JOHN TSITSIKLIS: And we're going to continue today with our discussion of classical statistics. We'll start with a quick review of what we discussed last time, and then talk about two topics that cover a lot of statistics that are happening in the real world. So two basic methods. One is the method of linear regression, and the other one is the basic methods and tools for how to do hypothesis testing.

OK, so these two are topics that any scientifically literate person should know something about. So we're going to introduce the basic ideas and concepts involved. So in classical statistics we basically have essentially a family of possible models about the world.

So the world is the random variable that we observe, and we have a model for it, but actually not just one model, several candidate models. And each candidate model corresponds to a different value of a parameter theta that we do not know. So in contrast to Bayesian statistics, this theta is assumed to be a constant that we do not know. It is not modeled as a random variable, there's no probabilities associated with theta.

We only have probabilities about the X's. So in this context what is a reasonable way of choosing a value for the parameter? One general approach is the maximum likelihood approach, which chooses the theta for which this quantity is largest. So what does that mean intuitively? I'm trying to find the value of theta under which the data that I observe are most likely to have occurred.

So is the thinking is essentially as follows. Let's say I have to choose between two choices of theta. Under this theta the X that I observed would be very unlikely. Under that theta the X that I observed would have a decent probability of occurring. So I chose the latter as my estimate of theta.

It's interesting to do the comparison with the Bayesian approach which we did discuss last time, in the Bayesian approach we also maximize over theta, but we maximize a quantity in which the relation between X's and thetas run the opposite way.

Here in the Bayesian world, Theta is a random variable. So it has a distribution. Once we observe the data, it has a posterior distribution, and we find the value of Theta, which is most likely under the posterior distribution.

As we discussed last time when you do this maximization now the posterior distribution is given by this expression. The denominator doesn't matter, and if you were to take a prior, which is flat-- that is a constant independent of Theta, then that term would go away. And syntactically, at least, the two approaches look the same.

So syntactically, or formally, maximum likelihood estimation is the same as Bayesian estimation in which you assume a prior which is flat, so that all possible values of Theta are equally likely.

Philosophically, however, they're very different things. Here I'm picking the most likely value of Theta. Here I'm picking the value of Theta under which the observed data would have been more likely to occur. So maximum likelihood estimation is a general purpose method, so it's applied all over the place in many, many different types of estimation problems.

There is a special kind of estimation problem in which you may forget about maximum likelihood estimation, and come up with an estimate in a straightforward way. And this is the case where you're trying to estimate the mean of the distribution of X, where X is a random variable. You observe several independent identically distributed random variables X1 up to Xn. All of them have the same distribution as this X.

So they have a common mean. We do not know the mean we want to estimate it. What is more natural than just taking the average of the values that we have observed? So you generate lots of X's, take the average of them, and you expect that this is going to be a reasonable estimate of the true mean of that random variable. And indeed we know from the weak law of large numbers that this estimate converges in probability to the true mean of the random variable.

The other thing that we talked about last time is that besides giving a point estimate we may want to also give an interval that tells us something about where we might believe theta to lie. And 1-alpha confidence interval is in interval generated based on the data. So it's an interval from this value to that value. These values are written with capital letters because they're random, because they depend on the data that we have seen. And this gives us an interval, and we would like this interval to have the property that theta is inside that interval with high probability.

So typically we would take 1-alpha to be a quantity such as 95% for example. In which case we have a 95% confidence interval. As we discussed last time it's important to have the right interpretation of what's 95% means.

What it does not mean is the following-- the unknown value has 95% percent probability of being in the interval that we have generated. That's because the unknown value is not a random variable, it's a constant. Once we generate the interval either it's inside or it's outside, but there's no probabilities involved.

Rather the probabilities are to be interpreted over the random interval itself. What a statement like this says is that if I have a procedure for generating 95% confidence intervals, then whenever I use that procedure I'm going to get a random interval, and it's going to have 95% probability of capturing the true value of theta.

So most of the time when I use this particular procedure for generating confidence intervals the true theta will happen to lie inside that confidence interval with probability 95%. So the randomness in this statement is with respect to my confidence interval, it's not with respect to theta, because theta is not random.

