PHILIPPE RIGOLLET: We're talking about tests. And to be fair, we spend most of our time talking about new jargon that we're using. The main goal is to take a binary decision, yes and no. So just so that we're clear and we make sure that we all speak the same language, let me just remind you what the key words are for tests.

So the first thing is that we split theta in theta 0 and theta 1. Both are included in theta, and they are disjoint. So I have my set of possible parameters. And then I have theta 0 is here, theta 1 is here. And there might be something that I leave out.

And so what we're doing is, we have two hypotheses. So here's our hypothesis testing problem. And it's h0 theta belongs to theta 0 versus h1 theta belongs to theta 1. This guy was called the null, and this guy was called the alternative.

And why we give them special names is because we saw that they have an asymmetric role. The null represents the status quo, and data is here to bring evidence against this guy. And we can really never conclude that h0 is true because all we could conclude is that h1 is not true, or may not be true. So that was the first thing.

The second thing was the hypothesis. The third thing is, what is a test? Well, psi, it's a statistic, and it takes the data, and it maps it into 0 or 1.

And I didn't really mention it, but there's some things such as called randomized tests, which is, well, if I cannot really make a decision, I might as well flip a coin. That tends to be biased, but that's really-- I mean, think about it in practice. You probably don't want to make decisions based on flipping a coin.

And so what people typically do-- this is happening, typically, at one specific value. So rather than flipping a coin for this very specific value, what people typically do is they say, OK, I'm going to side with h0 because that's the most conservative choice I can make. So in a way, they think of flipping this coin, but always falling on heads, say.

So associated to this test was something called, well, the rejection region r psi, which is just the set of data x1 xn such that psi of x1 xn is equal to 1. So that means we rejected h0 when the test is 1. And those are the set of data points that actually are going to lead me to reject

the test.

And then the things that we're actually, slightly, a little more important and really peculiar to test, specific to tests, were the type I and type II error. So the type I error arises when-- so type I error is when you reject, whereas h0 is correct. And the type II error is the opposite, so it's failed to reject, whereas h1 is correct-- h is correct, yeah. So those are the two types of errors you can make.

And we quantified their probability of type I error. So alpha psi is the probability-- so that's the probability of type I error. So psi is just the probability for theta that psi rejects and that's defined for theta and theta 0, so for different values of theta 0. So h0 being correct means there exists a theta in theta 0 for which that actually is the right distribution. So for different values of theta, I might make different errors.

So if you think, for example, about the coin example, I'm testing if the coin is biased towards heads or biased towards tails. So if I'm testing whether p is larger than 1/2 or less than 1/2, then when the true p-- let's say our h0 is larger than 1/2. When p is equal to 1, it's actually very difficult for me to make a mistake, because I only see heads.

So when p is getting closer to 1/2, I'm going to start making more and more probability of error. And so the type II error-- so that's the probability of type II-- is denoted by beta psi. And it's the function, well, that does the opposite and, this time, is defined for theta in theta 1.

And finally, we define something called the power, pi of psi. And this time, this is actually a number. And so this number is equal to the maximum over theta n theta 0. I mean, that could be a supremum, but think of it as being a maximum of p theta of psi is equal-- sorry, that's n0, right? Give me one sec. No, sorry, that's the min.

So this is not making a mistake. Theta 0 is in theta 2 So if theta is in theta 1 and I conclude 1, so this is a good thing. I want this number to be large. And I'm looking at the worst house-- what is the smallest value this number can be? So what I want to show you a little bit is a picture.

So now I'm going to take theta, and think of it as being a p. So I'm going to take p for some variable in the experiment. So p can range between 0 and 1, that's for sure.

And what I'm going to try to test is whether p is less than 1/2 or larger than 1/2. So this is going to be, let's say, theta 0. And this guy here is theta 1. Just trying to give you a picture of what

those guys are. So I have my y-axis, and now I'm going to start drawing number. All these things-- this function, this function, and this number-- are all numbers between 0 and 1.

So now I'm claiming that-- so when I move from left to right, what is my probability of rejecting going to do? So what I'm going to plot is the probability under theta. The first thing I want to plot is the probability under theta that psi is equal to 1. And let's say psi-- think of psi as being just this indicator that square root on n xn bar minus p over square root xn bar 1 minus xn bar is larger than some constant c for a probability chosen c.

So what we choose is that c is in such a way that, at 1/2, when we're testing for 1/2, what we wanted was this number to be equal to alpha, basically. So we fix this alpha number so that this guy-- so if I want alpha of psi of theta less than alpha given in advanced-- so think of it as being equal to, say, 5%. So I'm fixing this number, and I want this to be controlled for all theta and theta 0.

So if you're going to give me this budget, well, I'm actually going to make it equal where I can. If you're telling me you can make it equal to alpha, we know that if I increase my type I error, I'm going to decrease my type II error. If I start putting everyone in jail or if I start letting everyone go free, that's what we were discussing last time.

So since we have this trade-off and you're giving me a budget for one guy, I'm just going to max it out. And where am I going to max it out? Exactly at 1/2 at the boundary. So this is going to be 5%.

So what I know is that since alpha of theta is less than alpha for all theta in theta 0-- sorry, that's for theta 0, that's where alpha is defined. So for theta and theta 0, I knew that my function is going to look like this. It's going to be somewhere in this rectangle. Everybody agrees?

So this function for this guy is going to look like this. When I'm at 0, when p is equal to 0, which means I only observe 0's, then I know that p is going to be 0, and I will certainly not conclude that p is equal to 1. This test will never conclude that p is equal to 1-- that p is larger than 1/2, just because xn bar is going to be equal to 0.

Well, this is actually not well-defined, so maybe I need to do something-- put it equal to 0 if xn bar is equal to 0. So I guess, basically, I get something which is negative, and so it's never going to be larger than what I want. And so here, I'm actually starting at 0.

So now, this is this function here that increases-- I mean, it should increase smoothly. This function here is alpha psi of theta-- or alpha psi of p, let's say, because we're talking about p. Then it reaches alpha here.

