PHILIPPE RIGOLLET: --124. If I were to repeat this 1,000 times, so every one of those 1,000 times they collect 124 data points and then I'd do it again and do it again and again, then in average, the number I should get should be close to the true parameter that I'm looking for. The fluctuations that are due to the fact that I get different samples every time should somewhat vanish. And so what I want is to have a small bias, hopefully a 0 bias. If this thing is 0, then we see that the estimator is unbiased.

So this is definitely a property that we are going to be looking for in an estimator, trying to find them to be unbiased. But we'll see that it's actually maybe not enough. So unbiasedness should not be something you lose your sleep over.

Something that's slightly better is the risk, really the quadratics risk, which is expectation of-- so if I have an estimator, theta hat, I'm going to look at the expectation of theta hat n minus theta squared. And what we showed last time is that we can actually-- by inserting in there, adding and removing the expectation of theta hat, we actually get something where this thing can be decomposed as the square of the bias plus the variance, which is just the expectation of theta hat minus its expectation squared. That came from the fact that when I added and removed the expectation of theta hat in there, the cross-terms cancel.

All right. So that was the bias squared, and this is the variance. And so for example, if the quadratic risk goes to 0, then that means that theta hat converges to theta in the L2 sense. And here we know that if we want this to go to 0, since it's the sum of two positive terms, we need to have both the bias that goes to 0 and the variance that goes to 0, so we need to control both of those things.

And so there is usually an inherent trade-off between getting a small bias and getting a small variance. If you reduce one too much, then the variance of the other one is going to-- then the other one is going to increase, or the opposite. That happens a lot, but not so much, actually, in this class.

So let's just look at a couple of examples. So am I planning-- yeah. So examples. So if I do, for example, X1, Xn, there are iid Bernoulli. And I'm going to write it theta so that we keep the same notation. Then theta hat, what is the theta hat that we proposed many times? It's just X bar, Xn bar, the average of Xi's.

So what is the bias of this guy? Well, to know the bias, I just have to remove theta from the expectation. What is the expectation of Xn bar? Well, by linearity of the expectation, it's just the average of the expectations.

But since all my Xi's are Bernouilli with the same theta, then each of this guy is actually equal to theta. So this thing is actually theta, which means that this isn't biased, right?

Now, what is the variance of this guy? So if you forgot the properties of the variance for sum of independent random variables, now it's time to wake up. So we have the variance of something that looks like 1 over n, the sum from i equal 1 to n of Xi.

So it's of the form variance of a constant times a random variable. So the first thing I'm going to do is pull out the constant. But we know that the variance leaves on the square scale, so when I pull out a constant outside of the variance, it comes out with a square. The variance of a times X is a-squared times the variance of X, so this is equal to 1 over n squared times the variance of the sum.

So now we want to always do what we want to do. So we have the variance of the sum. We would like somehow to say that this is the sum of the variances. And in general, we are not allowed to say that, but we are because my Xi's are actually independent. So this is actually equal to 1 over n squared sum from i equal 1 to n of the variance of each of the Xi's. And that's by independence, so this is basic probability.

And now, what is the variance of Xi's where again they're all the same distribution, so the variance of Xi is the same as the variance of X1. And so each of those guys has variance what? What is the variance of a Bernoulli? We've said it once. It's theta times 1 minus theta.

And so now I'm going to have the sum of n times a constant, so I get n times the constant divided by n squared, so one of the n's is going to cancel. And so the whole thing here is actually equal to theta, 1 minus theta divided by n.

So if I'm interested in the quadratic risk-- and again, I should just say risk, because this is the only risk we're going to be actually looking at. Yeah. This parenthesis should really stop here. I really wanted to put quadratic in parenthesis.

So the risk of this guy is what? Well, it's the expectation of x bar n minus theta squared. And we know it's the square of the variance, so it's the square of the bias, which we know is 0, so it's 0 squared plus the variance, which is theta, 1 plus theta-- 1 minus theta divided by n. So it's just theta, 1 minus theta divided by n.

So this is just summarizing the performance of an estimator, which is the random variable. I mean, it's complicated. If I really wanted to describe it, I would just tell you the entire distribution of this random variable.

But now what I'm doing is I'm saying, well, let's just take this random variable, remove theta from it, and see how small the fluctuations around theta-- the squared fluctuations around theta are in expectation. So that's what the quadratic risk is doing.

And in a way, this decomposition, as the sum of the bias square and the variance, is really telling you that-- it is really accounting for the bias, which is, well, even if I had an infinite amount of observations, is this thing doing the right thing? And the other thing is actually the variance, so for finite number of observations, what are the fluctuations?

All right. Then you can see that those things, bias and variance, are actually very different. So I don't have any colors here, so you're going to have to really follow the speed-- the order in which I draw those curves.

All right. So let's find-- I'm going to give you three candidate estimators, so-- estimators for theta. So the first one is definitely Xn bar. That will be a good candidate estimator.

The second one is going to be 0.5, because after all, why should I bother if it's actually going to be-- right? So for example, if I ask you to predict the score of some candidate in some election, then since you know it's going to be very close to 0.5, you might as well just throw 0.5 and you're not going to be very far from reality. And it's actually going to cost you 0 time and $0 to come up with that. So sometimes maybe just a good old guess is actually doing the job for you.

Of course, for presidential elections or something like this, it's not very helpful if your prediction is telling you this. But if it was something different, that would be a good way to generate some close to 1/2. For a coin, for example, if I give you a coin, you never know. Maybe it's slightly biased. But the good guess, just looking at it, inspecting it, maybe there's something crazy happening with the structure of it, you're going to guess that it's 0.5 without trying to collect information.

And let's find another one, which is, well, you know, I have a lot of observations. But I'm recording couples kissing, but I'm on a budget. I don't have time to travel all around the world and collect some people. So really, I'm just going to look at the first couple and go home. So my other estimator is just going to be X1. I just take the first observation, 0, 1, and that's it.

So now I'm going-- I want to actually understand what the behavior of those guys is. All right. So we know-- and so we know that for this guy, the bias is 0 and the variance is equal to theta, 1 minus theta divided by n.

What is the bias of this guy, 0.5?

AUDIENCE: 0.5.

AUDIENCE: 0.5 minus theta?

