[SQUEAKING]

[RUSTLING]

[CLICKING]

SARA ELLISON: OK. So last time, we finished up with the example of the auction example, the sort of extended auction example, and that was a little bit of a sort of side trip into auction theory. But we'll go back now to probability and pick up where we left off. And we were talking about moments of distributions and about expectation in particular.

So what if instead of wanting to know a certain feature of a distribution of x, say, the expectation of x, we instead are interested in that feature, say, the expectation of some function of x. So for instance, we know what the distribution of x is. We really care about y, this other random variable y, which is equal to g of x.

And maybe we don't care about the entire distribution of y, we just care about some feature of it, like the expectation. How can we find the expectation of y? Well, we know one sort of surefire way to do it. We can figure out how y is distributed. We know how to do that, right?

So if we know how x is distributed, and we know y is a function of x, then we just use our knowledge of functions of random variables to figure that out. OK? And then we can compute the expectation of that new distribution no problem. But there might be an easier way.

In a lot of cases, this can be mathematically messy or difficult, and maybe we don't care about the whole distribution. Maybe we only care about the expectation, so we don't really need the PDF of y. All we care about is the expectation of y. And so in cases like that, it might be just easier to use this formula.

So what this tells us is that we can compute the expectation of y where y is a function of another random variable x in this fashion here. So this is the sort old way that we knew how to compute it, but it might be cumbersome or too much trouble.

But in fact, all we need to do is use this formula, OK? So we just integrate over the support of x, g of x, and then the PDF of x. OK?

So now let me do an example where we're going to do that exact calculation. OK. So this example is called the Saint Petersburg paradox. It is a classic example/paradox in probability theory been sort of taught to students of probability theory for centuries, probably.

And this example is first discussed by 18th century Swiss mathematician Nicolas Bernoulli and published in the Saint Petersburg Academy proceedings in 1738. That's where it got its name. OK? The name Bernoulli. Does that sound familiar? Yeah?

I can't remember if I defined a Bernoulli distribution in this class or not, but a Bernoulli distribution is just a special case of the binomial distribution. So remember, binomial distribution we can think of as n coin flips where the probability of a success on each coin flip is p and they're independent, where Bernoulli distribution is just one coin flip.

So it's just a binomial where n equals 1. OK? That's a Bernoulli distribution. OK. So here's the game. I'm going to propose a game to you, and I'm going to ask you how much you'd be willing to pay me to play this game. OK?

I flip a fair coin until it comes up heads. And if the number of flips necessary is x, I pay you 2 to the x dollars. OK? So how much would you be willing to pay me to play this game? Does anyone want to venture?

I'm not going to hold you to it, by the way. You're not committing to play this game for me. But how much do you think you might be willing to pay me to play the game? Yeah?

AUDIENCE: I would find the expectation of x.

SARA ELLISON: Oh, no, without doing any calculations. We'll do the calculations in a second. Yeah?

AUDIENCE: I would pay $1 because there's a 50/50 chance that I could double [INAUDIBLE].

SARA ELLISON: Yeah. There's also a small chance you could get $1,000, though, right? Or more than $1,000, you know? But OK, fine. You'd be willing to pay $1. Yeah. Any other?

AUDIENCE: [INAUDIBLE] numbers first?

SARA ELLISON: No, you're giving away the punch line. But you're absolutely right. So we'll come back to your answer in a second. To be honest, to be perfectly honest, if I said, how much would you be willing to pay me to play this game, you'd give me a number much less than infinity to play the game, right?

OK? So maybe I get someone who's willing to pay me $1. Maybe I get someone who's willing to pay me $5, something like that. OK? So let's actually compute what the expected winnings of this game are. OK? Oh, and by the way, what's this distribution? I can't remember.

So you saw it on a problem set, right? I can't remember whether the problem set or whether in any other time I've told you what the name of it is. It's called the geometric distribution, OK? So it's just a coin flip until you get 1 success. OK? And how many coin flips are there until you get to that success?

OK. So it makes sense-- I think, at least I'll make that claim initially. It makes sense that you should be willing to pay your expected winnings for this game. OK? So let's calculate the expected winnings and see what you guys are willing see if you'd be willing to pay me that to play this game.

So let x be the number of flips required, and note that x has this geometric distribution with probability 0.5. I can look up in some table of distributions what the expectation of x is, and it's equal to 2. I can also calculate it. The calculation that I would go through is a little fussy, so I won't bother doing it.

But but you you can calculate what this expectation is. OK? Or you could just look it up in a table of PDFs. OK? And we'll need that figure in a second. OK? And then we define y, a new random variable y, to be equal to the winnings in this game. OK?

