PHILIPPE RIGOLLET: So today, we're going to close this chapter, this short chapter, on Bayesian inference. Again, this was just an overview of what you can do in Bayesian inference. And last time, we started defining what's called Jeffreys priors. Right? So when you do Bayesian inference, you have to introduce a prior on your parameter. And we said that usually, it's something that encodes your domain knowledge about where the parameter could be. But there's also some principle way to do it, if you want to do Bayesian inference without really having to think about it.

And for example, one of the natural priors were those non-informative priors, right? If you were on a compact set, it's a uniform prior of this set. If you're on an infinite set, you can still think of taking the [? 01s ?] prior. And that's called an [INAUDIBLE] That's always equal to 1. And that's an improper prior if you are an infinite set or proportional to one.

And so another prior that you can think of, in the case where you have a Fisher information, which is well-defined, is something called Jefferys prior. And this prior is a prior which is proportional to square root of the determinant of the Fisher information matrix. And if you're in one dimension, it's basically proportional to a square root of the Fisher information coefficient, which we know, for example, is the asymptotic variance of the maximum likelihood estimator. And it turns out that it's basically. So square root of this thing is basically one over the standard deviation of the maximum likelihood estimator.

And so you can compute this, right? So you can compute for the maximum likelihood estimator. We know that the variance is going to be p1 minus p in the Bernoulli statistical experiment. So you get this one over the square root of this thing. And for example, in the Gaussian setting, you actually have the Fisher information, even in the multi-variate one, is actually going to be something like the identity matrix. So this is proportional to 1. It's the improper prior that you get, in this case, OK? Meaning that, for the Gaussian setting, no place where you center your Gaussian is actually better than any other.

All right. So we basically left on this slide, where we saw that Jeffreys prior satisfy a reparametrization [INAUDIBLE] invariant by transformation of your parameter, which is a desirable property. And the way, it says that, well, if I have my prior on theta, and then I suddenly decide that theta is not the parameter I want to use to parameterize my problem, actually what I want is phi of theta. So think, for example, as theta being the mean of a

Gaussian, and phi of theta as being mean to the cube. OK? This is a one-to-one map phi, right?

So for example, if I want to go from theta to theta cubed, and now I decide that this is the actual parameter that I want, well, then it means that, on this parameter, my original prior is going to induce another prior. And here, it says, well, this prior is actually also Jeffreys prior. OK? So it's essentially telling you that, for this new parametrization, if you take Jeffreys prior, then you actually go back to having exactly something that's of the form's [INAUDIBLE] of determinant of the Fisher information, but this thing with respect to your new parametrization All right.

And so why is this true? Well, it's just this change of variable theorem. So it's essentially telling you that, if you call-- let's call p-- well, let's go pi tilde of eta prior over eta. And you have pi of theta as the prior over theta, than since eta is of the form phi of theta, just by change of variable, so that's essentially a probability result. It says that pi tilde of eta is equal to pi of eta times d pi of theta times d theta over d eta and-- sorry, is that the one?

Sorry, I'm going to have to write it, because I always forget this. So if I take a function-- OK. So what I want is to check. OK, so I want the function of eta that I can here. And what I know is that this is h of phi of theta. All right? So sorry, eta is phi of theta, right? Yeah. So what I'm going to do is I'm going to do the change of variable, theta is phi inverse of eta. So eta is phi of theta, which means that d eta is equal to d-- well, to phi prime of theta d theta.

So when I'm going to write this, I'm going to get integral of h. Actually, let me write this, as I am more comfortable writing this as e with respect to eta of h of eta. OK? So that's just eta according to being drawn from the prior. And I want to write this as the integral of he of eta times some function, right? So this is the integral of h of phi of theta pi of theta d theta.

Now, I'm going to do my change of variable. So this is going to be the integral of h of eta. And then pi of phi of-- so theta is phi inverse of eta. And then d theta is phi prime of theta d theta, OK?

And so what is pi of phi theta? So this thing is proportional. So we're in, say, dimension 1, so it's proportional of square root of the Fisher information. And the Fisher information, we know, is the expectation of the square of the derivative of the log likelihood, right? So this is square root of the expectation of d over d theta of log of-- well, now, I need the density. Well, let's just call it l of theta. And I want this to be taken at phi inverse of eta squared.

And then what I pick up is the-- so I'm going to put everything under the square. So I get phi prime of theta squared d theta. OK? So now, I have the expectation of a square. This does not depend, so this is-- sorry, this is I of theta. This is the expectation of l of theta of an x, right? That's for some variable, and the expectation here is with respect to x. That's just the definition of the Fisher information.

So now I'm going to squeeze this guy into the expectation. It does not depend on x. It just acts as a constant. And so what I have now is that this is actually proportional to the integral of h eta times the square root of the expectation with respect to x of what? Well, here, I have d over d theta of log of theta. And here, this guy is really d eta over d theta, right? Agree?

