[SQUEAKING]

[RUSTLING]

[CLICKING]

SARA ELLISON: And today we continue. We just started with functions of random variables last time. That's what we'll do all day today. And then on Monday, we'll continue with functions of random variables and also start talking about moments of distributions.

And in the recitation, as I always say, we're flexible. We want to hear from you, what you'd like to have covered in the recitation, but the basic idea or the basic plan for this week will probably be to go over the problem set questions. So was it super hard?

AUDIENCE: Yeah.

SARA ELLISON: OK. Was it mostly the scraping exercise that was very difficult, time consuming? Yeah? Yeah.

AUDIENCE: I think if you know there's a server you could use, it would have been helpful. So for example, when I just tried it on my computer instead of on the server, the internet connection would be lost. And then because it took like four hours to run, I'd have to do it again. So I learned that there was a server that doesn't crash that I could run it on.

SARA ELLISON: Yes, that's good for us to know too because I can't say that we knew that, or at least certainly I didn't know that ahead of time, but good.

AUDIENCE: It-- In that case, it was fine.

SARA ELLISON: OK. So I do want to make a couple of comments about the problem set. So this is a heterogeneous bunch. There are some people who are here and have had probability courses before. Probability is not a prerequisite, so a number of you have never had probability before. There are some of you have had lots of coding experience, some of you who have had no coding experience.

And so we're trying to reach a balance. We want the people who feel comfortable with the theoretical part to stretch themselves and learn some of the data skills. We want the people who are more familiar with the data skills part to really understand where these statistics are coming from from a theoretical perspective. So it is a balancing act.

And since this is a new course, sometimes we miscalibrate things a little bit, but we're very happy to get your feedback. Let us know if things are overwhelming. We're happy to make adjustments along the way.

And one adjustment that I did want to propose was that we could change the problem sets. We'll still have six problem sets, but we could change them so that we'll only count your top four scores on the problem sets, and so that might take a little bit of the pressure off. Is that amenable to everyone? OK, good. So we'll go ahead.

We'll renormalize them, so the problem sets will still count for as much, but we'll just count your top four scores. Yep.

AUDIENCE: Also, I basically learned a lot from doing the problem set.

SARA ELLISON: You did.

AUDIENCE: Yeah, that was a good thing.

SARA ELLISON: Good, yes. So in case you didn't hear, he said that he learned a lot from doing the problem set and that's the whole point, so I'm happy to hear that. Great. OK, good.

So what I want to do today to start out is to talk a little bit about the empirical project. And we'll talk more about it as the semESTHER progresses, we'll fill in a little bit of the details, but I think I want to spend 10 minutes at the beginning to of lecture today to talk about how we view the empirical project, what the objectives of it are, what the structure of it is going to be, what the expectations are going to be, et cetera. And quite happy to take questions, if any arise about the empirical project.

So as it says in the syllabus, and I've reminded you here, it's going to be due Monday, May 2nd, so that's pretty near the end of the semESTHER. The idea behind the empirical project is twofold-- we want to help you improve your practical skills in data gathering, data manipulation, data analysis, et cetera, but we also just want to let you work in a concentrated way on something that interests you. So we try to pull problem set questions and examples in class from lots of different areas, but we know that you all come into this class with a set of interests, and we want you to be able to spend a little time working on what you're interested in.

And so I guess what our view of the project is that there really could be two different models, and you can choose which model is more, I don't know, you're more comfortable with. So I'll describe the two models.

The basic idea is one model for this project is that you can start with a data set that you find interesting, and just roll up your sleeves, and get your hands dirty, and find out lots of interesting things about the data set. And the second is to start with a question, start with maybe an economic question, or a political question, or something like that that is answerable with data, and then go out and try to gather data, perhaps from various sources, to try to answer your question.

So let me fill in some of the details and give you some examples. So let's suppose you wanted to write an empirical project of the first type, where you're starting with a data set. So the idea here is that you scrape, or you download, or you gather, or you otherwise obtain a data set that you find interesting.

There may be variables that you need to generate that aren't in your data set. I'll give you an example in just a second. You will then create summary statistics and perform diagnostics on the data set, and we want you to do that both numerically and graphically. We want you to graph the data, we want you to make histograms, we want you to compute various statistics involving the data, and it may make sense for you to only focus on part of your data set. If there's enough interesting stuff going on in just a subset of variables, that's perfectly fine.

When you're investigating your data set, look for aspects that are unusual, or unexpected, or particularly important, and document those carefully. These aspects could include unexpected correlations that you run across, unusual distributions of particular variables, surprising summary statistics.

And then the final output is a three to five page I'll call it data section on your data and findings, and the idea here is just to describe in detail how you collected the data, where the data came from, what the structure of the data set is, and then present these interesting or unusual aspects of the data. And we'll be more specific closer to the date about what the write up should include. So right now just think three to five pages with lots of detail about the data set, but we can be more specific later.

So let me give you an example. Let's suppose you had a particular interest in the papacy. And so you went back and you gathered data on popes going back to the late 1300s, which was when the popes move back to Rome after being in France. And so you cared about how old they were when they were elected, how long they lived after election. For the most part, that corresponds to the length of the pontificate because almost always they serve until death. And you gather all of this data about the papacy.

Then you start digging into the data and you notice-- and you also gather data on, you may have to create a new variable on the birthplace of the pope and the distance of the birthplace to Rome. So you create that variable because you think there might be something interesting in there, and you find out that, in fact, there's a positive correlation between the distance of the pope's birthplace from Rome and the age at which he was elected. So maybe not something you would have expected.