And so that's just equal to 2 raised to the x power. OK? So we've got two random variables x. This one has a geometric distribution. And y-- we don't know exactly what that distribution of y is, but we can calculate its expectation with the formula I just put up.

OK? So in particular, this is the discrete analog of the formula we just had up a couple of slides ago. If we're interested in the expectation of y where y is a function of random variable x, and we have the PDF of the random variable x, and we don't have the PDF of the random variable y, we can just use this formula. OK? So that's what we'll do.

And so let's plug in. This is the PDF of x. It's just one half raised to the x power. Right? On your problem set, that's what you got when you did a problem similar to this. And then we multiply it by the function r of x, which is 2 to the x, and we sum up over all possible values of x.

So all possible values of x-- I mean, we have to sum-- this is an infinite sum because there is some tiny probability that we could just be flipping this coin forever before we get a head, right? And so we're adding this up from x equals 1 to infinity. And when we multiply this out, we just get that that's 1. We're adding up 1. Infinite number of times, we get infinity. OK?

So this is maybe kind of a surprising result. And certainly, if you buy my argument that you should be willing to pay your expected winnings to play this game, this result doesn't sound right at all because no one's willing to pay me anything close to an infinite amount of money to play this game.

OK. So that's the paradox. But is it really a paradox? And the answer is economists-- not to economists, OK? So economists know that people have diminishing marginal utility of money. This is equivalent to being risk averse. So this is another way of describing a utility function with risk aversion.

So economists know that people have a diminishing marginal utility of money. So in other words, their valuation of additional money decreases as the amount of money they have increases. So if I win $50,000 in a lottery, that's going to mean a lot more to me than it would to Michael Bloomberg. I mean, it's just a tiny, little rounding error in his wealth.

And so this is a well-known, well-documented facet of people's utility functions or valuation of money. They care less about an increment of money the more money they have. So what we really should have been doing is instead of calculating expected winnings for this game, we should have been calculating our valuation of the expected winnings.

So basically, what diminishing marginal utility of money implies is that your utility function, as a function of money-- the amount of money you have-- is going to be increasing, but at a decreasing rate. So that's what diminishing marginal utility is going to imply. So let's just come up with some arbitrary functional form that looks kind of like this. And that, we'll say, is our valuation of the winnings of this game. And so what I came up with was a log of y. And you could use many other different kind of functions that have this general shape.

So now, we have a third random variable, z. And z is going to be the valuation of the winnings. And just we're assuming it's log of y. And then if I plug in x here, it's just log of 2 to the x. And so now, let's calculate. Let's use the formula again and calculate the expectation of z and see what we get.

So expectation of z is just equal to the sum from x equals 1 to infinity of log of 2 to the x-- that's the function that we've defined as z-- times the PDF of x, which is 1/2 to the x still. And we can rewrite this as log of 2 times the sum of x to the-- times 1/2 to the x. And then if we just use this formula for an infinite series, we get that that is equal to 2 times log of 2. And that's a lot less than infinity.

So that's the Saint Petersburg Paradox. But just keep in mind, it's only a paradox unless a little bit of economics. And then it makes perfect sense. OK. Questions? No? OK. So now, back to expectation.

So we've seen we've seen the definition of expectation. And we have talked about how to compute expectations. We've talked about how to-- we just saw an example how to compute expectation of a function of a random variable. And it's going to be useful for us to list a whole bunch of properties of expectation that will make expectation-- computing expectations easier and, in particular, is going to make computing expectations of functions of certain random variables a lot easier.

So let's go through the list of expectations, properties of expectation. So the first one is just that-- and this may seem pretty obvious. The expectation of a constant as opposed to a random variable is just equal to that constant. And I mean, that seems not very useful. And it seems obvious. But in fact, we'll use this fact implicitly all the time. I'll remind you, oh, this thing is a constant. So it can come outside the expectation. Let's see. Or its expectation is equal to itself.

The second property is that, if we have a linear transformation of the random variable x-- so y is equal to ax plus b-- then the expectation of y is just equal to that same linear function of the expectation of x. Number 3-- suppose we have a bunch of random variables, x1 through xn, and y is equal to the sum of those random variables. Then the expectation of y is equal to the sum of the expectations. So this is also going to be super useful. We'll use it many times.

And I want to point out something very important about this property, which is that I haven't said anything about the x's. So in particular, I haven't said that the x's needed to be independent. And in fact, they don't need to be independent. So this is true with any x1 through xn.