So now, what I'm really left by-- so I get d over d theta times d-- sorry, times d theta over d eta. so that's just d over d eta of log of eta x. And then this guy is now becoming d eta, right? OK, so this was a mess. This is a complete mess, because I actually want to use phi. I should not actually introduce phi at all. I should just talk about d eta over d theta type of things. And then that would actually make my life so much easier. OK.

I'm not going to spend more time on this. This is really just the idea, right? You have square root of a square in there. And then, when you do your change of variable, you just pick up a square. You just pick up something in here. And so you just move this thing in there. You get a square. It goes inside the square. And so your derivative of the log likelihood with respect to theta becomes a derivative of the log likelihood with respect to eta. And that's the only thing that's happening here. I'm just being super sloppy, for some reason. OK.

And then, of course, now, what you're left with is that this is really just proportional. Well, this is actually equal. Everything is proportional, but this is equal to the Fisher information tilde with respect to eta now. Right? You're doing this with respect to eta. And so that's your new prior with respect to eta. OK.

So one thing that you want to do, once you have-- so remember, when you actually compute your posterior rate, rather than having-- so you start with a prior, and you have some observations, let's say, x1 to xn. When you do Bayesian inference, rather than spitting out just some theta hat, which is an estimator for theta, you actually spit out an entire posterior distribution-- pi of theta, given x1 xn. OK?

So there's an entire distribution on the [INAUDIBLE] theta. And you can actually use this to perform inference, rather than just having one number. OK? And so you could actually build confidence regions from this thing. OK. And so a Bayesian confidence interval-- so if your set of parameters is included in the real line, then you can actually-- it's not even guaranteed to be to be an interval. So let me call it a confidence region, so a Bayesian confidence region, OK? So it's just a random subspace. So let's call it r, is included in theta.

And when you have the deterministic one, we had a definition, which was with respect to the randomness of the data, right? That's how you actually had a random subset. So you had a random confidence interval. Here, it's actually conditioned on the data, but with respect to the randomness that you actually get from your posterior distribution. OK? So such that the probability that your theta belongs to this confidence region, given x1 xn is, say, at least 1 minus alpha. Let's just take it equal to 1 minus alpha. OK so that's a confidence region at level 1 minus alpha.

OK, so that's one way. So why would you actually-- when I actually implement Bayesian inference, I'm actually spitting out that entire distribution. I need to summarize this thing to communicate it, right? I cannot just say this is this entire function. I want to know where are the regions of high probability, where my perimeter is supposed to be? And so here, when I have this thing, what I actually want to have is something that says, well, I want to summarize this thing into some subset of the real line, in which I'm sure that the area under the curve, here, of my posterior is actually 1 minus alpha.

And there's many ways to do this, right? So one way to do this is to look at level sets. And so rather than actually-- so let's say my posterior looks like this. I know, for example, if I have a Gaussian distribution, I can actually take my posterior to be-- my posterior is actually going to be Gaussian. And what I can do is to try to cut it here on the y-axis so that now, the area under the curve, when I cut here, is actually 1 minus alpha.

OK, so I have some threshold tau. If tau goes to plus infinity, then I'm going to have that this area under the curve here is going to--

AUDIENCE: [INAUDIBLE]

PHILIPPE RIGOLLET: Well, no. So the area under the curve, when tau is going to plus infinity, think of the small, the when tau is just right here.

AUDIENCE: [INAUDIBLE]

PHILIPPE RIGOLLET: So this is actually going to 0, right? And so I start here. And then I start going down and down and down and down, until I actually get something which is going down to 1 plus alpha. And if tau is going down to 0, then my area under the curve is going to-- if tau is here, I'm cutting nowhere. And so I'm getting 1, right? Agree?

Think of, when tau is very close to 0, I'm cutting [? s ?] s very far here. And so I'm getting some area under the curve, which is almost everything. And so it's going to 1-- as tau going down to 0. Yeah?

AUDIENCE: Does this only work for [INAUDIBLE]

PHILIPPE RIGOLLET: No, it does not. I mean-- so this is a picture. So those two things work for all of them, right? But when you have a [? bimodal, ?] actually, this is actually when things start to become interesting, right? So when we built a frequentist confidence interval, it was always of the form x bar plus or minus something. But now, if I start to have a posterior that looks like this, what I'm going to start cutting off, I'm going to have two-- I mean, my confidence region is going to be the union of those two things, right?

And it really reflects the fact that there is this bimodal thing. It's going to say, well, with hyperbole, I'm actually going to be either here or here. Now, the meaning here of a Bayesian confidence region and the confidence interval are completely distinct notions, right? And I'm going to work out on example with you so that we can actually see that sometimes-- I mean, both of them, actually you can come up with some crazy paradoxes. So since we don't have that much time, I will actually talk to you about why, in some instances, it's actually a good idea to think of Bayesian confidence intervals rather than frequentist ones.

So before we go into more details about what those Bayesian confidence intervals are, let's remind ourselves what does it mean to have a frequentist confidence interval? Right? OK. So when I have a frequentist confidence interval, let's say something like x bar n to minus 1.96 sigma over root n and x bar n plus 1.96 sigma over root n, so that's the confidence interval that you get for the mean of some Gaussian with known variants to be equal to sigma square, OK.

