[SQUEAKING]

[RUSTLING]

[CLICKING]

ESTHER DUFLO: I'll talk a bit about human subjects, and I also have a set of slides from that, but you can have them for reference. So I wanted to give you first a little bit of background of why this whole-- how the whole regulation about the use of humans as a subject of research has come about.

And then explain what are the big principles that we are operating on. And then finally, explain what this means for you right now and in the future if you continue doing human research with human subjects.

So the history of it is really-- until reasonably recently, and we're talking Second World War, end of the Second World War, there was no regulation of using humans as subject of research. People could do whatever it is they felt that was morally appropriate by their own standard of moral.

And that, of course, led to a number of excesses. The two most famous ones are Nazi research in the camp, which was exposed and everybody discussed in great detail during the Nuremberg trials after the war. But perhaps less obvious, although it was also relatively famous at the time and you may have heard of it, is the Tuskegee trials in America, in the US.

Although they are called trials, they were really not trials at all. They were a study of African-American men in who were sick with-- ill with syphilis. And the object of the study was to find out what would be the course of syphilis if left untreated in a context where we actually knew how to treat syphilis-- there was already a medication available for syphilis, but these people were actually not treated in order to see what is the natural course-- natural progression of the disease, what it does to the body, et cetera.

The people were lured with a false advertisement, that they would be tested and they would be-- and they would be given cure. And they weren't, and a bunch of people died while in the trial, obviously since they were sick with syphilis and no one treated them. And then when this came about, when this became public, this created a huge outcry, naturally.

There is any number of things that are wrong here. One is not treating people when there is a known cure that's already approved. One is lying to people when they come. The second is-- and the third one is addressing-- using as a group-- a group of-- they were sharecroppers. Very impoverished people who were not in a position to understand what was discussed, to advocate for themselves or to question what the researchers were doing.

After that, after-- and so with these-- these are the two that were the most discussed. I'm sure there were a bunch of other abuse of research at the same time, but this prompted a sort of international reflection on how to treat human subjects-- humans as a subject of research, which culminated in something called the Belmont protocol, which are international agreement of how to protect humans as a subject of research, which, in the US, are then translated as a set of regulations with a complicated codes that I don't remember, which tells us how to deal with subjects.

And basically, there are three principles, three important principles. One is beneficence, one is justice, and one is respect. Before I get into more detail into any of these subjects, I want to talk about what is research and what is a human subject, and that's important because research is defined as a systematic intervention, including research development, testing, and evaluation designed to develop or contribute to generalized knowledge activities.

That means definition constitute research for purposes of the HHS regulation whether or not they are conducted or supported under a program, which is considered research for other purposes. As you can guess, I didn't write this up. This is like a legal language.

For example, some demonstration and service program may include research activities. So for example, if you do some work with an NGO, the Y Combinator is launching a project on basic income. They are doing it themselves. In principle, that's research. I hope that they know that, but I don't know, because they are trying to find out-- knowledge that can be used to this kind of issue down the line, even though they are not funded by research funds and they are not in an academic institution.

But what this means, for example, is Facebook and Amazon or Google can experiment on you all they want without any human subject approval as this particular regulation because they are not producing research. For example, you don't need a human subject approval to conduct an experiment on how to display jars of-- pots of-- jars of jams in a supermarket because all you're trying to find out is how you should do it yourself and you're not trying to publish it in the newspaper. But the moment where it becomes research, that's when you need the human subject approval.

So for example, there is a study on testing-- jam testing in a supermarket where the researchers want to know whether people are more likely to buy jam after they have tested 10-- versus after they have tested two to test the theory of choice overload. You would imagine, this is minimal risk, which is true, but what-- which she still needed approval because she was planning to publish this, she being Radha Iyengar, who wrote this study, was planning to publish this in a journal.

So that's-- so the setting, in a sense, is irrelevant, but what you want to do, what the output needs to be is relevant. So when there was a big outcry apropos Facebook, the research that Facebook was doing and then-- I don't know if it was Tinder or OkCupid was doing some things as well by modifying the profile that you get to see and the like, there was some confusion whether it's an OK thing to do in general, that if people don't feel OK about it, they can just exit at some level.

And whether it was OK at the moment where it was discussed in the newspaper and presented in research project. And I think one issue in the Facebook, experiment was that no one had obtained human subject clearance even though it was clearly research. It was meant to be published.

So as I said, the key principle of the Belmont Report, and what every subject application with Human Subjects Committee will try to assess, is those three principles-- respect for persons, beneficence, and justice.

So how do you respect people? By asking them whether they want to be in the study. So it's informed consent. Protecting privacy and maintaining confidentiality. And then for people who are particularly subject to undue influence, that includes students, for example, in a university, there are special provision to particularly protect them.

Beneficence required to assess the risk-benefit analysis, including the study design. So it doesn't mean that you cannot do research that involves any risk, but that the risk needs to be commensurate to the potential benefit to the subject themselves and to society in general. So it could be that there is no benefit for the subject themselves. For example, when you do a survey, there's nothing in it for the subject, but you have to make the case about the importance of the knowledge to be gained to society.

And actually, it is fair game for a Human Subject Committee to tell you, I think your research is kind of stupid. So I'm not going to give approval because there is no gain to society to learn the answer to your question. I've never seen them really do that because generally they don't go into the weeds of assessing the quality of the research project, but theoretically, it is there. It is completely acceptable.

AUDIENCE: Can I ask a--

ESTHER DUFLO: Yeah.

AUDIENCE: What about studies where informed consent renders the study undoable?

ESTHER DUFLO: Yes. So you can request a waiver of informed consent in two circumstances. One is where the informed consent is the only thing that would lead the subject to the answer. So it would endanger confidentiality in a sense.

So for example, if you're planning to do a survey in a railway station when people are waiting for their train and you just want to ask questions-- for example, you give them a list of names and say, does this name sound to you that it's a name of an African American person or a Caucasian person? So you just want to find out how people names relate to people. So you have no interest in the identity of the person who gave you the answer. So you just write down the answer.

