PHILIPPE RIGOLLET: So again, before we start, there is a survey online if you haven't done so. I would guess at least one of you has not. Some of you have entered their answers and their thoughts, and I really appreciate this. It's actually very helpful. So it seems that the course is going fairly well from what I've read so far. So if you don't think this is the case, please enter your opinion and tell us how we can make it better.

One of the things that was said is that I speak too fast, which is absolutely true. I just can't help it. I get so excited, but I will really do my best. I will try to. I think I always start OK. I just end not so well.

So last time we talked about this chi squared distribution, which is just another distribution that's so common that it deserves its own name. And this is something that arises when we sum the squares of independent standard Gaussian random variables. And in particular, why is that relevant? It's because if I look at the sample variance, then it is a chi square distribution, and the parameter that shows up is also known as the degrees of freedom, is the number of observations of minus one.

And so as I said, this chi squared distribution has an explicit probability density function, and I tried to draw it. And one of the comments was also about my handwriting, so I will actually not rely on it for detailed things. So this is what the chi squared with one degree of freedom would look like. And really, what this is is just the distribution of the square of a standard Gaussian. I'm summing only one, so that's what it is.

Then when I go to 2, this is what it is-- 3, 4, 5, 6, and 10. And as I move, you can see this thing is becoming flatter and flatter, and it's pushing to the right. And that's because I'm summing more and more squares, and in expectation we just get one every time. So it really means that the mass is moving to infinity. In particular, a chi squared distribution with n degrees of freedom is going to infinity as n goes to infinity.

Another distribution that I asked you to think about-- anybody looked around about the student t-distribution, what the history of this thing was? So I'll tell you a little bit. I understand if you didn't have time. So the t-distribution is another common distribution that is so common that it will be used and will have its table of quintiles that are drawn at the back of the book.

Now, remember, when I mentioned the Gaussian, I said, well, there are several values for alpha that we're interested in. And so I wanted to draw a table for the Gaussian. We had something that looked like this, and I said, well, q alpha over 2 to get alpha over 2 to the right of this number. And we said that there is a table for this things, for common values of theta.

Well, if you try to envision what this table will look like, it's actually a pretty sad table, because it's basically one list of numbers. Why would I call it a table? Because all I need to tell you is something that looks like this. If I tell you this is alpha and this is q alpha over 2 and then I say, OK, basically the three alphas that I told you I care about are something like 1%, 5%, and 10%, then my table will just give me q alpha over 2.

So that's alpha, and that's q alpha over 2. And that's going to tell me that-- I don't remember this one, but this guy is 1.96. This guy is something like 2.45. I think this one is like 1.65 maybe. And maybe you can be a little finer, but it's not going to be an entire page at the back of the book.

And the reason is because I only need to draw these things for d1 standard Gaussian when the parameters are 0 for the mean and 1 for the variance. Now, if I'm actually doing this for the the chi squared, I basically have to give you one table per values of the degrees of freedom, because those things are different. There is no way I can take-- for Gaussian's, if you give me a different mean, I can substract it and make it back to be a standard Gaussian.

For the chi squared, there is no such thing. There is no thing that just takes the chi squared with d degrees of freedom, nd, and turns it into, say, a chi square with one degree of freedom. This just does not happen.

So the word is standardized. Make it a standard chi squared. There is no such thing as standard chi squared. So what it means is that I'm going to need one row like that for each value of the number of degrees of freedom. So that will certainly fill a page at the back of a book-- maybe even more.

I need one per sample size. So if I want to go from simple size 1 to 1,000, I need 1,000 rows. So now the student distribution is one that arises where it looks very much like the Gaussian distribution, and there's a very simple reason for that, is that I take a standard Gaussian and I divide it by something. That's how I get the student.

What do I divide it with? Well, I take an independent chi square-- I'm going to call it v-- and I

want it to be independent from z. And I'm going to divide z by root v over d. So I start with a chi squared v.

So this guy is chi squared d. I start with z, which is n 0, 1. I'm going to assume that those guys are independent. In my t-distribution, I'm going to write a T. Capital T is z divided by the square root of v over d.

Why would I want to do this? Well, because this is exactly what happens when a divide not by the true variance, a Gaussian, but by its empirical variance. So let's see why in a second. So I know that if you give me some random variable-- let's call it x, which is N mu sigma squared-- then I can do this. x minus mu divided by sigma.

I'm going to call this thing z, because this thing actually has some standard Gaussian distribution. I have standardized x into something that I can read the quintiles at the back of the book. So that's this process that I want to do.

Now, to be able to do this, I need to know what mu is, and I need to know what sigma is. Otherwise I'm not going to be able to make this operation. mu I can sort of get away with, because remember, when we're doing confidence intervals we're actually solving for mu. So it was good that mu was there. When we're doing hypothesis testing, we're actually plugging in here the mu that shows up in h0.

So that was good. We had this thing. Think of mu as being p, for example. But this guy here, we don't necessarily know what it is. I just had to tell you for the entire first chapter, assume you have Gaussian random variables and that you know what the variance is.

And the reason why I said assume you know it-- and I said sometimes you can read it on the side of the box of measuring equipment in the lab. That was just the way I justified it, but the real reason why I did this is because I would not be able to perform this operation if I actually did not know what sigma was. But from data, we know that we can form this estimator Sn, which is 1 over n, sum from i equals 1 to n of Xi, minus X bar squared. And this thing is approximately equal to sigma squared.

That's the sample variance, and it's actually a good estimator just by the law of large number, actually. This thing, by the law of large number, as n goes to infinity-- well, let's say it in probability goes to sigma squared by the law of large number. So it's a consistent estimator of sigma squared. So now, what I want to do is to be able to use this estimator rather than using

sigma.

And the way I'm going to do it is I'm going to say, OK, what I want to form is x minus mu divided by Sn this time. I don't know what the distribution of this guy is. Sorry, it's square root of Sn. This is sigma squared.

