6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013 Transcript – Lecture 25

PROFESSOR: OK, if you have not yet done it, please take a moment to go through the course evaluation website and enter your comments for the class. So what we're going to do today to wrap things up is we're going to go through a tour of the world of hypothesis testing. See a few examples of hypothesis tests, starting from simple ones such as the one the setting that we discussed last time in which you just have two hypotheses, you're trying to choose between them.

But also look at more complicated situations in which you have one basic hypothesis. Let's say that you have a fair coin and you want to test it against the hypotheses that your coin is not fair, but that alternative hypothesis is really lots of different hypothesis. So is my coin fair? Is my die fair? Do I have the correct distribution for random variable, and so on. And I'm going to end up with a few general comments about this whole business.

So the sad thing in simple hypothesis testing problems is the following-- we have two possible models, and this is the classical world so we do not have any prior probabilities on the two hypotheses. Usually we want to think of these hypotheses as not being completely symmetrical, but rather one is the default hypothesis, and usually it's referred to as the null hypothesis. And you want to check whether the null hypothesis is true, whether things are normal as you would have expected them to be, or whether it turns out to be false, in which case an alternative hypothesis would be correct.

So how does one go about it? No matter what approach you use, in the end you're going to end up doing the following. You have the space of all simple observations that you may obtain. So when you do the experiment you're going to get an X vector, a vector of data that's somewhere.

And for some vectors you're going to decide that you accept H. Note for some vectors that you reject H0 and you accept H1. So what you will end up doing is that you're going to have some division of the space of all X's into two parts, and one part is the rejection region, and one part is the acceptance region. So if you fall in here you accept H0, if you fall here you'd reject H0.

So to design a hypothesis test basically you need to come up with a division of your X space into two pieces. So the figuring out how to do this involves two elements. One element is to decide what kind of shape so I want for my dividing curve? And having chosen the shape of the dividing curve, where exactly do I put it?

So if you were to cut this space using, let's say, a straight cut you might put it here, or you might put it there, or you might put it there. Where exactly are you going to put it? So let's look at those two steps. The first issue is to decide the general shape of your rejection region, which is the structure of your test. And the way this is done for the case of two hypothesis is by writing down the likelihood ratio between the two hypothesis.

So let's call that quantity l of X. It's something that you can compute given the data that you have. A high value of l of X basically means that this probability here tends to be bigger than this probability. It means that the data that you have seen are quite likely to have occurred under H1, but less likely to have occurred under H0.

So if you see data that they are more plausible, can be better explained, under H1, then this ratio is big, and you're going to choose in favor of H1 or reject H0. That's what you do if you have discrete data. You use the PMFs. If you have densities, in the case of continues data, again you consider the ratio of the two densities.

So a big l of X is evidence that your data are more compatible with H1 rather than H0. Once you accept this kind of structure then your decision is really made in terms of that single number. That is, you had your data that was some kind of vector, and you condense your data into a single number-- a statistic as it's called-- in this case the likelihood ratio, and you put the dividing point somewhere here call it Xi. And in this region you accept H1, in this region you accept H0.

So by committing ourselves to using the likelihood ratio in order to carry out the test we have gone from this complicated picture of finding a dividing line in x-space, to a simpler problem of just finding a dividing point on the real line. OK, how are we going?

So what's left to do is to choose this threshold, Xi. Or as it's called, the critical value, for making our decision. And you can place it anywhere, but one way of deciding where to place it is the following-- look at the distribution of this random variable, l of X. It's has a certain distribution under H0, and it has some other distribution under H1.

If I put my threshold here, here's what's going to happen. When H0 is true, there is this much probability that I'm going to end up making an incorrect decision. If H0 is true there's still a probability that my likelihood ratio will be bigger than Xi, and that's the probability of making an incorrect decision of this particular type. That is of making a false rejection of H0.

Usually one sets this probability to a certain number, alpha. For example alpha being 5 %. And once you decide that you want this to be 5 %, that determines where this number Psi(Xi) is going to be.

So the idea here is that I'm going to reject H0 if the data that I have seen are quite incompatible with H0. if they're quite unlikely to have occurred under H0. And I take this level, 5%. So I see my data and then I say well if H0 was true, the probability that I would have seen data of this kind would be less than 5 %.

