PHILIPPE RIGOLLET: So yes, before we start, this chapter will not be part of the midterm. Everything else will be, so all the way up to goodness of fit tests. And there will be some practice exams that will be posted in the recitation section of the course. And that will be-- you will be working on. So the recitation tomorrow will be a review session for the midterm. I'll send an announcement by email.

So going back to our estimator, we showed that the least squares estimator in the case where we had some Gaussian observations. So we had something that looked like this-- y was equal to some matrix x times beta plus some epsilon. This was an equation that was happening in r to the n for n observations. And then we wrote the least squares estimator beta hat.

And for the purpose from here on, you see that you have this normal distribution, this Gaussian p variant distribution. That means that, at some point, we've made the assumption that epsilons were n and dimensional 0 identity of rn times sigma squared, which I kept on forgetting about last time. I will try not to do that this time.

And so from this, we derived a bunch of properties of this least squares estimator, beta hat. And in particular, the key thing that everything was built on was that we could write beta hat as the true unknown beta plus some multivariate Gaussian that was centered, but had a weird covariant structure. So that was definitely p dimensional.

And it was sigma squared times x-- so that's x transpose x. And that's inverse. And the way we derived that was by having a lot of-- at least one cancellation between x transpose x and x transpose x inverse.

So this is the basis for inference in linear regression. So in a way, that's correct, because what happened is that we used the fact that x beta hat-- once we have this beta, x beta hat is really just a projection of y onto the linear span of the columns of x, or the column span of x. And so in particular, those things-- y minus x beta hats-- are called residuals.

So that's the vector of residuals. What's the dimension of this vector?

AUDIENCE: n by 1.

PHILIPPE RIGOLLET: n by 1. So those things, we can write as epsilon hat. There's an estimate for this epsilon because we just put a hat on beta. And from this one, we could actually build an unbiased estimator of sigma hat squared, and that was this guy.

And we showed that, indeed, the right normalization for this was n minus p, because y minus x beta hat to norm is actually a chi squared with n minus p degrees of freedom. And so that's up to this scaling by sigma squared. So that's what we came up with.

And something I told you, which follows from Cochran's theorem-- we did not go into details about this. But essentially, since one of them corresponds to projection onto the linear span of the columns of x, and the other one corresponds to projection onto the orthogonal of this guy, and we're in a Gaussian case, things that are orthogonal are actually independent in a Gaussian case. So from a geometric point of view, you can sort of understand everything. You think of your subspace of the linear span of the x's, sometimes you project onto this guy, sometimes you project onto its orthogonal.

Beta hat corresponds to projection onto the linear span. Epsilon hats correspond to a projection onto the orthogonal. And those things tend to be independent, and that's what you have that beta hat is independent of sigma hat squared. So it's really just a statement about two linear spaces being orthogonal with respect to each other.

So we left on this slide last time. And what I claim is that this thing here is actually-- oh, yeah-- the other thing we want to use. So that's good for beta hat.

But since we don't know what sigma squared is-- if we knew what sigma squared is, that would totally be enough for us. But we also need this extra thing-- that sigma squared hat squared over sigma squared follows-- and there's an n minus p. This follows a chi squared with n minus p degrees of freedom. And sigma hat squared is independent of beta hat.

So that's going to be something we need. So that's useful if sigma squared if unknown. And again, sometimes it might be known if you're using some sort of measurement device for which it's written on the side of the box.

So from these two things, we're going to be able to do inference And inference, we said there's three pillars to inference. The first one is estimation, and we've been doing that so far. We've constructed this least squares estimator, which happens to be the maximum likelihood estimator in the Gaussian case.

The two other things we do in inference are confidence intervals. And we can do confidence intervals. We're not going to do much because we're going to talk about their sort of cousin, which are tests. And that's really where the statistical inference comes into.

And here, we're going to be interested in a very specific kind of test for linear regression. And those are tests of the form beta j-- so the j-th coefficient of beta is equal to 0, and that's going to be our null hypothesis, versus h1 where beta j is, say, not equal to 0. And for the purpose of regression, unless you have lots of domain-specific knowledge, it won't be beta j positive or beta j negative. It's really non-0 that's interesting to you.

So why would I want to do this test? Well, if I expand this thing where I have y is equal to x beta plus epsilon-- so what happens if I look, for example, at the first coordinates? So I have that y is actually-- so say, y1 is equal to beta 1 plus beta 2 x 1.

Well, that's actually complicated. Let me write it like this-- beta 0 plus beta 1 x1 plus beta p minus 1 xp minus 1 plus epsilon. And that's true for all i's. So this is beta 1 times 1. That was our first coordinate.

So that's just expanding this-- going back to the scalar form rather than going to the matrix vector form. That's what we're doing. When I write y is equal to x beta plus epsilon, I assume that each of my y's can be represented as a linear combination of the x's, the first one being 1 plus some epsilon i.

Everybody agrees with this? What does it mean for beta j to be equal to 0? Yeah?

AUDIENCE: That xj's not important.

PHILIPPE RIGOLLET: Yeah, that xj doesn't even show up in this thing. So if beta j is equal to 0, that means that, essentially, we can remove the j's coordinate, xj, from all observations. So for example, I'm a banker, and I'm trying to predict some score-- let's call it y-- without the noise. So I'm trying to predict what is going to be your score.

And that's something that should be telling me how likely you are to reimburse your loan on time or do you have late payments. Or actually, maybe these days bankers are actually looking at how much late fees will I be collecting from you. Maybe that's what they are more after rather than making sure that you reimburse everything. So they're trying to maximize this number of late fees.

