6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013 Transcript – Lecture 6

PROFESSOR: OK so let's start. So today, we're going to continue the subject from last time. And the subject is random variables. As we discussed, random variables basically associate numerical values with the outcomes of an experiment. And we want to learn how to manipulate them.

Now to a large extent, what's going to happen, what's happening during this chapter, is that we are revisiting the same concepts we have seen in chapter one. But we're going to introduce a lot of new notation, but really dealing with the same kind of stuff. The only difference where we go beyond the new notation, the new concept in this chapter is the concept of the expectation or expected values. And we're going to learn how to manipulate expectations.

So let us start with a quick review of what we discussed last time. We talked about random variables. Loosely speaking, random variables are random quantities that result from an experiment.

More precisely speaking, mathematically speaking, a random variable is a function from the sample space to the real numbers. That is, you give me an outcome, and based on that outcome, I can tell you the value of the random variable. So the value of the random variable is a function of the outcome that we have.

Now given a random variable, some of the numerical outcomes are more likely than others. And we want to say which ones are more likely and how likely they are. And the way we do that is by writing down the probabilities of the different possible numerical outcomes.

Notice here, the notation. We use uppercase to denote the random variable. We use lowercase to denote real numbers.

So the way you read this, this is the probability that the random variable, capital X, happens to take the numerical value, little x. This is a concept that's familiar from chapter one. And this is just the new notation we will be using for that concept.

It's the Probability Mass Function of the random variable, capital X. So the subscript just indicates which random variable we're talking about. And it's the probability assigned to a particular outcome.

And we want to assign such probabilities for all possibly numerical values. So you can think of this as being a function of little x. And it tells you how likely every little x is going to be.

Now the new concept we introduced last time is the concept of the expected value for random variable, which is defined this way. You look at all the possible outcomes. And you form some kind of average of all the possible numerical values over the random variable capital X. You consider all the possible numerical values, and you form an average. In fact, it's a weighted average where, to every little x, you assign a weight equal to the probability that that particular little x is going to be realized.

Now, as we discussed last time, if you have a random variable, you can take a function of a random variable. And that's going to be a new random variable. So if capital X is a random variable and g is a function, g of X is a new random variable.

You do the experiment. You get an outcome. This determines the value of X. And that determines the value of g of X.

So the numerical value of g of X is determined by whatever happens in the experiment. It's random. And that makes it a random variable.

Since it's a random variable, it has an expectation of its own. So how would we calculate the expectation of g of X? You could proceed by just using the definition, which would require you to find the PMF of the random variable g of X. So find the PMF of g of X, and then apply the formula for the expected value of a random variable with known PMF.

But there is also a shortcut, which is just a different way of doing the counting and the calculations, in which we do not need to find the PMF of g of X. We just work with the PMF of the original random variable. And what this is saying is that the average value of g of X is obtained as follows.

You look at all the possible results, the X's, how likely they are. And when that particular X happens, this is how much you get. And so this way, you add these things up. And you get the average amount that you're going to get, the average value of g of X, where you average over the likelihoods of the different X's.

Now expected values have some properties that are always true and some properties that sometimes are not true. So the property that is not always true is that this would be the same as g of the expected value of X. So in general, this is not true. You cannot interchange function and expectation, which means you cannot reason on the average, in general.

But there are some exceptions. When g is a linear function, then the expected value for a linear function is the same as that same linear function of the expectation. So for linear functions, so for random variable, the expectation behaves nicely.

So this is basically telling you that, if X is degrees in Celsius, alpha X plus b is degrees in Fahrenheit, you can first do the conversion to Fahrenheit and take the average. Or you can find the average temperature in Celsius, and then do the conversion to Fahrenheit. Either is valid.

So the expected value tells us something about where is the center of the distribution, more specifically, the center of mass or the center of gravity of the PMF, when you plot it as a bar graph. Besides the average value, you may be interested in knowing how far will you be from the average, typically. So let's look at this quantity, X minus expected value of X.

This is the distance from the average value. So for a random outcome of the experiment, this quantity in here measures how far away from the mean you happen to be. This quantity inside the brackets is a random variable.

