PHILIPPE RIGOLLET: We're talking about goodness-of-fit tests. Goodness-of-fit tests are, does my data come from a particular distribution? And why would we want to know this? Well, maybe we're interested in, for example, knowing if the zodiac signs of the Fortune 500 CEOs are uniformly distributed. Or maybe we actually have slightly more-- slightly deeper endeavors, such as understanding if you can actually apply the t-test by testing normality of your sample.

All right? So we saw that there's the main result-- the main standard test for this. It's called the Kolmogorov-Smirnov test that people use quite a bit. It's probably one of the most used tests out there. And there's other versions of it that I mentioned passing by. There's the Cramer-von Mises, and there's the Anderson-Darling test.

Now, how would you pick one of such tests? Well, they're always are going to-- they're always going to have their advantages and disadvantages. And Kolmogorov-Smirnov is definitely the most widely used because-- well, I guess because it's a natural notion of distance between functions. You just look for each point how far they can be, and you just look at the farthest they can be everywhere.

Now, Cramer-von Mises involves L2 distance. So if you're not used to Hilbert spaces or notions of Euclidean spaces, at least it's a little more complicated. And then Anderson-Darling is definitely even more complicated.

Now, each of these tests is going to be more powerful against other alternatives. So unless you can really guess which alternative you're expecting to see, which you probably don't, because, again, you're in a case where you want to typically declare H0 to be the correct one, then it's really a matter of tossing a coin. Maybe you can run all three of them and just sleep better at night, because all three of them have failed to reject, for example. All right?

So as I mentioned, one of the maybe primary goals to test goodness of fit is to be able to check whether we can apply Student's test, right, and if the Student distribution is actually a valid distribution. And for that, we need to have normally distributed data.

Now, as I said several times, normally distributed, it's not a specific distribution. It's a family of distributions that's indexed by means and variances. And the way I would want to test if a distribution is normally distributed is, well, I would just look at the most natural normal distribution or Gaussian distribution that my data could follow.

That means that's the Gaussian distribution that has the same mean as my data and the same empirical variance as my data, right? And so I'm going to be given some points x1, xn, and I'm going to be asking, are those Gaussian? That means this is equivalent to, say, are they N mu sigma square for some mu sigma squared?

And of course, the natural choice is to take mu hat to be-- mu to be equal to mu hat, which is xn bar. And sigma squared to be sigma squared hat to be, well, Sn hat-- Sn-- what we wrote Sn, which is 1/n sum from i equal 1 to n of xi minus xn bar squared. OK?

So this is definitely the natural one you would want to test. And maybe you could actually just close your eyes and just stuff that in a Kolmogorov-Smirnov test. OK? So here, there's a few things that don't work.

The first one is that Donsker's theorem does not work anymore, right? Donsker's theorem was the one that told us that, properly normalized, this thing would actually converge to the supremum of a Brownian bridge, which is not true. So that's one problem.

But there's actually an even bigger problem is that this distribution, we will check in a second, actually does not-- is pivotal itself, right, the statistic is pivotal. It does not have a distribution that depends on the known parameters, which is sort of nice, at least under the null.

However, the distribution is not the same as the one that had fixed mu and sigma. The fact that they come from some random variables is actually distorting the distribution itself. And in particular, the quantiles are going to be distorted, and we hinted at that last time.

So one other thing I need to tell you, though, is that this thing actually-- so I know there's some-- oh, yeah, that's where there's a word missing. So we compute the quantiles for this test statistic. And so what I need to promise to you is that these quantiles do not depend on any unknown parameter, right? I mean, it's not clear, right?

So I want to test whether my data has some Gaussian distribution. So under the null, all I know is that my xi's are Gaussian with some mean mu and some variance sigma, which I don't know. So it could be the case that when I try to understand the distribution of this quantity under the null, it depends on mu and sigma, which I don't know.

So we need to check that this is the case. And what's actually our redemption here is actually going to be the supremum. The supremum is going to basically allow us to, say, sup out mu and sigma square. So let's check that, right?

So what I'm interested in is this quantity, supremum over t and R of the difference between Fn of t and, what I write, phi mu hat sigma squared of t. So phi mu hat sigma hat squared-- sorry, sigma hat squared-- is the CDF of some Gaussian with mean mu hat and variance sigma hat squared.

And so in particular, this thing here, phi hat of mu hat-- sorry, phi hat of mu hat sigma hat squared of t is the probability that some x is less than t, where x follows some N mu hat sigma hat squared.

So what it means is that by just the translation and scaling trig that we typically do for Gaussian to turn it into some standard Gaussian, that implies that there exists some z, which is standard Gaussian this time, so mean 0 and variance 1, such that x is equal to sigma hat x-- sorry, z plus mu hat.

Agreed? That's basically saying that x has some Gaussian with mean mu and variance sigma squared. And I'm not going to say the hats every single time, OK? So OK, so that's what it means. So in particular, maybe I shouldn't use x here, because x is going to be my actual data. So let me write y. OK?

So now what is this guy here? It's basically-- so phi hat. So this implies that phi mu hat sigma hat squared of t is equal to the probability that sigma hat z plus mu hat is less than t, which is equal to the probability that z is less than t minus mu hat divided by sigma hat, right?

But now when z is the standard normal, this is really just the cumulative distribution function of a standard Gaussian but evaluated at a point which is not t, but t minus mu hat divided by sigma hat. All right? So in particular, what I know-- so from this what I get-- well, maybe I'll remove that, it's going to be annoying-- I know that phi mu hat sigma hat squared-- sorry-- phi mu hat sigma hat squared of t is simply phi of, say, 0, 1.

And that's just the notation. Usually we don't put those, but here it's more convenient. So it's phi 0, 1 of t minus mu hat divided by sigma hat. OK? That's just something you can quickly check. There's this nice way of writing the cumulative distribution function for any mean and any variance in terms of the cumulative distribution function with mean 0 and variance 1. All right? Not too complicated.

All right. So I know what I'm going to say is that, OK, I have this sup here. So what I can write is that this thing here is equal to the sup routine R of 1/n. Let me write what Fn is-- sum from i equal 1 to n of the indicator that xi is less than t minus phi 0, 1 of t minus mu hat divided by sigma hat. OK?

I actually want to make a change of variable so that this thing I'm going to call mu-- u, sorry. OK? And so I'm going to make my life easier, and I'm going to make it appear here. And so I'm just going to replace this by indicator that xi minus mu hat divided by sigma hat less than t minus mu hat divided by sigma hat, which is sort of useless at this point. I'm just making my formula more complicated.

But now I see something here that shows up, and I will call it u, and this is another u. OK? So now what it means is that suping over t, when t ranges from negative infinity to plus infinity, the new range is from negative infinity to plus infinity, right? So this sup, I can actually write-- this suping t I can write as the sup in u, as the indicator that xi minus mu hat divided by sigma hat is less than u minus phi 0, 1 of u.

Now, let's pause for one second. Let's see where we're going. What we're trying to show that this thing does not depend on the unknown parameters, say, mu and sigma, which are the mean and the variance of x under the null. To do that, we basically need to make only quantities that are sort of invariant under these values.