And they collect a bunch of things about you-- definitely your credit score, but maybe your zip code, profession, years of education, family status, a bunch of things. One might be your shoe size. And they want to know-- maybe shoe is actually a good explanation for how much fees they're going to be collecting from you.

But as you can imagine, this would be a controversial thing to bring, and people might want to test for their shoe size is a good idea. And so they would just look at the j corresponding to shoe size and test whether shoe size should appear or not in this formula. And that's essentially the kind of thing that people are going to do.

Now, if I do genomics and I'm trying to predict the size, the girth, of a pumpkin for a competition based on some available genomic data, then I can test whether gene j, which is called-- I don't know-- pea snap 24-- they always have these crazy names-- appears or not in this formula. Is the gene pea snap 24 going to be important or not for the size of the final pumpkin? So those are definitely the important things. And definitely, we want to put beta j not equal to 0 as the alternative because that's where scientific discovery shows up.

And so to do that, well, we're in a Gaussian set-up, so we know that even if we don't know what sigma hat is, we can actually call for a t-test. So how did we build the t-test in general? Well, we had something that looked like-- so before, what we had was something that looked like theta hat was equal to theta plus some n0 and something that depended on n, maybe, something like this-- sigma squared over n. So that's what it looked like.

Now what we have is that beta hat is equal to beta plus some n, but this time, it's p variant, and then x transpose x inverse sigma squared. So it's actually very similar, except that the matrix x transpose x inverse is now replacing just this number, 1/n, but it's playing the same role. So in particular, this implies that for every j from 1 to p, what is the distribution of beta hat j? Well, beta hat j is actually equal to-- so all I have to do-- so this is a system of p equations, and all I have to do is to read the j through.

So it's telling me here, I'm going to read beta hat j. Here, I'm going to read beta j. And here, I need to read, what is the distribution of the j-th coordinates of this guy? So this is a Gaussian vector, so we need to understand what its definition is.

So how do I do this? Well, the observation that's actually useful for this-- maybe I shouldn't use the word observation in a stats class, so let's call it claim. The interesting claim is that if I have a vector-- let's call it v-- then vj is equal to v transpose ej where ej is the vector with 0, 0, 0, and then the 1 on the j-th coordinate, and then 0 elsewhere. That's the j-th coordinate. So that's the j-th vector of the canonical basis of rp.

So now that I have this form, I can see that, essentially, beta hat j is just ej transpose this np0 sigma squared x transpose x inverse. And now, I know what the distribution of the inner product between a Gaussian and a deterministic vector is. What is it? It's a Gaussian.

So all I have to check is that ej transpose np0 sigma squared x transpose x inverse-- well, this is equal in distribution to what? Well, this is going to be a one-dimensional thing. A then your product is just a real number. So it's going to be some Gaussian. The mean is going to be 0 in a product with ej, which is 0.

What is the variance of this guy? We actually used this, except that ej was not a vector, but it was a matrix. So what we do is we, to see-- so the rule is that v transpose, say, n mu sigma is some n v transpose mu, and then v transpose sigma v. That's the rule for Gaussian vectors. There's just the property of Gaussian vectors.

So what do we have here? Well, ej plays the role of v. And sigma squared x transpose x inverse is the role of sigma. So here, I'm left with ej transpose-- let me pull out the sigma squared here.

But this thing is, what happens if I take a matrix, I premultiply it by this vector ej, and I postmultiply it by this vector ej? I'm claiming that this corresponds to only one single element of this matrix. Which one is it?

AUDIENCE: j.

PHILIPPE RIGOLLET: j's diagonal element. So this thing here is nothing but x transpose x inverse, and then the j-th diagonal element is jj. Now, I cannot go any further. x transpose x inverse can be a complicated matrix, and I do not know how to express jj's diagonal element much better than this.

Well, no, actually, I don't. It involves basically all the coefficients. Yeah?

AUDIENCE: [INAUDIBLE] second j come from, so I get why ej transpose [INAUDIBLE]. Where did the--

PHILIPPE RIGOLLET: From this rule?

AUDIENCE: [INAUDIBLE]

PHILIPPE RIGOLLET: So you always pre- and postmultiply when you talk about the covariance, because if you did not, it would be a vector and not a scalar, for one. But in general, think of v as a matrix. It's still true even in v is a matrix that's compatible with the premultiplying by some Gaussian. Any other question? Yeah?

AUDIENCE: When you say claim a vector v, what is vector v?

PHILIPPE RIGOLLET: So for any vector v--

AUDIENCE: OK.

PHILIPPE RIGOLLET: Any other question? So now we've identified that the j-th coefficient of this Gaussian, which I can represent from the claim as ej transpose this guy, is also a Gaussian that's centered. And its variance, now, is sigma squared times the j-th diagonal element of x transpose x inverse. So the conclusion is that beta hat j is equal to beta j plus some n. And I'm going to emphasize the fact that now it's one-dimensional with mean 0 and covariance sigma squared x transpose x inverse inverse jj.

Now, if you look at the last line of the second board and the first line on the first board, those are basically the same thing. Beta hat j is my theta hat. Beta j is my theta. And the variance sigma squared over n is now sigma squared times this [? jj's ?] element.

Now, the inverse suggests that it looks like the inverse of n. So those things are going to-- we're going to want to think of those guys as being some sort of 1/n kind of statement. So from this, the fact that those two things are the same leads us to believe that we are now equipped to perform the task that we're trying to do, because under the null hypothesis, beta j is known it's equal to 0, so I can remove it.