Now, when I go on the other side, I'm actually looking at beta. When I'm on theta 1, the function that matters is the probability of type II error, which is beta of psi. And this beta of psi is actually going to increase.

So beta of psi is what? Well, beta of psi should also-- sorry, that's the probability of being equal to alpha. So what I'm going to do is I'm going to look at the probability of rejecting. So let me draw this functional all the way. It's going to look like this.

Now here, if I look at this function here or here, this is the probability under theta that psi is equal to 1. And we just said that, in this region, this function is called alpha of psi. In that region, it's not called alpha of psi. It's not called anything. It's just the probability of rejection. So it's not any error, it's actually what you should be doing.

What we're looking at in this region is 1 minus this guy. We're looking at the probability of not rejecting. So I need to actually, basically, look at the 1 minus this thing, which here is going to be 95%. So I'm going to do 95%.

And this is my probability. Ability And I'm just basically drawing the symmetric of this guy. So this here is the probability under theta that psi is equal to 0, which is 1 minus p theta that psi is equal to 1. So it's just 1 minus the wide curve. And it's actually, by definition, equal to beta of psi of theta.

Now, where do I read pi psi? What is pi psi on this picture? Is pi psi a number or a function?

AUDIENCE: Number.

PHILIPPE RIGOLLET: It's a number, right? It's the minimum of a function. What is this function? It's the probability under theta that theta is equal to 1. I drew this entire function for between theta 0 and theta 1. I drew-- this is this entire white curve. This is this probability.

Now I'm saying, look at the smallest value this probability can take on the set theta 1. What is this? This guy. This is where my pi-- this thing here is pi psi, and so it's equal to 5%.

So that's for this particular test, because this test has a continuous curve for this psi. And so if I

want to make sure that I'm at 5% when I come to the right of the theta 0, if it touches theta 1, then I'd better have 5% on the other side if the function is continuous. So basically, if this function is increasing, which will be the case for most tests, and continuous, then what's going to happen is that the level of the test, which is alpha, is actually going to be equal to the power of the test.

Now, there's something I didn't mention, and I'm just mentioning it passing by. Here, I define the power itself. This function, this entire white curve here, is actually called the power function-- this thing. That's the entire white curve. And what you could have is tests that have the entire curve which is dominated by another test.

So here, if I look at this test-- and let's assume I can build another test that has this curve. Let's say it's the same here, but then here, it looks like this. What is the power of this test?

AUDIENCE: It's the same.

PHILIPPE RIGOLLET: It's the same. It's 5%, because this point touches here exactly at the same point. However, for any other value than the worst possible, this guy is doing better than this guy. Can you see that? Having a curve higher on the right-hand side is a good thing because it means that you tend to reject more when you're actually in h1. So this guy is definitely better than this guy.

And so what we say, in this case, is that the test with the dashed line is uniformly more powerful than the other tests. But we're not going to go into those details because, basically, all the tests that we will describe are already the most powerful ones. In particular, this guy is-- there's no such thing. All the other guys you can come up with are going to actually be below. So we saw a couple tests, then we saw how to pick this threshold, and we defined those two things.

AUDIENCE: Question.

PHILIPPE RIGOLLET: Yes?

AUDIENCE: But in that case, the dashed line, if it were also higher in the region of theta 0, do you still consider it better?

PHILIPPE RIGOLLET: Yeah.

AUDIENCE: OK.

PHILIPPE RIGOLLET: Because you're given this budget of 5%. So in this paradigm where you're given the-- actually, if the dashed line was this dashed line, I would still be happy. I mean, I don't care what this thing does here, as long as it's below 5%.

But here, I'm going to try to discover. Think about, again, the drug discovery example. You're trying to find-- let's say you're a scientist and you're trying to prove that your drug works.

What do you want to see? Well, FDA puts on you this constraint that your probability of type I error should never exceed 5%. You're going to work under this assumption.

But what you're going to do is, you're going to try to find a test that will make you find something as often as possible. And so you're going to max this constraint of 5%. And then you're going to try to make this curve, that means-- this is, basically, this number here, for any point here, is the probability that you publish your paper. That's the probability that you can release to market your drug. That's the probability that it works.

And so you want this curve to be as high as possible. You want to make sure that if there's evidence in the data that h1 is the truth, you want to squeeze as much of this evidence as possible. And the test that has the highest possible curve is the most powerful one.

Now, you have to also understand that having two curves that are on top of each other completely, everywhere, is a rare phenomenon. It's not always the case that there is a test that's uniformly more powerful than any other test. It might be that you have some trade-off, that it might be better here, but then you're losing power here. Maybe it's-- I mean, things like this. Well, actually, maybe it should not go down.

But let's say it goes like this, and then, maybe, this guy goes like this. Then you have to, basically, make an educated guess whether you think that the theta you're going to find is here or is here, and then you pick your test. Any other question? Yes?

AUDIENCE: Can you explain the green curve again? That's just the type II error?

PHILIPPE RIGOLLET: So the green curve is-- exactly. So that's beta psi of theta. So it's really the type II error. And it's defined only here. So here, it's not a definition, it's really I'm just mapping it to this point. So it's defined only here, and it's the probability of type II error.

So here, it's pretty large. I'm making it, basically, as large as I could because I'm at the boundary, and that means, at the boundary, since the status quo is h0, I'm always going to go for h0 if I don't have any evidence, which means that what's going to pay is the type II error that's going to basically pay this. Any other question?

So let's move on. So did we do this? No, I think we stopped here, right? I didn't cover that part.

So as I said, in this paradigm, we're going to actually fix this guy to be something. And this thing is actually called the level of the test. I'm sorry, this is, again, more words. Actually, the good news is that we split it into two lectures.

So we have, what is a test? What is a hypothesis? What is the null? What is the alternative? What is the type I error? What is the type II error?

And now, I'm telling you there's another thing. So we define the power, which was some sort of a lower bound on the-- or it's 1 minus the upper bound on the type II error, basically. And so it's alternative-- so the power is the smallest probability of rejecting when you're in the null. And it's alternative when you're in theta 1, so that's my power. I looked here, and I looked at the smallest value.