PHILIPPE RIGOLLET: 0.5 minus theta, right. So the bias, 0.5 minus theta. What is the variance of this guy? What is the variance of 0.5?

AUDIENCE: It's 0.

PHILIPPE RIGOLLET: 0. Right. It's just a deterministic number, so there's no fluctuations for this guy.

What is the bias? Well, X1 is actually-- just for simplicity, I can think of it as being X1 bar, the average of itself, so that wherever I saw an n for this guy, I can replace it by 1 and that will give me my formula. So the bias is still going to be 0. And the variance is going to be equal to theta, 1 minus theta.

So now I have those three estimators. Well, if I compare X1 and Xn bar, then clearly I have 0 bias in both cases. That's good. And I have the variance that's actually n times smaller when I use my n observations than when I don't.

So those two guys, on these two fronts, you can actually look at the two numbers and say, well, the first number is the same. The second number is better for the other guy, so I will definitely go for this guy compared to this guy. So this guy is gone.

But not this guy. Well, if I look at the bias, the variance is 0. It's always beating the variance of this guy. And if I look at the bias, it's actually really not that bad. It's 0.5 minus theta. In particular, if theta is 0.5, then this guy is strictly better.

And so you can actually now look at what the quadratic risk looks like. So here, what I'm going to do is I'm going to take my true theta-- so it's going to range between 0 and 1. And we know that those two things are functions of theta, so I can only understand them if I plot them as functions of theta.

And so now I'm going to actually plot-- the y-axis is going to be the risk. So what is the risk of the estimator of 0.5? This one is easy. Well, it's 0 plus the square of 0.5 minus theta.

So we know that at theta, it's actually going to be 0. And then it's going to be a square. So at 0, it's going to be 0.25. And at 1, it's going to be 0.25 as well. So it looks like this.

Well, actually, sorry. Let me put the 0.5 where it should be. OK. So this here is the risk of 0.5. And we'll write it like this.

So when theta is very close to 0.5, I'm very happy. When theta gets farther, it's a little bit annoying. And then here, I want to plot the risk of this guy.

So now the thing with the risk of this guy is that it will depend on n. So I will just pick some n that I'm happy with just so that I can actually draw a curve. Otherwise, I'm going to have to plot one curve per value of n.

So let's just say, for example, that n is equal to 10. And so now I need to plot the function theta, 1 minus theta divided by 10. We know that theta, 1 minus theta is a curve that goes like this. It takes value at 1/2. It thinks value 1/4. That's the maximum. And then it's 0 at the end.

So really, if n is equal to 1, this is what the variance looks like. The bias doesn't count in the risk. Yeah.

AUDIENCE: [INAUDIBLE]

PHILIPPE RIGOLLET: Sure. Can you move? All right. Are you guys good?

All right. So now I have this picture. And I know I'm going up to 25. And there's a place where those curves cross.

So if you're sure-- let's say you're talking about presidential election, you know that those things are going to be really close. Maybe you're actually better by predicting 0.5 if you know it's not going to go too far. But that's for one observation, so that's the risk of X1.

But if I look at the risk of Xn, all I'm doing is just crushing this curve down to 0. So as n increases, it's going to look more and more like this. It's the same curve divided by n. And so now I can just start to understand that for different values of thetas, now I'm going to have to be very close to theta is equal to 1/2 if I want to start saying that Xn bar is worse than the naive estimator 0.5. Yeah.

AUDIENCE: Sorry. I know you explained a little bit before, but can you just-- what is an intuitive definition of risk? What is it actually describing?

PHILIPPE RIGOLLET: So either you can-- well, when you have an unbiased estimator, it's simple. It's just telling you it's the variance, because the theta that you have over there is really-- so in the definition of the risk, the theta that you have here if you're unbiased is really the expectation of theta hat. So that's really just the variance.

So the risk is really telling you how much fluctuations I have around my expectation if unbiased. But actually here, it's telling you how much fluctuations I have in average around theta. So if you understand the notion of variance as being--

AUDIENCE: [INAUDIBLE]

PHILIPPE RIGOLLET: What?

AUDIENCE: Like variance on average.

PHILIPPE RIGOLLET: No.

AUDIENCE: No.

PHILIPPE RIGOLLET: It's just like variance.

AUDIENCE: Oh, OK.

PHILIPPE RIGOLLET: So when you-- I mean, if you claim you understand what variance is, it's telling you what is the expected squared fluctuation around the expectation of my random variable. It's just telling you on average how far I'm going to be. And you take the square because you want to cancel the signs. Otherwise, you're going to get 0.

AUDIENCE: Oh, OK.

PHILIPPE RIGOLLET: And here it's saying, well, I really don't care what the expectation of theta hat is. What I want to get to is theta, so I'm looking at the expectation of the squared fluctuations around theta itself. If I'm unbiased, it coincides with the variance. But if I'm biased, then I have to account for the fact that I'm really not computing the--

AUDIENCE: OK. OK. Thanks.

PHILIPPE RIGOLLET: OK? All right. Are there any questions?

So here, what I really want to illustrate is that the risk itself is a function of theta most of the times. And so for different thetas, some estimators are going to be better than others. But there's also the entire range of estimators, those that are really biased, but the bias can completely vanish.

And so here, you see you have no bias, but the variance can be large. Or you have 0 bias-- you have a bias, but the variance is 0. So you can actually have this trade-off and you can find things that are in the entire range in general.

So those things are actually-- those trade-offs between bias and variance are usually much better illustrated if we're talking about multivariate parameters. If I actually look at a parameter which is the mean of some multivariate Gaussian, so an entire vector, then the bias is going to-- I can make the bias bigger by, for example, forcing all the coordinates of my estimator to be the same. So here, I'm going to get some bias, but the variance is actually going to be much better, because I get to average all the coordinates for this guy.

And so really, the bias/variance trade-off is when you have multiple parameters to estimate, so you have a vector of parameters, a multivariate parameter, the bias increases when you're trying to pull more information across the different components to actually have a lower variance. So the more you average, the lower the variance. That's exactly what we've illustrated. As n increases, the variance decreases, like 1 over n or theta, 1 minus theta over n. And so this is how it happens in general.

In this class, it's mostly one-dimensional parameter estimation, so it's going to be a little harder to illustrate that. But if you do, for example, non-parametric estimation, that's all you do. There's just bias/variance trade-offs all the time. And in between, when you have high-dimensional parametric estimation, that happens a lot as well.

