PROFESSOR: OK, good morning. So today, we're going to have a fairly packed lecture. We are going to conclude with chapter two, discrete random variables. And we will be talking mostly about multiple random variables. And this is also the last lecture as far as quiz one is concerned. So it's going to cover the material until today, and of course the next recitation and tutorial as well.

OK, so we're going to review quickly what we introduced at the end of last lecture, where we talked about the joint PMF of two random variables. We're going to talk about the case of more than two random variables as well. We're going to talk about the familiar concepts of conditioning and independence, but applied to random variables instead of events. We're going to look at the expectations once more, talk about a few properties that they have, and then solve a couple of problems and calculate a few things in somewhat clever ways.

So the first point I want to make is that, to a large extent, whatever is happening in our chapter on discrete random variables is just an exercise in notation. There is stuff and concepts that you are already familiar with-- probabilities, probabilities of two things happening, conditional probabilities. And all that we're doing, to some extent, is rewriting those familiar concepts in new notation. So for example, this is the joint PMF of two random variable. It gives us, for any pair or possible values of those random variables, the probability that that pair occurs simultaneously. So it's the probability that simultaneously x takes that value, and y takes that other value.

And similarly, we have the notion of the conditional PMF, which is just a list of the -- condition of -- the various conditional probabilities of interest, conditional probability that one random variable takes this value given that the other random variable takes that value. Now, a remark about conditional probabilities. Conditional probabilities generally are like ordinary probabilities. You condition on something particular. So here we condition on a particular y. So think of little y as a fixed quantity. And then look at this as a function of x. So given that y, which we condition on, given our new universe, we're considering the various possibilities for x and the probabilities that they have.

Now, the probabilities over all x's, of course, needs to add to 1. So we should have a relation of this kind. So they're just like ordinary probabilities over the different x's in a universe where we are told the value of the random variable y. Now, how are these related? So we call these the marginal, these the joint, these the conditional. And there are some relations between these. For example, to find the marginal from the joint, it's pretty straightforward. The probability that x takes a particular value is the sum of the probabilities of all of the different ways that this particular value may occur.

What are the different ways? Well, it may occur together with a certain y, or together with some other y, or together with some other y. So you look at all the possible y's that can go together with this x, and add the probabilities of all of those pairs for which we get this particular value of x. And then there's a relation between that connects these two probabilities with the conditional probability. And it's this relation. It's nothing new. It's just new notation for writing what we already know, that the probability of two things happening is the probability that the first thing happens, and then given that the first thing happens, the probability that the second one happened.

So how do we go from one to the other? Think of A as being the event that X takes the value, little x, and B being the event that Y takes the value, little y. So the joint probability is the probability that these two things happen simultaneously. It's the probability that X takes this value times the conditional probability that Y takes this value, given that X took that first value. So it's the familiar multiplication rule, but just transcribed in our new notation. So nothing new so far.

OK, why did we go through this exercise and this notation? It's because in the experiments where we're interested in the real world, typically there's going to be lots of uncertain quantities. There's going to be multiple random variables. And we want to be able to talk about them simultaneously. Okay. Why two and not more than two? How about three random variables? Well, if you understand what's going on in this slide, you should be able to kind of automatically generalize this to the case of multiple random variables.

So for example, if we have three random variables, X, Y, and Z, and you see an expression like this, it should be clear what it means. It's the probability that X takes this value and simultaneously Y takes that value and simultaneously Z takes that value. I guess that's an uppercase Z here, that's a lowercase z. And if I ask you to find the marginal of X, if I tell you the joint PMF of the three random variables and I ask you for this value, how would you find it?

Well, you will try to generalize this relation here. The probability that x occurs is the sum of the probabilities of all events that make X to take that particular value.

So what are all the events? Well, this particular x can happen together with some y and some z. We don't care which y and z. Any y and z will do. So when we consider all possibilities, we need to add here over all possible values of y's and z's. So consider all triples, x, y, z. Fix x and consider all the possibilities for the remaining variables, y and z, add these up, and that gives you the marginal PMF of X. And then there's other things that you can do. This is the multiplication rule for two events.

