MITOCW | watch?v=TSkDZbGS94k

PHILIPPE RIGOLLET: Doesn't to run Flash Player. So I had to run them on Chrome.

All right, so let's move on to our second chapter. And hopefully, in this chapter, you will feel a little better if you felt like it was going a bit fast in the first chapter. And the main reason we actually went fast, especially in terms of confidence intervals. Some of you came and asked me what you mean by this is a confidence interval? What does it mean that it's not happening in there with probability 95%, et cetera?

I just went really fast so that you could see why I didn't want to give you a first week doing probability only without understanding what the statistical context for it was. So hopefully, all these things that we've done in terms of probability, you actually know why we've been doing them. And so we're basically going to go back to what we're doing, maybe start with some statistical setup. But the goal of this lecture is really going to go back again to what we've seen from a purely statistical perspective. All right?

So the first thing we're going to do is explain why we're doing statistical modeling, right? So in practice, if you have data, if you observe a bunch of points-- in here, I gave you some numbers, for example. So here's the partial data sets with the number of siblings, including self, that were collected from college students a few years back.

So I was teaching a class like yours and actually asked students to go and fill out some Google form and tell me a bunch of things. And one of the questions was, including yourself, how many siblings do you have? And so they gave me this list of numbers, right? And there's many ways I can think of this list of numbers, right? I could think of it as being just a discrete distribution on the set of numbers between 1-- I know there's not going to be an answer which is less than 1, unless, well, someone doesn't understand the question.

But all the answers I should get are positive integers-- 1, 2, 3, et cetera. And there probably is an upper bound, but I don't know it on the top of my head. So maybe I should say 100. Maybe I should say 15. It depends, right? And so I think the largest number I got for this was 6.

All right? So here you can see you have pretty standard families, you know, lots of 1s, 2s, and 3s. What statistical modeling is doing is to try to compress this information that I could actually describe in a very naive way. So let's start with the basic usual statistical set up, right? So I will start with many of the boards that look like x1, xn, random variables.

And what I'm going to assume, as we said typically is that those guys are IID. And they have some distribution, all right? So they all share the same distribution. And the fact that their IID is so that I can actually do statistics. Statistics means looking at the global averaging thing so that I can actually get a sense of what the global behavior is for the population, right?

If I start assuming that those things are not identically distributed-- they all live on their own-- that my sequence of number is your number of siblings-- the shoe size of this person-- the depth of the Charles River, and I start measuring a bunch of stuff. There's nothing I can actually get together. I need to have something that's cohesive. And so here, I collected some data that was cohesive.

And so the goal here-- the first thing is to say what is the distribution that I actually have here, right? So I could actually be very general. I could just say at some distribution p. And let's so those are random variables, not random vectors, right? I could collect entire vectors about students, but let's say those are just random variables.

And so now I can start making assumptions on this distribution p, right? What can I say about a distribution? Well, maybe if those numbers are continues, for example, I could assume they have a density-- a probability density function. That's already an assumption. Maybe I could start to assume that they're probability density function is smooth. That's another assumption.

Maybe I could actually assume that it's piecewise constant. That's even better, right? And those things make my life simpler and simpler, because what I do by making the successive assumptions is reducing the degrees of freedom of the space in which I am actually searching the distribution. And so what we actually want is to have something which is small enough so we can actually have some averaging going on. But we also want something which is big enough that it can actually express. It has chances of actually containing a distribution that makes sense for us.

So let's start with the simplest possible example, which is when the xi's belong to 0, 1. And as I said, here, we don't have a choice. The distribution of those guys has to be Bernoulli. And since they are IID, they all share the same p. So that's definitely the simplest possible thing I could think of. They are just Bernoulli p.

And so all I would have to figure out in this case is p. And this is the simplest case. And unsurprisingly, it has the simplest answer, right? We will come back to this example when we study maximum likelihood estimators or method of moments estimators by method of moments. But at the end of the day, the things that we did-- the things that we will do are always the naive estimator you would come up with is what is the proportion of 1. And this will be, in pretty much all respects, the best estimator you can think of.

All right? So then we're going to try to assess this performance. And we saw how to do that in the first chapter as well. So this problem here somehow is completely understood. We'll come back to it. But there's nothing fancy that is going to happen. But now, I could have some more complicated things. For example, in the example of the students now, my xi's belong to the sequence of integers 1, 2, 3, et cetera, OK, which is also denoted by n, maybe without 0 if you want to put 0 in that, right? So the positive integers.

Or I could actually just maybe put some prior knowledge about how humans have time to have families. But maybe some people thought of their college mates as being their brothers and sisters. And one student would actually put 465 siblings, because we're all good friends. Or maybe they actually think that all their Facebook contacts are actually their siblings.

And so you never know what's going to happen. So maybe you want to account for this, but maybe you know that people are reasonable, and they will actually give you something like this. Now intuitively, maybe you would say, well, why would you bother doing this if you're not really sure about the 20? But I think that probably all of you intuitively guess that this is probably a good idea to start putting this kind of assumption rather than allowing for any number in the first place, because this eventually will be injected into the precision of our estimators. If I allow anything, it's going to be more complicated for me to get an accurate estimator.

If I know that the numbers are either 1 or 2, then I'm actually going to be slightly more accurate as well. All right? Because I know that, for example, somebody put the 5, I can remove it. Then it's not going to actually corrupt with my estimator. All right, so now, let's say we actually agree that we have numbers. And here I put seven numbers, OK? So I just said, well, let's assume that the numbers I'm going to get are going to be 1 all the way to say this number that I denote by larger than or equal to 7, which is a placeholder for any numbers larger than 7, OK? Because I know maybe I don't want to distinguish between people that have 9 or 25 siblings.

OK, and so now, this is a distribution on seven possible values-- the discrete distributions. And you know from your probability class that the way you describe this distribution is using the probability mass function. OK, or PMF-- So that's how we describe a discrete distribution. And the PMF is just a list of numbers, right?

So as I wrote here, you have a list of numbers. And here, you wrote the possible value that your random variable can take. And here you rate the probability that your random variable takes this value. So the possible values being 1, 2, 3 all the way to larger than or equal to 7. And then I'm trying to estimate those numbers. Right?

If I give you those numbers, at least up to this you know compression of all numbers that are equal to 7, you have the full description of your distribution. And that is the ultimate goal of statistics, right? The ultimate goal of statistics is to say what distribution your data came from, because that's basically the best you're going to be able to do.