OK. So I'm just going to go quickly through those two remaining slides, because we've actually seen them. But I just wanted you to have somewhere a formal definition of what a confidence interval is. And so we fixed a statistical model for n observations, X1 to Xn.

The parameter theta here is one-dimensional. Theta is a subset of the real line, and that's why I talk about intervals. An interval is a subset of the line. If I had a subset of R2, for example, that would no longer be called an interval, but a region, just because-- well, that's just we can say a set, a confidence set. But people like to say confidence region.

So an interval is just a one-dimensional conference region. And it has to be an interval as well.

So a confidence interval of level 1 minus alpha-- so we refer to the quality of a confidence interval is actually called it's level. It takes value 1 minus alpha for some positive alpha. And so the confidence level-- the level of the confidence interval is between 0 and 1. The closer to 1 it is, the better the confidence interval. The closer to 0, the worse it is.

And so for any random interval-- so a confidence interval is a random interval. The bounds of this interval depends on random data. Just like we had X bar plus/minus 1 over square root of n, for example, or 2 over square root of n, this X bar was the random thing that would make fluctuate those guys.

And so now I have an interval. And now I have its boundaries, but now the boundaries are not allowed to depend on my unknown parameter. Otherwise, it's not a confidence interval, just like an estimator that depends on the unknown parameter is not an estimator. The confidence interval has to be something that I can compute once I collect data.

And so what I want is that-- so there's this weird notation. The fact that I write theta-- that's the probability that I contains theta. You're used to seeing theta belongs to I. But here, I really want to emphasize that the randomness is in I. And so the way you actually say it when you read this formula is the probability that I contains theta is at least 1 minus alpha.

So it better be close to 1. You want 1 minus alpha to be very close to 1, because it's really telling you that whatever random variable I'm giving you, my error bars are actually covering the right theta. And I want this to be true.

But I want this-- since I don't know what my confidence-- my parameter of theta is, I want this to hold true for all possible values of the parameters that nature may have come up with from. So I want this-- so there's theta that changes here, so the distribution of the interval is actually changing with theta hopefully. And theta is changing with this guy. So regardless of the value of theta that I'm getting, I want that the probability that it contains the theta is actually larger than 1 minus alpha.

So I'll come back to it in a second. I just want to say that here, we can talk about asymptotic level. And that's typically when you use central limit theorem to compute this guy. Then you're not guaranteed that the value is at least 1 minus alpha for every n, but it's actually in the limit larger than 1 minus alpha.

So maybe for each fixed n it's going to be not true. But for as no goes to infinity, it's actually going to become true. If you want this to hold for every n, you actually need to use things such as Hoeffding's inequality that we described at some point, that hold for every n.

So as a rule of thumb, if you use the central limit theorem, you're dealing with a confidence interval with asymptotic level 1 minus alpha. And the reason is because you actually want to get the quintiles of the normal-- the Gaussian distribution that comes from the central limit theorem. And if you want to use Hoeffding's, for example, you might actually get away with a confidence interval that's actually true even non-asymptotically. It's just the regular confidence interval.

So this is the formal definition. It's a bit of a mouthful. But we actually-- the best way to understand them is to build them.

Now, at some point I said-- and I think it was part of the homework-- so here, I really say the probability the true parameter belongs to the confidence interval is actually 1 minus alpha. And so that's because here, this confidence interval is still a random variable.

Now, if I start plugging in numbers instead of the random variables X1 to Xn, I start putting 1, 0, 0, 1, 0, 0, 1, like I did for the kiss example, then in this case, the random interval is actually going to be 0.42, 0.65. And this guy, the probability that theta belongs to it is not 1 minus alpha. It's either 0 if it's not in there or it's 1 if it's in there.

So here is the example that we had. So just let's look at back into our favorite example, which is the average of Bernoulli random variables, so we studied that maybe that's the third time already. So the sample average, Xn bar, is a strongly consistent estimator of p. That was one of the properties that we wanted.

Strongly consistent means that as n goes to infinity, it converges almost surely to the true parameter. That's the strong law of large number. It is consistent also, because it's strongly consistent, so it also converges in probability, which makes it consistent.

It's unbiased. We've seen that. We've actually computed its quadratic risk.

And now what I have is that if I look at-- thanks to the central limit theorem, we actually did this. We built a confidence interval at level 1 minus alpha-- asymptotic level, sorry, asymptotic level 1 minus alpha.

And so here, this is how we did it. Let me just go through it again. So we know from the central limit theorem-- so the central limit theorem tells us that Xn bar minus p divided by square root of p1 minus p, square root of n converges in distribution as n goes to infinity to some standard normal distribution.

So what it means is that if I look at the probability under the true p, that's square root of n, Xn bar minus p divided by square root of p1 minus p, it's less than Q alpha over 2, where this is the definition of the quintile. Then this guy-- and I'm actually going to use the same notation, limit as n goes to infinity, this is the same thing. So this is actually going to be equal to 1 minus alpha.

That's exactly what I did last time. This is by definition of the quintile of a standard Gaussian and of a limit in distribution. So the probabilities computed on this guy in the limit converges to the probability computed on this guy. And we know that this is just the probability that the absolute value of sum n 0, 1 is less than Q alpha over 2.

And so in particular, if it's equal, then I can put some larger than or equal to, which guarantees my asymptotic confidence level. And I just solve for p. So this is equivalent to the limit as n goes to infinity of the probability that theta is between Xn bar minus Q alpha over 2 divided by-- times square root of p1 minus p divided by square root of n, Xn bar plus q alpha over 2, square root of p1 minus p divided by square root of n is larger than or equal to 1 minus alpha.

And so there you go. I have my confidence interval. Except that's not, right? We just said that the bounds of a confidence interval may not depend on the unknown parameter. And here, they do.

And so we actually came up with two ways of getting rid of this. Since we only need this thing-- so this thing, as we said, is really equal. Every time I'm going to make this guy smaller and this guy larger, I'm only going to increase the probability. And so what we do is we actually just take the largest possible value for p1 minus p, which makes the interval as large as possible.

And so now I have this. I just do one of the two tricks. I replace p1 minus p by their upper bound, which is 1/4. As we said, p1 minus p, the function looks like this. So I just take the value here at 1/2.