We saw back in chapter one that there's a multiplication rule when you talk about more than two events. And you can write a chain of conditional probabilities. We can certainly do the same in our new notation. So let's look at this rule up here. Multiplication rule for three random variables, what does it say? The probability of three things happening simultaneously, X, Y, Z taking specific values, little x, little y, little z, that probability is the probability that the first thing happens, that X takes that value.

Given that X takes that value, we multiply it with the probability that Y takes also a certain value. And now, given that X and Y have taken those particular values, we multiply with a conditional probability that the third thing happens, given that the first two things happen. So this is just the multiplication rule for three events, which would be probability of A intersection B intersection C equals-- you know the rest of the formula. You just rewrite this formula in PMF notation. Probability of A intersection B intersection C is the probability of A, which corresponds to this term, times the probability of B given A, times the probability of C given A and B.

So what else is there that's left from chapter one that we can or should generalize to random variables? Well, there's the notion of independence. So let's define what independence means. Instead of talking about just two random variables, let's go directly to the case of multiple random variables. When we talked about events, things were a little complicated. We had a simple definition for independence of two events. Two events are independent if the probability of both is equal to the product of the probabilities. But for three events, it was kind of messy. We needed to write down lots of conditions.

For random variables, things in some sense are a little simpler. We only need to write down one formula and take this as the definition of independence. Three random variables are independent if and only if, by definition, their joint probability mass function factors out into

individual probability mass functions. So the probability that all three things happen is the product of the individual probabilities that each one of these three things is happening. So independence means mathematically that you can just multiply probabilities to get to the probability of several things happening simultaneously.

So with three events, we have to write a huge number of equations, of equalities that have to hold. How can it be that with random variables we can only manage with one equality? Well, the catch is that this is not really just one equality. We require this to be true for every little x, y, and z. So in some sense, this is a bunch of conditions that are being put on the joint PMF, a bunch of conditions that we need to check. So this is the mathematical definition. What is the intuitive content of this definition? The intuitive content is the same as for events. Random variables are independent if knowing something about the realized values of some of these random variables does not change our beliefs about the likelihood of various values for the remaining random variables.

So independence would translate, for example, to a condition such as the conditional PMF of X , given y, should be equal to the marginal PMF of X. What is this saying? That you have some original beliefs about how likely it is for X to take this value. Now, someone comes and tells you that Y took on a certain value. This causes you, in principle, to revise your beliefs. And your new beliefs will be captured by the conditional PMF, or the conditional probabilities. Independence means that your revised beliefs actually will be the same as your original beliefs. Telling you information about the value of Y doesn't change what you expect for the random variable X.

Why didn't we use this definition for independence? Well, because this definition only makes sense when this conditional is well-defined. And this conditional is only well-defined if the events that Y takes on that particular value has positive probability. We cannot condition on events that have zero probability, so conditional probabilities are only defined for y's that are likely to occur, that have a positive probability.

Now, similarly, with multiple random variables, if they're independent, you would have relations such as the conditional of X, given y and z, should be the same as the marginal of X. What is this saying? Again, that if I tell you the values, the realized values of random variables Y and Z, this is not going to change your beliefs about how likely x is to occur. Whatever you believed in the beginning, you're going to believe the same thing afterwards. So it's important to keep that intuition in mind, because sometimes this way you can tell whether random variables are

independent without having to do calculations and to check this formula.

OK, so let's check our concepts with a simple example. Let's look at two random variables that are discrete, take values between one and for each. And this is a table that gives us the joint PMF. So it tells us the probability that X equals to 2 and Y equals to 1 happening simultaneously. It's an event that has probability 1/20. Are these two random variables independent? You can try to check a condition like this. But can we tell directly from the table?

If I tell you a value of Y, could that give you useful information about X? Certainly. If I tell you that Y is equal to 1, this tells you that X must be equal to 2. But if I tell you that Y was equal to 3, this tells you that, still, X could be anything. So telling you the value of Y kind of changes what you expect or what you consider possible for the values of the other random variable. So by just inspecting here, we can tell that the random variables are not independent.