Then you document that papal longevity has been increasing over the years. Not particularly surprising, but maybe an important thing to note. And then you decide that you look at some kind of maybe a standardized distribution or histogram of the density of papal lifespans and it looks like a Weibull distribution, perhaps, and so you fit a Weibull distribution to that density and use it to estimate things like the probability that Pope Francis's pontificate will last for more than 10 years.

And then the write up would include all of this information. So that's one model that you can pursue for your empirical project. Questions? No. Do keep in mind, if you ask more questions, that slows me down, so I don't know if that's a good thing or a bad thing, but do keep in mind that there is that direct causality.

OK let's suppose instead you want to do the second type of project, where you have a question that you want to answer with some data. So the idea here is that you articulate a question, and we can help you articulate the question, but you articulate a question which you're likely to be able to answer with some available data. Pardon me?

ESTHER: Always some data they can obtain, scrape--

SARA ELLISON: Oh, right, exactly. Yes, so available data, meaning data that either you have, or you can scrape, or you can download, or you can borrow, or et cetera.

So then you'll have to perform a search, perhaps using resources that ESTHER mentioned, for relevant data sets. And it may be that you can answer your question with data that you download entirely from the Department of Energy's website, or maybe you have to go to the Census website, and the Department of Energy's website, and the NBER website to get some other data set or something like that.

After obtaining the relevant data, then you'll perform an analysis that allows you to answer your question. The analysis could be a comparison of means, it could be a linear regression, it could be something else, and it's probably going to be coupled with some hypothesis testing. So this is basically how social scientists try to answer questions is they build statistical models, and then they perform hypothesis tests on the parameters of those models, and that's how they come up with empirical answers to their questions. And then, again, you'll write a three to five page report describing your question and how your analysis helped to answer it.

So in this second type of project, the focus, although you'll be doing many of the same things as the first type of project, the focus will be more on your question and the analysis that you use to answer it. Maybe a little bit less focus on the structure of the data set and interesting things that you find about the data set.

So let me give you an example of a-- oh, and I will also mention we'll be more specific about this write up as well. So let me walk you through an example of this kind of project.

Let's suppose you're interested in knowing whether hosting the Olympics is an economic benefit or a burden to a city. So you go on Wikipedia and you know that Wikipedia lists all the cities that have hosted the Olympics in the modern era. But in addition, they also list all the finalist cities, the ones that came close to winning the Olympic bid, but didn't quite make it.

So you have, in some sense, a group that was treated, meaning they hosted the Olympics, and a control group, ones that came close to hosting the Olympics, but didn't quite make it. And so you want to see what the effect of that treatment of hosting the Olympics was on the cities.

So then you go gather data on economic outcomes of these two groups of cities, say, budget deficits and unemployment rates. Maybe you can think of other interesting outcome measures. You might have to gather those from different sources.

But you gather those for all the host and the finalist cities and say, you don't know how long the economic impacts of it might last, so you gather them five years after the Olympics and 10 years after the Olympics. And then you compare the outcomes of the host cities and the finalist cities, so you could just be doing a comparison of means.

It could be that there are other city characteristics you would like to control for in your analysis, and we'll be very specific about how to do that later on in the semester. And so you might want to perform regressions instead of just comparisons of means, but that's the idea here. Yes.

AUDIENCE: My question is more along the lines of how far into the rabbit hole should we go in the sense that, if we're doing these comparisons of cities, there are any number of different factors that could impact. Let's say you're looking at Beijing versus London and there are different growth trends for the cities, so should you control for other cities within the nation other than the ones that you're comparing specifically to the Olympics to see whether or not the trends are valid.

SARA ELLISON: So that's an excellent question, but do keep in mind that your write up is going to be three to five pages long. And we can give you guidance along the way about exactly how sophisticated and how in-depth the analysis has to be, but when you're limited to a three to five page write up, that suggests that the analysis is going to be fairly simple because you're just not going to be able to explain everything you've done and all of the data gathering.

I don't mean to suggest that the issues that you raise are not important ones, but they might be beyond the scope of what we're expecting for this paper, if that makes sense.

ESTHER: It might be a good idea to have a discussion section at the end, which includes these kind of doubts you might have had on your own analysis [INAUDIBLE]. This is not meant to be your Nobel Prize address yet, but if you are able to say, this is what the next step would be to make the analysis more robust or more informative, this is what I would do.

SARA ELLISON: Yes. Yeah, I'm glad Esther brought that up, because that's an important thing to keep in mind that no empirical paper is perfect. These papers in particular are not going to be perfect. And often a very important piece of doing a project like this is recognizing what the shortcomings are and discussing them in the paper even if you can't address them directly. Yes.

AUDIENCE: So if we are already working on an empirical project, is it possible to just repeat a part of that project.

SARA ELLISON: Yeah, so I think we would be amenable to people bringing in data sets or projects from research or other classes. We would want to have a conversation with you and make sure that what you're doing wasn't exactly what you were doing, say, for another class. But absolutely, we want to encourage you to, if you have a data set, you're interested in it, you think that there are interesting questions you can answer with it, we want to encourage you to be able to use that. Yeah, absolutely.

[NO SPEECH]

OK, a few additional notes about the paper. Initial steps of the project will be incorporated into future problem sets, and this, we do it for two reasons. First of all, it helps you spread out the tasks a little bit so you're not doing the entire empirical project in two weeks, the end. And it's also going to give us an opportunity to give you feedback. So if it seems like what you're proposing is just undoable or something like that, we'll know sooner rather than later and we can steer you in a better direction.

