PROFESSOR: So for the last three lectures we're going to talk about classical statistics, the way statistics can be done if you don't want to assume a prior distribution on the unknown parameters.

Today we're going to focus, mostly, on the estimation side and leave hypothesis testing for the next two lectures. So where there is one generic method that one can use to carry out parameter estimation, that's the maximum likelihood method. We're going to define what it is.

Then we will look at the most common estimation problem there is, which is to estimate the mean of a given distribution. And we're going to talk about confidence intervals, which refers to providing an interval around your estimates, which has some properties of the kind that the parameter is highly likely to be inside that interval, but we will be careful about how to interpret that particular statement.

Ok. So the big framework first. The picture is almost the same as the one that we had in the case of Bayesian statistics. We have some unknown parameter. And we have a measuring device. There is some noise, some randomness.

And we get an observation, X, whose distribution depends on the value of the parameter. However, the big change from the Bayesian setting is that here, this parameter is just a number. It's not modeled as a random variable. It does not have a probability distribution. There's nothing random about it. It's a constant. It just happens that we don't know what that constant is.

And in particular, this probability distribution here, the distribution of X, depends on Theta. But this is not a conditional distribution in the usual sense of the word.

Conditional distributions were defined when we had two random variables and we condition one random variable on the other. And we used the bar to separate the X from the Theta. To make the point that this is not a conditioned distribution, we use a different notation. We put a semicolon here.

And what this is meant to say is that X has a distribution. That distribution has a certain parameter. And we don't know what that parameter is.

So for example, this might be a normal distribution, with variance 1 but a mean Theta. We don't know what Theta is. And we want to estimate it. Now once we have this setting, then your job is to design this box, the estimator.

The estimator is some data processing box that takes the measurements and produces an estimate of the unknown parameter. Now the notation that's used here is as if X and Theta were one-dimensional quantities.

But actually, everything we say remains valid if you interpret X and Theta as vectors of parameters. So for example, you may obtain several measurements, X1 up to 2Xn. And there may be several unknown parameters in the background.

Once more, we do not have, and we do not want to assume, a prior distribution on Theta. It's a constant. And if you want to think mathematically about this situation, it's as if you have many different probabilistic models.

So a normal with this mean or a normal with that mean or a normal with that mean, these are alternative candidate probabilistic models. And we want to try to make a decision about which one is the correct model.

In some cases, we have to choose just between a small number of models. For example, you have a coin with an unknown bias. The bias is either 1/2 or 3/4. You're going to flip the coin a few times.

And you try to decide whether the true bias is this one or is that one. So in this case, we have two specific, alternative probabilistic models from which we want to distinguish.

But sometimes things are a little more complicated. For example, you have a coin. And you have one hypothesis that my coin is unbiased. And the other hypothesis is that my coin is biased. And you do your experiments. And you want to come up with a decision that decides whether this is true or this one is true.

In this case, we're not dealing with just two alternative probabilistic models. This one is a specific model for the coin. But this one actually corresponds to lots of possible, alternative coin models.

So this includes the model where Theta is 0.6, the model where Theta is 0.7, Theta is 0.8, and so on. So we're trying to discriminate between one model and lots of alternative models.

How does one go about this? Well, there's some systematic ways that one can approach problems of this kind. And we will start talking about these next time.

So today, we're going to focus on estimation problems. In estimation problems, theta is a quantity, which is a real number, a continuous parameter. We're to design this box, so what we get out of this box is an estimate.

Now notice that this estimate here is a random variable. Even though theta is deterministic, this is random, because it's a function of the data that we observe. The data are random. We're applying a function to the data to construct our estimate.

So, since it's a function of random variables, it's a random variable itself. The distribution of Theta hat depends on the distribution of X. The distribution of X is affected by Theta. So in the end, the distribution of your estimate Theta hat will also be affected by whatever Theta happens to be.

Our general objective, when designing estimators, is that we want to get, in the end, an error, an estimation error, which is not too large. But we'll have to make that specific. Again, what exactly do we mean by that?

So how do we go about this problem? One general approach is to pick a Theta, under which the data that we observe, that this is the X's, our most likely to have occurred.

So I observe X. For any given Theta, I can calculate this quantity, which tells me, under this particular Theta, the X that you observed had this probability of occurring. Under that Theta, the X that you observe had that probability of occurring. You just choose that Theta, which makes the data that you observed most likely.