Given that I saw those data, that suggests that H0 is not true, and I end up rejecting H0. Now of course there's the other type of error probability. If I put my threshold here, if H1 is true but my likelihood ratio falls here I'm going to make a mistake of the opposite kind. H1 is true, but my likelihood ratio turned out to be small, and I decided in favor of H0.

This is an error of the other kind, this probability of error we call beta. And you can see that there's a trade-off between alpha and beta. If you move your threshold this way alpha become smaller, but beta becomes larger. And the general picture is, in your trade-off, depending on where you put your threshold is as follows-- you can make this beta to be 0 if you put your threshold out here, but in that case you are certain that you're going to make a mistake of the opposite kind. So beta equals 0, alpha equals 1 is one possibility. Beta equals 1 alpha equals 0 is the other possibility if you send your thresholds complete to the other side. And in general you're going to get a trade-off curve of some sort.

And if you want to use a specific value of alpha, for example alpha being 0.05, then that's going to determine for you the probability for beta. Now there's a general, and quite important theorem in statistics, which were are not proving. And which tells us that when we use likelihood ratio tests we get the best possible trade-off curve.

You could think of other ways of making your decisions. Other ways of cutting off your x-space into a rejection and acceptance region. But any other way that you do it is going to end up with some probabilities of error that are going to be above this particular curve.

So the likelihood ratio test turns out to give you the best possible way of dealing with this trade-off between alpha and beta. We cannot minimize alpha and beta simultaneously, there's a trade-off between them. But at least we would like to have a test that deals with this trade-off in the best possible way.

For a given value of alpha we want to have the smallest possible value of beta. And as the theorem is that the likelihood ratio tests do have this optimality property. For a given value of alpha they minimize the probability of error of a different kind.

So let's make all these concrete and look at the simple example. We have two normal distributions with different means. So under H0 you have a mean of 0. Under H1 you have a mean of 1. You get your data, you actually get several data drawn from one of the two distributions. And you want to make a decision, which one of the two is true?

So what you do is you write down the likelihood ratio. The density for a vector of data, if that vector was generated according to H0 -- which is this one, and the density if it was generated according to H1. Since we have multiple data the density of a vector is the product of the densities of the individual elements.

Since we're dealing with normals we have those exponential factors. A product of exponentials gives us an exponential of the sum. I'll spare you the details, but this is the form of the likelihood ratio.

The likelihood ratio test tells us that we should calculate this quantity after we get your data, and compare with a threshold. Now you can do some algebra here, and simplify. And by tracing down the inequalities you're taking logarithms of both sides, and so on. One comes to the conclusion that using a test that has a threshold on this ratio is equivalent to calculating this quantity, and comparing it with a threshold.

Basically this quantity here is monotonic in that quantity. This being larger than the threshold is equivalent to this being larger than the threshold. So this tells us the general structure of the likelihood ratio test in this particular case.

And it's nice because it tells us that we can make our decisions by looking at this simple summary of the data. This quantity, this summary of the data on the basis of which we make our decision is called a statistic. So you take your data, which is a multi-dimensional vector, and you condense it to a single number, and then you make a decision on the basis of that number.

So this is the structure of the test. If I get a large sum of Xi's this is evidence in favor of H1 because here the mean is larger. And so I'm going to decide in favor of H1 or reject H0 if the sum is bigger than the threshold. How do I choose my threshold? Well I would like to choose my threshold so that the probability of an incorrect decision when H0 is true the probability of a false rejection equals to a certain number. Alpha, such as for example 5 %.

So you're given here that this is 5 %. You know the distribution of this random variable, it's normal. And you want to find the threshold value that makes this to be true. So this is a type of problem that you have seen several times. You go to the normal tables, and you figure it out. So the sum of the Xi's has some distribution, it's normal. So that's the distribution of the sum of the Xi's.

And you want this probability here to be alpha. For this to happen what is the threshold value that makes this to be true? So you know how to solve problems of this kind using the normal tables. A slightly different example is one in which you have two normal distributions that have the same mean -- let's take it to be 0 -- but they have a different variance.

So it's sort of natural that here, if your X's that you see are kind of big on either side you would choose H1. If your X's are near 0 then that's evidence for the smaller variance you would choose H0. So to proceed formally you again write down to the form of the likelihood ratio.