The fourth property is kind of a combination of number 2 and number 3. And that's just saying, let's say you have a linear combination, an arbitrary linear combination of a bunch of x-- a bunch of x random variables. Then the expectation of that linear combination is the linear combination of the expectations. Number 5-- if x and y are independent, then we have that expectation of xy, the product of x and y, is equal to the product of the expectations.

So we've been talking about expectation. Expectation will be the most important moment of a distribution that we are going to be concerned with. But it's not the only moment of a distribution that we care about. So in addition to describing the location or center of a distribution, we often would like to describe how spread out it is. And there's a moment for that. And it's called variance.

So here is the definition of the variance of the random variable x. So it's just equal to the expectation of x minus mu quantity squared. And here, mu-- I'm just using mu to stand for the expectation of x. So we're creating this new random variable, which is just equal to the squared deviation between x and its expectation.

And then we're taking the expectation of that new random variable. And that's the variance. I'll give you some more examples that-- we'll deal with variance in a variety of ways, where I think you'll get a good intuitive sense of what it is.

A note about terminology and notation-- we often denote the variance of x with the Greek symbol sigma squared. And note also that variance is an expectation. So many of the properties of variance will follow from the fact that it's an expectation.

So just like we went through the properties of expectation, we'll go through a series of properties of variance as well. So the first one is that variance of any random variable is less than or equal to 0. So why is that? Just look at the definition of variance and see that the thing that's inside the expectation is always going to be non-zero. So you're taking the expectation of something that's never negative. Then it's not going to be negative either.

The variance of a, where a is a constant, is equal to 0l. Again, you just think back to the formula for variance. And we remember that expectation of a constant is equal to that constant. So then you just plug in that constant. And you get that the variance of a constant is going to be equal to 0.

So remember that if we took a random variable x and we transformed it linearly, then the expectation of that transformed random variable was the linear transformation of the expectation? Well, something different happens with the variance of that random variable.

So in particular, if we take a random variable and we transform it linearly, the variance is just multiplied by the square of the coefficient on x. And it's not affected at all by this additive constant, b. So what's going on here? Well, basically, the variance-- remember, think back to the interpretation I gave you for variance or the motivation I gave you for variance a second ago, which is that it measured how spread out a distribution is.

And so if you have a measure of how spread out a distribution is and you just shift the distribution by a constant, you would hope that that measure doesn't change. And it doesn't. But if you multiply that distribution, or multiply that random variable by a constant-- a, in this case-- then, in fact, the measure of how spread out it is does change because when you multiply a random variable by a constant and it spreads it out or shrinks it down. And in fact, for variance, that's the factor by which it spreads it out or shrinks it down. It's the square of the multiplicative constant. Does that make sense?

Oh, I guess I just said this. In other words, shift the distribution and its variance doesn't change. So shifting corresponds to adding a constant to it. You shift the distribution, its variance doesn't change. You shrink or spread out a distribution and its variance changes by the square of the multiplicative factor.

Property number 4-- so let's suppose we have a bunch of random variables, x. And we add them up to create a new random variable, y. The variance of y is equal to the sum of the variances of the x's. But that's only true if the x's are independent. So remember when we had a very similar property for expectation? We did not require independence there. We require independence here.

AUDIENCE: Why is that?

SARA ELLISON: Why is that? I'm not sure if I can give you a good-- I mean, I can show you mathematically why that's true. But I mean, I'm not sure I could give you a good intuition. Yeah. Actually, there probably is a geometric intuition I could give you. But I might have to think about it for a couple of minutes, so it would be coherent. OK.

And then property number 5 is, again, a combination of 3 and 4. We have an arbitrary linear transformation of the x's. And the variance of the resulting random variable is just going to be equal to the sum of the variances where each variance is multiplied by the square of the multiplicative factor in the linear combination. And again here, the x's have to be independent for this to be true.

And then finally, the sixth property is that the variance of x is equal to the expectation of x squared minus the expectation of x quantity squared. And this is pretty easy to prove if you just start from the definition of variance. It's not too difficult to show this. And this can be a pretty useful property when we're computing-- if you need to compute the variance of a random variable.

Sometimes, computing the expectation-- sorry, the expectation squared and the expectation of x-- the expectation of x squared and the expectation of x quantity squared can be an easier thing to do than computing the variance. So anyhow, this can be a useful formula for that reason.

AUDIENCE: [INAUDIBLE] calculating expectation of x squared we just use the previous formula [INAUDIBLE]?

SARA ELLISON: Yep, yep, yep. You can do it in a variety of ways. One is you can figure out what the distribution of x squared is and compute its expectation. Or we can use the formula that we saw at the beginning of the lecture. Yep. OK. Questions about properties of variance? OK.