It's interesting to compare this maximum likelihood estimate with the estimates that you would have, if you were in a Bayesian setting, and you were using maximum approach theory probability estimation.

In the Bayesian setting, what we do is, given the data, we use the prior distribution on Theta. And we calculate the posterior distribution of Theta given X. Notice that this is sort of the opposite from what we have here.

This is the probability of X for a particular value of Theta, whereas this is the probability of Theta for a particular X. So it's the opposite type of conditioning. In the Bayesian setting, Theta is a random variable. So we can talk about the probability distribution of Theta.

So how do these two compare, except for this syntactic difference that the order X's and Theta's are reversed? Let's write down, in full detail, what this posterior distribution of Theta is. By the Bayes rule, this conditional distribution is obtained from the prior, and the model of the measurement process that we have. And we get to this expression.

So in Bayesian estimation, we want to find the most likely value of Theta. And we need to maximize this quantity over all possible Theta's.

First thing to notice is that the denominator is a constant. It does not involve Theta. So when you maximize this quantity, you don't care about the denominator. You just want to maximize the numerator.

Now, here, things start to look a little more similar. And they would be exactly of the same kind, if that term here was absent, it the prior was absent. The two are going to become the same if that prior was just a constant.

So if that prior is a constant, then maximum likelihood estimation takes exactly the same form as Bayesian maximum posterior probability estimation. So you can give this particular interpretation of maximum likelihood estimation.

Maximum likelihood estimation is essentially what you have done, if you were in a Bayesian world, and you had assumed a prior on the Theta's that's uniform, all the Theta's being equally likely.

Okay. So let's look at a simple example. Suppose that the Xi's are independent, identically distributed random variables, with a certain parameter Theta. So the distribution of each one of the Xi's is this particular term.

So Theta is one-dimensional. It's a one-dimensional parameter. But we have several data. We write down the formula for the probability of a particular X vector, given a particular value of Theta. But again, when I use the word, given, here it's not in the conditioning sense. It's the value of the density for a particular choice of Theta.

Here, I wrote down, I defined maximum likelihood estimation in terms of PMFs. That's what you would do if the X's were discrete random variables.

Here, the X's are continuous random variables, so instead of I'm using the PDF instead of the PMF. So this a definition, here, generalizes to the case of continuous random variables. And you use F's instead of X's, our usual recipe.

So the maximum likelihood estimate is defined. Now, since the Xi's are independent, the joint density of all the X's together is the product of the individual densities. So you look at this quantity. This is the density or sort of probability of observing a particular sequence of X's.

And we ask the question, what's the value of Theta that makes the X's that we observe most likely? So we want to carry out this maximization. Now this maximization is just a calculational problem.

We're going to do this maximization by taking the logarithm of this expression. Maximizing an expression is the same as maximizing the logarithm. So the logarithm of this expression, the logarithm of a product is the sum of the logarithms. You get contributions from this Theta term. There's n of these, so we get an n log Theta.

And then we have the sum of the logarithms of these terms. It gives us minus Theta. And then the sum of the X's. So we need to maximize this expression with respect to Theta.

The way to do this maximization is you take the derivative, with respect to Theta. And you get n over Theta equals to the sum of the X's. And then you solve for Theta. And you find that the maximum likelihood estimate is this quantity.

Which sort of makes sense, because this is the reciprocal of the sample mean of X's. Theta, in an exponential distribution, we know that it's 1 over (the mean of the exponential distribution). So it looks like a reasonable estimate.

So in any case, this is the estimates that the maximum likelihood estimation procedure tells us that we should report. This formula here, of course, tells you what to do if you have already observed specific numbers. If you have observed specific numbers, then you observe this particular number as your estimate of Theta.

If you want to describe your estimation procedure more abstractly, what you have constructed is an estimator, which is a box that's takes in the random variables, capital X1 up to Capital Xn, and produces out your estimate, which is also a random variable. Because it's a function of these random variables and is denoted by an upper case Theta, to indicate that this is now a random variable.

So this is an equality about numbers. This is a description of the general procedure, which is an equality between two random variables. And this gives you the more abstract view of what we're doing here.

