound says that the probability of at least one head is at most p times n, right? That's this expectation that I just wrote down over here. p times n was the expected value for the number of heads.

So in particular, the probability that at least one head is at most p times n. So as soon as p is bigger than 1 over n, this is not a useful upper bound, right? The true answer, actual probability there's at least one head, this is 1 minus 1 minus p to the n. Easier to compute the probability that there are 0 heads, that's the same as saying that there are n tails.

So this is the probability of a tails, 1 minus p. Raise it to the n-th power, we get the probability of n tails. We take 1 minus it, we get the probability that there are not n tails.

In other words, there's at least one head. So these are close for very small p, but not so great for-- even for small p, maybe not so great. I don't know. All right, so this is an example where the Union bound is not so helpful.

I'll come to other examples in a moment. Let me tell you a nice property on the other side, which I'll call Murphy's law. It's not the official name, but you've probably heard of some form of Murphy's law in the past. Whatever can go wrong will go wrong.

This is a formalization of that. So if the n events we have are mutually independent-- this is the first time I'm making this assumption other than that example-- then the probability of the union is at least 1 minus 1 over e to the sum of the probabilities of the events.

OK, this is without any n terminology, but maybe a nicer way is to write it in terms of n probability that at least one event occurs is at least 1 minus 1 over e to the expectation of n. Again, by this theorem, relating the expectation of n to the sum of the-- it is just the sum of those probabilities.

So the Union bound gives an upper bound. We want to know the chance of some event occurring or some failure happening. This tells us it's at most something. And this tells us it's at least something.

But this theorem assumes mutual independence, so doesn't always apply. But in this example, it does apply. So we can apply Murphy's law here. And it gives us the probability of at least one head is-- wrong page. At least 1 minus 1 over e to the p n.

All right, again, the expectation is p n. So 1 minus 1 over e to the n. So the bounds tell us it's somewhere between this thing that's kind of close to 1, especially if this exponent is large. So the point is here, this denominator increases exponentially with expectation.

So that means this is going to be very small if expectation of n is large. And so that means this is very close to 1. And so when the expectation is large, the Union bound doesn't tell us anything useful. It says it's at most 1 or whatever.

But also, Murphy's law says it's actually pretty close to 1. So maybe the Union bound wasn't so bad after all. That's the idea.

I mean, in particular, we also know that it's less than or equal to 1. So those two facts combined are maybe OK when the expectation is large. So what can go wrong will go wrong with high probability, assuming mutual independence and assuming expectation is large. I guess that's the formal version of Murphy's law.

OK, cool. Let's do a particular example of this that I think is instructive, which is, suppose that each event occurs with probability 1 over n. So this is the same thing. But I'm just plugging in p equals 1 over n.

Then with the Union bound, I get the probability of the union-- I'll write it one more time-- is at most 1, exactly. So it's useless. Murphy's law tells me the probability is at least 1 minus 1 over e to the 1, which is about 0.63. So they're pretty close. They're within a factor of 2 of each other, 1 and 0.63.

And the right answer-- well, I guess is this thing, 1 minus 1 over n to the n. I'm not sure it has a nice formula. It depends on n, which one is more correct. But they're pretty close. So they're both pretty correct.

Here is maybe another example. So we did two examples at the end of last class with randomized phone returning at some terrible school where you get a randomly selected phone when you go home. And then we talked about a version with a turntable or a lazy Susan, where either everyone gets their phone back or no one does. And it was kind of weird, because in expectation, these were the same.

The expected number of phones that went back to the correct person was 1 independent of how many people and phones there were, as long as the number of phones equals the number of people, which is kind of funny. Expectation was always 1. So yeah, this is an example, if you recall, we wrote down, there was an indicator random variable-- I forget what we called it.

Let's say I sub i-- which was, does the i-th person get back their phone. And we computed the probability of that equaling 1 here and here. And in both cases, it was 1 over n. So these are actually both examples of this picture. OK, I should say, if mutually independent.

In this example, I didn't want to assume mutual independence. In all cases, this holds if the events are mutually independent, because here, the events are not necessarily mutually independent, right? In fact, they're not. In both of these examples, they're different levels of independence, I would say. But they're both special cases of this property.

The probability of each event is exactly 1 over n. The probability each person gets their phone back is 1 over n, both with random permutation of the phones and with random rotation of the phones. And here, I will give you the actual answers, the probability that any phone is returned correctly, that's this probability of the union of the different events.