Second note to point out is that the project is meant to replace approximately two problem sets. It may take more time to do than two problem sets, and so keep that in mind. So you might want to start thinking about the project maybe even before we start asking you about it on problem sets.

And finally, for the graduate students in the class, we'll basically ask you to do a project that essentially combines both of those exercises that I described resulting in a six to eight page paper. And you can certainly come to talk to us if you would like. So basically, whereas the people who've signed up for 1431 would have more of the focus on describing and documenting a data set or more focus on answering a specific question, the graduate students would try to do both in a longer paper.

And by the way, I should say I've sort of given you two different models for this paper, starting with the data set model and starting with a question model. That doesn't mean that those are mutually exclusive. So if your project ends up having some of the flavor of each of those, being somewhere in the spectrum between those two models that I described, that's perfectly fine as well. OK. OK. Questions?

ESTHER: One addition.

SARA ELLISON: One addition, OK.

ESTHER: If you're planning to collect your own data by interviewing people, you need to obtain human subject approval. So that's not actually very difficult, but it means it's one step. So you will need to fill an application to the human subjects committee on the use of human subjects and we would need to sign it. And you need to give a little bit of time for them to actually approve it. So that builds in some time that's not compressible, so think about it early.

SARA ELLISON: And can we put something up on the course website, just a little document about the procedure, the steps you take?

ESTHER: I think I should actually do a little part in the lecture on human subjects. I will do it at some point or another. Maybe I should sooner rather than later.

SARA ELLISON: And if I'm not mistaken, the lead time could be as long as a month. Is that right?

ESTHER: [INAUDIBLE] just data collection. [INAUDIBLE]. But yeah, it could be. It could count month to we don't have to call in favor [INAUDIBLE] of our project. But a lot of those will be expedited, which means that the chairman can approve them. But in case that's actually a more involved project, [INAUDIBLE].

SARA ELLISON: So basically, just to clarify, the only people who have to worry about the human subjects approval would be people who are gathering their own data from maybe doing a survey, for instance, or an experiment with your fraternity brothers or something like that.

ESTHER: Using data that is restricted use for a reason or another, in which case the people who are giving you the data will know. [INAUDIBLE]. So they would tell you. It's pretty unlikely [INAUDIBLE] mostly for people who gathering data, actually interviewing. Copying data on the web [INAUDIBLE]. [INAUDIBLE] you want to ask your friends to get access to their Facebook contacts or whatnot, that's actually also gathering data, even though it might not be from a questionnaire [INAUDIBLE].

SARA ELLISON: So that would be subject to human subjects approval. OK, questions? Nope. OK, good. OK, back to functions of random variables,

So this is actually something I said last time before I did the two short examples, but I'll reiterate it now. There are various methods that we can use to figure out the distribution of function or a transformation of random variables, and the methods that are available to us depend on aspects of the problem. So they depend on whether the original random variable is continuous or discrete, whether there's just one random variable or a random vector, whether the function is a one to one transformation, an invertible transformation, or not.

And we could spend a lot of time going through all of these different methods. We don't have the luxury to do that in this class, but it's important for you to know that they exist, and so you can go look them up in textbooks. Not for this class, but for other purposes, if you need to figure out how the function of a random variable is distributed. And we'll focus on some examples, some interesting and important examples, in class instead of doing a more general treatment.

So I'm going to go through one fairly useful method for figuring out the distribution of functions of random variables, and here's the method. So we start with a random variable, x, and it has a PDF and we know the PDF. And we want the distribution of y, which is some function, h, of the random variable x.

Then one way we can find it is by first finding the CDF of y. How do we find the CDF of y? We integrate the PDF of x over the region x such that h of x is less than or equal to y. And then once we have the CDF of y, then if it is continuous-- typically it will be if x is continuous-- but if it is continuous, then we just take the derivative. If it's not continuous, then there are other ways to find the PDF from a discrete CDF, but we'll focus on the continuous case mostly.

So this probably seems a little abstract, and let me just emphasize a couple of things. So first, we find the CDF by integrating over the appropriate region. That part might seem a little abstract in particular, and we'll talk about how to find what the appropriate region is. And then take the derivative to find the PDF. We already knew that part. So let's do an example.

There is a typo in the slides. That should be a small f sub x of little x, not a capital F. So that is the PDF of x, not the CDF of x. And that typo is propagated on other slides as well, but I'll try to be clear about that.

So we have the distribution of x, the PDF of x, and we have the function, h of x, which, in this case, is just x squared, and we want to figure out what the distribution of y is. OK, so recall that we needed to quote integrate over the appropriate region. Well, that's easy for me to say, but not always clear how to do that, so I'm going to argue in steps what the appropriate region is.

But before we do that, first note that the support of x is -1 to 1. That's just listed right up there in the description of the PDF for x between -1 and 1. And then if we look at what values y can take on if x takes on those values and y is equal to x squared, then we can derive the induced support of y, and that's just the unit interval. Does everyone see where that came from? So keep that in the back of your mind. We will use that again in a couple of slides.

OK so what is the PDF of y? Well, let's start. Remember the slide I had up a few slides ago? It said, we find the CDF by integrating over the appropriate region, then we take the derivative to find the PDF.

So let's find the CDF first, and I'm going to do this in steps. So first of all, this is the first step, first of all, we're going to note. We're just going to write down the definition of the CDF of y. That's what we're trying to find. We're going to remind us what the CDF of y is. It's just simply the probability that Y is less than or equal to little y. Just the definition of the CDF.