In the absence of informed consent, there is nothing that links the person who gave the answer to your data. In that case, you will generally say, let me not obtain informed consent, so then I really have zero risk for the person because no one will ever know even who answered. So that's one reason why you would not ask for informed consent.

The other is when it makes the study infeasible, for example, because people should not know what the study is about because that would change their behavior. So for example, if you have slight, subtle manipulation of how things are presented to people-- if you start by showing them a lengthy survey about-- with the phone number of the guy in the Human Subjects Department, that's going to somewhat undo the benefit of doing this subtle manipulation.

So for this reason, you can say, look, I'm really-- there's really no risk here. Let me do it. Then what Human Subjects Committee might require from you is that you provide a debrief to the subject later. And that, even you can sometimes fight back if it's not feasible, but in general, when it's feasible, that's what they'll ask for.

So for example, at the end, you will send an email saying, hey, you were part of this study. We were interested in knowing whether-- when people receive a daily email about the weather, they feel more optimistic or more pessimistic, whatever. This is what we found. 55% of people who received the weather felt good about the day, and 45 felt bad, and it was exactly the same thing in the no email group, so our conclusion is that it makes no difference. If you could send this kind of email. So some humans subjects will require that.

So none of that is that-- so there is a form that you fill up for your study. None of them-- but there is also then someone who looks at it at MIT where the Human Subject Committee is very, very organized, you will receive questions before the meeting meets. They ask you to send your application a month before-- if it's not an exempt application, which I'll discuss in a minute, they ask you to send your form a month before the meeting.

They look at it before they send you question and iterate with you so that by the time you come to the meeting, it's pretty much going to be approved unless there is something really tricky about it. So these questions would have been-- this back-and-forth would have taken place.

If you're doing something very simple, for example, just collecting data-- so if you collect-- if you're doing a survey, you do not need a full human subject approval, you can be considered to be exempt. What exempt means is that it's going to be reviewed by the leader of the committee and not by the full committee.

So typically, for you, in your case, when if you collect your own data or do simple experiments, you can probably be exempt. Regardless, what this means for you is if you want to do anything involving human subjects, you have to take the human subject training. You'll find a link to that on the MIT page because they will not even consider your application if you've not been trained. It takes about an hour to go over all the screens and answer the questions, so it's something definitely worth doing.

And then you can submit your application. Just in case you might not be exempt, it's better to send it a month before you have deadlines, but in general, you could be approved much faster than that. Any questions on this set of issues? Yes?

AUDIENCE: Are there different types of-- because one of my friends who did the COUCHES training, he said there were 16 modules and one module took them two hours. And so I was like, for a 14-week project, should I have to go through all the training and--

ESTHER DUFLO: I don't know. I have to do this training all the time because you have to redo it regularly. It doesn't take me that long. So maybe the reality is in between. But it's like-- basically, it's a set of slides, and you-- it's makes sense. It's sensible.

AUDIENCE: You can do the whole thing like an hour or two. You just click through the screens and answer--

AUDIENCE: [INAUDIBLE] question?

AUDIENCE: Yeah. I just did it recently again for the second time. I think it took me an hour and a half.

ESTHER DUFLO: Yeah. I'm on the Human Subject Committee-- not for MIT, but for the NBER. So these issues are known by me, but still, I think it's not very difficult. And it's informative anyway because it is-- some of-- the way they ask the questions, a bit-- it's not obvious. So it's a bit-- like if you took a class in analytical philosophy, some of them are a little bit like this kind of problems that you might ask yourself. So it's pretty interesting.

No, maybe he was very, very, very, very careful, your friend. It shouldn't take that long. Any other questions? So what I want to talk about today is special distributions. So what is so special about the special distributions?

So some distributions are special because they are related to others. So they help-- they are building blocks from one to the other. And in fact, a bunch of nice charts showing the relationship between various distributions. Some distributions are special because they can be used to model a wide variety of random phenomena.

The normal distribution being chief among them, but for example, the Poisson distribution we are going to see today to model the arrival of events is also very useful. Sometimes this is the case because there is a fundamental underlying principle-- that's true for the normal distribution, there is a reason why a lot of things end up-- a lot of variable we use in social science end up looking like they are normally distributed, and we'll discuss that reason in detail on Wednesday, I think.

And some cases, it's just because a particular family is just-- can describe-- you have very few parameters that can describe a lot of-- that can describe a lot of random variables. So it helps when we go to estimation, the underlying, you can say that with-- by estimating one parameter, we can map potentially a lot of phenomenon. We like that to be explaining-- to be trying to fitting distribution with as few parameters as possible.

So I don't really work on statistics, but I work-- I do some work on social networks. And on social networks, every other week, there is a new measure of centrality. So it's a little bit like that with distributions. Like as far as I can tell, like there is always new candidate special distributions. But to be really special, distributions must be mathematically elegant, and they should rise in settings that are applications-- real applications that are interesting and diverse. So I'm never going to be the judge of what is special distribution, but that's what people look at it.

So today is going to be more of a reference class. We're going to go over a lot of distribution. The mean and the variance of the distribution is going to be there, and we are going to show you how they look like. It's going to be less conceptual than what we've done before, but you can then use it for later.

So we've seen many of them. Some of them we've not necessarily named them, but we've seen many of them along the way. This is stuff we've seen. And let's go over them. So we'll start with the discrete distribution.

So the fundamental building block, of course, is the Bernoulli distribution, which is when there are two possible outcomes-- you're flipping a coin, or you're wondering whether someone was treated or not treated, or you're wondering whether someone is Republican or Democrat, then the probability of failure one of the two outcomes you're going to name success; one of the two outcome you're going to name failure. So let's call the probability of success p and the probability of failure q, which is 1 minus p.