In this case-- well, it's easier to compute the probability that no one gets it correctly. This is like the no repeated serial numbers in the dollar bill, kind of. So for the first person, there are n minus 1 choices out of n, where they get the wrong phone back. And then the second person, there's n minus 2 choices out of n minus 1 for them to not get their correct phone back, and so on.

And get my cancellation chalk. n minus 1's cancel, n minus 2's cancel. n minus 3's cancel, and so on. We're left with 1 minus 1 over n, OK, which is in between these two values. So that's good.

And it's actually very close to the Union bound, closer to the Union bound than the other one. So I would say, yeah, I don't know. These are pretty independent is my fancy analysis. Yeah, maybe they're independent if you look at it the right way. It depends how you define the problem exactly.

Over here, what's the probability that any phone gets returned correctly? This is actually a really easy one. I rotate the phones on a table. What's the chance that anyone gets their phone back correctly? Yeah?

AUDIENCE: 1 over n.

ERIK DEMAINE: 1 over n. There are just n possible outcomes here. And one of them, everyone gets their phone back. So capital N equals little n. And in the other situation, zero people get their phone back.

So this seems to violate this lower bound, right? We said it should be at least 0.63. So in order to do that, it means that they're very dependent. I mean, it's very hard for events to be more dependent than them being equal, in fact.

The first person gets their phone back if and only if the second person gets their phone back, if and only if the third person gets their phone back. So indeed, the events are equal to each other here, the same single outcome. And so that's why this is breaking.

OK, I wanted to prove Murphy's law. And then we'll come back to one more example. We are we at? Over here. Like, where does this exponential come from? It's an approximation. So it comes from one of these fun facts that we've used before.

So just like in this last example, instead of computing the probability that at least one thing happens, let's compute the probability that zero of the events happen, and then do 1 minus it. So I'm going to compute the probability complement of this event is that n equals 0, so no events occur. OK, in other words, the probability that E 1 does not occur and E 2 does not occur and E n does not occur.

Well, the E i's are independent, are mutually independent. And one of the things we mentioned and maybe proved in recitation, I forget, back when we were doing the independence lecture, is that if events are independent, then also their complements are independent. If E 1 and E 2 are independent from each other, then E 1 complement is independent from E 2 complement. And so from the product rule, this probability of intersections for n mutually independent events is just the product of the probabilities.

What else? So this probability is, of course, 1 minus the probability of E i. And now we're going to use this fun fact, which is, this is at most E to the minus probability of E i by this fact that 1 minus x is less than or equal to E to the minus x for x non-negative.

We used this a few lectures ago, I think. This was the Taylor series expansion of E to the minus x. You take the first two terms. It's a very good approximation when x is small, which is kind of an interesting case.

We're thinking of these probabilities individually as being small, but there might be lots of events. Then this would be a good approximation. But it's always true that it's at less than or equal to in all cases. So this is fine.

And so now we have a product of exponentials. And by properties of exponents, we can pull the exponential out. This is less than or equal to e to this sum minus sum i equals 1 to n of probability of E i. OK, which if you stare at it long enough, that's exactly what we got here. e to the minus is the same as 1 over e to the thing in this sum.

And because we are computing the complement event, we have to do a 1 minus out front. And instead of an upper bound, this is an upper bound on the failure, this is a lower bound on success if success is one of the events occurring. OK? Cool.

So that's where the exponentials come from. It's a way to-- products of 1 minus something is ugly. There's no nice way to simplify that. But if we use this approximation, then we get a product of exponentials and life is good.

So for fun, as we're winding down the class, I think it's time to enjoy a music video. So the application is we're going to think about, we have some probabilities of failures. And here, this is a really cool Rube Goldberg machine made by the rock band, OK Go. And there are about over a hundred different components in this Rube Goldberg machine. So what you want to think about is what's the probability of at least one of those components failing in a single uncut video?

[VIDEO PLAYBACK]

[OK GO, "THIS TOO SHALL PASS"]

- You can't keep letting it get you down

And you can't keep dragging dead weight around

If there ain't that much to lug around

You better run like hell when you hit the ground

When the morning comes

When the morning comes

You can't stop these kids from dancing

Why would you want to

Especially when you are already getting yours

'Cause if your mind don't move and your knees don't bend

Well don't go blaming the kids again

When the morning comes

When the morning comes

When the morning comes

When the morning comes

When the morning comes

When the morning comes

Let it go, this too shall pass

Let it go, this too shall pass

You know you can't keep letting it get you down

You can't keep letting it get you down

This too shall pass

You know you can't keep letting it get you down

You know you can't keep letting it get you down

This too shall pass

When the morning comes

You can't keep letting it get you down

When the morning comes

You can't keep letting it get you down

You can't keep letting it get you down

You can't keep letting it get you down

You can't keep letting it get you down

[CLATTERING]

When the morning comes.