Now admittedly, if I started looking at the fraction of 1s, and the fraction of 2s, and the fraction of 3s, et cetera, I would actually eventually get those numbers-- just like looking at the fraction of 1s gave me a good estimate for p in the Bernoulli case, it would do the same in this case, right? It's a pretty intuitive idea. It's just the law of large numbers. Everybody agrees with that? If I look at the proportion of 1s, the proportion of 2s, the proportion of 3s, that should actually give me something that gets closer and closer, as my sample size increases to what I want.

But the problem is if my sample size is not huge, here I have seven numbers to estimate. And if I have 20 observations, the ratio is not really in my favor-- 20 observations to estimate seven parameters-- some of them are going to be pretty off, typically the ones with the large values. If you have only 20 students, look at the list of numbers. I don't know how many numbers I have, but it probably is close to 20-- maybe 15 or something.

And so if you look at this list, nobody's actually-- nobody has four or more siblings, right? There's no such person. So that means that eventually from this data set, my estimates-- so those numbers I denote by say p1, p2, p3, et cetera-- those estimates p4 hat would be equal to what from this data? 0, right? And p5 hat equal to 0 and p6 hat would be equal to 0. And p larger than or equal to 7 hat would be equal to 0.

That would be my estimate from this data set. So maybe this is not-- maybe I want to actually pull some information from the people who have less siblings to try to make a guess, which is probably slightly better for the larger values, right? It's pretty clear that in average, there is more than 0-- the proportion of the population of households that have four children or more is definitely more than 0, all right?

So it means that my data set is not representative of what I'm going to try to do is to find a model that tries to use the data they have for the smaller values that I can observe and just push it up to the other ones. And so what we can do is to just reduce those parameters into something that's understood. And this is part of the modeling that I talked about in the first place.

Now, how do you succinctly describe a number of something? Well, one thing that you do is the Poisson distribution, right? Why do Poisson? There's many reasons. Again, that's part of statical modeling. But once you know that you have number of something that can be modeled by a Poisson, why not try a Poisson, right? You could just fit a Poisson. And the Poisson is something that looks like this. And I guess you've all seen it.

But if x follows a Poisson distribution with parameter lambda, than the probability that x is equal to little x is equal to lambda to the x over factorial x e to the minus lambda. OK? And if you did the sheet that I gave you on the first day, you can check those numbers. So this is, of course, for x equals 0, 1, et cetera, right? So x is in natural integers.

And if you sum from x equals 0 to infinity, this thing you get is e to the lambda. And so they cancel, and you have some which is equal to 1, which is indeed a PMF. But what's key about this PMF is that it never takes value 0. Like this thing is always strictly positive. So whatever value of lambda I find from this data will give me something that's certainly more interesting than just putting the value 0.

But more importantly, rather than having to estimate seven parameters and, as a consequence, to actually have to estimate 1, 2, 3, 4 of them being equal to 0, I have only one parameter to estimate which is lambda. The problem with doing this is that now lambda may not be just something as simple as computing the average number. Right?

In this case, it will. But in many instances, it's actually not clear that this parametrization with lambda that I chose-- I'm going to be able to estimate lambda just by computing the average number that I get. It will be the case. But if it's not, remember this example of the exponential we did in the last lecture-- we could use the delta method and things like that to estimate them.

All right, so here's modeling 101. So the purpose of modeling is to restrict the base of possible distributions to a subspace that's actually plausible, but much simpler for me to estimate. So we went from all distributions on seven parameters, which is a large space-- that's a lot of things-- to something which is just one number. This number is positive. Any question about the purpose of doing this?

OK, so we're going to have to do a little bit of formalism now. And so if we want to talk-- this is a statistics classroom. I'm not going to want to talk about the Poisson model specifically every single time. I'm going to want to talk about generic models. And then you're going to build to plug in your favorite word-- Poisson, binomial, exponential, uniform-- all these words that you've seen, you're going to be able to plug in there. But we're just going to have some generic notations and some generic terminology for a statistical model.

All right? So here is the formal definition. So I'm going to go through it with you. OK, so the definition is that of a statistical model. OK? Sorry, that's a statistical experiment, I should say.

So a statistical experiment is actually just a pair-- E. And that's a set-- and a family of distributions P theta, where theta ranges in some set capital theta. OK? So I hope you're up to date with your Greek letters. So the small theta is the capital theta. And enough of us-- I don't have the handwriting. So if you don't see something, just ask me.

And so this thing now-- so each of this guy is a probability distribution. All right? So for example, this could be a Poisson with parameter theta or a Bernoulli with parameter theta-- OK, or an exponential with parameter-- I don't know-- 1 over theta squared if you want.

OK, but they're just indexed by theta. But for each theta, this completely describes the distribution. It could be more complicated. This theta should be a pair-- could be a pair-- a mu sigma square. And that could actually give you some n mu sigma squared.

OK so anything where you can actually-- rather than actually giving you a full distribution, I can compress into a parameter. But it could be worse. It could be this guy here. Right? Theta could be p1-- p larger than or equal to 7. And my distribution could just be something that has PMF-- p1-- p larger than 7. That's another parameter. This one is seven dimensional. This one is two dimensional. And all these guys are just one dimensional. All these guys are parameters. Is that clear?

What's important here is that once they give you theta, you know exactly all the probabilities associated with this random variable. You know its distribution perfectly. So this is the definition. Is that clear? Is there a question about this distribution-- about this definition, sorry?

All right. So really, the key thing is the statistical model associated to a statistical experiments. OK? So let's just see some examples. It's probably just better because, again, the formalism is never really clear. Actually, that's the next slide.

OK, so there's two things we need to assume. OK, so the purpose of a statistical model is once I estimate the parameter, I actually know exactly what distribution it has, OK? So it means that I could potentially have several parameters that give me the same distribution that would still be fine, because I could estimate one guy. Or I could estimate the other guy. And I would still recover the underlying distribution of my data.

The problem is that this creates really annoying theoretical problems, like things don't work, the algorithms won't work, the guarantees won't work. And so what we typically assume is that the model is so-called well-specified. Sorry, that's not well specified. I'm jumping ahead of myself. OK, well-specified means that your data-- the distribution of your data is actually one of those guys. OK?

