CH5230: System Identification

Frisher's information and properties of estimators

Part 07

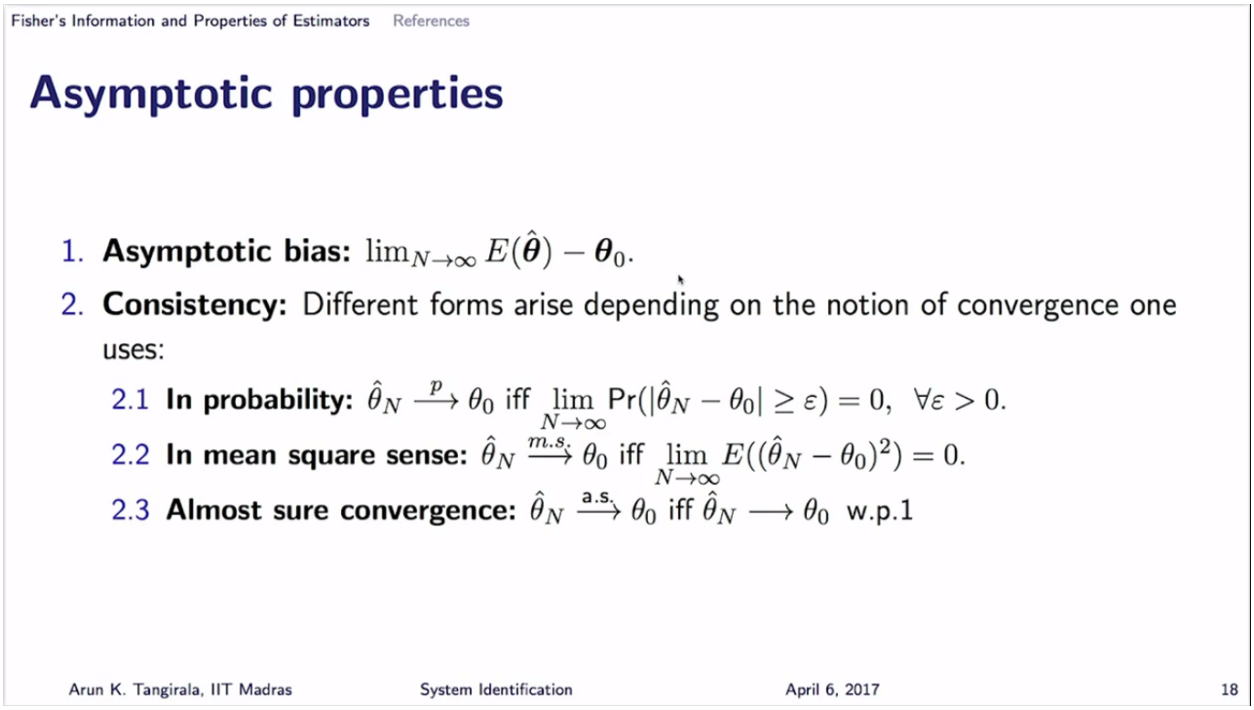
I apologize for the delay. So I just wanted to quickly now conclude on the properties of estimators and more on the least square estimators. The TAs have sent you links to the video lectures on methods of estimation. I expect you to, as I've said earlier to sit through those videos and understand the methods of estimation. What I'll do in the classes do a quick review of those methods and go through a few examples in MATLAB. So let's get going. Yesterday we talked extensively about efficiency, Cramér–Rao's inequality. How to derive the most efficient estimator without resorting to any other method per say, like least squares and so on. And then we talked about best linear unbiased estimator because the minimum-variance unbiased estimatorin general may not exist. That is it may be an ideal dream that you cannot realise. Whereas a best linear unbiased estimator can be built. Only when the observation errors are Gaussian and white then the blue coincides with the MVUE.

So those were the statistical properties that we reviewed. The other set of properties as I've mentioned always is the set of asymptotic properties. The difference between statistical and asymptotic properties is that when we look at statistical properties namely bias variance and mean square error, we fixed the sample size and then we walk across the realisations and ask how they estimator performs across the realizations. The asymptotic properties also look at that but they also, they tend to look at the sample behaviour. How? The estimatorperforms as it is presented with more and more observations. Whereas the statistical properties are evaluated for a fixed sample size. And in that sense asymptotic properties are of more interest particularly in situations where you can afford to have large samples. I must tell you that there are many situations where the number of samples is quite limited or observations is quite limited. And you must be dealing with a small sample size. Then these asymptotic properties do not come into picture. Because they are specifically meant for the large sample case.

In engineering arena, you can think of, or you can afford to think of having large samples but in many other domains such as biological, systems biology, or maybe you know, market surveys or anything to do with health and medicine and so on. It's very difficult to think of the large sample case, in which case we should not be thinking of the asymptotic property. So you should be aware as to what properties one should be interested in depending on the scenario. So the main asymptotic property of interest as the lectures must told is consistency which basically looks at how close theta hat gets to, whether theta hat gets converges to theta not as N goes to infinity.

Now the only thing that you have to remember of course that is asymptomatic bias but consistency takes care of asymptomatic bias as well. That's why I'm only focusing on consistency. So what we mean my consistency is, the other way of reading consistency is convergence whether the parameter estimates converge to the truth.