So this is what I would take. And I could think of Slutsky, maybe, something like this that would tell me, well, just use that and pretend it's a Gaussian. And we'll see how actually it's sort of valid to do that, because Slutsky tells us it is valid to do that. But what we can also do is to say, well, this is actually equal to x minus mu, divided by sigma, which I knew what the distribution of this guy is. And then what I'm going to do is I'm going to just-- well, I'm going to cancel this effect, sigma over square root Sn.

So I didn't change anything. I just put the sigma here. So now what I know what I know is that this is some z, and it has some standard Gaussian distribution. What is this guy?

Well, I know that Sn-- we wrote this here. Maybe I shouldn't have put those pictures, because now I keep on skipping before and after. We know that Sn times n divided by sigma squared is actually chi squared n minus 1.

So what do I have here? I have that chi squared-- so here I have something that looks like 1 over square root of Sn divided by sigma squared. This is what this guy is if I just do some more writing. And maybe I actually want to make my life a little easier. I'm actually going to plug in my n here, and so I'm going to have to multiply by square root of n here. Everybody's with me?

So now what I end up with is something that looks like this, where I have-- here I started with x. I should really start with Xn bar minus mu times square root of n. That's what the central limit theorem would tell me. I need to work with the average rather than just one observation. So if I start with this, then I pick up a square root of n here.

So if I had the sigma here, I would know that this thing is actually-- Xn bar minus mu divided by sigma times the square root of n would be a standard Gaussian. So if I put Xn bar here, I really need to put this thing that goes around the Xn bar. That's just my central limit theorem that says if I average, then my variance has shrunk by a factor 1 over n.

Now, I can still do this. That was still fine. And now I said that this thing is basically this guy. So what I know is that this thing is a chi squared with n minus 1 degrees of freedom, so this guy

here is chi squared with n minus 1 degrees of freedom.

Let me call this thing v in the spirit of what was used there and in the spirit of what is written here. So this guy was called v, so I'm going to call this v. So what I can write is that square root of n Xn bar minus mu divided by square root of Sn is equal to z times square root of n divided by square root of v. Everybody's with me here?

Which I can rewrite as z times square root of v divided by n And if you look at what the definition of this thing is, I'm almost there. What is the only thing that's wrong here? This is a student distribution, right?

So there's two things. The first one was that they should be independent, and they actually are independent. That's what Cochran's theorem tells me, and you just have to count on me for this. I told you already that Sn was independent of Xn bar. So those two guys are independent, which implies that the numerator and denominator here are independent. That's what Cochran's theorem tells us.

But is this exactly what I should be seeing if I wanted to have my sample variance, if I want to have to write this? Is this actually the definition of a student distribution? Yes? No. So we see z divided by square root of v over d. That looks pretty much like it, except there's a small discrepancy.

What is the discrepancy? There's just the square root of n minus 1 thing. So here, v has n minus 1 degrees of freedom. And in the definition, if the v has d degrees of freedom, I divide it by d, not by d minus 1 or not by d plus 1, actually, in this case.

So I have this extra thing. Well, there's two ways I can address this. The first one is by saying, well, this is actually equal to z over square root of v divided by n minus 1 times square root of n over n minus 1. I can always do that and say for n large enough this thing is actually going to be pretty small, or I can take account for it.

Or for any n you give me, I can compute this number. And so rather than having a t-distribution, I'm going to have a t-distribution time this deterministic number, which is just a function of my number of observations. But what I actually want to do instead is probably use a slightly different normalization, which is just to say, well, why do I have to define Sn-- where was my Sn?

Yeah, why do I have to define Sn tend to be divided by n? Actually, this is a biased estimator, and if I wanted to be unbiased, I can actually just put an n minus 1 here. You can check that. You can expend this thing and compute the expectation. You will see that it's actually not sigma squared, but n over n minus 1 sigma squared.

So you can actually just make it unbiased. Let's call this guy tilde, and then when I put this tilde here what I actually get is s tilde here and s tilde here. I need actually to have n minus 1 here to have this s tilde be a chi squared distribution. Yes?

AUDIENCE: [INAUDIBLE] defined this way so that you--

PHILIPPE RIGOLLET: So basically, this is what the story did. So the story was, well, rather than using always the central limit theorem and just pretending that my Sn is actually the true sigma squared, since this is something I'm going to do a lot, I might as well just compute the distribution, like the quintiles for this particular distribution, which clearly does not depend on any unknown parameter. d is the only parameter that shows up here, and it's completely characterized by the number of observations that you have, which you definitely know.

And so people said, let's just be slightly more accurate. And in a second, I'll show you how the distribution of the T-- so we know that if the sample size is large enough, this should not have any difference with the Gaussian distribution. I mean, those two things should be the same because we've actually not paid attention to this discrepancy by using empirical variance rather than true so far. And so we'll see what the difference is, and this difference actually manifests itself only in small sample sizes.

So those are things that matter mostly if you have less than, say, 50 observations. Then you might want to be slightly more precise and use t-distribution rather than Gaussian. So this is just a matter of being slightly more precise. If you have more than 50 observations, just drop everything and just pretend that this is the true one.

Any other questions? So now I have this thing, and so I'm on my way to changing this guy. So here now, I have not root n but root n minus 1. So I have a z.

So this guy here is S. Yet Where did I get my root n from in the first place? Yeah, because I wanted this guy. And so now what I am left with is Xn minus mu divided by Sn tilde, which is the new one, which is now indeed of the form z v root n minus 1, which now I can write it as z v minus 1.

And so now I have exactly what I want, and so this guy is n 0, 1. And this guy is chi squared with n minus 1 degrees of freedom. And so now I'm back to what I want. So rather than using Sn to be the empirical variance where I just divide my normatizations by n, if I use n minus 1, I'm perfect.

Of course, I can still use n and do this multiplying by root n minus 1 over n at the end. But that just doesn't make as much sense. Everybody's fine with what this T n distribution is doing and why this last line is correct? So that's just basically because it's been defined so that this is actually happening. That was your question, and that's really what happened.

