The following content is provided under a Creative Commons license. Your support will help MIT OpenCourseWare continue to offer high quality educational resources for free. To make a donation or view additional materials from hundreds of MIT courses, visit MIT OpenCourseWare at ocw.mit.edu.

PROFESSOR: It involves real phenomena out there. So we have real stuff that happens. So it might be an arrival process to a bank that we're trying to model.

This is a reality, but this is what we have been doing so far. We have been playing with models of probabilistic phenomena. And somehow we need to tie the two together.

The way these are tied is that we observe the real world and this gives us data. And then based on these data, we try to come up with a model of what exactly is going on. For example, for an arrival process, you might ask the model in question, is my arrival process Poisson or is it something different? If it is Poisson, what is the rate of the arrival process? Once you come up with your model and you come up with the parameters of the model, then you can use it to make predictions about reality or to figure out certain hidden things, certain hidden aspects of reality, that you do not observe directly, but you try to infer what they are. So that's where the usefulness of the model comes in.

Now this field is of course tremendously useful. And it shows up pretty much everywhere. So we talked about the polling examples in the last couple of lectures. This is, of course, a real application.

You sample and on the basis of the sample that you have, you try to make some inferences about, let's say, the preferences in a given population. Let's say in the medical field, you want to try whether a certain drug makes a difference or not. So people would do medical trials, get some results, and then from the data somehow you need to make sense of them and make a decision. Is the new drug useful or is it not? How do we go systematically about the question of this type?

A sexier, more recent topic, there's this famous Netflix competition where Netflix gives you a huge table of movies and people. And people have rated the movies, but not everyone has watched all of the movies in there. You have some of the ratings.

For example, this person gave a 4 to that particular movie. So you get the table that's partially

filled. And the Netflix asks you to make recommendations to people.

So this means trying to guess. This person here, how much would they like this particular movie? And you can start thinking, well, maybe this person has given somewhat similar ratings with another person.

And if that other person has also seen that movie, maybe the rating of that other person is relevant. But of course it's a lot more complicated than that. And this has been a serious competition where people have been using every heavy, wet machinery that there is in statistics, trying to come up with good recommendation systems.

Then the other people, of course, are trying to analyze financial data. Somebody gives you the sequence of the values, let's say of the SMP index. You look at something like this and you can ask questions. How do I model these data using any of the models that we have in our bag of tools? How can I make predictions about what's going to happen afterwards, and so on?

On the engineering side, anywhere where you have noise inference comes in. Signal processing, in some sense, is just an inference problem. You observe signals that are noisy and you try to figure out exactly what's happening out there or what kind of signal has been sent.

Maybe the beginning of the field could be traced a few hundred years ago where people would observe, make astronomical observations of the position of the planets in the sky. They would have some beliefs that perhaps the orbits of planets is an ellipse. Or if it's a comet, maybe it's a parabola, hyperbola, don't know what it is. But they would have a model of that.

But, of course, astronomical measurements would not be perfectly exact. And they would try to find the curve that fits these data. How do you go about choosing this particular curve on the base of noisy data and try to do it in a somewhat principled way?

OK, so questions of this type-- clearly the applications are all over the place. But how is this related conceptually with what we have been doing so far? What's the relation between the field of inference and the field of probability as we have been practicing until now?

Well, mathematically speaking, what's going to happen in the next few lectures could be just exercises or homework problems in the class in based on what we have done so far. That

means you're not going to get any new facts about probability theory. Everything we're going to do will be simple applications of things that you already do know.

So in some sense, statistics and inference is just an applied exercise in probability. But actually, things are not that simple in the following sense. If you get a probability problem, there's a correct answer.

There's a correct solution. And that correct solution is unique. There's no ambiguity.

The theory of probability has clearly defined rules. These are the axioms. You're given some information about probability distributions.

You're asked to calculate certain other things. There's no ambiguity. Answers are always unique.