Why? Because capital X is random. And what we have here is capital X, which is random, minus a number. Remember, expected values are numbers.

Now a random variable minus a number is a new random variable. It has an expectation of its own. We can use the linearity rule, expected value of something minus something else is just the difference of their expected value. So it's going to be expected value of X minus the expected value over this thing.

Now this thing is a number. And the expected value of a number is just the number itself. So we get from here that this is expected value minus expected value. And we get zero.

What is this telling us? That, on the average, the assigned difference from the mean is equal to zero. That is, the mean is here. Sometimes X will fall to the right. Sometimes X will fall to the left. On the average, the average distance from the mean is going to be zero, because sometimes the realized distance will be positive, sometimes it will be negative. Positives and negatives cancel out.

So if we want to capture the idea of how far are we from the mean, just looking at the assigned distance from the mean is not going to give us any useful information. So if we want to say something about how far we are, typically, we should do something different. One possibility might be to take the absolute values of the differences. And that's a quantity that sometimes people are interested in. But it turns out that a more useful quantity happens to be the variance of a random variable, which actually measures the average squared distance from the mean.

So you have a random outcome, random results, random numerical value of the random variable. It is a certain distance away from the mean. That certain distance is random.

We take the square of that. This is the squared distance from the mean, which is again random. Since it's random, it has an expected value of its own. And that expected value, we call it the variance of X. And so we have this particular definition.

Using the rule that we have up here for how to calculate expectations of functions of a random variable, why does that apply? Well, what we have inside the brackets here is a function of the random variable, capital X. So we can apply this rule where g is this particular function. And we can use that to calculate the variance, starting with the PMF of the random variable X. And then we have a useful formula that's a nice shortcut, sometimes, if you want to do the calculation.

Now one thing that's slightly wrong with the variance is that the units are not right, if you want to talk about the spread a of a distribution. Suppose that X is a random variable measured in meters. The variance will have the units of meters squared. So it's a kind of a different thing.

If you want to talk about the spread of the distribution using the same units as you have for X, it's convenient to take the square root of the variance. And that's something that we define. And we call it to the standard deviation of X, or the standard deviation of the distribution of X. So it tells you the amount of spread in your distribution. And it is in the same units as the random variable itself that you are dealing with.

And we can just illustrate those quantities with an example that's about as simple as it can be. So consider the following experiment. You're going to go from here to New York, let's say, 200 miles.

And you have two alternatives. Either you'll get your private plane and go at a speed of 200 miles per hour, constant speed during your trip, or otherwise, you'll decide to walk really, really slowly, at the leisurely pace of one mile per hour. So you pick the speed at random by doing this experiment, by flipping a coin.

And with probability one-half, you do one thing. With probably one-half, you do the other thing. So your V is a random variable.

In case you're interested in how much time it's going to take you to get there, well, time is equal to distance divided by speed. So that's the formula. The time itself is a random variable, because it's a function of V, which is random. How much time it's going to take you depends on the coin flip that you do in the beginning to decide what speed you are going to have.

OK, just as a warm up, the trivial calculations. To find the expected value of V, you argue as follows. With probability one-half, V is going to be one. And with probability one-half, V is going to be 200. And so the expected value of your speed is 100.5.

If you wish to calculate the variance of V, then you argue as follows. With probability one-half, I'm going to travel at the speed of one, whereas, the mean is 100.5. So this is the distance from the mean, if I decide to travel at the speed of one.

We take that distance from the mean squared. That's one contribution to the variance. And with probability one-half, you're going to travel at the speed of 200, which is this much away from the mean. You take the square of that.

OK, so approximately how big is this number? Well, this is roughly 100 squared. That's also 100 squared. So approximately, the variance of this random variable is 100 squared.

Now if I tell you that the variance of this distribution is 10,000, it doesn't really help you to relate it to this diagram. Whereas, the standard deviation, where you take the square root, is more interesting. It's the square root of 100 squared, which is a 100.