Then f x of p is p to the x q to the 1 minus x of-- if x is 0 or 1. And of course, it's not defined otherwise-- it's 0 otherwise. So this is just the chance that you get a tail or a head by doing a coin. And p is-- with a coin-- with a fair coin, of course, p would be a half.

The mean of Bernoulli distribution is variable-- that is, a Bernoulli distribution is simply p. That is something that's extremely useful, actually. It's going to come back again and again. You would think that's the simplest thing on the planet, and it is the simplest thing of the planet, but it's very useful in economics.

Partly when we think-- in particular when we start thinking about people-- when people are assigned to treatment and control, which is what I work with randomized control trial, people are assigned with treatment and control, we're keeping use-- making use of this very simple result all the time. Where does it come from? Well, the expectation is simply the probability that x is 1 times the probability that that happens. It should be a p, not a 1 here. Plus the probability that x is 0 times 0, and that's directly p.

And we can do the same work with p squared. And also, the expectation of x squared is also p for the same reason. It's p times 1 squared plus 1 minus p times 0, which is, again, p. And combine-- and that allows us to calculate-- combining these two allows us to calculate the variance of a binomial, which is p times q.

So it can tell you-- so it tells you, for example, that the Bernoulli-- so it tells you, for example, that the Bernoulli distribution that have the largest variance are the one that are with p, the half. So this is our friend Bernoulli. We're going to use it a lot today, but we're also going to use it a lot-- I use it all the time in my work.

Binomial we've seen already, we've seen as well. If X1 to Xn are iid random variables, which are all Bernoulli, then the sum of these Xk's is distributed as binomial distribution with parameter n and p. So the Bernoulli distribution you can write as B of-- binomial of 1p.

So in words, what does this mean? It means that the binomial distribution is the number of successes in a sequence of n independent success-failure trials, all of which yield success with probability p. We've seen the density function. It's binomial nk times pk times 1 minus p n minus k for x equals 0, 1, 2, 3, n, otherwise 0. It's defined over this-- it has a positive probability over this. This x is otherwise 0.

Well, just as a reminder point and case, n factorial of n divided by factorial of k factorial n minus k, It's n choose k. I'll give you the-- so the expectation-- the variance are easy to compute from what you've seen last time because expectation is a linear operator. We are summing n-values that have the same expectation p, so the expectation is np. And for the same reason, the probability of the property of the variance, the variance of x is simply going to be n times pq. n times variables that have pq.

What did-- what was our example for a binomial distribution? The example we chose-- we used in-- the same that I used in the first lecture where we introduced it? Oh yeah, many heads and tails, but in like a-- no, we are coming to the pizza toppings there. They will be back. They will be back-- Steph Curry, exactly.

So here-- so sorry. That's a binomial. That's two binomials, actually. So what did we-- just to see how they look like. So these are two binomials, one with a p of 0.2 and 1 with a p of 0.05. You can see that they are-- and in both cases, we take 40 though.

By the way, these ones I drew-- I drew in LaTeX. I discovered while writing this lecture notes that you can draw distributions in LaTeX. So if anybody has an interest, I can post the source code of how you do this lectures, and it actually makes nice graphs-- I mean, I don't know, maybe you don't think they are nice at all.

But so these are two examples, one with a p of 0.05, this one will see again; and one with a p of 0.02. So you can see, it's symmetric. And it's almost bell-shape, although, of course, it's a discrete distribution, so it's a funny-looking bell. We'll come back to that in a minute.

So Steph Curry. Do they follow-- so this is how binomial looks like. And this is the success out of the first 20 attempts. So this is what-- roughly what it should be. So do you think it's looking binomial?

AUDIENCE: Looks skewed to the left.

ESTHER DUFLO: Yeah. It's a little bit-- well, maybe it could be. There is some chance maybe it could be. But maybe it's a bit skewed to the left. So it doesn't look that binomial. And the thing is that-- so I'll give you two pieces of information. It's unlikely to me because-- so a binomial can be a sum of Bernoulli. Of course, some of binomial, as long as they have the same p.

But the success is made of the success of the 3-point shots, and the success of the 3-point shots. And this is how the 3-point shot looks like. A success on the 3-point shot looks like. And it's definitely quite skewed. And this is a 2-point, which skewed the other way. So this is a sum of two distribution, neither of which is binomial. It seems very unlikely that it's actually binomial.

We'll go back to-- we'll go back to think what these guys might be, actually, but all that to say that it is useful to know a little bit the theory of how things come. Sometimes the data is not sufficient. Maybe you could test that. We can not yet because we haven't done estimation yet.

But when you do estimation, you'd be able to test whether this looks like it's distributed like a binomial. I don't know with 56 observation, 56 games, whether you would really have the answer, but here, we're bringing the theory to say, well, it can't be true, actually, when you think of it. So that's for the binomial.

The hypergeometric is our friend, the pizza toppings. When you-- the binomial distribution we use to model when you want to model the number of successes in the sample of size n with replacement, when you're retrying all the time. So that's why the pizza topping was the number of vegetarian toppings on your pizza, you're not-- once you've picked artichokes, you don't put back and can pick artichoke again or your pizza will be very artichokey and not very delicious.

So the typical textbook example is the number of red balls you take from an urn. And the example we used as a running example in this class is the number of vegetarian toppings on pizza where they could be toppings that are either vegetarian and non-vegetarian. So here are-- let A be the number of successes and B the number of failures. You might want to define big N as the sum of A plus B.

And N the number of draws. Then this is the distribution of an hypergeometric of a X random variable that is distribution as-- I realize, I forgot to specify the-- I forgot to tell you over which x. So x is like 1, 2, 3, et cetera till n, and otherwise 0. So it's a binomial Ak, binomial B, and minus x over binomial A plus B over n.

So it is somewhat related to the binomial, which you can see very well when you look at the formula for the expectation and for the variance. A over A plus B is the number of successes divided by the sum of success plus failure. So it's a number of vegetarian toppings available over the-- overall. So you can think of it as a pB as the-- p being the fraction, the probability of a success.