In statistical questions, it's no longer the case that the question has a unique answer. If I give you data and I ask you what's the best way of estimating the motion of that planet, reasonable people can come up with different methods. And reasonable people will try to argue that's my method has these desirable properties but somebody else may say, here's another method that has certain desirable properties. And it's not clear what the best method is.

So it's good to have some understanding of what the issues are and to know at least what is the general class of methods that one tries to consider, how does one go about such problems. So we're going to see lots and lots of different inference methods. We're not going to tell you that one is better than the other. But it's important to understand what are the concepts between those different methods.

And finally, statistics can be misused really badly. That is, one can come up with methods that you think are sound, but in fact they're not quite that. I will bring some examples next time and talk a little more about this.

So, they want to say, you have some data, you want to make some inference from them, what many people will do is to go to Wikipedia, find a statistical test that they think it applies to that situation, plug in numbers, and present results. Are the conclusions that they get really justified or are they misusing statistical methods?

Well, too many people actually do misuse statistics and conclusions that people get are often false. So it's important to, besides just being able to copy statistical tests and use them, to

understand what are the assumptions between the different methods and what kind of guarantees they have, if any. All right, so we'll try to do a quick tour through the field of inference in this lecture and the next few lectures that we have left this semester and try to highlight at the very high level the main concept skills, and techniques that come in. Let's start with some generalities and some general statements.

One first statement is that statistics or inference problems come up in very different guises. And they may look as if they are of very different forms. Although, at some fundamental level, the basic issues turn out to be always pretty much the same.

So let's look at this example. There's an unknown signal that's being sent. It's sent through some medium, and that medium just takes the signal and amplifies it by a certain number.

So you can think of somebody shouting. There's the air out there. What you shouted will be attenuated through the air until it gets to a receiver. And that receiver then observes this, but together with some random noise.

Here I meant S. S is the signal that's being sent. And what you observe is an X.

You observe X, so what kind of inference problems could we have here? In some cases, you want to build a model of the physical phenomenon that you're dealing with. So for example, you don't know the attenuation of your signal and you try to find out what this number is based on the observations that you have.

So the way this is done in engineering systems is that you design a certain signal, you know what it is, you shout a particular word, and then the receiver listens. And based on the intensity of the signal that they get, they try to make a guess about A. So you don't know A, but you know S. And by observing X, you get some information about what A is.

So in this case, you're trying to build a model of the medium through which your signal is propagating. So sometimes one would call problems of this kind, let's say, system identification. In a different version of an inference problem that comes with this picture, you've done your modeling.

You know your A. You know the medium through which the signal is going, but it's a communication system. This person is trying to communicate something to that person. So you send the signal S, but that person receives a noisy version of S. So that person tries to reconstruct S based on X.

So in both cases, we have a linear relation between X and the unknown quantity. In one version, A is the unknown and we know S. In the other version, A is known, and so we try to infer S.

Mathematically, you can see that this is essentially the same kind of problem in both cases. Although, the kind of practical problem that you're trying to solve is a little different. So we will not be making any distinctions between problems of the model building type as opposed to models where you try to estimate some unknown signal and so on. Because conceptually, the tools that one uses for both types of problems are essentially the same.

OK, next a very useful classification of inference problems-- the unknown quantity that you're trying to estimate could be either a discrete one that takes a small number of values. So this could be discrete problems, such as the airplane radar problem we encountered back a long time ago in this class. So there's two possibilities-- an airplane is out there or an airplane is not out there.

And you're trying to make a decision between these two options. Or you can have other problems would you have, let's say, four possible options. You don't know which one is true, but you get data and you try to figure out which one is true.

In problems of these kind, usually you want to make a decision based on your data. And you're interested in the probability of making a correct decision. You would like that probability to be as high as possible.

Estimation problems are a little different. Here you have some continuous quantity that's not known. And you try to make a good guess of that quantity. And you would like your guess to be as close as possible to the true quantity.