Now the second thing we're going to do is we're going to plug in the function for y. So here, in this first line, I had Y less than or equal to little in the probability statement and I'm substituting in x squared less than or equal to little y in the probability statement. I can do that.

Then, the next step is we solve within that probability statement for x. So we just have that that's equal to the probability that x is between negative square root of y, little y, and positive square root of little y. Just solve for x. And now it should become clear what the relevant region to integrate the PDF over is.

So what we have here is a probability statement involving a random variable whose PDF we have, and so all we'd need to do is just integrate from negative square root of y up to square root of y the PDF of x. There we have it. And then once we do that integral, then we get positive square root of little y. So that's just what the PDF integrates to with those limits.

Now, note I put here this is for y between 0 and 1. Remember that a few slides ago I said that we would be using the fact that the induced support of y was the unit interval 0 to 1, and here's where we'll remind ourselves of that. So this is the CDF of y, and actually the next slide has it in more detail. Let me put that up.

So the CDF of y is equal to the square root of y over the induced support, over 0, 1. And CDF, as is true with all CDFs, is going to be equal to 0 for values less than the support and it's equal to 1 for values greater than support. That's just what a CDF looks like, remember? So it starts out at 0, then it accumulates distribution over the support, and then it goes to 1.

So if you have, in this case, we have a finite support, so the CDF is just going to-- I'll put a picture of it up. I think I have a picture of it. Oh, well, anyhow, it's going to be 0, and it's going to take this shape over the unit interval, and then it goes to 1.

And then so we've got the CDF. We're almost home free. We have a CDF of a random a continuous random variable. All we do is take the derivative to get the PDF, and that's what the PDF is.

[NO SPEECH]

And I also point out this is where we use the fact of the induced support. And that is what the PDF looks like.

[NO SPEECH]

So let's go back and just think about what we did. We started with the PDF of x, and then we figured out the appropriate region to integrate that PDF over to get the CDF of the function of x, and we did that in steps. And then once we had the CDF of the function of x of y, then we could just take the derivative to get the PDF. We'll see a few more examples as well.

[NO SPEECH]

OK, so the four examples that I'm going to go through in a fair amount of detail today, the first one is a linear transformation of a single random variable. The second is something called the probability integral transformation. The third is something called a convolution, and we've actually seen an example of a convolution earlier in the semester. And then the fourth one is order statistics,

OK, so let's start with a linear transformation. So linear transformations come up all the time. So I think it's important for us to have in the back of our mind or have formulas available for us to just plug into to figure out what the distribution of a linear transformation of a random variable is.

So why do linear transformations of random variables come up all the time? Well, maybe we have a random variable that's measured in the wrong units. So we have the length of Steph Curry shots in feet, but we want them in meters, or vice versa. So that's just a linear transformation.

Maybe we have a formula that dictates a linear relationship between two variables, and we know how one is distributed. So if we know how one is distributed, then we should be able to figure out how the other one is distributed if it's just a linear relationship. So for instance, the number of heating degree days in the month of February could be approximated as 28 times 65 minus the average high temperature. So if we know the average high temperature, the distribution of the average high temperature, then we can figure out the distribution of number of heating degree days in the month of February.

And then maybe we have a situation where there's some theory, maybe some economic theory, that predicts a linear relationship between two variables. So linear transformations or linear functions of random variables do come up quite regularly.

So let x have a PDF, little f of x, and let y be equal to this linear transformation, ax plus b. And we're not going to allow a to be equal to 0, because if we allowed a to be equal to 0, then that gives us this degenerate random variable. Basically it just gives us a point mass, which we're not really interested in that.

So how is y distributed? OK, well, we're going to do the same thing. We're going to go step by step and figure out the relevant region of the support of x to integrate over to get the CDF of y.

So we first write down, remind ourselves the definition of the CDF. So f sub y is just equal to the probability that big Y is less than or equal to little y, by definition. Now we plug in the function. So that's just equal to the probability that ax plus b is less than or equal to little y.

And here we have to split it up into two branches because the solving for x is different if a is less than 0 or if a is greater than 0. So if a is greater than 0, then that's just equal to the probability that x is less than or equal to y minus b over a. And if a is less than 0, it's just the less than or equal to sign gets switched.

And so now we've got-- actually, before I put that out, let me just emphasize what we're going to do. Remember, we've got the PDF of x, and we have two probability statements here. We only care about one of them, depending on whether a is positive or negative. But we've got these probability statements involving x and involving some other function of variables or whatever.

And so all we have to do is integrate the PDF over the region described by that probability statement. Is everyone on board with that? And what does that look like? Well, in the first case, we just integrate. We're asking, what's the probability that x is less than or equal to some value, and so we just integrate from negative infinity up to that value the PDF of x. That's it.

And then the other one, it sort of gets switched around. And by the way, I can write this, I can integrate from this value up to infinity, but I could also rewrite this-- it'll be handy to do this-- as 1 minus this integral. Because the integral over the entire support is going to be equal to 1.

[NO SPEECH]

OK, so I have these two expressions for the CDF, and all I have to do is take the derivative to get the PDF. And so I take the derivative, again, in these two different branches, and the derivative-- I'll flip back up-- the derivative of this with respect to y is just equal to that expression. And sort of similarly, I'm taking the derivative of this second expression with respect to y to get that.

And now we're at this point. We're done, basically, but we can rewrite it in maybe a slightly more convenient way by using absolute value of a so we don't have to carry around these two different branches. So we can just write the expression this way.