[CLATTERING]

[CHEERING]

ERIK DEMAINE: So good.

[CHEERING]

[END PLAYBACK]

ERIK DEMAINE: So if you haven't watched all of OK Go's music videos, I highly recommend, instead of studying for the final, you should watch them. Amazing stuff. Actually, I've watched this video like 10 times in the last day, and I only just now realized how many references there are to other OK Go music videos. So if you watch them all, you'll get extra appreciation. Then watch this one again, and you'll see them all.

Cool, so I happen to be friends in particular with Damian, who's the lead singer. And I was chatting with him, like, how did you make this work? Because in many talks that he gives, like he has a Ted Talk that's linked in the lecture notes.

He said, OK, there are about 130 components in the setup. So if you imagine each fails with probability, say, 10%, reasonable. If you engineer something well, it should work like 90% of the time. I don't know if you've ever built a Rube Goldberg machine. Even that's difficult to engineer.

It's very easy to get more like 50%, 70% success rate. But let's say you can get 90% success rate. Then what's the probability that there is a failure? Number of fails is at least 1.

Well, it's the same thing we had before. 1 minus 1 minus p to the n. This is the actual. We won't use any approximation here. It turns out to be 99.99989%, which is exactly 1 in 1,000,000 chance that this video will come out good. So that's terrible.

And then, in his talks, he goes on, it's like, yeah, so, we've got to take chances and just explore and not worry about failing, is the main point of his talk. But then, he never explains like well how did you make the video? If there's a 1 in 1,000,000 chance that it comes out right. Did you do a million takes?

And I don't know how many takes they did. They didn't remember. But he said, the secret for it working was to do the math after you've made the video instead of before. Maybe a little unsatisfying.

But on reflection, I mean, you could work out some other probabilities, right? If you say, OK, there's only 1% of chance of failure, then the probability that the whole thing succeeds is-- or the probability of failure overall is 73%. Still not very good, but doable. Like, you could do four takes, and then one of them will be good in expectation.

But 99% reliability is challenging. It's only when you get to-- if you get to 99.9% success rate, then this starts to get impressive, only a 12% chance that something fails. But apparently, in hindsight, what worked well for this video is because the effects, the interactions get bigger and bigger towards the end of the video, because they just wanted to make them more and more impressive with bowling balls and giant things and huge interactions, huge interactions actually fail less often, because something like a bowling ball is less affected by air currents in the room and stuff like that. So the bowling balls were super reliable.

They were probably down here in the very, very, maybe several 9's effectiveness. Whereas, some of the very early effects, in particular, the dominoes which opened the video-- I'll just play that again because I love watching this video. They failed almost every single time. So this very first move, apparently dominoes are extremely unpredictable.

So doing this in a way-- they said they set up the dominoes roughly 6 billion times. I think that's an exaggeration, but that's I guess the other answer is to do 6 billion takes, and then you'll be set. And they got one take with everything in it, which is pretty amazing, despite the odds. All right, so that was just a fun diversion. But I love the opportunity to talk about Rube Goldberg machines and stare at them.

All right, let's do some more fun things about expectation products. So it turns out there's a product rule, just like for probability, but for expectation. It relies on the random variables being independent. And then it's exactly what you'd expect.

The expectation of the product is equal to product of the expectations, so similar to probability. In fact, this is really just a generalized version of the product rule for probability. Maybe I'll give a quick sketch of the proof. What's my time? Yeah, let's do it.

So this expectation of x product y, if I just write out the definition, we just take all the possible outcomes. We don't know anything about them. But x times y is a function. We evaluate it at that outcome.

This is the same thing as taking x of the outcome and multiplying by y of the outcome. That's the definition of this product function. And then we multiply-- yeah, we multiply by the probability of outcome, sum them up. That's the deal.

Just like with this formula, we can rearrange the terms in that sum to where, here it was by x value, but we could do it by x and y value. So we can rewrite this as a sum over x and a sum over y, in other words, a sum over pairs x comma y, of this thing.