How does one construct confidence intervals? There's various ways of going about it, but in the case where we're dealing with the estimation of the mean of a random variable doing this is straightforward using the central limit theorem. Basically we take our estimated mean, that's the sample mean, and we take a symmetric interval to the left and to the right of the sample mean.

And we choose the width of that interval by looking at the normal tables. So if this quantity, 1-alpha is 95% percent, we're going to look at the 97.5 percentile of the normal distribution. Find the constant number that corresponds to that value from the normal tables, and construct the confidence intervals according to this formula. So that gives you a pretty mechanical way of going about constructing confidence intervals when you're estimating the sample mean.

So constructing confidence intervals in this way involves an approximation. The approximation is the central limit theorem. We are pretending that the sample mean is a normal random variable. Which is, more or less, right when n is large. That's what the central limit theorem tells us.

And sometimes we may need to do some extra approximation work, because quite often we do not know the true value of sigma. So we need to do some work either to estimate sigma from the data. So sigma is, of course, the standard deviation of the X's. We may want to estimate it from the data, or we may have an upper bound on sigma, and we just use that upper bound.

So now let's move on to a new topic. A lot of statistics in the real world are of the following flavor. So suppose that X is the SAT score of a student in high school, and Y is the MIT GPA of that same student. So you expect that there is a relation between these two. So you go and collect data for different students, and you record for a typical student this would be their SAT score, that could be their MIT GPA. And you plot all this data on an (X,Y) diagram.

Now it's reasonable to believe that there is some systematic relation between the two. So people who had higher SAT scores in high school may have higher GPA in college. Well that may or may not be true. You want to construct a model of this kind, and see to what extent a relation of this type is true.

So you might hypothesize that the real world is described by a model of this kind. That there is a linear relation between the SAT score, and the college GPA. So it's a linear relation with some parameters, theta0 and theta1 that we do not know.

So we assume a linear relation for the data, and depending on the choices of theta0 and theta1 it could be a different line through those data. Now we would like to find the best model of this kind to explain the data. Of course there's going to be some randomness. So in general it's going to be impossible to find a line that goes through all of the data points.

So let's try to find the best line that comes closest to explaining those data. And here's how we go about it. Suppose we try some particular values of theta0 and theta1. These give us a certain line. Given that line, we can make predictions.

For a student who had this x, the model that we have would predict that y would be this value. The actual y is something else, and so this quantity is the error that our model would make in predicting the y of that particular student. We would like to choose a line for which the predictions are as good as possible. And what do we mean by as good as possible? As our criteria we're going to take the following.

We are going to look at the prediction error that our model makes for each particular student. Take the square of that, and then add them up over all of our data points. So what we're looking at is the sum of this quantity squared, that quantity squared, that quantity squared, and so on. We add all of these squares, and we would like to find the line for which the sum of these squared prediction errors are as small as possible.

So that's the procedure. We have our data, the X's and the Y's. And we're going to find theta's the best model of this type, the best possible model, by minimizing this sum of squared errors. So that's a method that one could pull out of the hat and say OK, that's how I'm going to build my model. And it sounds pretty reasonable.

And it sounds pretty reasonable even if you don't know anything about probability. But does it have some probabilistic justification? It turns out that yes, you can motivate this method with probabilistic considerations under certain assumptions. So let's make a probabilistic model that's going to lead us to these particular way of estimating the parameters.

So here's a probabilistic model. I pick a student who had a specific SAT score. And that could be done at random, but also could be done in a systematic way. That is, I pick a student who had an SAT of 600, a student of 610 all the way to 1,400 or 1,600, whatever the right number is. I pick all those students.

And I assume that for a student of this kind there's a true model that tells me that their GPA is going to be a random variable, which is something predicted by their SAT score plus some randomness, some random noise. And I model that random noise by independent normal random variables with 0 mean and a certain variance.