So what we know is that the meaning of this is the probability that theta belongs to this is equal to 95%, right? And this, more generally, you can think of being q alpha over 2. And what you're going to get is 1 minus alpha here, OK? So what does it mean here? Well, it looks very much like what we have here, except that we're not conditioning on x1 xn. And we should not. Because there was a question like that in the midterm-- if I condition on x1 xn, this probability is either 0 or 1. OK? Because once I condition-- so here, this probability, actually, here is with respect to the randomness in x1 xn.

So if I condition-- so let's build this thing, r freq, for frequentist. Well, given x1 xn-- and actually, I don't need to know x1 xn really. What I need to know is what xn bar is. Well, this thing now is what? It's 1, if theta is in r, and it's 0, if theta is not in r, right? That's all there is. This is a deterministic confidence interval, once I condition x1 xn. So I have a number. The average is maybe 3. And so I get 3. Either theta is between 3 minus 0.5 or in 3 plus 0.5, or it's not. And so there's basically-- I mean, I write it as probability, but it's really not a probalistic statement. It's either it's true or not. Agreed?

So what does it mean to have a frequentist confidence interval? It means that if I were-- and here, where the word frequentist comes from, it says that if I repeat this experiment over and over, meaning that on Monday, I collect a sample of size n, and I build a confidence interval, and then on Tuesday, I collect another sample of size n, and I build a confidence interval, and on Wednesday, I do this again and again, what's going to happen is the following. I'm going to have my true theta that lives here.

And then on Monday, this is the confidence interval that I build. OK, so this is the real line. The true theta is here, and this is the confidence interval I build on Monday. All right? So x bar was here, and this is my confidence interval.

On Tuesday, I build this confidence interval maybe. x bar was closer to theta, but smaller. But then on Wednesday, I build this confidence interval. I'm not here. It's not in there. And that's this case. Right? It happens that it's just not in there. And then on Thursday, I build another one. I almost miss it, but I'm in there, et cetera. Maybe here, Here, I miss again. And so what it means to have a confidence interval-- so what does it mean to have a confidence interval at 95%?

AUDIENCE: [INAUDIBLE]

PHILIPPE RIGOLLET: Yeah, so it means that if I repeat this the frequency of times-- hence, the word frequentist-- at which I'm actually going to overlap that, I'm actually going to contain theta, should be 95%. That's what frequentist means. So it's just a matter of trusting that.

So on one given thing, one given realization of your data, it's not telling you anything. [INAUDIBLE] it's there or not. So it's not really something that's actually something that assesses the confidence of your decision, such as data is in there or not. It's something that assesses the confidence you have in the method that you're using. If you were you repeat it over and again, it'd be the same thing. It would be 95% of the time correct, right?

So for example, we know that we could build a test. So it's pretty clear that you can actually build a test for whether theta is equal to theta naught or not equal to theta naught, by just checking whether theta naught is in a confidence interval or not. And what it means is that, if you actually are doing those tests at 5%, that means that 5% of the time, if you do this over and again, 5% of the time you're going to be wrong.

I mentioned my wife does market research. And she does maybe, I don't know, 100,000 tests a year. And if they do all of them at 1%, then it means that 1% of the time, which is a lot of time, right? When you do 100,000 a year, it's 1,000 of them are actually wrong. OK, I mean, she's actually hedging against the fact that 1% of them that are going to be wrong. That's 1,000 of them that are going to be wrong.

Just like, if you do this 100,000 times at 95%, 5,000 of those guys are actually not going to be the correct ones. OK? So I mean, it's kind of scary. But that's the way it is. So that's with the frequentist interpretation of this is.

Now, as I mentioned, when we started this Bayesian chapter, I said, Bayesian statistics converge to-- I mean, Bayesian decisions and Bayesian methods converge to frequentist methods. When the sample size is large enough, they lead to the same decisions. And in general, they need not be the same, but they tend to actually, when the sample size is large enough, to have the same behavior.

Think about, for example, the posterior that you have when you have in the Gaussian case, right? We said that, in the Gaussian case, what you're going to see is that it's as if you had an extra observation which was essentially given by your prior. OK? And now, what's going to happen is that, when this just one observation among n plus 1, it's really going to be totally drawn, and you won't see it when the sample size grows larger.

So Bayesian methods are particularly useful when you have a small sample size. And when you have a small sample size, the effect of the prior is going to be bigger. But most importantly, you're not going to have to repeat this thing over and again. And you're going to have a meaning. You're going to have to have something that has a meaning for this particular data set that you have.

When I said that the probability that theta belongs to r-- and here, I'm going to specify the fact that it's a Bayesian confidence region, like this one-- this is actually conditionally on the data that you've collected. It says, given this data, given the points that you have-- just put in some numbers, if you want, in there-- it's actually telling you the probability that theta belongs to this Bayesian thing, to this Bayesian confidence region. Here, since I have conditioned on x1 xn, this probability is really just with respect to theta drawn from the prior, right?