And then I also want to introduce standard deviation. So standard deviation we're not going to use as much as variance. But sometimes, it's convenient for our measure of dispersion of a distribution to have the same units as the random variable itself. And so for this reason-- so variance, the units of variance are the units of the random variable squared. And here, what we do is we just define the standard deviation to be the square root of the variance. And that's going to have the same units as the random variable itself.

So in a lot of ways, standard deviation and variance are-- I mean, obviously, one's a function of the other. But they're essentially equivalent ways of measuring the dispersion of a distribution. And sometimes, it's easier to use standard deviation. And sometimes, it's easier to use variance. OK. So just so you know the definition.

OK. Now, we can do something similar. So at the very beginning of the lecture, I gave you the formula for how to calculate the expectation of a function of a random variable. And I said-- well, I said, sometimes that's a lot easier to do than actually figuring out what the PDF of that function the random variable is, and then using that to calculate the expectation.

Well, sometimes, we might want a shortcut for calculating variance as well. And we can basically just apply the results of expectation of a function of a random variable and the fact that the variance is, in fact, an expectation to get a formula for the variance of a function of a random variable.

So variance of y is just equal to the expectation of y squared minus the expectation of y quantity squared. And that's one of the properties I just told you. But let's suppose we don't want to calculate the expectation-- we don't want to calculate this distribution, this PDF.

And so what we want to do instead is just plug in the formula, r of x, and calculate the expectation of that instead. And so using the formula for expectation of a function of a random variable, we get that the variance is equal to this minus this. So this is just combining two-- combining the definition of variance and this property of variance I just showed you with the expectation of a function of a random variable, combining all those things together to get this formula.

So now, let me define another quantity that, especially as we move into linear regression and other kinds of models, other kinds of models that include multiple random variables. This is a concept that's going to be very useful for us. And the concept is conditional expectation.

So what is a conditional expectation? A conditional expectation is just simply the expectation of a conditional distribution. So remember a couple of lectures ago, we figured out we had a joint distribution and we talked about computing a conditional distribution? It was just taking a slice and then blowing it up, normalizing it so that it integrated to 1. So that's a conditional distribution.

And the notion of a conditional expectation is going to be useful going forward. And that is just the expectation of that conditional distribution. So in some sense, it isn't-- it's not quite a new concept. It's just the expectation of a distribution we knew already existed.

But if we think of the conditional distribution more broadly as a function of the conditioning variable, then we can also think of the conditional expectation as a function of this conditioning variable. And that's going to be a useful thing for us when we're building these linear models.

So this is the definition. It's just expectation of y given x is equal to the integral of y times the conditional distribution of y given x-- so nothing particularly new there. The thing that's new about this is how I want you to think about this quantity.

So note that expectation of y given x is a function of the random variable x. If I plug in a specific realization for x, it's no longer a function of x. If we leave it in this general form, then x-- it's a function of the random variable x. And what do we know about functions of random variables? They're also random variables. So x is a random variable. That means the expectation of y conditional on x is a random variable.

And then, as I said, if we just plug in a particular realization for x, then it's just a number. Then it's just an expectation, just a number. So this distinction might seem a little odd to you. What I'm going to do is I'm going to give you a couple of laws involving conditional expectation. And then we'll do an example. And I hope the example will not only make this distinction a little clearer, but also give you some idea of why this might be a useful concept.

So the first law that I'm going to tell you is the law of iterated expectations. So think now of the expectation of y conditional on x being a random variable because it's a function of random variables. And so we can talk about its expectation because it itself is a random variable. So the expectation of the expectation of y given x is equal to the expectation of y. That's the result here.

Does this seem mysterious? Yeah? Yeah, it is a little mysterious. The proof is actually pretty straightforward. But I don't think it's necessary for me to show it. But the proof is not the hard part. The hard part is wrapping your mind around the fact that we're treating this conditional expectation as a random variable. It's a function of a random variable. So it is a random variable. And so we can also take the expectation of it. So it's a little-- I don't know. It's a little subtle, I guess. But like I said, I'll do an example in a couple of minutes. And I hope that will help.

OK. And well, secondly, the definition of conditional variance follows from that of variance and conditional expectation. And then the second law, the law of total variance-- the second law I'm going to tell you is that the variance of the expectation of y given x is equal to the expectation of the variance of y given x-- or sorry, plus the expectation of the variance of y given x is equal to the variance of y, the unconditional variance of y.

So now, we have two laws that tell us how to compute unconditional expectation and variance of y when we just have conditional moments. Here are the two laws. So just keep those in the back of your mind. You don't have to memorize them. But keep those in the back of your mind as we go through this next example.