So what is this student t-distribution? Where does the name come from? Well, it does not come from Mr. T. And if you know who Mr. T was-- you're probably too young for that-- he was our hero in the 80s.

And it comes from this guy. His name is Sean William Gosset-- 1908. So that was back in the day. And this guy actually worked at the Guinness Brewery in Dublin, Ireland. And Mr. Guinness back then was a bit of a fascist, and he didn't want him to actually publish papers.

And so what he had to do is to use a fake name to do that. And he was not very creative, and he used a name "student." Because I guess he was a student of life. And so here's the guy, actually. So back in 1908, it was actually not difficult to put your name or your pen name on a distribution.

So what does this thing look like? How does it compare to the standard normal distribution? You think it's going to have heavier or lighter tails compared to the standard distribution, the Gaussian distribution? Yeah, because they have extra uncertainty in the denominator, so it's actually going to make things wiggle a little wider.

So let's start with a reference, which is the standard normal distribution. So that's my usual bell-shaped curve. And this is actually the t-distribution with 50 degrees of freedom. So right now, that's probably where you should just stand up and leave, because you're like, why are we wasting our time?

Those are actually pretty much the same thing, and it is true. If you have 50 observations, both the central limit theorem-- so here one of the things that you need to know is that if I want to talk about t-distribution for, say, eight observations, I need those observations to be Gaussian for real. There's no central limit theorem happening at eight observations.

But really, what this is telling me is not that the central limit theorem kicks in. It's telling me what are the asymptotics that kick in? The law of large number, right? This is exactly this guy. That's here. When I write this statement, what this picture is really telling us is that for n is equal to 50, I'm at the limit already almost. There's virtually no difference between using the left-hand side or using sigma squared.

And now I start reducing. 40, I'm still pretty good. We can start seeing that this thing is actually losing some mass on top, and that's because it's actually pushing it to the left and to the right in the tails. And then we keep going, keep going, keep going.

So that's at 10. When you're at 10, there's not much of a difference. And so you can start seeing difference when you're at five, for example. You can see the tails become heavier.

And the effect of this is that when I'm going to build, for example, a confidence interval to put the same amount of mass to the right of some number-- let's say I'm going to look at this q alpha over 2-- I'm going to have to go much farther, which is going to result in much wider confidence intervals to 4, 3, 2, 1. So that's the t1. Obviously that's the worst. And if you ever use the t1 distribution, please ask yourself, why in the world are you doing statistics based on one observation?

But that's basically what it is. So now that we have this t-distribution, we can define a more sophisticated test than just take your favorite estimator and see if it's far from the value you're currently testing. That was our rationale to build a test before. And the first test that's non-trivial is a test that exploits the fact that the maximum likelihood estimator, under some technical condition, has a limit distribution which is Gaussian with mean 0 when properly centered and a covariance matrix given by the Fisher information matrix. Remember this Fisher information matrix?

And so this is the setup that we have. So we have, again, an i.i.d. sample. Now I'm going to assume that I have a d-dimensional parameter space, theta. And that's why I talk about Fisher information matrix-- and not just Fisher information. It's a number. And I'm going to consider two hypotheses.

So you're going to have h0, theta is equal to theta 0. h1, theta is not equal to theta 0. And this is basically what we thought when we said, are we testing if a coin is fair or unfair. So fair was p equals 1/2, and fair was p different from 1/2. And here I'm just making my life a bit easier.

So now, I have this maximum likelihood estimate that I can construct. Because let's say I know what p theta is, and so I can build a maximum likelihood estimator. And I'm going to assume that these technical conditions that ensure that this maximum likelihood properly standardized converges to some Gaussian are actually satisfy, and so this thing is actually true.

So the theorem, the way I stated it-- if you're a little puzzled, this is not the way I stated it. And the first time, the way we stated it was that theta hat mle minus theta not-- so here I'm going to place myself under the null hypothesis, so here I'm going to say under h0. And honestly, if you have any exercise on tests, that's the way that it should start.

What is the distribution under h0? Because otherwise you don't know what this guy should be. So you have this, and what we showed is that this thing was going in distribution as n goes to infinity to some normal with mean 0 and covariance matrix, which was i of theta, which was here for the true parameter. But here I'm under h0, so there's only one true parameter, which is theta 0.

This was our limiting central limit theorem for-- I mean, it's not really central limited theorem; limited theorem for the maximum likelihood estimator. Everybody remembers that part? The line before said, under technical conditions, I guess. So now, it's not really stated in the same way. If you look at what's on the slide, here I don't have the Fisher information matrix, but I really have the identity of rd.

If I have a random variable x, which has some covariance matrix sigma, how do I turn this thing into something that has covariance matrix identity? So if this was a sigma squared, well, the thing I would do would be divide by sigma, and then I would have a 1, which is also known as the identity matrix of r1. Now, what is this? This was root of sigma squared.

So what I'm looking for is the equivalent of taking sigma and dividing by the square root of sigma, which-- obviously those are matrices-- I'm certainly not allowed to do. And so what I'm going to do is I'm actually going to do the following. Sigma 1 over root of sigma squared can be written as sigma to the negative 1/2. And this is actually the same thing here. So I'm going to write it as sigma to the negative 1/2, and now this guy is actually well-defined.

So this is a positive symmetric matrix, and you can actually define the square root by just taking the square root of its eigenvalues, for example. And so you get sigma 1/2 equals and follows n0 identity. And in general, I'm going to see something that looks like sigma 1/2

negative 1/2 sigma sigma negative 1/2.

And I have minus 1/2 plus 1 minus 1/2. This whole thing collapses to 0, and it's actually the identity. So that's the actual rule. So if you're not familiar, this is basic multivariate Gaussian distribution computations. Take a look at it.

If you feel like you don't need to look at it but you the basic maneuver, it's fine as well. We're not going to go much deeper into that, but those are part of the thing that are sort of standard manipulations about standard Gaussian vectors. Because obviously, standard Gaussian vectors arise from this theorem a lot.