And I have to deal with the sigma squared. If sigma squared is known, then I can just perform a regular Gaussian test using Gaussian quintiles. And if sigma squared is unknown, I'm going to just divide by sigma squared and multiply by sigma hat, and then I'm going to basically get my t-test.

Actually, for the purpose of your exam, I really suggest that you understand every single word I'm going to be saying now, because this is exactly the same thing that you're expected to know from other courses, because right now, I'm just going to apply exactly the same technique that we did for the single parameter estimation. So what do we have now is that under h0, beta j is equal to 0. Therefore, beta hat j follows some n0 sigma squared.

Just like I do in the slide, I'm going to call this gamma j. So gamma j is this x transpose x inverse j-th diagonal element. So that implies that beta hat j over sigma-- oh, was it a square root? Yeah, sigma square root of gamma j follows some n0 1.

So I can form my test statistic, which to be reject if the absolute value of beta hat j divided by sigma square root gamma j is larger than what? Can somebody tell me what I want this to be larger than to reject?

AUDIENCE: q alpha.

PHILIPPE RIGOLLET: q alpha. Everybody agrees? Of what? Of this guy, where the standard notation that this is the quintile. Everybody agrees?

AUDIENCE: It's alpha over 2 I think. I think alpha's--

PHILIPPE RIGOLLET: Alpha over 2. So not everybody should be agreeing. Thank you, you're the first one to disagree with yourself, which is probably good. It's alpha over 2 because of the absolute value.

I want to just be away from this guy, and that's because I have-- so the alpha over 2-- the sanity check should be that h1 is beta j not equal to 0. So that works if sigma is known, because I need to know sigma to be able to build my test. So if sigma is unknown, well, I can tell you, use this test, but you're going to be like, OK, when I'm going to have to plug in some numbers, I'm going to be stuck. But if sigma is unknown, we have sigma hat squared as an estimator. So let me write sigma squared here.

So in particular, beta hat divided by sigma hat squared times square root gamma j-- something I can compute. Sorry, that's beta hat j. I can compute that thing. Agreed?

Now I have sigma hat j. What I need to do is to be able to compute the distribution of this thing. So I know the distribution of beta hat j over the square root of gamma j. That's some Gaussian 0, 1.

I don't know exactly what the distribution of sigma hat j squared is, but what I know is that that was actually written, maybe, here is that n minus p sigma hat squared over sigma squared follows some chi squared with n minus p degrees of freedom, and that it's actually independent of beta hat j. It's independent of beta hat, so it's independent of each of its coordinates. That was part of your homework where you had to-- some of you were confused by the fact that-- I mean, if you're independent of some big thing, you're independent of all the smaller components of this big thing.

That's basically what you need to know. And so now I can just write this as-- this is beta hat j divided by-- so now I want to make this guy appear, so it's beta hat j sigma squared over sigma squared-- sigma hat squared over sigma squared times n minus p divided by the square root of gamma j. So that's what I want to see. Yeah?

AUDIENCE: Why do you have to stick the hat in the denominator? Shouldn't it be sigma?

PHILIPPE RIGOLLET: Yeah, so I write this. I decide to write this. I could have put a Mickey Mouse here. It just wouldn't make sense. I just decided to take this thing.

AUDIENCE: OK.

PHILIPPE RIGOLLET: OK. So now, let-- so I take this guy, and now, I'm going to rewrite it as something I want, because if you don't know what sigma is-- sorry, that's not sigm-- you mean the square?

AUDIENCE: Yeah.

PHILIPPE RIGOLLET: Oh, thank you. Yes, that's correct. [LAUGHS] OK, so you don't know what's sigma is, you replace it by sigma hat. That's the most natural thing to do.

You just now want to find out what the distribution of this guy is. So this is not exactly what I had. To be able to get this, I need to divide by sigma squared-- sorry, I need to--

AUDIENCE: Square root.

PHILIPPE RIGOLLET: I'm sorry.

AUDIENCE: Do we need a square root of the sigma hat [INAUDIBLE].

PHILIPPE RIGOLLET: That's correct now. And now I have that-- sorry, I should not write it like that. That's not what I want. What I want is this.

And to be able to get this guy, what I need is sigma over sigma hat square root. And then I need to make this thing show up. So I need to have this n minus p show up in the denominator.

So to be able to get it, I need to multiply the entire thing by the square root of n minus p. So this is just a tautology. I just squeezed in what I wanted.

But now this whole thing here, this is actually of the form beta hat j divided by sigma over square root gamma j, and then divided by square root of sigma hat squared over sigma squared. No, I don't want to divide it by square root of minus p, sorry. And now it's times n minus p divided by n minus p.

And what is the distribution of this thing here? So I'm going to keep going here. So the distribution of this thing here is what? Well, this numerator, what is this distribution?

AUDIENCE: [INAUDIBLE]

PHILIPPE RIGOLLET: Yeah, n0 1. It's actually still written over there. So that's our n0 1. What is the distribution of this guy? Sorry, I don't think you have color again. So what is the distribution of this guy? This is still written on the board.

AUDIENCE: Chi squared.

PHILIPPE RIGOLLET: It's the chi squared that I have right here. So that's a chi squared n minus p divided by n minus p degrees of freedom. The only thing I need to check is that those two guys are independent, which is also what I have from here. And so that implies that beta hat j divided by sigma hat square root of gamma j, what is the distribution of this guy?

[INTERPOSING VOICES]

PHILIPPE RIGOLLET: n minus p. Was that crystal clear for everyone? Was that so simple that it was boring to everyone? OK, good. That's where the point at which you should be.