So I tried to make this thing invariant under anything, and it's just really something that depends on nothing. It's the CDF. It doesn't depend on sigma hat and mu hat anymore. But sigma hat and mu hat will depend on mu and sigma, right? I mean, they're actually good estimators of those guys, so they should be pretty close to them. And so I need to make sure that I'm not actually doing anything wrong here.

So the key thing here is going to be to observe that 1/n sum from i equal 1 to n of indicator of xi minus u hat divided by sigma hat less than u, which is the first term that I have in this absolute value, well, this is what-- well, this is equal to 1/n sum from i equal 1 to n of indicator that-- well, now under the null, which is that x follows N mu sigma squared, for some mu and sigma squared that are unknown.

But they are here. They exist. I just don't know what they are. Then xi minus mu can be written as sigma zi plus mu minus mu hat divided by sigma hat, where z is equal to x minus mu divided by sigma, right? That's just the same trick that I wrote here. OK? Everybody agree? So I just standardize-- sorry, z-- yeah, so zi is xi minus mu i minus mu divided by sigma. All right? Just a standardization.

So now once I write this, I can actually divide everybody by sigma. Right? So I just divided on top here and in the bottom here. So now what I need to check is that the distribution of this guy does not depend on mu or sigma. That's what I claim.

What is the distribution of this indicator? It's a Bernoulli, right? And so if I want to understand its distribution, all I need to do is to compute its expectation, which is just the probability that this thing happens.

But the probability that this thing happens is actually now depending on mu and sigma. And the reason is that mu is what? Well, it's x bar-- sorry, yeah, so mu hat-- sorry, is xn bar. So mu hat minus mu, which under the null follows N mu sigma square over n, right? That's the property of the average.

So when I do mu hat minus mu divided by sigma, this thing is what distribution? It's still a normal. It's a linear transformation of a normal. What are the parameters?

AUDIENCE: 0, 1/n.

PHILIPPE RIGOLLET: Yeah, 0, 1/n. But this does not depend on mu or sigma, right? Now, I need to check that this guy does not depend on mu or sigma. What is the distribution of sigma hat over sigma?

AUDIENCE: It's a chi-square, right?

PHILIPPE RIGOLLET: Yeah, it is a chi-square. So this is actually-- sorry, sigma hat squared divided by sigma squared is a chi-square with n minus 1 degrees of freedom. Does not depend on mu or sigma.

AUDIENCE: [INAUDIBLE]

AUDIENCE: [INAUDIBLE]

AUDIENCE: Or sigma hat squared over sigma squared?

PHILIPPE RIGOLLET: Yeah, thank you. So this is actually divided by it. So maybe this guy. Let's write it like that. This is the proper way of writing it. Thank you. Right? So now I have those two things. Neither of them depends on mu or sigma. I these two things. There's just one more thing to check. What is it?

AUDIENCE: That they're independent?

PHILIPPE RIGOLLET: That they're independent, right? Because the dependence in mu and sigma could be hidden in the covariance. It could be the case that the marginal distribution of mu does not depend on mu or sigma, that the marginal distribution of sigma-- of mu hat does not depend on mu and sigma. The marginal distribution of sigma hat does not depend on mu or sigma, but their correlation could depend on mu and sigma.

But we also have that if I look at-- so if I look at-- so since mu hat is independent of sigma hat, it means that the joint distribution of mu hat divided by sigma and sigma hat divided by sigma does not depend on blah, blah, blah, on mu and sigma. OK? Agree? It's not in the individual ones, and it's not in the way they interact with each other. It's nowhere.

AUDIENCE: [INAUDIBLE] independence be [INAUDIBLE] theorem?

PHILIPPE RIGOLLET: Yeah, covariance theorem, right. So that's something we've been using over and again. That's all under the null. If my data is not Gaussian, nothing actually holds. I just use the fact that under the null I'm Gaussian for some mean mu and variance sigma squared.

But that's all I care about. When I'm designing a test, I only care about the distribution under the null, at least to control the type I error. Then to control the type II error, then I cross my fingers pretty hard. OK?

So now this basically implies what's written on the board, that this distribution, this test statistic, does not depend on any unknown parameters. It's just something that's pivotal. In particular, I could go at the back of a book and check if there's a table for the quantiles of these things, and indeed there are.

This is the table that you see. So actually, this is not even in a book. This is in Lilliefors original paper, 1967, as you can tell from the typewriting. And he actually probably was rolling some dice from his office back in the day and was checking that this was-- he simulated it, and this is how he computed those numbers.

And here you also have some limiting distribution, which is not the sup of a Brownian motion over 0, 1 of-- sorry, of a Brownian bridge over 0, 1, which is the one that you would see for the Kolmogorov-Smirnov test, but it's something that's slightly different.

And as I said, these numbers are actually typically much smaller than the numbers you would get, right? Remember, we got something that was about 0.5, I think, or maybe 0.41, for the Kolmogorov-Smirnov test at the same entrance, which means that using Kolmogorov-Lilliefors test it's going to be harder for you not to reject for the same data.

It might be the case that in one case you reject, and in the other one you fail to reject. But the ordering is always that if you fail to reject with Kolmogorov-Lilliefors, you will fail to reject with Kolmogorov-Smirnov, right? There's always one. So that's why people tend to close their eyes and prefer Kolmogorov-Smirnov because it just makes their life easier. OK?

So this is called Kolmogorov-Lilliefors. I think there's actually an E here-- sorry, an I before the E. Doesn't matter too much. OK? Are there any questions? Yes?

AUDIENCE: Is there like a place you can point to like [INAUDIBLE]

PHILIPPE RIGOLLET: Yeah.

AUDIENCE: [INAUDIBLE].

PHILIPPE RIGOLLET: So the fact that it's actually a different distribution is that here-- so if I actually knew what mu and sigma were, I would do exactly the same thing. But here, rather than having this average with mu and sigma, I would just have the-- with mu hat and sigma hat, I would just have the average with mu and sigma. OK?

So what it means is that the key thing is that what I would compare is the 1/n sum of some Bernoullis with parameter. And the parameter here would be the probability that mu-- xi minus mu over sigma is less than u, which is just the probability that phi-- sorry, it's a Bernoulli with probability F of t. Well, let me write what it is, right? So that's minus phi 0, 1 of t. OK?

So that's for the K-S test, and then I sup over t, right? That's what I would have had, because this is actually exactly the right thing. Here I would remove the true mean. I would divide by the true standard deviation. So that would actually end up being a standard Gaussian, and that's why I'm allowed to use phi 0, 1 here. Agreed? And these are Bernoullis because they're just indicators.

What happens in the Kolmogorov-Lilliefors test? Well, here the Bernoulli, the only thing that's going to change is this guy, right? They still have a Bernoulli. It's just that the parameters of the Bernoulli are weird. The parameters of the Bernoulli looks like it's-- it becomes the probability that some N(0, 1) plus some N(0, 1/n), right, divided by some square root of chi-squared n minus 1 divided by n is less than t.

And those things are independent, but those guys are not necessarily independent, right? And so why is this probability changing? Well, because this denominator is actually fluctuating a lot. So that actually makes this probability different. And so that's basically where it comes from, right?