So some vocabulary-- so well-specified means that for my observations x, there exists a theta in capital theta such that x follows p sub theta. I should put a double bar. OK, so that's what well-specified means.

So that means that the distribution of your actual data is just one of those guys. This is a bit strong of an assumption. It's strong in the sense that-- I don't know if you've heard of this sense, which I don't know, I can tell you who it's attributed to, but that probably means that this person did not come up with it. But I said that all models are wrong, but some of them are useful.

All right, so all models are wrong means that maybe it's not true that this Poisson distribution that I assume for the number of siblings for college students-- maybe that's not perfectly correct. Maybe there's a spike at three, right? Maybe there's a spike at one, because you know, maybe those are slightly more educated families. They have less children. Maybe this is actually not exactly perfect.

But it's probably good enough for our purposes. And when we make this assumption, we're actually assuming that the data really comes from a Poisson model. There is a lot of research that goes on about misspecified models and that tells you how well you're doing in the model that's the closest to the actual distribution. So that's pretty much it. Yeah?

AUDIENCE: [INAUDIBLE].

PHILIPPE RIGOLLET: So my data-- so it's always the way I denote one of the generic observations, right? So my observations are x1, xn. And they're IID with distribution p-- always. So x is just one of those guys. I don't want to write x5 or x4. They're IID. So they all have the same distribution.

So OK-- no, no, no. They're all IID. So they all have the same p data. They'll have the same p, which means they'll have the same p data. So I can pick any one of them. So I'd just remove the index just so we're clear. OK? So when I write x, I just mean think of x1. Right they're an idea. I can pick whichever I want. I'm not going to write x1. It's going to be weird. OK? Is that clear? OK.

So this particular theta is called the true parameter. Sometimes since we're going to want some variable theta, we might denote it by theta star as opposed to theta hat, which is always our estimator. But I'll keep it to be theta for now.

And so the aim of this physical experiment is to estimate theta so that once I actually plug in theta in the form of my distribution, for example, I could plug in theta here. So theta here was actually lambda. So once I estimate this guy, I would plug it in, and I would know the probability that my random variable takes any value, by just putting the lambda hat and the lambda hat here. OK?

So my goal is going to be to estimate this guy so that I can actually compute those distributions. But actually, we'll see, for example, when we talk about regression that this parameter actually has a meaning in many instances. And so just knowing the parameter itself intuitively or say more-- let's say more so than just computing probabilities, will actually tell us something about the process.

For example, we're going to run linear regression. And when we do linear regression, there's going to be some coefficients in the linear regression. And the value of this coefficient is actually telling me what is the sensitivity of the response that I'm looking at to this particular input. All right? So just knowing if this number is larger or if this number is small is actually going to be useful for us to just look at this guy.

All right? So there's going to be some instances where it's going to be important. Sometimes we're going to want to know if this parameter is larger or smaller than something or if it's equal to something or not equal to something. And those things are also important-- for example, if theta actually measures the true-- right? So theta is the true unknown parameter-- true efficacy of a drug.

OK? Let's say I want to know what the true efficacy of a drug is. And what I'm going to want to know is maybe it's a score. Maybe I'm going to want to know if theta is larger than 2. Maybe I want to know if theta is the average number of siblings. Is this true number larger than 2 or not? Right? Maybe I am interested in knowing if college students come from-- so maybe from a sociological perspective, I'm interested in knowing if college students come from households with more than two children.

All right, so those can be the questions that I may ask myself. I'm going to want to know maybe theta is going to be equal to 1/2 or not. So maybe for a drug efficacy, is it completely standard-- maybe for elections. Is the proportion of the population that is going to vote for this particular candidate equal to 0.5? Or is it different from 0.5?

OK, and I can think of different things. When I'm talking about the regression, I'm going to want to test if this coefficient is actually 0 or not, because if it's 0, it means that the variable that's in front of it actually goes out. And so those are things we're testing. Actually having this very specific yes/no answer is going to give me a huge intuition or huge understanding of what's going on in the phenomenon that I observe. But actually, since the questions are so precise, it's going to be much more-- I'm going to be much better at answering them rather than giving you an estimate for theta with some confidence around it.

All right, it's sort of the same principle as trying to reduce. What you're trying to do as a statistician is to inject as much knowledge about the question and about the problem that you can so that the data has to do a minimal job. And henceforth, you actually need less data. So from now on, we will always assume-- and this is because this is an intro stats class-- you will always assume that theta-- the subset of parameters is a subset of r to the d.

That means that theta is a vector with at most a finite number of coordinates. Why do I say this? Well, this is called a parametric model. So it's called a parametric model or sometimes parametric statistics. Actually, we don't really talk about parametric statistics. But we talk a lot about nonparametric statistics or a non-parametric model. Can somebody think of a model which is non-parametric?

For example, in the siblings example, if I did not cap the number of siblings to 7, but I let this list go to infinity, I would have an infinite number of parameters to estimate. Very likely, the last ones would be 0. But still, I would have an infinite number of parameters to estimate. So this would not be a parametric model if I just let this list of things to be estimated to be infinite.

But there's other classes that are actually infinite and cannot represented by vectors. For example, function-- right? If I tell you my model, pf, is just the distribution of x's, the probability distributions, that have density f, right? So what I know is that the density is non-negative and that it integrates to one, right? That's all I know about densities.

Well f is not something you're going to be able to describe with a finite number of values, right? All possible functions is the huge set. It's certainly not representable by 10 numbers. And so non-parametric estimation is typically when you actually want to parametrize this by a large class of functions. And so for example, histograms is the prime tool of non-parametric estimation, because when you fit a histogram to data, you're trying to estimate the density of your data, but you're not trying to represent it as a finite number of points. That's really-- I mean, effectively, you have to represent it, right? So you actually truncate somewhere and just say those things are not going to matter. All right?

But really the key thing is that this is non-parametric where you have a potentially infinite number of parameters. Whereas we're going to only talk about finites. And actually finite in the overwhelming majority of cases is going to be 1. So theta is going to be a subset of r1. OK, we're going to be interested in estimating one parameter just like the parameter of a Poisson or the parameter of an exponential-- the parameter of Bernoulli. But for example, really, we're going to be interested in estimating mu and sigma square for the normal.

