

The following content is provided under a Creative Commons license. Your support will help MIT OpenCourseWare continue to offer high quality educational resources for free. To make a donation or view additional materials from hundreds of MIT courses, visit MIT OpenCourseWare at [ocw.mit.edu](http://ocw.mit.edu).

**PROFESSOR:** So today's agenda is to say a few more things about continuous random variables. Mainly we're going to talk a little bit about inference. This is a topic that we're going to revisit at the end of the semester. But there's a few things that we can already say at this point.

And then the new topic for today is the subject of derived distributions. Basically if you know the distribution of one random variable, and you have a function of that random variable, how to find a distribution for that function.

And it's a fairly mechanical skill, but that's an important one, so we're going to go through it. So let's see where we stand. Here is the big picture. That's all we have done so far. We have talked about discrete random variables, which we described by probability mass function. So if we have multiple random variables, we describe them with the a joint mass function.

And then we define conditional probabilities, or conditional PMFs, and the three are related according to this formula, which is, you can think of it either as the definition of conditional probability. Or as the multiplication rule, the probability of two things happening is the product of the probabilities of the first thing happening, and then the second happening, given that the first has happened.

There's another relation between this, which is the probability of  $x$  occurring, is the sum of the different probabilities of the different ways that  $x$  may occur, which is in conjunction with different values of  $y$ . And there's an analog of all that in the continuous world, where all you do is to replace  $p$ 's by  $f$ 's, and replace sums by integrals. So the formulas all look the same. The interpretations are a little more subtle, so the  $f$ 's are not probabilities, they're probability densities. So they're probabilities per unit length, or in the case of joint PDF's, these are probabilities per unit area. So they're densities of some sort.

Probably the more subtle concept to understand what it really is the conditional density. In some sense, it's simple. It's just the density of  $X$  in a world where you have been told the value of the random variable  $Y$ . It's a function that has two arguments, but the best way to think about it is to say that we fixed  $y$ . We're told the value of the random variable  $Y$ , and we look at it as a function of  $x$ . So as a function of  $x$ , the denominator is a constant, and it just looks like the joint density. when we keep  $y$  fixed. So it's really a function of one argument, just the argument  $x$ . And it has the same shape as the joint's density when you take that slice of it.

So conditional PDFs are just slices of joint PDFs.

There's a bunch of concepts, expectations, variances, cumulative distribution functions that apply equally well for to both universes of discrete or continuous random variables. So why is probability useful? Probability is useful because, among other things, we use it to make sense of the world around us. We use it to make inferences about things that we do not see directly. And this is done in a very simple manner using the base rule. We've already seen some of that, and now we're going to revisit it with a bunch of different variations.

**And the variations** come because sometimes our random variable are discrete, sometimes they're continuous, or we can have a combination of the two. So the big picture is that there's some unknown random variable out of there, and we know the distribution that's random variable. And in the discrete case, it's going to be given by PMF. In the continuous case, it's given a PDF. Then we have some phenomenon, some noisy phenomenon or some measuring device, and that measuring device produces observable random variables  $Y$ .

We don't know what  $x$  is, but we have some beliefs about how  $X$  is distributed. We observe the random variable  $Y$ . We need a model of this box. And the model of that box is going to be either a PMF, for the random variable  $Y$ . And that model tells us, if the true state of the world is  $X$ , how do we expect to  $Y$  to be distributed? That's for the case where  $Y$  is this discrete. If  $Y$  is a continuous, you might instead have a density for  $Y$ , or something of that form.

So in either case, this should be a function that's known to us. This is our model of the measuring device. And now having observed  $y$ , we want to make inferences about  $x$ . What does it mean to make inferences? Well the most complete answer in the inference problem is to tell me the probability distribution of the unknown quantity.

But when I say the probability distribution, I don't mean this one. I mean the probability distribution that takes into account the measurements that you got. So the output of an inference problem is to come up with the distribution of  $X$ , the unknown quantity, given what we have already observed. And in the discrete case, it would be an object like that. If  $X$  is continuous, it would be an object of this kind.

OK, so we're given conditional probabilities of this type, and we want to get conditional distributions of the opposite type where the order of the conditioning is being reversed. So the starting point is always a formula such as this one. The probability of  $x$  happening, and then  $y$  happening given that  $x$  happens. This is the probability that a particular  $x$  and  $y$  happen simultaneously.