So it is almost looking like-- the expectation would be almost like np if we define p as the probability of vegetarian over the total sum of toppings. But the variance is a little different. So this part of it looks like n times p times q, but then there is this guy, too. So what is this guy doing? For us, it's going to make the variance bigger or smaller?

It's going to make it smaller. Accounting for the fact that every time we remove a vegetarian topping, there is a smaller sample, so there are fewer options. And of course, as n becomes very large, then this term is-- what happens to this term? It becomes smaller and it goes to 1. It becomes smaller and smaller and smaller over time.

So when big N becomes much larger than n. So when the sample is very large and you're drawing a few, it's basically saying the replacement doesn't matter anymore.

AUDIENCE: Could you explain what k and x are?

ESTHER DUFLO: So x is-- the k is probably a-- what is k? There's no k here.

AUDIENCE: --the formula [INAUDIBLE].

ESTHER DUFLO: Yeah, yeah, yeah. k is a typo.

AUDIENCE: Oh.

ESTHER DUFLO: k is a typo. There is a number of these-- I keep switching between k and x. People use both conventions, so I got mixed up. I'm trying to stick with x because that's what I used earlier. And I'm starting from things that have k, so I try to-- sometimes there would be also floating y's, so I apologize for that [INAUDIBLE]. Anything that's a k or x or y will typically be interchangeable. This should be an x.

So it's basically telling us that it's the number of combinations where I have picked A vegetarian toppings. And the number of combinations where I have picked B, the rest, the complement of A, non-vegetarian toppings, divided by all of the possibilities for picking exactly those ones.

So when n is-- when the sample size is large and the sample-- and the number of trials you have is small relative to that, then they are going to look more and more and more similar. Basically, you can ignore that there is replacement or not replacement. So typically, for example, when you're going to draw-- when you're drawing a-- doesn't matter. Go back to that.

So we have asked ourselves, how is a random variable distributed if it's describes just the probability-- if it describes success or failure? And if it's describe the number of success out of a given number of attempts, then the same thing without the replacement-- so say the number of vegetarian toppings that I pick, if I pick a given number of toppings, a number of toppings.

The last thing we could ask is-- or the next thing we could ask is, how many trials do I need to perform to achieve r successes? So you might ask yourself, why would I ever ask this particular question? So what would be a practical example-- what would be an example where you would be interested in this very question, the number of trials to achieve r successes? We haven't done an example of that yet.

AUDIENCE: I have to make a team of people who speak a particular language out of lots of people. I pick them at random and ask them what language do you speak, and they say, whatever, Hindi or Chinese. And so they are on the team; otherwise, they are not.

ESTHER DUFLO: Yeah. It could give-- exactly. It would tell me how many people I need to pick in order to have my fully constituted team, exactly. Yeah?

AUDIENCE: If you're taking a survey or you just want to [? do ?] in a study [INAUDIBLE] number of people [INAUDIBLE] survey [INAUDIBLE].

ESTHER DUFLO: Yes, exactly. So it would be very useful for polling. Exactly. Or you could ask yourself-- you could set r equal 1. For example, the first failure. And you could define-- the first-- you could define success as the machine breaking down. It could be how long will the machine go in expectation before breaking down? So in engineering, that's something that they will need all the time. They will need to know all the time like-- what's the-- how is it distributed, the number of times that I can run my machine until it's actually going to fall off? Become sick.

So that is something that people use also in engineering or in studies of linking networks together, et cetera because once if you have several of these machines together, that are linked together, then you can also start thinking about issues like, what if they all fail? So depending on what you're trying to study, it could be, what is-- how many times of use will it take till the entire system breaks down? How many supports do I need for my machines, et cetera.

So this is distributed in the-- here, I had the maximum number of y and x and k is mixed up, but I think I cleaned them. This is p to the r q, which is 1 minus p, to the power of x minus r multiplied by x minus-- binomial of x minus 1, r minus 1. So let me try to give you an intuition for that.

So we're trying to get exactly r-value and then stop. So what we know is that the last one has to be a success. So we know that in-- we've done a number of successes. So we've done x minus 1 trial. And finally, 1 success. It's x-- yeah, for x, we've done x minus 1 trials. And then finally, 1 is a success. The other thing we know is, out of this, how many of those have been successes?

AUDIENCE: r minus 1.

ESTHER DUFLO: r minus 1. Exactly. So there has been-- so we've done x minus 1 trial, out of which we got r minus 1 successes, and finally, we get one the last thing is a success. So it happens-- that one happens with what probability? p. So that one happens with probability p. So we know that we are going to multiply something that is here by p.

And then to find out what's the probability of that happening, we say, well, for any sequence that has r minus 1 successes and the rest is failure to happen, it's probability p to the power r minus 1 q to the power x minus r to get x minus-- to get p successes out of-- I need to look this side of the room-- out of x minus 1 trials.

For any sequence that has-- any of this sequence happen with that probability. And then how many such sequences are there? Well, any way to arrange r minus 1 successes out of our x minus 1 trials, so we can put it-- so that gives us this number. So this is the number of possible arrangements. And each of them happens with that probability. And then we have another p here, which gives you the formula that's over there. So that's how you go from the Bernoulli to negative binomial.

Some textbook-- some people define it differently, but for our purpose in this class, we're always going to define it this way. But some people define it differently, and then there are-- but use the same r, so you need to be careful. Some people define it as the number of failures needed to achieve r successes. So instead of being the entire number of trials, it's just a failure. So I didn't write down the formula so that do not confuse you, but you could find it this way as well. Of course, it's just a re-normalization of the indices.

The geometric distribution is the same thing with r. So it's the number of failures before the first success. And as we were discussing as an example, most of these things-- these failures are actually lack of failures. So this failure is the number of times where the machine doesn't break down until it actually breaks down, is a geometric distribution. So it's a binomial distribution with r equal 1.