Okay. What's the other concept we introduced in chapter one? We introduced the concept of conditional independence. And conditional independence is like ordinary independence but applied to a conditional universe where we're given some information. So suppose someone tells you that the outcome of the experiment is such that X is less than or equal to 2 and Y is larger than or equal to 3. So we are given the information that we now live inside this universe.

So what happens inside this universe? Inside this universe, our random variables are going to have a new joint PMF which is conditioned on the event that we were told that it has occurred. So let A correspond to this sort of event here. And now we're dealing with conditional probabilities. What are those conditional probabilities? We can put them in a table. So it's a two by two table, since we only have two possible values. What are they going to be?

Well, these probabilities show up in the ratios 1, 2, 2, and 4. Those ratios have to stay the same. The probabilities need to add up to one. So what should the denominators be since these numbers add up to nine? These are the conditional probabilities. So this is the conditional PMF in this example. Now, in this conditional universe, is x independent from y? If I tell you that y takes this value, so we live in this universe, what do you know about x? What you know about x is at this value is twice as likely as that value. If I condition on y taking this value, so we're living here, what do you know about x? What you know about x is that this value is twice as likely as that value.

So it's the same. Whether we live here or we live there, this x is twice as likely as that x. So the conditional PMF in this new universe, the conditional PMF of X given y, in the new universe is

the same as the marginal PMF of X, but of course in the new universe. So no matter what y is, the conditional PMF of X is the same. And that conditional PMF is 1/3 and 2/3. This is the conditional PMF of X in the new universe no matter what y occurs.

So Y does not give us any information about X, doesn't cause us to change our beliefs inside this little universe. And therefore the two random variables are independent. Now, the other way that you can verify that we have independence is to find the marginal PMFs of the two random variables. The marginal PMF of X, you find it by adding those two terms. You get 1/3. Adding those two terms, you get 2/3. Marginal PMF of Y, you find it, you add these two terms, and you get 1/3. And the marginal PMF of Y here is going to be 2/3.

And then you ask the question, is the joint the product of the marginals? And indeed it is. This times this gives you 1/9. This times this gives you 2/9. So the values in the table with the joint PMFs is the product of the marginal PMFs of X and Y in this universe, so the two random variables are independent inside this universe. So we say that they're conditionally independent. All right.

Now let's move to the new topic, to the new concept that we introduce in this chapter, which is the concept of expectations. So what are the things to know here? One is the general idea. The way to think about expectations is that it's something like the average value for random variable if you do an experiment over and over, and if you interpret probabilities as frequencies. So you get x's over and over with a certain frequency -- P(x) -- a particular value, little x, gets realized. And each time that this happens, you get x dollars. How many dollars do you get on the average? Well, this formula gives you that particular average.

So first thing we do is to write down a definition for this sort of concept. But then the other things you need to know is how to calculate expectations using shortcuts sometimes, and what properties they have. The most important shortcut there is is that, if you want to calculate the expected value, the average value for a random variable, you do not need to find the PMF of that random variable. But you can work directly with the x's and the y's. So you do the experiment over and over. The outcome of the experiment is a pair (x,y). And each time that a certain (x,y) happens, you get so many dollars.

So this fraction of the time, a certain (x,y) happens. And that fraction of the time, you get so many dollars, so this is the average number of dollars that you get. So what you end up, since it is the average, then that means that it corresponds to the expected value. Now, this is

something that, of course, needs a little bit of mathematical proof. But this is just a different way of accounting. And it turns out we give you the right answer. And it's a very useful shortcut.

Now, when we're talking about functions of random variables, in general, we cannot speak just about averages. That is, the expected value of a function of a random variable is not the same as the function of the expected values. A function of averages is not the same as the average of a function. So in general, this is not true. But what it's important to know is to know the exceptions to this rule. And the important exceptions are mainly two. One is the case of linear functions of a random variable. We discussed this last time. So the expected value of temperature in Celsius is, you first find the expected value of temperature in Fahrenheit, and then you do the conversion to Celsius. So whether you first average and then do the conversion to the new units or not, it shouldn't matter when you get the result.