But this is also equal to the probability that  $y$  happens, and then that  $x$  happens, given that  $y$  has happened. And you take this expression and send one term to the denominator of the other side, and this gives us the base rule for the discrete case. Which is this one that you have already seen, and you have played with it.

So this is what the formula looks like in the discrete case. And the typical example where both random variables are discrete is the one we discussed some time ago.  $X$  is, let's say, a binary variable, or whether an airplane is present up there or not.  $Y$  is a discrete measurement, for example, whether our radar beeped or it didn't beep. And we make inferences and calculate the probability that the plane is there, or the probability that the plane is not there, given the measurement that we have made.

And of course  $X$  and  $Y$  do not need to be just binary. They could be more general discrete random variables. So how does the story change in the continuous case? First, what's a possible application of the continuous case? Well, think of  $X$  as being some signal that takes values over a continuous range. Let's say  $X$  is the current through a resistor. And then you have some measuring device that measures currents, but that device is noisy, it gets hit, let's say for example, by Gaussian noise.

And the  $Y$  that you observe is a noisy version of  $X$ . But your instruments are analog, so you measure things on a continuous scale. What are you going to do in that case? Well the inference problem, the output of the inference problem, is going to be the conditional distribution of  $X$ . What do you think your current is based on a particular value of  $Y$  that you have observed?

So the output of our inference problem is, given the specific value of  $Y$ , to calculate this entire function as a function of  $x$ , and then go and plot it. How do we calculate it? You go through the same calculation as in the discrete case, except that all of the  $x$ 's gets replaced by  $p$ 's. In the continuous case, it's equally true that the joint's density is the product of the marginal density with the conditional density. So the formula is still valid with just a little change of notation. So we end up with the same formula here, except that we replace  $x$ 's with  $p$ 's.

So all of these functions are known to us. We have formulas for them. We fix a specific value of  $y$ , we plug it in, so we're left with a function of  $x$ . And that gives us the posterior distribution. Actually there's also a denominator term that's not necessarily given to us, but we can always calculate it if we have the marginal of  $X$ , and we have the model for measuring device. Then we can always find the marginal distribution of  $Y$ . So this quantity, that number, is in general a known one, as well, and doesn't give us any problems.

So to complicate things a little bit, we can also look into situations where our two random variables are of different kinds. For example, one random variable could be discrete, and the other it might be continuous. And there's two versions. Here one version is when  $X$  is discrete, but  $Y$  is continuous. What's an example of this?

Well suppose that I send a single bit of information so my  $X$  is 0 or 1. And what I measure is  $Y$ , which is  $X$  plus, let's say, Gaussian noise. This is the standard example that shows up in any textbook on communication, or signal processing. You send a single bit, but what you observe is a noisy version of that bit.

You start with a model of your  $x$ 's. These would be your prior probabilities. For example, you might believe that either 0 or 1 are equally likely, in which case your PMF gives equal weight to two possible values. And then we need a model of our measuring device. This is one specific model. The general model would have a shape such as follows.  $Y$  has a distribution, its density. And that density, however, depends on the value of  $X$ .

So when  $x$  is 0, we might get a density of this kind. And when  $x$  is 1, we might get the density of a different kind. So these are the conditional densities of  $y$  in a universe that's specified by a particular value of  $x$ .

And then we go ahead and do our inference. OK, what's the right formula for doing this inference? We need a formula that's sort of an analog of this one, but applies to the case where we have two random variables of different kinds. So let me just redo this calculation here. Except that I'm not going to have a probability of taking specific values. It will have to be something a little different. So here's how it goes.

Let's look at the probability that  $X$  takes a specific value that makes sense in the discrete case, but for the continuous random variable, let's look at the probability that it takes values in some little interval. And now this probability of two things happening, I'm going to write it as a product. And I'm going to write this as a product in two different ways. So one way is to say that this is the probability that  $X$  takes that value and then given that  $X$  takes that value, the probability that  $Y$  falls inside that interval.