So you could probably convince yourself very quickly that this only makes those guys closer. And why does it make those guys closer? No, sorry. It makes those guys farther, right? And it makes those guys farther for a very clear reason, is that the expectation of this Bernoulli is exactly that guy.

Here I think it's going to be true as well that the expectation of this Bernoulli is going to be that guy, but the fluctuations are going to be much bigger than just the phi of the Bernoulli. Because the first thing I do is I have a random parameter from my Bernoulli, and then I flip the Bernoulli.

So fluctuations are going to be bigger than a Bernoulli. And so when I take the sup, I'm going to have to [INAUDIBLE] them. So it makes things farther apart, which makes it more likely for you to reject. Yeah?

AUDIENCE: You also said that if you compare the same-- if you compare the table and you set at the same level, the Lilliefors is like 0.2, and for the Smirnov is at 0.4.

PHILIPPE RIGOLLET: Yeah.

AUDIENCE: OK. So it means that Lilliefors is harder not to reject?

PHILIPPE RIGOLLET: It means that Lilliefors is harder not to reject, yes, because we reject when we're larger than the number. So the number being smaller with the same data, we might be, right? So basically, it looks like this. What we run-- so here we have the distribution for the-- so let's say this is the density for K-S.

And then we have the density for Kolmogorov-Lilliefors, K-L. OK? And what the density of K-L looks like, it looks like this, right? And so if I want to squeeze in alpha here, I'm going to have to squeeze in-- and I squeeze in alpha here, then this is the quantile of order 1 minus alp-- well, let's say alpha of the K-L. And this is the quantile alpha of K-S.

So now you give me data, and what I do with it, I check whether they're larger than this number. So if I apply K-S, I check whether I'm larger or smaller than this thing. But if I apply Kolmogorov-Lilliefors, I check whether I'm larger or smaller than this thing.

So over this entire range of values for my test statistic-- because it is the same test statistic, I just plugged in mu hat and sigma hat-- for this entire range, the two tests have different outcomes. And this is a big range in practice, right? I mean, it's between-- I mean, it's pretty much at scale here. OK? Any other-- yeah?

AUDIENCE: [INAUDIBLE] when n goes to infinity, the two tests become the same now, right?

PHILIPPE RIGOLLET: Hmmm.

AUDIENCE: Looking at that formula--

PHILIPPE RIGOLLET: Yeah, They should become the same very far. Let me see, though, because-- right. So here we have 8-- so here we have, say, for 0.5, we get 0.886. And for-- oh, I don't have it. Yeah, actually, sorry. So you're right. You're totally right.

This is the Brownian bridge values. Because in the limit by, say, Slutsky-- sorry, I'm lost. Yeah, these are the values that you get for the Brownian bridge. Because in the limit by Slutsky, this thing is going to have no fluctuation, and this thing is going to have no fluctuation. So they're just going to be pinned down, and it's going to look like as if I did not replace anything.

Because in the limit, I know those guys much faster-- the mu hat and sigma hat converge much faster to mu and sigma than the distribution itself, right? So those are actually going to be negligible. You're right. Actually even, I didn't have-- these are actually the numbers I showed you for the bridge, the Brownian bridge, last time, because I didn't have it for the Kolmogorov-Smirnov one. OK?

So there's actually-- so those are numerical ways of checking things, right? I give you data. You just crank the Kolmogorov-Smirnov test. Usually you press a 5 on MATLAB. But let's say you actually compute this entire thing, and there's a number that comes out, and you decide whether it's large enough or small enough.

Of course, statistical software is going to make your life even simpler by spitting out a p-value, because you can-- I mean, if you can compute quantiles, you can also when compute p-values. And so your life is just fairly easy. You just have red is bad, green is good, and then you can go.

The problem is that those are numbers you want to rely on. But let's say you actually reject. Let's say you reject. Your p-value is actually just like slightly below 5%. So you can say, well, maybe I'm just going to change my p-value-- my threshold to 1%, but you might want to see what's happening.

And for that you need a visual diagnostic. Like, how do I check if something departs from being normal, for example? How do I check if a distribution-- why is a distribution not a uniform distribution? Why is a distribution not an exponential distribution? There's many, many, right?

If I have an exponential distribution and half of my values are negative, for example, well, there's like pretty obvious reasons why it should not be exponential. But it could be the case that it's just the tails are little heavier or there's more concentration at some point. Maybe it has two modes. There's things like this.

But the real thing, we don't believe that the Gaussian is so important because it looks like this close to 0. What we like about the Gaussian is that the tails here decay at this rate-- exponential minus x squared over 2 that we described in the maybe first lecture. And in particular, if there were like kinks around here, it wouldn't matter too much. This is not what makes issues for the Gaussian.

And so what we want is to have a visual diagnostic that tells us if the tails of my distribution are comparable to the tails of a Gaussian one, for example. And those are what's called quantile-quantile plots, and in particular-- or QQ plots. And the basic QQ plots we're going to be using are the ones that are called normal QQ plots that are comparing your data to a Gaussian distribution, or a normal distribution.

But in general, you could be comparing your data to any distribution you want. And the way you do this is by comparing the quantiles of your data, the empirical quantiles, to the quantiles of the actual distribution you're trying to compare yourself to.

So this, in a way, is a visual way of performing these goodness-of-fit tests. And what's nice about visual is that there's room for debate. You can see something that somebody else cannot see, and you can always-- because you want to say that things are Gaussian. And we'll see some examples where you can actually say it if you are good at debate, but it's actually going to be clearly not true.

All right. So this is a quick and easy check. That's something I do all the time. You give me data, I'm just going to run this. One of the first things I do so I can check if I can start entering the Gaussian world without compromising myself too much.

And the idea is to say, well, if F is close to-- if F-- if my data comes from an F, and if I know that Fn is close to F, then rather than computing some norm, some number that tells me how far they are, summarizing how far they are, I could actually plot the two functions and see if they're far apart.

So let's think for one second what this kind of a plot would look like. Well, I would go between 0 and 1. That's where everything would happen. Let's say my distribution is the Gaussian distribution. So this is the CDF of N(0, 1). And now I have this guy that shows up, and remember we had this piecewise constant. Well, OK, let's say we get something like this. We get a piecewise constant distribution for Fn, right?

Just from this, and even despite my bad skills at drawing, it's clear that it's going to be hard for you to distinguish those two things, even for a fairly large amount of points. Because the problem is going to happen here, and those guys look pretty much the same everywhere you are here. You're going to see differences maybe in the middle, but we don't care too much about those differences.

And so what's going to happen is that you're going to want to compare those two things. And this is basically you have the information you want, but visually it just doesn't render very well because you're not scaling things properly.

And the way we actually do it is by flipping things around. And rather than comparing the plot of F to the plot of Fn, we compare the plot of Fn inverse to the plot of F inverse. Now, if F goes from the real line to the interval 0, 1, F inverse goes from 0, 1 to the whole real line. So what's going to happen is that I'm going to compare things on some intervals, which is the-- which are the entire real line.