The other property that turns out to be true when you talk about multiple random variables is that expectation still behaves linearly. So let X, Y, and Z be the score of a random student at each one of the three sections of the SAT. So the overall SAT score is X plus Y plus Z. This is the average score, the average total SAT score. Another way to calculate that average is to look at the first section of the SAT and see what was the average. Look at the second section, look at what was the average, and so the third, and add the averages. So you can do the averages for each section separately, add the averages, or you can find total scores for each student and average them.

So I guess you probably believe that this is correct if you talk just about averaging scores. Since expectations are just the variation of averages, it turns out that this is also true in general. And the derivation of this is very simple, based on the expected value rule. And you can look at it in the notes. So this is one exception, which is linearity. The second important exception is the case of independent random variables, that the product of two random variables has an expectation which is the product of the expectations. In general, this is not true. But for the case where we have independence, the expectation works out as follows. Using the expected value rule, this is how you calculate the expected value of a function of a random variable. So think of this as being your g(X, Y) and this being your g(little x, y).

So this is something that's generally true. Now, if we have independence, then the PMFs factor out, and then you can separate this sum by bringing together the x terms, bring them outside the y summation. And you find that this is the same as expected value of X times the

expected value of Y. So independence is used in this step here. OK, now what if X and Y are independent, but instead of taking the expectation of X times Y, we take the expectation of the product of two functions of X and Y? I claim that the expected value of the product is still going to be the product of the expected values.

How do we show that? We could show it by just redoing this derivation here. Instead of X and Y, we would have g(X) and h(Y), so the algebra goes through. But there's a better way to think about it which is more conceptual. And here's the idea. If X and Y are independent, what does it mean? X does not convey any information about Y. If X conveys no information about Y, does X convey information about h(Y)? No. If X tells me nothing about Y, nothing new, it shouldn't tell me anything about h(Y).

Now, if X tells me nothing about h of h(Y), could g(X) tell me something about h(Y)? No. So the idea is that, if X is unrelated to Y, doesn't have any useful information, then g(X) could not have any useful information for h(Y). So if X and Y are independent, then g(X) and h(Y) are also independent. So this is something that one can try to prove mathematically, but it's more important to understand conceptually why this is so. It's in terms of conveying information.

So if X tells me nothing about Y, X cannot tell me anything about Y cubed, or X cannot tell me anything by Y squared, and so on. That's the idea. And once we are convinced that g(X) and h(Y) are independent, then we can apply our previous rule, that for independent random variables, expectations multiply the right way. Apply the previous rule, but apply it now to these two independent random variables. And we get the conclusion that we wanted.

Now, besides expectations, we also introduced the concept of the variance. And if you remember the definition of the variance, let me write down the formula for the variance of aX. It's the expected value of the random variable that we're looking at minus the expected value of the random variable that we're looking at. So this is the difference of the random variable from its mean. And we take that difference and square it, so it's the squared distance from the mean, and then take expectations of the whole thing.

So when you look at that expression, you realize that a can be pulled out of those expressions. And because there is a squared, when you pull out the a, it's going to come out as an a-squared. So that gives us the rule for finding the variance of a scale or product of a random variable. The variance captures the idea of how wide, how spread out a certain distribution is. Bigger variance means it's more spread out.

Now, if you take a random variable and the constants to it, what does it do to its distribution? It just shifts it, but it doesn't change its width. So intuitively it means that the variance should not change. You can check that mathematically, but it should also make sense intuitively. So the variance, when you add the constant, does not change. Now, can you add variances is the way we added expectations? Does variance behave linearly? It turns out that not always. Here, we need a condition. It's only in special cases-- for example, when the two random variables are independent-- that you can add variances. The variance of the sum is the sum of the variances if X and Y are independent.

The derivation of this is, again, very short and simple. We'll skip it, but it's an important fact to remember. Now, to appreciate why this equality is not true always, we can think of some extreme examples. Suppose that X is the same as Y. What's going to be the variance of X plus Y? Well, X plus Y, in this case, is the same as 2X, so we're going to get 4 times the variance of X, which is different than the variance of X plus the variance of X.

