**MITOCW | watch?v=bFZ-0FH5hfs**

**PHILIPPE RIGOLLET:** So today WE'LL actually just do a brief chapter on Bayesian statistics. And there's entire courses on Bayesian statistics, there's entire books on Bayesian statistics, there's entire careers in Bayesian statistics. So admittedly, I'm not going to be able to do it justice and tell you all the interesting things that are happening in Bayesian statistics. But I think it's important as a statistician to know what it is, how it works, because it's actually a weapon of choice for many practitioners. And because it allows them to incorporate their knowledge about a problem in a fairly systematic manner. So if you look at like, say the Bayesian statistics literature, it's huge. And so here I give you sort of a range of what you can expect to see in Bayesian statistics from your second edition of a traditional book, something that involves computation, some things that involve risk thinking.

And there's a lot of Bayesian thinking. There's a lot of things that you know talking about sort of like philosophy of thinking Bayesian. This book, for example, seems to be one of them. This book is definitely one of them. This one represents sort of a wide, a broad literature on Bayesian statistics, for applications for example, in social sciences. But even in large scale machine learning, there's a lot of Bayesian statistics happening, particular using something called Bayesian parametrics, or hierarchical Bayesian modeling. So we do have some experts at MIT in the c-cell.

Tamara Broderick for example, is a person who does quite a bit of interesting work on Bayesian parametrics. And if that's something you want to know more about, I urge you to go and talk to her. So before we go into more advanced things, we need to start with what is the Bayesian approach. What do Bayesians do, and how is it different from what we've been doing so far?

So to understand the difference between Bayesians and what we've been doing so far is, we need to first put a name on what we've been doing so far. It's called frequentist statistics. Which usually Bayesian versus frequentist statistics, by versus I don't mean that there is naturally in opposition to them. Actually, often you will see the same method that comes out of both approaches.

So let's see how we did it, right. The first thing, we had data. We observed some data. And we assumed that this data was generated randomly. The reason we did that is because this would allow us to leverage tools from probability. So let's say by nature, measurements, you do a survey, you get some data. Then we made some assumptions on the data generating process.

For example, we assumed they were iid. That was one of the recurring things. Sometimes we assume it was Gaussian. If you wanted to use say, T-test. Maybe we did some nonparametric statistics. We assume it was a smooth function or maybe linear regression function. So those are our modeling. And this was basically a way to say, well, we're not going to allow for any distributions for the data that we have. But maybe a small set of distributions that indexed by some small parameters, for example. Or at least remove some of the possibilities. Otherwise, there's nothing we can learn.

And so for example, this was associated to some parameter of interest, say data or beta in the regression model. Then we had this unknown problem and this unknown thing, a known parameter. And we wanted to find it. We wanted to either estimate it or test it, or maybe find a confidence interval for the subject.

So, so far I should not have said anything that's new. But this last sentence is actually what's going to be different from the Bayesian part. And particular, this unknown but fixed things is what's going to be changing. In the Bayesian approach, we still assume that we observe some random data. But the generating process is slightly different. It's sort of a two later process. And there's one process that generates the parameter and then one process that, given this parameter generates the data.

So what the first layer does, nobody really believes that there's some random process that's happening, about generating what is going to be the true expected number of people who turn their head to the right when they kiss. But this is actually going to be something that brings us some easiness for us to incorporate what we call prior belief. We'll see an example in a second.

But often, you actually have prior belief of what this parameter should be. When we, say least squares, we looked over all of the vectors in all of R to the p, including the ones that have coefficients equal to 50 million. Those are things that we might be able to rule out. We might be able to rule out that on a much smaller scale.

For example, well I'm not an expert on turning your head to the right or to the left. But maybe you can rule out the fact that almost everybody is turning their head in the same direction, or almost everybody is turning their head to another direction.

So we have this prior belief. And this belief is going to play say, hopefully less and less important role as we collect more and more data. But if we have a smaller amount of data, we might want to be able to use this information, rather than just shooting in the dark. And so the idea is to have this prior belief. And then, we want to update this prior belief into what's called the posterior belief after we've seen some data.

Maybe I believe that there's something that should be in some range. But maybe after I see data, it's comforting me in my beliefs. So I'm actually having maybe a belief that's more. So belief encompasses basically what you think and how strongly you think about it. That's what I call belief.

So for example, if I have a belief about some parameter theta, maybe my belief is telling me where theta should be and how strongly I believe in it, in the sense that I have a very narrow region where theta could be. The posterior beliefs, as well, you see some data. And maybe you're more confident or less confident about what you've seen. Maybe you've shifted your belief a little bit. And so that's what we're going to try to see, and how to do this in a principal manner.

To understand this better, there's nothing better than an example. So let's talk about another stupid statistical question. Which is, let's try to understand p. Of course, I'm not going to talk about politics from now on. So let's talk about p, the proportion of women in the population. And so what I could do is to collect some data, X1, Xn and assume that they're Bernoulli with some parameter, p unknown.

So p is in 0, 1. OK, let's assume that those guys are iid. So this is just an indicator for each of my collected data, whether the person I randomly sample is a woman, I get a one. If it's a man, I get a zero. Now the question is, I sample these people randomly. I do you know their gender. And the frequentist approach was just saying, OK, let's just estimate p hat being Xn bar. And then we could do some tests.

So here, there's a test. I want to test maybe if p is equal to 0.5 or not. That sounds like a pretty reasonable thing to test. But we want to also maybe estimate p. But here, this is a case where we definitely prior belief of what p should be. We are pretty confident that p is not going to be 0.7. We actually believe that we should be extremely close to one half, but maybe not exactly.