And so now, it has a slightly different meaning. It's just telling you that when-- it's really making a statement about where the regions of hyperability of your posterior are. Now, why is that useful? Well, there's actually an interesting story that goes behind Bayesian methods. Anybody knows the story of the *USS* I think it's *Scorpion?* Do you know the story?

So that was an American vessel that disappeared. I think it was close to Bermuda or something. But you can tell the story of the Malaysian Airlines, except that I don't think it's such a successful story. But the idea was essentially, we're trying to find where this thing happened. And of course, this is a one-time thing. You need something that works once. You need something that works for this particular vessel. And you don't care, if you go to the Navy, and you tell them, well, here's a method. And for 95 out of 100 vessels that you're going to lose, we're going to be able to find it. And they want this to work for this particular one.

And so they were looking, and they were diving in different places. And suddenly, they brought in this guy. I forget his name. I mean, there's a whole story about this on Wikipedia. And he started collecting the data that they had from different dives and maybe from currents. And he started to put everything in. And he said, OK, what is the posterior distribution of the location of the vessel, given all the things that I've seen?

And what have you seen? Well, you've seen that it's not here, it's not there, and it's not there. And you've also seen that the currents were going that way, and the winds were going that way. And you can actually put some modeling traits to understand this. Now, given this, for this particular data that you have, you can actually think of having a two-dimensional density that tells you where it's more likely that the vessel is.

And where are you going to be looking for? Well, if it's a multimodal distribution, you're just going to go to the highest mode first, because that's where it's the most likely to be. And maybe it's not there, so you're just going to update your posterior, based on the fact that it's not there, and do it again. And actually, after two dives, I think, he actually found the thing.

And that's exactly where Bayesian statistics start to kick in. Because you put a lot of knowledge into your model, but you also can actually factor in a bunch of information, right? The model, he had to build a model that was actually taking into account and currents, and when. And what you can have as a guarantee is that, when you talk about the probability that this vessel is in this location, given what you've observed in the past, it actually has some sense. Whereas, if you were to use a frequentist approach, then there's no probability. Either it's underneath this position or it's not, right? So that's actually where it start to make sense.

And so you can actually build this. And there's actually a lot of methods that are based on, for search, that are based on Bayesian methods. I think, for example, the Higgs boson was based on a lot of Bayesian methods, because this is something you need to find [INAUDIBLE], right? I mean, there was a lot of prior that has to be built in. OK.

So now, you build this confidence interval. And the nicest way to do it is to use level sets. But again, just like for Gaussians, I mean, if I had, even in the Gaussian case, I decided to go at x bar plus or minus something, but I could go at something that's completely asymmetric. So what's happening is that here, this method guarantees that you're going to have the narrowest possible confidence intervals. That's essentially what it's telling you, OK? Because every time I'm choosing a point, starting from here, I'm actually putting as much area under the curve as I can. All right.

So those are called Bayesian confidence [? interval. ?] Oh yeah, and I promised you that we're going to work on some example that actually gives a meaning to what I just told you, with actual numbers. So this is something that's taken from Wasserman's book. And also, it's coming from a paper, from a stats paper, from [? Wolpert ?] and I don't know who, from the '80s. And essentially, this is how it works.

So assume that you have n equals 2 observations. And you have y1, so those observations are y1-- no, sorry, let's call them x1, which is theta, plus epsilon 1 and x2, which is theta plus epsilon 2, where epsilon 1 and epsilon 2 are iid. And the probability that epsilon i is equal to plus 1 is equal to the probability that epsilon i is equal to minus 1 is equal to 1/2. OK, so it's just the uniform sign plus minus 1, OK?

Now, let's think about so you're trying to do some inference on theta. Maybe you actually want to find some inference on theta that's actually based on-- and that's based only on the x1 and x2. OK? So I'm going to actually build a confidence interval. But what I really want to build is a-- but let's start thinking about how I would find an estimator for those two things.

Well, what values am I going to be getting, right? So I'm going to get either theta plus 1 or theta minus 1. And actually, I can get basically four different observations, right? Sorry, four different pairs of observations-- plus plus theta minus 1. Agreed? Those are the four possible observations that I can get. Agreed? Either they're both equal to plus 1, both equal to minus 1, or one of the two is equal to plus 1, the other one to minus 1, or the epsilons. OK. So those are the four observations I can get.

So in particular, if they take the same value, and you know it's either theta plus 1 or theta minus 1, and if they take a different value, I know one of them is theta plus 1, and one is actually theta minus 1. So in particular, if I take the average of those two guys, when they take different values, I know I'm actually getting theta right.

So let's build a confidence region. OK, so I'm actually going to take a confidence region, which is just a singleton. And I'm going to say the following. Well, if x1 is equal to x2, I'm just going to take x1 minus 1, OK? So I'm just saying, well, I'm never going to able to resolve whether it's plus 1 or minus 1 that actually gives me the best one, so I'm just going to take a dive and say, well, it's just plus 1. OK?