So that expression would give us twice the variance of X. But actually now it's 4 times the variance of X. The other extreme would be if X is equal to -Y. Then the variance is the variance of the random variable, which is always equal to 0. Now, a random variable which is always equal to 0 has no uncertainty. It is always equal to its mean value, so the variance, in this case, turns out to be 0.

So in both of these cases, of course we have random variables that are extremely dependent. Why are they dependent? Because if I tell you something about Y, it tells you an awful lot about the value of X. There's a lot of information about X if I tell you Y, in this case or in that case. And finally, a short drill. If I tell you that the random variables are independent and you want to calculate the variance of a linear combination of this kind, then how do you argue? You argue that, since X and Y are independent, this means that X and 3Y are also independent. X has no information about Y, so X has no information about -Y. X has no information about -Y, so X should not have any information about -3Y. So X and -3Y are independent.

So the variance of Z should be the variance of X plus the variance of -3Y, which is the variance of X plus 9 times the variance of Y. The important thing to note here is that no matter what happens, you end up getting a plus here, not a minus. So that's the sort of important thing to remember in this type of calculation. So this has been all concepts, reviews, new concepts and all that. It's the usual fire hose. Now let's use them to do something useful finally.

So let's revisit our old example, the binomial distribution, which counts the number of successes in independent trials of a coin. It's a biased coin that has a probability of heads, or probability of success, equal to p at each trial. Finally, we can go through the exercise of calculating the expected value of this random variable. And there's the way of calculating that expectation that would be the favorite of those people who enjoy algebra, which is to write down the definition of the expected value. We add over all possible values of the random variable, over all the possible k's, and weigh them according to the probabilities that this particular k occurs. The probability that X takes on a particular value k is, of course, the binomial PMF, which is this familiar formula.

Clearly, that would be a messy and challenging calculation. Can we find a shortcut? There's a very clever trick. There's lots of problems in probability that you can approach really nicely by breaking up the random variable of interest into a sum of simpler and more manageable random variables. And if you can make it to be a sum of random variables that are just 0's or 1's, so much the better. Life is easier. Random variables that take values 0 or 1, we call them indicator variables. They indicate whether an event has occurred or not.

In this case, we look at each coin flip one at a time. For the i-th flip, if it resulted in heads or a success, we record it 1. If not, we record it 0. And then we look at the random variable. If we take the sum of the Xi's, what is it going to be? We add one each time that we get a success, so the sum is going to be the total number of successes. So we break up the random variable of interest as a sum of really nice and simple random variables.

And now we can use the linearity of expectations. We're going to find the expectation of X by finding the expectation of the Xi's and then adding the expectations. What's the expected value of Xi? Well, Xi takes the value 1 with probability p, and takes the value 0 with probability 1-p. So the expected value of Xi is just p. So the expected value of X is going to be just n times p. Because X is the sum of n terms, each one of which has expectation p, the expected value of the sum is the sum of the expected values. So I guess that's a pretty good shortcut for doing this horrendous calculation up there.

So in case you didn't realize it, that's what we just established without doing any algebra. Good. How about the variance of X, of Xi? Two ways to calculate it. One is by using directly the formula for the variance, which would be -- let's see what it would be. With probability p, you get a 1. And in this case, you are so far from the mean. That's your squared distance from

the mean. With probability 1-p, you get a 0, which is so far away from the mean. And then you can simplify that formula and get an answer.

How about a slightly easier way of doing it. Instead of doing the algebra here, let me indicate the slightly easier way. We have a formula for the variance that tells us that we can find the variance by proceeding this way. That's a formula that's generally true for variances. Why is this easier? What's the expected value of Xi squared? Backtrack. What is Xi squared, after all? It's the same thing as Xi. Since Xi takes value 0 and 1, Xi squared also takes the same values, 0 and 1. So the expected value of Xi squared is the same as the expected value of Xi, which is equal to p. And the expected value of Xi squared is p squared, so we get the final answer, p times (1-p).

If you were to work through and do the cancellations in this messy expression here, after one line you would also get to the same formula. But this sort of illustrates that working with this formula for the variance, sometimes things work out a little faster. Finally, are we in business? Can we calculate the variance of the random variable X as well? Well, we have the rule that for independent random variables, the variance of the sum is the sum of the variances. So to find the variance of X, we just need to add the variances of the Xi's. We have n Xi's, and each one of them has variance p_n times (1-p). And we are done.