Maybe this population is not the population in the world. But maybe this is the population of, say some college and we want to understand if this college has half women or not. Maybe we know it's going to be close to one half, but maybe we're not quite sure. We're going to want to integrate that knowledge.

So I could integrate it in a blunt manner by saying, discard the data and say that p is equal to one half. But maybe that's just a little too much. So how do I do this trade off between adding the data and combining it with this prior knowledge?

In many instances, essentially what's going to happen is this one half is going to act like one new observation. So if you have five observations, this is just the sixth observation, which will play a role. If you have a million observations, you're going to have a million and one. It's not going to play so much of a role. That's basically how it goes. But, definitely not always because we'll see that if I take my prior to be a point minus one half here, it's basically as if I was discarding my data.

So essentially, there's also your ability to encompass how strongly you believe in this prior. And if you believe infinitely more in the prior than you believe in the data you collected, then it's not going to act like one more observation. The Bayesian approach is a tool to one, include mathematically our prior. And our prior belief into statistical procedures. Maybe I have this prior knowledge. But if I'm a medical doctor, it's not clear to me how I'm going to turn this into some principal way of building estimators.

And the second goal is going to be to update this prior belief into a posterior belief by using the data. How do I do this? And at some point, I sort of suggested that there's two layers. One is where you draw the parameter at random. And two, once you have the parameter, conditionless parameter, you draw your data.

Nobody believed this actually is happening, that nature is just rolling dice for us and choosing parameters at random. But what's happening is that, this idea that the parameter comes from some random distribution actually captures, very well, this idea that how you would encompass your prior. How would you say, my belief is as follows?

Well here's an example about p. I'm 90% sure that p is between 0.4 and 0.6. And I'm 95% sure that p is between 0.3 and 0.8. So essentially, I have this possible value of p. And what I know is that, there's 90% here between 0.4 and 0.6. And then I have 0.3 and 0.8. And I know that I'm 95% sure that I'm in here.

If you remember, this sort of looks like the kind of pictures that I made when I had some Gaussian, for example. And I said, oh here we have 90% of the observations. And here, we have 95% of the observations. So in a way, if I were able to tell you all those ranges for all possible values, then I would essentially describe a probability distribution for p.

And what I'm saying is that, p is going to have this kind of shape. So of course, if I tell you only two twice this information that there's 90% I'm here, and I'm between here and here. And 95%, I'm between here and here, then there's many ways I can accomplish that, right. I could have something that looks like this, maybe. It could be like this. There's many ways I can have this.

Some of them are definitely going to be mathematically more convenient than others. And hopefully, we're going to have things that I can parameterize very well. Because if I tell you this is this guy, then there's basically one, two three, four, five, six, seven parameters. So I probably don't want something that has seven parameters. But maybe I can say, oh, it's a Gaussian and I all I have to do is to tell you where it's centered and what the standard deviation is.

So the idea of using this two layer thing, where we think of the parameter p as being drawn from some distribution, is really just a way for us to capture this information. Our prior belief being, well there's this percentage of chances that it's there. But the percentage of this chance, I'm not I'm deliberately not using probability here. So it's really a way to get close to this.

That's why I say, the true parameter is not random. But the Bayesian approach does as if it was random. And then, just spits out a procedure out of this thought process, this thought experiment. So when you practice Bayesian statistics a lot, you start getting automatisms. You start getting some things that you do without really thinking about it. just like when you you're a statistician, the first thing you do is, can I think of this data as being Gaussian for example?

When you're Bayesian you're thinking about, OK I have a set of parameters. So here, I can describe my parameter as being theta in general, in some big space parameter of theta. But what spaces did we encounter? Well, we encountered the real line. We encountered the interval 0, 1 for Bernoulli's And we encountered some of the positive real line for exponential distributions, etc.

And so what I'm going to need to do, if I want to put some prior on those spaces, I'm going to have to have a usual set of tools for this guy, usual set of tools for this guy, usual sort of tools for this guy. And by usual set of tools, I mean I'm going to have to have a family of distributions that's supported on this.

So in particular, this is the speed in which my parameter that I usually denote by p for Bernoulli lives. And so what I need is to find a distribution on the interval 0, 1 just like this guy. The problem with the Gaussian is that it's not on the interval 0, 1. It's going to spill out in the end. And it's not going to be something that works for me.

And so the question is, I need to think about distributions that are probably continuous. Why would I restrict myself to discrete distributions that are actually convenient and for Bernoulli, one that's actually basically the main tool that everybody is using is the so-called beta distribution.

So the beta distribution has two parameters. So x follows a beta with parameters a and b if it has a density, f of x is equal to x to the a minus 1. 1 minus x to the b minus 1, if x is in the interval 0, 1 and 0 for all other x's. OK?

Why is that a good thing? Well, it's a density that's on the interval 0, 1 for sure. But now I have these two parameters and a set of shapes that I can get by tweaking those two parameters is incredible. It's going to be a unimodal distribution. It's still fairly nice. It's not going to be something that goes like this and this.

Because if you think about this, what would it mean if your prior distribution of the interval 0, 1 had this shape? It would mean that, maybe you think that p is here or maybe you think that p is here, or maybe you think that p is here. Which essentially means that you think that p can come from three different phenomena. And there's other models that are called mixers for that, that directly account for the fact that maybe there are several phenomena that are aggregated in your data set.

But if you think that your data set is sort of pure, and that everything comes from the same phenomenon, you want something that looks like this, or maybe looks like this, or maybe is sort of symmetric. You want to get all this stuff. Maybe you want something that says, well if I'm talking about p being the probability of the proportion of women in the whole world, you want something that's probably really spiked around one half. Almost the point math, because you know let's agree that 0.5 is the actual number.