So here are some statistical models. All right? So I'm going to go through them with you. So if I tell you I observe-- I'm interested in understanding-- I'm still [INAUDIBLE] I'm interested in understanding the proportion of people who kiss by bending their head to the right. And for that, I collected n observations. And I'm interested in making some inference in the statistical model. My question to you is, what is the statistical model?

Well, if you want to read the statistical model, you're going to have to write this E-- oh, sorry, I never told you what E was. OK, well actually just go to the examples, and then you'll know what E is. So you're going to have to write to me an E and a p theta, OK? So let's start with the

Bernoulli trials.

So this e here is called the sample space. And in the normal people's words, it just means the space or the set in which x and-- back to your question, x is just a generic observation lips. OK, and hopefully, this is the smallest you can think of. OK, so for example, for Bernoulli trials, I'm going to observe a sequence of 0's and 1's. So my experiment is going to be-- as written on the board, is going to be 1, 0, 1. And then the probability distributions are going to be, well, it's just going to be the Bernoulli distributions indexed by p, right?

So rather than writing p sub p, I'm going to write it as Bernoulli p, because it's clear what I mean when I write that. Is everybody happy? Actually, I need to tell you something more. This is a family of distributions. So I need p. And maybe I don't want to have to p that's a value 0, 1, right? It doesn't make sense. I would probably not look at this problem if I anticipated that everybody would kiss to the right. And everybody would kiss to the left.

So I am going to assume that p is in 0, 1, but does not have 0 and 1. OK? So that's the statistical model for a Bernoulli trial.

OK, now the next one, what do we have? Exponential. OK? OK, so when I have exponential distributions, what is the support of the exponential distribution? What value is it going to take? 0 to infinity, right? So what I have is that my first space is the value that my random variables can take. So its-- well, actually I can remove the 0 again-- 0 to plus infinity.

And then the family of distributions that I have are exponential with parameter lambda. And again, maybe you've seen me switching from p, to lambda, to theta, to mu, to sigma square. Honestly you can do whatever you want. But its just that it's customary to have this particular group of letters. OK?

And so the parameters of an exponential are just positive numbers. OK? And that's my exponential model. What is the third one? Can somebody tell me? Poisson, OK? OK, so Poisson-- is a Poisson random verbal discrete or continuous? Go back to your probability.

All right, so the answer being the opposite of continuous-- good job. All right, so it's going to be-- what value can a Poisson take? All the natural integers, right? So 0, 1, 2, 3, all the way to infinity. We don't have any control of this. So I'm going to write this as n without 0. I think in the slides, it's n-star maybe. Actually, no, you can take value 0. I'm sorry. This actually takes value 0 quite a lot. That's typically, in many instances, actually the mode.

So it's n, and then I'm going to write it as Poisson with parameter-- well, here it's again lambda as a parameter. And lambda can take any positive value. OK? And that's where you can actually see that the model that we had for the siblings-- right? So let me actually just squeeze in the siblings model here.

So that was the bad model that I had in the first place when I actually kept this. Let's say we just kept it at 7. Forget about larger than or equal to 7. We just assumed it was 7. What was our sample space? We said 7. So it's 1, 2, to 7, right? Those were the possible values that this thing would take.

And then what was my-- what's my parameter space? So it's going to be a nightmare write. But I'm going to write it. OK, so I'm going to write it as something like the probability that x is equal to k is equal to p sub k. OK? And that's going to be for p. OK, so that's for all k's, right? Or for k equal 1 to 7. And here the index is the set of parameters p1 to pk.

And I know a little more about those guys, right? I know there are going to be non-negative-- PJ non-negative. And I know that they sum to 1. OK, so maybe writing this, you start seeing why we like those Poisson, exponential, and short notation, because I actually don't have to write the PMF of a Poisson. The Poisson is really just this. But I call it Poisson so I don't have to rewrite this all the time.

And so here, I did not use a particular form. So I just have this thing, and that's what it is. The set of parameters is the set of positive numbers of-- p1 to p7, pk-- and sum to 7, right? And so this as just a list of numbers that are non-negative and sum up to 1. So that's my parameter space.

OK? So here that's my theta. This whole thing here-- this is my capital theta. OK? So that's just the set of parameters that theta-- the set of parameters that theta is allowed to take.

OK, and finally, we're going to end with the star of all, and that's the normal distribution. And in the normal distribution, you still have also some flexibility in terms of choices, because then naturally, the normal distribution is parametrized by-- the normal distribution is parametrized by two parameters, right? Mean and variance.

So what values can a Gaussian random variable take? An entire real line, right? And the set of parameters that it can take it-- so this is going to be n, mu, sigma square. And mu is going to be positive. And stigma square is going-- sorry, m is going to be an r. And sigma square is going to be positive.

OK, so again here, that's the way you're supposed to write it. If you really want to identify what theta is, well, theta formally is the set of mu sigma square such that-- well, in r times 0 infinity, right? That's just to be formal, but this does the job just fine. OK? You don't have to be super formal.

OK, that's not three. That's like five. Actually, I just want to write another one. Let's call it 5-bit. And 5-bit is just Gaussian with known variants. And this arises a lot in labs when you have measurement error-- when you actually receive your measurement device. This thing has been tested by the manufacturer so much that it actually comes in on the side of the box. It says that the standard deviation of your measurements is going to be 0.23.

OK, and actually why you do this is because we can brag about accuracy, right? That's how they sell you this particular device. And so you actually know exactly what sigma square is. So once you actually get your data in the lab, you actually only have to estimate mu, because stigma comes on the label. So now, what is your statistical model? Well, the numbers are collecting still in r.

But now, the models that I have is n, mu, sigma squared. But the parameter space is not mu, and r, and sigma positive. It's just mu and r. And to be a little more emphatic about this, this is enough to describe it, right? Because if sigma is the sigma that was specified by the manufacturer, then this is the sigma you want. But you can actually write sigma is equal to-- sigma square is equal to sigma square manufacturer.

Right? You can just fix it to be this particular value. Or maybe you don't want to write that index that's the manufacturer. And so you just say, well, the sigma-- when I write n squared what I mean is the sigma square from the manufacturer. Yeah?

AUDIENCE: [INAUDIBLE]

PHILIPPE RIGOLLET: Yeah. For a particular measuring device? You know, you're in a lab, and you have some measuring device. I don't know-- something that measures tensile strength of something. And it's just going to measure something. And it will naturally make errors. But it's been tested so much by the manufacturer and calibrated by them. They know it's not going to be perfect. But they knew exactly what error it was making, because they've actually tried it on things for which they exactly knew what the tensile strength was. OK? Yeah.