And then what values should I be looking at those things at? Well, technically for F, if F is continuous I could look at F inverse for any value that I please, right? So I have F. And if I want to look at F inverse, I pick a point here and I look at the value that it gives me, and that's F inverse of, say, u, right, if this is u. And I could pick any value I want, I'm going to be able to find it.

The problem is that when I start to have this piecewise constant thing, I need to decide what value I assign for anything that's in between two jumps, right? And so I can choose whatever I want, but in practice it's just going to be things that I myself decide. Maybe I can decide that this is the value. Maybe I can decide that the value is here.

But for all these guys, I'm going to pretty much decide always the same value, right? If I'm in between-- for this value u, for this jump the jump is here. So for this value, I'm going to be able to decide whether I want to go above or below, but it's always this value that's going to come out.

So rather than picking values that are in between, I might as well just pick only values for which this is the value that it's going to get. And those values are exactly 1/n, 2/n, 3/n, 4/n. It's all the way to n/n, right? That's exactly where the flat parts are. We know we jump from 1/n every time.

And so that's exactly the recipe. It says look at those values, 1/n, 2/n, 3/n until, say, n minus 1 over n. And for those values, compute the inverse of both the empiricial CDF and the true CDF. Now, for the empirical CDF, it's actually easy. I just told you this is basically where the points-- where the jumps occur. And the jumps occur where? Well, exactly at my observations.

Now, remember I need to sort those observations to talk about them. So the one that occurs for the i-th jump is the i-th largest observation, which we denoted by X sub (i). Remember? We had this formula that we said, well, we have x1, xn. These are my data.

And what I'm going to sort them into is x sub (1), which is less than or equal to x sub (2), which is less than x sub (n). OK? So we just ordered them from smallest to largest. And then now we've done that, we just put this parenthesis notation. So in particular, Fn inverse of i/n is the location where the i-th jumps occur, which is the i-th largest observation. OK?

So for this guy, these values, the y-axes are actually fairly easy. I know it's basically my ordered observations. The x-values are-- well, that depends on the function F I'm trying to test. If it's the Gaussian, it's just the quantile of order 1 minus 1/n, right?

It's this Q1 minus 1/n here that I need to compute. It's the inverse of the cumulative distribution function, which, given the formula for F, you can actually compute or maybe estimate fairly well. But it's something that you can find in tables. Those are basically quantiles. Inverse of CDFs are quantiles, right?

And so that's basically the things we're interested in. That's why it's called quantile-quantile. Those are sometimes referred to as theoretical quantiles, the one we're trying to test, and empirical quantiles, the one that corresponds to the empirical CDF.

And so I'm plotting a plot where the x-axis is quantile. The y-axis is quantile. And so I call this plot a quantile-quantile plot, or QQ plot, because, well, just say 10 times quantile-quantile, and then you'll see why. Yeah?

AUDIENCE: [INAUDIBLE] have to have the [INAUDIBLE]?

PHILIPPE RIGOLLET: Well, that's just-- we're back to the-- we're back to the goodness-of-fit test, right? So if you look-- so you don't do it yourself. That's the simple answer. You don't-- I'm just telling you how those plots are going to be seen spit out from a software are going to look like.

Now, depending on the software, there's a different thing that's happening. Some softwares are actually plotting F with the right-- let's say you want to do normal, as you asked. So some software are just going to use F to be with mu hat and sigma hat, and that's fine. Some software are actually not going to do this. They're just going to use a Gaussian. But then they're going to actually have a different reference point.

So what do we want to see here? What should happen if all these points-- if all my points actually come from F, from a distribution that has CDF F? What should happen? What should I see? Well, since Fn should be close to F, Fn inverse should be close to F inverse, which means that this point should be close to that point. This point should be close to that point.

So ideally, if I actually pick the right F, I should see a plot that looks like this, something where all my points are very close to the line y is equal to x, right? And I'm going to have some fluctuations, but something very close to this.

Now, that's if F is exactly the right one. If F is not exactly the right one, in particular, in the case of a Gaussian one, if I actually plotted here the quantiles-- so if I plotted F 0, 1 of t, right? So let's say those are the ones I actually plot, but I really don't know what-- mu hat is not 0 and sigma hat is not 0. And so this is not the one I should be getting.

Since we actually know that phi of mu hat sigma hat squared t is equal to phi 0, 1 of t minus mu hat divided by sigma hat, there's just this change of axis, which is actually very simple. This change of axis is just a simple translation scaling, which means that this line here is going to be transformed into another line with a different slope and a different intercept.

And so some software will actually decide to go with this curve and just show you what the reference curve should be, rather than actually putting everything back onto the 45-degree curve.

AUDIENCE: So if you get any straight line?

PHILIPPE RIGOLLET: Any straight line, you're happy. I mean, depending on the software. Because if the software actually really rescaled this thing to have mu hat and sigma square and you find a different line, a different straight line, this is bad news, which is not going to happen actually. It's impossible that happens, because you actually-- well, it could. If it's crazy, it could. It shouldn't be very crazy.

OK. So let's see what R does for us, for example. So here in R, R actually does this funny trick where-- so here I did not actually plot the lines. I should actually add the lines. So the command is like qqnorm of my sample, right?

And that's really simple. I just stack all my data into some vector, say, x. And I say qqnorm of x, and it just spits this thing out. OK? Very simple. But I could actually add another command, which I can't remember. I think it's like qqline, and it's just going to add the line on top of it.

But if you see, actually what R does for us, it's actually doing the translation and scaling on the axes themselves. So it actually changes the x and y-axis in such a way that when you look at your picture and you forget about what the meaning of the axes are, the relevant straight line is actually still the 45-degree line. It's Because it's actually done the change of units for you.

So you don't have to even see the line. You know that, in your mind, that this is basically-- the reference line is still 45 degree because that's the way the axes are made. But if I actually put my axes, right-- so here, for example, it goes from-- let's look at some-- well, OK, those are all square. Yeah, and that's probably because they actually have-- the samples are actually from a standard normal.

So I did not make my life very easy to illustrate your question, but of course, I didn't know you were going to ask it. Next time, let's just prepare. Let's script more. We'll see another one in the next plot.

But so here what you expect to see is that all the plots should be on the 45-degree line, right? This should be the right one. And if you see, when I start having 10,000 samples, this is exactly what's happening. So this is as good as it gets. This is an N(0, 1) plotted against the theoretical quantile of an N(0, 1). As good as it gets.

And if you see, for the second one, which is 50, sample size of size-- sample of size 50, there is some fudge factor, right? I mean, those things-- doesn't look like there's a straight line, right? It sort of appears that there are some weird things happening here at the lower tail.

And the reason why this is happening is because we're trying to compare the tails, right? When I look at this picture, the only thing that goes wrong somehow is always at the tip, because those are sort of rare and extreme values, and they're sort of all over the place. And so things are never really super smooth and super clean.

So this is what your best shot is. This is what you will ever hope to get. So size 10, right, so you have 10 points. Remember, we actually-- well, I didn't really tell you how to deal with the extreme cases. Because the problem is that F inverse of 1 for the true F is plus infinity. So you have to make some sort of weird boundary choices to decide what F inverse of 1 is, and it's something that's like somewhere.