But now there are a bunch of probabilities of outcomes that have the same x and y. So we're just going to write the probability that X equals little x and big Y equals little y, and multiply that by x times y. OK, that's rearranging terms with the same x and y value like we did last class. And then this probability is a product.

By the product rule for probabilities, x and y are independent means that these two events are independent for all x and y. And so we can rewrite this as probability of X equaling little x times probability of Y equaling little y. And now we have a double sum over some things involving x and some things involving y.

Sorry for the pun. I didn't mean to say "sum" all the time there. Things involving x and things involving y. And we know from way back in our summation formulas, we can separate out the summation over x from the summation over y, take their product, and if you do that, you get exactly this formula with x and with y.

And so I didn't write it here, but this is x in the range of x. This is y in the range of y. And so that's why this works. So it's really just using product rule for probabilities.

Cool, let's do an example with that. Suppose we've invented a new game, which is we roll two independent six-sided dice. And then the score is the product of those two dice. Not a typical game, usually some.

But let's say we take the product. What is the expectation of that product? Well, if they're independent, we know it's the product of the expectations. And expectation of a die roll, we covered last time. It's 3.5.

So this is 3.5 squared, which is 12.25. OK, weird number, but there it is. Whereas, if I take the expectation of D 1 times D 1, also known as D 1 squared, D 1 is not independent from D 1. In fact, it's the most dependent it could be. They're the same random variable.

I claim I do not get 12.25, even though each of these individually has an expectation of 3.5. Let's just compute that or get an idea how we would compute it. I think I just used the definition we have, or really the summing over x in range of x. So we have six possible outcomes for D 1, and we want the probability that D 1 equals little d times little d.

And that's the second formula for expectation. This is just 1/6 for pair of dice. I guess I didn't mention fair, but I meant to say, fair dice. Sorry, I didn't get this right. I'm missing a product. This is this thing. So sum over the different die rolls.

But now we have d squared here And I can never remember the formula for sum of squares. But we get 1/6 times that sum of squares. 1 squared plus 2 squared plus 3 squared up to 6 squared, which turns out to be 15.1666, so bigger. Kind of weird, but there you go.

All right, so this was product rule for two random variables. Of course, if you take the product of n random variables you can do the same thing. If there's product of n X i's here, you just get a product of n expectations of X i's So when things are independent, life is good.

What about division? We spend all this time adding and multiplying probabilities and expectations, can we divide them? The short answer is no. Never divide random variables, ever. I'll show you why. Don't divide.

OK, one reason is expectation of 1 over x. If we could do this, we could multiply it with something and end up dividing. Does not equal 1 over expectation of x, in general, or probably usually.

And here's one example where this is particularly striking. They're very different. Suppose you have x. This is almost an indicator random variable. But instead of 1 or 0, it's going to be 1 or minus 1. So let's say it's 1 with probability 1/2. And it's minus 1 with probability 1/2.

OK, what's the expectation? 0. It's just the average, because they're equally likely. So 1 over the expectation is not a very good number. It's like infinity or something. That's bad, 1 over 0.

Whereas, if I take expectation of 1 over x, well, 1 over 1 is just 1. 1 over minus 1 is also minus 1. Like, 1 over doesn't do anything to x. This expectation of 1 over x is expectation of x, which is 0.

So this thing is like infinity or just undefined. And this thing is 0. So they're very different from each other. And in some sense, you should just never do this. This is a reasonable quantity to think about. Don't think about this one. Although that seems backwards, given what we showed here.

Here's a real-world example. This is based on a paper in computer architecture, where they're trying to argue that one architecture-- I don't know if it was P 1 or P 2. I think this is a little divorced from the original example.

So there were two different processor architectures for solving three different possible workloads. These were benchmarks that they could evaluate them on. And processor 1 on workload 1 took 10 seconds to complete, processor 2 took 16 seconds, and so on. There's a table here.

You run each of these processors for each of these workloads, and you see what you get. Let's say these are each equally likely, 1/3 probability of occurring. And you want to know which one is better. And this is one of those things where you think, well, either I should take the expectation of the ratios or I should take the ratio of the expectations.

Let's compute the expectations individually first. So if I add these up, I get 30. There's three of them. So the expectation is 10. And here, it's also 10. OK, so in fact, the ratio of the expectations, which I think is a reasonable thing to do-- so let's just call this P 1 performance over expectation of P 2 performance is 1. So in fact, these are equally good in expectation.