So you want something that says, OK maybe I'm wrong. But I'm sure I'm not going to be really that way off. So you want something that's really pointy. But if it's something you've never checked, and again I can not make references at this point, but something where you might have some uncertainty that should be around one half. Maybe you want something that a little more allows you to say, well, I think there's more around one half. But there's still some fluctuations that are possible.

And in particular here, I talk about p, where the two parameters a and b are actually the same. I call them a. One is called scale. The other one is called shape.

Oh sorry, this is not a density. So it actually has to be normalized. When you integrate this guy, it's going to be some function that depends on a and b, actually depends on this function through the beta function. Which is this combination of gamma function, so that's why it's called beta distribution. That's the definition of the beta function when you integrate this thing anyway. You just have to normalize it. That's just a number that depends on the a and b.

So here, if you take a equal to b, you have something that essentially is symmetric around one half. Because what does it look like? Well, so my density f of x, is going to be what? It's going to be my constant times x, times one minus x to a minus one. And this function, x times 1 minus x looks like this.

We've drawn it before. That was something that showed up as being the variance of my Bernoulli. So we know it's something that takes its maximum at one half. And now I'm just taking a power of this guy. So I'm really just distorting this thing into some fairly symmetric manner.

This distribution that we actually take for p. I assume that p, the parameter, notice that this is kind of weird. First of all, this is probably the first time in this entire course that something has a distribution when it's actually a lower case letter. That's something you have to deal with, because we've been using lower case letters for parameters. And now we want them to have a distribution. So that's what's going to happen.

This is called the prior distribution. So really, I should write something like f of p is equal to a constant times p, 1 minus p, to the n minus 1. Well no, actually I should not because then it's confusing.

One thing in terms of notation that I'm going to write, when I have a constant here and I don't want to make it explicit. And we'll see in a second why I don't need to make it explicit. I'm going to write this as f of x is proportional to x 1 minus x to the n minus 1. That's just to say, equal to some constant that does not depend on x times this thing.

So if we continue with our experiment where I'm drawing this data, X1 to Xn, which is Bernoulli p, if p has some distribution it's not clear what it means to have a Bernoulli with some random parameter. So what I'm going to do is, then I'm going to first draw my p. Let's say I get a number, 0.52. And then, I'm going to draw my data conditionally on p. So here comes the first and last flowchart of this class.

So nature first draws p. p follows some data on a, a. Then I condition on p. And then I draw X1, Xn that are iid, Bernoulli p. Everybody understand the process of generating this data? So you first draw a parameter, and then you just flip those independent biased coins with this particular p. There's this layered thing.

Now conditionally p, right so here I have this prior about p which was the thing. So this is just the thought process again, it's not anything that actually happens in practice. This is my way of thinking about how the data was generated. And from this, I'm going to try to come up with some procedure.

Just like, if your estimator is the average of the data, you don't have to understand probability to say that my estimator is the average of the data. Anyone outside this room understands that the average is a good estimator for some average behavior. And they don't need to think of the data as being a random variable, et cetera. So same thing, basically.

In this case, you can see that the posterior distribution is still a beta. What it means is that, I had this thing. Then, I observed my data. And then, I continue and here I'm going to update my prior into some posterior distribution, pi. And here, this guy is actually also a beta. My posterior distribution, p, is also a beta distribution with the parameters that are on this slide. And I'll have the space to reproduce them.

So I start the beginning of this flowchart as having p, which is a prior. I'm going to get some observations and then, I'm going to update what my posterior is. This posterior is basically something that's, in business statistics was beautiful is as soon as you have this distribution, it's essentially capturing all the information about the data that you want for p.

And it's not just the point. It's not just an average. It's actually an entire distribution for the possible values of theta. And it's not the same thing as saying, well if theta hat is equal to Xn bar, in the Gaussian case I know that this is some mean, mu. And then maybe it has varying sigma squared over n. That's not what I mean by, this is my posterior distribution. This is not what I mean.

This is going to come from this guy, the Gaussian thing and the central limit theorem. But what I mean is this guy. And this came exclusively from the prior distribution. If I had another prior, I would not necessarily have a beta distribution on the output.

So when I have the same family of distributions at the beginning and at the end of this flowchart, I say that beta is a conjugate prior. Meaning I put in beta as a prior and I get beta as [INAUDIBLE] And that's why betas are so popular.

Conjugate priors are really nice, because you know that whatever you put in, what you're going to get in the end is a beta. So all you have to think about is the parameters. You don't have to check again what the posterior is going to look like, what the PDF of this guy is going to be. You don't have to think about it. You just have to check what the parameters are.

And there's families of conjugate priors. Gaussian gives Gaussian, for example. There's a bunch of them. And this is what drives people into using specific priors as opposed to others. It has nice mathematical properties. Nobody believes that p is really distributed according to beta. But it's flexible enough and super convenient mathematically.

Now let's see for one second, before we actually go any further. I didn't mention A and B are both in here, A and B are both positive numbers. They can be anything positive. So here what I did is that, I updated A into a plus the sum of my data, and b into b plus n minus the sum of my data. So that's essentially, a becomes a plus the number of ones. Well, that's only when I have a and a. So the first parameters become itself plus the number of ones. And the second one becomes itself plus the number of zeros.

And so just as a sanity check, what does this mean? If a it goes to zero, what is the beta when a goes to 0? We can actually read this from here. Actually, let's take a goes to-- no. Sorry, let's just do this. I'll do it when we talk about non-informative prior, because it's a little too messy.