So the polling problem was of this type. There was an unknown fraction f of the population that had some property. And you try to estimate f as accurately as you can.

So the distinction here is that usually here the unknown quantity takes on discrete set of values. Here the unknown quantity takes a continuous set of values. Here we're interested in the probability of error.

Here we're interested in the size of the error. Broadly speaking, most inference problems fall either in this category or in that category. Although, if you want to complicate life, you can also

think or construct problems where both of these aspects are simultaneously present.

OK, finally since we're in classification mode, there is a very big, important dichotomy into how one goes about inference problems. And here there's two fundamentally different philosophical points of view, which is how do we model the quantity that is unknown?

In one approach, you say there's a certain quantity that has a definite value. It just happens that they don't know it. But it's a number. There's nothing random about it. So think of trying to estimate some physical quantity.

You're making measurements, you try to estimate the mass of an electron, which is a sort of universal physical constant. There's nothing random about it. It's a fixed number. You get data, because you have some measuring apparatus.

And that measuring apparatus, depending on what that results that you get are affected by the true mass of the electron, but there's also some noise. You take the data out of your measuring apparatus and you try to come up with some estimate of that quantity theta. So this is definitely a legitimate picture, but the important thing in this picture is that this theta is written as lowercase. And that's to make the point that it's a real number, not a random variable.

There's a different philosophical approach which says, well, anything that I don't know I should model it as a random variable. Yes, I know. The mass of the electron is not really random. It's a constant.

But I don't know what it is. I have some vague sense, perhaps, what it is perhaps because of the experiments that some other people carried out. So perhaps I have a prior distribution on the possible values of Theta.

And that prior distribution doesn't mean that the nature is random, but it's more of a subjective description of my subjective beliefs of where do I think this constant number happens to be. So even though it's not truly random, I model my initial beliefs before the experiment starts. In terms of a prior distribution, I view it as a random variable. Then I observe another related random variable through some measuring apparatus. And then I use this again to create an estimate.

So these two pictures philosophically are very different from each other. Here we treat the unknown quantities as unknown numbers. Here we treat them as random variables.

When we treat them as a random variables, then we know pretty much already what we should be doing. We should just use the Bayes rule. Based on X, find the conditional distribution of Theta. And that's what we will be doing mostly over this lecture and the next lecture.

Now in both cases, what you end up getting at the end is an estimate. But actually, that estimate is what kind of object is it? It's a random variable in both cases. Why?

Even in this case where theta was a constant, my data are random. I do my data processing. So I calculate a function of the data, the data are random variables.

So out here we output something which is a function of a random variable. So this quantity here will be also random. It's affected by the noise and the experiment that I have been doing.

That's why these estimators will be denoted by uppercase Thetas. And we will be using hats. Hat, usually in estimation, means an estimate of something.

All right, so this is the big picture. We're going to start with the Bayesian version. And then the last few lectures we're going to talk about the non-Bayesian version or the classical one.

By the way, I should say that statisticians have been debating fiercely for 100 years whether the right way to approach statistics is to go the classical way or the Bayesian way. And there have been tides going back and forth between the two sides. These days, Bayesian methods tend to become a little more popular for various reasons. We're going to come back to this later.

All right, so in Bayesian estimation, what we got in our hands is Bayes rule. And if you have Bayes rule, there's not a lot that's left to do. We have different forms of the Bayes rule, depending on whether we're dealing with discrete data, And discrete quantities to estimate, or continuous data, and so on.

In the hypothesis testing problem, the unknown quantity Theta is discrete. So in both cases here, we have a P of Theta. We obtain data, the X's. And on the basis of the X that we observe, we can calculate the posterior distribution of Theta, given the data.

So to use Bayesian inference, what do we start with? We start with some priors. These are our initial beliefs about what Theta that might be. That's before we do the experiment.