So this is our usual multiplication rule for multiplying probabilities, but I can use the multiplication rule also in a different way. It's the probability that  $Y$  falls in the range of interest. And then the probability that  $X$  takes the value of interest given that  $Y$  satisfies the first condition. So this is something that's definitely true. We're just using the multiplication rule. And now let's translate it into PMF or PDF notation.

So the entry up there is the PMF of  $X$  evaluated at  $x$ . The second entry, what is it? Well probabilities of little intervals are given to us by densities. But we are in the conditional universe where  $X$  takes on a particular value. So it's going to be the density of  $Y$  given the value of  $X$  times  $\delta$ . So probabilities of little intervals are given by the density times the length of the little interval, but because we're working in the conditional universe, it has to be the conditional density.

Now let's try the second expression. This is the probability that the  $Y$  falls into the little interval. So that's the density of  $Y$  times  $\delta$ . And then here we have an object which is the conditional probability  $X$  in a universe where the value of  $Y$  is given to us.

Now this relation is sort of approximate. This is true for very small  $\delta$  in the limit. But we can cancel the  $\delta$ s from both sides, and we're left with a formula that links together PMFs and PDFs. Now this may look terribly confusing because there's both  $p$ 's and  $f$ 's involved. But the logic should be clear. If a random variable is discrete, it's described by PMF. So here we're talking about the PMF of  $X$  in some particular universe.  $X$  is discrete, so it has a PMF.

Similarly here.  $Y$  is continuous so it's described by a PDF. And even in the conditional universe where I tell you the value of  $X$ ,  $Y$  is still a continuous random variable, so it's been described by a PDF. So this is the basic relation that links together PMF and PDFs. In this mixed world. And now in this inequality, you can take this term and send it to the new denominator to the other side. And what you end up with is the formula that we have up here.

And this is a formula that we can use to make inferences about the discrete random variable  $X$  when we're told the value of the continuous random variable  $Y$ . The probability that  $X$  takes on a particular value has something to do with the prior. And other than that, it's proportional to this quantity, the conditional of  $Y$  given  $X$ . So these are the quantities that we plotted here.

Suppose that the  $x$ 's are equally likely in your prior, so we don't really care about that term. It tells us that the posterior of  $X$  is proportional to that particular density under the given  $x$ 's. So in this picture, if I were to get a particular  $y$  here, I would say that  $x = 1$  has a probability that's proportional to this quantity.  $x = 0$  has a probability that's proportional to this quantity.

So the ratio of these two quantities gives us the relative odds of the different  $x$ 's given the  $y$  that we have observed.

So we're going to come back to this topic and redo plenty of examples of these kinds towards the end of the class, when we spend some time dedicated to inference problems. But already at this stage, we sort of have the basic skills to deal with a lot of that. And it's useful at this point to pull all the formulas together.

So finally let's look at the last case that's remaining. Here we have a continuous phenomenon that we're trying to measure, but our measurements are discrete. What's an example where this might happen?

So you have some device that emits light, and you drive it with a current that has a certain intensity. You don't know what that current is, and it's a continuous random variable. But the device emits light by sending out individual photons. And your measurement is some other device that counts how many photons did you get in a single second.

So if we have devices that emit a very low intensity you can actually start counting individual photons as they're being observed. So we have a discrete measurement, which is the number of problems, and we have a continuous hidden random variable that we're trying to estimate. What do we do in this case?

Well we start again with a formula of this kind, and send the  $p$  term to the denominator. And that's the formula that we use there, except that the roles of  $x$ 's and  $y$ 's are interchanged. So since here we have  $Y$  being discrete, we should change all the subscripts. It would be  $p_Y f_X$  given  $y f_X$ , and  $P(Y \text{ given } X)$ . So just change all those subscripts. Because now what we're used to be continuous became discrete, and vice versa.

Take that formula, send the other terms to the denominator, and we have a formula for the density, or  $X$ , given the particular measurements for  $Y$  that we have obtained.

In some sense that's all there is in Bayesian inference. It's using these very simple one line formulas. But why are there people then who make their living solving inference problems? Well, the devil is in the details. As we're going to discuss, there are some real world issues of how exactly do you design your  $f$ 's, how do you model your system, then how do you do your calculations.

This might not be always easy. For example, there's certain integrals or sums that have to be evaluated, which may be hard to do and so on. So this object is a lot of richer than just these formulas. On the other hand, at the conceptual level, that's the basis for Bayesian inference, that these are the basic concepts.