All right. So what can we tell about our estimate? Is it good or is it bad? So we should look at this particular random variable and talk about the statistical properties that it has.

What we would like is this random variable to be close to the true value of Theta, with high probability, no matter what Theta is, since we don't know what Theta is.

Let's make a little more specific the properties that we want. So we cook up the estimator somehow. So this estimator corresponds, again, to a box that takes data in, the capital X's, and produces an estimate Theta hat.

This estimate is random. Sometimes it will be above the true value of Theta. Sometimes it will be below. Ideally, we would like it to not have a systematic error, on the positive side or the negative side. So a reasonable wish to have, for a good estimator, is that, on the average, it gives you the correct value.

Now here, let's be a little more specific about what that expectation is. This is an expectation, with respect to the probability distribution of Theta hat. The probability distribution of Theta hat is affected by the probability distribution of the X's. Because Theta hat is a function of the X's.

And the probability distribution of the X's is affected by the true value of Theta. So depending on which one is the true value of Theta, this is going to be a different expectation. So if you were to write this expectation out in more detail, it would look something like this.

You need to write down the probability distribution of Theta hat. And this is going to be some function. But this function depends on the true Theta, is affected by the true Theta. And then you integrate this with respect to Theta hat.

What's the point here? Again, Theta hat is a function of the X's. So the density of Theta hat is affected by the density of the X's. The density of the X's is affected by the true value of Theta. So the distribution of Theta hat is affected by the value of Theta.

Another way to put it is, as I've mentioned a few minutes ago, in this business, it's as if we are considering different possible probabilistic models, one probabilistic model for each choice of Theta. And we're trying to guess which one of these probabilistic models is the true one.

One way of emphasizing the fact that this expression depends on the true Theta is to put a little subscript here, expectation, under the particular value of the parameter Theta. So depending on what value the true parameter Theta takes, this expectation will have a different value.

And what we would like is that no matter what the true value is, that our estimate will not have a bias on the positive or the negative sides. So this is a property that's desirable.

Is it always going to be true? Not necessarily, it depends on what estimator we construct. Is it true for our exponential example? Unfortunately not, the estimate that we have in the exponential example turns out to be biased.

And one extreme way of seeing this is to consider the case where our sample size is 1. We're trying to estimate Theta. And the estimator from the previous slide, in that case, is just 1/X1. Now X1 has a fair amount of density in the vicinity of 0, which means that 1/X1 has significant probability of being very large.

And if you do the calculation, this ultimately makes the expected value of 1/X1 to be infinite. Now infinity is definitely not the correct value. So our estimate is biased upwards. And it's actually biased a lot upwards.

So that's how things are. Maximum likelihood estimates, in general, will be biased. But under some conditions, they will turn out to be asymptotically unbiased.

That is, as you get more and more data, as your X vector is longer and longer, with independent data, the estimate that you're going to have, the expected value of your estimator is going to get closer and closer to the true value. So you do have some nice asymptotic properties, but we're not going to prove anything like this.

Speaking of asymptotic properties, in general, what we would like to have is that, as you collect more and more data, you get the correct answer, in some sense. And the sense that we're going to use here is the limiting sense of convergence in probability, since this is the only notion of convergence of random variables that we have in our hands.

This is similar to what we had in the pollster problem, for example. If we had a bigger and bigger sample size, we could be more and more confident that the estimate that we obtained is close to the unknown true parameter of the distribution that we have.

So this is a desirable property. If you have an infinitely large amount of data, you should be able to estimate an unknown parameter more or less exactly. So this is it desirable property of estimators.

It turns out that maximum likelihood estimation, given independent data, does have this property, under mild conditions. So maximum likelihood estimation, in this respect, is a good approach.

So let's see, do we have this consistency property in our exponential example? In our exponential example, we used this quantity to estimate the unknown parameter Theta. What properties does this quantity have as n goes to infinity?

Well this quantity is the reciprocal of that quantity up here, which is the sample mean. We know from the weak law of large numbers, that the sample mean converges to the expectation. So this property here comes from the weak law of large numbers.

In probability, this quantity converges to the expected value, which, for exponential distributions, is 1/Theta. Now, if something converges to something, then the reciprocal of that should converge to the reciprocal of that. That's a property that's certainly correct for numbers.