We have a model of the experimental aparatus. And the model of the experimental apparatus tells us if this Theta is true, I'm going to see X's of that kind. If that other Theta is true, I'm going to see X's that they are somewhere else. That models my apparatus.

And based on that knowledge, once I observe I have these two functions in my hands, we have already seen that if you know those two functions, you can also calculate the denominator here. So all of these functions are available, so you can compute, you can find a formula for this function as well. And as soon as you observe the data, that X's, you plug in here the numerical value of those X's. And you get a function of Theta. And this is the posterior distribution of Theta, given the data that you have seen.

So you've already done a fair number of exercises of these kind. So we not say more about this. And there's a similar formula as you know for the case where we have continuous data. If the X's are continuous random variable, then the formula is the same, except that X's are described by densities instead of being described by a probability mass functions.

OK, now if Theta is continuous, then we're dealing with estimation problems. But the story is once more the same. You're going to use the Bayes rule to come up with the posterior density of Theta, given the data that you have observed.

Now just for the sake of the example, let's come back to this picture here. Suppose that something is flying in the air, and maybe this is just an object in the air close to the Earth. So because of gravity, the trajectory that it's going to follow it's going to be a parabola.

So this is the general equation of a parabola. Zt is the position of my objects at time t. But I don't know exactly which parabola it is. So the parameters of the parabola are unknown quantities.

What I can do is to go and measure the position of my objects at different times. But unfortunately, my measurements are noisy. What I want to do is to model the motion of my object. So I guess in the picture, the axis would be t going this way and Z going this way.

And on the basis of the data that they get, these are my X's. I want to figure out the Thetas. That is, I want to figure out the exact equation of this parabola.

Now if somebody gives you probability distributions for Theta, these would be your priors. So this is given. We need the conditional distribution of the X's given the Thetas.

Well, we have the conditional distribution of Z, given the Thetas from this equation. And then by playing with this equation, you can also find how is X distributed if Theta takes a particular value.

So you do have all of the densities that you might need. And you can apply the Bayes rule. And at the end, your end result would be a formula for the distribution of Theta, given to the X that you have observed-- except for one sort of computation, or to make things more interesting.

Instead of these X's and Theta's being single random variables that we have here, typically those X's and Theta's will be multi-dimensional random variables or will correspond to multiple ones. So this little Theta here actually stands for a triplet of Theta0, Theta1, and Theta2. And that X here stands here for the entire sequence of X's that we have observed.

So in reality, the object that you're going to get at to the end after inference is done is a function that you plug in the values of the data and you get the function of the Theta's that tells you the relative likelihoods of different Theta triplets.

So what I'm saying is that this is no harder than the problems that you have dealt with so far, except perhaps for the complication that's usually in interesting inference problems. Your Theta's and X's are often the vectors of random variables instead of individual random variables.

Now if you are to do estimation in a case where you have discrete data, again the situation is no different. We still have a Bayes rule of the same kind, except that densities gets replaced by PMF's. If X is discrete, you put a P here instead of putting an f.

So an example of an estimation problem with discrete data is similar to the polling problem. You have a coin. It has an unknown parameter Theta. This is the probability of obtaining heads. You flip the coin many times. What can you tell me about the true value of Theta?

A classical statistician, at this point, would say, OK, I'm going to use an estimator, the most reasonable one, which is this. How many heads did they obtain in n trials? Divide by the total number of trials. This is my estimate of the bias of my coin.

And then the classical statistician would continue from here and try to prove some properties and argue that this estimate is a good one. For example, we have the weak law of large numbers that tells us that this particular estimate converges in probability to the true

parameter. This is a kind of guarantee that's useful to have. And the classical statistician would pretty much close the subject in this way.

What would the Bayesian person do differently? The Bayesian person would start by assuming a prior distribution of Theta. Instead of treating Theta as an unknown constant, they would say that Theta would speak randomly or pretend that it would speak randomly and assume a distribution on Theta.