How do we do this? How did I get this posterior distribution, given the prior? How do I update This well this is called Bayesian statistics. And you've heard this word, Bayes before. And the way you've heard it is in the Bayes formula. What was the Bayes formula? The Bayes formula was telling you that the probability of A, given B was equal to something that depended on the probability of B, given A. That's what it was.

You can actually either remember the formula or you can remember the definition. And this is what p of A and B divided by p of B. So this is p of B, given A times p of A divided by p of B. That's what Bayes formula is telling you. Agree?

So now what I want is to have something that's telling me how this is going to work. What is going to play the role of those events, A and B? Well one is going to be, this is going to be the distribution of my parameter of theta, given that I see the data. And this is going to tell me, what is the distribution of the data, given that I know what my parameter if theta is.

But that part, if this is theta and this is the parameter of theta, this is what we've been doing all along. The distribution of the data, given the parameter here was n iid Bernoulli p. I knew exactly what their joint probability mass function is.

Then, that was what? So we said that this is going to be my data and this is going to be my parameter. So that means that, this is the probability of my data, given the parameter. This is the probability of the parameter. What is this? What did we call this? This is the prior. It's just the distribution of my parameter.

Now what is this? Well, this is just the distribution of the data, itself. This is essentially the distribution of this, if this was indeed not conditioned on p. So if I don't condition on p, this data is going to be a bunch of iid, Bernoulli with some parameter. But the perimeter is random, right.

So for different realization of this data set, I'm going to get different parameters for the Bernoulli. And so that leads to some sort of convolution. It's not really a convolution in this case, but it's like some sort of composition of distributions. I have the randomness that comes from here and then, the randomness that comes from realizing the Bernoulli.

That's just the marginal distribution. It actually might be painful to understand what this is, right. In a way, it's sort of a mixture and it's not super nice. But we'll see that this actually won't matter for us. This is going to be some number. It's going to be there. But it will matter for us, what it is. Because it actually does not depend on the parameter. And that's all that matters to us.

Let's put some names on those things. This was very informal. So let's put some actual names on what we call prior. So what is the formal definition of a prior, what is the formal definition of a posterior, and what are the rules to update it?

So I'm going to have my data, which is going to be X1, Xn. Let's say they are iid, but they don't actually have to. And so I'm going to have given, theta. And when I say given, it's either given like I did in the first part of this course in all previous chapters, or conditionally on. If you're thinking like a Bayesian, what I really mean is conditionally on this random parameter. It's as if it was a fixed number.

They're going to have a distribution, X1, Xn is going to have some distribution. Let's assume for now it's a PDF, pn of X1, Xn. I'm going to write theta like this. So for example, what is this? Let's say this is a PDF. It could be a PMF. Everything I say, I'm going to think of them as being PDF's. I'm going to combine PDF's with PDF's, but I could combine PDF it PMF, PMF with PDF's or PMF with PMF. So everywhere you see a D could be an M.

Now I have those things. So what does it mean? So here is an example. X1, Xn or iid, and theta 1. Now I know exactly what the joint PDF of this thing is. It means that pn of X1, Xn given theta is equal to what? Well it's 1 over 2pi to the power n e, to the minus sum from i equal 1 to n of xi minus theta squared divided by 2. So that's just the joint distribution of n iid and theta 1, random variables.

That's my pn given theta. Now this is what we denoted by f sub theta before. We had the subscript before, but now we just put a bar in theta because we want to remember that this is actually conditioned on theta. But this is just notation. You should just think of this as being, just the usual thing that you get from some statistical model.

Now, that's going to be pn. Theta has prior distribution, pi. For example, so think of it as either PDF or PMF again. For example, pi of theta was what? Well it was some constant times theta to the a minus 1, 1 minus theta to a minus 1. So it has some prior distribution, and that's another PMF.

So now I'm given the distribution of my, x is given theta and given the distribution of my theta. I'm given this guy. That's this guy. I'm given that guy, which is my pi. So that's my pn of X1, Xn given theta. That's my pi of theta. Well, this is just the integral of pn of X1, Xn times pi of theta, d theta, over all possible sets of theta. That's just when I integrate out my theta, or I compute the marginal distribution, I did this by integrating. That's just basic probability, conditional probabilities. Then if I had the PMF, I would just sum over the values of thetas.

Now what I want is to find what's called, so that's the prior distribution, and I want to find the posterior distribution. It's pi of theta, given X1, Xn. If I use Bayes' rule I know that this is pn of X1, Xn, given theta times pi of theta. And then it's divided by the distribution of those guys, which I will write as integral over theta of pn, X1, Xn, given theta times pi of theta, d theta.

Everybody's with me, still? If you're not comfortable with this, it means that you probably need to go read your couple of pages on conditional densities and conditional PMF's from your probably class. There's really not much there. It's just a matter of being able to define those quantities, f density of x, given y. This is just what's called a conditional density. You need to understand what this object is and how it relates to the joint distribution of x and y, or maybe the distribution of x or the distribution of y.

But it's the same rules. One way to actually remember this is, this is exactly the same rules as this. When you see a bar, it's the same thing as the probability of this and this guy. So for densities, it's just a comma divided by the second the probably the second guy. That's it. So if you remember this, you can just do some pattern matching and see what I just wrote here.

Now, I can compute every single one of these guys. This something I get from my modeling. So I did not write this. It's not written in the slides. But I give a name to this guy that was my prior distribution. And that was my posterior distribution. In chapter three, maybe what did we call this guy? The one that does not have a name and that's in the box. What did we call it?