AUDIENCE: [INAUDIBLE]

PHILIPPE RIGOLLET: This?

AUDIENCE: [INAUDIBLE]

PHILIPPE RIGOLLET: Oh, like that's pointing to-- 5 prime? OK? And we can come up with other examples, right? So for example, here's another one. So the names don't really matter, right? I call it the siblings model. But you won't find the siblings model in the textbook, right? So I wouldn't worry too much. But for example, let's say you have something-- so let's call it 6.

You have-- I don't know-- a truncated-- and that's the name I just came up with. But it's actually not exactly describing what I want. But let's say I observe y, which is the indicator of x larger than say 5 when x follows some exponential with parameter lambda. OK? This is what I get to observe. I only observe if my waiting time was more than five minutes, because I see somebody coming out of the Kendall Station being really upset.

And that's all I record is I've been waiting for more than five minutes. And that's all I get to record. OK? That happens a lot. These are called censored data. I should probably not call it truncated, but this should be censored. OK? You see a lot of censored data when you ask people how much they make. They say, well, more than five figures. And that's all they want to tell you. OK?

And so you see a lot of censored data in survival analysis, right? You are trying to understand how long your patients are going to live after some surgery, OK? And maybe you're not going to keep people alive, and you're not going to actually be in touch in their family every day and ask them, is the guy still alive?

And so what you can do is just you ask people maybe five years after your study and say, please, come in. And you will just happen to have some people say, well, you know, the person is deceased. And you will only be able to know that the person deceased less than five years ago. But you only see what happens after that, OK?

And so this is this truncated and censored data. It happens all the time just because you don't have the ability to do better than that. So this could happen here. So what is my physical experiment, right? So here, I should probably write this like this, because I just told you that my observations are going to be x, but there is some unknown y. I will never get to see this y. I only get to see the x.

What is my statistical experiment? Please help me. So is it the real line? My sample space-- is it the real line? Sorry, who does not know what this means? I'm sorry. OK. So this is called an indicator. So I read it as-- if I wrote well, that would be one with a double bar. You can also write i if you prefer if you don't feel like writing one in double bars.

And it's one of say-- I'm going to write it like that-- 1 of a is equal to 1 if a is true and 0 if a is false. OK? So that means that if y is larger than 4, this thing is 1. And if y is not larger than 5, this thing is 0. OK. So that's called an indicator-- indicator function.

It was very useful to just turn anything into a 0, 1. So now that I'm here, what is my sample space? 0, 1. Well, whatever this thing I did not tell you was taking value with the thing you should have-- if I end up telling you that is taking value 6 or 7 that would be your sample space, OK?

OK, so it takes values 0, 1. And then what is the probability here? What should I write here? What should you write without even thinking? Yeah. So let's assume there's two seconds before the end of the exam. You're going to write Bernoulli. And that's where you're going to start checking if I'm going to give you extra time, OK? So you write Bernoulli without thinking, because it's taking value 0, 1. So you just write Bernoulli, but you still have to tell me what possible parameters this thing is taking, right?

So I'm going to write it p, because I don't know. And then p take value-- OK, so sorry. I could write it like that. Right? That would be perfectly valid, but actually no more. It's not any p. The p is the probability that an exponential lambda is larger than 5. And maybe I want to have lambda as a parameter.

OK, so what I need to actually compute is, what is the probability that y is larger than 5-- when y is this exponential lambda, which means that what I need to compute is the integral between 5 and infinity of-- what is it? 1 over lambda. How did I define it in this class? Did I change it-- what?

AUDIENCE: [INAUDIBLE].

PHILIPPE RIGOLLET: Yeah, right, right, right. Yeah. Lambda e to the minus lambda x dx, right? So that's what I need to compute. What is this? Yeah, so what is the value of this integral? Can you take appropriate measures?

AUDIENCE: [INAUDIBLE]

PHILIPPE RIGOLLET: OK? And again, you can cancel this, right? So when I'm going to integrate this guy, those guys are going to cancel. I'm going to get 0 for infinity. I'm going to get a 5 for this guy. And well, I know it's going to be positive number, so I'm not really going to bother with the signs, because I know that's what it should be.

OK, so I get e to the minus 5 lambda. And so that means that I can actually write this like that-- and now parametrize this thing by lambda positive. OK? So what I did here is I changed the parametrization from p to lambda. Why? Well, because maybe if I know this is happening, maybe I am actually interested in reporting lambda to MBTA, for example.

Maybe I'm actually trying to estimate 1 over lambda, so that I know it is-- well, lambda is actually the intensity of arrival of my Poisson process, right? I have a Poisson process. That's how my trains are coming in. And so I'm interested in lambda. So I will parametrize things by lambda. So the thing I get is lambda. You can play with this, right? I mean, I could parametrize this by 1 over lambda and put 1 over lambda here if I want it. But you know, the context of your problem will tell you exactly how to parametrize this.

OK? So what else did I want to tell you? OK, let's do a final one. By the way, are you guys OK with Poisson exponential, Bernoulli's-- I don't know, binomial, normal-- all these things. I'm not going to go back to it, but I'm going to use them heavily. So just spend five minutes on Wikipedia if you forgot about what those things are.

Usually, you must have seen them the in your probability class. So they should not be a crazy name. And again, I'm not expecting you. I don't remember what the density of an exponential is. So it would be pretty unfair of me to actually ask you to remember what it is. Even for the Gaussian, I don't expect you to remember what it is. But I want you to remember that if I add 5 to a Gaussian, then I have a Gaussian with me and mu plus 5 if I multiply it by something, right? You need to know how to operate those things. But knowing complicated densities is definitely not part of the game. OK?

So let's do a final one. I don't know what number I have now. I'm going to just do uniform. That's another one. Everybody knows what uniform is? So it's uniform, right? So I'm going to have x, which my observations are going to be uniform on the interval 0 theta, right? So if I want to define a uniform distribution for a random variable, I have to tell you which interval or which set I want it to be uniform on. And so here I'm telling you is the interval 0 theta. And so what is going to be my sample space?