So now I pre-multiplied my sigma to minus minus 1/2. Now of course, I'm doing all of this in the asymptotics, and so I have this effect. So if I pre-multiply everything by sigma to the 1/2, sigma being the Fisher information matrix at theta 0, then this is actually equivalent to saying that square root of n-- so now i of theta now plays the role of sigma-- times theta hat mle minus theta not goes in distribution as n goes to infinity to some multivariate standard Gaussian and 0 identity of rd.

And here, to make sure that we're talking about a multivariate distribution, I can put a d here-- so just so we know we're talking about the multivariate, though it's pretty clear from the context, since the covariance matrix is actually a matrix and not a number. Michael?

AUDIENCE: [INAUDIBLE].

PHILIPPE RIGOLLET: Oh, yeah. Right. Thanks. So Yeah, you're right. So that's a minus and that's a plus. Thanks.

So yeah, anybody has a way to remember whether it's inverse Fisher information or Fisher information as a variance other than just learning it? It is called information, so it's really telling me how much information I have. So when a variance increases, I'm getting less and less information, and so this thing should actually be 1 over a variance. The notion of information is 1 over a notion of variance.

So now I just wrote this guy like this, and the reason why I did this is because now everything on the right-hand side does not depend on any known parameter. There's 0 and identity. Those two things are just absolute numbers or absolute quantities, which means that this thing-- I call this quantity here-- what was the name that I used? Started with a "p." Pivotal.

So this is a pivotal quantity, meaning that its distribution, at least asymptotic distribution, does not depend on any unknown parameter. Moreover, it is indeed a statistic, because I can actually compute it. I know theta 0 and I know theta hat mle. One thing that I did, and you should actually complain about this, is on the board I actually used i of theta not.

And on the slides, it says i of theta hat. And it's exactly the same thing that we did before. Do I want to use the variance as a way for me to check whether I'm under the right assumption or not? Or do I actually want to leave that part and just plug in the theta hat mle, which should go to the true one eventually? Or do I actually want to just plug in the theta 0?

So this is exactly playing the same role as whether I wanted to see square root of Xn bar 1 minus Xn bar in the denominator of my test statistic for p, or if I wanted to see square root of 0.5, 1 minus 0.5 when I was testing if p was equal to 0.5. So this is really a choice that's left up to you, and that's something you can really choose the two. And as we said, maybe this guy is slightly more precise, but it's not going to extend to the case where theta 0 is not reduced to one single number. Any questions?

So now we have our pivotal distribution, so from there this is going to be my test statistic. I'm going to use this as a test statistic and declare that if this thing is too large, n absolute value-- because this is really a way to quantify how far theta hat is from theta 0. And since theta hat should be close to the true one, when this thing is large in absolute value, it means that the true theta should be far from theta 0. So this is my new test statistic.

Now, I said it should be far, but this is a vector. So if I want a vector to be far, two vectors to be far, I measure their norm. And so I'm going to form the Euclidean norm of this guy. So if I look at the Euclidean norm of n-- and Euclidean norm is the one you know-- I'm going to take its square.

Let me now put a 2 here. So that's just the Euclidean norm, and so the norm of vector x is just x transpose x. In the slides, the transpose is denoted by prime. Wow, that's hard to say. Put prime in quotes.

That's a statistic standard that people do. They put prime for transpose. Everybody knows what the transpose is? So I just make it flat and I do it like this, and then that means that's actually equal to the sum of the coordinates Xi squared. And that's what you know.

But here, I'm just writing it in terms of vectors. And so when I run to write this, this is

equivalent, this is equal to-- well, the square root of n is going to pick up the square. So I get square root of n times square root of n.

So this guy is just 1/2. So 1/2 times 1/2 is going to give me 1, and so I get theta hat mle minus theta. And then I have e of theta not. And then I get theta hat mle minus theta not.

And so by definition, I'm going to say that this is my test statistic Tn. And now I'm going to have a test that rejects if Tn is large, because Tn is really measuring the distance between theta hat and theta 0. So my test now is going to be psi, which rejects.

So it says 1 if Tn is larger than some threshold T. And how do I pick this T? Well, by controlling my type I error-- sorry, the c by controlling my type I error. So to choose c, what we have to check is that p under theta not-- so here it's theta not-- that I reject so that psi is equal to 1.

I want this to be equal to alpha, right? That's how I maximize my type I error under the budget that's actually given to me, which is alpha. So that's actually equivalent to checking whether p not of Tn is larger than c.

And so if I want to find the c, all I need to know is what is the distribution of Tn when theta is equal to theta not? Whatever this distribution is-- maybe it has some weird density like this-- whatever this distribution is, I'm just going to be able to pick this number, and I'm going to take this quintile alpha, here alpha, and I'm going to reject if I'm larger than alpha-- whatever this guy is. So to be able to do that, I need to know what is the distribution of Tn when theta is equal to theta 0.

What is this distribution? What is Tn? It's the norm squared of this vector. What is this vector? What is the asymptotic distribution of this vector? Yes?

AUDIENCE: [INAUDIBLE].

PHILIPPE RIGOLLET: Just look one board up. What is this asymptotic distribution of the vector for which we're taking the norm squared? It's right here. It's a standard Gaussian multivariate.

So when I look at the norm squared-- so if z is a standard Gaussian multivariate, then the norm of z squared, by definition of the norm squared, is the sum of the Zi squared. That's just the definition of the norm. But what is this distribution?

AUDIENCE: Chi-squared.

PHILIPPE RIGOLLET: That's a chi-square, because those guys are all of variance 1. That's what the diagonal tells me-- only ones. And they're independent because they have all these zeros outside of the diagonal. So really, this follows some chi-squared distribution. How many degrees of freedom?

Well, the number of them that I sell, d. So now I have found the distribution of Tn under this guy. And that's true because this is true under h0. If I was not under h0, again, I would need to take another guy here.

How did I use the fact that theta is equal to theta 0 when I centered by theta 0? And that was very important. So now what I know is that this is really equal-- why did I put 0 here? So this here is actually equal.