All right, so now let's change gear and move to the new subject, which is the topic of finding the distribution of a functional for a random variable. We call those distributions derived distributions, because we're given the distribution of  $X$ . We're interested in a function of  $X$ . We want to derive the distribution of that function based on the distribution that we already know.

So it could be a function of just one random variable. It could be a function of several random variables. So one example that we are going to solve at some point, let's say you have to run the variables  $X$  and  $Y$ . Somebody tells you their distribution, for example, is a uniform of the square. For some reason, you're interested in the ratio of these two random variables, and you want to find the distribution of that ratio.

You can think of lots of cases where your random variable of interest is created by taking some other unknown variables and taking a function of them. And so it's legitimate to care about the distribution of that random variable.

A caveat, however. There's an important case where you don't need to find the distribution of that random variable. And this is when you want to calculate the expectations. If all you care about is the expected value of this function of the random variables, you can work directly with the distribution of the original random variables without ever having to find the PDF of  $g$ .

So you don't do unnecessary work if it's not needed, but if it's needed, or if you're asked to do it, then you just do it.

So how do we find the distribution of the function? As a warm-up, let's look at the discrete case. Suppose that  $X$  is a discrete random variable and takes certain values. We have a function  $g$  that maps  $x$ 's into  $y$ 's. And we want to find the probability mass function for  $Y$ .

So for example, if I'm interested in finding the probability that  $Y$  takes on this particular value, how would they find it? Well I ask, what are the different ways that these particular  $y$  value can happen? And the different ways that it can happen is either if  $x$  takes this value, or if  $X$  takes that value. So we identify this event in the  $y$  space with that event in the  $x$  space. These two events are identical.  $X$  falls in this set if and only if  $Y$  falls in that set.

Therefore, the probability of  $Y$  falling in that set is the probability of  $X$  falling in that set. The probability of  $X$  falling in that set is just the sum of the individual probabilities of the  $x$ 's in this set. So we just add the probabilities of the different  $x$ 's where the summation is taken over all  $x$ 's that leads to that particular value of  $y$ .

Very good. So that's all there is in the discrete case. It's a very nice and simple. So let's transfer these methods to the continuous case.

Suppose we are in the continuous case. Suppose that  $X$  and  $Y$  now can take values anywhere. And I try to use same methods and I ask, what is the probability that  $Y$  is going to take this value? At least if the diagram is this way, you would say this is the same as the probability that  $X$  takes this value. So I can find the probability of  $Y$  being this in terms of the probability of  $X$  being that.

Is this useful? In the continuous case, it's not. Because in the continuous case, any single value has 0 probability. So what you're going to get out of this argument is that the probability  $Y$  takes this value is 0, is equal to the probability that  $X$  takes that value which also 0.

That doesn't help us. We want to do something more. We want to actually find, perhaps, the density of  $Y$ , as opposed to the probabilities of individual  $y$ 's. So to find the density of  $Y$ , you might argue as follows. I'm looking at an interval for  $y$ , and I ask what's the probability of falling in this interval. And you go back and find the corresponding set of  $x$ 's that leads to those  $y$ 's, and equate those two probabilities.

The probability of all of those  $y$ 's collectively should be equal to the probability of all of the  $x$ 's that map into that interval collectively. And this way you can relate the two.

As far as the mechanics go, in many cases it's easier to not to work with little intervals, but instead to work with cumulative distribution functions that used to work with sort of big intervals. So you can instead do a different picture. Look at this set of  $y$ 's. This is the set of  $y$ 's that are smaller than a certain value. The probability of this set is given by the cumulative distribution of the random variable  $Y$ .

Now this set of  $y$ 's gets produced by some corresponding set of  $x$ 's. Maybe these are the  $x$ 's that map into  $y$ 's in that set. And then we argue as follows. The probability that the  $Y$  falls in this interval is the same as the probability that  $X$  falls in that interval. So the event of  $Y$  falling here and the event of  $X$  falling there are the same, so their probabilities must be equal. And then I do the calculations here. And I end up getting the cumulative distribution function of  $Y$ . Once I have the cumulative, I can get the density by just differentiating.