OK, the intuition why you should do this is, if you assume that these workloads are equally likely, then yeah, you should be computing how long will your process take given a randomly chosen workload according to that distribution. So you should be summing all these up before you do any kind of ratio. But what the paper did was take the expectation of the ratio.

And with these particular carefully chosen numbers, but they roughly match those of the paper, you get a funny fact, which is if you take the expectation of P 1 over P 2-- this is a ton of arithmetic, so I will not do it-- we get 10 over 9. So it looks like these are running times, so smaller is better. That looks like P 2 is better than P 1. But if you take the expected ratio of P 2 over P 1, it's also 10 over 9.

So clearly, P 1 is better than P 2. Why is this happening? Because expectation of ratios is just something you should never do and is meaningless. Or taken a different way, if you want to lie with statistics, here is a great way to do it.

Take expectation of ratios, you could support any story you like. At least with these, if you're lucky and the numbers are close, then you get this funny behavior. This is what you should do if you don't want to lie and you want to compare performance.

Don't divide random variables, I think, is all you'll have to do with the dividing random variables. Cool, but multiplying is good as long as they're mutually independent. Any questions? Cool.

So the last topic for today is variance. So expectation was a way to take a whole function, a random variable, and reduce it down to one number. Now we're going to talk about how to take a whole function and reduce it down to two numbers, expectation and variance.

And I'll say a little bit why these are maybe the first two numbers you might care about. There are more, of course, you might care about the whole distribution. And in some sense, what we've been doing, what we did with these guys, the union bound of the Murphy's law, was trying to capture a little more than just the expectation, but just trying to understand how much is at the zero spot versus everything else.

And next class, on Tuesday, we'll talk about a more generalized form of that, where there's some cutoff, and we want to know how much probability is out beyond the cutoff. Those are called tails of a distribution. But in the rest of today, we're just going to look at one more number that measures some kind of variance. It tries to measure how good of an approximation is the expectation. That's I think a good way to think about it.

And I'm sure you've seen variance of some form before. But maybe motivating example is investment, a.k.a. gambling. Sorry for the dig. But let's consider an investment process X i, which is plus i with probability 1/2.

Going to be a generalized version of the one I just did. And minus i with probability-- sorry, not i, 1/2. With probability 1/2. OK, so X 1 is like I invest $1 and 50% chance I double my money, 50% chance I lose my money. X 100 is I put in $100, and 50% chance I double my money, 50% chance I lose all my money. Or I could put in a billions dollars.

Are these exactly the same investment strategies? I mean, you would probably say, when i is larger, this is a higher risk investment. If I put in a billion dollars, that's very scary. I could potentially lose a billion dollars. Of course, I could also win a billion dollars, so that's like high risk, high reward.

But I invest $1, yeah, I'd play that game. Why not? It's just for fun. OK, so if we look at the expectation, we learn nothing. Expectation of all of these is 0, just like it was in the i equals 1 case. Variance is going to try to capture something involving i.

OK, let me define variance. It's kind of a big formula, but not too big as you stare at it long enough. So many brackets though, alternating square and round. All right, so what we're going to do is take the deviation from the mean, X minus the expectation of X. So this is a random variable.

It's just like X. But we subtract the expectation of X. And by linearity of expectation, that means-- this may be worth writing down-- the expectation of X minus-- the expectation of this braced quantity here is 0, because linearity says, well, this is expectation of X minus the expectation of the expectation of X.

But the expectation of the expectation of X is just the expectation of X, because while X is a random variable, I mean, expectation of X is also a random variable, but it doesn't change, right? Expectation of X is a single number. You reroll your dice, the expectation doesn't change. Only the value of X can change because it's a function. But expectation of X is a constant thing, so its expectation is itself.

OK, so these cancel. OK, so in other words, I'm taking X, which has some expectation. But then I'm shifting it over. So it's now its expectation is 0. I'd like to center things on 0 additively. Then I'm going to square it, which is maybe a little weird. We'll talk about why in a second.

And then I take the expectation of that, because while this thing is constant, this thing is not. So if your expectation happens to be 0, this is just the expectation of X squared. Whereas, before we were taking the expectation of X, now we're taking expectation of X squared. And for X squared to be nice, we're going to first shift it so that its mean is 0. OK, that was a lot of words.

In this example, the expectation was already 0, and so the variance of X i is just the expectation of X i squared, which is-- get this right. Oh, yeah, fun thing about square is it treats positive and negative numbers the same. So what is X i squared? It's always i squared, no matter how the coin flip occurred.