**AUDIENCE:** [INAUDIBLE]

**PHILLIPE RIGOLLET:** It is the joint distribution of the Xi's. And we gave it a name.

**AUDIENCE:** [INAUDIBLE]

**PHILLIPE RIGOLLET:** It's the likelihood, right? This is exactly the likelihood. This was the likelihood of theta. And this is something that's very important to remember, and that really reminds you that these things are really not that different. Maximum likelihood estimation and Bayesian estimation, because your posterior is really just your likelihood times something that's just putting some weights on the thetas, depending on where you think theta should be.

If I had, say a maximum likelihood estimate, and my likelihood and theta looked like this, but my prior and theta looked like this. I said, oh I really want thetas that are like this. So what's going to happen is that, I'm going to turn this into some posterior that looks like this. So I'm just really waiting, this posterior, this is a constant that does not depend on theta right? Agreed? I integrated over theta, so theta is gone. So forget about this guy.

I have basically, that the posterior distribution up to scaling, because it has to be a probability density and not just anything any function that's positive, is the product of this guy. It's a weighted version of my likelihood. That's all it is. I'm just weighing the likelihood, using my prior belief on theta. And so given this guy a natural estimator, if you follow the maximum likelihood principle, would be the maximum of this posterior. Agreed? That would basically be doing exactly what maximum likelihood estimation is telling you.

So it turns out that you can. It's called Maximum A Posteriori, and I won't talk much about this, or MAP. That's Maximum a Posteriori. So it's just the theta hat is the arc max of pi theta, given X1, Xn. And it sounds like it's OK. I'll give you a density and you say, OK I have a density for all values of my parameters. You're asking me to summarize it into one number. I'm just going to take the most likely number of those guys.

But you could summarize it, otherwise. You could take the average. You could take the median. You could take a bunch of numbers. And the beauty of Bayesian statistics is that, you don't have to take any number in particular. You have an entire posterior distribution. This is not only telling you where theta is, but it's actually telling you the difference if you actually give as something that gives you the posterior.

Now, let's say the theta is p between 0 and 1. If my posterior distribution looks like this, or my posterior distribution looks like this, then those two guys have one, the same mode. This is the same value. And their symmetric, so they'll also have the same mean. So these two posterior distributions give me the same summary into one number. However clearly, one is much more confident than the other one. So I might as well just spit it out as a solution.

You can do even better. People actually do things, such as drawing a random number from this distribution. Say, this is my number. That's kind of dangerous, but you can imagine you could do this.

This is what works. That's what we went through. So here, as you notice I don't care so much about this part here. Because it does not depend on theta. So I know that given the product of those two things, this thing is only the constant that I need to divide so that when I integrate this thing over theta, it integrates to one. Because this has to be a probability density on theta. I can write this and just forget about that part. And that's what's written on the top of this slide.

This notation, this sort of weird alpha, or I don't know. Infinity sign propped to the right. Whatever you want to call this thing is actually just really emphasizing the fact that I don't care. I write it because I can, but you know what it is. In some instances, you have to compute the integral. In some instances, you don't have to compute the integral.

And a lot of Bayesian computation is about saying, OK it's actually really hard to compute this integral, so I'd rather not doing it. So let me try to find some methods that will allow me to sample from the posterior distribution, without having to compute this. And that's what's called Monte-Carlo Markov chains, or MCMC, and that's exactly what they're doing. They're just using only ratios of things, like that for different thetas. And which means that if you take ratios, the normalizing constant is gone and you don't need to find this integral.

So we won't go into those details at all. That would be the purpose of an entire course on Bayesian inference. Actually, even Bayesian computations would be an entire course on its own. And there's some very interesting things that are going on there, the interface of stats and computation.

So let's go back to our example and see if we can actually compute any of those things. Because it's very nice to give you some data, some formulas. Let's see if we can actually do it. In particular, can I actually recover this claim that the posterior associated to a beta prior with a Bernoulli likelihood is actually giving me a beta again?

What was my prior? So p was following a beta AA, which means that p, the density. That was pi of theta. Well I'm going to write this as pi of p-- was proportional to p to the A minus 1 times 1 minus p to the A minus 1.

So that's the first ingredient I need to complete my posterior. I really need only two, if I wanted to bound up to constant. The second one was p hat. We've computed that many times. And we had even a nice compact way of writing it, which was that pn of X1, Xn, given the parameter p. So the joint density of my data, given p, that's my likelihood. The likelihood of p was what? Well it was p to the sum of Xi's. 1 minus p to the n minus some of the Xi's.

Anybody wants me to parse this more? Or do you remember seeing that from maximum likelihood estimation? Yeah?

**AUDIENCE:** [INAUDIBLE]

**PHILLIPE RIGOLLET:** That's what conditioning does.

**AUDIENCE:** [INAUDIBLE] previous slide. [INAUDIBLE] bottom there, it says D pi of t. Shouldn't it be dt pi of t?

**PHILLIPE RIGOLLET:** So D pi of T is a measure theoretic notation, which I used without thinking. And I should not because I can see it upsets you. D pi of T is just a natural way to say that I integrate against whatever I'm given for the prior of theta. In particular, if theta is just the mix of a PDF and a point mass, maybe I say that my p takes value 0.5 with probability 0.5. And then is uniform on the interval with probability 0.5.

For this, I neither have a PDF nor a PMF. But I can still talk about integrating with respect to this, right? It's going to look like, if I take a function f of T, D pi of T is going to be one half of f of one half. That's the point mass with probability one half, at one half. Plus one half of the integral between 0 and 1, of f of TDT.

This is just the notation, which is actually funnily enough, interchangeable with pi of DT. But if you have a density, it's really just the density pi of TDT. If pi is really a density, but that's when it's when pi is and measure and not a density. Everybody else, forget about this. This is not something you should really worry about at this point. This is more graduate level probability classes. But yeah, it's called measure theory. And that's when you think of pi as being a measure in an abstract fashion. You don't have to worry whether it's a density or not, or whether it has a density.