OK. So I have a former student. And he moved to New York City after graduating from MIT. And he started an innovation incubator. So suppose he's been doing this for a few years. In reality, I think he's only been doing it for about a year and a half. But suppose he's been doing it for a few years and has kept track of the number of patents produced every year in his incubator.

And he knows that the expectation of N-- here, N is the number of patents. The expectation of N is equal to 2. And the variance of N is also equal to 2. So he doesn't know the distribution of patents. He just knows what the expectation and the variance is. And then let's also suppose that each patent is a commercial success with probability 0.2. We'll assume independence across patents.

Now, suppose there are five patents this year coming out of his incubator. What is the probability that three are commercial successes? So how would we even think about approaching this problem? So all I told you is a couple of moments of the distribution with which the patents are generated.

And then I also told you this piece, which is going to be crucial. What's the probability that three out of the five patents produced this year are going to be commercial successes? Any guesses on how we might proceed?

AUDIENCE: [INAUDIBLE]

SARA ELLISON: So your answer was almost entirely correct. But it was more detailed than I was looking for. But that's fine. We'll get to that detail in a second. The insight that you had that let you come up with that calculation was that you figured out what the conditional distribution of successes conditional on N was. You might not have even realized that's what you were doing, but that's what you did.

So this was the important insight, that S conditional on N being equal to some little n is just equal to-- is binomial with parameters n and 0.2. Where did that come from? Let's go back. So each patent is a commercial success with probability 0.2. And we can assume independence.

So each patent is like a coin flip where the probability of a success is 0.2. We're assuming independence. What is that? That's a binomial distribution. So basically, just based on this, based on my verbal description, you had to make the leap, the-- I don't know-- intellectual leap that what I was saying is that successes conditional on number of patents has a binomial distribution with parameters n and 0.2.

So in this case, what I'm saying is the question is asking, what's the probability that there are three successes given that there are five patents? And that's just the probability that a binomial random variable, with parameters n and 0.2-- or 5 and 0.2, I guess-- is equal to 3. So the probability that S is equal to 3 given that N is equal to 5 is just equal to-- I'm just plugging into the binomial formula here. Make sense? Sort of? OK.

So now, we can answer. So now that we've had this insight, we've figured out what the conditional distribution of successes conditional on patents is. We've had that insight. And now, we can do a lot more with that. Oh, and by the way, that's equal to 5%. Suppose there are five patents this year. What's the expected number of commercial successes?

AUDIENCE: The same thing for every number of successes from [INAUDIBLE].

SARA ELLISON: Yes, that's right. So there are a couple of different ways you could do this. So you could actually just go through and figure out the probability for 0 successes, 1, success, 2 successes, and then use the expectation formula and compute it. That's not exactly how I did it. But that would work perfectly fine.

So what I did instead is I knew, off the top of my head, that the expectation of a binomial random variable was equal to np, where the parameters the two parameters are n and p. I happen to know that. Proving that is actually, again, sort of a fussy calculation. But you can-- actually, I'll just go ahead. But you can compute the expectation in general of a Bernoulli random variable, and then add it up n times.

So the expectation of one coin flip being a success is 0.2. That's the Bernoulli. And then since there are five of them, you just add that up five times using one of the properties of expectation that we saw before. So this gives you a different way to get the expectation.

And the third way you could get the expectation is you just know this is binomial and you look up the expectation for binomial in a book. That's a third way you can do it. And so you get np. And then, in this case, n is equal to 5 and p is equal to 0.2. You multiply them and you get 1. OK? Makes sense?

Now, what's the unconditional expected number of commercial successes? So all the calculations we've done so far are conditional on n being equal to 5. But we want to know-- next year, we don't know how many patents are going to be produced. And what we want to know is-- maybe my student's putting together his budget for the incubator. And he's trying to figure out how much to charge. And he gets a certain percentage of commercial successes and so forth.

So he has to be able to compute the unconditional expected number of commercial successes because he doesn't know what n next year is going to be. OK. Yes?

AUDIENCE: He knows the expected number of patents is [INAUDIBLE].

SARA ELLISON: Yes, exactly. That's exactly right. He will use that. But how does he use that?

AUDIENCE: [INAUDIBLE]?

SARA ELLISON: More or less, yeah.

AUDIENCE: [INAUDIBLE].

SARA ELLISON: Yes. But how did you guys come up with that?

AUDIENCE: The expected number of patents [INAUDIBLE].

SARA ELLISON: I think what you're trying to say is the law of iterated expectations. So yes, exactly. OK. So how do we use the law of iterated-- so now, does this now make the law of iterated expectations feel a little bit more-- make a little bit more sense? Because that's the calculation you did in your head without even realizing you were using the law of iterated expectations.