So this is a specific probabilistic model, and now I can think about doing maximum likelihood estimation for this particular model. So to do maximum likelihood estimation here I need to write down the likelihood of the y's that I have observed. What's the likelihood of the y's that I have observed?

Well, a particular w has a likelihood of the form e to the minus w squared over (2 sigma-squared). That's the likelihood of a particular w. The probability, or the likelihood of observing a particular value of y, that's the same as the likelihood that w takes a value of y minus this, minus that. So the likelihood of the y's is of this form. Think of this as just being the w_i-squared.

So this is the density -- and if we have multiple data you multiply the likelihoods of the different y's. So you have to write something like this. Since the w's are independent that means that the y's are also independent. The likelihood of a y vector is the product of the likelihoods of the individual y's. The likelihood of every individual y is of this form. Where w is y_i minus these two quantities.

So this is the form that the likelihood function is going to take under this particular model. And under the maximum likelihood methodology we want to maximize this quantity with respect to theta0 and theta1. Now to do this maximization you might as well consider the logarithm and maximize the logarithm, which is just the exponent up here. Maximizing this exponent because we have a minus sign is the same as minimizing the exponent without the minus sign. Sigma squared is a constant. So what you end up doing is minimizing this quantity here, which is the same as what we had in our linear regression methods.

So in conclusion you might choose to do linear regression in this particular way, just because it looks reasonable or plausible. Or you might interpret what you're doing as maximum likelihood estimation, in which you assume a model of this kind where the noise terms are normal random variables with the same distribution -- independent identically distributed.

So linear regression implicitly makes an assumption of this kind. It's doing maximum likelihood estimation as if the world was really described by a model of this form, and with the W's being random variables. So this gives us at least some justification that this particular approach to fitting lines to data is not so arbitrary, but it has a sound footing.

OK so then once you accept this formulation as being a reasonable one what's the next step? The next step is to see how to carry out this minimization. This is not a very difficult minimization to do. The way it's done is by setting the derivatives of this expression to 0. Now because this is a quadratic function of theta0 and theta1-- when you take the derivatives with respect to theta0 and theta1-- you get linear functions of theta0 and theta1. And you end up solving a system of linear equations in theta0 and theta1. And it turns out that there's very nice and simple formulas for the optimal estimates of the parameters in terms of the data.

And the formulas are these ones. I said that these are nice and simple formulas. Let's see why. How can we interpret them? So suppose that the world is described by a model of this kind, where the X's and Y's are random variables. And where W is a noise term that's independent of X. So we're assuming that a linear model is indeed true, but not exactly true. There's always some noise associated with any particular data point that we obtain.

So if a model of this kind is true, and the W's have 0 mean then we have that the expected value of Y would be theta0 plus theta1 expected value of X. And because W has 0 mean there's no extra term. So in particular, theta0 would be equal to expected value of Y minus theta1 expected value of X.

So let's use this equation to try to come up with a reasonable estimate of theta0. I do not know the expected value of Y, but I can estimate it. How do I estimate it? I look at the average of all the y's that I have obtained. so I replace this, I estimate it with the average of the data I have seen.

Here, similarly with the X's. I might not know the expected value of X's, but I have data points for the x's. I look at the average of all my data points, I come up with an estimate of this expectation. Now I don't know what theta1 is, but my procedure is going to generate an estimate of theta1 called theta1 hat. And once I have this estimate, then a reasonable person would estimate theta0 in this particular way.

So that's how my estimate of theta0 is going to be constructed. It's this formula here. We have not yet addressed the harder question, which is how to estimate theta1 in the first place. So to estimate theta0 I assumed that I already had an estimate for a theta1.

OK, the right formula for the estimate of theta1 happens to be this one. It looks messy, but let's try to interpret it. What I'm going to do is I'm going to take this model for simplicity let's assume that they're the random variables have 0 means. And see how we might estimate how we might try to estimate theta1.

Let's multiply both sides of this equation by X. So we get Y times X equals theta0 plus theta0 times X plus theta1 times X-squared, plus X times W. And now take expectations of both sides. If I have 0 mean random variables the expected value of Y times X is just the covariance of X with Y.