And then, if they're different, then here, I can do much better. I'm going to actually just think the average. OK? Now, what I claim is that this is a confidence region-- and by default, when I don't mention it, this is a frequentist confidence region-- at level 75%. OK? So let's just check that. To check that this is correct, I need to check that the probability under the realization of x1 and x2, that theta belongs, is one of those two guys, is actually equal to 0.75. Yes?

AUDIENCE: What are the [INAUDIBLE]

PHILIPPE RIGOLLET: Well, it's just the frequentist confidence interval that does not need to be an interval. Actually, in this case, it's going to be an interval. But that's just what it means. Yeah, region for Bayesian was just because-- I mean, the confidence intervals, when we're frequentist, we tend to make them intervals, because we want-- but when you're Bayesian, and you're doing this level set thing, you cannot really guarantee, unless its [INAUDIBLE] is going to be an interval. So region is just a way to not have to say interval, in case it's not.

OK. So I have this thing. So what I need to check is the probability that theta is in one of those two things, right? So what I need to find is the probability that theta is an [INAUDIBLE] Well, x1 minus 1 and x1 is not equal to x2. And those are disjoint events, so it's plus the probability that theta is in x1 plus x2 over 2 and x1-- sorry, that's equal. That's different. OK.

And OK, just before we actually finish the computation, why do I have 75%? 75% is 3/4. So it means that we have four cases. And essentially, I did not account for one case. And it's true. I did not account for this case, when the both of the epsilon i's are equal to minus 1. Right? So this is essentially the one I'm not going to be able to account for. And so we'll see that in a second.

So in this case, we know that everything goes great. Right? So in this case, this is-- OK. Well, let's just start from the first line. So the first line is the probability that theta is equal to x1 minus 1 and those two are equal. So this is the probability that theta is equal to-- well, this is theta plus epsilon 1 minus 1. And epsilon 1 is equal to epsilon 2, right? Because I can remove the theta from here, and I can actually remove the theta from here, so that this guy here is just epsilon 1 is equal to 1.

So when I intersect with this guy, it's actually the same thing as epsilon 1 is equal to 1, as well-- episilon 2 is equal to 1, as well, OK? So this first thing is actually equal to the probability that epsilon 1 is equal to 1 and epsilon 2 is equal to 1, which is equal to what?

AUDIENCE: [INAUDIBLE]

PHILIPPE RIGOLLET: Yeah, 1/4, right? So that's just the first case over there. They're independent. Now, I still need to do the second one. So this case is what? Well, when those things are equal, x1 plus x2 over 2 is what? Well, I get theta plus theta over 2. So that's just equal to the probability that epsilon 1 plus epsilon 2 over 2 is equal to 0 and epsilon 1 is different from epsilon 2. Agreed? I just removed the thetas from these equations, because I can. They're just on both sides every time.

OK. And so that means what? That means that the second part-- so this thing is actually equal to 1/4 plus the probability that epsilon 1 and epsilon 2 over 2 is equal to 0. I can remove the 2. So this is just the probability that one is 1, and the other one is minus 1, right? So that's equal to the probability that epsilon 1 is equal to 1 and epsilon 2 is equal to minus 1 plus the probability that epsilon 1 is equal to minus 1 and epsilon 2 is equal to plus 1, OK? Because they're disjoint events. So I can break them into the sum of the two.

And each of those guys is also one of the atomic part of it. It's one of the basic things. And so each of those guys has probability 1/4. And so here, we can really see that we accounted for everything, except for the case when epsilon 1 was equal to minus 1, and epsilon 2 was equal to minus 1. So this is 1/4. This is 1/4. So the whole thing is equal to 3/4.

So now, what we have is that the probability that epsilon 1 is in-- so the probability that data belongs to this confidence region is equal to 3/4. And that's very nice. But the thing is some people are sort of-- I mean, it's not super nice to be able to see this, because, in a way, I know that, if I observe x1 and x2 that are different, I know for sure that theta, that I actually got the right theta, right? That this confidence interval is actually happening with probability 1.

And the problem is that I do not know-- I cannot make this precise with the notion of frequentist confidence intervals. OK? Because frequentist confidence intervals have to account for the fact that, in the future, it might not be the case that x1 and x2 are different. So Bayesian confidence regions, by definition-- well, they're all gone-- but they are conditioned on the data that I have. And so that's what I want. I want to be able to make this statement conditionally and the data that I have. OK.

So if I want to be able to make this statement, if I want to build a Bayesian confidence region, I'm going to have to put a prior on theta. So without loss of generality-- I mean, maybe with-- but let's assume that pi is a prior on theta. And let's assume that pi of j is strictly positive for all integers j equal, say, 0-- well, actually, for all j in the integers, positive or negative. OK.

So that's a pretty weak assumption on my prior. I'm just assuming that theta is some integer. And now, let's build our Bayesian confidence region. Well, if I want to build a Bayesian confidence region, I need to understand what my posterior is going to be. OK? And if I want to understand what my posterior is going to be, I actually need to build a likelihood, right? So we know that it's the product of the likelihood and of the prior divided by-- OK.