And I can look at this side and say, well, what is the largest probability that I make a type I error? Again, this largest probability is the level of the test. So this is alpha equal, by definition, to the maximum for theta in theta 0 of alpha psi of theta.

So here, I just put the level itself. As you can see, here, it essentially says that if I'm of level of 5%, I'm also of level 10%, I'm also of level 15%. So here, it's really an upper bound. Whatever you guys want to take, this is what it is.

But as we said, if this number is 4.5%, you're losing in your type II error. So if you're allowed to have-- if this maximum here is 4.5% and FDA told you you can go to 5%, you're losing in your type II error. So you actually want to make sure that this is the 5% that's given to you. So the way it works is that you give me the alpha, then I'm going to go back, pick c that depends on alpha here, so that this thing is actually equal to 5%.

And so of course, in many instances, we do not know the probability. We do not know how to compute the probability of type I error. This is a maximum value for the probability of type I error. We don't know how to compute it.

I mean, it might be a very complicated random variable. Maybe it's a weird binomial. We could compute it, but it would be painful. But we know how to compute is its asymptotic value.

Just because of the central limit theorem, convergence and distribution tells me that the probability of type I error is basically going towards the probability that some Gaussian is in some region. And so we're going to compute, not the level itself, but the asymptotic level. And that's basically the limit as n goes to infinity of alpha psi of theta. And then I'm going to make the max here.

So how am I going to compute this? Well, if I take a test that has rejection region of the form tn-- because it depends on the data, that's tn of x1 xn-- my observation's larger than some number c. Of course, I can almost always write tests like that, except that sometimes, it's going to be an absolute value, which essentially means I'm going away from some value. Maybe, actually, I'm less than something, but I can always put a negative sign in front of everything.

So this is not without much of generality. So this includes something that looks like-- something is larger than the constants, so that means-- which is equivalent to-- well, let me write that as tq, because then that means that-- so that's tn. But this actually encompasses the fact that qn is larger than c or qn is less than c and n minus c. So that includes this guy.

That also includes qn less than c, because this is equivalent to qn is larger than minus c. And minus qn is-- and so that's going to be my tn. So I can actually encode several type of things-- rejection regions.

So here, in this case, I have a rejection region that looks like this, or a rejection region that looks like this, or a rejection region that looks like this. And here, I don't really represent it for the whole data, but maybe for the average, for example, or the normalized average. So if I write this, then-- yeah.

And in this case, this tn that shows up is called test statistic. I mean, this is not set in stone. Here, for example, q could be the test statistic. It doesn't have to be minus q itself that's the test statistic.

So what is the test statistic? Well, it's what you're going to build from your data and then compare to some fixed value. So in the example we had here, what is our test statistic? Well, it's this guy. This was our test statistic.

And is this thing a statistic? What are the criteria for a statistic? What is the statistic? I know you know the answer.

AUDIENCE: Measurable function.

PHILIPPE RIGOLLET: Yeah, it's a measurable function of the data that does not depend on the parameter. Is this guy a statistic?

AUDIENCE: It's not.

PHILIPPE RIGOLLET: Let's think again. When I implemented the test, what did I do? I was able to compute my test. My test did not depend on some unknown parameter.

How did we do it? We just plugged in 0.5 here, remember? That was the value for which we computed it, because under h0, that was the value we're seeing.

And if theta 0 is actually an entire set, I'm just going to take the value that's the closest to h1. We'll see that in a second. I mean, I did not guarantee that to you.

But just taking the worst type I error and bounded by alpha is equivalent to taking p and taking the value of p that's the closest to theta 1, which is completely intuitive. The worst type I error is going to be attained for the p that's the closest to the alternative. So even if the null is actually just an entire set, it's as if it was just the point that's the closest to the alternative.

So now we can compute this, because there's no unknown parameters that shows up. We replace p by 0.5. And so that was our test statistic. So when you're building a test, you want to first build a test statistic, and then see what threshold you should be getting.

So now, let's go back to our example where we want to have-- we have x1 xn, their IID [INAUDIBLE] p. And I want to test if p is 1/2 versus p not equal to 1/2, which, as I said, is what you want to do if you want to test if a coin is fair.

And so here, I'm going to build a test statistic. And we concluded last time that-- what do we want for this statistic? We want it to have a distribution which, under the null, does not depend on the parameters, a distribution that I can actually compute quintiles of.

So what we did is, we said, well, if I look at-- the central limit theorem tells me that square root of n xn bar minus p divided by-- so if I do central limit theorem plus Slutsky, for example, I'm going to have square root. And we've had this discussion whether we want to use Slutsky or

not here. But let's assume we're taking Slutsky wherever we can. So this thing tells me that, by the central limit theorem, as n goes to infinity, this thing converges in distribution to some n01.

Now, as we said, this guy is not something we know. But under the null, we actually know it. And we can actually replace it by 1/2. So this thing holds under h0. When I write under h0, it means when this is the truth.

So now I have something that converges to something that has no dependence on anything I don't know. And in particular, if you have any statistics textbook, which you don't because I didn't require one-- and you should be thankful, because these things cost $350. Actually, if you look at the back, you actually have a table for a standard Gaussian.

I could have anything else here. I could have an exponential distribution. I could have a-- I don't know-- well, we'll see the chi squared distribution in a minute. Any distribution from which you can actually see a table that somebody actually computed this thing for which you can actually draw the pdf and start computing whatever probability you want on them, then this is what you want to see at the right-hand side.

This is any distribution. It's called pivotal. I think we've mentioned that before. Pivotal means it does not depend on anything that you don't know.

And maybe it's easy to compute those things. Probably, typically, you need a computer to simulate them for you because computing probabilities for Gaussians is not an easy thing. We don't know how to solve those integrals exactly, we have to do it numerically.

So now I want to do this test. My test statistic will be declared to be what? Well, I'm going to reject if what is larger than some number? The absolute value of this guy.