But you still want to put like 10 dots, right? 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 dots. So I have 10 observations, you will see 10 dots. I have 50 observations, you will see 50 dots, right, because I have-- there are 1/n, 2/n, 3/n all the way to n/n. I didn't tell you the last one.

OK. So this is when things go well, and this is when things should not go well. OK? So here, actually, the distribution is a Student's t with 15 degrees of freedom, which should depart somewhat from a Gaussian distribution. The tails should be heavier.

And what you can see is basically the following, is that for 10 you actually see something that's crazy, right, if I do 10 observations. But if I do 50 observations, honestly, it's kind of hard to say that it's different from the standard normal.

So you could still be happy with this for 100. And then this is what's happening for 10,000. And even here it's not the beautiful straight line, but it feels like you would be still tempted to conclude that it's a beautiful straight line.

So let's try to guess. So basically, there's-- for each of those sides there's two phenomena. Either it goes like this or it goes like this, and then it goes like this or it goes like this. Each side corresponds to the left tail, all the smallest values. So that's the left side. And that's the right side-- corresponds to the large values. OK?

And so basically you can actually think of some sort of a table that tells you what your QQ plot looks like. And so let's say it looks-- so we have our reference 45-degree line. So let's say this is the QQ plot. That could be one thing. This could be the QQ plot where I have another thing. Then I can do this guy, and then I do this guy. So this is like this. OK? So those are the four cases. OK?

And here what's changing is the right tail, and here what's changing is the-- and when I go from here to here, what changes is the left tail. Is that true? No, sorry. What changes here is the right tail, right? It's this part that changes from top to bottom. So here it's something about right tail, and here that's something about left tail. Everybody understands what I mean when I talk about tails? OK.

And so here it's just going to be a question of whether the tails are heavier or lighter than the Gaussian. Everybody understand what I mean when I say heavy tails and light tails? OK. So right, so heavy tails just means that basically here the tails of this guy are heavier than the tails of this guy. So it means that if I draw them, they're going to be above. Actually, I'm going to keep this picture because it's going to be very useful for me.

When I plug the quantiles at the same-- so let's look at the right tail, for example. Right here my picture is for right tails. When I look at the quantiles of my theoretical distribution-- so here you can see the bottom curve we have the theoretical quantiles, and those are the empirical quantiles.

If I look to the right here, are the theoretical quantiles larger or smaller than the empirical quantiles? Let me phrase it the other-- are the empirical quantiles larger or smaller than the theoretical quantiles?

AUDIENCE: This is a graph of quantiles, right? So if it's [INAUDIBLE] it should be smaller.

PHILIPPE It should be smaller, right? On this line, they are equal. So if I see the empirical quantile

RIGOLLET: showing up here, it means that here the empirical quantile is less than the theoretical quantile. Agree? So that means that if I look at this thing-- and that's for the same values, right? So the quantiles are computed for the same values i/n.

So it means that the empirical quantiles should be looking-- so that should be the empirical quantile, and that should be the theoretical quantile. Agreed? Those are the smaller values for the same alpha. So that implies that the tails-- the right tail, is it heavy or lighter-- heavier or lighter than the Gaussian?

AUDIENCE: Lighter.

PHILIPPE RIGOLLET: Lighter, right? Because those are the tails of the Gaussian. Those are my theoretical quantiles. That means that this is the tail of my empirical distribution. So they are actually lighter. OK? So here, if I look at this thing, this means that the right tail is actually light. And by light, I mean lighter than Gaussian. Heavy, I mean heavier than Gaussian. OK?

OK, now we can probably do the entire thing. Well, if this is light, this is going to be heavy, right? That's when I'm above the curve. Exercise-- is this light or is this heavy, the first column? And it's OK. It should take you at least 30 seconds.

AUDIENCE: [INAUDIBLE] different column?

PHILIPPE RIGOLLET: Yeah, this column, right? So this is something that pertains-- this entire column is going to tell me whether the fact that this guy is above, does this mean that I have lighter or heavier left tails?

AUDIENCE: Well, on the left, it's heavier.

PHILIPPE RIGOLLET: On the left, it's heavier. OK. I don't know. Actually, I need to draw a picture. You guys are probably faster than I am.

AUDIENCE: [INTERPOSING VOICES].

PHILIPPE RIGOLLET: Actually, let me check how much randomness is-- who says it's lighter? Who says it's heavier?

AUDIENCE: Yeah, but we're biased.

AUDIENCE: [INAUDIBLE]

PHILIPPE RIGOLLET: Yeah, OK.

AUDIENCE: [INAUDIBLE]

PHILIPPE RIGOLLET: All right. So let's see if it's heavier. So we're on the left tail, and so we have one looks like this, one looks like that, right? So we know here that I'm looking at this part here. So it means that here my empirical quantile is larger than the theoretical quantile. OK? So are my tails heavier or lighter? They're lighter. That was a bad bias.

AUDIENCE: [INAUDIBLE]

PHILIPPE RIGOLLET: Right? It's below, so it's lighter. Because the problem is that larger for the negative ones means that it's smaller [INAUDIBLE], right? Yeah?

AUDIENCE: Sorry but, what exactly are these [INAUDIBLE]? If this is the inverse-- if this is the inverse CDF, shouldn't everything-- well, if this is the inverse CDF, then you should only be inputting values between 0 and 1 in it. And--

PHILIPPE RIGOLLET: Oh, did I put the inverse CDF?

AUDIENCE: Like on the previous slide, I think.

PHILIPPE RIGOLLET: No, the inverse CDF, yeah, so I'm inputting--

AUDIENCE: Oh, you're [INAUDIBLE].

PHILIPPE RIGOLLET: Yeah, so it's a scatter plot, right? So each point is attached-- each point is attached 1/n, 2/n, 3/n. Now, for each point I'm plotting, that's my x-value, which maps a number between 0 and 1 back onto the entire real line, and my y-value is the same. OK?

So what it means is that those two numbers, this is in the-- this lives on the entire real line, not on the interval. This lives on the entire real line, not in the interval. And so my QQ plots take values on the entire real line, entire real line, right?

So you think of it as a parameterized curve, where the time steps are 1/n, 2/n, 3/n, and I'm just like putting a dot every time I'm making one step. OK? OK, so what did we say? That was lighter, right?

AUDIENCE: [INAUDIBLE]

PHILIPPE RIGOLLET: OK? One of my favorite exercises is, here's a bunch of densities. Here's a bunch of QQ plots. Map the correct QQ plot to its own density. All right? And there won't be mingled lines that allow you to do that, then you just have to follow, like at the back of cereal boxes.

All right. Are there any questions? So one thing-- there's two things I'm trying to communicate here is if you see a QQ plot, now you should understand, one, how it was built, and two, whether it means that you have heavier tails or lighter tails.

Now, let's look at this guy. What should we see? We should see heavy on the left and heavy on the right, right? We know that this should be the case. So this thing actually looks like this, and it sort of does, right? If I take this line going through here, I can see that this guy's tipping here, and this guy's dipping here.