If it was plus i, the square of that is i squared. If it was minus i, the square of that is i squared. So this is part of the magic of variance that we treat negative and positive numbers the same. And so the expectation of i squared, i squared doesn't depend on anything. So it's just i squared.

And for this reason, we also often define the square root of variance to be the standard deviation, which we usually write sigma sub x here is going to be square root of the variance of x. So you can use whichever of these you want. I mean, they're the same up to a square root or a square. But it's the second number, sigma sub X i here, is equal to i. So while expectation told us nothing about the risk of our investment, sigma told us exactly what the risk is in this very simple setup.

OK, let's go on to the next topic, which is why squared. I'll try to give you some intuition. OK, why squared? There are many possible answers to this. But probably the simplest thing is to try what if I didn't put this square here and I left the formula otherwise the same? What if I did expectation of X minus expectation of X? What would happen then? Yeah?

AUDIENCE: 0.

ERIK DEMAINE: 0. I wrote it right here. Expectation-- I was saying, oh, this shifted things over. So the expectation was 0. So if you remove the square, you just are going to get 0. This would be a useless number.

AUDIENCE: Right.

ERIK DEMAINE: And so we want to put something here, one possible answer. Now, there are different things you might try to do. One would be expectation of X minus expectation of X to the p for some number p. What we just did was 2.

Another thing you might try to do to fix it is absolute value. And just for fun we could also raise that to the p-th power. And so these are natural things to try, I would argue, especially if you know anything about norms or geometry and stuff. And it turns out, all of these are interesting. Question?

AUDIENCE: I have a question about your formula in the example for various [INAUDIBLE].

ERIK DEMAINE: Yeah.

AUDIENCE: Are you setting X i squared equal to i squared or are you setting the expectation of [INAUDIBLE].

ERIK DEMAINE: I wasn't setting anything. I was claiming that X i squared as a random variable is always i squared, because X i is either plus i or minus i. And in both cases, the square of that is i squared. So this is inside the expectation, yeah. And therefore, the expectation doesn't have anything to do. It's a constant, not much of a random variable.

OK, back to these guys. These have names. This is for p greater than 1, this is the p-th central moment. And this one, for p greater than or equal to 1, is the p-th probably central. This one's not talked about nearly as much, but I would call it the absolute moment. And they're all interesting.

For this one, p equals 2 is called variance. That's the one we're talking about. p equals 3, with some normalization, which I'm not going to do here, is called the skew of the distribution. p equals 4, it's got a really fancy name that's just showing off my English knowledge. It's called kurtosis. I assume people looked at other moments too. Although, usually, they're just called first moment, second moment, third moment, whatever.

These are all very interesting quantities. They capture different things about the distribution. Variance is how far from the mean are you on average, in some sense. Skew is how to the left or to the right is your distribution.

Kurtosis, I have no idea. I'm not a statistician. But they all have some meaning. And you can do the same thing with absolute moment. You can also do it without this normalization term. Then you'd remove the word "central."

These are all interesting, cool quantities. So there's no particular reason 2 is special. And in fact, if you use absolute values, you can even go down to 1 and you get a reasonably interesting thing. But often, we really like the version without absolute value because it's smooth, because it's infinitely differentiable, whereas, absolute value is not differentiable. So it's just uglier to work with analytically.

So historically, this is why we go here. I think this is also a very interesting measure. And you could use it, just we're not going to use it in this class. OK, they give you different things, because the square is exaggerating larger differences.

If you square something that's big, it has a bigger impact than if you square something that's small. So if you care about deviation from the mean and you really you care a little more about big deviations than small deviations, then raising this to some power is a meaningful thing. And that's why we do it.

There's some other reasons. In statistics world, if you know that X is normally distributed, which actually is kind of common, even in our world. If you take a sum of a bunch of Bernoulli trials-- this has some name, right? Now the Bernoulli distribution.

So Bernoulli trials are coin flips. You take a sum of them, you get a binomial distribution. I should know that. Then there's something called the central limit theorem that says that's very close to a normal distribution. Normal distribution is Gaussian bell curve.

And it turns out, if the expectation and the variance of your distribution and it's normal, then you know everything about that distribution. So because normal distributions are frequently occurring in nature, both in real life and in mathematical life, like when we sum a bunch of coin flips, this expectation and variance are often the first two things you should care about, maybe they're the only things you need to know if you want to understand your distribution.