So my test statistic is going to be square root of n minus 0.5 divided by square root of xn bar 1 minus xn bar. That's my test statistic, absolute value of this guy, because I want to reject either when this guy is too large or when this guy is too small. I don't know ahead whether I'm going to see p larger than 1/2 or less than 1/2.

So now I need to compute c such that the probability that tn is larger than c. So that's the probability under p, which is unknown. I want this probability to be less than some level alpha, asymptotically.

So I want the limit of this guy to be less than alpha, and that's the level of my test. So that's the

given level. So I want this thing to happen. Now, what I know is that this limit-- actually, I should say given asymptotic level.

So what is this thing? Well, OK, that's the probability that something that looks like under p. So under p, this guy-- so what I know is that tn is square root of n minus xn bar minus 0.5 divided by square root of xn bar 1 minus xn bar exceeds. Is this true that as n to infinity, this probability is the same as the probability that the absolute value of a Gaussian exceeds c of a standard Gaussian? Is this true?

AUDIENCE: The absolute value of the standard Gaussian.

PHILIPPE RIGOLLET: Yeah, the absolute. So you're saying that this, as n becomes large enough, this should be the probability that some absolute value of n01 exceeds c, right?

AUDIENCE: Yes.

PHILIPPE RIGOLLET: So I claim that this is not correct. Somebody tell me why.

AUDIENCE: Even in the limit it's not correct?

PHILIPPE RIGOLLET: Even in the limit, it's not correct.

AUDIENCE: OK.

PHILIPPE RIGOLLET: So what do you see?

AUDIENCE: It's because, at the beginning, we picked the worst possible true parameter, 0.5. So we don't actually know that this 0.5 is the mean.

PHILIPPE RIGOLLET: Exactly. So we pick this 0.5 here, but this is for any p. But what is the only p I can get? So what I want is that this is true for all p in theta 0. But the only p that's in theta 0 is actually p is equal to 0.5.

So yes, what you said was true, but it required to specify p to be equal to 0.5. So this, in general, is not true. But it happens to be true if p belongs to theta 0, which is strictly equivalent to p is equal to 0.5, because theta 0 is really just this one point, 0.5.

So now, this becomes true. And so what I need to do is to find c such that this guy is equal to what? I mean, let's just follow. So I want this to be less than alpha. But then we said that this was equal to this, which is equal to this.

So all I want is that this guy is less than alpha. But we said we might as well just make it equal to alpha if you allow me to make it as big as I want, as long as it's less than alpha.

AUDIENCE: So this is a true statement.

PHILIPPE RIGOLLET: So this is a true statement. But it's under this condition.

AUDIENCE: Exactly.

PHILIPPE RIGOLLET: So I'm going to set it equal to alpha, and then I'm going to try to solve for c. So what I'm looking for is a c such that if I draw a standard Gaussian-- so that's pdf of some n01-- I want the probability that the absolute value of my Gaussian exceeding this guy-- so that means being either here or here. So that's minus c and c.

I want the sum of those two things to be equal to alpha. So I want the sum of these areas to equal alpha. So by symmetry, each of them should be equal to alpha over 2.

And so what I'm looking for is c such that the probability that my n01 exceeds c, which is just this area to the right, now, equals alpha, which is equivalent to taking c, which is q equals alpha over 2, and that's q alpha over 2 by definition of q alpha over 2. That's just what q alpha over 2 is. And that's what the tables at the back of the book give you.

Who has already seen a table for Gaussian probabilities? What it does, it's just a table. I mean, it's pretty ancient. I mean, of course, you can actually ask Google to do it for you now. I mean, it's basically standard issue. But back in the day, they actually had to look at tables.

And since the values alphas were pretty standard, the values alpha that people were requesting were typically 1%, 5%, 10%, all you could do is to compute these different values for different values of alpha. That was it. So there's really not much to give you.

So for the Gaussian, I can tell you that alpha is equal to-- if alpha is equal to 5%, then q alpha over 2, q 2.5% is equal to 1.96, for example. So those are just fixed numbers that are functions of the Gaussian. So everybody agrees?

We've done that before for our confidence intervals. And so now we know that if I actually plug in this guy to be q alpha over 2, then this limit is actually equal to alpha. And so now I've actually constrained this.

So q alpha over 2 here for alpha equals 5%, as I said, is 1.96. So in the example 1, the number that we found was 3.54, I think, or something like that, 3.55 for t. So if we scroll back very quickly, 3.45-- that was example 1. Example two-- negative 0.77.

So if I look at tn in example 1, tn was just the absolute value of 3.45, which-- don't pull out your calculators-- is equal to 3.45. Example 2, absolute value of negative 0.77 was equal to 0.77. And so all I need to check is, is this number larger or smaller than 1.96? That's what my test ends up being.

So in example 1, 3.45 being larger than 1.96, that means that I reject. Fairness of my coins, in example 2, 0.77 being smaller than 1.96-- what do I do? I fail to reject.

So here is a question. In example 1, for what level alpha would psi alpha-- OK, so here, what's going to happen if I start decreasing my level? When I decrease my level, I'm actually making this area smaller and smaller, which means that I push this c to the right.

So now I'm asking, what is the smallest c I should pick so that now, I actually do not reject h0? What is the smallest c I should be taking here? What is the smallest c?

So c here, in the example I gave you for 5%, was 1.96. What is the smallest c I should be taking so that now, this inequality is reversed? 3.45. I ask only trivial questions, don't be worried. So 3.45 is the smallest c that I'm actually willing to tolerate.

So let's say this was my 5%. If this was 2.5-- if here, let's say, in this picture, alpha is 5%, that means maybe I need to push here. And this number should be what? So this is going to be 1.96. And this number here is going to be 3.45, clearly to scale.

And so now, what I want to ask you is, well, there's two ways I can understand this number 3.45. It is the number 3.45, but I can also try to understand what is the area to the right of this guy. And if I understand what the area to the right of this guy is, this is actually some alpha prime over 2.