So in the end, I need c such that the probability-- and here I'm not going to put a theta 0. I'm just talking about the possibility of the random variable that I'm going to put in there. It's a chi-square with d degrees of freedom [INAUDIBLE] is equal to alpha.

I just replaced the fact that this guy, Tn, under the distribution was just a chi-square. And this distribution here is just really referring to the distribution of a chi-square. There's no parameters here. And now, that means that I look at my chi-square distribution. It sort of looks like this.

And I'm going to pick some alpha here, and I need to read this number q alpha. And so here what I need to do is to pick this q alpha here, for c. So take c to be q alpha, the quintile of order 1 minus alpha of a chi-squared distribution with this d degree of freedom.

And why do I say 1 minus alpha? Because again, the quintiles are usually referring to the area that's to the left of them by-- well, actually, it's by a convention. However, in statistics, we only care about the right tail usually, so it's not very convenient for us. And that's why rather than calling this guy s sub 1 minus alpha all the time, I write it q alpha.

So now you have this q alpha, which is the 1 minus alpha quintile, or quintile of order 1 minus alpha of chi squared d. And so now I need to use a table. For each d, this thing is going to take a different value, and this is why I can not just spit out a number to you like I spit out 1.96.

Because if I were able to do that, that would mean that I would remember an entire column of this table for each possible value of d, and that I just don't know. So you need just to look at tables, and this is what it will tell you. Often software will do that, too. You don't have to search through tables.

And so just as a remark is that this test, Wald's test, is also valid when I have this sort of other alternative that I could see quite a lot-- if I actually have what's called a one-sided alternative. By the way, this is called Wald's test-- so taking Tn to be this thing. So this is Wald's test.

Abraham Wald was a famous statistician in the early 20th century, who actually was at Columbia for quite some time. And that was actually at the time where statistics were getting very popular in India, so he was actually traveling all over India in some dinky planes. And one of them crashed, and that's how he died-- pretty young.

But actually, there's a huge school of statistics now in India thanks to him. There's the Indian Statistical Institute, which is actually a pretty big thing and trans the best statisticians. So this is called Wald's test, and it's actually a pretty popular test. Let's just look back a second.

So you can do the other alternatives, as I said, and for the other alternatives you can actually do this trick where you put theta 0 as well, as long as you take the theta 0 that's the closest to the alternative. You just basically take the one that's the least favorable to you-- the alternative, I mean. So what is this thing doing? If you did not know anything about statistics and I told you here's a vector-- that's the mle vector, theta hat mle.

So let's say this theta hat mle takes the values, say-- so let's say theta hat mle takes values, say, 1.2, 0.9, and 2.1. And then testing h0, theta is equal to 1, 1, 2, versus theta is not equal to the same number. That's what I'm testing.

So you compute this thing and you find this. If you don't know any statistics, what are you going to do? You're just going to check if this guy goes to that guy, and probably what you're going to do is compute something that looks like the norm squared between those guys-- so the sum.

So you're going to do 1.2 minus 1 squared plus 0.9 minus 1 squared plus 2.1 minus 2 squared and check if this number is large or not. Maybe you are going to apply some stats to try to understand how those things are, but this is basically what you are going to want to do. What Wald's test is telling you is that this average is actually not what you should be doing.

It's telling you that you should have some sort of a weighted average. Actually, it would be a weighted average if I was guaranteed that my Fisher information matrix was diagonal. If my Fisher information matrix is diagonal, looking at this number minus this guy, transpose i, and

then this guy minus this, that would look like I have some weight here, some weight here, and some weight here. Sorry, that's only three.

So if it has non-zero numbers on all of its nine entries, then what I'm going to see is weird cross-terms. If I look at some vector pre-multiplying this thing and post-multiplying this thing-- so if I look at something that looks like this, x transpose i of theta not, x transpose-- think of x as being theta hat mle minus theta-- so if I look at what this guy looks like, it's basically a sum over i and j of Xi, Xj, i, theta not Ij. And so if none of those things are 0, you're not going to see a sum of three terms that are squares, but you're going to see a sum of nine cross-products.

And it's just weird. This is not something standard. So what is Wald's test doing for you? Well, it's saying, I'm actually going to look at all the directions all at once. Some of those directions are going to have more or less variance, i.e., less or more information.

And so for those guys, I'm actually going to use a different weight. So what you're really doing is putting a weight on all directions of the space at once. So what this Wald's test is doing-- by squeezing in the Fisher information matrix, it's placing your problem into the right geometry. It's a geometry that's distorted and where balls become ellipses that are distorted in some directions and shrunk in others, or depending on if you have more variance or less variance in those directions.

Those directions don't have to be aligned with the axes of your coordinate system. And if they were, then that would mean you would have a diagonal information matrix, but they might not be. And so there's this weird geometry that shows up. There is actually an entire field, admittedly a bit dormant these days, that's called information geometry, and it's really doing differential geometry on spaces that are defined by Fisher information matrices.

And so you can do some pretty hardcore-- something that I certainly cannot do-- differential geometry , just by playing around with statistical models and trying to understand with the geometry of those models are. What does it mean for two points to be close in some curved space? So that's basically the idea. So this thing is basically curving your space.

So again, I always feel satisfied when my estimator on my test does not involve just computing an average and checking if it's big or not. And that's not what we're doing here. We know that this theta hat mle can be complicated-- CF problem set, too, I believe. And we know that this Fisher information matrix can also be pretty complicated.

So here, your test is not going to be trivial at all, and that requires understanding the mathematics behind it. I mean, it all built upon this theorem that I just erased, I believe, which was that this guy here inside this norm was actually converging to some standard Gaussian.

So there's another test that you can actually use. So Wald's test is one option, and there's another option. And just like maximum likelihood estimation and method of moments would sometimes agree and sometimes disagree, those guys are going to sometimes agree and sometimes disagree. And this test is called the likelihood ratio test.

So let's parse those words-- "likelihood," "ratio," "test." So at some point, I'm going to have to take the likelihood of something divided by the likelihood of some other thing and then work with this. And this test is just saying the following. Here's the simplest principle you can think of.