So now we have-- I mean, if x is a continuous random variable and y is a linear transformation of x where a is not equal to 0, here we have a formula that we can always just plug into to get the PDF of x. Questions?

[NO SPEECH]

OK, second example-- probability integral transformation. So I like this example for two reasons. One, because it's a little and it kind of makes you think in a way you might not be used to thinking. And the second is that it's super useful. Practically speaking, the probability integral transformation is used by lots of people all the time in all kinds of situations.

So let x be continuous and have PDF of little f of x and CDF big F sub x. And let y be equal to big F sub x of x. How is y distributed?

So there's something a little strange about this, right? Why would we use a CDF? We're used to thinking of CDF as a function that describes the distribution of a random variable and that that's what they are, but here we're using it to actually transform a random variable, which may seem like a strange thing to do.

But why not? It's a function. It's a function just like any other function, so we can consider it as something that we can use to transform a random variable, and it turns out there's a reason to do this that you'll see in a minute. It turns out that the result is very useful.

So now you have to switch the way you're thinking about big F and think of it as just any old function that might transform a random variable.

[NO SPEECH]

OK, so first let's note what the induced support of this new random variable is. So whatever the support of x is, the support of x can be anything, y is going to live on 0, 1. Why is that?

AUDIENCE: It's a CDF.

SARA ELLISON: It's a CDF, exactly. CDFs always are between 0 and 1.

[NO SPEECH]

Also note that F sub x is invertible, or at least note that I'm claiming that F sub x is invertible. So for F sub x to be invertible, we've noted before that it's non-decreasing, but in fact, if x is a continuous random variable over a connected set, then F sub x is going to be invertible. And we'll assume that in the description of the problem.

So we have a CDF. It always will take on values between 0 and 1 because that's what CDFs do, and it's going to be invertible because we're assuming these regularity conditions.

So how is y distributed? Well, let's just proceed the way that we've done in the last couple of examples. So we'll remind ourselves first what the definition of a CDF is. A CDF of y is just equal to the probability that big Y is less than or equal to little y.

Now we plug in the function. So that's equal to the probability that big F sub x of big X is less than or equal to little y. And now we solve for x. And that's why we need that big F is invertible to be able to do this. So now we have a probability statement that involves x and involves some function of little y.

So we actually don't, in this case, need to do the integration. We just note that this is the definition of the CDF of x evaluated at F inverse of little y. And we have the CDF of x, so we just put it there. So that's getting from the second step down to this step here.

And what is big F sub x of big F sub x inverse of little y? It's little y. And remember, the induced support of little y is the unit interval because CDFs always live on the unit interval.

OK, I'm seeing some puzzled looks. Do you just need a minute for this to sink in or do you have specific questions? I have more to say about this, by the way, but the next slide this calculation goes away, so I want you to ask questions you have about this before it goes away. Yep.

AUDIENCE: Can you give examples of when this would be used.

SARA ELLISON: Yes, I will, but coming up. Is this OK or do you want me to just go through the logic one more time? Yeah, OK.

OK, so we've got this function, capital F sub x. Don't think about it as a CDF. You'll have to go back and forth. You'll have to think about it as a CDF in a second. But right now, just think of it as a function that we can use to transform a random variable just like any other function.

So what we do is we start out using the multiple steps to figure out what the CDF of this transformed random variable is going to be, and so we just remind ourselves that the CDF of y is equal to the probability that y is less than or equal to little y, and that's equal to the probability-- now we're going to plug in the function for y-- that is F sub x of big X.

And now the next step we solve for x just like we've done before and we get that it's equal to the probability that x is less than or equal to-- just think of this as some function over here. Well, what is that by definition? That is the CDF of big X evaluated at that function. And so we write it this way. And then since we're taking F of F inverse of little y, then the answer is just little y.

[NO SPEECH]

This seem a little mysterious, I'll talk a little bit more about it and I hope I add a little bit of meat to this discussion. Let's see. Oh, right, so actually, this is a very important piece. We have to recognize what random variable has a CDF that looks like this. So I will tell you, it's a random variable we've seen before. It's a random variable.

AUDIENCE: Uniform.

SARA ELLISON: Yeah, a uniform random variable. So let me draw a picture to convince you of that. So does everyone recognize this as the CDF of a uniform 0, 1 random variable? So remember, the PDF of a uniform 0, 1 random variable looks like that. If we integrate this, then we just get 0 up until this point and then we get a line equaling y-- this is y here-- up until this 0.1 and then we get 1 after this.

You agree? OK so let me go back for a second. So we figured out what the CDF of y is, and it's just equal to little y, and that tells us that y has a uniform 0, 1 distribution. That's the CDF. The only random variable that has that CDF is a uniform 0, 1 random variable.

So what's the sort of bottom line of this example? A continuous random variable transformed by its own CDF will always have a uniform 0, 1 distribution. Any old continuous random variable. You take any continuous random variable, you transform it by its own CDF, it becomes a uniform 0, 1.

So how could this be useful? What's the practical application of something like this? I happen to think it's pretty cool, but that doesn't tell you why it's useful.

But before I tell you why it's useful, let's also think about whether this can work going the other way. So can we transform a uniform 0, 1 random variable by the inverse of a CDF and get a random variable with that CDF?

And I will tell you without making you puzzle over this for too long that the answer is yes. So we need to have some regularity conditions, like the random variable is continuous on a connected set, and so forth.

Again, pretty cool. Now, interesting, but how can this be useful?

[NO SPEECH]

OK, so let me talk for just a minute about performing computer simulations. Have any of you written computer code to do simulations before? Did you have to use random draws from some distribution? Do you remember what distribution?