And the standard deviation, indeed, gives us a sense of how spread out this distribution is from the mean. So the standard deviation basically gives us some indication about this spacing that we have here. It tells us the amount of spread in our distribution.

OK, now let's look at what happens to time. V is a random variable. T is a random variable. So now let's look at the expected values and all of that for the time.

OK, so the time is a function of a random variable. We can find the expected time by looking at all possible outcomes of the experiment, the V's, weigh them according to their probabilities, and for each particular V, keep track of how much time it took us. So if V is one, which happens with probability one-half, the time it takes is going to be 200. If we travel at speed of one, it takes us 200 time units.

And otherwise, if our speed is equal to 200, the time is one. So the expected value of T is once more the same as before. It's 100.5.

So the expected speed is 100.5. The expected time is also 100.5. So the product of these expectations is something like 10,000.

How about the expected value of the product of T and V? Well, T times V is 200. No matter what outcome you have in the experiment, in that particular outcome, T times V is total distance traveled, which is exactly 200. And so what do we get in this simple example is that the expected value of the product of these two random variables is different than the product of their expected values.

This is one more instance of where we cannot reason on the average. So on the average, over a large number of trips, your average time would be 100. On the average, over a large number of trips, your average speed would be 100.

But your average distance traveled is not 100 times 100. It's something else. So you cannot reason on the average, whenever you're dealing with non-linear things. And the non-linear thing here is that you have a function which is a product of stuff, as opposed to just linear sums of stuff.

Another way to look at what's happening here is the expected value of the time. Time, by definition, is 200 over the speed. Expected value of the time, we found it to be about a 100. And so expected value of 200 over V is about a 100.

But it's different from this quantity here, which is roughly equal to 2, and so 200. Expected value of V is about 100. So this quantity is about equal to two. Whereas, this quantity up here is about 100.

So what do we have here? We have a non-linear function of V. And we find that the expected value of this function is not the same thing as the function of the expected value. So again, that's an instance where you cannot interchange expected values and functions. And that's because things are non-linear.

OK, so now let us introduce a new concept. Or maybe it's not quite a new concept. So we discussed, in chapter one, that we have probabilities. We also have conditional probabilities.

What's the difference between them? Essentially, none. Probabilities are just an assignment of probability values to give different outcomes, given a particular model.

Somebody comes and gives you new information. So you come up with a new model. And you have a new probabilities. We call these conditional probabilities, but they taste and behave exactly the same as ordinary probabilities.

So since we can have conditional probabilities, why not have conditional PMFs as well, since PMFs deal with probabilities anyway. So we have a random variable, capital X. It has a PMF of its own.

For example, it could be the PMF in this picture, which is a uniform PMF that takes for possible different values. And we also have an event. And somebody comes and tells us that this event has occurred.

The PMF tells you the probability that capital X equals to some little x. Somebody tells you that a certain event has occurred that's going to make you change the probabilities that you assign to the different values. You are going to use conditional probabilities.

So this part, it's clear what it means from chapter one. And this part is just the new notation we're using in this chapter to talk about conditional probabilities. So this is just a definition.

So the conditional PMF is an ordinary PMF. But it's the PMF that applies to a new model in which we have been given some information about the outcome of the experiment. So to make it concrete, consider this event here.

Take the event that capital X is bigger than or equal to two. In the picture, what is the event A? The event A consists of these three outcomes.

OK, what is the conditional PMF, given that we are told that event A has occurred? Given that the event A has occurred, it basically tells us that this outcome has not occurred. There's only three possible outcomes now.

In the new universe, in the new model where we condition on A, there's only three possible outcomes. Those three possible outcomes were equally likely when we started. So in the conditional universe, they will remain equally likely.

Remember, whenever you condition, the relative likelihoods remain the same. They keep the same proportions. They just need to be re-scaled, so that they add up to one.

So each one of these will have the same probability. Now in the new world, probabilities need to add up to 1. So each one of them is going to get a probability of 1/3 in the conditional universe.

So this is our conditional model. So our PMF is equal to 1/3 for X equals to 2, 3 and 4. All right.