You're going to have to understand the notion of likelihood in the context of statistics. You just have to understand the meaning of the word "likelihood." This test is just saying if I want to test h0, theta is equal to theta 0, versus theta is equal to theta 1, all I have to look at is whether theta 0 is more or less likely than theta 1.

And I have an exact number that spits out. Given a theta 0 or a theta 1 and given data, I can put in this function called the likelihood, and they tell me exactly how likely those things are. And so all I have to check is whether one is more likely than the other, and so what I can do is form the likelihood of theta, say, 1 divided by the likelihood of theta 0 and check if this thing is larger than 1. That would mean that this guy is more likely than that guy. That's a natural way to proceed.

Now, there's one caveat here, which is that when I do hypothesis testing and I have this asymmetry between h0 and h1, I still need to be able to control what my probably of type I error is. And here I basically have no knob. This is something if you give me data in theta 0 and theta 1 I can compute to you and spit out the yes/no answer.

But I have no way of controlling the type II and type I error, so what we do is that we replace this 1 by some number c. And then we calibrate c in such a way that the type I error is exactly at level alpha. So for example, if I want to make sure that my type I error is always 0, all I have to do is to say that this guy is actually never more likely than that guy, meaning never reject. And so if I let c go to infinity, then this is actually going to make my type I error go to zero.

But if I let c go to negative infinity, then I'm always going to conclude that h1 is the right one. So I have this straight off, and I can turn this knob by changing the values of c and get different results. And I'm going to be interested in the one that maximizes my chances of rejecting the null hypothesis while staying under my alpha budget of type I error.

So this is nice when I have two very simple hypotheses, but to be fair, we've actually not seen any tests that correspond to real-life example. Where theta 0 was of the form am I equal to, say, 0.5 or am I equal to 0.41, we actually sort of suspected that if somebody asked you to perform this test, they've sort of seen the data before and they're sort of cheating. So it's typically something am I equal to 0.5 or not equal to 0.5 or am I equal to 0.5 or larger than 0.5. But it's very rare that you actually get only two points to test-- am I this guy or that guy?

Now, I could go on. There's actually a nice mathematical theory, something called the Neyman-Pearson lemma that actually tells me that this test, the likelihood ratio test, is the test, given the constraint of type I error, that will have the smallest type II error. So this is the ultimate test. No one should ever use anything different.

And we could go on and do this, but in a way, it's completely irrelevant to practice because you will never encounter such tests. And I actually find students that they took my class as sophomores and then they're still around a couple of years later and they're doing, and they're like, I have this testing problem and I want to use likelihood ratio test, the Neyman-Pearson one, but I just can't because it just never occurs. This just does not happen.

So here, rather than going into details, let's just look at what building on this principle we can actually make a test that will work. So now, for simplicity, I'm going to assume that my alternatives-- so now, I still have a d dimensional vector theta. And what I'm going to assume is that the null hypothesis is actually only testing if the last coefficients from r plus 1 to d are fixed numbers.

So in this example, where I have theta was equal-- so if I have d equals 3, here's an example. h0 is theta 2 equals 1, and theta 3 equals 2. That's my h0, but I say I don't actually care about what theta 1 is going to be. So that's my null hypothesis.

I'm not going to specify right now what the alternative is. That's what the null is. And in particular, this null is actually not of this form. It's not restricting it to one point. It's actually restricting it to an infinite amount of points. Those are all the vectors of the form theta 1 1, 2 for all theta 1 in, say, r. That's a lot of vectors, and so it's certainly not like it's equal to one

specific vector.

So now, what I'm going to do is I'm actually going to look at the maximum likelihood estimator, and I'm going to say, well, the maximum likelihood estimator, regardless of anything, is going to be close to. reality. Now, if you actually tell me ahead of time that the true parameter is of this form, I'm not going to maximize over all three coordinates of theta. I'm just going to say, well, I might as well just set the second one to 1, the third one to 2, and just optimize for this guy.

So effectively, you can say if you're telling me that this is the reality, I can compute a constrained maximum likelihood estimator which is constrained to look like what you think reality is. So this is what the maximum likelihood estimator is. That's the one that's maximizing, say, here the log likelihood over the entire space of candidate vectors, of candidate parameters. But this partial one, this is the constraint mle. That's the one that's actually not maximizing our real thetas, but only over the thetas that are plausible under the null hypothesis.

So in particular, if I look at ln of this constraint thing theta hat n c compared to ln, theta hat-- let's say n mle, so we know which one-- which one is bigger? The first one is bigger. So why?

AUDIENCE: [INAUDIBLE].

PHILIPPE RIGOLLET: So the second one is maximized over a larger space. Right. So I have this all of theta, which are all the parameters I can take, and let's say theta 0 is this guy. I'm maximizing a function over all these things. So if the true maximum is this here, then the two things are equal, but if the maximum is on this side, then the one on the right is actually going to be larger. They're maximizing over a bigger space, so this guy has to be less than this guy.

So maybe it's not easy to see. So let's say that this is theta and this is theta 0 and now I have a function. The maximum over theta 0 is this guy here, but the maximum over the entire space is here.

So the maximum over a larger space has to be larger than the maximum over a smaller space. It can be equal, but the one in the bigger space can be even bigger. However, if my true theta actually did belong to theta 0-- if h0 was true-- what would happen? Well, if theta 0 is true, then theta isn't theta 0, and since the maximum likelihood should be close to theta, it should be the case that those two things should be pretty similar. I should be in a case not in

this kind of thing, but more in this kind of position, where the true maximum is actually attained at theta 0.

And in this case, they're actually of the same size, those two things. If it's not true, then I'm going to see a discrepancy between the two guys. So my test is going to be built on this intuition that if h0 is true, the values of the likelihood at theta hat mle and at the constraint mle should be pretty much the same. But if theta hat-- if it's not true, then the likelihood of the mle should be much larger than the likelihood of the constrained mle.