So again the density of an X vector under H0 is this one. It's the product of the densities of each one of the Xi's. Product of normal densities gives you a product of exponentials, which is exponential of the sum, and that's the expression that you get.

Under the other hypothesis the only thing that changes is the variance. And the variance, in the normal distribution, shows up here in the denominator of the exponent. So you put it there. So this is the general structure of the likelihood ratio test. And now you do some algebra. These terms are constants comparing this ratio to a constant is the same as just comparing the ratio of the exponentials to a constant.

Then you take logarithms, you want to compare the logarithm of this thing to a constant. You do a little bit of algebra, and in the end you find that the structure of the test is to reject H0 if the sum of the squares of the Xi's is bigger than the threshold.

So by committing to a likelihood ratio test you are told that you should be making it your decision according to a rule of this type. So this fixes the shape or the structure of the decision region, of the rejection region. And the only thing that's left, once more, is to pick this threshold in order to have the property that the probability of a false rejection is equal to say 5 %.

So that's the probability that H0 is true, but the sum of the squares accidentally happens to be bigger than my threshold. In which case I end up deciding H1. How do I find the value of Xi prime? Well what I need to do is to look at the picture, more or less of this kind, but now I need to look at the distribution of the sum of the Xi's squared.

Actually the sum of the Xi's squared is a non-negative random variable. So it's going to have a distribution that's something like this. I look at that distribution, and once more I want this tail probability to be alpha, and that determines where my threshold is going to be. So that's again a simple exercise provided that you know the distribution of this quantity. Do you know it? Well we don't really know it, we have not dealt with this particular distribution in this class. But in principle you should be able to find what it is.

It's a derived distribution problem. You know the distribution of Xi, it's normal. Therefore, by solving a derived distribution problem you can find the distribution of Xi squared. And the Xi squared's are independent of each other, because the Xi's are independent. So you want to find the distribution of the sum of random variables with known distributions. And since they're independent, in principle, you can do this using the convolution formula.

So in principle, and if you're patient enough, you will be able to find the distribution of this random variable. And then you plot it or tabulate it, and find where exactly is the 95th percentile of that distribution, and that determines your threshold. So this distribution actually turns out to have a nice and simple closed-form formula.

Because this is a pretty common test, people have tabulated that distribution. It's called the chi-square distribution. There's tables available for it. And you look up in the tables, you find the 95th percentile of the distribution, and this way you determine your threshold.

So what's the moral of the story? The structure of the likelihood ratio test tells you what kind of decision region you're going to have. It tells you that for this particular test you should be using the sum of the Xi squared's as your statistic, as the basis for making your decision. And then you need to solve a derived distribution problem to find the probability distribution of your statistic. Find the distribution of this quantity under H0, and finally, based on that distribution, after you have derived it, then determine your threshold.

So now let's move on to a somewhat more complicated situation. You have a coin, and you are told that I tried to make a fair coin. Is it fair?

So you have the hypothesis, which is the default-- the null hypothesis-- that the coin is fair. But maybe it isn't. So you have the alternative hypothesis that your coin is not fair. Now what's different in this context is that your alternative hypothesis is not just one specific hypothesis.

Your alternative hypothesis consists of many alternatives. It includes the hypothesis that p is 0.6. It includes the hypothesis that p is 0.51. It includes the hypothesis that p is 0.48, and so on.

So you're testing this hypothesis versus all this family of alternative hypothesis. What you will end up doing is essentially the following-- you get some data. That is, you flip the coin a number of times. Let's say you flip it 1,000 times. You observe some outcome. Let's say you saw 472 heads.

And you ask the question if this hypothesis is true is this value really possible under that hypothesis? Or would it be very much of an outlier? If it looks like an extreme outlier under this hypothesis then I reject it, and I accept the alternative. If this number turns out to be something within the range that you would have expected then you keep, or accept your null hypothesis.

OK so what does it mean to be an outlier or not? First you take your data, and you condense them to a single number. So your detailed data actually would have been a sequence of heads/tails, heads/tails and all that. Any reasonable person would tell you that you shouldn't really care about the exact sequence of heads and tails. Let's just base our decision on the number of heads that we have observed.