So now I have this, I can read the quintiles of this guy. So my test statistic becomes-- well, my rejection region, I reject if the absolute value of this new guy exceeds the quintile of order alpha over 2, but this time, of a tn minus p. And now you can actually see that the only difference between this test and that test, apart from replacing sigma by sigma hat, is that now I've moved from the quintiles of a Gaussian to the quintiles of a tn minus p.

What's actually interesting, from this perspective, is that the tn minus p, we know, has heavier tails than the Gaussian, but if the number of degrees of freedom reaches, maybe, 30 or 40, they're virtually the same. And here, the number of degrees of freedom is not given only by n, but it's n minus p. So if I have more and more parameters to estimate, this will result in some heavier, heavier tails, and that's just to account for the fact that it's harder and harder to estimate the variance when I have a lot of parameters. That's basically where it's coming from.

So now let's move on to-- well, I don't know what because this is not working anymore. So this is the simplest test. And actually, if you run any statistical software for least squares, the output in any of them will look like this.

You will have a sequence of rows. And you're going to have an estimate for beta 0, an estimate for beta 1, et cetera. Here, you're going to have a bunch of things. And on this row, you're going to have the value here, so that's going to be what's estimated by least squares.

And then the second line immediately is going to be, well, either the value of this thing-- so let's call it t. And then there's going to be the p value corresponding to this t. This is something that's just routinely coming out because-- oh, and then there's, of course, the last line for people who cannot read numbers that's really just giving you little stars. They're not stickers, but that's close to it.

And that's just saying, well, I have three stars, I'm very significantly different from 0's. If I have 2 stars, I'm moderately differently from 0. And if I have 1 star, it means, well, just give me another $1,000 and I will sign that it's actually different from 0. So that's basically the kind of outputs. Everybody sees what I mean by that?

So what I mean, what I'm trying to emphasize here, is that those things are so routine when you run linear aggression, because people stuff in maybe-- even if you have 200 observations, you're going to stuff in maybe 20 variables-- p equals 20. That's still a big number to interpret what's going on. And it's nice for you if you can actually trim some fat out.

And so the problem is that when you start doing this, and then this, and then this, and then this, the probability that you make a mistake in your test, the probably that you erroneously reject the null here is 5%. Here, it's 5%. Here, it's 5%. Here, it's 5%. And at some point, if things happen with 5% chances and you keep on doing them over and over again, they're going to start to happen. So you can see that basically what's happening is that you actually have an issue is that if you start repeating those tests, you might not be at 5% error at some point.

And so what do you do to prevent from that, if you want to test all those beta j's simultaneously, you have to do what's called the Bonferroni correction. And the Bonferroni correction follows from what's called a union bound. A union bound is actually-- so if you're a computer scientist, you're very familiar with it. If you're a mathematician, that's just, essentially, the third axiom of probability that you see, that the probability of the union is less than the sum of the probabilities. That's the union bound.

And you, of course, can generalize that to more than 2. And that's exactly what you're doing here. So let's see how we would want to perform Bonferroni correction to control the probability that they're all equal to 0 at the same time.

So recall-- so if I want to perform this test over there where I want to test h0, that beta j is equal to 0 for all j in some subset s. So think of s included in 1p. You can think of it as being all of 1 of p if you want. It really doesn't matter. s is something that's given to you. Maybe you want to test the subset of them, but maybe you want to test all of them.

Versus h1, beta j is not equal to 0 for some j in s. That's a test that tests all these things at once. And if you actually look at this table all at once, implicitly, you're performing this test for all of the rows, for s equal 1 to p. You will do that. Whether you like it or not, you will.

So now let's look at what the probability of type I error looks like. So I want the probability of type 1 error, so that's the probably when h0 is true. Well, so let me call psi j the indicator that, say, beta j hat over sigma hat square root gamma j exceeds q alpha over 2 of tn minus p.

So we know that those are the tests that I perform. Here, I just add this extra index j to tell me that I'm actually testing the j-th coefficient. So what I want is the probability that under the null so that those are all equal to 0 that beta j's-- that I will reject to the alternative for one of them.

So that's psi 1 is equal to 1 or psi 2 is equal to 1, all the way to psi-- well, let's just say that this is the entire thing, because it's annoying. I mean, you can check the slide if you want to do it more generally. But psi p is equal to-- or, or-- everybody agrees that this is the probability of type I error? So either I reject this one, or this one, or this one, or this one, or this one. And that's exactly when I'm going to reject at least one of them.

So this is the probability of type I error. And what I want is to keep this guy less than alpha. But what I know is to control the probability that this guy is less than alpha, that this guy is less than alpha, that this guy is less than alpha. In particular, if all these guys are disjoint, then this could really be the sum of all these probabilities.

So in the worst case, if psi j equals 1 intersected with psi k equals 1 is the empty set, so that means those are called disjoint sets. You've seen this terminology in probability, right? So if those sets are disjoint, for all of them, for all j different from k, then this probability-- well, let me write it as star-- then star is equal to, well, the probability under h0 that psi 1 is equal to 1 plus the probability under h0 that psi p is equal to 1.

Now, if I use this test with this alpha here, then this probability is equal to alpha. This probability is also equal to alpha. So the probably of type I error is actually not equal to alpha. It's equal to?

AUDIENCE: p alpha.

PHILIPPE RIGOLLET: p alpha. So what is the solution here? Well, it's to run those guys not with alpha, but with alpha over p. And if they do this, then this guy is equal to alpha over p, this guy is equal to alpha over p. And so when I get those things, I get p times alpha over p, which is just alpha.