And that means that if I actually fix this level alpha prime, that would be exactly the tipping point at which I would go from accepting to rejecting. So I knew, in terms of absolute

thresholds, 3.45 is the trivial answer to the question. That's the tipping point, because I'm comparing a number to 3.45.

But now, if I try to map this back and understand what level would have been giving me this particular tipping point, that's a number between 0 and 1. The smaller the number, the larger this number here, which means that the more evidence I have in my data against h0. And so this number is actually something called the p-value.

And so saying, for example 2, there's the tipping point alpha at which I go from failing to reject to rejecting. And that's exactly the number, the area under the curve, such that here, I see 0.77. And this is this alpha prime prime over 2. Alpha prime prime is clearly larger than 5%.

So what's the advantage of thinking and mapping back these numbers? Well, now, I'm actually going to spit out some number which is between 0 and 1. And that should be the only scale you should have in mind.

Remember, we discussed that last time. I was like, well, if I actually spit out a number which is 3.45, maybe you can try to think, is 3.45 a large number for a Gaussian? That's a number. But if I had another random variable that was not Gaussian, maybe it was a double exponential, you would have to have another scale in your mind.

Is 3.45 so large that it's unlikely for it to come from a double exponential. If I had a gamma distribution-- I can think of any distribution, and then that means, for each distribution, you would have to have scale in mind. So of course, you can have the Gaussian scale in mind. I mean, I have the Gaussian scale in mind.

But then, if I map it back into this number between 0 and 1, all the distributions play the same role. So whether I'm talking about if my limiting distribution is normal or exponential or gamma, or whatever you want, for all these guys, I'm just going to map it into one number between 0 and 1. Small number means lots of evidence against h1. Large number means lots of evidence against h0. Small number means very few evidence against h9.

And this is the only number you need to keep in mind. And the question is, am I willing to tolerate this number between 5%, 6%, or maybe 10%, 12%? And this is the only scale you have to have in mind. And this scale is the scale of p-values.

So the p-value is the tipping point in terms of alpha. In words, I can make it formal, because

tipping point, as far as I know, is not a mathematical term. So a p-value of a test is the smallest, potentially asymptotic level if I talk about an asymptotic p-value-- and that's what we do when we talk about central theorem-- at which the test rejects h0.

If I were to go any smaller-- sorry, it's the smallest level-- yeah, if I were to go any smaller, I would fail to reject. The smaller the level, the less likely it is for me to reject. And if I were to go any smaller, I would start failing to reject. And so it is a random number. It depends on what I actually observe.

So here, of course, I instantiated those two numbers, 3.45 and 0.77, as realizations of random variables. But if you think of those as being the random numbers before I see my data, this was a random number, and therefore, the area under the curve to the right of it is also a random area. If this thing fluctuates, then the area under the curve fluctuates.

And that's what the p-value is. That's what-- what is his name? I forget. John Oliver talks about when he talks about p-hacking. And so we talked about this in the first lecture.

So p-hacking is, how do I do-- oh, if I'm a scientist, do I want to see a small p-value or a large p-value?

AUDIENCE: Small.

PHILIPPE RIGOLLET: Small, right? Scientists want to see small p-values because small p-values equals rejecting, which equals discovery, which equals publications, which equals promotion. So that's what people want to see. So people are tempted to see small p-values. And what's called p-hacking is, well, find a way to cheat. Maybe look at your data, formulate your hypothesis in such a way that you will actually have a smaller p-value than you should have.

So here, for example, there's one thing I did not insist on because, again, this is not a particular course on statistical thinking, but one thing that we implicitly did was set those theta 0 and theta 1 ahead of time. I fixed them, and I'm trying to test this.

This is to be contrasted with the following approach. I draw my data. So I draw-- I run this experiment, which is probably going to get me a publication in nature. I'm trying to test if a coin is fair.

And I draw my data, and I see that there's 13 out of 30 of my observations that are heads. That means that, from this data, it looks like p is less than 1/2. So if I look at this data and then

decide that my alternative is not p not equal to 1/2, but rather p less than 1/2, that's p-hacking.

I'm actually making my p-value strictly smaller by first looking at the data, and then deciding what my alternative is going to be. And that's cheating, because all the things we did, we're assuming that this 0.5, or the alternative, was actually a fixed-- everything was deterministic. The only randomness came from the data.

But if I start looking at the data and designing my experiment or my alternatives and null hypothesis based on the data, it's as if I started putting randomness all over the place. And then I cannot control it because I don't know how it just intermingles with each other. So that was for the John Oliver moment.

So the p-value is nice. So maybe I mentioned that, before, my wife works in market research. And maybe every two years, she seems to run into a statistician in the hallway, and she comes home and says, what is a p-value again?

And for her, a p-value is just the number in an Excel spreadsheet. And actually, small equals good and large equals bad. And that's all she needs to know at this point. Actually, they do the job for her-- small is green, large is red. And so for her, a p-value is just green or red.

But so what she's really implicitly doing with this color code is just applying the golden rule. What the statisticians do for her in the Excel spreadsheet is that they take the numbers for the p-values that are less than some fixed level. So depending on the field in which she works-- so she works for pharmaceutical companies-- so the p-values are typically compared-- the tests are usually performed at level 1%, rather than 5%.

So 5% is maybe your gold standard if you're doing sociology or trying to-- I don't know-- release a new blueberry flavor for your toothpaste. Something that's not going to change the life of people, maybe you're going to run at 5%. It's OK to make a mistake. See, people are just going to feel gross, but that's about it, whereas here, if you have this p-value which is less than 1%, it might be more important for some drug discovery, for example.

And so let's say you run at 1%. And so what they do in this Excel spreadsheet is that all the numbers that are below 1% show up in green and all the numbers that are above 1% show up in red. And that's it. That's just applying the golden rule. If the number is green, reject. If the number is red, fail to reject. Yeah?

AUDIENCE: So going back to example 2 where the prior example where you want to cheat by looking after

beta and then formulating, say, theta 1 to be p less than 1/2.

PHILIPPE RIGOLLET: Yeah.