So for example, if you don't know they need anything more, you might assume that any value for the bias of the coin is as likely as any other value of the bias of the coin. And this way so the probability distribution that's uniform. Or if you have a little more faith in the manufacturing processes that's created that coin, you might choose your prior to be a distribution that's centered around 1/2 and sits fairly narrowly centered around 1/2.

That would be a prior distribution in which you say, well, I believe that the manufacturer tried to make my coin to be fair. But they often makes some mistakes, so it's going to be, I believe, it's approximately 1/2 but not quite. So depending on your beliefs, you would choose an appropriate prior for the distribution of Theta. And then you would use the Bayes rule to find the probabilities of different values of Theta, based on the data that you have observed.

So no matter which version of the Bayes rule that you use, the end product of the Bayes rule is going to be either a plot of this kind or a plot of that kind. So what am I plotting here? This axis is the Theta axis. These are the possible values of the unknown quantity that we're trying to estimate.

In the continuous case, theta is a continuous random variable. I obtain my data. And I plot for the posterior probability distribution after observing my data. And I'm plotting here the probability density for Theta. So this is a plot of that density.

In the discrete case, theta can take finitely many values or a discrete set of values. And for each one of those values, I'm telling you how likely is that the value to be the correct one, given the data that I have observed. And in general, what you would go back to your boss and report after you've done all your inference work would be either a plot of this kinds or of that kind.

So you go to your boss who asks you, what is the value of Theta? And you say, well, I only have limited data. That I don't know what it is. It could be this, with so much probability.

There's probability.

OK, let's throw in some numbers here. There's probability 0.3 that Theta is this value. There's probability 0.2 that Theta is this value, 0.1 that it's this one, 0.1 that it's this one, 0.2 that it's that one, and so on.

OK, now bosses often want simple answers. They say, OK, you're talking too much. What do you think Theta is? And now you're forced to make a decision. If that was the situation and you have to make a decision, how would you make it? Well, I'm going to make a decision that's most likely to be correct. If I make this decision, what's going to happen?

Theta is this value with probability 0.2, which means there's probably 0.8 that they make an error if I make that guess. If I make that decision, this decision has probably 0.3 of being the correct one. So I have probably of error 0.7.

So if you want to just maximize the probability of giving the correct decision, or if you want to minimize the probability of making an incorrect decision, what you're going to choose to report is that value of Theta for which the probability is highest. So in this case, I would choose to report this particular value, the most likely value of Theta, given what I have observed. And that value is called them maximum a posteriori probability estimate. It's going to be this one in our case.

So picking the point in the posterior PMF that has the highest probability. That's the reasonable thing to do. This is the optimal thing to do if you want to minimize the probability of an incorrect inference. And that's what people do usually if they need to report a single answer, if they need to report a single decision.

How about in the estimation context? If that's what you know about Theta, Theta could be around here, but there's also some sharp probability that it is around here. What's the single answer that you would give to your boss?

One option is to use the same philosophy and say, OK, I'm going to find the Theta at which this posterior density is highest. So I would pick this point here and report this particular Theta. So this would be my Theta, again, Theta MAP, the Theta that has the highest a posteriori probability, just because it corresponds to the peak of the density.

But in this context, the maximum a posteriori probability theta was the one that was most likely

to be true. In the continuous case, you cannot really say that this is the most likely value of Theta. In a continuous setting, any value of Theta has zero probability, so when we talk about densities. So it's not the most likely. It's the one for which the density, so the probabilities of that neighborhoods, are highest. So the rationale for picking this particular estimate in the continuous case is much less compelling than the rationale that we had in here.

So in this case, reasonable people might choose different quantities to report. And the very popular one would be to report instead the conditional expectation. So I don't know quite what Theta is.

Given the data that I have, Theta has this distribution. Let me just report the average over that distribution. Let me report to the center of gravity of this figure.