AUDIENCE: [INAUDIBLE]

PHILIPPE RIGOLLET: I'm sorry? 0 to theta. And then what is my probability distribution? My family of parameters? So well, I can write it like this, right? Uniform theta, right? And theta let's say is positive. Can somebody tell me what's wrong with what I wrote? This makes no sense. Tell me why. Yeah? Yeah, this set depends on theta, and why is that a problem?

AUDIENCE: [INAUDIBLE]

PHILIPPE RIGOLLET: There is no theta. Right now, there's the families of theta. Which one did you pick here? Right, this is just something that's indexed by theta, but I could have very well written it as, you know, just not being Greek for a second, I could have just written this as t rather than theta. That would be the same thing. And then what the hell is theta?

There's no such thing as theta. We don't know what the parameter is. This parameter should move with everyone. And so that means that I actually am not allowed to pick this theta. I'm actually-- just for the reason that there is no parameter to put on the left side-- there should not be, right? So you just said, well, there's a problem because the parameter is on the left-hand side. But there's not even a parameter. I'm describing the family of possible parameters. There is no one that you can actually plug it in.

So this should really be 1. And I'm going to go back to writing this as theta because that's pretty standard. Is that clear for everyone. I cannot just pick one and put it in there and just take the-- before I run my experiments, I could potentially get numbers that are all the way up to 1, because I don't know what theta is going to be ahead of time.

Now, if somebody promised to me that theta was going to be less than 0.5, that would be-- sorry, why do I put 1 here? I could put theta between 0 and 1. But if somebody is going to promise me, for example, if theta is going to be less than 1, then you expect to put 0, 1. All right? Is that clear?

OK, so now you know how to answer the question-- what is the statistical model? And again, within the scope of this class, you will not be asked to just come up with a model right that will just tell you. Poisson would be probably be a good idea here. And then you would just have to trust me that indeed it would be a good idea.

All right, so what I started talking about 20 minutes ago-- so it's definitely ahead of myself is the notion-- so that's when I was talking about well-specified. Remember, well-specified says that the true distribution is one of the distributions in this parametric families of distribution. The true distribution of my siblings is actually a Poisson with some parameters. And all I need to figure out is what this parameter is.

When I started saying that, I said, well, but then that could be that there are several parameters that give me the same distribution, right? It could be the case that Poisson 5 and Poisson 17 are exactly the same distributions when I started putting those numbers in the formula which I erased, OK? So it could be the case that two different numbers would give me exactly the same probabilities.

And in this case, we see that the model is not identifiable. I mean, the parameter is not identifiable. I cannot identify the parameter, even if you actually gave me an infinite amount of data, which means that I could actually estimate exactly the PMF. I might not be able to go back, because there would be several candidates, and I would not be able to tell you which one it was in the first place.

OK? So what we want is that this function-- theta maps to p theta is injective. And that really can be fancy. What I really mean is that if theta is different from theta prime, then p of theta is different from p of theta prime. Or, if you prefer to think about the contrapositive of this, this is the same as saying that if p theta gives me the same distribution as theta prime, then that implies that theta must be equal to the theta prime. The logic of those two things are equivalent, right? So that's what this means. So this is-- we say that the parameter is identifiable or identified-- it doesn't really matter-- in this model.

And this is something we're going to want. OK? So in all the examples that I gave you, those parameters are completely identified. Right? If I tell you-- I mean, if those things are in probability box, it means that they were probably thought through, right? So when I say exponential lambda, I'm really talking about one specific distribution and not-- there's not another lambda going to give you exactly the same distribution.

OK so that was the case. And you can check that, but it's a little annoying. So I would probably not do it. But rather than doing this, let me just give you some examples where it would not be the case. Again, here's an example, if I take xi-- so now I'm back to just using this indicator function-- but now for a Gaussian. So what I observe is x is the indicator that y is, what did we say? Positive. OK?

So this is a Bernoulli random variable, right? And it has some parameter p. But p now is going to depend-- sorry, and here y is n mu sigma square. So the p, the probability that this thing is positive, is actually-- I don't think I put the 0. Oh, yeah, because I have mu. OK, so this distribution-- this p the probability that it's positive is just the probably that some Gaussian is positive. And it will depend on mu and sigma, right? Because if I draw a 0, and I draw my Gaussian around mu, then the probability of this Bernoulli being 1 is really the area under the curve here.

Right? And this thing-- well, if mu is very large, it's going to become very large. If mu is very small, it's going to become very small. And if sigma changes, it's also going to effect-- is that clear for everyone? But we can actually compute this, right? So the parameter p that I'm looking for here as a function of mu and sigma is simply the probability that some y is non-negative, which is the probability that y minus mu divided by sigma is larger than minus mu divided by sigma.

But when you study probability, is that some operation you were used to making? Removing the mean and dividing by the standard deviation? What is the effect of doing that on the Gaussian random variable? Yeah, so you normalize it, right? And you standardize it. You make it a standard Gaussian. You remove the mean. The mean 0 is Gaussian. And you remove the variance for it to become 1.

So when you have a Gaussian, remove the mean and divide by the standard deviation, it becomes a standard Gaussian-- which this thing has n , 0, 1 distribution, which is the one you can read the quintiles of at the end of the book. Right? And that's exactly what we did. OK? So now you have the probability that some standard Gaussian exceeds negative mu over sigma, which I can write in terms of the cumulative distribution function, capital phi-- like we did in the first lecture. So if I do this cumulative distribution function, what is this probability in terms of phi? [INAUDIBLE]?

AUDIENCE: [INAUDIBLE].

PHILIPPE RIGOLLET: Well, that's what your name tag says. 1 minus--

AUDIENCE: [INAUDIBLE].

PHILLIPPE RIGOLLET: 1 minus mu of sigma. What happens with phi in our-- do you think I defined this for fun? 1 minus phi of mu over sigma, right? Right? Because this is 1 minus the probability that it's less than this. And this is exactly the definition of the cumulative distribution function. So in particular, this thing only depends on mu over sigma. Agreed? So in particular, if I had 2 mu over 2 sigma, p would remain unchanged. If I have 12 mu over 12 sigma, this thing would remain unchanged, which means that p does not change if I scale mu and sigma by the same factor.

So there's no way just by observing x, even an infinite times, so that I can actually get exactly what p is. I'm never going to be able to get mu and sigma separately. All I'm going to be able to get is mu over sigma. So here, we say that mu sigma-- the parameter mu sigma-- or actually each of them individually-- those guys-- they're not identifiable.