AUDIENCE: So how would you achieve your goal by changing the theta--

PHILIPPE RIGOLLET: By achieving my goal, you mean letting ethics aside, right?

AUDIENCE: Yeah, yeah.

PHILIPPE RIGOLLET: Ah, you want to be published.

AUDIENCE: Yeah.

PHILIPPE RIGOLLET: [LAUGHS] So let me teach you how, then. So well, here, what do you do? You want to-- at the end of the day, a test is only telling you whether you found evidence in your data that h1 was more likely than h0, basically.

How do you make h1 more likely? Well, you just basically target h1 to be what it is-- what the data is going to make it more likely to be. So if, for example, I say h1 can be on both sides, then my data is going to have to take into account fluctuations on both sides, and I'm going to lose a factor or two somewhere because things are not symmetric.

Here is the ultimate way of making this work. I'm going back to my example of flipping coins. And now, so here, what I did is, I said, oh, this number 0.43 is actually smaller than 0.5, so I'm just going to test whether I'm 0.5 or I'm less than 0.5.

But here is something that I can promise you I did not make the computation will reject. So here, this one actually-- yeah, this one fails to reject. So here is one that will certainly reject. h0 is 0.5, p is 0.5, h1p is 0.43.

Now, you can try, but I can promise you that your data will tell you that h1 is the right one. I mean, you can check very quickly that this is really extremely likely to happen. Actually, what am I-- no, actually, that's not true, because here, the test that I derive that's based on this kind of stuff, here at some point, somewhere under some layers, I assume that all our tests are going to have this form.

But here, this is only when you're trying to test one region versus another region next to it, or one point versus a region around it, or something like this, whereas for this guy, there's another test that could come up with, which is, what is the probability that I get 0.43, and what is the probability that I get 0.5? Now, what I'm going to do is, I'm going to just conclude it's whichever has the largest probability. Then maybe I'm going to have to make some adjustments so that the level is actually 5%.

But I can make this happen. I can make the level be 5% and always conclude this guy, but I would have to use a different test. Now, the test that I described, again, those tn larger than c are built in to be tests that are resilient to these kind of manipulations because they're oblivious towards what the alternative looks like. I mean, they're just saying it's either to the left or to the right, but whether it's a point or an entire half-line doesn't matter.

So if you try to look at your data and just put the data itself into your hypothesis testing problem, then you're failing the statistical principle. And that's what people are doing. I mean, how can I check? I mean, of course, here, it's going to be pretty blatant if you publish a paper that looks like this.

But there's ways to do it differently. For example, one way to do it is to just do mult-- so typically, what people do is they do multiple hypothesis testing. They're doing 100 tests at a time.

Then you have random fluctuations every time. And so they just pick the one that has the random fluctuations that go their way. I mean, sometimes it's going in your way, and sometimes it's going the opposite way, so you just pick the one that works for you. We'll talk about multiple hypothesis testing soon if you want to increase your publication count. There's actually papers-- I think it was a big news that some papers, I think, in psychology or psychometrics papers that actually refused to publish p-values now.

Where were we? Here's the golden rule. So one thing that I like to show is this thing, just so you know how you apply the golden rule and how you apply the standard tests.

So the standard paradigm is the following. You have a black box, which is your test. For my wife, this is the 4th floor of the building. That's where the statisticians sit.

What she sends there is data-- let's say x1 xn. And she says, well, this one is about

toothpaste, so here's a level-- let's say 5%. What the 4th floor brings back is that answer-- yes, no, green, red, just an answer. So that's the standard testing. You just feed it the data and the level at which you want to perform the test, maybe asymptotic, and it spits out a yes, no answer.

What p-value does, you just feed it the data itself. And what it spits out is the p-value. And now it's just up to you. I mean, hopefully your brain has the computational power of deciding whether a number is larger or smaller than 5% without having to call a statistician for this. And that's what it does. So now we're on 1 scale.

Now, I see some of you nodding when I talk about p-hacking, so that means you've seen p-values. If you've seen more than 100 p-values in your life, you have an entire scale. A good p-value is less than 10 to the minus 4. That's the ultimate sweet spot.

Actually, statistical software spits out an output which says less than 10 to the minus 4. But then maybe you want a p-val-- if you tell me my p-value was 4.65, then I will say, you've been doing some p-hacking until you found a number that was below 5%. That's typically what people will do.

But if you tell me-- if you're doing the test, if you're saying, I published my result, my test at 5% said yes, that means that maybe you're p-value was 4.99, or you're p-value was 10 to the minus 4, I will never know. I will never know how much evidence you had against the null. But if you tell me what the p-value is, I can make my own decision.

I don't have to tell me whether it's a yes, no. You tell me it's 4.99, I'm going to say, well, maybe yes, but I'm going to take it with a grain of salt. And so that's why p-values are good numbers to have in mind.

Now, I should, as if it was like an old trick that you start mastering when you're 45 years old. No, it's just, how small is the number between 0 and 1? That's really what you need to know. Maybe on the log scale-- if it's 10 to the minus 1, 10 to the minus 2, 10 to the minus 3, et cetera-- that's probably the extent of the mastery here.

So this traditional standard paradigm that I showed is actually commonly referred to as the Neyman-Pearson paradigm. So here, it says name Neyman-Pearson's theory, so there's an entire theory that comes with it. But it's really a paradigm.

It's a way of thinking about hypothesis testing that says, well, if I'm not going to be able to

optimize both my type I and type II error, I'm actually going to lock in my type I error below some level and just minimize the type II error under this constraint. That's what the Neyman-Pearson paradigm is. And it sort of makes sense for hypothesis testing problems.

Now, if you were doing some other applications with multi-objective optimization, you would maybe come up with something different. For example, machine learning is not performing typically under Neyman-Pearson paradigm. So if you do spam filtering, you could say, well, I want to constrain the probability as much as I can of taking somebody's important emails and throwing them out as spam, and under this constraint, not send too much spam to that person. That sort of makes sense for spams.