AUDIENCE: A Gaussian distribution.

SARA ELLISON: A Gaussian distribution, OK. Yeah, so a normal distribution. So suppose that we want to write a computer simulation. We're interested in the spread of some virus over time in a school population, for instance. And we've got lots of pieces that go into our simulation, and some of them we have actual data on, and others we have to make assumptions about the distribution of certain variables to feed into the simulation.

So to perform the simulation, we might need random draws from a uniform distribution to model the proportion of the school population that was infected initially with the virus. We might need random draws from an exponential distribution to model the physical proximity of children during PE class. Maybe that gets fed into our computer simulation. And maybe we need random draws from a beta distribution to model humidity inside the school on different days. Maybe the humidity affects how transmissible the virus is or something like that.

So we want to feed all these different pieces into our computer simulation, but we need random draws. We have reason to believe that the humidity follows a beta distribution, say, but we need random draws from the beta distribution for the simulation.

Well, what happens if the computer language that we're using only generates random draws from a uniform 0, 1? So that's not actually so uncommon. There are lots of computer languages that may also generate Gaussian random draws because those are pretty common, but there may not be, in the statistical package or the computer language that you're using, there might not be any way to generate random draws from a beta distribution, or any way to generate random draws from an exponential distribution.

Well, I've just given you the tools to do it, so how is that exactly? Yep.

AUDIENCE: So there is an active [INAUDIBLE] published that time. Basically, what is that exactly? I'll take a sample of students and say that they are infected.

SARA ELLISON: Yeah, so I guess I haven't necessarily thought through all of the details of how one would write a computer simulation like this, but one idea I had was that you might want to start with different proportions of the population having the virus and see how quickly or slowly it spreads. And so then you might want to take a random draw from a uniform distribution and say, oh, we'll start with 3% of the school population having them.

And I should say, also I sort of said this just a second ago, random number generators, tables of random digits, many other sources of random and pseudo random numbers are giving you uniform random numbers. So if all you have is uniform random numbers, but you want random draws from a beta distribution or random draws from an exponential, then this transformation is how to do it.

So specifically, if you knew or could look up in a book the CDFs of whatever distributions you were interested in, say, the exponential and the beta, you could then compute the inverse CDFs and then use those functions to transform the random draws from a uniform 0, 1 into random draws from, say, exponential and beta distributions. So there may be a question on the next problem set asking you to generate random draws from distributions other than the uniform distribution, and now you know how to do it.

AUDIENCE: Can you just draw the plots and explain step by step what you would draw? So CDF of the exponential.

SARA ELLISON: So off the top of my head, I don't know what the CDF of the-- I mean, I can draw it, I think. So here's what the exponential distribution looks like. And so the CDF-- this is the PDF of exponential. So the CDF is going to look something like this. It's going to be 0 and then it's going to asymptote out to 1.

I don't remember if I tried to write down the formula. I'd probably screw it up or something. But you can find the CDF in any statistics textbook or probability textbook. So you just take that CDF, you figure out what the inverse of that CDF is, and that's the function you use to transform uniform random draws into random draws from the exponential.

AUDIENCE: So the inverse also has a plot.

SARA ELLISON: Yeah.

AUDIENCE: The inverse can also be bothered.

SARA ELLISON: That's right.

AUDIENCE: And the inverse of that CDF is a uniform distribution?

SARA ELLISON: Sorry, if you take a uniform-- no, no, this is great because I'm sure you're not the only one who's a little puzzled by this. You take a uniform distribution, you transform it by the inverse of this CDF, you get an exponential. And then it follows that if you have random draws, random numbers, generated from a uniform distribution, and you use the inverse of the CDF to transform each of those numbers, then those will be random draws from an exponential distribution.

AUDIENCE: How would we write that mathematically? F inverse of-- Can you give a picture?

SARA ELLISON: I'm not sure exactly what you're asking. So how would I write?

AUDIENCE: Same thing that you said that we take the inverse of an exponential and then transform the uniform random variable. If you could write F inverse--

SARA ELLISON: To be honest, I don't know off the top of my head what the CDF of the exponential--

AUDIENCE: [INAUDIBLE].

SARA ELLISON: Oh, so it's just-- let's see, sorry. No, where did I put it. Well, I guess, yes. I guess I didn't write the equation here, so I have it in words. We just transform a uniform 0, 1 random variable by the inverse of the CDF and we get a random variable with that CDF.

So we can just take if x is uniform and y is equal to-- let's see. I'm going to call this z to avoid confusion. So we have a new random variable, y, and this is a function of x. And in particular, it's this function of x. We take the inverse of where? We take the inverse of the CDF of z and transform x by that, where z is distributed exponential.

[NO SPEECH]

Then y will have an exponential distribution.

[NO SPEECH]

So you guys might not be able to see the board as well, but do you have questions? Everything's OK? So we have x uniform. We transform x in this particular way, where this is the inverse, CDF of an exponential, then the resulting random variable y is going to have an exponential distribution. Yes.

AUDIENCE: Taking inverse just means 1 over the [INAUDIBLE]?

SARA ELLISON: No, the inverse function, the inverse function.

AUDIENCE: And so now y becomes a exponential random data?

SARA ELLISON: y becomes an exponential random variable, yes.

AUDIENCE: A generator function.

SARA ELLISON: Yes, so if then what you do is you have random draws from a uniform, and you transform each random draw by this function, they'll be random draws from an exponential.