But the parameter mu over sigma is identifiable. So if I wanted to write a statistical model in which the parameter is identifiable-- I would write 0, 1 Bernoulli. And then I would write 1 minus phi over mu over sigma. And then I would take two parameters, which are mu and r and sigma squared positive. So let's write sigma positive. Right?

No, this is not identifiable. I cannot write those two guys as being two things different. Instead, what I want to write is 0, 1, Bernoulli 1 minus-- and now my parameter-- I forgot this-- my parameter is mu over sigma. Can somebody tell me where mu over sigma lives? What values can this thing take? Any real value, right?

OK, so now I've done this definitely out of convenience, right? Because that was the only thing I was able to identify-- the ratio of mu over sigma. But it's still something that has some meaning. It's the normalized mean. It really tells me what the mean is compared to the standard deviation. So in some models, in reality, in some real applications, this actually might have a good meaning.

It's just telling me how big the mean is compared to the standard fluctuations of this model. But

I won't be able to get more than that. Agreed? All right? So now that we've set a parametric model, let's try to see what our goals are going to be. OK? So now we have a sample and a statistical model. And we want to estimate the parameter theta, and I could say, well, you know what? I don't have time for this analysis. Collecting data is going to take me a while.

So I'm just going to mmm-- and I'm going to say that mu over sigma is 4. And I'm just going to give it to you. And maybe you will tell me, yeah, it's not very good, right? So we need some measure of performance of a given parameter. We need to be able to evaluate if eyeballing the problem is worse than actually collecting a large amount of theta. We need to know if even if I come up with an estimator that actually sort of uses the data, does it make an efficient use of the data?

Would I actually need 10 times more observations to achieve the same accuracy? To be able to answer these questions, well, I need to define what accuracy means. And accuracy is something that sort of makes sense. It says, well, I want theta. I have to be close to theta. And the theta is a random variable. So I'm going to have to understand what it means for a random variable to be close to a deterministic number. And so, what is a parameter estimator, right? So I have an estimator, and I said it's a random variable.

And the formal definition-- so an estimator is a measurable function of the data. So when I write theta hat, and that will typically be my notation for an estimator, right? I should really write theta hat of x1 xn. OK? That's what an estimator is. If you want to know an estimator is, this is a measurable function of the data. And it's actually also known as a statistic.

And you know, if you're interested in, you know, I see every day I think when I have like, you know, a dinner with normal people. And they say I'm a statistician. Oh, yeah, I really like baseball. And they talk to me about batting averages. That's not what I do. But for them, that's what it is, and that's because in a way, that's what a statistic is. A batting average is a statistic.

OK, and so here are some examples. You can take the average xn bar. You can take the maximum of your observation. That's the statistics. You can take the first one. You can take the first one plus log of 1 plus the absolute value of the last one. You can do whatever you want that will be an estimator. Some of them are clearly going to be bad. But that's still a statistic, and you can do this.

Now, when I say measurable, I always have-- so you know, graduate students sometimes ask me like, yeah, how do I know if this estimator is measurable or not. And usually, my answer is, well, if I give you data, can you compute it. And they say, yeah, I'm like, well, then it's measurable. That's a very good rule to check if you can actually-- if something is actually measurable.

When is this thing non-measurable? It's when it's implicitly defined. OK, and in particular, the things that give you problems are-- sup or inf. Anybody knows what a sup or an inf is? It's like a max or a min. But it's not always attained. OK, so if I have x1. So if I look at the infimum of the function f of x for x on the real line and f of x, sorry, let's say x on the 1 infinity. And f of x is equal to 1 over x. Right? Then the infimum is the smallest value we can take except that it doesn't really take it at 0 right, because 1 over x is going to 0.

But it's never really getting there. So we just called the inf 0. But it's not the value that it ever takes. And these things might actually be complicated to compute. And so that's when you actually have problems, right? When the limit is not-- you're not really quite reaching the limit. You won't have this problem in general, but just so you know, an estimator is not really anything. It has to actually be measurable.

OK, so the first thing we want to know I mentioned it-- so an estimator is a statistic which does not depend on theta, of course. So if I give you the data, you have to be able to compute it. And that probably should not require not knowing any known parameters. OK, so an estimator is said to be consistent. When my data-- when I collect more and more data, this thing is getting closer and closer to the true parameter.

All right? And we said that eyeballing and saying that it's going to be 4 is not really something that's probably going to be consistent. But they can have things that are consistent but that are converging to theta at different speeds. OK? And we know also that this is a random variable. It converges to something. And there might be some different notions of convergence that kick in.

And actually there are. And we say that it's weakly convergent if it converges in probability and strongly convergent if it converges almost [INAUDIBLE]. OK? And this is just vocabulary. It won't make a big difference. OK? So we will typically say it's consistent with any of the two.

AUDIENCE: [INAUDIBLE].

PHILIPPE RIGOLLET: Well, so in parametric statistics, it's actually a little difficult to come up with. But in non-parametric ones, I could just say, if I had xi, yi, and I know that yi is f of xi plus noise s1i. And I know that f belongs to some class of function, let's say-- [INAUDIBLE] class of smooth functions-- it's massive. And now, I'm going to actually find the following estimator. I'm going to take the average. So I'm going to do least squares, right?

So I just check. I'm trying to minimize the distance of each of my f of xi to my yi. And now, I want to find the smallest of them. So if I look at the infimum here, then the question is-- so that could be-- well, that's not really an estimator for f. But it's an estimator for the smallest possible value. And so for example, this is actually an estimator for the variance of sigma square. This might not be attained, and this might not be measurable if f is massive?

All right, so that's the infimum over some class f of x. OK? So those are all voice things that are defined implicitly. If it's an average, for example, it's completely measurable. OK?

Any other question? OK, so we know that the first thing we might want to check, and that's definitely something we want about estimators that is consistent, because all consistency tells us is that just as I collect more and more data, my estimator is going to get closer and closer to the parameter.

There's other things we can look at. For each possible value of n-- now, right now, I have a finite number of observations-- 25. And I want to know something about my estimator. The first thing I want to check is maybe if in average, right? So this is a random variable. Is this random variable in average going to be close to theta or not?

And so the difference how far I am from theta is actually called the bias. So the bias of an estimator is the expectation of theta hat minus the value that I hope it gets, which is theta. If this thing is equal to 0, we say that theta hat is unbiased. And unbiased estimators are things that people are looking for in general.