I have assumed that my random variables have 0 means, so the expectation of this is 0. This one is going to be the variance of X, so I have theta1 times variance of X. And since I'm assuming that my random variables have 0 mean, and I'm also assuming that W is independent of X this last term also has 0 mean.

So under such a probabilistic model this equation is true. If we knew the variance and the covariance then we would know the value of theta1. But we only have data, we do not necessarily know the variance and the covariance, but we can estimate it.

What's a reasonable estimate of the variance? The reasonable estimate of the variance is this quantity here divided by n, and the reasonable estimate of the covariance is that numerator divided by n.

So this is my estimate of the mean. I'm looking at the squared distances from the mean, and I average them over lots and lots of data. This is the most reasonable way of estimating the variance of our distribution.

And similarly the expected value of this quantity is the covariance of X with Y, and then we have lots and lots of data points. This quantity here is going to be a very good estimate of the covariance. So basically what this formula does is-- one way of thinking about it-- is that it starts from this relation which is true exactly, but estimates the covariance and the variance on the basis of the data, and then using these estimates to come up with an estimate of theta1.

So this gives us a probabilistic interpretation of the formulas that we have for the way that the estimates are constructed. If you're willing to assume that this is the true model of the world, the structure of the true model of the world, except that you do not know means and covariances, and variances. Then this is a natural way of estimating those unknown parameters.

All right, so we have a closed-form formula, we can apply it whenever we have data. Now linear regression is a subject on which there are whole courses, and whole books that are given. And the reason for that is that there's a lot more that you can bring into the topic, and many ways that you can elaborate on the simple solution that we got for the case of two parameters and only two random variables.

So let me give you a little bit of flavor of what are the topics that come up when you start looking into linear regression in more depth. So in our discussions so far we made the linear model in which we're trying to explain the values of one variable in terms of the values of another variable. We're trying to explain GPAs in terms of SAT scores, or we're trying to predict GPAs in terms of SAT scores.

But maybe your GPA is affected by several factors. For example maybe your GPA is affected by your SAT score, also the income of your family, the years of education of your grandmother, and many other factors like that. So you might write down a model in which I believe that GPA has a relation, which is a linear function of all these other variables that I mentioned. So perhaps you have a theory of what determines performance at college, and you want to build a model of that type.

How do we go about in this case? Well, again we collect the data points. We look at the i-th student, who has a college GPA. We record their SAT score, their family income, and grandmother's years of education. So this is one data point that is for one particular student.

We postulate the model of this form. For the i-th student this would be the mistake that our model makes if we have chosen specific values for those parameters. And then we go and choose the parameters that are going to give us, again, the smallest possible sum of squared errors. So philosophically it's exactly the same as what we were discussing before, except that now we're including multiple explanatory variables in our model instead of a single explanatory variable.

So that's the formulation. What do you do next? Well, to do this minimization you're going to take derivatives once you have your data, you have a function of these three parameters. You take the derivative with respect to the parameter, set the derivative equal to 0, you get the system of linear equations. You throw that system of linear equations to the computer, and you get numerical values for the optimal parameters.

There are no nice closed-form formulas of the type that we had in the previous slide when you're dealing with multiple variables. Unless you're willing to go into matrix notation. In that case you can again write down closed-form formulas, but they will be a little less intuitive than what we had before. But the moral of the story is that numerically this is a procedure that's very easy. It's a problem, an optimization problem that the computer can solve for you. And it can solve it for you very quickly. Because all that it involves is solving a system of linear equations.

Now when you choose your explanatory variables you may have some choices. One person may think that your GPA a has something to do with your SAT score. Some other person may think that your GPA has something to do with the square of your SAT score. And that other person may want to try to build a model of this kind.

Now when would you want to do this? ? Suppose that the data that you have looks like this. If the data looks like this then you might be tempted to say well a linear model does not look right, but maybe a quadratic model will give me a better fit for the data. So if you want to fit a quadratic model to the data then what you do is you take X-squared as your explanatory variable instead of X, and you build a model of this kind.