Now whenever you have a probabilistic model involving a random variable and you have a PMF for that random variable, you can talk about the expected value of that random variable. We defined expected values just a few minutes ago. Here, we're dealing with a conditional model and conditional probabilities.

And so we can also talk about the expected value of the random variable X in this new universe, in this new conditional model that we're dealing with. And this leads us to the definition of the notion of a conditional expectation. The conditional expectation is nothing but an ordinary expectation, except that you don't use the original PMF.

You use the conditional PMF. You use the conditional probabilities. It's just an ordinary expectation, but applied to the new model that we have to the conditional universe where we are told that the certain event has occurred.

So we can now calculate the condition expectation, which, in this particular example, would be 1/3. That's the probability of a 2, plus 1/3 which is the probability of a 3 plus 1/3, the probability of a 4. And then you can use your calculator to find the answer, or you can just argue by symmetry. The expected value has to be the center of gravity of the PMF we're working with, which is equal to 3.

So conditional expectations are no different from ordinary expectations. They're just ordinary expectations applied to a new type of situation or a new type of model. Anything we might know about expectations will remain valid about conditional expectations.

So for example, the conditional expectation of a linear function of a random variable is going to be the linear function of the conditional expectations. Or you can take any formula that you might know, such as the formula that expected value of X is equal to the-- sorry-- expected value of g of X is the sum over all X's of g of X times the PMF of X. So this is the formula that we already know about how to calculate expectations of a function of a random variable.

If we move to the conditional universe, what changes? In the conditional universe, we're talking about the conditional expectation, given that event A has occurred. And we use the conditional probabilities, given that A has occurred.

So any formula has a conditional counterpart. In the conditional counterparts, expectations get replaced by conditional expectations. And probabilities get replaced by conditional probabilities. So once you know the first formula and you know the general idea, there's absolutely no reason for you to memorize a formula like this one. You shouldn't even have to write it on your cheat sheet for the exam, OK?

OK, all right, so now let's look at an example of a random variable that we've seen before, the geometric random variable, and this time do something a little more interesting with it. Do you remember from last time what the geometric random variable is? We do coin flips. Each time

there's a probability of P of obtaining heads. And we're interested in the number of tosses we're going to need until we observe heads for the first time.

The probability that the random variable takes the value K, this is the probability that the first K appeared at the K-th toss. So this is the probability of K minus 1 consecutive tails followed by a head. So this is the probability of having to weight K tosses.

And when we plot this PMF, it has this kind of shape, which is the shape of a geometric progression. It starts at 1, and it goes all the way to infinity. So this is a discrete random variable that takes values over an infinite set, the set of the positive integers.

So it's a random variable, therefore, it has an expectation. And the expected value is, by definition, we'll consider all possible values of the random variable. And we weigh them according to their probabilities, which leads us to this expression.

You may have evaluated that expression some time in your previous life. And there are tricks for how to evaluate this and get a closed-form answer. But it's sort of an algebraic trick. You might not remember it. How do we go about doing this summation?

Well, we're going to use a probabilistic trick and manage to evaluate the expectation of X, essentially, without doing any algebra. And in the process of doing so, we're going to get some intuition about what happens in coin tosses and with geometric random variables. So we have two people who are going to do the same experiment, flip a coin until they obtain heads for the first time.

One of these people is going to use the letter Y to count how many heads it took. So that person starts flipping right now. This is the current time. And they are going to obtain tails, tails, tails, until eventually they obtain heads. And this random variable Y is, of course, geometric, so it has a PMF of this form.

OK, now there is a second person who is doing that same experiment. That second person is going to take, again, a random number, X, until they obtain heads for the first time. And of course, X is going to have the same PMF as Y.

But that person was impatient. And they actually started flipping earlier, before the Y person started flipping. They flipped the coin twice. And they were unlucky, and they obtained tails both times. And so they have to continue.

Looking at the situation at this time, how do these two people compare? Who do you think is going to obtain heads first? Is one more likely than the other?

So if you play at the casino a lot, you'll say, oh, there were two tails in a row, so a head should be coming up sometime soon. But this is a wrong argument, because coin flips, at least in our model, are independent. The fact that these two happened to be tails doesn't change anything about our beliefs about what's going to be happening here.