Or, I can use Slutsky and say that if I replace p by Xn bar, that's the same as just replacing p by Xn bar here. And by Slutsky, we know that this is actually converging also to some standard Gaussian. We've seen that when we saw Slutsky as an example.

And so those two things-- actually, just because I'm taking the limit and I'm only caring about the asymptotic confidence level, I can actually just plug in consistent quantities in there, such as Xn bar where I don't have a p. And that gives me another confidence interval.

All right. So this by now, hopefully after doing it three times, you should really, really be comfortable with just creating this confidence interval. We did it three times in class. I think you probably did it another couple times in your homework. So just make sure you're comfortable with this.

All right. That's one of the basic things you would want to know. Are there any questions? Yes.

AUDIENCE: So Slutsky holds for any single response set p. But Xn converges [INAUDIBLE].

PHILIPPE RIGOLLET: So that's not Slutsky, right?

AUDIENCE: That's [INAUDIBLE].

PHILIPPE RIGOLLET: So Slutsky tells you that if you-- Slutsky's about combining two types of convergence. So Slutsky tells you that if you actually have one Xn that converges to X in distribution and Yn that converges to Y in probability, then you can actually multiply Xn and Yn and get that the limit in distribution is the product of X and Y, where X is now a constant.

And here we have the constant, which is 1. But I did that already, right? Using Slutsky to replace it for the-- to replace P by Xn bar, we've done that last time, maybe a couple of times ago, actually. Yeah.

AUDIENCE: So I guess these statements are [INAUDIBLE].

PHILIPPE RIGOLLET: That's correct.

AUDIENCE: So could we like figure out [INAUDIBLE] can we set a finite [INAUDIBLE].

PHILIPPE RIGOLLET: So of course, the short answer is no. So here's how you would go about thinking about which method is better. So there's always the more conservative method.

The first one, the only thing you're losing is the rate of convergence of the central limit theorem. So if n is large enough so that the central limit theorem approximation is very good, then that's all you're going to be losing. Of course, the price you pay is that your confidence interval is wider than it would be if you were to use Slutsky for this particular problem, typically wider. Actually, it is always wider, because Xn bar-- 1 minus Xn bar is always less than 1/4 as well. And so that's the first thing you-- so Slutsky basically adds your relying on the central limit-- your relying on the asymptotics again.

Now of course, you don't want to be conservative, because you actually want to squeeze as much from your data as you can. So it depends on how comfortable and how critical it is for you to put valid error bars. If they're valid in the asymptotics, then maybe you're actually going to go with Slutsky so it actually gives you slightly narrower confidence intervals and so you feel like you're a little more-- you have a more precise answer.

Now, if you really need to be super-conservative, then you're actually going to go with the P1 minus P. Actually, if you need to be even more conservative, you are going to go with Hoeffding's so you don't even have to rely on the asymptotics level at all. But then you're confidence interval becomes twice as wide and twice as wide and it becomes wider and wider as you go. So depends on-- I mean, there's a lot of data in statistics which is gauging how critical it is for you to output valid error bounds or if they're really just here to be indicative of the precision of the estimator you gave from a more qualitative perspective.

AUDIENCE: So the error there is [INAUDIBLE]?

PHILIPPE RIGOLLET: Yeah. So here, there's basically a bunch of errors. There's one that's-- so there's a theorem called Berry-Esseen that quantifies how far this probability is from 1 minus alpha, but the constants are terrible. So it's not very helpful, but it tells you as n grows how smaller this thing grows-- becomes smaller.

And then for Slutsky, again you're multiplying something that converges by something that fluctuates around 1, so you need to understand how this thing fluctuates. Now, there's something that shows up. Basically, what is the slope of the function 1 over square root of X1 minus X around the value you're interested in? And so if this function is super-sharp, then small fluctuations of Xn bar around this expectation are going to lead to really high fluctuations of the function itself.

So if you're looking at-- if you have f of Xn bar and f around say the true P, if f is really sharp like that, then if you move a little bit here, then you're going to move really a lot on the y-axis. So that's what the function here-- the function you're interested in is 1 over square root of X1 minus X. So what does this function look like around the point where you think P is the true parameter? Its derivative really is what matters.

OK? Any other question. OK. So it's important, because now we're going to switch to the real let's do some hardcore computation type of things. All right.

So in this chapter, we're going to talk about maximum likelihood estimation. Who has already seen maximum likelihood estimation? OK. And who knows what a convex function is? OK.

So we'll do a little bit of reminders on those things. So those things are when we do maximum likelihood estimation, likelihood is the function, so we need to maximize a function. That's basically what we need to do.

And if I give you a function, you need to know how to maximize this function. Sometimes, you have closed-form solutions. You can take the derivative and set it equal to 0 and solve it. But sometimes, you actually need to resort to algorithms to do that. And there's an entire industry doing that. And we'll briefly touch upon it, but this is definitely not the focus of this class.

OK. So before diving directly into the definition of the likelihood and what is the definition of the maximum likelihood estimator, what I'm going to try to do is to give you an insight for what we're actually doing when we do maximum likelihood estimation.

So remember, we have a model on a sample space E and some candidate distributions P theta. And really, your goal is to estimate a true theta star, the one that generated some data, X1 to Xn, in an iid fashion. But this theta star is really a proxy for us to know that we actually understand the distribution itself. The goal of knowing theta star is so that you can actually know what P theta star. Otherwise, it has-- well, sometimes we said it has some meaning itself, but really you want to know what the distribution is.

And so your goal is to actually come up with the distribution-- hopefully that comes from the family P theta-- that's close to P theta star. So in a way, what does it mean to have two distributions that are close? It means that when you compute probabilities on one distribution, you should have the same probability on the other distribution pretty much.

So what we can do is say, well, now I have two candidate distributions. So if theta hat leads to a candidate distribution P theta hat, and this is the true theta star, it leads to the true distribution P theta star according to which my data was drawn. That's my candidate. As a statistician, I'm supposed to come up with a good candidate, and this is the truth.

And what I want is that if you actually give me the distribution, then I want when I'm computing probabilities for this guy, I know what the probabilities for the other guys are. And so really what I want is that if I compute a probability under theta hat of some interval a, b, it should be pretty close to the probability under theta star of a, b. And more generally, if I want to take the union of two intervals, I want this to be true. If I take just 1/2 lines, I want this to be true from 0 to infinity, for example, things like this.