And this is exactly what this test is doing. So that's the likelihood ratio test. So rather than looking at the ratio of the likelihoods, we look at the difference of the log likelihood, which is really the same thing. And there is some weird normalization factor, too, that shows up here.

And this is what we get. So if I look at the likelihood ratio test, so it's looking at two times ln of theta hat mle minus ln of theta hat mle constrained. And this is actually the test statistic. So we've actually decided that this statistic is what? It's non-negative, right?

We've also decided that it should be close to zero if h0 is true and of course then maybe far from zero if h0 is not true. So what should be the natural test based on Tn? Let me just check that it's-- well, it's already there. So the natural test is something that looks like indicator that Tn is larger than c.

And you should say, well, again? I mean, we just did that. I mean, it is basically the same thing that we just did. Agreed?

But the Tn now is different. The Tn is the difference of log likelihoods, whereas before the Tn was this theta hat minus theta not transpose identity of Fisher information matrix theta hat minus theta not. And this, there's no reason why this guy should be of the same form.

Now, if I have a Gaussian model, you can check that those two things are actually exactly the same. But otherwise, they don't have any reason to be. And now, what's happening is that under some technical conditions-- if h0 is true, now what happens is that if I want to calibrate c, what I need to do is to look at what is the c such that this guy is equal to alpha? And that's for the distribution of T under the knob.

But there's not only one. The null hypothesis here was actually just the family of things. It was not just one vector. It was an entire family of vectors, just like in this example. So if I want my type I error to be constrained over the entire space, what I need to make sure of is that the

maximum overall theta n theta not is actually equal to alpha.

Agreed? Yeah?

AUDIENCE: [INAUDIBLE].

PHILIPPE RIGOLLET: So not equal. In this case, it's going to be not equal. I mean, it can really be anything you want. It's just you're going to have a different type II error.

I guess here we're sort of stuck in a corner. We built this T, and it has to be small under the null. And whatever not the null is, we just hope that it's going to be large. So even if I tell you what the alternative is, you're not going to change anything about the procedure.

So here, q alpha-- so what I need to know is that if h0 is true, then Tn in this case actually converges to some chi-square distribution. And now here, the number of degrees of freedom is kind of weird. But actually, what it should tell you is, oh, finally, I know when you call this parameter degrees of freedom rather than dimension or just d parameter. It's because here what we did is we actually pinned down everything, but r-- sorry, we pinned down everything but r coordinates of this thing.

And so now I'm actually wondering why-- did I make a mistake here? I think this should be chi square with r degrees of freedom. Let me check and send you an update about this, because the number of degrees of freedom, if you talk to normal people they will tell you that here the number of degrees of freedom is r.

This is what's allowed to move, and that's what's called degrees of freedom. The rest is pinned down to being something. So here, this chi-square should be a chi-squared r. And that's something you just have to believe me. Anybody guess what theorem is going to tell me this?

In some cases, it's going to be Cochran's theorem-- just something that tells me that thing's [INAUDIBLE]. Now, here, I use the very specific form of the null alternative. And so for those of you who are sort of familiar with linear algebra, what I did here is h0 consists in saying that theta belongs to an r dimensional linear space. It's actually here, the r dimensional linear space of vectors, that have the first r coordinates that can move and the last coordinates that are fixed to some number.

Actually, it's undefined space because it doesn't necessarily go through zero. And so I have this defined space that has dimension r, and if I were to constrain it to any other r dimensional

space, that would be exactly the same thing. And so to do that, essentially what you need to do is to say, if I take any matrix that's say, invertible-- let's call it u-- and then so h0 is going to be something like of the form u times theta and now I look only at the coordinates r plus 1 2d, then I want to fix those guys to some numbers.

So I want to call them theta, so let's call them tau. So it's going to be tau r plus 1, all the way to tau d. So this is not part of the requirements, but just so you know, it's really not a matter of keeping only some coordinates. Really, what matters is the dimension in the sense of linear subspaces of the problem, and that's what determines what your degrees of freedom are.

So now that we know what the asymptotic distribution is under the null, then we know basically that we know how which table we need to pick our q alpha from. And here, again, the table is a chi-squared table, but here, the number of degrees of freedom is this weird d minus r degrees of freedom thing. I just said it was r. I'm just checking, actually, if I'm-- it's r. It's definitely r.

So here we've made tests. We're testing if r parameter theta was explicitly in some set or not. By explicitly, I mean we're saying, is theta like this or is theta not like this? Is theta equal to theta not or is theta not equal to theta not? Are the last coordinates of theta equal to those fixed numbers, or are they not?

There was something I was stating directly about theta. But there's going to be some instances where you actually want to test something about a function of data, not data itself. For example, is the difference between the first coordinate of theta and the second coordinate of theta positive? That's definitely something you might want to test, because maybe theta 1 is-- let me try to think of some good example.

I don't know. Maybe theta 1 is your drawing accuracy with the right hand and theta 2 is the drawing accuracy with the left hand, and I'm actually collecting data on young children to be able to test early on whether they're going to be left-handed or right-handed, for example. And so I want to just compare those two with respect to each other, but I don't necessarily need to know what the absolute score for this handwriting skills are.

So sometimes it's just interesting to look at the difference of things or maybe the sum, say the combined effect. Maybe this is my two measurements of blood pressure, and I just want to talk about the average blood pressure. And so I can make a linear combination of those two, and so those things implicitly depend on theta.

And so I can generically encapsule them in some test of the form g of theta is equal to 0 versus g of theta is not equal to 0. And sometimes, in the first test that we saw, g of theta was just the identity or maybe the identity minus 0.5. If g of theta is theta minus 0.5, that's exactly what we've been testing. If g of theta is theta minus 0.5 and theta is p, the parameter of a coin, this is exactly of this form. So this is a simple one, but then there's more complicated ones we can think of.

So how can I do this? Well, let's just follow a recipe. So we traced back. We were trying to build a test statistic which was pivotal.

We wanted to have this thing that had nothing that depended on the parameter, and the only thing we had for that that we built in our chi-square test one is basically some form of central limit theorem. Maybe it's for the maximum likelihood estimator. Maybe it's for the average, but it's basically some form of asymptotic normality of the estimator. And that's what we started from every single time.