So let's use the law of iterated expectations. The unconditional expectation of S is equal to the expectation of the expectation of S conditional on N. N is a random variable here. We don't know what the realization is. We're treating it as a random variable. So that is just equal to-- well, we plug in the formula for the expectation of a binomial. It's just equal to Np. But here, I'm using capital N to emphasize the fact that this is still a random variable. Yep?

AUDIENCE: Does this theorem hold for joint distributions or distributions that has variablest that are so [INAUDIBLE]?

SARA ELLISON: Yeah. I mean, this whole-- yes. I mean, this holds for-- all you need is the conditional distribution, yeah. Yeah, it's general. You don't have to have random variables that are independent. Well, yeah, I mean, in this particular example, I'm using independence to get the binomial. But that's not relevant to the law of iterated expectations.

So here, I'm still using capital N because we're still thinking of capital N as a random variable. So the expectation of capital N times p is just equal to-- well, we saw some properties of expectation that tell us if we have a constant p and we're multiplying a random variable by that constant, well, the expectation of that new random variable is just-- we can bring the constant outside the expectation. What property was that? I don't know. Property 3 of expectations or something.

And so we do that. So p, we just plug in p is equal to 0.2. And that comes outside the expectation because it's just a constant. We're allowed to do that. And then we have expectation of n. We were told at the beginning of the problem that my student knows the expectation of n. He can plug that in. We're done. OK? Makes sense?

He'd also like to know something about how volatile his income is going to be next year. So he cares about the expectation. But he also cares maybe about the probability that he's going to get no income or probability he's going to get a ton of income. And the variance is going to tell us something about that. It tells us how spread out the distribution of his income for next year is going to be. And so how do we calculate the unconditional variance of number of commercial successes?

AUDIENCE: The law of total variance.

SARA ELLISON: The law of total variance, exactly. OK. There we go. So the variance of S is just equal to the variance of the expectation of S conditional on N plus the expectation of the variance of S conditional on N. Somehow, that sounded funny. But I think I said it correctly.

So now, we're going to do something similar to what we did before. Actually, in this case, I put in a little piece that you probably don't know off the top-- I knew it off the top of my head because I've been teaching probability for years. You probably don't know off the top of your head that the variance of a binomial random variable is N times p times 1 minus p.

So here, we plug in the expectation of the binomial random variable N times p, again keeping N as a capital to emphasize that it's a random variable. And then we plug in the variance of a binomial. So capital N times p times 1 minus p. And then we take the variance of that first thing, the expectation of the second thing. We use now properties of variance and expectation to pull things outside of the variance and the expectation.

So here, we pull out the 0.2. The p is equal to 0.2. We pull it out the front. But we've got to square it. Remember that? So we've got 0.2 squared times the variance of N plus-- and then we just bring out p times 1 minus p out of the expectation because it's just a constant. We can do that. So we get 0.2 times 1 minus 0.2 times expectation of N. And then you plug in what the unconditional variance of N and the unconditional expectation of N is into the formula. And you get 0.4. Does this seem like magic?

AUDIENCE: Could you say something about the intuition behind the law of total variance?

SARA ELLISON: Probably not. The intuition behind the law of total variance?

AUDIENCE: Why does it make sense that it has those two parts?

SARA ELLISON: No. I don't think I can. Does anyone have any intuition for the law of total variance? Maybe I'll ponder that. I mean, it's a very reasonable question. But it's not something I've ever formed an intuition about. So I just think, oh, it's useful. And I can plug in and use it. Yeah?

AUDIENCE: I think that the expectation of the variance itself is kind of like a mean kind of thing. And the variance of expectation is how much it's deviating from the mean. And if you add those two together--

SARA ELLISON: So you're saying it's some kind of a decomposition? Yeah. I mean, that's got to be the case. It's some specific kind of decomposition of the variance coming from different sources. Yeah, yeah, that's probably right. Yeah. And whether I can say anything more specific now, I'm not sure. OK?

So we've been talking. We've talked a fair amount about moments of single-- of univariate distributions. So we talked about expectation. And we talked about variance and mentioned standard deviation. But we often are interested in the relationship between random variables. That's what we do a lot of in econometrics. That's what happens in multivariate statistics, et cetera.

And we have an important moment of joint distributions to describe one aspect of the relationship between random variables. And that's covariance. And basically, covariance is a way to describe how closely associated two random variables are. When we get to properties, that will give you a little bit more information about how to interpret covariance.

But the way to think about covariance, I suppose, is just that if two random variables are independent-- so if two random variables are independent-- we'll see in a second-- their covariance is equal to 0. And if two random variables are very closely related, then they're going to have a high covariance.