But you're not talking about convergence of numbers. We're talking about convergence in probability, which is a more complicated notion.

Fortunately, it turns out that the same thing is true, when we deal with convergence in probability. One can show, although we will not bother doing this, that indeed, the reciprocal of this, which is our estimate, converges in probability to the reciprocal of that. And that reciprocal is the true parameter Theta.

So for this particular exponential example, we do have the desirable property, that as the number of data becomes larger and larger, the estimate that we have constructed will get closer and closer to the true parameter value.

And this is true no matter what Theta is. No matter what the true parameter Theta is, we're going to get close to it as we collect more data.

Okay. So these are two rough qualitative properties that would be nice to have. If you want to get a little more quantitative, you can start looking at the mean squared error that your estimator gives.

Now, once more, the comment I was making up there applies. Namely, that this expectation here is an expectation with respect to the probability distribution of Theta hat that corresponds to a particular value of little theta.

So fix a little theta. Write down this expression. Look at the probability distribution of Theta hat, under that little theta. And do this calculation. You're going to get some quantity that depends on the little theta.

And so all quantities in this equality here should be interpreted as quantities under that particular value of little theta. So if you wanted to make this more explicit, you could start throwing little subscripts everywhere in those expressions.

And let's see what those expressions tell us. The expected value squared of a random variable, we know that it's always equal to the variance of this random variable, plus the expectation of the random variable squared. So the expectation value of that random variable, squared.

This equality here is just our familiar formula, that the expected value of X squared is the variance of X plus the expected value of X squared. So we apply this formula to X equal to Theta hat minus Theta.

Now, remember that, in this classical setting, theta is just a constant. We have fixed Theta. We want to calculate the variance of this quantity, under that particular Theta. When you add or subtract a constant to a random variable, the variance doesn't change. This is the same as the variance of our estimator.

And what we've got here is the bias of our estimate. It tells us, on the average, whether we fall above or below. And we're taking the bias to be b squared. If we have an unbiased estimator, the bias term will be 0.

So ideally we want Theta hat to be very close to Theta. And since Theta is a constant, if that happens, the variance of Theta hat would be very small. So Theta is a constant. If Theta hat has a distribution that's concentrated just around own little theta, then Theta hat would have a small variance.

So this is one desire that have. We're going to have a small variance. But we also want to have a small bias at the same time.

So the general form of the mean squared error has two contributions. One is the variance of our estimator. The other is the bias. And one usually wants to design an estimator that simultaneously keeps both of these terms small.

So here's an estimation method that would do very well with respect to this term, but badly with respect to that term. So suppose that my distribution is, let's say, normal with an unknown mean Theta and variance 1.

And I use as my estimator something very dumb. I always produce an estimate that says my estimate is 100. So I'm just ignoring the data and report 100. What does this do?

The variance of my estimator is 0. There's no randomness in the estimate that I report. But the bias is going to be pretty bad. The bias is going to be Theta hat, which is 100 minus the true value of Theta.

And for some Theta's, my bias is going to be horrible. If my true Theta happens to be 0, my bias squared is a huge term. And I get a large error.

So what's the moral of this example? There are ways of making that variance very small, but, in those cases, you pay a price in the bias. So you want to do something a little more delicate, where you try to keep both terms small at the same time.

So these types of considerations become important when you start to try to design sophisticated estimators for more complicated problems. But we will not do this in this class. This belongs to further classes on statistics and inference.

For this class, for parameter estimation, we will basically stick to two very simple methods. One is the maximum likelihood method we've just discussed. And the other method is what you would do if you were still in high school and didn't know any probability.

You get data. And these data come from some distribution with an unknown mean. And you want to estimate that the unknown mean. What would you do? You would just take those data and average them out.

So let's make this a little more specific. We have X's that come from a given distribution. We know the general form of the distribution, perhaps. We do know, perhaps, the variance of that distribution, or, perhaps, we don't know it. But we do not know the mean.

And we want to estimate the mean of that distribution. Now, we can write this situation. We can represent it in a different form. The Xi's are equal to Theta. This is the mean. Plus a 0 mean random variable, that you can think of as noise.

So this corresponds to the usual situation you would have in a lab, where you go and try to measure an unknown quantity. You get lots of measurements. But each time that you measure them, your measurements have some extra noise in there. And you want to kind of get rid of that noise.