So all I do is, rather than running each of the tests with probability of error-- so that's a test at level alpha over p. That's actually very stringent. If you think about it for 1 second, even if you have only 5 variables-- p equals 5-- and you started with the tests, you wanted to do your tests at 5%.

It forces you to do the test at 1% for each of those variables. If you have 10 variables, I mean, that start to be very stringent. So it's going to be harder and harder for you to conclude to the alternative.

Now, one thing I need to tell you is that here I said, if they are disjoint, then those probabilities are equal. But if they are not disjoint, the union bound tells me that the probability of the union is less than the sum of the probabilities. And so now I'm not exactly equal to alpha, but I'm bounded by alpha.

And that's why Bonferroni correction, people are not super comfortable with, is because, in reality, you never think that those tests are going to be giving you completely disjoint things. I mean, why would it be? Why would it be that if this guy is equal to 1, then all the other ones are equal to 0? Why would it make any sense?

So this is definitely conservative, but the problem is that we don't know how to do much better. I mean, we have a formula that tells you the probability of the union as some crazy sum that looks at all the intersection and all these little things. I mean, it's the generalization of p of a or b is equal to p of a plus p of b minus probability of the intersection.

But if you start doing this for more than 2, it's super complicated. The number of terms grows really fast. But most importantly, even if you go here, you still need to control the probability of the intersection.

And those tests are not necessarily independent. If they were independent, then that would be easy. The probably of the intersection would be the product of the probabilities. But those things are super correlated, and so it doesn't really help.

And so we'll see, when we talk about high-dimensional stats towards the end, that there's something called false discovery rate, which is essentially saying, listen, if I want to control this thing, if I really define my probability of type I error as this, I want to make sure that I never make this kind of error, I'm doomed. This is just not going to happen. But I can revise what my goals are in terms of errors that I make, and then I will actually be able to do. And what people are looking at is false discovery rate. And this is called family-wise error rate, which is a stronger thing to control.

So this trick that consists in replacing alpha by alpha over the number of times you're going to be performing your test, or alpha over the number of terms in your union, is actually called the Bonferroni correction. And that's something you use when you have what's called-- another key word here is multiple testing, when you're trying to do multiple tests simultaneously. And if s is not of p, well, you just divide by the number of tests that you are actually making.

So if s is of size k for some k less than p, you just divide alpha by k and not by p, of course. I mean, you can always divide by p, but you're going to make your life harder for no reason. Any question about Bonferroni correction?

So one thing that is maybe not as obvious as the test beta j equal to 0 versus beta j not equal to 0-- and in particular, what it means is that it's not going to come up as a software output without even you requesting it because this is so standard that it's just coming out. But there's other tests that you might think of that might be more complicated and more tailored to your particular problem. And those tests are of the form g times beta is equal to some lambda.

So let's see, the test we've just done, beta j equals 0 versus beta j not equal to 0, is actually equivalent to ej transpose beta equals 0 versus ej transpose beta not equal to 0. That was our claim. But now I don't have to stop here. I don't have to multiply by a vector and test if it's equal to 0.

I can actually replace this by some general matrix g and replace this guy by some general vector lambda. And I'm not telling you what the dimensions are because they're general. I can take whatever I want. Take your favorite matrix, as long as the right side of the matrix can be multiplying beta, and lambda, take it as the number of rows of g, and then you can do that.

I can always formulate this test. What will this test encompass? Well, those are kind of weird tests. So you can think of things like, I want to test if beta 2 plus beta 3 are equal to 0, for example.

Maybe I want to test if beta 5 minus 2 beta 6 is equal to 23. Well, that's weird. But why would you want to test if beta 2 plus beta 3 is equal to 0? Maybe you don't want to know if the-- you know that the effect of some gene is not 0. Maybe you know that this gene affects this trait, but you want to know if the effect of this gene is canceled by the effect of that gene. And this is the kind of stuff that you're going to be testing for that.

Now, this guy is much more artificial, and I don't have a bedtime story to tell you around this. So those things can happen and can be much more complicated. Now, here, notice that the matrix g has one row for both of the examples. But if I want to test if those two things happen at the same time, then I actually can take a matrix.

Another matrix that can be useful is g equals the identity of rp and lambda is equal to 0. What am I doing here in this case? What is this test testing? Sorry, this test. Yeah?

AUDIENCE: Whether or not beta is 0.

PHILIPPE RIGOLLET: Yeah, we're testing if the entire vector beta is equal to 0, because g times beta is equal to beta, and we're asking whether it's equal to 0. So the thing is, when you want to actually test if beta is equal to 0, you're actually testing if your entire model, everything you're doing in life, is just junk. This is just telling you, actually, forget about this y is x beta plus epsilon. y is really just epsilon. There's nothing. There's just some big noise with some big variants, and there's nothing else.

So turns out that the statistical software output that I wrote here spits out an answer to this question. Just the last line, usually, is doing this test. Does your model even make sense? And it's probably for people to check whether they actually just mix their two data sets. Maybe they're actually trying to predict-- I don't know-- some credit score from genomic data, and so just want to make sure, maybe, that's not the right thing.

So it turns out that the machinery is exactly the same as the one we've just taken. So we actually start from here. So let me pull this up. So we start from here. Beta hat was equal to beta plus this guy.

And the first thing we did was to say, well, beta j is equal to this thing because, well, beta j was just ej times beta. So rather than taking ej here, let me just take g. Now, we said that for any vector-- well, that was trivial.