So let me move on, although it seems like you still have questions lingering. So I'm happy to do another example involving this probability integral transformation next time, but I'll move on to our next example.

[NO SPEECH]

Next example is a convolution. So a convolution has a specific meaning in math, and it has actually even a more specific meaning in the context of probability. And in particular, it refers to the sum of independent random variables.

So we've already seen one example where we cared about the sum of independent random variables, although we didn't know at the time that they were independent random variables. What was that example?

AUDIENCE: The pills.

SARA ELLISON: Yes, the pills. The headache example, exactly. So we were interested in the sum there because I could take the two pills sequentially, so the distribution of the sum of their effective lives was of interest.

And there are lots of cases where we care about the sum of random variables. So questions can arise in many contexts. The total value of two investments, for instance. Let's say I have my 401(k) invested in one place and my husband has his 401(k) invested in another place, well, it's all going to of get mixed together at some point, so we don't really care about each one separately. We care about the sum of the two random variables that represent the values in those two accounts.

Maybe we care about the total number of successes in two independent sets of trials. So there's some set of trials that's happening in one lab, and another set in another lab, and we split them up because we didn't have the capacity in any one lab to do all of the trials, but what we really care about is the sum. And so there are lots of cases where sum of independent random variables arises.

And I should also say that what I'm going to do here generalizes naturally in two ways. So I'm going to talk about the sum of two independent random variables, but it generalizes naturally to the sum of N independent random variables, and then it also generalizes naturally to the linear function of independent random variables. So what I'm doing here is sort of even more general than what I'm going to show you, but I'm just going to do the simple version.

So let x be continuous with PDF f sub x, and y is also continuous with PDF f sub y, and they're independent. Then let Z be their sum. What is the PDF of Z?

So remember in the headache example, we actually computed what the PDF of Z was. You not remember that, but what we did is we cared about, we hadn't even talked about functions of random variables yet, but what we cared about was the sum of them. When I asked, what if I take the pills sequentially, we cared about whether the sum was going to be greater than or less than or equal to some number.

So we found the region of the xy plane over which we should integrate, we drew the line x plus y is equal to little z, or something like that, and then we integrate it over that region of the xy plane, and then that gave us a probability that x plus y was less than little z. And that is, in fact, the definition of the CDF of the sum of those two random variables.

So we'll proceed similarly to the headache example. Oh, I should say one difference is that in the headache example, I just gave you the joint PDF. We hadn't even introduced independent random variables yet, so I just gave you the joint PDF. But in fact, they were independent.

We can easily get the joint PDF here because we know the random variables are independent, so we just multiply the marginal PDFs to get the joint. Then once we do that, it's exactly the same as the headache example.

[NO SPEECH]

So I think this is what I said just a minute ago. We set up the double integral to get the probability that x plus y was less than or equal to little z, and that is the definition of the CDF of this new random variable, Z, and then we took the derivative to get the PDF. And so I won't go through the math again. You can look back at the headache example, if you're interested, but that method works, and there are other methods that could also work.

And so what we get is we get that the CDF is equal to this expression. Now the difference, remember that in the headache example I gave you the joint PDF, and here we're leaving it general instead of specifying a particular joint PDF. I say each one of these random variables has their independent and they have their own PDFs, and so the joint PDF is just this. So it'll look a little different from the headache example, but basically this is what one of the steps of the headache example looked like.

[NO SPEECH]

And so once we have the CDF, then we can just take the derivative and get the PDF. Yes.

AUDIENCE: [INAUDIBLE] of the equation isn't z minus 1 because that's [INAUDIBLE].

SARA ELLISON: Yes, exactly, yeah. Yeah, so if you can go back to the headache example and see the pictures that we drew of the support and the joint PDF, and how we drew the line where x plus y is equal to Z, and you'll get the same limits of integration. I guess the other difference between this and the headache example is that in the headache example, the random variables only took on non-negative values, and so the limits of integration here started at 0 instead of negative infinity.

[NO SPEECH]

So now, based on this example, we have a formula that we can just plug into for sums of independent random variables, and this is the convolution formula.

[NO SPEECH]

OK, last example. This one may take a while, so we might not get through all of it today, but that's all right. So I told you before that the uniform was my favorite distribution. Well, order statistics are my favorite function of random variables, so that may be enough motivation for you guys to be excited about this example.

If it's not, then I will tell you that order statistics can be very useful in economic modeling. We're going to see an example of that next time involving auctions. And they're also the basis for some important estimators. So there's a lot of important applications of order statistics.

[NO SPEECH]

So let x1 through xn, x sub 1 through x sub n, be continuous, independent, identically distributed. And since they're identically distributed, they all have the same PDF. That's just the f sub capital X.

And I should say that having a set of independent and identically distributed random variables, this comes up all the time. So when we start talking about statistics, we will often use this as our first assumption.

And so because it comes up a lot, we have different names for it. We can abbreviate independent identically distributed as IID, so we'll see that. You'll see that in many contexts in probability and statistics.

And furthermore, a group of IID random variables is also called a random sample. So I'll use those terms interchangeably-- IID, random variables, or a random sample. Sometimes I'll say both and say IID random sample. That happens.

And so we've got this set of random variables. They're independent. They're identically distributed. Because they're independent, remember we don't need to be told what their joint distribution is. We only need their marginals because we can always find the joint because they're independent.