The way to try to get rid of the measurement noise is to collect lots of data and average them out. This is the sample mean. And this is a very, very reasonable way of trying to estimate the unknown mean of the X's.

So this is the sample mean. It's a reasonable, plausible, in general, pretty good estimator of the unknown mean of a certain distribution. We can apply this estimator without really knowing a lot about the distribution of the X's.

Actually, we don't need to know anything about the distribution. We can still apply it, because the variance, for example, does not show up here. We don't need to know the variance to calculate that quantity.

Does this estimator have good properties? Yes, it does. What's the expected value of the sample mean? If the expectation of this, it's the expectation of this sum divided by n. The expected value for each one of the X's is Theta. So the expected value of the sample mean is just Theta itself.

So our estimator is unbiased. No matter what Theta is, our estimator does not have a systematic error in either direction. Furthermore, the weak law of large numbers tells us that this quantity converges to the true parameter in probability. So it's a consistent estimator. This is good.

And if you want to calculate the mean squared error corresponding to this estimator. Remember how we defined the mean squared error? It's this quantity. Then it's a calculation that we have done a fair number of times by now.

The mean squared error is the variance of the distribution of the X's divided by n. So as we get more and more data, the mean squared error goes down to 0.

In some examples, it turns out that the sample mean is also the same as the maximum likelihood estimate. For example, if the X's are coming from a normal distribution, you can write down the likelihood, do the maximization with respect to Theta, you'll find that the maximum likelihood estimate is the same as the sample mean.

In other cases, the sample mean will be different from the maximum likelihood. And then you have a choice about which one of the two you would use. Probably, in most reasonable situations, you would just use the sample mean, because it's simple, easy to compute, and has nice properties.

All right. So you go to your boss. And you report and say, OK, I did all my experiments in the lab. And the average value that I got is a certain number, 2.37. So is that the informative to your boss?

Well your boss would like to know how much they can trust this number, 2.37. Well, I know that the true value is not going to be exactly that. But how close should it be? So give me a range of what you think are possible values of Theta.

So the situation is like this. So suppose that we observe X's that are coming from a certain distribution. And we're trying to estimate the mean. We get our data. Maybe our data looks something like this.

You calculate the mean. You find the sample mean. So let's suppose that the sample mean is a number, for some reason take to be 2.37. But you want to convey something to your boss about how spread out these data were.

So the boss asks you to give him or her some kind of interval on which Theta, the true parameter, might lie. So the boss asked you for an interval. So what you do is you end up reporting an interval.

And you somehow use the data that you have seen to construct this interval. And you report to your boss also the endpoints of this interval. Let's give names to these endpoints, Theta_n- and Theta_n+. The ends here just play the role of keeping track of how many data we're using.

So what you report to your boss is this interval as well. Are these Theta's here, the endpoints of the interval, lowercase or uppercase? What should they be? Well you construct these intervals after you see your data.

You take the data into account to construct your interval. So these definitely should depend on the data. And therefore they are random variables. Same thing with your estimator, in general, it's going to be a random variable. Although, when you go and report numbers to your boss, you give the specific realizations of the random variables, given the data that you got.

So instead of having just a single box that produces estimates. So our previous picture was that you have your estimator that takes X's and produces Theta hats. Now our box will also be producing Theta hats minus and Theta hats plus. It's going to produce an interval as well.

The X's are random, therefore these quantities are random. Once you go and do the experiment and obtain your data, then your data will be some lowercase x, specific numbers. And then your estimates and estimator become also lower case.

What would we like this interval to do? We would like it to be highly likely to contain the true value of the parameter. So we might impose some specs of the following kind.

I pick a number, alpha. Usually that alpha, think of it as a probability of a large error. Typical value of alpha might be 0.05, in which case this number here is point 0.95.

And you're given specs that say something like this. I would like, with probability at least 0.95, this to happen, which says that the true parameter lies inside the confidence interval.

Now let's try to interpret this statement. Suppose that you did the experiment, and that you ended up reporting to your boss a confidence interval from 1.97 to 2.56. That's what you report to your boss.

And suppose that the confidence interval has this property. Can you go to your boss and say, with probability 95%, the true value of Theta is between these two numbers? Is that a meaningful statement?