So the thing we need to know is, what is this thing? Well, this thing here, what is this guy? It's also normal and the mean is 0. Again, that's just using properties of Gaussian vectors.

And what is the covariance matrix? Let's call these guys sigma so that you can make an answer, you can formulate an answer. So what is the distribution of-- what is the covariance of g times some Gaussian 0 sigma?

AUDIENCE: g sigma g transpose.

PHILIPPE RIGOLLET: g sigma g transpose, right? So that's gx transpose x inverse g transpose. Now, I'm not going to be able to go much farther. I mean, I made this very acute observation that ej transpose the matrix times ej is the j-th angle element.

Now, if I have a general matrix, the price to pay is that I cannot just shrink this thing any further because I'm trying to be abstract. And so I'm almost there. The only thing that happened last time is that when this was ej under h0, 0, we knew that this was equal to 0 under the null. But under the null, what is this equal to?

AUDIENCE: Lambda.

PHILIPPE RIGOLLET: Lambda, which I know. I mean, I wrote my thing. And in the couple instances I just showed you, including this one over there on top, lambda was equal to 0. But in general, it can be any lambda. But what's key about this lambda is that I actually know it. That's the hypothesis I'm formulating.

So now I'm going to have to be a little more careful when I want to build the distribution of g beta hat. I need to actually subtract this lambda. So now we go from this, and we say, well, g beta hat minus lambda follows some np0 sigma squared g x transpose x inverse g transpose. So that's true.

Let's assume-- let's go straight to the case when we don't know what sigma is. So what I'm going to be interested in is g beta hat minus lambda divided by sigma hat. And that's going to follow some Gaussian that has this thing, gx transpose x inverse g transpose.

So now, what did I do last time? So clearly, the quintiles of this distribution is-- well, OK, what is the size of this distribution? Well, I need to tell you that g is an-- what did I take here?

AUDIENCE: 1 divided by sigma, not sigma hat.

PHILIPPE RIGOLLET: Oh, yeah, you're right. So let me write it like this. Well, let me write it like this-- sigma squared over sigma.

So let's forget about the size of g now. Let's just think of any general g. When g was a vector, what was nice is that this guy was just the scalar number, just one number.

And so if I wanted to get rid of this in the right-hand side, all I had to do was to divide it by this thing. We called it gamma j. And we just had to divide by square root of gamma j, and that would be gone.

Now I have a matrix. So I need to get rid of this matrix somehow because, clearly, the quintiles of this distribution are not going to be written in the back of a book for any value of g and any value of x. So I need to standardize before I can read anything out of a table.

So how do we do it? Well, we just form this guy here. So what we know is that if-- so here's the claim, again, another claim about Gaussian vector. If x follows some n0 sigma, then x transpose sigma inverse x follows some chi squared.

And here, it's going to depend on what is the dimension here. So if I make this a k by k, a k-dimensional Gaussian vector, this is x squared k. Where have we used that before? Yeah?

AUDIENCE: Wald's test.

PHILIPPE RIGOLLET: Wald's test, that's exactly what we used. Wald's test had a chi squared that was showing up. And the way we made it show up was by taking the asymptotic variance, taking its inverse, which, in this framework, was called--

AUDIENCE: Fisher.

PHILIPPE RIGOLLET: Fisher information. And then we pre- and postmultiply by this thing. So this is the key. And so now, it tells me exactly, when I start from this guy that has this multivariate Gaussian, it tells me how to turn it into something that has a distribution which is pivotal. Chi squared k is completely pivotal, does not depend on anything I don't know.

The way I go from here is by saying, well, now, I look at g beta hat minus lambda transpose, and now I need to look at the inverse of the matrix over there. So it's gx transpose x inverse g inverse g beta hat minus lambda. This guy is going to follow-- well, here, I need to actually divide by sigma in this case-- if g is k times p. So what I mean here is just that's the same k. The k that shows up is the number of constraints that I have in my tests.

So now, if I go from here to using sigma hat, the key thing to observe is that this guy is actually not a Gaussian. I'm not going to have a student t-distribution that shows up. So that implies that if I take the same thing, so now I just go from sigma to sigma hat, then this thing is of the form-- well, this chi squared k divided by the chi squared that shows up in the denominator of the t-distribution, which is square root of-- oh, I should not divide by sigma-- so this is sigma squared, right?

AUDIENCE: Yeah.

PHILIPPE RIGOLLET: So this is sigma squared. So this is of the form divided by chi squared n minus p divided by n minus p. So that's the same denominator that I saw in my t-test.

The numerator has changed, though. The numerator is now this chi squared and no longer a Gaussian. But this distribution is actually pivotal, as long as we can guarantee that there's no hidden parameter in the correlation between the two chi squares.

So again, as all statements of independence in this class, I will just give it to you for free. Those two things, I claim-- so OK, let's say admit these are independent. We're almost there. This could be a distribution that's pivotal.

But there's something that's a little unbalanced with it is that this guy is divided by its number of degrees of freedom, but this guy is not divided by its number of degrees of freedom. And so we just have to make the extra step that if I divide this guy by k, and this guy is a chi squared divided by k, if I divide this guy by k, then I get this guy divided by k. And now it looks-- I mean, it doesn't change anything. I've just divided by a fixed number. But it just looks more elegant-- is the ratio of two independent chi squared that are individually divided by the number of degrees of freedom.

And this has a name, and it's called a Fisher or F-distribution. So unlike William Gosset, who was not allowed to use his own name and used the name student, Fisher was allowed to use his own name, and that's called the Fisher distribution. And the Fisher distribution has now 2 parameters, a set of 2 degrees of freedom-- 1 for the numerator and 1 for the denominator.