So this is the general cookbook procedure that we will be using to calculate it derived distributions.

We're interested in a random variable  $Y$ , which is a function of the  $x$ 's. We will aim at obtaining the cumulative distribution of  $Y$ . Somehow, manage to calculate the probability of this event. Once we get it, and what I mean by get it, I don't mean getting it for a single value of little  $y$ . You need to get this for all little  $y$ 's. So you need to get the function itself, the cumulative distribution. Once you get it in that form, then you can calculate the derivative at any particular point. And this is going to give you the density of  $Y$ .

So a simple two-step procedure. The devil is in the details of how you carry the mechanics. So let's do one first example. Suppose that  $X$  is a uniform random variable, takes values between 0 and 2. We're interested in the random variable  $Y$ , which is the cube of  $X$ . What kind of distribution is it going to have?

Now first notice that  $Y$  takes values between 0 and 8. So  $X$  is uniform, so all the  $x$ 's are equally likely. You might then say, well, in that case, all the  $y$ 's should be equally likely. So  $Y$  might also have a uniform distribution. Is this true? We'll find out.

So let's start applying the cookbook procedure. We want to find first the cumulative distribution of the random variable  $Y$ , which by definition is the probability that the random variable is less than or equal to a certain number. That's what we want to find. What we have in our hands is the distribution of  $X$ . That's what we need to work with. So the first step that you need to do is to look at this events and translate it, and write it in terms of the random variable about which you know you have information.

So  $Y$  is  $X$  cubed, so this event is the same as that event. So now we can forget about the  $y$ 's. It's just an exercise involving a single random variable with a known distribution and we want to calculate the probability of some event.

So we're looking at this event.  $X$  cubed being less than or equal to  $Y$ . We massage that expression so that's it involves  $X$  directly, so let's take cubic roots of both sides of this inequality. This event is the same as the event that  $X$  is less than or equal to  $Y$  to the  $1/3$ . Now with a uniform distribution on  $[0,2]$ , what is that probability going to be?

It's the probability of being in the interval from 0 to  $y$  to the  $1/3$ , so it's going to be in the area under the uniform going up to that point. And what's the area under that uniform?

So here's  $x$ . Here is the distribution of  $X$ . It goes up to 2. The distribution of  $X$  is this one. We want to go up to  $y$  to the  $1/3$ . So the probability for this event happening is this area. And the area is equal to the base, which is  $y$  to the  $1/3$  times the height. What is the height?

Well since the density must integrate to 1, the total area under the curve has to be 1. So the height here is  $1/2$ , and that explains why we get the  $1/2$  factor down there.

So that's the formula for the cumulative distribution. And then the rest is easy. You just take derivatives. You differentiate this expression with respect to  $y$   $1/2$  times  $1/3$ , and  $y$  drops by one power. So you get  $y$  to  $2/3$  in the denominator.

So if you wish to plot this, it's  $1/y$  to the  $2/3$ . So when  $y$  goes to 0, it sort of blows up and it goes on this way. Is this picture correct the way I've drawn it? What's wrong with it?

**[? AUDIENCE:]** Something. ?]

**PROFESSOR:** Yes.  $y$  only takes values from 0 to 8. This formula that I wrote here is only correct when the preview picture applies. I took my  $y$  to the  $1/3$  to be between 0 and 2. So this formula here is only correct for  $y$  between 0 and 8. And for that reason, the formula for the derivative is also true only for a  $y$  between 0 and 8. And any other values of  $y$  are impossible, so they get zero density. So to complete the picture here, the PDF of  $y$  has a cut-off of 8, and it's also 0 everywhere else.

And one thing that we see is that the distribution of  $Y$  is not uniform. Certain  $y$ 's are more likely than others, even though we started with a uniform random variable  $X$ .

All right. So we will keep doing examples of this kind, a sequence of progressively more interesting or more complicated. So that's going to continue in the next lecture. You're going to see plenty of examples in your recitations and tutorials and so on. So let's do one that's pretty similar to the one that we did, but it's going to add to just a small twist in how we do the mechanics.

OK so you set your cruise control when you start driving. And you keep driving at the constants based at the constant speed. Where you set your cruise control is somewhere between 30 and 60. You're going to drive a distance of 200. And so the time it's going to take for your trip is 200 over the setting of your cruise control. So it's  $200/V$ .