And in this figure, the center of gravity would probably be somewhere around here. And that would be a different estimate that you might choose to report. So center of gravity is something around here. And this is a conditional expectation of Theta, given the data that you have.

So these are two, in some sense, fairly reasonable ways of choosing what to report to your boss. Some people might choose to report this. Some people might choose to report that. And a priori, if there's no compelling reason why one would be preferable than other one, unless you set some rules for the game and you describe a little more precisely what your objectives are.

But no matter which one you report, a single answer, a point estimate, doesn't really tell you the whole story. There's a lot more information conveyed by this posterior distribution plot than any single number that you might report. So in general, you may wish to convince your boss that's it's worth their time to look at the entire plot, because that plot sort of covers all the possibilities. It tells your boss most likely we're in that range, but there's also a distinct change that our Theta happens to lie in that range.

All right, now let us try to perhaps differentiate between these two and see under what circumstances this one might be the better estimate to perform. Better with respect to what? We need some rules. So we're going to throw in some rules.

As a warm up, we're going to deal with the problem of making an estimation if you had no information at all, except for a prior distribution. So this is a warm up for what's coming next,

which would be estimation that takes into account some information.

So we have a Theta. And because of your subjective beliefs or models by others, you believe that Theta is uniformly distributed between, let's say, 4 and 10. You want to come up with a point estimate.

Let's try to look for an estimate. Call it c, in this case. I want to pick a number with which to estimate the value of Theta. I will be interested in the size of the error that I make. And I really dislike large errors, so I'm going to focus on the square of the error that they make.

So I pick c. Theta that has a random value that I don't know. But whatever it is, once it becomes known, it results into a squared error between what it is and what I guessed that it was. And I'm interested in making a small air on the average, where the average is taken with respect to all the possible and unknown values of Theta.

So the problem, this is a least squares formulation of the problem, where we try to minimize the least squares errors. How do you find the optimal c? Well, we take that expression and expand it. And it is, using linearity of expectations-- square minus 2c expected Theta plus c squared-- that's the quantity that we want to minimize, with respect to c.

To do the minimization, take the derivative with respect to c and set it to 0. So that differentiation gives us from here minus 2 expected value of Theta plus 2c is equal to 0. And the answer that you get by solving this equation is that c is the expected value of Theta.

So when you do this optimization, you find that the optimal estimate, the things you should be reporting, is the expected value of Theta. So in this particular example, you would choose your estimate c to be just the middle of these values, which would be 7.

OK, and in case your boss asks you, how good is your estimate? How big is your error going to be? What you could report is the average size of the estimation error that you are making. We picked our estimates to be the expected value of Theta. So for this particular way that I'm choosing to do my estimation, this is the mean squared error that I get. And this is a familiar quantity. It's just the variance of the distribution.

So the expectation is that best way to estimate a quantity, if you're interested in the mean squared error. And the resulting mean squared error is the variance itself. How will this story change if we now have data as well? Now having data means that we can compute posterior distributions or conditional distributions. So we get transported into a new universe where

instead the working with the original distribution of Theta, the prior distribution, now we work with the condition of distribution of Theta, given the data that we have observed.

Now remember our old slogan that conditional models and conditional probabilities are no different than ordinary probabilities, except that we live now in a new universe where the new information has been taken into account. So if you use that philosophy and you're asked to minimize the squared error but now that you live in a new universe where X has been fixed to something, what would the optimal solution be? It would again be the expectation of theta, but which expectation? It's the expectation which applies in the new conditional universe in which we live right now.

So because of what we did before, by the same calculation, we would find that the optimal estimates is the expected value of X of Theta, but the optimal estimate that takes into account the information that we have. So the conclusion, once you get your data, if you want to minimize the mean squared error, you should just report the conditional estimation of this unknown quantity based on the data that you have.