So how do we define covariance? Well, we define it as the expectation of x minus mu sub x times y minus mu sub y-- so that function, the expectation of that function of random variables. And we often denote it with a sigma sub xy. I'm not sure-- I mean-- well, that's how we-- I should just say that's how we often denote it. You might think it might make sense to denote it sigma squared sub xy. But this is how we often denote it.

And then we also have a standardized version of this called correlation. So correlation-- we'll often use rho to denote correlation. So either rho of xy-- or sometimes, rho sub xy, we use that notation as well. And that's just equal to the covariance divided by the square root of the variance of x times the square root of the variance of y.

So we'll see some things about properties of correlation and its relationship with covariance in a second. Well, I guess, here's the first bit of that. This is just terminology. We say that the random variables x and y are positively correlated if rho is greater than 0. And we say they're negatively correlated if rho is less than 0. And we say that they're uncorrelated if rho is equal to 0.

And for some reason, we don't have similar terminology with covariance. But it actually doesn't matter. This is sort of equivalent to what-- we don't call something uncovariated or something like that. We just call it uncorrelated. OK. So let's go through some properties of covariance and correlation as well.

So first of all, you might have noticed from the definition of covariance that it was very similar to the definition of variance, but it involved two variables. And in fact, the covariance, if you plug in x and itself into the covariance formula, you just get the variance of x. You might also notice from the definition that you can switch the places of x and y and you're still going to get the same value.

So when we saw properties of variance, there was a property of variance at the very end that I said is sometimes useful for calculating variance and it's pretty easy to prove. Well, this is sort of the counterpart for covariance. So you can show-- it's not too difficult. You can show that the covariance of x and y is equal to the expectation of x times y minus the expectation of x times the expectation of y. So remember, the counterpart for that for variance was that the variance of x was equal to the expectation of x squared minus the expectation of x quantity squared.

If we have two random variables, x, and they're independent, that implies that their covariance is equal to 0. So if you have two random variables whose covariance is equal to 0, you that does not, in fact, imply that they're independent. Although, in most cases that we're going to see in this class, random variables with covariance 0 will be independent. But the implication definitely does not go the other way.

So this is a property having to do with linear transformations of random variables. So you take x and transform it, linearly transform it, and take y and linearly transform it using different constants. And the covariance of x and y gets multiplied by the coefficients on x and y. The additive constants, b and d, don't do anything to change the covariance.

So remember, when we saw properties of variance, we said that if x-- well, we actually saw this in a more general form for x1 up through xn. But if we only had two x's, we saw a property that said if x1 and x2 are independent, then the variance of the-- sorry, the variance of the sum of the x's is equal to the sum of the variances, but only if they're independent. Here, this gives us a formula for-- at least in the case of two random variables, for calculating the variance of the sum when they're not independent.

The seventh property I want to mention is that rho is always less than or equal to 1 in absolute value. So rho goes from negative 1 to positive 1. And actually, it can be very handy to have. So covariance can be greater than 1 or less than negative 1. And sometimes, it's handy to have this units-free moment that describes how closely associated two random variables are.

And in particular, if the absolute value of rho is equal to 1-- so it's either equal to 1 or negative 1-- then that implies that y is a linear transformation of x. And actually, the implication goes the other way as well. So basically, if you have two random variables and they're perfectly correlated, either with correlation coefficient 1 or negative 1, then you know that there is a linear relationship between those two random variables.

OK. So now, with all of these definitions, properties, tools, and things like that under our belt, I'm going to give you a little preview of regression. So still, we're going to talk about-- in the next couple of minutes, we're going to talk about linear regression. It's still going to be in the context of probability.

So we're not we're not actually talking about estimating the parameters of a linear regression yet. We'll get there. We're going to talk about linear regression in the context of probability. But I hope this is going to give you a little bit of a sense for what's coming up in this class, what we're going to be covering in the weeks to come.

So let's suppose we have two random variables, x and y. We're going to denote the expectation of x as mu sub x and the expectation of y is mu sub y. And likewise, we're going to denote the variances of x and y in this sort of standard way. And rho sub xy, the correlation, is just equal to the standard definition.

Well, I just told you that if the correlation was equal to 1-- actually, I didn't tell you this, but this is true. I told you something close to this. If the correlation is equal to 1, then y is equal to a plus bx with b positive. And if rho is equal to negative 1, then y is equal to a plus bx b negative. So I told you that there is a linear relationship if the absolute value of this is equal to 1. So this just gives you a little more information.

If rho is strictly less than 1 in absolute value, then we can't write y and x as a linear combination of each other. But what we can do is we can write y as a linear function of x plus another random variable. And we'll call that other random variable u.