So everybody is OK with this?

Now I need to compute my posterior. And as I said, my posterior is really just the product of the likelihood weighted by the prior. Hopefully, at this stage of your application, you can multiply two functions. So what's happening is, if I multiply this guy with this guy, p gets this guy to the power this guy plus this guy.

And then 1 minus p gets the power n minus some of Xi's. So this is always from I equal 1 to n. And then plus A minus 1 as well. This is up to constant, because I still need to solve this. And I could try to do it. But I really don't have to, because I know that if my density has this form, then it's a beta distribution. And then I can just go on Wikipedia and see what should be the normalization factor.

But I know it's going to be a beta distribution. It's actually the beta with parameter. So this is really my beta with parameter, sum of Xi, i equal 1 to n plus A minus 1. And then the second parameter is n minus sum of the Xi's plus A minus 1. I just wrote what was here. What happened to my one? Oh no, sorry. Beta has the power minus 1. So that's the parameter of the beta. And this is the parameter of the beta.

Beta is over there, right? So I just replace A by what I see. A is just becoming this guy plus this guy and this guy plus this guy. Everybody is comfortable with this computation?

We just agreed that beta priors for Bernoulli observations are certainly convenient. Because they are just conjugate, and we know that's what is going to come out in the end. That's going to be a beta as well. I just claim it was convenient. It was certainly convenient to compute this, right? There was certainly some compatibility when I had to multiply this function by that function. And you can imagine that things could go much more wrong, than just having p to some power and p to some power, 1 minus p to some power, when it might just be some other power. Things were nice.

Now this is nice, but I can also question the following things. Why beta, for one? The beta tells me something. That's convenient, but then how do I pick A? I know that A should definitely capture the fact that where I want to have my p most likely located. But it also actually also captures the variance of my beta.

And so choosing different As is going to have different functions. If I have A and B, If I started with the beta with parameter. If I started with a B here, I would just pick up the B here. Agreed? And that would just be a symmetric. But they're going to capture mean and variance of this thing. And so how do I pick those guys?

If I'm a doctor and you're asking me, what do you think the chances of this drug working in this kind of patients is? And I have to spit out the parameters of a beta for you, it might be a bit of a complicated thing to do. So how do you do this, especially for problems? So by now, people have actually mastered the art of coming up with how to formulate those numbers.

But in new problems that come up, how do you do this? What happens if you want to use Bayesian methods, but you actually do not know what you expect to see? To be fair, before we started this class, I hope all of you had no idea whether people tend to bend their head to the right or to the left before kissing. Because if you did, well you have too much time on your hands and I should double your homework.

So in this case, maybe you still want to use the Bayesian machinery. Maybe you just want to do something nice. It's nice right, I mean it worked out pretty well. What if you want to do? Well you actually want to use some priors that carry no information, that basically do not prefer any theta to another theta.

Now, you could read this slide or you could look at this formula. We just said that this pi here was just here to weigh some thetas more than others, depending on their prior belief. If our prior belief does not want to put any preference towards some thetas than to others, what do I do?

**AUDIENCE:** [INAUDIBLE]

**PHILLIPE RIGOLLET:** Yeah, I remove it. And the way to remove something we multiply by, is just replace it by one. That's really what we're doing. If this was a constant not depending on theta, then that would mean that we're not preferring any theta. And we're looking at the likelihood. But not as a function that we're trying to maximize, but it is a function that we normalize in such a way that it's actually a distribution.

So if I have pi, which is not here, this is really just taking the like likelihood, which is a positive function. It may not integrate to 1, so I normalize it so that it integrates to 1. And then I just say, well this is my posterior distribution. Now I could just maximize this thing and spit out my maximum likelihood estimator.

But I can also integrate and find what the expectation of this guy is. I can find what the median of this guy is. I can sample data from this guy. I can build, understand what the variance of this guy is. Which is something we did not do when we just did maximum likelihood estimation because given a function, all we cared about was the arc max of this function.

These priors are called uninformative. This is just replacing this number by one or by a constant. Because it still has to be a density. If I have a bounded set, I'm just looking for the uniform distribution on this bounded set, the one that puts constant one over the size of this thing.

But if I have an invalid set, what is the density that takes a constant value on the entire real line, for example? What is this density?

**AUDIENCE:** [INAUDIBLE]

**PHILLIPE RIGOLLET:** Doesn't exist, right? It just doesn't exist. The way you can think of it is a Gaussian with the variance going to infinity, maybe, or something like this. But you can think of it in many ways. You can think of the limit of the uniform between minus T and T, with T going to infinity. But this thing is actually zero. There's nothing there.

You can actually still talk about this. You could always talk about this thing, where you think of this guy as being a constant, remove this thing from this equation, and just say, well my posterior is just the likelihood divided by the integral of the likelihood over theta. And if theta is the entire real line, so be it. As long as this integral converges, you can still talk about this stuff.

This is what's called an improper prior. An improper prior is just a non-negative function defined in theta, but it does not have to integrate neither to one, nor to anything. If I integrate the function equal to 1 on the entire real line, what do I get? Infinity. It's not a proper prior, and it's called and improper prior.

And those improper priors are usually what you see when you start to want non-informative priors on infinite sets of datas. That's just the nature of it. You should think of them as being the uniform distribution of some infinite set, if that thing were to exist.

Let's see some examples about non-informative priors. If I'm in the interval 0, 1 this is a finite set. So I can talk about the uniform prior on the interval 0, 1 for a parameter, p of a Bernoulli. If I want to talk about this, then it means that my prior is p follows some uniform on the interval 0, 1. So that means that f of x is 1 if x is in 0, 1.