So F- of Fisher distribution-- so F is equal to the ratio of a chi squared p/p and a chi squared q/q. So that's Fpq P-q where the 2 chi squareds are independent. Is that clear what I'm defining here? So this is basically what plays the role of t-distributions when you're testing more than 1 parameter at a time.

So you basically replace-- the normal that was in the numerator, you replace it by chi squared because you're testing if 2 vectors are simultaneously close. And the way you do it is by looking at their squared norm. And that's how the chi squared shows up.

Quick remark-- are those things really very different? How can I relate a chi squared with a t-distribution? Well, if t follows, say, a t-- I don't know, let's call it q. So that means that t, let me look at-- t is some n01 divided by the square root of a chi squared q/q. That's the distribution of t.

So if I look at the square of the-- the distribution of t squared-- let me put it here-- well, that's the square of some n01 divided by chi squared q/q. Agreed? I just removed the square root here, and I took the square of the Gaussian. But what is the distribution of a square of a Gaussian?

AUDIENCE: Chi squared with 1 degree.

PHILIPPE RIGOLLET: Chi squared with 1 degree of freedom. So this is a chi squared with 1 degree of freedom. And in particular, it's also a chi squared with 1 degree of freedom divided by 1.

So t-squared, in the end, has an F-distribution with 1 and q degrees of freedom. So those two things are actually very similar. The only thing that's going to change is that, since we're actually looking at, typically, absolute values of t when we do our tests, it's going to be exactly the same thing. These quintiles of one guy are going to be, essentially, the square root of the quintiles of the other guy. That's all it's going to be.

So if my test is psi is equal to the indicator that t exceeds q alpha over 2 of tq, for example, then it's equal to the indicator that t-squared exceeds q squared alpha over 2 tq, because I had the absolute value here, which is equal to the indicator that t squared is greater than q alpha over 2. And now this time, it's an F1q. So in a way, those two things belong to the same family. They really are a natural generalization of each other. I mean, at least the F-test is a generalization of the t-test.

And so now I can perform my test just like it's written here. I just formed this guy, and then I perform against the quintile of an F-test. Notice, there's no absolute value-- oh, yeah, I forgot, this is actually q alpha because the F-statistic is already positive. So I'm not going to look between left and right, I'm just going to look whether it's too large or not. So that's by definition.

So you can check-- if you look at a table for student and you look at a table for F1q, one it just going to-- you're going to have to move from one column to the other because you're going to have to move from alpha over 2 to alpha, but one is going to be squared root of the other one, just like the chi squared is the square of the Gaussian. I mean, if you look at the chi squared 1 degree of freedom, you will see the same thing as the Gaussians.

So I'm actually going to start with the last one because you've been asking a few questions about why is my design deterministic. So there's many answers. Some are philosophical.

But one that's actually-- well, there's the one that says everything you cannot do if you don't have a condition-- if you don't have x, because all of the statements that we made here, for example, just the fact that this is chi squared, if those guys start to be random variables, then it's clearly not going to be a chi squared. I mean, it cannot be chi squared when those guys are deterministic and when they are random. I mean, things change. So that's just maybe [INAUDIBLE] check statement.

But I think the one that really matters is that-- remember when we did the t-test, we had this gamma j that showed up. Gamma j was playing the role of the variance. So here, the variance, you never think of-- I mean, we'll talk about this in the Bayesian setup, but so far, we haven't thought of the variance as a random variable.

And so here, your x's really are the parameters of your data. And the diagonal elements of x transpose x inverse actually tell you what the variance is. So that's also one reason why you should think of your x as being a deterministic number. They are, in a way, things that change the geometry of your problem. They just say, oh, let me look at it from the perspective of x.

Actually, for that matter, we didn't really spend much time commenting on what is the effect of x onto gamma. So remember, gamma j, so that was the variance parameter. So we should try to understand what x's lead to big variance and what x's lead to small variance. That would be nice.

Well, if this is the identity matrix-- let's say identity over n, which is the natural thing to look at, because we want this thing to scale like 1/n-- then this is just 1/n. We're back to the original case. Yes?

AUDIENCE: Shouldn't that be inverse?

PHILIPPE RIGOLLET: Yeah, thank you. x inverse, yes. So if this is the identity, then, well, the inverse is-- let's say just this guy here is n times this guy. So then the inverse is 1/n. So in this case, that means that gamma j is equal to 1/n and we're back to the theta hat theta case, the basic one-dimensional thing.

What does it mean for a matrix for when I take its-- yeah, so that's of dimension p. But when I take its transpose-- so forget about the scaling by n right now. This is just a matter of scaling things. I can always multiply my x's so that I have this thing that shows up. But when I have a matrix, if I look at x transpose x and I get something which is the identity, how do I call this matrix?

AUDIENCE: Orthonormal?

PHILIPPE RIGOLLET: Orthogonal, yeah. Orthonormal or orthogonal. So you call this thing an orthogonal matrix. And when it's an orthogonal matrix, what it means is that the-- so this matrix here, if you look at the matrix xx transpose, the entries of this matrix are the inner products between the columns of x. That's what's happening.

You can write it, and you will see that the entries of this matrix are linear products. If it's the identity, that means that you get some 1's and a bunch of 0's, it means that all the inner products between 2 different columns is actually 0. What it means is that this matrix x is an orthonormal basis for your space. The columns form an orthonormal basis. So they're basically as far from each other as they can.