So the statement is, the tentative statement is, with probability 95%, the true value of Theta is between 1.97 and 2.56. Well, what is random in that statement? There's nothing random. The true value of theta is a constant. 1.97 is a number. 2.56 is a number.

So it doesn't make any sense to talk about the probability that theta is in this interval. Either theta happens to be in that interval, or it happens to not be. But there are no probabilities associated with this. Because theta is not random.

Syntactically, you can see this. Because theta here is a lower case. So what kind of probabilities are we talking about here? Where's the randomness? Well the random thing is the interval. It's not theta.

So the statement that is being made here is that the interval, that's being constructed by our procedure, should have the property that, with probability 95%, it's going to fall on top of the true value of theta.

So the right way of interpreting what the 95% confidence interval is, is something like the following. We have the true value of theta that we don't know. I get data. Based on the data, I construct a confidence interval. I get my confidence interval. I got lucky. And the true value of theta is in here.

Next day, I do the same experiment, take my data, construct a confidence interval. And I get this confidence interval, lucky once more. Next day I get data. I use my data to come up with an estimate of theta and the confidence interval.

That day, I was unlucky. And I got a confidence interval out there. What the requirement here is, is that 95% of the days, where we use this certain procedure for constructing confidence intervals, 95% of those days, we will be lucky. And we will capture the correct value of theta by your confidence interval.

So it's a statement about the distribution of these random confidence intervals, how likely are they to fall on top of the true theta, as opposed to how likely they are to fall outside. So it's a statement about probabilities associated with a confidence interval. They're not probabilities about theta, because theta, itself, is not random.

So this is what the confidence interval is, in general, and how we interpret it. How do we construct a 95% confidence interval? Let's go through this exercise, in a particular example.

The calculations are exactly the same as the ones that you did when we talked about laws of large numbers and the central limit theorem. So there's nothing new calculationally but it's, perhaps, new in terms of the language that we use and the interpretation.

So we got our sample mean from some distribution. And we would like to calculate a 95% confidence interval. We know from the normal tables, that the standard normal has 2.5% on the tail, that's after 1.96.

Yes, by this time, the number 1.96 should be pretty familiar. So if this probability here is 2.5%, this number here is 1.96.

Now look at this random variable here. This is the sample mean. Difference, from the true mean, normalized by the usual normalizing factor. By the central limit theorem, this is approximately normal. So it has probability 0.95 of being less than 1.96.

Now take this event here and rewrite it. This the event, well, that Theta hat minus theta is bigger than this number and smaller than that number. This event here is equivalent to that event here.

And so this suggests a way of constructing our 95% percent confidence interval. I'm going to report the interval, which gives this as the lower end of the confidence interval, and gives this as the upper end of the confidence interval

In other words, at the end of the experiment, we report the sample mean, which is our estimate. And we report also, an interval around the sample mean. And this is our 95% confidence interval.

The confidence interval becomes smaller, when n is larger. In some sense, we're more certain that we're doing a good estimation job, so we can have a small interval and still be quite confident that our interval captures the true value of the parameter.

Also, if our data have very little noise, when you have more accurate measurements, you're more confident that your estimate is pretty good. And that results in a smaller confidence interval, smaller length of the confidence interval. And still you have 95% probability of capturing the true value of theta.

So we did this exercise by taking 95% confidence intervals and the corresponding value from the normal tables, which is 1.96.

Of course, you can do it more generally, if you set your alpha to be some other number. Again, you look at the normal tables. And you find the value here, so that the tail has probability alpha over 2.

And instead of using these 1.96, you use whatever number you get from the normal tables. And this tells you how to construct a confidence interval.

Well, to be exact, this is not necessarily a 95% confidence interval. It's approximately a 95% confidence interval. Why is this? Because we've done an approximation. We have used the central limit theorem.

So it might turn out to be a 95.5% confidence interval instead of 95%, because our calculations are not entirely accurate. But for reasonable values of n, using the central limit theorem is a good approximation. And that's what people almost always do.

So just take the value from the normal tables. Okay, except for one catch. I used the data. I obtained my estimate. And I want to go to my boss and report this theta minus and theta hat, which is the confidence interval.

What's the difficulty? I know what n is. But I don't know what sigma is, in general. So if I don't know sigma, what am I going to do?