I want this to be true for all of them at once. And so what I do is that I write A for a probability event. And I want that P hat of A is close to P star of A for any event A in the sample space. Does that sound like a reasonable goal for a statistician? So in particular, if I want those to be close, I want the absolute value of their difference to be close to 0.

And this turns out to be-- if I want this to hold for all possible A's, I have all possible events, so I'm going to actually maximize over these events. And I'm going to look at the worst possible event on which theta hat can depart from theta star. And so rather than defining it specifically for theta hat and theta star, I'm just going to say, well, if you give me two probability measures, P theta and P theta prime, I want to know how close they are. Well, if I want to measure how close they are by how they can differ when I measure the probability of some event, I'm just looking at the absolute value of the difference of the probabilities and I'm just maximizing over the worst possible event that might actually make them differ. Agreed?

That's a pretty strong notion. So if the total variation between theta and theta prime is small, it means that for all possible A's that you give me, then P theta of A is going to be close to P theta prime of A, because if-- let's say I just found the bound on the total variation distance, which is 0.01. All right. So that means that this is going to be larger than the max over A of P theta minus P theta prime of A, which means that for any A-- actually, let me write P theta hat and P theta star, like we said, theta hat and theta star.

And so if I have a bound, say, on the total variation, which is 0.01, that means that P theta hat-- every time I compute a probability on P theta hat, it's basically in the interval P theta star of A, the one that I really wanted to compute, plus or minus 0.01. This has nothing to do with confidence interval. This is just telling me how far I am from the value of actually trying to compute.

And that's true for all A. And that's key. That's where this max comes into play. It just says, I want this bound to hold for all possible A's at once.

So this is actually a very well-known distance between probability measures. It's the total variation distance. It's extremely central to probabilistic analysis. And it essentially tells you that every time-- if two probability distributions are close, then it means that every time I compute a probability under P theta but I really actually have data from P theta prime, then the error is no larger than the total variation.

OK. So this is maybe not the most convenient way of finding a distance. I mean, how are you going-- in reality, how are you to compute this maximum over all possible events? I mean, it's just crazy, right? There's an infinite number of them. It's much larger than the number of intervals, for example, so it's a bit annoying.

And so there's actually a way to compress it by just looking at the basically function distance or vector distance between probability mass functions or probability density functions. So I'm going to start with the discrete version of the total variation. So throughout this chapter, I will make the difference between discrete random variables and continuous random variables. It really doesn't matter. All it means is that when I talk about discrete, I will talk about probability mass functions. And when I talk about continuous, I will talk about probability density functions.

When I talk about probability mass functions, I talk about sums. When I talk about probability density functions, I talk about integrals. But they're all the same thing, really.

So let's start with the probability mass function. Everybody remembers what the probability mass function of a discrete random variable is. This is the function that tells me for each possible value that it can take, the probability that it takes this value. So the Probability Mass Function, PMF, is just the function for all x in the sample space tells me the probability that my random variable is equal to this little value. And I will denote it by P sub theta of X.

So what I want is, of course, that the sum of the probabilities is 1. And I want them to be non-negative. Actually, typically we will assume that they are positive. Otherwise, we can just remove this x from the sample space.

And so then I have the total variation distance, I mean, it's supposed to be the maximum overall sets of-- of subsets of E, such that the probability of A minus probability of theta prime of A-- it's complicated, but really there's this beautiful formula that tells me that if I look at the total variation between P theta and P theta prime, it's actually equal to just 1/2 of the sum for all X in E of the absolute difference between P theta X and P theta prime of X.

So that's something you can compute. If I give you two probability mass functions, you can compute this immediately. But if I give you just the densities and the original distribution, the original definition where you have to max over all possible events, it's not clear you're going to be able to do that very quickly.

So this is really the one you can work with. But the other one is really telling you what it is doing for you. It's controlling the difference of probabilities you can compute on any event. But here, it's just telling you, well, if you do it for each simple event, it's little x. It's actually simple.

Now, if we have continuous random variables-- so by the way, I didn't mention, but discrete means Bernoulli. Binomial, but not only those that have finite support, like Bernoulli has support of size 2, binomial NP has support of size n-- there's n possible values it can take-- but also Poisson. Poisson distribution can take an infinite number of values, all the positive integers, non-negative integers.

And so now we have also the continuous ones, such as Gaussian, exponential. And what characterizes those guys is that they have a probability density. So the density, remember the way I use my density is when I want to compute the probability of belonging to some event A. The probability of X falling to some subset of the real line A is simply the integral of the density on this set. That's the famous area under the curve thing.

So since for each possible value, the probability at X-- so I hope you remember that stuff. That's just probably something that you must remember from probability. But essentially, we know that the probability that X is equal to little x is 0 for a continuous random variable, for all possible X's. There's just none of them that actually gets weight.

So what we have to do is to describe the fact that it's in some little region. So the probability that it's in some interval, say, a, b, this is the integral between A and B of f theta of X, dx. So I have this density, such as the Gaussian one. And the probability that I belong to the interval a, b is just the area under the curve between A and B. If you don't remember that, please take immediate remedy.

So this function f, just like P, is non-negative. And rather than summing to 1, it integrates to 1 when I integrate it over the entire sample space E. And now the total variation, well, it takes basically the same form. I said that you essentially replace sums by integrals when you're dealing with densities. And here, it's just saying, rather than having 1/2 of the sum of the absolute values, you have 1/2 of the integral of the absolute value of the difference.

Again, if I give you two densities and if you're not too bad at calculus, which you will often be, because there's lots of them you can actually not compute. But if I gave you, for example, two Gaussian densities, exponential minus x squared, blah, blah, blah, and I say, just compute the total variation distance, you could actually write it as an integral. Now, whether you can actually reduce this integral to some particular number is another story. But you could technically do it. So now, you have actually a handle on this thing and you could technically ask Mathematica, whereas asking Mathematica to take the max over all possible events is going to be difficult.

All right. So the total variation has some properties. So let's keep on the board the definition that involves, say, the densities. So think Gaussian in your mind. And you have two Gaussians, one with mean theta and one with mean theta prime. And I'm looking at the total variation between those two guys.