The expectation of a geometric distribution, q over p, and the variance is q over pi square. So the sum of r independent geometric is a negative binomial. And in fact, the sum of a number of negative binomial is also a negative binomial. So you can relate them to each other, and that means that you can tell us what is the expectation of a negative binomial.

AUDIENCE: [INAUDIBLE] and the expectation is 1 over p. The variance is 1 minus p over--

ESTHER DUFLO: Yeah, yeah, it shouldn't be-- so I think it might be related to the scaling over-- it might be related. So first, let me check that I get it wrong, and I get it right. And secondly, if that's the case, it might be because they define it as a-- they define it differently. Let me check that for you.

So I believe this is correct. Why would they say that-- yeah, they define it by the number of-- they define it differently. They define it with this alternative way. It's the number of-- so that's why you need to be careful. It's going-- naturally it's going to change the-- naturally it's going to change the expectation. The number-- if I count the total number of trials, I should, by the way, write you the formula.

AUDIENCE: They also have-- there's two columns in the Wikipedia page, 1 over p, and the other column says 1 minus p over p. And you define q as 1 minus p.

ESTHER DUFLO: Yeah. I'm glad that you are checking in real-time. By the way, the expectation for-- the expectation for the negative is rq. over p. And the variance-- actually, I don't know, what is more conventional?

AUDIENCE: I would have said that the number of trials--

ESTHER DUFLO: This number of trials is more conventional?

AUDIENCE: --the number of failures. For the geometric, maybe you see number of--

ESTHER DUFLO: Failures, yeah. Because--

AUDIENCE: But then to make it consistent with the typical definition for the negative binomial and make the sum of geometrics [INAUDIBLE] the negative--

ESTHER DUFLO: Binomial, you need to define as a number--

AUDIENCE: The number of trials.

AUDIENCE: So is that the expected value of a negative binomial?

ESTHER DUFLO: Yeah, it's here. So I should write, some textbook people and Wikipedia define it as the number of-- although that could change. By tomorrow, one of you could have fixed it. And then I fixed it, I don't know what changed it. I don't know, it's not an error. So that I had written the-- I had written out the formula, and then we decided that that would create confusion, so I removed it.

So the Poisson distribution is a very nice one. The Poisson distribution expresses the probability of a given number of events occurring over a fixed interval of time if you can count the event in whole number, if the occurrence are independent, and if the average frequency of occurrences for a time period is known. So what could be-- I'll put that more formally in a moment, but what would describe-- what could be nicely described by a Poisson distribution?

AUDIENCE: Radioactive decay.

ESTHER DUFLO: Sorry?

AUDIENCE: Radioactive decay.

ESTHER DUFLO: Radioactivity-- radioactive decay? Are you asking? You're way past my expertise level. I don't even know it would be a discrete thing.

AUDIENCE: Not discrete.

ESTHER DUFLO: It has to be like a number of-- think of a number of things. Yeah?

AUDIENCE: Positive [INAUDIBLE].

ESTHER DUFLO: Yes, although if you had-- yes, I guess if you had-- yeah, that's a discrete-- positive period of the case of the flu in a day, for example, you could, exactly.

AUDIENCE: Number of the buses coming to the bus station.

ESTHER DUFLO: Number of?

AUDIENCE: Buses. How many [INAUDIBLE] bus station [INAUDIBLE].

ESTHER DUFLO: Yeah, it could be, although it might not be-- it might change a lot over time. So it might not have might-- you want to be much more specific, which is between a period of time because it might be clumped otherwise. Yes?

AUDIENCE: Could you do something like number of hurricanes in a hurricane season or those who may not be in the [INAUDIBLE]?

ESTHER DUFLO: I think that there might not be independent across season, but for one season in particular, you could say this is the arrival rate. What do you think?

AUDIENCE: I didn't hear what he said.

ESTHER DUFLO: Hurricanes. Number of hurricanes in a hurricane season. You could say, yes, you could say for a particular season, the probability that they are-- it could be Poisson, yeah.

AUDIENCE: I mean, we would have to know a little bit more about meteorology to be able to answer that because we have to assume that these are independent events, and maybe, in fact, meteorologists know that they're very connected, related.

ESTHER DUFLO: Definitely not independent-- within season, there are higher season and they are new season, non-new season, et cetera, but within that, maybe for a period of time within a season, they are-- they could be considered independent. I don't know, but that's exactly the right question. Yeah.

AUDIENCE: [INAUDIBLE]

ESTHER DUFLO: Yes. The number of-- the number of goals scored in a match where people could describe as Poisson. Yeah?

AUDIENCE: Typical example is that I'm waiting for the train, whether it comes at [? any ?] given time.

ESTHER DUFLO: The number-- it has to be the number of trains showing up. That's what we discussed with the bus. I don't know whether it is that clear because it's going to-- so I guess you could-- if you restrict it over a given time period, you could, but it depends a lot from time period to other. So it's not-- there are many more trains at 6:00 in the morning than at than at noon. Yeah?

So something that would not be Poisson for example is the number of people showing up at a cafeteria during a given interval of time. First of all because people tend to arrive together, so they are not independent. The events are not independent. And then second of all, it's going to depend a lot of-- and so people might arrive together. So you can have two people in the exact same moment on time, which we are going to rule out.

And second, they are coming more at some time than some others, so they are not independent. There are more people coming over time. That's why with the bus and the train, I was trying to specify that you need to know, within a particular-- within peak time, maybe it's independent, but for the hour, definitely-- for the whole day, definitely it isn't because there are more train coming up at some time than some others. You had a point? No.

AUDIENCE: If you see the number of students coming to the class from 10:30 or 10:55, is that because they're independent--

ESTHER DUFLO: Well, people-- yeah. So the problem you might have is people might arrive together. Two friends might arrive together, and we're ruling that one of the person Poisson is going to rule this out. Another? Yeah?

AUDIENCE: [INAUDIBLE] radioactive decay would just be the number of decay of that?