So using some kind of reasoning which could be mathematical, or intuitive, or involving artistry-- you pick a one-dimensional, or scalar summary of the data that you have seen. In this case, the summary of the data is just the number of heads that's a quite reasonable one. And so you commit yourself to make a decision on the basis of this quantity.

And you ask the quantity that I'm seeing does it look like an outlier? Or does it look more or less OK? OK, what does it mean to be an outlier? You want to choose the shape of this rejection region, but on the basis of that single number s. And again, the reasonable thing to do in this context would be to argue as follows-- if my coin is fair I expect to see n over 2 heads. That's the expected value.

If the number of heads I see is far from the expected number of heads then I consider this to be an outlier. So if this number is bigger than some threshold Xi. I consider it to be an outlier, and then I'm going to reject my hypothesis.

So we picked our statistic. We picked the general form of how we're going to make our decision, and then we pick a certain significance, or confidence level that we want. Again, this famous 5% number. And we're going to declare something to be an outlier if it lies in the region that has 5% or less probability of occurring.

That is I'm picking my rejection region so that if H0 is true under the default, or null hypothesis, there's only 5% chance that by accident I fall there, and the thing makes me think that H1 is going to be true.

So now what's left to do is to pick the value of this threshold. This is a calculation of the usual kind. I want to pick my threshold, my Xi number so that the probability that s is further from the mean by an amount of Xi is less than 5%. Or that the probability of being inside the acceptance region-- so that the distance from the default is less than my threshold. I want that to be 95%.

So this is an equality that you can get using the central limit theorem and the normal tables. There's 95% probability that the number of heads is going to be within 31 from the correct mean. So the way the exercise is done of course, is that we start with this number, 5%. Which translates to this number 95%. And once we have fixed that number then you ask the question what number should we have here to make this equality to be true?

It's again a problem of this kind. You have a quantity whose distribution you know. Why do you know it? The number of heads by the central limit theorem is approximately normal. So this here talks about the normal distribution. You set your alpha to be 5%, and you ask where should I put my threshold so that this probability of being out there is only 5%?

Now in our particular example the threshold turned out to be 31. This number turned out was just 28 away from the correct mean. So these distance was less than the threshold. So we end up not rejecting H0.

So we have our rejection region. The way we designed it is that when H0 is true there's only a small chance, 5%, that we get to data out of there. Data that we would call an outlier. If we see such an outlier we reject H0. If what we see is not an outlier as in this case, where that distance turned out to be kind of small, then we do not reject H0.

An interesting little piece of language here, people generally prefer to use this terminology-- to say that H0 is not rejected by the data. Instead of saying that H0 is accepted. In some sense they're both saying the same thing, but the difference is sort of subtle. When I say not rejected what I mean is that I got some data that are compatible with my hypothesis.

That is the data that I got do not falsify the hypothesis that I had, my null hypothesis. So my null hypothesis is still alive, and may be true. But from data you can never really prove that the hypothesis is correct. Perhaps my coin is not fair in some other complicated way.

Perhaps I was just lucky, and even though my coin is not fair I ended up with an outcome that suggests that it's fair. Perhaps my coin flips are not independent as I assumed in my model. So there's many ways that my null hypothesis could be wrong, and still I got data that tells me that my hypothesis is OK.

So this is the general way that things work in science. One comes up with a model or a theory. This is the default theory, and we work with that theory trying to find whether there are examples that violate the theory. If you find data and examples that violate the theory your theory is falsified, and you need to look for a new one.

But when you have your theory, really no amount of data can prove that your theory is correct. So we have the default theory that the speed of light is constant as long as we do not find any data that runs counter to it. We stay with that theory, but there's no way of really proving this, no matter how many experiments we do.

But there could be experiments that falsify that theory, in which case we need to do look for a new one. So there's a bit of an asymmetry here in how we treat the alternative hypothesis. H0 is the default which we'll accept until we see some evidence to the contrary. And if we see some evidence to the contrary we reject it. As long as we do not see evidence to the contrary then we keep working with it, but always take it with a grain of salt.

You can never really prove that a coin has a bias exactly equal to 1/2. Maybe the bias is equal to 0.50001, so the bias is not 1/2. But with an experiment with 1,000 coin tosses you wouldn't be able to see this effect.