There's nothing really different in models of this kind compared to models of that kind. They are still linear models because we have theta's showing up in a linear fashion. What you take as your explanatory variables, whether it's X, whether it's X-squared, or whether it's some other function that you chose. Some general function h of X, doesn't make a difference. So think of you h of X as being your new X. So you can formulate the problem exactly the same way, except that instead of using X's you choose h of X's.

So it's basically a question do I want to build a model that explains Y's based on the values of X, or do I want to build a model that explains Y's on the basis of the values of h of X. Which is the right value to use? And with this picture here, we see that it can make a difference. A linear model in X might be a poor fit, but a quadratic model might give us a better fit.

So this brings to the topic of how to choose your functions h of X if you're dealing with a real world problem. So in a real world problem you're just given X's and Y's. And you have the freedom of building models of any kind you want. You have the freedom of choosing a function h of X of any type that you want.

So this turns out to be a quite difficult and tricky topic. Because you may be tempted to overdo it. For example, I got my 10 data points, and I could say OK, I'm going to choose an h of X. I'm going to choose h of X and actually multiple h's of X to do a multiple linear regression in which I'm going to build a model that's uses a 10th degree polynomial.

If I choose to fit my data with a 10th degree polynomial I'm going to fit my data perfectly, but I may obtain a model is does something like this, and goes through all my data points. So I can make my prediction errors extremely small if I use lots of parameters, and if I choose my h functions appropriately. But clearly this would be garbage.

If you get those data points, and you say here's my model that explains them. That has a polynomial going up and down, then you're probably doing something wrong. So choosing how complicated those functions, the h's, should be. And how many explanatory variables to use is a very delicate and deep topic on which there's deep theory that tells you what you should do, and what you shouldn't do.

But the main thing that one should avoid doing is having too many parameters in your model when you have too few data. So if you only have 10 data points, you shouldn't have 10 free parameters. With 10 free parameters you will be able to fit your data perfectly, but you wouldn't be able to really rely on the results that you are seeing.

OK, now in practice, when people run linear regressions they do not just give point estimates for the parameters theta. But similar to what we did for the case of estimating the mean of a random variable you might want to give confidence intervals that sort of tell you how much randomness there is when you estimate each one of the particular parameters.

There are formulas for building confidence intervals for the estimates of the theta's. We're not going to look at them, it would take too much time. Also you might want to estimate the variance in the noise that you have in your model. That is if you are pretending that your true model is of the kind we were discussing before, namely Y equals theta1 times X plus W, and W has a variance sigma squared. You might want to estimate this, because it tells you something about the model, and this is called standard error.

It puts a limit on how good predictions your model can make. Even if you have the correct theta0 and theta1, and somebody tells you X you can make a prediction about Y, but that prediction will not be accurate. Because there's this additional randomness. And if that additional randomness is big, then your predictions will also have a substantial error in them.

There's another quantity that gets reported usually. This is part of the computer output that you get when you use a statistical package which is called R-square. And its a measure of the explanatory power of the model that you have built linear regression. Using linear regression. Instead of defining R-square exactly, let me give you a sort of analogous quantity that's involved.

After you do your linear regression you can look at the following quantity. You look at the variance of Y, which is something that you can estimate from data. This is how much randomness there is in Y. And compare it with the randomness that you have in Y, but conditioned on X. So this quantity tells me if I knew X how much randomness would there still be in my Y?

So if I know X, I have more information, so Y is more constrained. There's less randomness in Y. This is the randomness in Y if I don't know anything about X.

So naturally this quantity would be less than 1, and if this quantity is small it would mean that whenever I know X then Y is very well known. Which essentially tells me that knowing x allows me to make very good predictions about Y. Knowing X means that I'm explaining away most of the randomness in Y.

So if you read a statistical study that uses linear regression you might encounter statements of the form 60% of a student's GPA is explained by the family income. If you read the statements of this kind it's really refers to quantities of this kind. Out of the total variance in Y, how much variance is left after we build our model?