Somebody gives you the distribution of  $V$ , and they tell you not only it's between 30 and 60, it's roughly equally likely to be anything between 30 and 60, so we have a uniform distribution over that range. So we have a distribution of  $V$ . We want to find the distribution of the random variable  $T$ , which is the time it takes till your trip ends.

So how are we going to proceed? We'll use the exact same cookbook procedure. We're going to start by finding the cumulative distribution of  $T$ . What is this? By definition, the cumulative distribution is the probability that  $T$  is less than a certain number. OK. Now we don't know the distribution of  $T$ , so we cannot to work with these event directly. But we take that event and translate it into  $T$ -space. So we replace the  $t$ 's by what we know  $T$  to be in terms of  $V$  or the  $v$ 's

All right. So we have the distribution of  $V$ . So now let's calculate this quantity. OK. Let's massage this event and rewrite it as the probability that  $V$  is larger or equal to  $200/T$ .

So what is this going to be? So let's say that  $200/T$  is some number that falls inside the range. So that's going to be true if  $200/T$  is bigger than 30, and less than 60. Which means that  $t$  is less than  $30/200$ . No,  $200/30$ . And bigger than  $200/60$ . So for  $t$ 's inside that range, this number  $200/t$  falls inside that range. This is the range of  $t$ 's that are possible, given the description of the problem the we have set up.

So for  $t$ 's in that range, what is the probability that  $V$  is bigger than this number? So  $V$  being bigger than that number is the probability of this event, so it's going to be the area under this curve. So the area under that curve is the height of the curve, which is  $1/3$  over 30 times the base. How big is the base? Well it's from that point to 60, so the base has a length of  $60$  minus  $200/t$ .

And this is a formula which is valid for those  $t$ 's for which this picture is correct. And this picture is correct if  $200/T$  happens to fall in this interval, which is the same as  $T$  falling in that interval, which are the  $t$ 's that are possible.

So finally let's find the density of  $T$ , which is what we're looking for. We find this by taking the derivative in this expression with respect to  $t$ . We only get one term from here. And this is going to be  $200/30$ ,  $1$  over  $t$  squared.



And this is the formula for the density for  $t$ 's in the allowed to range. OK, so that's the end of the solution to this particular problem as well. I said that there was a little twist compared to the previous one. What was the twist? Well the twist was that in the previous problem we dealt with the  $X$  cubed function, which was monotonically increasing. Here we dealt with the function that was monotonically decreasing. So when we had to find the probability that  $T$  is less than something, that translated into an event that  $V$  was bigger than something. Your time is less than something if and only if your velocity is bigger than something.

So for when you're dealing with the monotonically decreasing function, at some point some inequalities will have to get reversed.

Finally let's look at a very useful one. Which is the case where we take a linear function of a random variable. So  $X$  is a random variable with given distribution, and we can see there is a linear function. So in this particular instance, we take  $a$  to be equal to 2 and  $b$  equal to 5. And let us first argue just by picture.

So  $X$  is a random variable that has a given distribution. Let's say it's this weird shape here. And  $x$  ranges from -1 to +2. Let's do things one step at the time. Let's first find the distribution of  $2X$ . Why do you think you know about  $2X$ ? Well if  $x$  ranges from -1 to 2, then the random variable  $X$  is going to range from -2 to +4. So that's what the range is going to be.

Now dealing with the random variable  $2X$ , as opposed to the random variable  $X$ , in some sense it's just changing the units in which we measure that random variable. It's just changing the scale on which we draw and plot things. So if it's just a scale change, then intuition should tell you that the random variable  $X$  should have a PDF of the same shape, except that it's scaled out by a factor of 2, because our random variable of  $2X$  now has a range that's twice as large.

So we take the same PDF and scale it up by stretching the  $x$ -axis by a factor of 2. So what does scaling correspond to in terms of a formula? So the distribution of  $2X$  as a function, let's say, a generic argument  $z$ , is going to be the distribution of  $X$ , but scaled by a factor of 2.

So taking a function and replacing its arguments by the argument over 2, what it does is it stretches it by a factor of 2. You have probably been tortured ever since middle school to figure out when need to stretch a function, whether you need to put  $2z$  or  $z/2$ . And the one that actually does the stretching is to put the  $z/2$  in that place. So that's what the stretching does.