But honestly-- actually, I can't remember exactly, but t 15, if I plotted the density on top of the Gaussian, you can see a difference. But if I just gave it to you, it would be very hard for you to tell me if there's an actual difference between t 15 and Gaussian, right? Those things are actually very close.

And so in particular, here we're really trying to recognize what the shape is the fact-- right? So t 15 compared to a standard Gaussian was different, but t 15 compared to a Gaussian with a slightly larger variance is not going to actually-- you're not going to see much of a difference.

So in a way, such distributions are actually not too far from the Gaussian, and it's not too-- it's still pretty benign to conclude that this was actually a Gaussian distribution because you can just use the variance as a little bit of a buffer.

I'm not going to get really into how you would use a t-distribution into a t-test, because it's kind of like *Inception,* right? So but you could pretend that your data actually is t-distributed and then build a t-distribution from it, but let's not say that. Maybe that was a bad example.

But there's like other heavy-tailed distributions like Cauchy distribution, which doesn't even have a-- it's not even integrable because that's as heavy as the tails get. And this you can really tell it's going to look like this. It's going to be like pfft.

What does a uniform distribution look like? Like this? It's going to be-- it's going to look like a Gaussian one, right? So a uniform-- so this is my Gaussian. A uniform is basically going to look like this, one side take the right mean and the right variance, right?

So the tails are definitely lighter. They're 0. That's as lighter as it gets. So the light-light is going to look like this S shape. So an S-- light-tailed distribution has this S shape. OK?

What is the exponential going to look like? So the exponential is positively supported. It only has positive numbers. So there's no left tail. This is also as light as it gets. But the right tail, is it heavier or lighter than the Gaussian?

AUDIENCE: Heavier.

PHILIPPE RIGOLLET: It's heavier, right? It's only the case like e of the minus x rather e to the minus x squared. So it's heavier. So it means that on the left it's going to be light, and on the right it's going to be heavy. So it's going to be U-shaped. OK? That will be fine.

All right. Any other question? Again, two messages, like, more technical, and you can sort of fiddle with it by looking at it. You can definitely conclude that this is OK enough to be Gaussian for your purposes. Yeah?

AUDIENCE: So [INAUDIBLE]

PHILIPPE RIGOLLET: I did not hear the "if" at the beginning of your sentence.

AUDIENCE: I would want to be lighter tail, right, because that'll be-- it's easier to reject? Is that correct?

PHILIPPE RIGOLLET: So what is your purpose as a--

AUDIENCE: I want to-- I have some [INAUDIBLE] right? I want to be able to say I reject H0 [INAUDIBLE].

PHILIPPE RIGOLLET: Yes.

AUDIENCE: So if you wanted to make it easier to reject H0, then--

PHILIPPE RIGOLLET: Yeah, in a way that's true, right? So once you've actually factored in the mean and the variance, the only thing that actually-- right. So if you have Gaussian tails or lighter-- even lighter tails, then it's harder for you to explain deviations from randomness only, right?

If you have a uniform distribution and you see something which is-- if you're uniform on 0, 1 plus some number and you see 25, you know this number is not going to be 0, right? So that's basically as good as it gets. And there's basically some smooth interpolation if you have lighter tails.

Now, if you start having something that has heavy tails, then it's more likely that pure noise will generate large observations and therefore discovery. So yes, lighter tails is definitely the better-behaved noise. Let's put it this way. The lighter it is, the better behaved it is.

Now, this is good-- this is good for some purposes, but when you want to compute actual quantiles, like exact quantiles, then it is true in general that the quantiles of lighter-tail distributions are going to be dominated by the-- are going to be dominated by the-- let's say on the right tails, are going to be dominated by those of a heavy distribution. That is true.

But that's not always the case. And in particular, there's going to be some like sort of weird points where things are actually changing depending on what level you're actually looking at those things, maybe 5% or 10%, in which case things might be changing a little bit.

But if you started going really towards the tail, if you start looking at levels alpha which are 1% or 0.1%, it is true that it's always-- if you can actually-- so if you see something that looks light tail, you definitely do not want to conclude that it's Gaussian.

You want to actually change your modeling so that it makes your life even easier. And you actually factor in the fact that you can see that the noise is actually more benign than you would like it to be. OK? Stretching fingers, that's it? All right.

OK. So I want to-- I mentioned at some point that we had this chi-square test that was showing up. And I do not know what I did-- let's just-- oh, yeah. So we have this chi-square test that we worked on last time, right?

So the way I introduced the chi-square test is by saying, I am fascinated by this question. Let's check if it's correct, OK? Or something maybe slightly deeper-- let's check if juries in this country are representative of racial distribution.

But you could actually-- those numbers here come from a very specific thing. That was the uniform. That was our benchmark. Here's the uniform. And there was this guy, which was a benchmark, which was the actual benchmark that we need to have for this problem. And those things basically came out of my hat, right? Those are numbers that exist.

But in practice, you actually make those numbers yourself. And the way you do it is by saying, well, if I have a binomial distribution and I want to test if my data comes from a binomial distribution, you could ask this question, right?

You have a bunch of data. I did not promise to you that this was the sum of independent Bernoullis and [INAUDIBLE]. And then you can actually check that it's a binomial indeed, and you have binomial.

If you think about where you've encountered binomials, it was mostly when you were drawing balls from urns, which you probably don't do that much in practice. OK? And so maybe one day you want to model things as a binomial, or maybe you want to model it as a Poisson, as a limiting binomial, right?

People tell you photons arrive-- the rate of a photon hitting some surface is actually a Poisson distribution, right? That's where they arise a lot in imaging. So if I have a colleague who's taking pictures of the skies over night, and he's like following stars and it's just like moving around with the rotation of the Earth.

And he has to do this for like eight hours because he needs to get enough photons over this picture to actually arise. And he knows they arrive at like a Poisson process, and you know, chapter 7 of your probability class, I guess. And

And there's all these distributions outside the classroom you probably want to check that they're actually correct. And so the first one you might want to check, for example, is a binomial. So I give you a distribution, a binomial distribution on, say, K trials, and you have some number p.

And here, I don't know typically what p should be, but let's say I know it or estimate it from my data. And here, since we're only going to deal with asymptotics, just like it was the case for the Kolmogorov-Smirnov one, in the asymptotic we're going to be able to think of the estimated p as being a true p, OK, under the null at least.

So therefore, each outcome, I can actually tell you what the probability of a binomial-- is this outcome. For a given K and a given p, I can tell you exactly what a binomial should give you as the probability for the outcome. And that's what I actually use to replace the numbers 1/12, 1/12, 1/12, 1/12 or the numbers 0.72, 0.7, 0.12, 0.9. All these numbers I can actually compute using the probabilities of a binomial, right?

So I know, for example, that the probability that a binomial np is equal to, say, K is n choose K p to the K 1 minus p to the n minus K. OK? I mean, so these are numbers. If you give me p and you give me n, I can compute those numbers for all K from 0 to n.

And from this I can actually build a table. All right? So for each K-- 0. So K is here, and from 0, 1, et cetera, all the way to n, I can compute the true probability, which is the probability that my binomial np is equal to 0, the probability that my binomial is equal to 1, et cetera, all the way to n. I can compute those numbers. Those are actually going to be exact numbers, right? I just plug in the formula that I had.