So what's going to be happening to that person is they will be flipping independent coin flips. That person will also be flipping independent coin flips. And both of them wait until the first head occurs. They're facing an identical situation, starting from this time.

OK, now what's the probabilistic model of what this person is facing? The time until that person obtains heads for the first time is X. So this number of flips until they obtain heads for the first time is going to be X minus 2. So X is the total number until the first head. X minus 2 is the number or flips, starting from here.

Now what information do we have about that person? We have the information that their first two flips were tails. So we're given the information that X was bigger than 2. So the probabilistic model that describes this piece of the experiment is that it's going to take a random number of flips until the first head.

That number of flips, starting from here until the next head, is that number X minus 2. But we're given the information that this person has already wasted 2 coin flips. Now we argued that probabilistically, this person, this part of the experiment here is identical with that part of the experiment. So the PMF of this random variable, which is X minus 2, conditioned on this information, should be the same as that PMF that we have down there.

So the formal statement that I'm making is that this PMF here of X minus 2, given that X is bigger than 2, is the same as the PMF of X itself. What is this saying? Given that I tell you that you already did a few flips and they were failures, the remaining number of flips until the first head has the same geometric distribution as if you were starting from scratch. Whatever happened in the past, it happened, but has no bearing what's going to happen in the future. Remaining coin flips until a head has the same distribution, whether you're starting right now, or whether you had done some other stuff in the past.

So this is a property that we call the memorylessness property of the geometric distribution. Essentially, it says that whatever happens in the future is independent from whatever happened in the past. And that's true almost by definition, because we're assuming independent coin flips. Really, independence means that information about one part of the experiment has no bearing about what's going to happen in the other parts of the experiment.

The argument that I tried to give using the intuition of coin flips, you can make it formal by just manipulating PMFs formally. So this is the original PMF of X.

Suppose that you condition on the event that X is bigger than 3. This conditioning information, what it does is it tells you that this piece did not happen. You're conditioning just on this event.

When you condition on that event, what's left is the conditional PMF, which has the same shape as this one, except that it needs to be re-normalized up, so that the probabilities add up to one. So you take that picture, but you need to change the height of it, so that these terms add up to 1.

And this is the conditional PMF of X, given that X is bigger than 2. But we're talking here not about X. We're talking about the remaining number of heads. Remaining number of heads is X minus 2.

If we have the PMF of X, can we find the PMF of X minus 2? Well, if X is equal to 3, that corresponds to X minus 2 being equal to 1. So this probability here should be equal to that probability.

The probability that X is equal to 4 should be the same as the probability that X minus 2 is equal to 2. So basically, the PMF of X minus 2 is the same as the PMF of X, except that it gets shifted by these 2 units.

So this way, we have formally derived the conditional PMF of the remaining number of coin tosses, given that the first two flips were tails. And we see that it's exactly the same as the PMF that we started with. And so this is the formal proof of this statement here. So it's useful here to digest both these formal statements and understand it and understand the notation that is involved here, but also to really appreciate the intuitive argument what this is really saying.

OK, all right, so now we want to use this observation, this memorylessness, to eventually calculate the expected value for a geometric random variable. And the way we're going to do it is by using a divide and conquer tool, which is an analog of what we have already seen sometime before. Remember our story that there's a number of possible scenarios about the world? And there's a certain event, B, that can happen under any of these possible scenarios.

And we have the total probability theory. And that tells us that, to find the probability of this event, B, you consider the probabilities of B under each scenario. And you weigh those probabilities according to the probabilities of the different scenarios that we have. So that's a formula that we already know and have worked with.

What's the next step? Is it something deep? No, it's just translation in different notation.

This is the exactly same formula, but with PMFs. The event that capital X is equal to little x can happen in many different ways. It can happen under either scenario.

And within each scenario, you need to use the conditional probabilities of that event, given that this scenario has occurred. So this formula is identical to that one, except that we're using conditional PMFs, instead of conditional probabilities. But conditional PMFs, of course, are nothing but conditional probabilities anyway. So nothing new so far.