OK, so that's how you go about testing about whether your coin is fair. You can also think about testing whether a die is fair. So for a die the null hypothesis would be that every possible result when you roll the die has equal probability and equal to 1/6. And you also make the hypothesis that your die rolls are statistically independent from each other.

So I take my die, I roll it a number of times, little n, and I count how many 1's I got, how many 2's I got, how many 3's I got, and these are my data. I count how many times I observed a specific result in my die roll that was equal to sum i.

And now I ask the question-- the Ni's that I observed, are they compatible with my hypothesis or not? What does compatible to my hypothesis mean? Under the null hypothesis Ni should be approximately equal, or is equal in expectation to N times little Pi. And in our example this little Pi is of course 1/6.

So if my die is fair the number of ones I expect to see is equal to the number of rolls times 1/6. The number of 2's I expect to see is again that same number. Of course there's randomness, so I do not expect to get exactly that number. But I can ask how far away from the expected values was i?

If my capital Ni's turn to be very different from N/6 this is evidence that my die is not fair. If those numbers turn out to be close to N times 1/6 then I'm going to say there's no evidence that would lead me to reject this hypothesis. So this hypothesis remains alive.

So someone has come up with this thought that maybe the right statistic to use, or the right way of quantifying how far away are the Ni's from their mean is to look at this quantity. So I'm looking at the expected value of Ni under the null hypothesis. See what I got, take the square of this, and add it over all i's.

But also throw in these terms in the denominator. And why that term is there, that's a longer story. One can write down certain likelihood ratios, do certain Taylor Series approximations, and there's a Heuristic argument that justifies why this would be a good form for the test to use.

So there's a certain art that's involved in this step that some people somehow decided that it's a reasonable thing to do is to calcelate. Once you get your results to calculate this one-dimensional summary of your result, this is going to be your statistic, and compare that statistic to a threshold. And that's how you make your decision.

So by this point we have fixed the type of the rejection region that we're going to have. So we've chosen the qualitative structure of our test, and the only thing that's now left is to choose the particular threshold we're going to use. And the recipe, once more, is the same.

We want to set our threshold so that the probability of a false rejection is 5%. We want the probability that our data fall in here is only 5% when the null hypothesis is true. So that's the same as setting our threshold Xi so that the probability that our test statistic is bigger than that threshold. We want that probability to be only 0.05.

So to solve a problem of this kind what is it that you need to do? You need to find the probability distribution of capital T. So once more it's the same picture. You need to do some calculations of some sort, and come up with the distribution of the random variable T, where T is defined this way. You want to find this distribution under hypothesis H0.

Once you find what that distribution is then you can solve this usual problem. I want this probability here to be 5%. What should my threshold be? So what does this boil down to? Finding the distribution of capital T is in some sense a messy, difficult, derived distribution problem. From this model we know the distribution of the capital Ni's. And actually we can even write down the joint distribution of the capital Ni's.

In fact we can make an approximation here. Capital Ni is a binomial random variable. Let's say the number of 1's that I got in little N rolls off my die. So that's a binomial random variable. When little n is big this is going to be approximately normal. So we have normal random variables, or approximately normal minus a constant. They're still approximately normal. We take the squares of these, scale them so you can solve a derived distribution problem to find the distribution of this quantity.

You can do more work, more derived distribution work, and find the distribution of capital T. So this is a tedious matter, but because this test is used quite often, again people have done those calculations. They have found the distribution of capital T, and it's available in tables. And you go to those tables, and you find the appropriate threshold for making a decision of this type.

Now to give you a sense of how complicated hypothesis one might have to deal with let's make things one level more complicated. So here you can think this X is a discrete random variable. This is the outcome of my roll. And I had a model in which the possible values of my discrete random variables they have probabilities all equal to 1/6.

So my null hypothesis here was a particular PMF for the random variable capital X. So another way of phrasing what happened in this problem was the question is my PMF correct? So this is the PMF of the result of one die roll. You're asking the question is my PMF correct? Make it more complicated.

How about the question of the type is my PDF correct when I have continuous data? So I have hypothesized that's the probability distribution that I have is let's say a particular normal. I get lots of results from that random variable. Can I tell whether my results look like normal or not? What are some ways of going about it?