So if I look at P theta minus-- sorry. TV between P theta and P theta prime, this is equal to 1/2 of the integral between f theta, f theta prime. And when I don't write it-- so I don't write the X, dx but it's there. And then I integrate over E.

So what is this thing doing for me? It's just saying, well, if I have-- so think of two Gaussians. For example, I have one that's here and one that's here. So this is let's say f theta, f theta prime.

This guy is doing what? It's computing the absolute value of the difference between f and f theta prime. You can check for yourself that graphically, this I can represent as an area not under the curve, but between the curves. So this is this guy.

Now, this guy is really the integral of the absolute value. So this thing here, this area, this is 2 times the total variation. The scaling 1/2 really doesn't matter. It's just if I want to have an actual correspondence between the maximum and the other guy, I have to do this.

So this is what it looks like. So we have this definition. And so we have a couple of properties that come into this. The first one is that it's symmetric. TV of P theta and P theta prime is the same as the TV between P theta prime and P theta.

Well, that's pretty obvious from this definition. I just flip those two, I get the same number. It's actually also true if I take the maximum. Those things are completely symmetric in theta and theta prime. You can just flip them.

It's non-negative. Is that clear to everyone that this thing is non-negative? I integrate an absolute value, so this thing is going to give me some non-negative number. And so if I integrate this non-negative number, it's going to be a non-negative number. The fact also that it's an area tells me that it's going to be non-negative.

The nice thing is that if TV is equal to zero, then the two distributions, the two probabilities are the same. That means that for every A, P theta of A is equal to P theta prime of A. Now, there's two ways to see that.

The first one is to say that if this integral is equal to 0, that means that for almost all X, f theta is equal to f theta prime. The only way I can integrate a non-negative and get 0 is that it's 0 pretty much everywhere. And so what it means is that the two densities have to be the same pretty much everywhere, which means that the distributions are the same.

But this is not really the way you want to do this, because you have to understand what pretty much everywhere means-- which I should really say almost everywhere. That's the formal way of saying it. But let's go to this definition-- which is gone. Yeah. That's the one here.

The max of those two guys, if this maximum is equal to 0-- I have a maximum of non-negative numbers, their absolute values. Their maximum is equal to 0, well, they better be all equal to 0, because if one is not equal to 0, then the maximum is not equal to 0. So those two guys, for those two things to be-- for the maximum to be equal to 0, then each of the individual absolute values have to be equal to 0, which means that the probability here is equal to this probability here for every event A.

So those two things-- this is nice, right? That's called definiteness. The total variation equal to 0 implies that P theta is equal to P theta prime. So that's really some notion of distance, right? That's what we want. If this thing being small implied that P theta could be all over the place compared to P theta prime, that would not help very much.

Now, there's also the triangle inequality that follows immediately from the triangle inequality inside this guy. If I squeeze in some f theta prime prime in there, I'm going to use the triangle inequality and get the triangle inequality for the whole thing. Yeah?

AUDIENCE: The fact that you need two definitions of the [INAUDIBLE], is it something obvious or is it complete?

PHILIPPE RIGOLLET: I'll do it for you now. So let's just prove that those two things are actually giving me the same definition.

So what I'm going to do is I'm actually going to start with the second one. And I'm going to write-- I'm going to start with the density version. But as an exercise, you can do it for the PMF version if you prefer.

So I'm going to start with the fact that f-- so I'm going to write f of g so I don't have to write f and g. So think of this as being f sub theta, and think of this guy as being f sub theta prime. I just don't want to have to write indices all the time.

So I'm going to start with this thing, the integral of f of X minus g of X dx. The first thing I'm going to do is this is an absolute value, so either the number in the absolute value is positive and I actually kept it like that, or it's negative and I flipped its sign. So let's just split between those two cases.

So this thing is equal to 1/2 the integral of-- so let me actually write the set A star as being the set of X's such that f of X is larger than g of X. So that's the set on which the difference is going to be positive or the difference is going to be negative. So this, again, is equivalent to f of X minus g of X is positive.

OK. Everybody agrees? So this is the set I'm interested in. So now I'm going to split my integral into two parts, in A, A star, so on A star, f is larger than g, so the absolute value is just the difference itself. So here I put parenthesis rather than absolute value.

And then I have plus 1/2 of the integral on the complement. What are you guys used to to write the complement, to the C or the bar? To the C? And so here on the complement, then f is less than g, so this is actually really g of X minus f of X, dx.

Everybody's with me here? So I just said-- I mean, those are just rewriting what the definition of the absolute value is.

OK. So now there's nice things that I know about f and g. And the two nice things is that the integral of f is equal to 1 and the integral of g is equal to 1. This implies that the integral of f minus g is equal to what?

AUDIENCE: 0.

PHILIPPE RIGOLLET: 0. And so now that means that if I want to just go from the integral here on A complement to the integral on A-- or on A star, complement to the integral of A star, I just have to flip the sign. So that implies that an integral on A star complement of g of X minus f of X, dx, this is simply equal to the integral on A star of f of X minus g of X, dx.

All right. So now this guy becomes this guy over there. So I have 1/2 of this plus 1/2 of the same guy, so that means that 1/2 half of the integral between of f minus g absolute value-- so that was my original definition, this thing is actually equal to the integral on A star of f of X minus g of X, dx. And this is simply equal to P of A star-- so say Pf of A start minus Pg of A star.

Which one is larger than the other one?

AUDIENCE: [INAUDIBLE]

PHILIPPE RIGOLLET: It is. Just look at this board.

AUDIENCE: [INAUDIBLE]

PHILIPPE RIGOLLET: What?

AUDIENCE: [INAUDIBLE]

PHILIPPE RIGOLLET: The first one has to be larger, because this thing is actually equal to a non-negative number. So now I have this absolute value of two things, and so I'm closer to the actual definition. But I still need to show you that this thing is the maximum value. So this is definitely at most the maximum over A of Pf of A minus Pg of A.

That's certainly true. Right? We agree with this? Because this is just for one specific A, and I'm bounding it by the maximum over all possible A. So that's clearly true.

So now I have to go the other way around. I have to show you that the max is actually this guy, A star. So why would that be true?

Well, let's just inspect this thing over there. So we want to show that if I take any other A in this integral than this guy A star, it's actually got to decrease its value. So we have this function. I'm going to call this function delta.

And what we have is-- so let's say this function looks like this. Now it's the difference between two densities. It doesn't have to integrate-- it doesn't have to be non-negative. But it certainly has to integrate to 0.