So if only 40% of the variance of Y is left after we build our model, that means that X explains 60% of the variations in Y's. So the idea is that randomness in Y is caused by multiple sources. Our explanatory variable and random noise. And we ask the question what percentage of the total randomness in Y is explained by variations in the X parameter? And how much of the total randomness in Y is attributed just to random effects? So if you have a model that explains most of the variation in Y then you can think that you have a good model that tells you something useful about the real world.

Now there's lots of things that can go wrong when you use linear regression, and there's many pitfalls. One pitfall happens when you have this situation that's called heteroskedacisity. So suppose your data are of this kind. So what's happening here? You seem to have a linear model, but when X is small you have a very good model. So this means that W has a small variance when X is here.

On the other hand, when X is there you have a lot of randomness. This would be a situation in which the W's are not identically distributed, but the variance of the W's, of the noise, has something to do with the X's. So with different regions of our x-space we have different amounts of noise. What will go wrong in this situation? Since we're trying to minimize sum of squared errors, we're really paying attention to the biggest errors. Which will mean that we are going to pay attention to these data points, because that's where the big errors are going to be So the linear regression formulas will end up building a model based on these data, which are the most noisy ones. Instead of those data that are nicely stacked in order.

Clearly that's not to the right thing to do. So you need to change something, and use the fact that the variance of W changes with the X's, and there are ways of dealing with it. It's something that one needs to be careful about. Another possibility of getting into trouble is if you're using multiple explanatory variables that are very closely related to each other.

So for example, suppose that I tried to predict your GPA by looking at your SAT the first time that you took it plus your SAT the second time that you took your SATs. I'm assuming that almost everyone takes the SAT more than once. So suppose that you had a model of this kind.

Well, SAT on your first try and SAT on your second try are very likely to be fairly close. And you could think of coming up with estimates in which this is ignored. And you build a model based on this, or an alternative model in which this term is ignored, and you make predictions based on the second SAT. And both models are likely to be essentially as good as the other one, because these two quantities are essentially the same.

So in that case, your theta's that you estimate are going to be very sensitive to little details of the data. You change your data, you have your data, and your data tell you that this coefficient is big and that coefficient is small. You change your data just a tiny bit, and your theta's would drastically change. So this is a case in which you have multiple explanatory variables, but they're redundant in the sense that they're very closely related to each other, and perhaps with a linear relation. So one must be careful about the situation, and do special tests to make sure that this doesn't happen.

Finally the biggest and most common blunder is that you run your linear regression, you get your linear model, and then you say oh, OK. Y is caused by X according to this particular formula. Well, all that we did was to identify a linear relation between X and Y. This doesn't tell us anything. Whether it's Y that causes X, or whether it's X that causes Y, or maybe both X and Y are caused by some other variable that we didn't think about.

So building a good linear model that has small errors does not tell us anything about causal relations between the two variables. It only tells us that there's a close association between the two variables. If you know one you can make predictions about the other. But it doesn't tell you anything about the underlying physics, that there's some physical mechanism that introduces the relation between those variables.

OK, that's it about linear regression. Let us start the next topic, which is hypothesis testing. And we're going to continue with it next time.

So here, instead of trying to estimate continuous parameters, we have two alternative hypotheses about the distribution of the X random variable. So for example our random variable could be either distributed according to this distribution, under H0, or it might be distributed according to this distribution under H1. And we want to make a decision which distribution is the correct one?

So we're given those two distributions, and some common terminologies that one of them is the null hypothesis-- sort of the default hypothesis, and we have some alternative hypotheses-- and we want to check whether this one is true, or that one is true. So you obtain a data point, and you want to make a decision. In this picture what would a reasonable person do to make a decision? They would probably choose a certain threshold, Xi, and decide that H1 is true if your data falls in this interval. And decide that H0 is true if you fall on the side. So that would be a reasonable way of approaching the problem.

More generally you take the set of all possible X's, and you divide the set of possible X's into two regions. One is the rejection region, in which you decide H1, or you reject H0. And the complement of that region is where you decide H0.