ESTHER DUFLO: Yeah, maybe it's-- like, I really know even less about radioactive decay than about meteorology, so I don't even know what the object is. So the number of ideas that a research department might-- have a research department of a firm will have in a month might be a Poisson distribution. Every moment they could have an idea or not and they're coming in this independent way. So sometimes we model this as Poisson distribution. In fact, it's often modeled like that. I wonder whether it's appropriate, but it's often modeled like that.

AUDIENCE: [INAUDIBLE], is it fair to say [INAUDIBLE] you can't within a smaller distribution actually prove [INAUDIBLE] likely independence [INAUDIBLE] on the population? The reason I mention it is, referring to that example, within a specific [? lab ?] [INAUDIBLE] assume that these people have a lot of contact with each other. [? Then ?] if somebody [INAUDIBLE] basically [INAUDIBLE] the fact that people [INAUDIBLE] not necessarily independent.

But if you took, let's say, the entire economics department where you'd have not necessarily as much-- like, if somebody has an idea up here, it's not necessarily [INAUDIBLE], then could you make that assumption?

ESTHER DUFLO: Yeah. Even if it's one department, you could say, well, I'm going to consider the entire department as a-- so the question is whether they arrive in an independent way, you can say, well, a particular lab would have some streak of imagination at some point and they will start generate a ton of ideas, more at some other than some others, than it would not be a Poisson.

But with several people working independently-- so for each of them, it might be person. They are thinking sometimes-- something comes up, sometimes it doesn't. And then you could say for the department, it's a bunch of people working together, that might work. Then might work.

So let's put that in-- let's put this-- the discussion we're having informally in a more formal, mathematical way. So let's so-- the Poisson is a discrete distribution, although sometimes people draw it as a continuous-- or draw it continuously because when you have a lot of ends-- it starts looking almost continuous, but it's really a discrete distribution. It's about--

So it's Nt-- let Nt be an integer-valued random variable which satisfies some property. First, N0 is 0. Then s is smaller than t. And Nt minus Ns are independent, what does it mean? It means that arrival are independent over disjoint interval.

So for example, it rules out, in the case of students arriving at the cafeteria, it rules out the fact that there is a rush hour where a lot of people are coming then a non-rush hour. And then Ns and Nt-- plus s minus Nt have identical distribution, what does it say? What does it reflect? Yeah?

AUDIENCE: Particular time slot. They're identically distributed.

ESTHER DUFLO: Yeah, exactly. arrival time-- the number of arrival in an interval depends only on the length of that interval and not where that interval is located in the time period we're looking at. And then the distribution of-- as the interval-- as the interval becomes the entire period, you're getting-- the limit of that is gamma.

So we call gamma the arrival rate, and it's a constant for small intervals. So it's the limit of the interval becoming smaller and smaller and smaller and smaller is going to be gamma. Which means that in a period of length t, what's the arrival rate-- what's the number of arrival we can expect? Gamma t, exactly.

So if all of that is-- so that's already tell us what is happening. Then if all of that is true-- oh, I forgot to discuss the fifth postulate before I got there. The two things-- no simultaneous arrival. We rule out simultaneous arrival. So no people coming together. No two-- so for example, for ideas in the lab, it might be a problem if one idea immediately brings down a downstream idea.

If all of that is true, then we defined the probability that the PDF as the probability that Nt is k is gamma t to the k e to the power minus gamma t divided by k factorial. So gamma and t are always together. Gamma is the arrival rate for as the interval becomes smaller and smaller and smaller. And t is the length of the interval we are considering.

So we generally denote lambda is equal to gamma t. So if you check Wikipedia, it's going to be full of lambdas and it's not my typos. Lambda is equal to gamma t. It is the propensity that someone will arrive over-- the propensity to arrive over a fixed period-- over a fixed period of time.

So what is the expectation of x-- of n? You already gave me the-- give me the answer. It's gamma t over lambda. Exactly. The expectation-- so it's a very interesting distribution because its expectation is lambda or gamma t. And its variance is also lambda. So one parameter entirely describes this distribution.

Here is how it looks like. Probability for different lambdas. So I plotted it as a function of the lambdas since that's the-- and you can see, that with the small lambda, we have a lot of mass. It's towards the left-- it's quite skewed to the right. And as lambda becomes bigger and bigger, what does it start to look like? It's more-- less and less and less as we get more occurrence of-- or more-- as the event becomes more likely to occur during a fixed time period, it becomes less and less skewed.

Eventually what it looks like-- because it's discrete, it would say-- it looks more and more like a binomial distribution. So some property, its expectation is lambda. Its variance is lambda. You might wonder whether that's a bug or a feature.

In a lot of cases, it's a feature because it has fewer parameters to estimate, but in some cases, it might be a bug because you might not be able to-- you might want to increase the expectation of the distribution you're trying to work with without increasing the variance at the same time, in which case, you would use something else. You would use a distribution that allows-- that has a different-- that don't pin the variance and the mean together.

It is skewed because it cannot be negative. So if an event is not very likely, it's still the probability that it happens. It's never negative, so it is skewed. But as the lambda increases, and therefore, this constraint that it cannot be negative is less and less likely to bind, then it becomes more and more symmetric.

So these two things-- as you pointed out, as you increase lambda and you have a certain number of-- you have a lot of observation, it's starting to look more and more like a binomial. So you said a normal, but that's because, as we will see very soon, the binomial starts looking like a normal when you have enough trials.

So all these things start looking together-- and the two are, in fact, related. So suppose that you took an interval-- your interval 0, t. And you divided it in lots of small intervals. Small enough that the chance that two events happen together during that interval is almost 0.

Then the probability of success in each interval is lambda over n because you've divided that guy where the arrival rate for the entire period is lambda. So the probability within each interval is lambda over n. And the probability of n success over the period is, therefore, approximately binomial, because it's the number of successes in this n little independent trials that you are making-- that you are trying.