So let's assume that I have this, and I'm going to talk very abstractly. Let's assume that I start with an estimator. Doesn't have to be the mle. It doesn't have to be the average, but it's just something.

And I know that I have the estimator such that this guy converges in distribution to some n0, and I have some covariance matrix theta. Maybe it's not the Fisher information. Maybe that's something that's not as good as the mle, meaning that this is going to give me less information than the Fisher information, less accuracy.

And now I can actually just say, OK, if I know this about theta, I can apply the multivariate delta method, which tells me that square root of n, g of theta hat, minus g of theta goes in distribution to some n0. And then the price to pay in one dimension was multiplying the square root of the derivative, and we know that in multivariate dimensions pre-multiplying by the gradient, post-multiplying by the gradient. So I'm going to write delta g of theta transpose sigma-- sorry, not delta; nabla-- g of theta-- so gradient.

And here, I assume that g takes values into rk. That's what's written here. g takes value from d to k, but think of k as being 1 for now. So the gradient is really just a vector and not a matrix. That's your usual gradient for real valid functions. So effectively, if g takes values in dimension 1, what is the size of this matrix?

I only ask trivial questions. Remember, that's rule number one. It's one by one, right? And you can check it, because on this side those are just the difference between numbers. And it would be kind of weird if they had a covariance matrix at the end. I mean, this is a random variable, not a random vector.

So I know that this thing happens. And now, if I basically divide by the square root of this thing-- so for board I'm working with k is equal to 1 divided by square root of delta g of theta transpose sigma delta nabla-- sorry, g of theta-- then this thing should go to some standard normal random variable, standard normal distribution. I just divided by square root of the variance here, which is the usual thing.

Now, if you do not have a univariate thing, you do the same thing we did before, which is 3 multiplied by the covariance matrix to the negative 1/2-- so before this role was played by the inverse Fisher information matrix. That's why we ended up having i of theta to the 1/2, and now we just have this gamma, which is just this function that I wrote up there. That could be potentially k by k if g takes values into rk. Yes?

AUDIENCE: [INAUDIBLE].

PHILIPPE RIGOLLET: Yeah, the gradient of a vector is just the vector with all the derivatives with respect to each component, yes. So you know the word vector for derivatives, but not for vectors? I mean, the word gradient you use for one-dimensional? Yes, derivative in one dimension.

Now, of course, here, you notice there's something-- I actually have a little caveat here. I want this to have rank k. I want this to be invertible. I want this matrix to be invertible.

Even for the Fisher information matrix, I sort of need it to be invertible. Even for the original theorem, that was part of my technical condition, just so that I could actually write Fisher information matrix inverse. And so here, you can make your life easy and just assume that it's true all the time, because I'm actually writing in a fairly abstract way. But in practice, we're going to have to check whether this is going to be true for specific distributions.

And we will see an example towards the end of the chapter, the multinomial, where it's actually not the case that Fisher information matrix exists. The asymptotic covariance matrix, is not invertible, so it's not the inverse of a Fisher information matrix. Because to be the inverse of someone, you need to be invertible yourself.

And so now what I can do is apply Slutsky. So here, what I needed to have is theta, the true

theta. So what I can do is just put some theta hat in there, and so that's the gamma of theta hat that I see there.

And if theta is true, then g of theta is equal to 0. That's what we assume. That was our h0, was that under h0 g of theta is equal to 0. So the number I need to plug in here, I don't need to replace theta here. What I need to replace here is 0.

Now let's go back to what you were saying. Here you could say, let me try to replace 0 here, but there is no such thing. There is no g here. It's only the gradient of g. So this thing that says replace theta by theta 0 wherever you see it could not work here.

If g was invertible, I could just say that theta is equal to g inverse of 0 in the null and then I could plug in that value. But in general, it doesn't have to be invertible. And it might be a pain to invert g, even. I mean, it's not clear how you can invert all functions like that. And so here you just go with Slutsky, and you say, OK, I'm just going to put theta hat in there.

But this guy, I know I need to check whether it's 0 or not. Same recipe we did for theta, except we do it for g of theta now. And now I have my asymptotic thing. I know this is a pivotal distribution.

This might be a vector. So rather than looking at the matrix itself, I'm going to actually look at the norm-- rather than looking at the vectors, I'm going to look at their square norm. That gives me a chi square, and I reject when my test statistic, which is the norm square, exceeds the quintile of a chi square-- same as before, just doing on your own.

Before we part ways, I wanted to just mention one thing, which is look at this thing. If g was of dimension 1, the Euclidean norm in dimension 1 is just the absolute value of the number, right? Which means that when I am actually computing this, I'm looking at the square, so it's the square of something. So it means that this is the square of a Gaussian. And it's true that, indeed, the chi squared 1 is just the square of a Gaussian.

Sure, this is the tautology, but let's look at this test now. This test was built using Wald's theory and some pretty heavy stuff. But now if I start looking at Tn and I think of it as being just the absolute value of this quantity over there, squared, what I'm really doing is I'm looking at whether the square of some Gaussian exceeds the quintile of a chi squared of 1 degree of freedom, which means that this thing is actually equivalent-- completely equivalent-- to the test.

So if k is equal to 1, this is completely equivalent to looking at the absolute value of something and check whether it's larger than, say, q over 2-- well, than q alpha-- well, that's q alpha over 2-- so that the probability of this thing is actually equal to alpha. And that's exactly what we've been doing before. When we introduced tests in the first place, we just took absolute values, said, well, is the absolute value of a Gaussian in the limit. And so it's the same thing.

So this is actually equivalent to the probability that the norm squared is larger so that the chi squared of some normal-- and that's the q alpha of some chi squared with one degree of freedom. Those are exactly the two same tests. So in one dimension, those things just collapse into being one little thing, and that's because there's no geometry in one dimension. It's just one dimension, whereas if I'm in a higher dimension, then things get distorted and things can become weird.