Could that to be the full answer? Well there's a catch. If you stretch this function by a factor of 2, what happens to the area under the function? It's going to get doubled. But the total probability must add up to 1, so we need to do something else to make sure that the area under the curve stays to 1. So we need to take that function and scale it down by this factor of 2.

So when you're dealing with a multiple of a random variable, what happens to the PDF is you stretch it according to the multiple, and then scale it down by the same number so that you preserve the area under that curve. So now we found the distribution of  $2X$ .

How about the distribution of  $2X + 5$ ? Well what does adding 5 to random variable do? You're going to get essentially the same values with the same probability, except that those values all get shifted by 5. So all that you need to do is to take this PDF here, and shift it by 5 units. So the range used to be from -2 to 4. The new range is going to be from 3 to 9. And that's the final answer. This is the distribution of  $2X + 5$ , starting with this particular distribution of  $X$ .

Now shifting to the right by  $b$ , what does it do to a function? Shifting to the right to by a certain amount, mathematically, it corresponds to putting  $-b$  in the argument of the function. So I'm taking the formula that I had here, which is the scaling by a factor of  $a$ . The scaling down to keep the total area equal to 1. And then I need to introduce this extra term to do the shifting.

So this is a plausible argument. The proof by picture that this should be the right answer. But just in order to keep our skills tuned and refined, let us do this derivation in a more formal way using our two-step cookbook procedure. And I'm going to do it under the assumption that  $a$  is positive, as in the example that's we just did.

So what's the two-step procedure? We want to find the cumulative of  $Y$ , and after that we're going to differentiate. By definition the cumulative is the probability that the random variable takes values less than a certain number. And now we need to take this event and translate it, and express it in terms of the original random variables.

So  $Y$  is, by definition,  $aX + b$ , so we're looking at this event. And now we want to express this event in a clean form where  $X$  shows up in a straight way. Let's say I'm going to massage this event and write it in this form. For this inequality to be true,  $x$  should be less than or equal to  $(y - b) / a$ .

OK, now what is this? This is the cumulative distribution of  $X$  evaluated at the particular point. So we got a formula for the cumulative  $Y$  based on the cumulative of  $X$ . What's the next step? Next step is to take derivatives of both sides. So the density of  $Y$  is going to be the derivative of this expression with respect to  $y$ . OK, so now here we need to use the chain rule. It's going to be the derivative of the  $F$  function with respect to its argument. And then we need to take the derivative of the argument with respect to  $y$ .

What is the derivative of the cumulative? The derivative of the cumulative is the density itself. And we evaluate it at the point of interest. And then the chain rule tells us that we need to take the derivative of this with respect to  $y$ , and the derivative of this with respect to  $y$  is  $1/a$ . And this gives us the formula which is consistent with what I had written down here, for the case where  $a$  is a positive number.

What if  $a$  was a negative number? Could this formula be true? Of course not. Densities cannot be negative, right? So that formula cannot be true. Something needs to change. What should change? Where does this argument break down when  $a$  is negative?

So when I write this inequality in this form, I divide by  $a$ . But when you divide by a negative number, the direction of an inequality is going to change. So when  $a$  is negative, this inequality becomes larger than or equal to. And in that case, the expression that I have up there would change when this is larger than here. Instead of getting the cumulative, I would get 1 minus the cumulative of  $(y - b) / a$ .

So this is the probability that  $X$  is bigger than this particular number. And now when you take the derivatives, there's going to be a minus sign that shows up. And that minus sign will end up being here. And so we're taking the negative of a negative number, and that basically is equivalent to taking the absolute value of that number.

So all that happens when we have a negative  $a$  is that we have to take the absolute value of the scaling factor instead of the factor itself.

All right, so this general formula is quite useful for dealing with linear functions of random variables. And one nice application of it is to take the formula for a normal random variable, consider a linear function of a normal random variable, plug into this formula, and what you will find is that  $Y$  also has a normal distribution. So using this formula, now we can prove a statement that I had made a couple of lectures ago, that a linear function of a normal random variable is also linear. That's how you would prove it. I think this is it for today so.