Well, we saw in the previous slide that there is a methodology for deciding if your PMF is correct. So you could take your normal results, the data that you got from your experiment, and discretize them, and so now you're dealing with discrete data. And sort of used in previous methodology to solve a discrete problem of the type is my PDF correct?

So in practice the way this is done is that you get all your data, let's say data points of this kind. You split your space into bins, and you count how many you have in each bin. So you get this, and that, and that, and nothing. So that's a histogram that you get from the data that you have. Like the very familiar histograms that you see after each one of our quizzes.

So if you look at these histogram, and you ask does it look like normal? OK, we need a systematic way of going about it. If it were normal you can calculate the probability of falling in this interval. The probability of falling in that interval, probability of falling into that interval. So you would have expected values of how many results, or data points, you would have in this interval. And compare these expected values for each interval with the actual ones that you observed. And then take the sum of squares, and so on, exactly as in the previous slide. And this gives you a way of going about it.

This is a little messy. It gets hard to do because you have the difficult decision of how do you choose the bin size? If you take your bins to be very narrow you would get lots of bins with 0's, and a few bins that only have one outcome in them. It probably wouldn't feel right. If you choose your bins to be very wide then you're losing a lot of information. Is there some way of making a test without creating bins?

This is just to illustrate the clever ideas of what statisticians have thought about. And here's a really cute way of going about a test, whether my distribution is correct or not. Here we're essentially plotting a PMF, or an approximation of a PDF. And we ask does it look like the PDF we assumed?

Instead of working with PDFs let's work with cumulative distribution functions. So how does this go? The true normal distribution that I have hypothesized, the density that I'm hypothesizing-- my null hypothesis-- has a certain CDF that I can plot. So supposed that my hypothesis H0 is that the X's are normal with our standard normals, and I plot the CDF of the standard normal, which is the sort of continuous looking curve here.

Now I get my data, and I plot the empirical CDF. What's the empirical CDF? In the empirical CDF you ask the question what fraction of the data fell below 0? You get a number. What fraction of my data fell below 1? I get a number. What fraction of my data fell below 2, and so on.

So you're talking about fractions of the data that fell below each particular number. And by plotting those fractions as a function of this number you get something that looks like a CDF. And it's the CDF suggested by the data.

Now the fraction of the data that fall below 0 in my experiment is-- if my hypothesis were true-- expected to be 1/2. 1/2 is the value of the true CDF. I look at the fraction that I got, it's expected to be that number. But there's randomness, so it's might be a little different than that. For any particular value, the fraction that I got below a certain number-- the fraction of data that we're below, 2, its expectation is the probability of falling below 2, which is the correct CDF.

So if my hypothesis is true the empirical CDF that I get based on data should, when n is large, be very close to the true CDF. So a way of judging whether my model is correct or not is to look at the assumed CDF, the CDF under hypothesis H0. Look at the CDF that I constructed based on the data, and see whether they're close enough or not.

And by close enough, I mean I'm going to look at all the possible X's, and look at the maximum distance between those two curves. And I'm going to have a test that decides in favor of H0 if this distance is small, and in favor of H1 if this distance is large.

That still leaves me the problem of coming up with a threshold. Where exactly do I put my threshold? Because this test is important enough, and is used frequently people have made the effort to try to understand the probability distribution of this quite difficult random variable. One needs to do lots of approximations and clever calculations, but these have led to values and tabulated values for the probability distribution of this random variable.

And, for example, those tabulated values tell us that if we want 5% false rejection probability, then our threshold should be 1.36 divided by the square root of n. So we know where to put our threshold for this particular value. If we want this particular error or error probability to occur.

So that's about as hard and sophisticated classical statistics get. You want to have tests for hypotheses that are not so easy to handle. People somehow think of clever ways of doing tests of this kind. How to compare the theoretical predictions with the observed predictions with the observed data. Come up with some measure of the difference between theory and data, and if that difference is big, than you reject your hypothesis.

OK, of course that's not the end of the field of statistics, there's a lot more. In some ways, as we kept moving through today's lecture, the way that we constructed those rejection regions was more and more ad hoc. I pulled out of a hat a particular measure of fit between data and the model. And I said let's just use a test based on this.