The problem is that there's lots of unbiased estimators. And so it might be misleading to look for unbiasedness when that's not really the only thing you should be looking for. OK, so what does it mean to be unbiased? Maybe for this particular round of data you collected, you're actually pretty far from the true estimator. But one thing that actually-- what it means is that if I redid this experiment over, and over, and over again, and I averaged all the values of my estimators that I got, then this would actually be the right-- the true parameter.

OK. That's what it means. If I were to repeat this experiment, in average, I would actually get the right thing. But you don't get to repeat the experiment. OK, just a remark about estimators, look at this estimator-- xn bar. Right? Think of the kiss example. I'm looking at the average of my observations. And I want to know what the expectation of this thing is.

OK? Now, this guy is by linearity of the expectation, it is this, right? But my data is identically distributed. So in particular, all the xi's have the same expectation, right? Everybody agrees with this. When it's identically distributed, they'll get the same expectation. So what it means is that this guy's here-- they're all equal to the expectation of x1. Right?

So what it means is that these guys-- I have the average of the same number. So this is actually the expectation of x1. OK? And it's true. In the kiss example, this was p. And this is p-- the probability of turning your head right. OK? So those two things are the same. In particular, that means that xn bar and just x1 have the same bias.

So that should probably illustrate to you that bias is not something that really is telling you the entire picture, Right? I can take only one of my observations-- Bernoulli 0, 1. This thing will have the same bias as if I average 1,000 of them. But the bias is really telling you where I am in average. But it's really not telling me what fluctuations I'm getting.

And so if you want to start having fluctuations coming into the picture, we actually have to look at the risk or the quadratic risk of the estimator. And so the quadratic risk is the finest-- the expectation of the square distance between theta hat and theta. OK? So let's look at this.

So the quadratic risk-- sometimes it's denoted that people call it the l2 risk of theta hat, of course. I'm sorry for maintaining such an ugly board. [INAUDIBLE] this stuff.

OK, so I look at the square distance between theta hat and theta. This is still-- this is the function of a random variable. So it's a random variable as well. And now I'm looking at the expectation of this guy. That's the definition. I claimed that when this thing goes to 0, then my estimator is actually going to be consistent. Everybody agrees with this?

So if it goes to zero as n goes to infinity-- and here, I don't need to tell you what kind of convergence I have, because this is just the number, right? It's an expectation. So it's a regular, usual calculus-style convergence. Then that implies that theta hat is actually weakly consistent. What did I use to tell you this?

Yeah, this is the convergence in l2. This actually is strictly equivalent. This is by definition saying that theta hat converges in l2 to theta. And we know that convergence in l2 implies convergence in credibility to theta. That was the picture. We're going up. And this is actually equivalent to a consistency by definition-- a weak consistency.

OK, so this is actually telling you a little more because this guy here-- they are both unbiased. Theta xn bar is unbiased. X1 is unbiased. But x1 is certainly not consistent , because the more data I collect, I'm not even doing anything with it. I'm just taking the first data point you're giving to me. So they're both unbiased. But this one is not consistent. And this one we'll see is actually consistent. xn bar is consistent. And actually, we've seen that last time. And that's because of the? What guarantees the fact that xn bar is consistent?

AUDIENCE: The law of large numbers.

PHILIPPE RIGOLLET: The law of large numbers, right? Actually, it's strongly consistent if you have a strong [INAUDIBLE]. OK, so just in the last two minutes, I want to tell you a little bit about how this risk is linked to see, quadratic risk is equal to bias squared plus the variance. So let's see what I mean by this? So I'm going to forget about the absolute values that you have a square. I don't really need them.

If theta hat was unbiased, this thing would be the expectation of theta hat. It might not be the case. So let me see how I can actually see-- put the bias in there. Well, one way to do this is to see that this is equal to the expectation of theta hat minus the expectation of theta hat, plus the expectation of theta hat minus theta. OK? I just removed the same and added the same thing. So I didn't change anything.

Now, this guy is my bias, right? So now let me expand the square. So what I get is the expectation of the square of theta hat minus its expectation. I should put some square brackets-- plus two times the cross-product. So the cross-product is what expectation of theta hat minus the expectation of theta hat times expectation of theta hat minus theta. And then I have the last square.

Expectation of theta hat minus theta squared. OK? So square, cross-products, square. Everybody is with me? now this guy here-- if you pay attention, this thing is the expectation of some random variable. So it's a deterministic number. Theta is the true parameter. It's a deterministic number. So what I can do is pull out this entire thing out of the expectation like this and compute the expectation only with respect to that part. But what is the expectation of this thing?

It's zero, right? The expectation of theta hat minus the expectation of theta hat is 0. So this entire thing is equal 0. So now when I actually collect back my quadratic terms-- my two squared terms in this expansion-- what I get is that the expectation of theta hat minus theta squared is equal to the expectation of theta hat minus expectation of theta hat squared plus the square of expectation of theta hat minus theta. Right?

So those are just the two-- the first and the last term of the previous equality? Now, here I have the expectation of the square of the difference between a random variable and its expectation. This is otherwise known as the variance, right? So this is actually equal to the variance of theta hat. And well, this was the bias. We already said that's there. So this whole thing is the bias square.

OK? And hence the quadratic term is the sum of the variance and the squared bias. Why squared bias? Well, because otherwise, you would add dollars in dollars squared. So you need to add dollars squared and dollars squared so that this thing is actually homogeneous. So if x is in dollars, then the bias is in dollars, but the variance is in dollars squared. OK, and the square here forced you to put everything on the square scale.

All right, so what's nice is that if the quadratic risk goes to 0, then since I have the sum of two positive terms, both of them have to go to 0. That means that my variance is going to 0-- very little fluctuations. And my bias is also going to 0, which means that I'm actually going to be on target once I reduce my fluctuations, because it's one thing to reduce the fluctuations. But if I'm not on target, it's an issue, right?

For example, the estimator for the value 4 has no variance. Every time I'm going to repeat the experiments, I'm going to get 4, 4, 4, 4-- variance is 0. But the bias is bad. The bias is 4 minus theta. And if theta is far from 4, that's not doing very well. OK, so next week, we will-- we'll talk about what is a good estimate-- how estimators change if they have high variance or low variance or high bias and low bias. And we'll talk about confidence intervals as well.