And then I'm going to have some observed. So that's going to be p hat, 0, and that's basically the proportion of 0's, right? So here you have to remember it's not a one-time experiment like you do in probability where you say, I'm going to draw n balls from an urn, and I'm counting how many-- how many I have.

This is statistics. I need to be able to do this experiment many times so I can actually, in the end, get an idea of what the proportion of p's is. So you have not just one binomial, but you have n binomials. Well, maybe I should not use n twice. So that's why it's the K here, right? So I have a binomial [INAUDIBLE] at Kp and I just seize n of those guys.

And with this n of those guys, I can actually estimate those probabilities. And what I'm going to want to check is if those two probabilities are actually close to each other. But I already know how to do this. All right?

So here I'm going to test whether P is in some parametric family, for example, binomial or not binomial. And testing-- if I know that it's a binomial [INAUDIBLE], and I basically just have to test if P is the right thing. OK? Oh, sorry, I'm actually lying to you here. OK. I don't want to test if it's binomial. I want to test the parameter of the binomial here. OK? So I know-- no, sorry, [INAUDIBLE] sorry.

OK. So I want to know if I'm in some family, the family of binomials, or not in the family of binomials. OK? Well, that's what I want to do. And so here H0 is basically equivalent to testing if the pj's are the pj's that come from the binomial. And the pj's here are the probabilities that I get. This is the probability that I get j successes. That's my pj. That's j's value here.

OK? So this is the example, and we know how to do this. We construct p hat, which is the estimated proportion of successes from the observations. So here now I have n trials. This is the actual maximum likelihood estimator.

This becomes a multinomial experiment, right? So it's kind of confusing. We have a multinomial experiment for a binomial distribution. The binomial here is just a recipe to create some test probabilities. That's all it is. The binomial here doesn't really matter. It's really to create the test probabilities.

And then I'm going to define this test statistic, which is known as the chi-square statistic, right? This was the chi-square test. We just looked at sum of the square root of the differences. Inverting the covariance matrix or using the Fisher information with removing the part that was not invertible led us to actually use this particular value here, and then we had to multiply by n. OK? And that, we know, converges to what? A chi-square distribution.

So I'm not going to go through this again. I'm just telling you you can use the chi-square that we've seen, where we just came up with the numbers we were testing. Those numbers that were in this row for the true probabilities, we came up with them out of thin air.

And now I'm telling you you can actually come up with those guys from a binomial distribution or a Poisson distribution or whatever distribution you're happy with. Any question? So now I'm creating this thing, and I can apply the entire theory that I have for the chi-square and, in particular, that this thing converges to a chi-square.

But if you see, there's something that's different. What is different? The degrees of freedom. And if you think about it, again, the meaning of degrees of freedom. What does this word-- these words actually mean?

It means, well, to which extent can I play around with those values? What are the possible values that I can get? If I'm not equal to this particular value I'm testing, how many directions can I be different from this guy? And when we had a given set of values, we could be any other set of values, right?

So here, I had this-- I'm going to represent-- this is the set of all probability distributions of vectors of size K. So here, if I look at one point in this set, this is something that looks like p1 through pK such that their sum-- such that they're non-negative, and the sum p1 through pK is equal to 1. OK? So I have all those points here. OK?

So this is basically the set that I had before. I was testing whether I was equal to this one guy, or if I was anything else. And there's many ways I can be anything else. What matters, of course, is what's around this guy that I could actually confuse myself with. But there's many ways I can move around this guy. Agreed?

Now I'm actually just testing something very specific. I'm saying, well, now the piece that I have had to come from this-- have to be constructed from this formula, this parametric family P of theta. And there's a fixed way for-- let's say this is theta, so I have a theta here. There's not that many ways this can actually give me a set of probabilities, right? I have to move to another theta to actually start being confused.

And so here the number of degrees of freedom is basically, how can I move along this family? And so here, this is all the points, but there might be just the subset of the points that looks like this, just this curve, not the half of this thing. And those guys on this curve are the p thetas, and that's for all thetas when theta runs across data.

So in a way, this is just a much smaller dimensional thing. It's a much smaller object. Those are only the ones that I can create that are exactly of this very specific parametric form. And of course, not all are of this form. Not all probability PMFs are of this form. And so that is going to have an effect on what my PMF is going to be-- sorry, on what my-- sorry, what my degrees of freedoms are going to be.

Because when this thing is very small, that means when-- that's happening when theta is actually, say, a one-dimensional space, then there's still many ways I can escape, right? I can be different from this guy in pretty much every other direction, except for those two directions, just when I move from here or when I move in this direction.

But now if this thing becomes bigger, your theta is, say, two dimensional, then when I'm here it's becoming harder for me to not be that guy. If I want to move away from it, then I have to move away from the board. And so that means that the bigger the dimension of my theta, the smaller the degrees of freedoms that I have, OK, because moving out of this parametric family is actually very difficult for me.

So if you think, for example, as an extreme case, the parametric family that I have is basically all PMFs, all of them, right? So that's a stupid parametric family. I'm indexed by the distribution itself, but it's still finite dimensional.

Then here, I have basically no degrees of freedom. There's no way I can actually not be that guy, because this is everything I have. And so you don't have to really understand how the computation comes into the numbers of dimension and what I mean by dimension of this current space.

But really, what's important is that as the dimension of theta becomes bigger, I have less degrees of freedom to be away from this family. This family becomes big, and it's very hard for me to violate this. So it's actually shrinking the number of degrees of freedom of my chi-square. And that's all you need to understand.

When d increases, the number of degrees of freedom decreases. And I'd like to you to have an idea of why this is somewhat true, and this is basically the picture you should have in mind. OK. So now once I have done this, I can just construct. So here I need to check. So what is d in the case of the binomial?

AUDIENCE: 1.

PHILIPPE RIGOLLET: 1, right? It's just a one-dimensional thing. And for most of the examples we're going to have it's going to be one dimensional. So we have this weird thing. We're going to have K minus 2 degrees of freedom.

So now I have this thing, and I have this asymptotic. And then I can just basically use a test that has-- that uses the fact that the asymptotic distribution is this. So I compute my quantiles out of this.

Again, I made the same mistake. This should be q alpha, and this should be q alpha. So that's just the tail probability is equal to alpha when I'm on the right of q alpha. And so those are the tail probability of the appropriate chi-square with the appropriate number of degrees of freedom. And so I can compute p-values, and I can do whatever I want. OK? So then I just like [INAUDIBLE] my testing machinery.

OK? So now I know how to test if I'm a binomial distribution or not. Again here, testing if I'm a binomial distribution is not a simple goodness of fit. It's a composite one where I can actually-- there's many ways I can be a binomial distribution because there's as many as there is theta. And so I'm actually plugging in the theta hat, which is estimated from the data, right?

And here, since everything's happening in the asymptotics, I'm not claiming that Tn has a pivotal distribution for finite n. That's actually not true. It's going to depend like crazy on what the actual distribution is. But asymptotically, I have a chi-square, which obviously does not depend on anything [INAUDIBLE]. OK? Yeah?