So what is my likelihood? So my likelihood is the probability of x1 x2, given theta. Right? That's what the likelihood should be. And now let's say that actually, just to make things a little simpler, let us assume that x1 is equal to, I don't know, 5, and x2 is equal to 7. OK? So I'm not going to take the case where they're actually equal to each other, because I know that, in this case, x1 and x2 are different. I know I'm going to actually nail exactly what theta is, by looking at the average of those guys, right? Here, it must be that theta is equal to 6.

So what I want is to compute the likelihood at 5 and 7, OK? And what is this likelihood? Well, if theta is equal to 6, that's just the probability that I will observe 5 and 7, right? So what is the probability I observe 5 and 7? Yeah? 1?

AUDIENCE: 1/4.

PHILIPPE RIGOLLET: That's 1/4, right? As the probability, I have minus 1 for the first epsilon 1, right? So this is infinity 6. This is the probability that epsilon 1 is equal to minus 1, and epsilon 2 is equal to plus 1, which is equal to 1/4. So this probability is 1/4.

If theta is different from 6, what is this probability? So if theta is different from 6, since we know that we've only loaded the integers-- so if theta has to be another integer, what is the probability that I see 5 and 7?

AUDIENCE: 0.

PHILIPPE RIGOLLET: 0. So that's my likelihood. And if I want to know what my posterior is, well, it's just pi of theta times p of 5/6, given theta, divided by the sum over all T's, say, in Z. Right? So now, I just need to normalize this thing. So of pi of T, p of 4/6, given T. Agreed? That's just the definition of the posterior. But when I sum these guys, there's only one that counts, because, for those things, we know that this is actually equal to 0 for every T, except for when T is equal to 6. So this entire sum here is actually equal to pi of 6 times p of 5/6-- sorry, 5/7, of 5/7, given that theta is equal to 6, which we know is equal to 1/4.

And I did not tell you what pi of 6 was. But it's the same thing here. The posterior for any theta that's not 6 is actually going to be-- this guy's going to be equal to 0. So I really don't care what this guy is.

So what it means is that my posterior becomes what? It becomes the posterior pi of theta, given 5 and 7 is equal to-- well, when theta is not equal to 6, this is actually 0. So regardless of what I do here, I get something which is 0. And if theta is equal to 6, what I get is pi of 6 times p of 5/7, given 6, which I've just computed here, which is 1/4 divided by pi of 6 times 1/4. So it's the ratio of two things that are identical. So I get 1.

So now, my posterior tells me that, given that I observe 5 and 7, theta has to be 1 with probability-- has to be 6 with probability 1. So now, I say that this thing here-- so now, this is not something that actually makes sense when I talk about frequentist confidence intervals.

They don't really make sense, to talk about confidence intervals, given something. And so now, given that I observe 5 and 7, I know that the probability of theta is equal to 1. And in this sense, the Bayesian confidence interval is actually more meaningful.

So one thing I want to actually say about this Bayesian confidence interval is that it's-- I mean, here, it's equal to the value 1, right? So it really encompasses the thing that we want. But the fact that we actually computed it using the Bayesian posterior and the Bayesian rule did not really matter for this argument. All I just said was that it had a prior.

But just what I want to illustrate is the fact that we can actually give a meaning to the probability that theta is equal to 6, given that I see 5 and 7. Whereas, we cannot really in the other cases. And we don't have to be particularly precise in the prior and theta to be able to give theta this-- to give this meaning. OK? All right.

So now, as I said, I think the main power of Bayesian inference is that it spits out the posterior distribution, and not just the single number, like frequentists would give you. Then we can say decorate, or theta hat, or point estimate, with maybe some confidence interval. Maybe we can do a bunch of tests. But at the end of the day, we just have, essentially, one number, right? Then maybe we can understand where the fluctuations of this number are in a frequentist setup.

but the Bayesian framework is essentially giving you a natural method. And you can interpret it in terms of the probabilities that are associated to the prior. But you can actually also try to make some-- so a Bayesian, if you give me any prior, you're going to actually build an estimator from this prior, maybe from the posterior. And maybe it's going to have some frequentist properties. And that's what's really nice about [? Bayesians, ?] is that you can actually try to give some frequentist properties of Bayesian methods, that are built using Bayesian methodology.

But you cannot really go the other way around. If I give you a frequency methodology, how are you going to say something about the fact that there's a prior going on, et cetera? And so this is actually one of the things there's actually some research that's going on for this. They call it Bayesian posterior concentration. And one of the things-- so there's something called the Bernstein-von Mises theorem.

And those are a class of theorems, and those are essentially methods that tell you, well, if I actually run a Bayesian method, and I look at the posterior that I get-- it's going to be something like this-- but now, I try to study this in a frequentist point of view, there's actually a true parameter of theta somewhere, the true one. There's no prior for this guy. This is just one fixed number.