Now, if you're labeling cats versus dogs, that's probably not like you want to make sure that no more than 5% of the dogs are labeled cat because, I mean, it doesn't matter. So what you typically do is, you just sum up the two types of errors you can make, and you minimize the sum without putting any more weight on one or the other. So here's an example where doing a binary decision, one or two of the errors you can make, you don't have to actually be like that.

So this example here, I did not. The trivial test psi is equal to 0, what was it in the US trial court example? What is psi equals 0? That was concluding always to the null. What was the null?

AUDIENCE: Innocent.

PHILIPPE RIGOLLET: Innocent, right? That's the status quo. So that means that this guy never rejects h0. Everybody's going away free. So you're sure you're not actually going against the constitution because alpha is 0%, which is certainly less than 5%.

But the power, the fact that a lot of criminals go back outside in the free world is actually formulated in terms of low power, which, in this case, is actually 0. Again, the power is the number between 0 and 1. Close to 0, good. Close to 1, bad.

Now, what is the definition of the p-value? That's going to be something-- it's a mouthful. The definition of the p-value is a mouthful. It's the tipping point. It is the smallest level at which blah, blah, blah, blah, blah. It's complicated to remember it.

Now, I think that my 6th explanation, my wife, after saying, oh, so it's the probability of making an error, I said, yeah, that's the probability of making an error because, of course, she can think probability of making an error small, good, large, bad. So that's actually a good way to

remember. I'm pretty sure that at least 50% of people who are using p-values out there think that the p-value is the probability of making an error.

Now, for all matters of purposes, if your goal is to just threshold the p-value, this is OK to have this in y. But when comes, at least until December 22, I would recommend trying to actually memorize the right definition for the p-value. So the idea, again, is fix the level and try to optimize the power.

So we're going to try to compute some p-values from now on. How do you compute the p-value? Well, you can actually see it from this picture over there.

One thing I didn't show on this picture-- so here, it was my q alpha over 2 that had alpha here, alpha over 2 here. That was my q alpha over 2. And I said, if tn is to the right of this guy, I'm going to reject. If tn is to the left of this guy, I'm going to fail to reject.

Pictorially, you can actually represent the p-value. It's when I replace this guy by tn itself. Sorry, that's p-value over 2. No, actually, that's p-value. So let me just keep it like that and put the absolute value here.

So if you replace the role of q alpha over 2, by your test statistic, the area under the curve is actually the p-value itself up to a scale because of the symmetric thing. So there's a good way to see, pictorially, what the p-value is. It's just the probability that some Gaussians-- it's just the probability that some absolute value of n01 exceeds tn. That's what the p-value is. Now, this guy has nothing to do with this guy, so this is really just 1 minus the Gaussian cdf of tn, and that's it. So that's how I would compute p-values.

Now, as I said, the p-value is a beauty because you don't have to understand the fact that your limiting distribution is a Gaussian. It's already factored in this construction. The fact that I'm actually looking at this cumulative distribution function of a standard Gaussian makes my p-value automatically adjust to what the limiting distribution is. And if this was the cumulative distribution function of a exponential, I would just have a different function here denoted by f, for example, and I would just compute a different value. But in the end, regardless of what the limiting value is, my p-value would still be a number between 0 and 1.

And so to illustrate that, let's look at other weird distributions that we could get in place of the standard Gaussian. And we're not going to see many, but we'll see one. And it's not called the chi squared distribution. It's actually called the Student's distribution, but it involves the chi

squared distribution as a building block. So I don't know if my phonetics are not really right there, so I try to say, well, it's chi squared. Maybe it's "kee" squared above, in Canada, who knows.

So for a positive integer, so there's only 1 parameter. So for the Gaussian, you have 2 parameters, which are mu and sigma squared. Those are real numbers. Sigma squared's positive.

Here, I have 1 integer parameter. Then the chi squared distribution with d degrees of freedom-- so the parameter is called a degree of freedom, just like mu is called the expected value and sigma squared is called the variance. Here, we call it degrees of freedom. You don't have to really understand why.

So that's the law that you would get-- that's the random variable you would get if you were to sum d squares of independent standard Gaussians. So I take the square of an independent random Gaussian. I take another one. I sum them, and that's a chi squared with 2 degrees of freedom. That's how you get it.

Now, I could define it using its probability density function. I mean, after all, this is the sum of positive random variables, so it is a positive random variable. It has a density on the positive real line.

And the pdf of chi squared with d degrees of freedom is what? Well, it's fd of x is-- what is it?-- x to the d/2 minus 1 e to the minus x/2. And then here, I have a gamma of d/2.

And the other one is, I think, 2 to the d/2 minus 1. No, 2 to the d/2. That's what it is. That's the density.

If you are very good at probability, you can make the change of variable and write your Jacobian and do all this stuff and actually check that this is true. I do not recommend doing that. So this is the density, but it's better understood like that. I think it was just something that you built from standard Gaussian.

So for example, an example of a chi squared with 2 degrees of freedom is actually the following thing. Let's assume I have a target like this. And I don't aim very well. And I'm trying to hit the center. And I'm not going to have, maybe, a deviation, which is standard Gaussian left, right and standard Gaussian north, south.

So I'm throwing, and then I'm here, and I'm claiming that this number here, by Pythagoras theorem, the square distance here is the sum of this square distance here, which is the square of a Gaussian by assumption. This is plus the square of this distance, which is the square of another independent Gaussian. I assume those are independent.

And so the square distance from this point to this point is the chi squared with 2 degrees of freedom. So this guy here is n01 squared. This is n01 squared. And so this guy here, this distance here, is chi squared with 2 degrees of freedom. I mean the square distance. I'm talking about square distances here.

So now you can see that, actually, Pythagoras is basically why chi squared [? arrives. ?] That's why it has its own name. I mean, I could define this random variable.

I mean, it's actually a gamma distribution. It's a special case of something called the gamma distribution. The fact that the special case has its own name is because there's many times what we're going to take sum of squares of independent Gaussians because Gaussians, the sum of squares is really the norm, the Euclidean norm squared, just by Pythagoras theorem. If I'm in higher dimension, I can start to sum more squared coordinates, and I'm going to measure the norm squared.