And so now I take this thing. And the A star, what is the set A star here? The set A star is the set over which the function delta is non-negative. So that's just the definition. A star was the set over which f minus g was positive, and f minus g was just called delta.

So what it means is that what I'm really integrating is delta on this set. So it's this area under the curve, just on the positive things. Agreed?

So now let's just make some tiny variations around this guy. If I take A to be larger than A star-- so let me add, for example, this part here. That means that when I compute my integral, I'm removing this area under the curve. It's negative. The integral here is negative. So if I start adding something to A, the value goes lower. If I start removing something from A, like say this guy, I'm actually removing this value from the integral.

So there's no way. I'm actually stuck. This A star is the one that actually maximizes the integral of this function. So we used the fact that for any function, say delta, the integral over A of delta is less than the integral over the set of X's such that delta of X is non-negative of delta of X, dx. And that's an obvious fact, just by picture, say. And that's true for all A. Yeah?

AUDIENCE: [INAUDIBLE] could you use like a portion under the axis as like less than or equal to the portion above the axis?

PHILIPPE RIGOLLET: It's actually equal. We know that the integral of f minus g-- the integral of delta is 0. So there's actually exactly the same area above and below. But yeah, you're right. You could go to the extreme cases. You're right.

No. It's actually still be true, even if there was-- if this was a constant, that would still be true. Here, I never use the fact that the integral is equal to 0.

I could shift this function by 1 so that the integral of delta is equal to 1, and it would still be true that it's maximized when I take A to be the set where it's positive. Just need to make sure that there is someplace where it is, but that's about it. Of course, we used this before, when we made this thing. But just the last argument, this last fact does not require that.

All right. So now we have this notion of-- I need the--

OK. So we have this notion of distance between probability measures. I mean, these things are exactly what-- if I were to be in a formal math class and I said, here are the axioms that a distance should satisfy, those are exactly those things. If it's not satisfying this thing, it's called pseudo-distance or quasi-distance or just metric or nothing at all, honestly. So it's a distance. It's symmetric, non-negative, equal to 0, if and only if the two arguments are equal, then it satisfies the triangle inequality.

And so that means that we have this actual total variation distance between probability distributions. And here is now a statistical strategy to implement our goal. Remember, our goal was to spit out a theta hat, which was close such that P theta hat was close to P theta star. So hopefully, we were trying to minimize the total variation distance between P theta hat and P theta star.

Now, we cannot do that, because just by this fact, this slide, if we wanted to do that directly, we would just take-- well, let's take theta hat equals theta star and that will give me the value 0. And that's the minimum possible value we can take.

The problem is that we don't know what the total variation is to something that we don't know. We know how to compute total variations if I give you the two arguments. But here, one of the arguments is not known. P theta star is not known to us, so we need to estimate it.

And so here is the strategy. Just build an estimator of the total variation distance between P theta and P theta star for all candidate theta, all possible theta in capital theta. Now, if this is a good estimate, then when I minimize it, I should get something that's close to P theta star.

So here's the strategy. This is my function that maps theta to the total variation between P theta and P theta star. I know it's minimized at theta star. That's definitely TV of P-- and the value here, the y-axis should say 0.

And so I don't know this guy, so I'm going to estimate it by some estimator that comes from my data. Hopefully, the more data I have, the better this estimator is. And I'm going to try to minimize this estimator now. And if the two things are close, then the minima should be close.

That's a pretty good estimation strategy. The problem is that it's very unclear how you would build this estimator of TV, of the Total Variation. So building estimators, as I said, typically consists in replacing expectations by averages. But there's no simple way of expressing the total variation distance as the expectations with respect to theta star of anything.

So what we're going to do is we're going to move from total variation distance to another notion of distance that sort of has the same properties and the same feeling and the same motivations as the total variation distance. But for this guy, we will be able to build an estimate for it, because it's actually going to be of the form expectation of something. And we're going to be able to replace the expectation by an average and then minimize this average.

So this surrogate for total variation distance is actually called the Kullback-Leibler divergence. And why we call it divergence is because it's actually not a distance. It's not going to be symmetric to start with.

So this Kullback-Leibler or even KL divergence-- I will just refer to it as KL-- is actually just more convenient. But it has some roots coming from information theory, which I will not delve into. But if any of you is actually a Core 6 student, I'm sure you've seen that in some-- I don't know-- course that has any content on information theory.

All right. So the KL divergence between two probability measures, P theta and P theta prime-- and here, as I said, it's not going to be the symmetric, so it's very important for you to specify which order you say it is, between P theta and P theta prime. It's different from saying between P theta prime and P theta.

And so we denote it by KL. And so remember, before we had either the sum or the integral of 1/2 of the distance-- absolute value of the distance between the PMFs and 1/2 of the absolute values of the distances between the probability density functions. And then we replace this absolute value of the distance divided by 2 by this weird function.

This function is P theta, log P theta, divided by P theta prime. That's the function. That's a weird function.

OK. So this was what we had. That's the TV. And the KL, if I use the same notation, f and g, is integral of f of X, log of f of X over g of X, dx. It's a bit different. And I go from discrete to continuous using an integral.

Everybody can read this. Everybody's fine with this. Is there any uncertainty about the actual definition here?

So here I go straight to the definition, which is just plugging the functions into some integral and compute. So I don't bother with maxima or anything. I mean, there is something like that, but it's certainly not as natural as the total variation. Yes?

AUDIENCE: The total variation, [INAUDIBLE].

PHILIPPE RIGOLLET: Yes, just because it's hard to build anything from total variation, because I don't know it. So it's very difficult. But if you can actually-- and even computing it between two Gaussians, just try it for yourself. And please stop doing it after at most six minutes, because you won't be able to do it. And so it's just very hard to manipulate, like this integral of absolute values of differences between probability density function, at least for the probability density functions we're used to manipulate is actually a nightmare.

And so people prefer KL, because for the Gaussian, this is going to be theta minus theta prime squared. And then we're going to be happy. And so those things are much easier to manipulate.

But it's really-- the total variation is telling you how far in the worst case the two probabilities can be. This is really the intrinsic notion of closeness between probabilities. So that's really the one-- if we could, that's the one we would go after.