Otherwise, there is actually not even a normalization. This thing integrates to 1. And so now if I look at my likelihood, it's still the same thing. So my posterior becomes theta X1, Xn. That's my posterior. I don't write the likelihood again, because we still have it-- well we don't have it here anymore. The likelihood is given here. Copy, paste over there.

The posterior is just this thing times 1. So you will see it in a second. So it's p to the power sum of the Xi's, one minus p to the power, n minus sum of the Xi's. And then it's multiplied by 1, and then divided by this integral between 0 and 1 of p, sum of the Xi's. 1 minus p, n minus sum of the Xi's. Dp, which does not depend on p. And I really don't care what the thing actually is. That's posterior of p.

And now I can see, well what is this? It's actually just the beta with parameters. This guy plus 1. And this guy plus 1. I didn't tell you what the expectation of a beta was. We don't know what the expectation of a beta is, agreed?

If I wanted to find say, the expectation of this thing that would be some good estimator, we know that the maximum of this guy-- what is the maximum of this thing? Well, it's just this thing, it's the average of the Xi's. That's just the maximum likelihood estimator for Bernoulli. We know it's the average.

Do you think if I take the expectation of this thing, I'm going to get the average? So actually, I'm not going to get the average. I'm going to get this guy plus this guy, divided by n plus 1. Let's look at what this thing is doing. It's looking at the number of ones and it's adding one. And this guy is looking at the number of zeros and it's adding one.

Why is it adding this one? What's going on here? This is going to matter mostly when the number of ones is actually zero, or the number of zeros is zero. Because what it does is just pushes the zero from non-zero.

And why is that something that this Bayesian method actually does for you automatically? It's because when we put this non-informative prior on p, which was uniform on the interval 0, 1. In particular, we know that the probability that p is equal to 0 is zero. And the probability p is equal to 1 is zero.

And so the problem is that if I did not add this 1 with some positive probability, I wouldn't be allowed to spit out something that actually had p hat, which was equal to 0. If by chance, let's say I have n is equal to 3, and I get only 0, 0, 0, that could happen with probability. 1 over pq, one over 1 minus pq. That's not something that I want.

And I'm using my priors. My prior is not informative, but somehow it captures the fact that I don't want to believe p is going to be either equal to 0 or 1. So that's sort of taken care of here. So let's move away a little bit from the Bernoulli example, shall we? I think we've seen enough of it.

And so let's talk about the Gaussian model. Let's say I want to do Gaussian inference. I want to do inference in a Gaussian model, using Bayesian methods. What I want is that Xi, X1, Xn, or say 0, 1 iid. Sorry, theta 1, iid conditionally on theta. That means that pn of X1, Xn, given theta is equal to exactly what I wrote before. So 1 square root to pi, to the n exponential minus one half sum of Xi minus theta squared.

So that's just the joint distribution of my Gaussian with mean data. And the another question is, what is the posterior distribution? Well here I said, let's use the uninformative prior, which is an improper prior. It puts weight on everyone. That's the so-called uniform on the entire real line. So that's certainly not a density. But it can still just use this. So all I need to do is get this divided by normalizing this thing.

But if you look at this, essentially I want to understand. So this is proportional to the exponential minus one half sum from I equal 1 to n of Xi minus theta squared. And now I want to see this thing as a density, not on the Xi's but on theta. What I want is a density on theta.

So it looks like I have chances of getting something that looks like a Gaussian. To have a Gaussian, I would need to see minus one half. And then I would need to see theta minus something here, not just the sum of something minus thetas. So I need to work a little bit more, to expand the square here. So this thing here is going to be equal to exponential minus one half sum from I equal 1 to n of Xi squared minus 2Xi theta plus theta squared.

Now what I'm going to do is, everything remember is up to this little sign. So every time I see a term that does not depend on theta, I can just push it in there and just make it disappear. Agreed? This term here, exponential minus one half sum of Xi squared, does it depend on theta? No. So I'm just pushing it here. This guy, yes. And the other one, yes.

So this is proportional to exponential sum of the Xi. And then I'm going to pull out my theta, the minus one half canceled with the minus 2. And then I have minus one half sum from I equal 1 to n of theta squared. Agreed?

So now what this thing looks like, this looks very much like some theta minus something squared. This thing here is really just n over 2 times theta. Sorry, times theta squared. So now what I need to do is to write this of the form, theta minus something. Let's call it mu, squared, divided by 2 sigma squared.

I want to turn this into that, maybe up to terms that do not depend on theta. That's what I'm going to try to do. So that's called completing the squaring. That's some exercises you do. You've done it probably, already in the homework. And that's something you do a lot when you do Bayesian statistics, in particular.

So let's do this. What is it going to be the leading term? Theta squared is going to be multiplied by this thing. So I'm going to pull out my n over 2. And then I'm going to write this as minus theta over 2. And then I'm going to write theta minus something squared. And this something is going to be one half of what I see in the cross-product.

I need to actually pull this thing out. So let me write it like that first. So that's theta squared. And then I'm going to write it as minus 2 times 1 over n sum from I equal 1 to n of Xi's times theta. That's exactly just a rewriting of what we had before. And that should look much more familiar. A squared minus 2 blap A, and then I missed something.

So this thing, I'm going to be able to rewrite as theta minus Xn bar squared. But then I need to remove the square of Xn bar. Because it's not here. So I just complete the square. And then I actually really don't care with this thing actually was, because it's going to go again in the little Alpha's sign over there. So this thing eventually is going to be proportional to exponential of minus n over 2 times theta of minus Xn bar squared.