And let's consider the following function. We'll call it y sub n, and it's the max of x sub 1 through x sub n. So that might seem kind of a funny thing to consider because I've told you these are identically distributed. They all have the same PDF. But if you think about, you have to maybe be a little more expansive in the way you're thinking about these random variables. And what x sub n is, it's how if you think of each one of these random variables is having a realization associated with it. So in a particular random sample, we have the set of random variables, but each one is a particular realization from the support or from the distribution. The largest realization is the nth order statistic.

[NO SPEECH]

We can also define similarly the first order statistic as the smallest value. The second order statistic is the second smallest value, et cetera. So we can have a whole set of order statistics if we want.

And so the question is, how is the nth order statistic distributed? Well, let's figure it out.

So we start the way that we often like to start, by reminding ourselves what the definition of the CDF is. So we say that the CDF of the nth order statistic is just equal to the probability that the nth order statistic is less than or equal to some value, little y. And since we have-- oh, so by definition, by the way we defined y sub n, we can then plug in this probability statement in the place of the second one. So we can say that's equal to the probability that x1 is less than-- sorry, that looks like y sub 1. It's actually y comma.

So the probability that x1 is less than or equal to y, and the probability that x2 is less than or equal to y, and so forth, up to the probability that x sub n is less than or equal to y. So is everyone on board with that step? That comes from the definition of max, basically. If something is the max, that means everything else has to be less than it.

[NO SPEECH]

But we know these guys are independent, so that big probability statement, that probability statement that involved intersection of lots of different events, we can write that as the product of those events instead of the probability of the intersection of this big complicated event.

So that's just equal to the probability that x1 is less than y times the probability x2 is less than y and so forth. So we get that from independence. And, well, what are those things? Those are, in fact, just the CDF of x evaluated at y multiplied by itself n times. That's what it is. And this is true because these are identically distributed, so each one of these capital F's is the same.

[NO SPEECH]

Well, that was easy, right? Well, now we've got the CDF of the nth order statistic. Let's take the derivative to get the PDF. And so we do that, and when we take the derivative, we get n times the CDF raised to the n minus 1 times the derivative of what was inside here, which is just the PDF.

[NO SPEECH]

Pretty cool. How is the first order statistic distributed? Well, we can do a similar calculation, and it's going to lead to the PDF of the first order statistic looking like this. And that probably won't take you too much convincing to see why that's true.

And in fact, we can do this with the second order statistic, third order statistic. Those get a little bit more complicated, as you can imagine, because we don't have a whole bunch of events of everything being less than. We have to consider the different permutations of ways that we can get n1 of the order statistics less than a value, and n2 of them greater than a value, and so forth. So the PDFs of all the intervening order statistics, we can write them down. They get more complicated.

[NO SPEECH]

So now we have the following. We have the PDF of the nth order statistic, and the PDF of the first order statistic, and let's see a specific example, what do these distributions look like if we have a random sample from say, my favorite distribution, a uniform 0, 1 distribution? Well, we're just going to plug into these formulas, and this is what we get. Oh, it's going to be a function of n, so we'll choose n equals 5. We plug into these formulas and we get these two PDFs.

[NO SPEECH]

And here are the pictures. So this is the PDF for the first order statistic and here's the PDF for the nth order statistic. Does this make sense?

AUDIENCE: Do these graphs approach some probability when n goes to infinity?

SARA ELLISON: Can we leave that for just a moment?

AUDIENCE: Sorry.

SARA ELLISON: No, no, no, that's fine. It's a good question, but I'll get to it. Actually, I might not get to it today, but I'm planning to address that exact question in a minute.

OK, so this is what the distribution of the first order statistic looks like and the nth order statistic. Does this make sense to you? Yeah, why not? So if you have five realizations from a uniform distribution and you look at the smallest one, well, first of all, the support of that distribution has to be 0, 1. Because that's the support of the underlying distribution, it has to be the same.

And there's going to be more probability towards 0, if you're looking at the distribution of the smallest value. And if you're looking at the distribution of the largest value out of 5, there's going to be more probability up towards 1.

So think of it like this. You have a random sample of size 5 from a uniform 0, 1 distribution. How is the smallest realization from that random sample distributed?

So you just take this random sample of 5 from a uniform distribution, and whatever the smallest one is, this formula gives us how that smallest value is distributed. And likewise, you get something like this. Same support, 0, 1, but with a probability concentrated near 0.

And the logic carries over to the largest value. You just are looking at how the largest one of these guys out of 5 is distributed. And it's going to look like that.

So what if n is larger than 5? What if n is really large? So would you like to step in?

AUDIENCE: [INAUDIBLE].

SARA ELLISON: So basically what's going to happen is it's going to get more and more concentrated. The distribution for the smallest one is going to be more and more concentrated towards 0. The distribution for the largest one is going to be more and more concentrated towards 1. And in the limit, in fact, you'll get their point masses at 0 and 1 for the first order statistic and the nth order statistic.

And I just drew a couple of pictures that maybe illustrate this. So here's where n is equal to, I'm not sure what it's equal to, maybe 12 or something like that. So if you just think about what the distribution of this highest realization and the lowest realization are relative to if there are only 5 random variables in the random sample, then it's just likely they're going to have more probability of being closer to the edge points.

And then as you get, I don't know how many that is, what is that, 40 or something like that, then the probability that the nth order statistic is close to 1 is really high, and the probability that the first order statistic is close to 0 is really high. It would be a very strange event if we had 40 observations from a uniform 0, 1 and none of them were down here, for instance.

So I went one minute over and I apologize for that. But we got through order statistics. And next time I'll do an example, what was it, the probability integral transformation that had you guys scratching your heads. So I'll do an example of that, and then we'll move on.

[NO SPEECH]