Sometimes people will compute them numerically, so that they can say, oh, here's the total variation distance I have between those two things. And then you actually know that that means they are close, because the absolute value-- if I tell you total variation is 0.01, like we did here, it has a very specific meaning. If I tell you the KL divergence is 0.01, it's not clear what it means.

OK. So what are the properties? The KL divergence between P theta and P theta prime is different from the KL divergence between P theta prime and P theta in general. Of course, in general, because if theta is equal to theta prime, then this certainly is true. So there's cases when it's not true.

The KL divergence is non-negative. Who knows the Jensen's inequality here? That should be a subset of the people who raised their hand when I asked what a convex function is. All right.

So you know what Jensen's inequality is. This is Jensen's-- the proof is just one step Jensen's inequality, which we will not go into details. But that's basically an inequality involving expectation of a convex function of a random variable compared to the convex function of the expectation of a random variable.

If you know Jensen, have fun and prove it. What's really nice is that if the KL is equal to 0, then the two distributions are the same. And that's something we're looking for.

Everything else we're happy to throw out. And actually, if you pay attention, we're actually really throwing out everything else. So they're not symmetric. It does satisfy the triangle inequality in general.

But it's non-negative and it's 0 if and only if the two distributions are the same. And that's all we care about. And that's what we call a divergence rather than a distance, and divergence will be enough for our purposes.

And actually, this asymmetry, the fact that it's not flipping-- the first time I saw it, I was just annoyed. I was like, can we just like, I don't know, take the average of the KL between P theta and P theta prime and P theta prime and P theta, you would think maybe you could do this. You just symmatrize it by just taking the average of the two possible values it can take.

The problem is that this will still not satisfy the triangle inequality. And there's no way basically to turn it into something that is a distance. But the divergence is doing a pretty good thing for us. And this is what will allow us to estimate it and basically overcome what we could not do with the total variation.

So the first thing that you want to notice is the total variation distance-- the KL divergence, sorry, is actually an expectation of something. Look at what it is here. It's the integral of some function against a density. That's exactly the definition of an expectation, right?

So this is the expectation of this particular function with respect to this density f. So in particular, if I call this is density f-- if I say, I want the true distribution to be the first argument, this is an expectation with respect to the true distribution from which my data is actually drawn of the log of this ratio.

So ha ha. I'm a statistician. Now I have an expectation. I can replace it by an average, because I have data from this distribution. And I could actually replace the expectation by an average and try to minimize here.

The problem is that-- actually the star here should be in front of the theta, not of the P, right? That's P theta star, not P star theta.

But here, I still cannot compute it, because I have this P theta star that shows up. I don't know what it is. And that's now where the log plays a role.

If you actually pay attention, I said you can use Jensen to prove all this stuff. You could actually replace the log by any concave function. That would be f divergent. That's called an f divergence. But the log itself is a very, very specific property, which allows us to say that the log of the ratio is the ratio of the log.

Now, this thing here does not depend on theta. If I think of this KL divergence as a function of theta, then the first part is actually a constant. If I change theta, this thing is never going to change. It depends only on theta star.

So if I look at this function KL-- so if I look at the function, theta maps to KL P theta star, P theta, it's of the form expectation with respect to theta star, log of P theta star of X. And then I have minus expectation with respect to theta star of log of P theta of x.

Now as I said, this thing here, this second expectation is a function of theta. When theta changes, this thing is going to change. And that's a good thing. We want something that reflects how close theta and theta star are.

But this thing is not going to change. This is a fixed value. Actually, it's the negative entropy of P theta star. And if you've heard of KL, you've probably heard of entropy. And that's what-- it's basically minus the entropy.

And that's a quantity that just depends on theta star. But it's just the number. I could compute this number if I told you this is n theta star 1. You could compute this.

So now I'm going to try to minimize the estimate of this function. And minimizing a function or a function plus a constant is the same thing. I'm just shifting the function here or here, but it's the same minimizer.

OK. So the function that maps theta to KL of P theta star to P theta is of the form constant minus this expectation of a log of P theta. Everybody agrees? Are there any questions about this? Are there any remarks, including I have no idea what's happening right now? OK. We're good? Yeah.

AUDIENCE: So when you're actually employing this method, how do you know which theta to use as theta star and which isn't?

PHILIPPE RIGOLLET: So this is not a method just yet, right? I'm just describing to you what the KL divergence between two distributions is. If you really wanted to compute it, you would need to know what P theta star is and what P theta is.

AUDIENCE: Right.

PHILIPPE RIGOLLET: And so here, I'm just saying at some point, we still-- so here, you see-- so now let's move onto one step. I don't know expectation of theta star. But I have data that comes from distribution P theta star.

So the expectation by the law of large numbers should be close to the average. And so what I'm doing is I'm replacing any-- I can actually-- this is a very standard estimation method. You write something as an expectation with respect to the data-generating process of some function. And then you replace this by the average of this function. And the law of large numbers tells me that those two quantities should actually be close.

Now, it doesn't mean that's going to be the end of the day, right. When we did Xn bar, that was the end of the day. We had an expectation. We replaced it by an average. And then we were gone.

But here, we still have to do something, because this is not telling me what theta is. Now I still have to minimize this average. So this is now my candidate estimator for KL, KL hat.

And that's the one where I said, well, it's going to be of the form of constant. And this constant, I don't know. You're right. I have no idea what this constant is. It depends on P theta star.

But then I have minus something that I can completely compute. If you give me data and theta, I can compute this entire thing. And now what I claim is that the minimizer of f or f plus-- f of X or f of X plus 4 are the same thing, or say 4 plus f of X. I'm just shifting the plot of my function up and down, but the minimizer stays exactly where it is.

If I have a function-- so now I have a function of theta. This is KL hat of P theta star, P theta. And it's of the form-- it's a function like this. I don't know where this function is. It might very well be this function or this function.

Every time it's a translation on the y-axis of all these guys. And the value that I translated by depends on theta star. I don't know what it is. But what I claim is that the minimizer is always this guy, regardless of what the value is. OK?

So when I say constant, it's a constant with respect to theta. It's an unknown constant. But it's with respect to theta, so without loss of generality, I can assume that this constant is 0 for my purposes, or 25 if you prefer.

All right. So we'll just keep going on this property next time. And we'll see how from here we can move on to-- the likelihood is actually going to come out of this formula. Thanks.