So here, we've got a random variable u that we tacked on here. We'll talk about the interpretation of it in a second. But what can we say about it? Can we say anything about how u behaves, or what its properties are, or anything like that? Well, it depends on how we define alpha and beta.

So let's suppose that just falling out of the sky we're told that we should define beta as rho sub xy times sigma sub y over sigma sub x. You don't have to worry about where this comes from at this point. And let's say we were told that we should define alpha as mu sub y minus beta mu sub x.

Then u has the following properties. Sorry, u defined as y minus alpha-- oh, I think that's supposed to be minus beta. Yeah, so u is equal to y minus alpha minus beta x has the following properties. The expectation of u is equal to 0. And the covariance between x and u is also equal to 0.

So you don't have to take my word for it. You can show these two things pretty easily using properties of expectation variance and covariance that we've seen. And so maybe that will be on the problem set that I need to put together this afternoon. So you may see that soon, OK?

But basically, if we define alpha and beta this way, then u has those properties. So how do we think about x, y, u? How do we think about their relationship in a case like this? Yeah?

AUDIENCE: What is the covariance between x and the function of x?

SARA ELLISON: So the covariance between x and a function of x? It just depends on the function. There's nothing I can say generally about it. Yeah, OK. So then in the case that I have discussed here, the alpha and beta have a particular name. They're called the regression coefficients in a bivariate regression.

The way that we think about the relationship among these random variables is that we think of alpha plus beta x as the part of y. We're decomposing the variation that we see in y. And we think of alpha plus beta x as the part of y that's explained by x and u as the part that's unexplained by x.

How do we get this interpretation? Well, in particular, notice that we've chosen alpha and beta so that covariance of x and u is equal to 0. The x and u are-- they have covariance 0. It means that they're completely uncorrelated. They don't have-- yeah, I don't know how else to explain it exactly.

And so the part of the variation that we see in y that is explained by this linear function of x is the-- I don't know. Well, how do I say this? We're decomposing the variation we see in y into the part that's explained by this linear function of x and this uncorrelated part. And we typically think of u, in a regression context, as being the error term.

So anyhow, you don't have to understand or completely have a clear intuition of what's going on here. But this is one way to think about linear regression. And we'll see other ways later on in the semester. OK? Questions? Nope? OK.

So I have two other quick things to go through. And these are two inequalities involving probabilities and distributions of random variables that do come in handy from time to time. The first one is called the Markov inequality. So let's suppose you have a random variable x. And it's always non-negative.

Then for any t, any constant t that's positive, the probability that x is greater than or equal to that constant t is less than or equal to the expectation of x over t. So basically, how much probability is out in the right tail of this random variable, x, is bounded by the expectation-- some function of the expectation of x, which I guess makes sense, right?

I mean, the expectation is a function of the probability density in all parts of the support. And so if you have a lot of probability out in the right tail, then that's going to pull the expectation out. So it's not surprising that there is this relationship between the probability in the right tail and the expectation.

So let me draw just a couple of pictures. So what the Markov inequality tells us-- let's suppose we have a uniform distribution over on the left. And the uniform distribution-- let's see. The Markov inequality tells us the probability that the uniform distribution-- or a random variable with that uniform distribution is going to be greater than t is bounded-- that slice of the distribution is bounded above by the expectation of that distribution divided by t.

And the same thing is true for any shape of non-negative random variable. So do keep in mind that the Markov inequality is only for random variables that are always non-negative. But this is going to be true for any non-negative random variable for which expectation exists.

The second inequality is the Chebyshev inequality. So for Chebyshev, we need that x is a random variable whose variance exists. But it doesn't need to be a non-negative random variable. Then for any constant t greater than 0, we have that the probability that x minus the expectation absolute value is greater than or equal to t is less than or equal to the variance of x over t squared.

So basically, what the Chebyshev inequality is doing is it's putting bounds on both tails of the distribution. And I have pictures for that as well. So it's putting bounds on both tails of the distribution. And those bounds are a function of the variance.

So basically, if you have a random variable with an expectation, whose expectation and variance exists-- basically, if you look more than t away from the expectation in both tails and you add up that probability that's more than t away from the expectation in both tails, that probability is going to be bounded above by the variance of x over t squared.

AUDIENCE: Can you use two Markov inequalities to define an asymmetric Chebyshev?

SARA ELLISON: I believe that is true. So the Chebyshev inequality can be derived from the Markov inequality. So my guess is you could derive other flavors of the Chebyshev inequality as well, yeah. Yeah. OK, so that's it. We'll call it a day. And we'll start talking about the sample mean next time.