Is it true that as my sample size is going to go to infinity, then this thing is going to concentrate around theta? And the rate of concentration of this thing, the size of this width, the standard deviation of this thing, is something that should decay maybe like 1 over square root of n, or something like this. And the rate of posterior concentration, when you characterize it, it's called the Bernstein-von Mises theorem.

And so people are looking at this in some non-parametric cases. You can do it in pretty much everything we've been doing before. You can do it for non-parametric regression estimation or density estimation. You can do it for, of course-- you can do it for sparse estimation, if you want.

OK. So you can actually compute the procedure and-- yeah. And so you can think of it as being just a method somehow. Now, the estimator I'm talking about-- so that's just a general Bayesian posterior concentration. But you can also try to understand what is the property of something that's extracted from this posterior. And one thing that we actually describe was, for example, well, given this guy, maybe it's a good idea to think about what the mean of this thing is, right?

So there's going to be some theta hat, which is just the integral of theta pi theta, given x1 xn-- so that's my posterior-- d theta. Right? So that's the posterior mean. That's the expected value with respect to the posterior distribution. And I want to know how does this thing behave, how close it is to a true theta if I actually am in a frequency setup. So that's the posterior mean.

But this is not the only thing I can actually spit out, right? This is definitely uniquely defined. If you give me a distribution, I can actually spit out its posterior mean. But I can also think of the posterior median. But now, if this is not continuous, you might have some uncertainty. Maybe the median is not uniquely defined, and so maybe that's not something you use as much. Maybe you can actually talk about the posterior mode.

All right, so for example, if you're posterior density looks like this, then maybe you just want to summarize your posterior with this number. So clearly, in this case, it's not such a good idea, because you completely forget about this mode. But maybe that's what you want to do. Maybe you want to focus on the most peak mode. And this is actually called maximum a posteriori.

As I said, maybe you want a sample from this posterior distribution. OK, and so in all these cases, these Bayesian estimators will depend on the prior distribution. And the hope is that, as the sample size grows, you won't see that again. OK.

So to conclude, let's just do a couple of experiments. So if I look at-- did we do this? Yes. So for example, so let's focus on the posterior mean. And we know-- so remember in experiment one-- [INAUDIBLE] example one, what we had was x1 xn that were [? iid, ?] Bernoulli p, and the prior I put on p was a beta with parameter aa. OK? And if I go back to what we computed, you can actually compute the posterior of this thing. And we know that it's actually going to be-- sorry, that was uniform? Where is-- yeah.

So what we get is that the posterior, this thing is actually going to be a beta with parameter a plus the sum, so a plus the number of 1s and a plus the number of 0s. OK? And the beta was just something that looked like-- the density was p to the a minus 1, 1 minus p. OK?

So if I want to understand the posterior mean, I need to be able to compute the expectation of a beta, and then maybe plug in a for a plus this guy and minus this guy. OK. So actually, let me do this. OK. So what is the expectation? So what I want is something that looks like the integral between 0 and 1 of p times a minus 1-- sorry, p times p a minus 1, 1 minus p, b minus 1. Do we agree that this-- and then there's a normalizing constant. Let's call it c. OK?

So this is what I need to compute. So that's c of a and b. Do we agree that this is the posterior mean with respect to a beta with parameters a and b? Right? I just integrate p against the density. So what does this thing look like? Well, I can actually move this guy in here. And here, I'm going to have a plus 1 minus 1. OK?

So the problem is that this thing is actually-- the constant is going to play a big role, right? Because this is essentially equal to c a plus 1b divided by c ab, where ca plus 1b is just the normalizing constant of a beta a plus 1 b. So I need to know the ratio of those two constants. And this is not something-- I mean, this is just a calculus exercise.

So in this case, what you get is-- sorry. In this case, you get-- well, OK, so we get essentially a divided by, I think, it's a plus b. Yeah, it's a plus b. So that's this quantity. OK? And when I plug in a to be this guy and b to be this guy, what I get is a plus sum of the xi. And then I get a plus this guy, a plus n minus this guy. So those two guys go away, and I'm left with 2a plus n, which does not work. No, that actually works. And so now what I do, I can actually divide and get this thing, over there.

OK. So what you can see, the reason why this thing has been divided is that you can really see that, as n goes to infinity, then this thing behaves like xn bar, which is our frequentist estimator. The effect of a is actually going away. The effect of the prior, which is completely captured by a, is going away as n goes to infinity. Is there any question? You guys have a question. What is it? Do you have a question?

AUDIENCE: Yeah, on the board, is that divided by some [INAUDIBLE] stuff?

PHILIPPE RIGOLLET: Is that divided by what?

AUDIENCE: That a over a plus b, and then you just expanded--

PHILIPPE RIGOLLET: Oh yeah, yeah, then I said that this is equal to this, right. So that's for a becomes a plus sum of the xi's, and b becomes a plus n minus sum of the xi's. OK. So that's just for the posterior one.

AUDIENCE: What's [INAUDIBLE]

PHILIPPE RIGOLLET: This guy?

AUDIENCE: Yeah.

PHILIPPE RIGOLLET: 2a.

AUDIENCE: 2a. Oh, OK.