There are attempts at more or less systematic ways of coming up with the general shape of rejection regions that have at least some desirable or favorable theoretical properties. Some more specific problems that people study-- instead of having a test, is this the correct PDF? Yes or no. I just give you data, and I ask you tell me, give me a model or a PDF for those data.

OK, my thoughts of this kind are of many types. One general method is you form a histogram, and then you take your histogram and plot a smooth line, that kind of fits the histogram. This still leaves the question of how do you choose the bins? The bin size in your histograms. How narrow do you take them? And that depends on how many data you have, and there's a lot of theory that tells you about the best way of choosing the bin sizes, and the best ways of smoothing the data that you have.

A completely different topic is in signal processing -- you want to do your inference. Not only you want it to be good, but you also want it to be fast in a computational way. You get data in real time, lots of data. You want to keep processing and revising your estimates and your decisions as they come and go.

Another topic that was briefly touched upon the last couple of lectures is that when you set up a model, like a linear regression model, you choose some explanatory variables, and you try to predict y from your X, these variables. You have a choice of what to take as your explanatory variables. Are there systematic ways of picking the right X variables to try to estimate a Y.

For example should I try to estimate Y on the basis of X? Or on the basis of X-squared? How do I decide between the two?

Finally, the rage these days has to do with anything big, high-demensional. Complicated models of complicated things, and tons and tons of data. So these days data are generated everywhere. The amounts of data are humongous. Also, the problems that people are interested in tend to be very complicated with lots of parameters.

So I need specially tailored methods that can give you good results, or decent results even in the face of these huge amounts of data, and possibly with computational constraints. So with huge amounts of data you want methods that are simple, but still can deliver for you meaningful answers.

Now as I mentioned some time ago, this whole field of statistics is very different from the field of probability. In some sense all that we're doing in statistics is probabilistic calculations. That's what the theory kind of does. But there's a big element of art.

You saw that we chose the shape of some decision regions or rejection regions in a somewhat ad hoc way. There's even more basic things. How do you organize your data? How do you think about which hypotheses you would like to test, and so on. There's a lot of art that's involved here, and there's a lot that can go wrong.

So I'm going to close with a note that you can take either as pessimistic or optimistic. There is a famous paper that came out a few years ago and has been cited about a 1,000 times or so. And the title of the paper is Why Most Published Research Findings Are False. And it's actually a very good argument why, in fields like psychology or the medical science and all that a lot of what you see published-- that yes, this drug has an effect on that particular disease-- is actually false, because people do not do their statistics correctly.

There's lots of biases in what people do. I mean an obvious bias is that you only published a result when you see something. So the null hypothesis is that the drug doesn't work. You do your tests, the drug didn't work, OK, you just go home and cry.

But if by accident that 5% happens, and even though the drug doesn't work, you got some outlier data, and it seemed to be working. Then you're excited, you publish it. So that's clearly a bias. That gets results to be published, even though they do not have a solid foundation behind them.

Then there's another thing, OK? I'm picking my 5%. So H0 is true there's a small probability that the data will look like an outlier, and in that case I published my result. OK it's only 5% -- it's not going to happen too often. But suppose that I go and do a 1,000 different tests? Test H0 against this hypothesis, test H0 against that hypothesis , test H0 against that hypothesis.

Some of these tests, just by accident might turn out to be in favor of H1, and again these are selected to be published. So if you do lots and lots of tests and in each one you have a 5% probability of error, when you consider the collection of all those tests, actually the probability of making incorrect inferences is a lot more than 5%.

One basic principle in being systematic about such studies is that you should first pick your hypothesis that you're going to test, then get your data, and do your hypothesis testing. What would be wrong is to get your data, look at them, and say OK I'm going now to test for these 100 different hypotheses, and I'm going to choose my hypothesis to be for features that look abnormal in my data.

Well, given enough data, you can always find some abnormalities just by chance. And if you choose to make a statistical test-- is this abnormality present? Yes, it will be present. Because you first found the abnormality, and then you tested for it. So that's another way that things can go wrong.

So the moral of this story is that while the world of probability is really beautiful and solid, you have your axioms. Every question has a unique answer that by now you can, all of you, find in a very reliable way. Statistics is a dirty and difficult business. And that's why the subject is not over. And if you're interested in it, it's worth taking follow-on courses in that direction. OK so have good luck in the final, do well, and have a nice vacation afterwards.