Then what I do is to take this formula here and multiply both sides by X and take the sum over all X's. What do we get on this side? We get the expected value of X.

What do we get on that side? Probability of A1. And then here, sum over all X's of X times P. That's, again, the same calculation we have when we deal with expectations, except that, since here, we're dealing with conditional probabilities, we're going to get the conditional expectation.

And this is the total expectation theorem. It's a very useful way for calculating expectations using a divide and conquer method.

We figure out the average value of X under each one of the possible scenarios. The overall average value of X is a weighted linear combination of the expected values of X in the different scenarios where the weights are chosen according to the different probabilities.

OK, and now we're going to apply this to the case of a geometric random variable. And we're going to divide and conquer by considering separately the two cases where the first toss was heads, and the other case where the first toss was tails. So the expected value of X is the probability that the first toss was heads, so that X is equal to 1, and the expected value if that happened.

What is the expected value of X, given that X is equal to 1? If X is known to be equal to 1, then X becomes just a number. And the expected value of a number is the number itself. So this first line here is the probability of heads in the first toss times the number 1.

So the probability that X is bigger than 1 is 1 minus P. And then we need to do something about this conditional expectation. What is it?

I can write it in, perhaps, a more suggested form, as expected the value of X minus 1, given that X minus 1 is bigger than 1. Ah. OK, X bigger than 1 is the same as X minus 1 being positive, this way. X minus 1 is positive plus 1.

What did I do here? I added and subtracted 1. Now what is this? This is the expected value of the remaining coin flips, until I obtain heads, given that the first one was tails.

It's the same story that we were going through down there. Given that the first coin flip was tails doesn't tell me anything about the future, about the remaining coin flips. So this expectation should be the same as the expectation faced by a person who was starting just now. So this should be equal to the expected value of X itself. And then we have the plus 1 that's come from there, OK?

Remaining coin flips until a head, given that I had a tail yesterday, is the same as expected number of flips until heads for a person just starting now and wasn't doing anything yesterday. So the fact that they I had a coin flip yesterday doesn't change my beliefs about how long it's going to take me until the first head. So once we believe that relation, than we plug this here. And this red term becomes expected value of X plus 1.

So now we didn't exactly get the answer we wanted, but we got an equation that involves the expected value of X. And it's the only unknown in that equation. Expected value of X equals to P plus (1 minus P) times this expression. You solve this equation for expected value of X, and you get the value of 1/P.

The final answer does make intuitive sense. If P is small, heads are difficult to obtain. So you expect that it's going to take you a long time until you see heads for the first time. So it is definitely a reasonable answer.

Now the trick that we used here, the divide and conquer trick, is a really nice one. It gives us a very good shortcut in this problem. But you must definitely spend some time making sure you understand why this expression here is the same as that expression there.

Essentially, what it's saying is that, if I tell you that X is bigger than 1, that the first coin flip was tails, all I'm telling you is that that person has wasted a coin flip, and they are starting all over again. So they've wasted 1 coin flip. And they're starting all over again. If I tell you that the first flip was tails, that's the only information that I'm basically giving you, a wasted flip, and then starts all over again.

All right, so in the few remaining minutes now, we're going to quickly introduce a few new concepts that we will be playing with in the next ten days or so. And you will get plenty of opportunities to manipulate them. So here's the idea.

A typical experiment may have several random variables associated with that experiment. So a typical student has height and weight. If I give you the PMF of height, that tells me something about distribution of heights in the class. I give you the PMF of weight, it tells me something about the different weights in this class.

But if I want to ask a question, is there an association between height and weight, then I need to know a little more how height and weight relate to each other. And the PMF of height individuality and PMF of weight just by itself do not tell me anything about those relations. To be able to say something about those relations, I need to know something about joint probabilities, how likely is it that certain X's go together with certain Y's. So these probabilities, essentially, capture associations between these two random variables. And it's the information I would need to have to do any kind of statistical study that tries to relate the two random variables with each other.

These are ordinary probabilities. This is an event. It's the event that this thing happens and that thing happens.