So that's a way of relating the two. And so that helps you give you the intuition for the result, that the limit of a binomial distribution as n-- the number of n goes to infinity is binomial. That only-- that doesn't work for if the lambda is too small because as you could see, it doesn't work if the lambda is-- if the Poisson distribution is so skewed, it could never look like a binomial. So it works only for this lambda has to be a little bit larger-- large enough.

So p is one of our lambda. Lambda is fixed, and n is a positive number. So the advantage of the Poisson distribution is that it's much easier to work with than a binomial. So when you have a small value of p, which means that that is going to give you-- if you have large values of p, the binomial distribution is better approximated by a normal, which we are going to see in a minute.

But if you have small values of p's, then the Poisson distribution can be used to-- become close to a-- can be used to describe a variable that's actually a really distributed as a binomial distribution and it's much easier to work with. I think we discussed all that.

Oh, and I want to go back to Steph Curry. So this is a number of 3-point shots made in a game. This is not a fraction anymore, this is a number of success of-- this is the number of successful shots. And this is the number of successful 2-point shots. Maybe it's Poisson. I don't know.

I don't have a theory that it would not be Poisson. The 2-point shot doesn't look so great. The 3-point shot, it has skewed and it starts by going low, and then it has a peak. Maybe it could be a Poisson distribution. Do you think it could be a Poisson distribution on theoretical grounds?

AUDIENCE: --in my head to--

ESTHER DUFLO: I went-- I tried to-- I tried to think of the postulate and thinking that maybe it could work. Certainly you don't do two at a time. It might not be independent.

AUDIENCE: It won't be independent, but it could be close, perhaps.

AUDIENCE: Each shot has a fixed probability, but each success in the Poisson distribution has a fixed probability p, right?

ESTHER DUFLO: Over an interval of probability-- gamma over a period, over a long period, it's over any interval, it's-- as the interval becomes smaller and smaller and smaller, it goes to gamma. So over any interval t, the probability of success is gamma t.

AUDIENCE: I'm trying to rationalize what Steve Curry's-- or Steph-- if his shots have one probability p. p where his shots being a success--

ESTHER DUFLO: So here, it's like maybe success is a little bit confusing here because I put the number of successful games, but just think the number of shots.

AUDIENCE: Yeah.

ESTHER DUFLO: We could also do the number of attempts. This is the number of actually successful shots. So the question is the number of-- so the question is, how many times does he actually put the ball into the basket in a game? Not out of what we were doing before with the Bernoulli. It's the number of successful shots out of the 20 shots.

So this had this idea of the probability of success. The probability of success is nowhere here. So this is just the number of shots that are being made. So in every single-- in every moment of the game, suppose that in every moment of the game, it could be throwing the ball into the basket. And as this-- if the entire game is-- how long is the basket game? It's two hours, then the lambda would be--

AUDIENCE: --15 minutes.

ESTHER DUFLO: The lambda would be gamma times 2 hours.

AUDIENCE: Wouldn't you expect the parameters to vary depending on the other team, like how good they are [INAUDIBLE].

ESTHER DUFLO: Yes. And this is average over 56 game. So it is possible that it's not independent because it's not independent by games that, there are better games and worse games. But people use it. Yeah. That's a-- people use it for soccer matches, soccer goals. So that seems to me to be similar, but you're right, that the Poisson arrival rate is going to depend on-- might depend on the other team. Joseph?

AUDIENCE: [INAUDIBLE] how the distribution changes as you increase the intervals. Like instead of doing it for a game, you do it points per quarter, points per minute, [INAUDIBLE].

ESTHER DUFLO: So the gamma remains-- the gamma is the parameter-- that's why we like-- it's conceptually cleaner to work with gamma because it's-- the gamma is the limit as this interval becomes to-- as interval tends to 0. So the gamma is a fixed parameter. Lambda is going to increase if you're looking by per quarter. For a given gamma, you're going to have a larger lambda if you're working during the entire match, and if you're working during the first half of the match or the second half, lambda is going to increase proportionally to that.

So if you're asking-- so the variant-- the expectation and the variance of the Poisson is necessarily going to depend on what period of time you're thinking about. So it's a 15-minute period, it's a 10-minute period, it's a one-second period. In one second, it would--

AUDIENCE: So if we do that-- like that graphic, would it look more like probably [INAUDIBLE]. more of a Poisson distribution?

ESTHER DUFLO: No. If it is, in fact, Poisson, as you change the length of the interval, it's going to still look like a Poisson, but with a lambda that is smaller and smaller. So for example, if I work with two hours, the expected number of shots over two hours is a gamma times two hours. The rate per whatever it is for the period of two hours. If I work for 15 minutes, it's going to be correspondingly like eight of that.

That makes sense. The number of arrivals is going to depend both on the arrival rate, which is gamma, and-- which is something that is not going to-- that we consider to be fixed. And the period of time over which we are counting the arrivals.

The explanation is related to the Poisson distribution, and it's a waiting time between two events in a Poisson process. So how long do you have to wait between two process to arrive? So this is the exponential we've seen before without motivating it particularly, but you've played with it.

AUDIENCE: Oh, the headache example.

ESTHER DUFLO: The headache example was the motivation. The exponential. So it's fx lambda e to the power of minus-- E to the minus lambda X. One interesting property of-- so this is the expectation of the exponential and the variance. One interesting property of the exponential distribution is that it's memoryless. So it doesn't matter what has happened before.

The probability that X is greater than t is the probability that X is greater than t plus h given that X is greater than h-- should be an h here. Oh, sorry. The probability that X is greater than h is the probability that X is greater than t plus h given that X is greater than t. So when the interval moves, it only depends on the current location of the interval, not on the past.

So the lambda distribution is a specific case of a gamma distribution, which I don't want to get into the detail, which is-- you can think it's the waiting time before a number of occurrences. And we're skipping gamma for today. So those were our discrete distribution. And now we only have two continuous distribution to go through, but they are quite-- we like them a lot.