And then, because I'm a geometer, I have to say because we live in a--

(SINGING) Because we're living in a Euclidian world--

In Euclidean world, you square things. And then you take the square root at the end. This is the norm. This is actually the distance between two points in size of S dimensions.

OK, I'm usually a two-dimensional geometer, maybe three dimensions. But if you're crazy, you can live all the way up in size of S dimensions, where you can represent a function as just like a big vector, very popular these days. Just you write down all the possible-- for every outcome, you write what X is.

That's a vector in S dimensions. And if you also write down for every possible outcome what expected value of X is, it's the same every time. It's mean, mean, mean, mean, mean, mean, mean, very mean.

And you take the distance between those two vectors, that's exactly variance. Or sorry, standard deviation. So if you live in Euclidean world, that's why 2 is natural. Of course, you can live in any p-norm and then you get all these other things.

All right, let me tell you some cool properties of invariance. Up here. Some of these I'll prove. Some of these I'll just state, and you'll prove them in recitation tomorrow. But we'll see some. We'll use them to analyze some examples.

OK, so the first one is translation in variance. So what I'd like to do is take some random variable X and add a constant to it. Now, when I add a constant to X, I change the expectation, right? It also shifts over by linearity.

With variance, it actually doesn't change. The variance of X plus c equals the variance of X. That's a nice property. Probably true of all of these definitions, all of the moments. But we'll just do it for variance. So let's prove it.

So as I said, the expectation, we need that in order to compute variance. Expectation is always what you should do first. By linearity, this is expectation of X plus expectation of c, but expectation of c is just c, because it doesn't change depending on the random event.

So variance of X plus c, we just plug in the definition here. So we want expectation of X minus the expectation of X. But now X is X plus c. So we want expectation of first term X is-- so I guess this is going to be for the square. Then the first X is actually X plus c.

And then we have minus the expectation of X plus c, and then that squared. And then we close the expectation. And this thing is what we just computed here.

So we have inside the parens here, we have X plus c minus expectation of X plus c in parentheses. And get my red chalk. The c's cancel. So we end up with X minus expectation of X squared and expected. And that's exactly the definition of variation of X, variance of X. So that's kind of nice.

OK, a couple more. One is, here is another way to compute variance. I can't say I have a lot of intuition for this one, but it works out nicely mathematically. Variance of X is the difference between the squared value expected and the expected squared. The square of the expectation versus the expectation of the square with this sign, it turns out to be right.

I guess this is related to whether you do things inside or outside the expectation. This is going to exaggerate larger values of X more. And this adds them all up first, and then squares. So it's less dramatic. And the difference in those dramas are exactly the variance, turns out. So we're not going to prove that. You'll prove that in recitation tomorrow. But let's do a simple example of it.

So suppose we have a biased coin flip where the probability of heads is p. And take an indicator random variable H. Or equivalently, this holds for every indicator random variable H. We know the expectation of H is just the probability of it equaling 1, which is p.

The variance, it's quite nice to compute with this formula, because again, with indicator random variables, they're always 0 or 1. And so when you square them, nothing happens. So this is, according to this formula, it's expectation of X squared of H squared minus expectation of H outside squared.

And H squared is equal to H, because if you take 0, you square it, you get 0. If you take 1, you square it, you get 1. So this thing is just expectation of H, which is p. And then we have minus expectation of H squared. Well, expectation of H, again, this is p. So this is p squared.

So p minus p squared, also known as p times 1 minus p, looks a little bit more natural. Its probability of heads times the probability of tails, turns out to be the variance in this case. If you look at standard deviation, it's the geometric mean between p and 1 minus p, which is kind of a nice number.

OK, next one. Do I want this example? And I might. Let's go over here.

So expectation had this nice linearity property. How about variance? Well, it has this nice non linearity property. It's not usually called non linearity of variance, but it's kind of linear-ish. So the nice part-- yeah, all right. Let's start with this part, the less nice part, the less linear part.

So while variance was translation invariant-- is that what I called it? Yeah. If I add something to X, it doesn't change the variance. If I multiply by something, it better change the variance. I mean, that's the whole point of the investment.

Say, if I multiply X i by a billion, that has higher risk. So variance should go up. And it's a little funny, because everything's squared. So it doesn't go up linearly. It goes up by c squared instead of c.

Of course, if you take standard deviation, it is linear in this special case. If you take a standard deviation of c times X, it's going to be c times standard deviation of X. OK, cool. But what if we add two random variables.