Here, there's a few options for what you can do. And the first option is familiar from what we did when we talked about the pollster problem. We don't know what sigma is, but maybe we have an upper bound on sigma.

For example, if the Xi's Bernoulli random variables, we have seen that the standard deviation is at most 1/2. So use the most conservative value for sigma. Using the most conservative value means that you take bigger confidence intervals than necessary.

So that's one option. Another option is to try to estimate sigma from the data. How do you do this estimation? In special cases, for special types of distributions, you can think of heuristic ways of doing this estimation.

For example, in the case of Bernoulli random variables, we know that the true value of sigma, the standard deviation of a Bernoulli random variable, is the square root of theta1 minus theta, where theta is the mean of the Bernoulli.

Try to use this formula. But theta is the thing we're trying to estimate in the first place. We don't know it. What do we do? Well, we have an estimate for theta, the estimate, produced by our estimation procedure, the sample mean.

So I obtain my data. I get my data. I produce the estimate theta hat. It's an estimate of the mean. Use that estimate in this formula to come up with an estimate of my standard deviation. And then use that standard deviation, in the construction of the confidence interval, pretending that this is correct.

Well the number of your data is large, then we know, from the law of large numbers, that theta hat is a pretty good estimate of theta. So sigma hat is going to be a pretty good estimate of sigma. So we're not making large errors by using this approach.

So in this scenario here, things were simple, because we had an analytical formula. Sigma was determined by theta. So we could come up with a quick and dirty estimate of sigma.

In general, if you do not have any nice formulas of this kind, what could you do? Well, you still need to come up with an estimate of sigma somehow. What is a generic method for estimating a standard deviation? Equivalently, what could be a generic method for estimating a variance?

Well the variance is an expected value of some random variable. The variance is the mean of the random variable inside of those brackets. How does one estimate the mean of some random variable?

You obtain lots of measurements of that random variable and average them out. So this would be a reasonable way of estimating the variance of a distribution. And again, the weak law of large numbers tells us that this average converges to the expected value of this, which is just the variance of the distribution.

So we got a nice and consistent way of estimating variances. But now, we seem to be getting in a vicious circle here, because to estimate the variance, we need to know the mean. And the mean is something we're trying to estimate in the first place.

Okay. But we do have an estimate from the mean. So a reasonable approximation, once more, is to plug-in, here, since we don't know the mean, the estimate of the mean. And so you get that expression, but with a theta hat instead of theta itself.

And this is another reasonable way of estimating the variance. It does have the same consistency properties. Why? When n is large, this is going to behave the same as that, because theta hat converges to theta.

And when n is large, this is approximately the same as sigma squared. So for a large n, this quantity also converges to sigma squared. And we have a consistent estimate of the variance as well. And we can take that consistent estimate and use it back in the construction of confidence interval.

One little detail, here, we're dividing by n. Here, we're dividing by n-1. Why do we do this? Well, it turns out that's what you need to do for these estimates to be an unbiased estimate of the variance. One has to do a little bit of a calculation, and one finds that that's the factor that you need to have here in order to be unbiased.

Of course, if you get 100 data points, whether you divide by 100 or divided by 99, it's going to make only a tiny difference in your estimate of your variance.

So it's going to make only a tiny difference in your estimate of the standard deviation. It's not a big deal. And it doesn't really matter. But if you want to show off about your deeper knowledge of statistics, you throw in the 1 over n-1 factor in there.

So now one basically needs to put together this story here, how you estimate the variance. You first estimate the sample mean. And then you do some extra work to come up with a reasonable estimate of the variance and the standard deviation. And then you use your estimate, of the standard deviation, to come up with a confidence interval, which has these two endpoints.

In doing this procedure, there's basically a number of approximations that are involved. There are two types of approximations. One approximation is that we're pretending that the sample mean has a normal distribution. That's something we're justified to do, by the central limit theorem. But it's not exact. It's an approximation.

And the second approximation that comes in is that, instead of using the correct standard deviation, in general, you will have to use some approximation of the standard deviation.

Okay so you will be getting a little bit of practice with these concepts in recitation and tutorial. And we will move on to new topics next week. But the material that's going to be covered in the final exam is only up to this point. So next week is just general education. Hopefully useful, but it's not in the exam.