So the picture here is that Theta is unknown. You have your apparatus that creates measurements. So this creates an X. You take an X, and here you have a box that does calculations. It does calculations and it spits out the conditional expectation of Theta, given the particular data that you have observed.

And what we have done in this class so far is, to some extent, developing the computational tools and skills to do with this particular calculation-- how to calculate the posterior density for Theta and how to calculate expectations, conditional expectations. So in principle, we know how to do this. In principle, we can program a computer to take the data and to spit out condition expectations.

Somebody who doesn't think like us might instead design a calculating machine that does something differently and produces some other estimate. So we went through this argument and we decided to program our computer to calculate conditional expectations. Somebody else came up with some other crazy idea for how to estimate the random variable. They came up with some function g and the programmed it, and they designed a machine that estimates Theta's by outputting a certain g of X.

That could be an alternative estimator. Which one is better? Well, we convinced ourselves that

this is the optimal one in a universe where we have fixed the particular value of the data. So what we have proved so far is a relation of this kind. In this conditional universe, the mean squared error that I get-- I'm the one who's using this estimator-- is less than or equal than the mean squared error that this person will get, the person who uses that estimator.

For any particular value of the data, I'm going to do better than the other person. Now the data themselves are random. If I average over all possible values of the data, I should still be better off. If I'm better off for any possible value X, then I should be better off on the average over all possible values of X.

So let us average both sides of this quantity with respect to the probability distribution of X. If you want to do it formally, you can write this inequality between numbers as an inequality between random variables. And it tells that no matter what that random variable turns out to be, this quantity is better than that quantity. Take expectations of both sides, and you get this inequality between expectations overall.

And this last inequality tells me that the person who's using this estimator who produces estimates according to this machine will have a mean squared estimation error that's less than or equal to the estimation error that's produced by the other person. In a few words, the conditional expectation estimator is the optimal estimator. It's the ultimate estimating machine. That's how you should solve estimation problems and report a single value. If you're forced to report a single value and if you're interested in estimation errors.

OK, while we could have told you that story, of course, a month or two ago, this is really about interpretation -- about realizing that conditional expectations have a very nice property. But other than that, any probabilistic skills that come into this business are just the probabilistic skills of being able to calculate conditional expectations, which you already know how to do.

So conclusion, all of optimal Bayesian estimation just means calculating and reporting conditional expectations. Well, if the world were that simple, then statisticians wouldn't be able to find jobs if life is that simple. So real life is not that simple. There are complications. And that perhaps makes their life a little more interesting.

OK, one complication is that we would deal with the vectors instead of just single random variables. I use the notation here as if X was a single random variable. In real life, you get several data. Does our story change? Not really, same argument-- given all the data that you have observed, you should still report the conditional expectation of Theta.

But what kind of work does it take in order to report this conditional expectation? One issue is that you need to cook up a plausible prior distribution for Theta. How do you do that? In a given application , this is a bit of a judgment call, what prior would you be working with. And there's a certain skill there of not making silly choices.

A more pragmatic, practical issue is that this is a formula that's extremely nice and compact and simple that you can write with minimal ink. But the behind it there could be hidden a huge amount of calculation. So doing any sort of calculations that involve multiple random variables really involves calculating multi-dimensional integrals.

And the multi-dimensional integrals are hard to compute. So implementing actually this calculating machine here may not be easy, might be complicated computationally. It's also complicated in terms of not being able to derive intuition about it. So perhaps you might want to have a simpler version, a simpler alternative to this formula that's easier to work with and easier to calculate.

We will be talking about one such simpler alternative next time. So again, to conclude, at the high level, Bayesian estimation is very, very simple, given that you have mastered everything that has happened in this course so far. There are certain practical issues and it's also good to be familiar with the concepts and the issues that in general, you would prefer to report that complete posterior distribution. But if you're forced to report a point estimate, then there's a number of reasonable ways to do it. And perhaps the most reasonable one is to just the report the conditional expectation itself.