This is just the notation that we will be using. It's called the joint PMF. It's the joint Probability Mass Function of the two random variables X and Y looked at together, jointly. And it gives me the probability that any particular numerical outcome pair does happen.

So in the finite case, you can represent joint PMFs, for example, by a table. This particular table here would give you information such as, let's see, the joint PMF evaluated at 2, 3. This is the probability that X is equal to 3 and, simultaneously, Y is equal to 3. So it would be that number here. It's 4/20.

OK, what is a basic property of PMFs? First, these are probabilities, so all of the entries have to be non-negative. If you adopt the probabilities over all possible numerical pairs that you could get, of course, the total probability must be equal to 1. So that's another thing that we want.

Now suppose somebody gives me this model, but I don't care about Y's. All I care is the distribution of the X's. So I'm going to find the probability that X takes on a particular value.

Can I find it from the table? Of course, I can. If you ask me what's the probability that X is equal to 3, what I'm going to do is to add up those three probabilities together.

And those probabilities, taken all together, give me the probability that X is equal to 3. These are all the possible ways that the event X equals to 3 can happen. So we add these, and we get the 6/20.

What I just did, can we translate it to a formula? What did I do? I fixed the particular value of X. And I added up the values of the joint PMF over all the possible values of Y.

So that's how you do it. You take the joint. You take one slice of the joint, keeping X fixed, and adding up over the different values of Y.

The moral of this example is that, if you know the joint PMFs, then you can find the individual PMFs of every individual random variable. And we have a name for these. We call them the marginal PMFs.

We have the joint that talks about both together, and the marginal that talks about them one at the time. And finally, since we love conditional probabilities, we will certainly want to define an object called the conditional PMF.

So this quantity here is a familiar one. It's just a conditional probability. It's the probability that X takes on a particular value, given that Y takes a certain value.

For our example, let's take little y to be equal to 2, which means that we're conditioning to live inside this universe. This red universe here is the y equal to 2 universe. And these are the conditional probabilities of the different X's inside that universe.

OK, once more, just an exercise in notation. This is the chapter two version of the notation of what we were denoting this way in chapter one. The way to read this is that it's a conditional PMF having to do with two random variables, the PMF of X conditioned on information about Y. We are fixing a particular value of capital Y, that's the value on which we are conditioning. And we're looking at the probabilities of the different X's.

So it's really a function of two arguments, little x and little y. But the best way to think about it is to fix little y and think of it as a function of X. So I'm fixing little y here, let's say, to y equal to 2. So I'm considering only this.

And now, this quantity becomes a function of little x. For the different little x's, we're going to have different conditional probabilities. What are those conditional probabilities?

OK, conditional probabilities are proportional to original probabilities. So it's going to be those numbers, but scaled up. And they need to be scaled so that they add up to 1.

So we have 1, 3 and 1. That's a total of 5. So the conditional PMF would have the shape zero, 1/5, 3/5, and 1/5. This is the conditional PMF, given a particular value of Y. It has the same shape as those numbers, where by shape, I mean try to visualize a bar graph.

The bar graph associated with those numbers has exactly the same shape as the bar graph associated with those numbers. The only thing that has changed is the scaling. Big moral, let me say in different words, the conditional PMF, given a particular value of Y, is just a slice of the joint PMF where you maintain the same shape, but you rescale the numbers so that they add up to 1.

Now mathematically, of course, what all of this is doing is it's taking the original joint PDF and it rescales it by a certain factor. This does not involve X, so the shape, is a function of X, has not changed. We're keeping the same shape as a function of X, but we divide by a certain number. And that's the number that we need, so that the conditional probabilities add up to 1.

Now where does this formula come from? Well, this is just the definition of conditional probabilities. Probability of something conditioned on something else is the probability of both things happening, the intersection of the two divided by the probability of the conditioning event.

And last remark is that, as I just said, conditional probabilities are nothing different than ordinary probabilities. So a conditional PMF must sum to 1, no matter what you are conditioning on. All right, so this was sort of quick introduction into our new notation. But you get a lot of practice in the next days to come.

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013