So this is the x-space of your data. In this example here, x was one-dimensional. But in general X is going to be a vector, where all the possible data vectors that you can get, they're divided into two types. If it falls in this set you'd make one decision. If it falls in that set, you make the other decision. OK, so how would you characterize the performance of the particular way of making a decision?

Suppose I chose my threshold. I may make mistakes of two possible types. Perhaps H0 is true, but my data happens to fall here. In which case I make a mistake, and this would be a false rejection of H0. If my data falls here I reject H0. I decide H1. Whereas H0 was true. The probability of this happening? Let's call it alpha.

But there's another kind of error that can be made. Suppose that H1 was true, but by accident my data happens to falls on that side. Then I'm going to make an error again. I'm going to decide H0 even though H1 was true. How likely is this to occur? This would be the area under this curve here. And that's the other type of error than can be made, and beta is the probability of this particular type of error.

Both of these are errors. Alpha is the probability of error of one kind. Beta is the probability of an error of the other kind. You would like the probabilities of error to be small. So you would like to make both alpha and beta as small as possible.

Unfortunately that's not possible, there's a trade-off. If I go to my threshold it this way, then alpha become smaller, but beta becomes bigger. So there's a trade-off. If I make my rejection region smaller one kind of error is less likely, but the other kind of error becomes more likely. So we got this trade-off.

So what do we do about it? How do we move systematically? How do we come up with rejection regions? Well, what the theory basically tells you is it tells you how you should create those regions. But it doesn't tell you exactly how. It tells you the general shape of those regions.

For example here, the theory who tells us that the right thing to do would be to put the threshold and make decisions one way to the right, one way to the left. But it might not necessarily tell us where to put the threshold. Still, it's useful enough to know that the way to make a good decision would be in terms of a particular threshold.

Let me make this more specific. We can take our inspiration from the solution of the hypothesis testing problem that we had in the Bayesian case. In the Bayesian case we just pick the hypothesis which is more likely given the data. The produced posterior probabilities using Bayesian rule, they're written this way.

And this term is the same as that term. They cancel out, then let me collect terms here and there. I get an expression here. I think the version you have in your handout is the correct one. The one on the slide was not the correct one, so I'm fixing it here.

OK, so this is the form of how you make decisions in the Bayesian case. What you do in the Bayesian case, you calculate this ratio. Let's call it the likelihood ratio. And compare that ratio to a threshold. And the threshold that you should be using in the Bayesian case has something to do with the prior probabilities of the two hypotheses.

In the non-Bayesian case we do not have prior probabilities, so we do not know how to set this threshold. But we're going to do is we're going to keep this particular structure anyway, and maybe use some other considerations to pick the threshold. So we're going to use a likelihood ratio test, that's how it's called in which we calculate a quantity of this kind that we call the likelihood, and compare it with a threshold.

So what's the interpretation of this likelihood? We ask-- the X's that I have observed, how likely were they to occur if H1 was true? And how likely were they to occur if H0 was true? This ratio could be big if my data are plausible they might occur under H1. But they're very implausible, extremely unlikely to occur under H0.

Then my thinking would be well the data that I saw are extremely unlikely to have occurred under H0. So H0 is probably not true. I'm going to go for H1 and choose H1. So when this ratio is big it tells us that the data that we're seeing are better explained if we assume H1 to be true rather than H0 to be true. So I calculate this quantity, compare it with a threshold, and that's how I make my decision.

So in this particular picture, for example the way it would go would be the likelihood ratio in this picture goes monotonically with my X. So comparing the likelihood ratio to the threshold would be the same as comparing my x to the threshold, and we've got the question of how to choose the threshold.

The way that the threshold is chosen is usually done by fixing one of the two probabilities of error. That is, I say, that I want my error of one particular type to be a given number, so I fix this alpha. And then I try to find where my threshold should be. So that this probability theta, probability out there, is just equal to alpha.

And then the other probability of error, beta, will be whatever it turns out to be. So somebody picks alpha ahead of time. Based on the probability of a false rejection based on alpha, I find where my threshold is going to be. I choose my threshold, and that determines subsequently the value of beta. So we're going to continue with this story next time, and we'll stop here.