PHILIPPE RIGOLLET: Right. So I get a plus a plus n. And then those two guys cancel. OK? And that's what you have here. So for a is equal to 1/2-- and I claim that this is Jeffreys prior. Because remember, Jeffreys was [INAUDIBLE] was square root and was proportional to the square root of p1 minus p, which I can write as p to the 1/2, 1 minus p to the 1/2. So it's just the case a is equal to 1/2. OK. So if I use Jeffreys prior, I just plug in a equals to 1/2, and this is what I get. OK?

So those things are going to have an impact again when n is moderately large. For large n, those things, whether you take Jeffreys prior or you take whatever a you prefer, it's going to have no impact whatsoever. But n is of the order of 10 maybe, then you're going to start to see some impact, depending on what a you want to pick. OK.

And then in the second example, well, here we actually computed the posterior to be this guy. Well, here, I can just read off what the expectation is, right? I mean, I don't have to actually compute the expectation of a Gaussian. It's just that xn bar. And so in this case, there's actually no-- I mean, when I have a non-informative prior for a Gaussian, then I have basically xn in bar.

As you can see, actually, this is an interesting example. When I actually look at the posterior, it's not something that cost me a lot to communicate to you, right? There's one symbol here, one symbol here, and one symbol here. I tell you the posterior is a Gaussian with mean xn bar and variance 1/n. When I actually turn that into a poster mean, I'm dropping all this information. I'm just giving you the first parameter.

So you can see there's actually much more information in the posterior than there is in the posterior mean. The posterior mean is just a point. It's not telling me how confident I am in this point. And this thing is actually very interesting. OK.

So you can talk about the posterior variance that's associated to it, right? You can talk about, as an output, you could give the posterior mean and posterior variance. And those things are actually interesting. All right.

So I think this is it. So as I said, in general, just like in this case, the impact of the prior is being washed away as the sample size goes to infinity. Just well, like here, there's no impact of the prior. It was an noninvasive one. But if you actually had an informative one, [? CF ?] homework-- yeah?

AUDIENCE: [INAUDIBLE]

PHILIPPE RIGOLLET: Yeah, so [? CF ?] homework, you would actually see an impact of the prior, which, again, would be washed away as your sample size increases. Here, it goes away. You just get xn bar over 1. And actually, in these cases, you see that the posterior distribution converges to-- sorry, the Bayesian estimator is asymptotically normal. This is different from the distribution of the posterior, right? This is just the posterior mean, which happens to be asymptotically normal. But the posterior may not have a-- I mean, here, the posterior is a beta, right? I mean, it's not normal. OK, so there's different-- those things are two different things. Your question?

AUDIENCE: What was the prior [INAUDIBLE]

PHILIPPE RIGOLLET: All 1, right? That was the improper prior.

AUDIENCE: OK. And so that would give you the same thing as [INAUDIBLE], not just the proportion.

PHILIPPE RIGOLLET: Well, I mean, yeah. So it's essentially telling you that-- so we said that, when you have a non-informative prior, essentially, the maximum likelihood is the maximum a posteriori, right? But in this case, there's so much symmetry, that it just so happens that the maximum in this thing is completely symmetric around its maximum. So it means that the expectation is equal to the maximum, to [INAUDIBLE] max. Yeah?

AUDIENCE: I read somewhere that one of the issues with Bayesian methods is that we choose the wrong prior, and it could mess up your results.

PHILIPPE RIGOLLET: Yeah, but hence, do not pick the wrong prior. I mean, of course, it would. I mean, it would mess up your res-- of course. I mean, you're putting extra information. But you could say the same thing by saying, well, the issue with frequentist method is that, if you mess up the choice of your likelihood, then it's going to mess up your output.

So here, you just have two chances of messing it up, right? You have the-- well, it's gone. So you have the product of the likelihood and the prior, and you have one more chance to-- but it's true, if you assume that the model is right, then, of course, finding the wrong prior could completely mess up things if your prior, for example, has no support on the true parameter. But if your prior has a positive weight on the true parameter as n goes to infinity-- I mean, OK, I cannot speak for all counterexamples in the world. But I'm sure, under minor technical conditions, you can guarantee that your posterior mean is going to converge to what you need it to converge to. Any other question?

All right. So I think this closes the more traditional mathematical-- not mathematical, but traditional statistics part of this class. And from here on, we'll talk about more multivariate statistics, starting with principal component analysis. So that's more like when you have multiple data. We started, in a way, to talk about multivariate statistics when we talked about multivariate regression. But we'll move on to principal component analysis.

I'll talk a bit about multiple testing. I haven't made my mind yet about what we'll talk really in December. But I want to make sure that you have a taste and a flavor of what is being interesting in statistics these days, especially as you go towards more [INAUDIBLE] learning type of questions, where really, the focus is on prediction rather than the modeling itself. We'll talk about logistic regression, as well, for example, which is generalized linear models, which is just the generalization in the case that y does not take value in the whole real line, maybe 0,1, for example, for regression. All right. Thanks.