So if you want to draw this picture, it looks like this. Again, it's the sum of positive numbers, so it's going to be on 0 plus infinity. That's fd. And so f1 looks like this, f2 looks like this. So the tails become heavier and heavier as d increases.

And then at [INAUDIBLE] to 3, it starts to have a different shape. It starts from 0 and it looks like this. And then, as d increases, it's basically as if you were to push this thing to the right. It's just like, psh, so it's just falling like a big blob. Everybody sees what's going on? So there's just this fat thing that's just going there.

What is the expected value of a chi squared? So it's the expected value of the sum of Gaussian random variables, squared. I know I said that.

AUDIENCE: So it's the sum of their second moments, right?

PHILIPPE RIGOLLET: Which is? Those are n01.

AUDIENCE: It's like-- oh, I see, 1.

PHILIPPE RIGOLLET: Yeah.

AUDIENCE: So n times 1 or d times 1.

PHILIPPE RIGOLLET: Yeah, which is d. So one thing you can check quickly is that the expected value of a chi squared is d. And so you see, that's why the mass is shifting to the right as d increases. It's just going there. Actually, the variance is also increasing. The variance is 2d.

So this is one thing. And so why do we care about this? In basic statistics, it's not like we actually have statistics much about throwing darts at high-dimensional boards. So what's happening is that if I look at the sample variance, the average of the sum of squared centered by their mean, then I can actually expend this as the sum of the squares minus the average squared It's just the same trick that we have for the variance-- second moment minus first moment square.

And then I claim that Cochran's theorem-- and I will tell you in a second what Cochran's theorem tells me is that this sample variance is actually-- so if I had only this-- look at those guys. Those guys are Gaussian with mean mu and variance sigma squared. Think for 1 second mu being 0 and sigma squared being 1.

Now, this part would be a chi squared with n degrees of freedom divided by n. Now I get another thing here, which is the square of something that looks like a Gaussian as well. So it looks like I have something else here, which looks also like a chi squared.

Now, Cochran's theorem is essentially telling you that those things are independent, and so that in a way, you can think of those guys as being, here, n degrees of freedom minus 1 degree of freedom. Now, here, as I said, this does not mean 0 and variance 1. The fact that it's not mean 0 is not a problem because I can remove the mean here and remove the mean here. And so this thing has the same distribution, regardless of what the actual mean is. So without loss of generality, I can assume that mu is equal to 0.

Now, the variance, I'm going to have to pay, because if I multiply all these numbers by 10, then this sn is going to multiplied by 100. So this thing is going to scale with the variance. And not surprisingly, it's scaling like the square of the variance.

So if I look at sn, it's distributed as sigma squared times the chi squared with n minus 1

degrees of freedom divided by n. And we don't really write that, because a chi squared times sigma squared divided by n is not a distribution, so we put everything to the left, and we say that this is actually a chi squared with n minus 1 degrees of freedom. So here, I'm actually dropping a fact on you, but you can see the building block.

What is the thing that's fuzzy at this point, but the rest should be crystal clear to you? The thing that's fuzzy is that removing this squared guy here is actually removing 1 degree of freedom. That should be weird, but that's what Cochran's theorem tells. It's essentially stating something about orthogonality of subspaces with the span of the constant vector, something like that. So you don't have to think about it too much, but that's what it's telling me.

But the rest, if you plug in-- so the scaling in sigma squared and in n, so that should be completely clear to you. So in particular, if I remove that part, it should be clear to you that this thing, if mean is 0, this thing is actually distributed. Well, if mu is 0, what is the distribution of this guy?

So I remove that part, just this part. So I have xi, which are n0 sigma squared. And I'm asking, what is the distribution of 1/n sum from i equal 1 to n of xi squared?

So it is the sum of their IID. So it's the sum of independent Gaussians, but not standard. So the first thing to make them standard is that I divide all of them by sigma squared.

Now, this guy is of the form zi squared where zi is n01. So now, this thing here has what distribution?

AUDIENCE: Chi squared n.

PHILIPPE RIGOLLET: Chi squared n. And now, sigma squared over n times chi squared n-- so if I have sigma squared divided by n times chi squared-- sorry, so n times n divided by sigma squared. So if I take this thing and I multiply it by n divided by sigma squared, it means I remove this term, and now I am left with a chi squared with n degrees of freedom. Now, the effect of centering with the sample mean here is only to lose 1 degree of freedom. That's it.

So if I want to do a test about variance, since this is supposedly a good estimator of variance, this could be my pivotal distribution. This could play the role of a Gaussian. If I want to know if my variance is equal to 1 or larger than 1, I could actually build a test based on this only statement and test if the variance is larger than 1 or not. Now, this is not asymptotic because I started with the very assumption that my data was Gaussian itself.

Now, just a side remark-- you can check that this chi squared 2, 2 is an exponential with 1/2 degrees of freedom, which is certainly not clear from the fact that z1 squared plus z2 squared is a chi squared with 2 degrees of freedom. if I give you the sum of the square of 2 independent Gaussian, this is actually an exponential. That's not super clear, right?

But if you look at what was here-- I don't know if you took notes, but let me rewrite it for you. So it was x to the d/2 minus 1 e to the minus x/2 divided by 2 to the d/2 gamma of d/2. So if I plug in d is equal to 2, gamma of 2/2 is gamma of 1, which is 1. It's factorial of 0. So it's 1, so this guy goes away.

2 to the d/2 is 2 to 1, so that's just 1. No, that's just 2. Then x the d/2 minus 1 is x the 0, goes away. And so I have x minus x/2 1/2, which is really, indeed, of the form lambda e to the minus lambda x for lambda is equal to 1/2, which was our exponential distribution.

Well, next week is, well, Columbus Day? So not next Monday-- so next week, we'll talk about Student's distribution. And so that was discovered by a guy who pretended his name was Student, but was not Student. And I challenge you to find why in the meantime.

So I'll see you next week. Your homework is going to be outside so we can release the room.