One we've seen-- we already spend a lot of time with, is the uniform distribution. The uniform distribution is-- I can plot the graphic first. It's the probability-- the probability distribution function is constant over an interval and it's 0 everywhere else. So it's 1 over b minus a over the interval a, b, and it's 0 everywhere else. So that's how it looks like.

That's the mean. You can calculate those very easily. And you can calculate the variance also. I put the one intermediate step. Write down the integral. You can calculate the E of X squared, and then you can calculate the variance. If a is 0 and b equal 1, we call the resulting distribution standard uniform. And one interesting aspect of the standard uniform is if you want a standard uniform, 1 minus u1 is likewise standard uniform.

We use it always. We use the uniform distribution a lot. An important one that we already discussed, and given-- in the interest of time, I'm not going to spend as much time as I have stuff on the graph, but I want you to go over the example on the graph because there is some R coding. An important one is randomly-generating numbers.

As you know, computers don't know how to generate random numbers, but they know how to take a-- draw from a random distribution. So the uniform distribution is very useful when we want to create any simulation. So for example, when you want to create a treated and controlled observation.

And as you know, it turns out, R is all of-- can-- is able to randomly sample from any distribution you'd like. But suppose that this somehow disappear, we know how to go from the lecture and I think from the problem sets, we know how to go from sampling from a uniform distribution to sampling directly-- to sampling from most-- many distributions you might be interested in.

I put in the code here, too, to show you that, in fact, it's true. So let's just go quickly over the exponential one. So in fact, that might have been the example in the problem set. This is the--

AUDIENCE: I think this example is--

ESTHER DUFLO: Yeah.

AUDIENCE: Yeah, in class, actually. Not the R.

ESTHER DUFLO: So this is the R part that-- this is the R equivalent that what we saw in class. Like literally, we just put down, if we have a uniform distribution, we take the inverse CDF, we have the formula for the inverse CDF and then we can sample from it. And when you do that, you do something that looks reasonably like an exponential function.

And the alternative, this is could tell-- you could-- instead of writing down the analytic formula, you could ask LaTeX to go from the uniform to the qx. For any distribution in R, the q gives you the-- it's the inverse quantile function. So the qx is the R formula for-- that we wrote out explicitly in the previous slides.

So you could do inverse PDF that way. And you would get something similar. Or if you were even wanted to go even faster, you could directly sample from the exponential distribution, which is called rex. So you can directly sample an exponential distribution called x. So rex is a sample from the exponential distribution. And it would also look pretty similar. That is fortunate.

Some example you go to, you can go through it at your leisure for Poisson, to go from a uniform distribution to a Poisson. It's not-- it has a bit of a problem towards the left. So there might be an error in this code. So maybe the work is find the error. And then this is what you do if you actually ask R to sample for you from the Poisson distribution, which it can do very well.

Another thing that is underlying this type of work is when you want to choose a random sample. So you have a list-- for example, this chooses a list of-- so the function sample in R samples from a list of 50 states either with replacement or without replacement. So we sample 25 states with replacement, without replacement. At the bottom of that, it's going to ask whether-- that's also what is underlying-- this is also a uniform distribution. And many other things that you can look at.

Let's look at the normal distribution. We'll introduce it now and then we'll use it a lot next time since we are probably directly going to that we can first have a look at it. You will know how it looks like. It looks bell-shaped. Symmetrical. It has thin tail, which is it doesn't have that many observation towards the top. It is-- and where does it come from?

Well, one way to think of where it comes from is that it's the limit of a binomial. So suppose that x is a binomial np, then for any number c and d, the probability that-- the standardized version of this binomial, which is x minus np, which is the mean of the binomial, divided by the square root of the variance of this binomial is between these two intervals.

The limit of that as n goes to infinity is the integral between c and d of this function, function that we call little phi. And that's the normal-- that's the normal-- the standardized normal distribution. We call little phi for this function the density. And we call a big phi for the CDF of it. This is a standard normal. It has an expectation of 0 and a variance of 1.

So here's like a graphical way of how the binomials, as you add-- as you make more and more, intervals will look more and more like a normal distribution. So this is a binomial with a p of 0.5. It doesn't have to be 0.5, but it works better with 0.5. Like when p's are very little, as we discussed before, you would approximate the binomial with a Poisson and not with a normal distribution.

So this is with a small n, with a larger n, with a bigger n. Of course, we need to start by standardizing them. So this is a standardized version. And eventually, it's going to look like-- as you make more and more and more and more interval of your standardized version, it's going to look normal.

So from a normal distribution-- so from the standard normal, you can go to-- you can have any other number of other normal distribution which have a mean mu and standard deviation sigma. So if z is the standard normal, then x equal mu plus sigma z is a normal distribution with a mean mu and sigma, the standard deviation of it, which is the square of the variance.

So this will have-- this will be distributed like over there where we just take phi. Instead of taking phi of x, we take phi of x minus mu over sigma. And we have to standardize by dividing by the standard deviation as well.

One useful property that will be an extremely useful property is that if X1 is normal, then X2-- then the X2 plus bX-- what is it? If X1 and X2 are normal, then X2 plus bX1 are also normal. What did I say? No. X2 equal a plus bX1, sorry. X2 equal a plus bX1 is also normal, which means a plus bX-- bE of X1 and variance b squared of variance of X1.

And more generally, if you have n iid random variable and you are taking the sum of them, then the sum of them is also a normal with the sum-- the expectation is the sum of the expectations and the variance is-- the sum of the sigma squared is the variance. We knew that-- we knew the mean and variance, but the theorem here, which I have not proven, but what's new here is that there are also normally distributed.

It's not just-- so when you take a sum of random of not-- of random variables that are distributed-- that have normal distribution, there are some-- also have normal distributions. And I think I can stop here because it's a very natural way to take it from here. It's just-- I plotted-- so maybe next time we should go over how we use the CDF, I can take five minutes. This is a normal distribution for you with the region-- what fraction of the density is within each of the regions. We'll take from that later. I took a long time doing this graph, so you look at it.