So this way, we have calculated both the mean and the variance of the binomial random variable. It's interesting to look at this particular formula and see what it tells us. If you are to plot the variance of X as a function of p, it has this shape. And the maximum is here at 1/2. p times (1-p) is 0 when p is equal to 0. And when p equals to 1, it's a quadratic, so it must have this particular shape. So what does it tell us? If you think about variance as a measure of uncertainty, it tells you that coin flips are most uncertain when your coin is fair. When p is equal to 1/2, that's when you have the most randomness.

And this is kind of intuitive. if on the other hand I tell you that the coin is extremely biased, p very close to 1, which means it almost always gives you heads, then that would be a case of low variance. There's low variability in the results. There's little uncertainty about what's going to happen. It's going to be mostly heads with some occasional tails. So p equals 1/2. Fair coin, that's the coin which is the most uncertain of all coins, in some sense. And it corresponds to the biggest variance. It corresponds to an X that has the widest distribution.

Now that we're on a roll and we can calculate such hugely complicated sums in simple ways,

let us try to push our luck and do a problem with this flavor, but a little harder than that. So you go to one of those old-fashioned cocktail parties. All males at least will have those standard big hats which look identical. They check them in when they walk in. And when they walk out, since they look pretty identical, they just pick a random hat and go home. So n people, they pick their hats completely at random, quote, unquote, and then leave. And the question is, to say something about the number of people who end up, by accident or by luck, to get back their own hat, the exact same hat that they checked in.

OK, first what do we mean completely at random? Completely at random, we basically mean that any permutation of the hats is equally likely. Any way of distributing those n hats to the n people, any particular way is as likely as any other way. So there's complete symmetry between hats and people. So what we want to do is to calculate the expected value and the variance of this random variable X. Let's start with the expected value.

Let's reuse the trick from the binomial case. So total number of hats picked, we're going to think of total number of hats picked as a sum of (0, 1) random variables. X1 tells us whether person 1 got their own hat back. If they did, we record a 1. X2, the same thing. By adding all X's is how many 1's did we get, which counts how many people selected their own hats. So we broke down the random variable of interest, the number of people who get their own hats back, as a sum of random variables. And these random variables, again, are easy to handle, because they're binary. The only take two values.

What's the probability that Xi is equal to 1, the i-th person has a probability that they get their own hat? There's n hats by symmetry. The chance is that they end up getting their own hat, as opposed to any one of the other n - 1 hats, is going to be 1/n. So what's the expected value of Xi? It's one times 1/n. With probability 1/n, you get your own hat, or you get a value of 0 with probability 1-1/n, which is 1/n.

All right, so we got the expected value of the Xi's. And remember, we want to do is to calculate the expected value of X by using this decomposition? Are the random variables Xi independent of each other? You can try to answer that question by writing down a joint PMF for the X's, but I'm sure that you will not succeed. But can you think intuitively? If I tell you information about some of the Xi's, does it give you information about the remaining ones? Yeah. If I tell you that out of 10 people, 9 of them got their own hat back, does that tell you something about the 10th person? Yes. If 9 got their own hat, then the 10th must also have gotten their own hat back.

So the first 9 random variables tell you something about the 10th one. And conveying information of this sort, that's the case of dependence. All right, so the random variables are not independent. Are we stuck? Can we still calculate the expected value of X? Yes, we can. And the reason we can is that expectations are linear. Expectation of a sum of random variables is the sum of the expectations. And that's always true. There's no independence assumption that's being used to apply that rule. So we have that the expected value of X is the sum of the expected value of the Xi's. And this is a property that's always true. You don't need independence. You don't care. So we're adding n terms, each one of which has expected value 1/n. And the final answer is 1.

So out of the 100 people who selected hats at random, on the average, you expect only one of them to end up getting their own hat back. Very good. So since we are succeeding so far, let's try to see if we can succeed in calculating the variance as well. And of course, we will. But it's going to be a little more complicated. The reason it's going to be a little more complicated is because the Xi's are not independent, so the variance of the sum is not the same as the sum of the variances.