Here, unfortunately, we need independence. So if X and Y are independent random variables, then the variance of the sum is equal to the sum of the variations. OK, so it's not as nice as we had with expectation, where it just always holds, but still pretty good.

So if you sum things and multiply them by things, it gets a little weird. You've got to square all of these coefficients. But these are the analogs of the things we have in expectation. This is, again, something you will do in lecture. So I'm going to just do an example-- two related examples.

OK, so let's say we have n mutually independent coin flips. And I want H to be the number of heads. This is what we were calling n before, and maybe on this board, H.

We've often called this H. We've counted this before. We've done the expectation twice now. We did it in the last lecture and we did it at the beginning of this lecture. The expectation of H is p times n if the probability of heads is p.

OK, what about the variance? And the idea, just like when we computed expectation, we split up H as a sum of n H i's, where H i was an indicator random variable for whether the i-th flip was a heads. H i is indicator random variable for i-th flip being heads. And so, then we just took the sum here.

And so each of them had a probability p of occurring. So we got n of them got p times n. Let's do the same thing with variance because we're assuming these coin flips are independent. With the expectation, we didn't have to. But with variance, we have to.

So we can use this property. This is going to be equal to, by this kind of linearity of variance, sum of variance of H i, for i equals 1 to n. Cool, so then we just need the variance of each individual coin flip, which we just did, right.

Bias coin flip, probability p of heads, variance was p times 1 minus p. So we get n times p times 1 minus p. I find it a little hard to have intuition for this, but it's a number.

Here's another example. Last example will give you some intuition for why this is interesting. Back to gambling, I mean, investment. Let's say you have n dollars you want to invest. Back to my great investment strategy here. Let's X n, now.

I could invest n dollars in this thing, and I get a variance of n squared. I could win n dollars, lose n dollars. But now suppose there's not just one investment like this. Suppose there are a whole bunch of investments, like X i. Let's call them stocks, right?

You have k different stocks. And I can choose, I'll invest so many dollars in this stock, so many dollars in this stock, so many dollars in this stock. And let's just assume, like this great investment, set up, that each of them will either double your money or you lose all your money.

Does this make a difference? Is it any different to invest in k different stocks equally versus investing in one stock, everything. OK, I claim they're different. So let's do it in general. And then we can plug in different values of k.

So let's say we have k different stocks. And we're going to equally split our investment of n dollars across those stocks. And let's say that each stock S i is, let's just do plus 1 and minus 1, simple.

Of course, you could generalize. Each of these occurs with probability 1/2, just like our previous investments. But now each stock is, let's say they're independent, mutually independent. This is probably not a reasonable assumption. Do not invest in the stock market with this advice.

But if we assume that they're mutually independent, then what we get, if we want the variance of our overall outcome, which is sum i equals 1 to n of n over k times S i, this is how many dollars-- I assume the basics of investment. If you invest more dollars, then you're going to either get more back or lose more, just like this X i. But I pulled it outside.

So if we invest n over k dollars in this stock, then we end up with this many dollars in the end. We add it up over all stocks, OK, now we want to know the variance of this investment strategy, which I'm going to call risk. So probably risk is standard deviation. But if variance is lower, so will the standard deviation. So the variation of some, if they're mutually independent, is exactly the sum of the variations of the insides, which is n over k times S i.

That was the first property. And then, the second one, if we want to pull this outside, it gets squared. So we get sum i equals 1 to n of n over k squared times variance of S i. Variance of S i is 1. I think this was an early example we did. In general, the variance of this X i, we're interested in the case i equals 1 over here.

The variance was i squared. So that's just 1. So we get the sum of something that doesn't depend on i. So we just get n of these things. We get n cubed over k squared. This is our variance. And if you take the square root, you get the standard deviation.

And if you hold n fixed-- of course, the more dollars you put in here, the more risk you expect. So it turns out to grow, if you're doing standard deviation, you get n to the 1.5 risk. So don't invest a lot of money maybe. And we get divided by k.

The cool thing is, the bigger k is, the smaller your risk. So I conjecture this is why financial people say, diversify your portfolio, because the more things you invest in, if they're mutually independent, it reduces your risk. That's kind of a fun conclusion to variance.

Of course, if they're mutually dependent, which stock market tends to be somewhat, then you don't get as much of a benefit. But you probably get some benefit by diversifying. So it's a good thing to do when you grow up and invest your money. All right, that's it for today. One more lecture on probability, Tuesday