AUDIENCE: So in general, for the binomial [INAUDIBLE] trials. But in the general case, the number of-- the size of our PMF is the number of [INAUDIBLE].

PHILIPPE RIGOLLET: Yeah.

AUDIENCE: So let's say that I was also uncertain about what K was so that I don't know how big my [INAUDIBLE] is. [INAUDIBLE]

PHILIPPE RIGOLLET: That is correct. And thank you for this beautiful segue into my next slide. So we can actually deal with the case not only where it's infinite, which would be the case of Poisson. I mean, nobody believes I'm going to get an infinite number of photons in a finite amount of time.

But we just don't want to have to say there's got to be a-- this is the largest possible number. We don't want to have to do that. Because if you start doing this and the probabilities become close to 0, things become degenerate and it's an issue.

So what we do is we bin. We just bin stuff. OK? And so maybe if I have a binomial distribution with, say, 200,000 possible values, then it's actually maybe not the level of precision I want to look at this. Maybe I want to bin. Maybe I want to say, let's just think of all things that are between 0 and 100 to be the same thing, between 100 and 200 the same thing, et cetera.

And so in fact, I'm actually going to bin. I don't even have to think about things that are discrete. I can even think about continuous cases. And so if I want to test if I have a Gaussian distribution, for example, I can just approximate that by some, say, piecewise constant function that just says that, well, if I have a Gaussian distribution like this, I'm going to bin it like this.

And I'm going to say, well, the probability that I'm less than this value is this. The probability that I'm between this and this value is this. The probability I'm between this and this value is this, and then this and then this, right?

And now I've turned-- I've discretized, effectively, my Gaussian into a PMF. The value-- this is p1. The value here is p1. This is p2. This is p3. This is p4. This is p5 and p6, right? I have discretized my Gaussian into six possible values. That's just the probability that they fall into a certain bin.

And we can do this-- if you don't know what K is, just stop at 10. You look at your data quickly and you say, well, you know, I have so few of them that are-- like I see maybe one 8, one 11, and one 15. Well, everything that's between 8 and 20 I'm just going to put it in one bin. Because what else are you going to do? I mean, you just don't have enough observations.

And so what we do is we just bin everything. So here I'm going to actually be slightly abstract. Our bins are going to be intervals Aj. So here-- they don't even have to be intervals. I could go crazy and just like call the bin this guy and this guy, right? That would make no sense, but I could do that.

And then I'm-- and of course, you can do whatever you want, but there's going to be some consequences in the conclusions that you can take, right? All you're going to be able to say is that my distribution does not look like it could be binned in this way. That's all you're going to be able to say.

So if you decide to just put all the negative numbers and the positive numbers, then it's going to be very hard for you to distinguish a Gaussian from a random variable that takes values of minus 1 and plus 1 only. You need to just be reasonable.

OK? So now I have my pj's become the probability that my random variable falls into bin j. So that's pj of theta under the parametric distribution. For the true one, whether it's parametric or not, I have a pj.

And then I have p hat j, which is the proportion of observations that falls in this bin. All right? So I have a bunch of observations. I count how many of them fall in this bin. I divide by n, and that tells me what my estimated probability for this bin is.

And theta hat, well, it's the same as before. If I'm in a parametric family, I'm just estimating theta hat, maybe the maximum likelihood estimator, plug it in, and estimate those pj's of theta hat. From this, I form my chi-square, and I have exactly the same thing as before. So the answer to your question is, yes, you bin. And it's the answer to even more questions.

So that's why there you can actually use the chi-square test to test for normality. Now here it's going to be slightly weaker, because there's only an asymptotic theory, whereas Kolmogorov-Smirnov and Kolmogorov-Lilliefors work actually even for finite samples. For the chi-square test, it's only asymptotic.

So you just pretend you actually know what the parameters are. You just stuff them into a theta, a mu hat, and sigma square hat. And you just go to-- you just cross your finger that n is large enough for everything to have converged by the time you make your decision. OK?

And then this is a copy/paste, with the same error actually as the previous slide, where you just build your test based on whether you exceed or not some quantile, and you can also compute some p-value. OK?

AUDIENCE: The error?

PHILIPPE RIGOLLET: I'm sorry?

AUDIENCE: What's the error?

PHILIPPE RIGOLLET: What is the error?

AUDIENCE: You said [INAUDIBLE] copy/paste [INAUDIBLE].

PHILIPPE RIGOLLET: Oh, the error is that this should be q alpha, right?

AUDIENCE: OK.

PHILIPPE RIGOLLET: I've been calling this q alpha. I mean, that's my personal choice, because I don't want to-- I only use q alpha. So I only use quantiles where alpha is to the right, so. That's what statisticians-- probabilists would use this notation.

OK. And so some questions, right? So of course, in practice you're going to have some issues which translate. I say, well, how do you pick this guy, this K? So I gave you some sort of a-- I mean, the way we discussed, right? You have 8 and 10 and 20, then it's ad hoc. And so depending on whether you want to stop K at 20 or if you want to bin those guys is really up to you.

And there's going to be some considerations about the particular problem at hand. I mean, is it coarse-- too coarse for your problem to decide that the observations between 8 and 20 are the same? It's really up to you. Maybe that's actually making a huge difference in terms of what phenomenon you're looking at.

The choice of the bins, right? So here there's actually some sort of rules, which are don't use only one bin and make sure there's actually-- don't use them too small so that there's at least one observation per bin, right? And it's basically the same kind of rules that you would have to build a histogram. If you were to build a histogram for your data, you still want to make sure that you bin in an appropriate fashion. OK?

And there's a bunch of rule of thumbs. Every time you ask someone, they're going to have a different rule of thumb, so just make your own. And then there's the computation of pj of theta, which might be a bit complicated because, in this case, I would have to integrate the Gaussian between this number and this number. So for this case, I could just say, well, it's the difference of the CDF in that value and that value and then be happy with it.

But you can imagine that you have some slightly more crazy distributions. You're going to have to somewhat compute some integrals that might be unpleasant for you to compute. OK? And in particular, I said the difference of the PDF between that value and that value of-- sorry, the CDF between that value and that value, it is true.

But it's not like you actually have tables that compute the CDF at any value you like, right? You have to sort of-- well, there might be but at some degree, but you are going to have to use a computer typically to do that. OK?

And so for example, you could do the Poisson. If I had time, if I had more than one minute, I would actually do it for you. But it's basically the same. The Poisson, you are going to have an infinite tail, and you just say, at some point I'm going to cut everything that's larger than some value. All right?

So you can play around, right? I say, well, if you have extra knowledge about what you expect to see, maybe you can cut at a certain number and then just fold all the largest values from K minus 1 to infinity so that you actually have-- you have everything into one large bin. OK? That's the entire tail.

And that's the way people do it in insurance companies, for example. They assume that the number of accidents you're going to have is a Poisson distribution. They have to fit it to you. They have to know-- or at least to your pool of insurance of injured people.

So they just slice you into what your character-- relevant characteristics are, and then they want to estimate what the Poisson distribution is. And basically, they can do a chi-square test to check if it's indeed a Poisson distribution.

All right. So that will be it for today. And so I'll be-- I'll have your homework--