So it's not enough to find the variances of the Xi's. We'll have to do more work. And here's what's involved. Let's start with the general formula for the variance, which, as I mentioned before, it's usually the simpler way to go about calculating variances. So we need to calculate the expected value for X-squared, and subtract from it the expectation squared. Well, we already found the expected value of X. It's equal to 1. So 1-squared gives us just 1. So we're left with the task of calculating the expected value of X-squared, the random variable X-squared. Let's try to follow the same idea. Write this messy random variable, X-squared, as a sum of hopefully simpler random variables.

So X is the sum of the Xi's, so you square both sides of this. And then you expand the right-hand side. When you expand the right-hand side, you get the squares of the terms that appear here. And then you get all the cross-terms. For every pair of (i,j) that are different, i different than j, you're going to have a cross-term in the sum. So now, in order to calculate the expected value of X-squared, what does our task reduce to? It reduces to calculating the expected value of this term and calculating the expected value of that term. So let's do them one at a time.

Expected value of Xi squared, what is it going to be? Same trick as before. Xi takes value 0 or

1, so Xi squared takes just the same values, 0 or 1. So that's the easy one. That's the same as expected value of Xi, which we already know to be 1/n. So this gives us a first contribution down here. The expected value of this term is going to be what? We have n terms in the summation. And each one of these terms has an expectation of 1/n. So we did a piece of the puzzle. So now let's deal with the second piece of the puzzle.

Let's find the expected value of Xi times Xj. Now by symmetry, the expected value of Xi times Xj is going to be the same no matter what i and j you see. So let's just think about X1 and X2 and try to find the expected value of X1 and X2. X1 times X2 is a random variable. What values does it take? Only 0 or 1? Since X1 and X2 are 0 or 1, their product can only take the values of 0 or 1. So to find the probability distribution of this random variable, it's just sufficient to find the probability that it takes the value of 1. Now, what does X1 times X2 equal to 1 mean? It means that X1 was 1 and X2 was 1. The only way that you can get a product of 1 is if both of them turned out to be 1's.

So that's the same as saying, persons 1 and 2 both picked their own hats. The probability that person 1 and person 2 both pick their own hats is the probability of two things happening, which is the product of the first thing happening times the conditional probability of the second, given that the first happened. And in words, this is the probability that the first person picked their own hat times the probability that the second person picks their own hat, given that the first person already picked their own. So what's the probability that the first person picks their own hat? We know that it's 1/n.

Now, how about the second person? If I tell you that one person has their own hat, and that person takes their hat and goes away, from the point of view of the second person, there's n - 1 people left looking at n - 1 hats. And they're getting just hats at random. What's the chance that I will get my own? It's 1/n - 1. So think of them as person 1 goes, picks a hat at random, it happens to be their own, and it leaves. You're left with n - 1 people, and there are n - 1 hats out there.

Person 2 goes and picks a hat at random, with probability 1/n - 1, is going to pick his own hat. So the expected value now of this random variable is, again, that same number, because this is a 0, 1 random variable. So this is the same as expected value of Xi times Xj when i different than j. So here, all that's left to do is to add the expectations of these terms. Each one of these terms has an expected value that's 1/n times (1/n - 1).

And how many terms do we have? How many of these are we adding up? It's n-squared - n. When you expand the quadratic, there's a total of n-squared terms. Some are self-terms, n of them. And the remaining number of terms is n-squared - n. So here we got n-squared - n terms. And so we need to multiply here with n-squared - n. And after you realize that this number here is 1, and you realize that this is the same as the denominator, you get the answer that the expected value of X squared equals 2. And then, finally going up to the top formula, we get the expected value of X squared, which is 2 - 1, and the variance is just equal to 1.

So the variance of this random variable, number of people who get their own hats back, is also equal to 1, equal to the mean. Looks like magic. Why is this the case? Well, there's a deeper explanation why these two numbers should come out to be the same. But this is something that would probably have to wait a couple of chapters before we could actually explain it. And so I'll stop here.