And so we know that if this is a density that's proportional to this guy, it has to be some n with mean, Xn bar. And variance, this is supposed to be 1 over sigma squared. This guy over here, this n. So that's really just 1 over n.

So the posterior distribution is a Gaussian centered at the average of my observations. And with variance, 1 over n. Everybody's with me? Why I'm saying this, this was the output of some computation. But it sort of makes sense, right? It's really telling me that the more observations I have, the more concentrated this posterior is. Concentrated around what? Well around this Xn bar.

That looks like something we've sort of seen before. But it does not have the same meaning, somehow. This is really just the posterior distribution. It's sort of a sanity check, that I have this 1 over n when I have Xn bar. But it's not the same thing as saying that the variance of Xn bar was 1 over n, like we had before.

As an exercise, I would recommend if you don't get it, just try pi of theta to be equal to some n mu 1. Here, the prior that we used was completely non-informative. What happens if I take my prior to be some Gaussian, which is centered at mu and it has the same variance as the other guys?

So what's going to happen here is that we're going to put a weight. And everything that's away from mu is going to actually get less weight. I want to know how I'm going to be updating this prior into a posterior. Everybody sees what I'm saying here? So that means that pi of theta has the density proportional to exponential minus one half theta minus mu squared.

So I need to multiply my posterior with this, and then see. It's actually going to be a Gaussian. This is also a conjugate prior. It's going to spit out another Gaussian. You're going to have to complete a square again, and just check what it's actually giving you. And so spoiler alert, it's going to look like you get an extra observation, which is actually equal to mu.

It's going to be the average of n plus 1 observations. The first n1's being X1 to Xn. And then, the last one being mu. And it sort of makes sense. That's actually a fairly simple exercise. Rather than going into more computation, this is something you can definitely do when you're in the comfort of your room.

I want to talk about other types of priors. The first thing I said is, there's this beta prior that I just pulled out of my hat and that was just convenient. Then there was this non-informative prior. It was convenient. It was non-informative, so if you don't know anything else maybe that's what you want to do.

The question is, are there any other priors that are sort of principled and generic, in the sense that the uninformative prior was generic, right? It was equal to 1, that's as generic as it gets. So is there anything that's generic as well? Well, there's this priors that are called Jeffrey's priors.

And Jeffrey's prior, which is proportional to square root of the determinant of the Fisher information of theta. This is actually a weird thing to do. It says, look at your model. Your model is going to have a Fisher information. Let's say it exists. Because we know it does not always exist. For example, in the multinomial model, we didn't have a Fisher information.

The determinant of a matrix is somehow measuring the size of a matrix. If you don't trust me, just think about the matrix being of size one by one, then the determinant is just the number that you have there. And so this is really something that looks like the Fisher information. It's proportional to the amount of information that you have at a certain point.

And so what my prior is saying well, I want to put more weights on those thetas that are going to just extract more information from the data. You can actually compute those things. In the first example, Jeffrey's prior is something that looks like this. In one dimension, Fisher information is essentially one the word variance. That's just 1 over the square root of the variance, because I have the square root.

And when I have the Jeffrey's prior, when I have the Gaussian case, this is the identity matrix that I would have in the Gaussian case. The determinant of the identities is 1. So square root of 1 is 1, and so I would basically get 1. And that gives me my improper prior, my uninformative prior that I had.

So the uninformative prior 1 is fine. Clearly, all the thetas carry the same information in the Gaussian model. Whether I translate it here or here, it's pretty clear none of them is actually better than the other. But clearly for the Bernoulli case, the p's that are closer to the boundary carry more information.

I sort of like those guys, because they just carry more information. So what I do is, I take this function. So p1 minus p. Remember, it's something that looks like this. On the interval 0, 1. This guy, 1 over square root of p1 minus p is something that looks like this. Agreed

What it's doing is sort of wants to push towards the piece that actually carry more information. Whether you want to bias your data that way or not, is something you need to think about. When you put a prior on your data, on your parameter, you're sort of biasing towards this idea your data.

That's maybe not such a good idea, when you have some p that's actually close to one half, for example. You're actually saying, no I don't want to see a p that's close to one half. Just make a decision, one way or another. But just make a decision. So it's forcing you to do that.

Jeffrey's prior, I'm running out of time so I don't want to go into too much detail. We'll probably stop here, actually. So Jeffrey's priors have this very nice property. It's that they actually do not care about the parameterization of your space. If you actually have p and you suddenly decide that p is not the right parameter for Bernoulli, but it's p squared. You could decide to parameterize this by p squared. Maybe your doctor is actually much more able to formulate some prior assumption on p squared, rather than p. You never know.

And so what happens is that Jeffrey's priors are an invariant in this. And the reason is because the information carried by p is the same as the information carried by p squared, somehow. They're essentially the same thing. You need to have one to one map. Where you basically for each parameter, before you have another parameter. Let's call Eta the new parameters.

The PDF of the new prior indexed by Eta this time is actually also Jeffrey's prior. But this time, the new Fisher information is not the Fisher information with respect to theta. But it's this Fisher information associated to this statistical model indexed by Eta. So essentially, when you change the parameterization of your model, you still get Jeffrey's prior for the new parameterization. Which is, in a way, a desirable property.

Jeffrey's prior is just an uninformative priors, or priors you want to use when you want a systematic way without really thinking about what to pick for your mile.

I'll finish this next time. And we'll talk about Bayesian confidence regions. We'll talk about Bayesian estimation. Once I have a posterior, what do I get? And basically, the only message is going to be that you might want to integrate against the posterior. Find the posterior, the expectation of your posterior distribution. That's a good point estimator for theta. We'll just do a couple of computation.