Now, if I start making those guys closer and closer, then I'm starting to have some issues. x transpose x is not going to be the identity. I'm going to start to have some non-0 entries. But if they all remain of norm 1, then-- oh, sorry, so that's for the inverse.

So I first start putting some stuff here, which is non-0, by taking my x's. Rather than having this, I move to this. Now I'm going to start seeing some non-0 entries.

And when I'm going to take the inverse of this matrix, the diagonal elements are going to start to blow up. Oh, sorry, the diagonals start to become smaller and smaller. So when I take the inverse-- no, sorry, the diagonal limits are going to blow up.

And so what it means is that the variance is going to blow up. And that's essentially telling you that if you get to choose your x's, you want to take them as orthogonal as you can. But if you don't, then you just have to deal with it, and it will have a significant impact on your estimation performance.

And that's what, also, routinely, statistical software is going to spit out this value here for you. And you're going to have-- well, actually square root of this value. And it's going to tell you, essentially-- you're going to know how much randomness, how much variation you have in this particular parameter that you're estimating.

So if gamma j is large, then you're going to have wide confidence intervals and your tests are not going to reject very much. And that's all captured by x. That's what's important. Everything, all of this, is completely captured by x.

Then, of course, there was the sigma squared that showed up here. Actually, it was here, even in the definition of gamma j. I forgot it. What is the sigma squared police doing?

And so this thing was here as well, and that's just exogenous. It comes from the noise itself. But there was this huge factor that came from the x's itself.

So let's go back, now, to reading this list in a linear fashion. So I mean, you're MIT students, you've probably heard that correlation does not imply causation many times. Maybe you don't know what it means. If you don't, that's OK, you just have to know the sentence.

No, what it means is that it's done because I decided that something was going to be the x and that something else was going to be the y, that whatever thing I'm getting, it means that x implies y. For example, even if I do genetics, genomics, or whatever, I mean, I implicitly assume that my genes are going to have an effect on my outside look. I could be the opposite.

I mean, who am I to say? I'm not a biologist. I don't know. I didn't open a biology book in 20 years. So maybe, if I start hitting my head with a hammer, I'm going to have changing my genetic material. Probably not, but that's why-- but causation definitely does not come from statistics.

So if you know that that's the different thing, it's actually going to-- it's not coming from there. So actually, I remember, once, I put an exam to students, and there was an old data set from police expenditures, I think, in Chicago in the '60s. And they were trying to understand-- no, it was on crime. It was the crime data set.

And they were trying-- so the y variable was just the rate of crime, and the x's were a bunch of things, and one of them was police expenditures. And if you rend the regression, you would find that the coefficient in front of police expenditure was a positive number, which means that if you increase police expenditures, that increases the crime. I mean, that's what it means to have a positive coefficient.

Everybody agrees with this fact? If beta j is 10, then it means that if I increase by $1 my police expenditure, I [INAUDIBLE] by 10 my crime, everything else being kept equal. Well, there were, I think, about 80% of the students that were able to explain to me that if you give more money to the police, then the crime is going to raise. Some people were like, well, police is making too much money, and they don't think about their work, and they become lazy. And I mean, people were really coming up with some crazy things.

And what it just meant is that, no, it's not causation. It's just, if you have more crime, you give more money to your police. That's what's happening. And that's all there is.

So just be careful when you actually draw some conclusions that causation is a very important thing to keep in mind. And in practice, unless you have external sources of reason for causality-- for example, genetic material and physical traits, we agree upon what the direction of the arrow of causality is here. There's places where you might not.

Now, finally, the normality on the noise-- everything we did today required normal Gaussian distribution on the noise. I mean, it's everywhere. There's some Gaussian, there's some chi squared. Everything came out of Gaussian.

And for that, we needed this basic formula for inference, which we derived from the fact that the noise was Gaussian itself. If we did not have that, the only thing we could write is, beta hat is this number, or this vector. We would not be able to say, the fluctuations of beta hat are this guy. We would not be able to do tests. We would not be able to build, say, confidence regions or anything.

And so this is an important condition that we need, and that's what statistical software assumes by default. But we now have a recipe on how to do those tests. We can do it either visually, if we really want to conclude that, yes, this is Gaussian, using our normal Q-Q plots. And we can also do it using our favorite tests.

What test should I be using to test that? With two names? Yeah?

AUDIENCE: Normal [INAUDIBLE].

PHILIPPE RIGOLLET: Not the 2 Russians. So I want a Russian and a Scandinavian person for this one. What's that?

AUDIENCE: Lillie-something?

PHILIPPE RIGOLLET: Yeah, Lillie-something. So Kolmogorov Lillie-something test. And [LAUGHS] so it's the Kolmogorov Lilliefors test. And because I'm testing if there Gaussian, and I'm actually not really making any-- I don't need to know what the variance is.

The mean is 0. We saw that at the beginning. It's 0 by construction, so we actually don't need to think about the mean being 0 itself. This just happens to be 0.

So we know that it's 0, but the variance, we don't know. So we just want to know if it belongs to the family of Gaussians, and so we need to Kolmogorov Lilliefors for that. And that's also one of the thing that's spit out by statistical software by default. When you run a linear regression, actually, it spits out both Kolmogorov-Smirnov and Kolmogorov Lilliefors, probably contributing to the widespread use of Kolmogorov-Smirnov when you really shouldn't.

So next time, we will talk about more advanced topics on regression. But I think I'm going to stop here for today. So again, tomorrow, sometime during the day, at least before the recitation, you will have a list of practice exercises that will be posted. And if you go to the optional recitation, you will have someone solving them