(Refer Slide Time: 04:05)



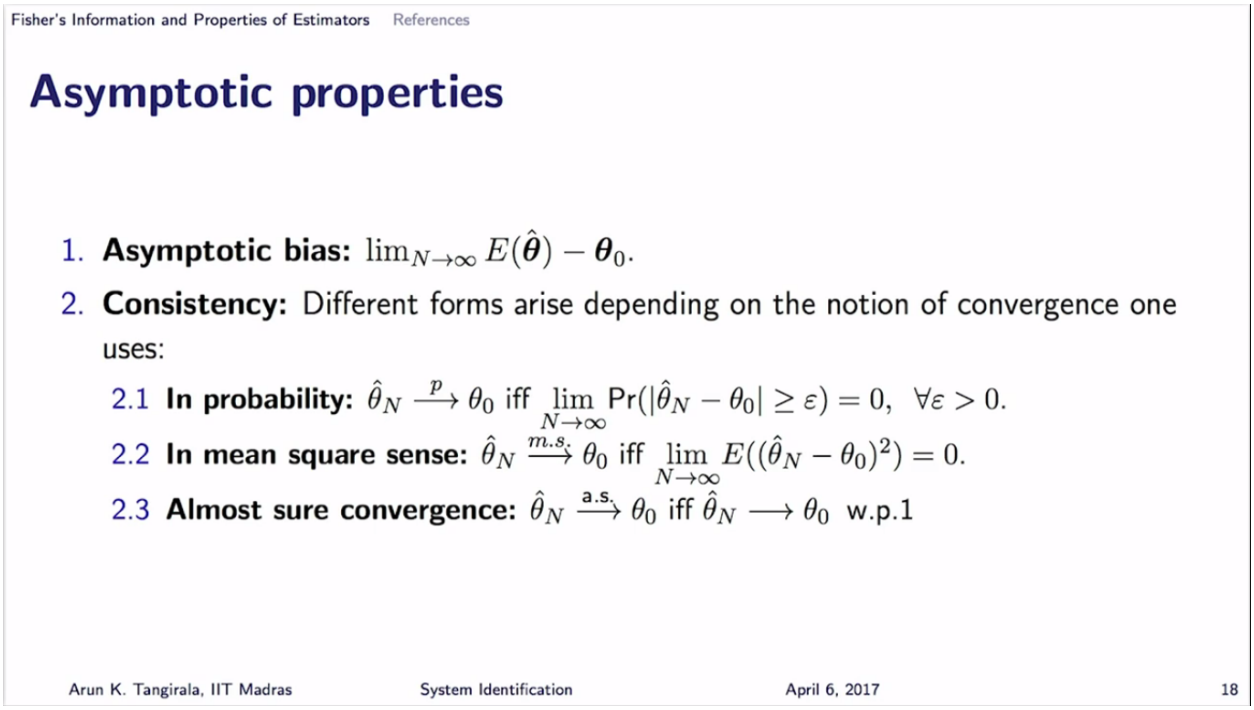
The only thing you have to remember is technically this is not such an easy problem because theta hat is a random variable and we assume truth to be deterministic. So we want the random variable that is a sequence of random variables. You should just remember this. Whether you understand the full technicality or not all you have to remember is, that you have a sequence of parameter estimates, assume a single parameter that you have as a function of the sample size. The subscript indicates the sample size. From one observation you build an estimate, from two observations you build another estimate, N observations you build another estimate and so on. So as I keep increasing the sample size, I would generate a sequence of parameter estimates. And the question is as N goes to infinity whether the sequence converges to the truth, that is the idea behind consistency. That's all.

Now, having made it look very simple. Technically I said,it's lot more involved because now you're dealing with a sequence of random variables. Remember theta hat is a random variable. It is not like your regular sequence one, half, one fourth and so on. It's not a sequence of deterministic numbers. You're looking at sequence of random variables, therefore special notions of convergence are required and those are the notions that I have described in the video lectures. You should sit through them. And depending now on how you define your convergence, you have different forms of consistency ranging from the weakest to the strongest.

So please sit through those videos and understand what are the different notions of convergence of a sequence of random variables. As I said, this notion of convergence is different from the regular convergence that we learn in maybe you know, when you are 11th and 12th standard or 1st year undergraduate mathematics course. The three different forms of convergence are convergence and probability, convergence in mean square error, and almost sure convergence. In the order of increasing rigor. And all three are talking about whether theta hat N gets close to theta not. Right. In fact, the probabilistic form of convergence says that theta hat N gets very close to theta not but does not necessarily go and sit at theta not. It is hovering within an epsilon radius of the truth that's how you-- and this Epsilon can be arbitrarily small but not zero.

Mean square error says that as N goes to infinity the distance of theta hat from theta not average distance goes to zero. So you're in fact talking of kind of the second moment of theta hat N with respect to theta not. In fact the reason it's called mean square because if you look at the statement, it says expectation of thetahat N minus theta not square. What is that? What is that expectation of theta hatN minus theta not to the whole square. What properties is it?

(Refer Slide Time: 07:48)



What is it?Variance.Seriously. We have discussed at length yesterday and day before. What is expectation of theta hat N minus theta not to the whole square. The answer is on the screen.Then why do you say variance? It's varianceonly if you're looking at unbiased estimators. So the mean square consistency is

asking whether the mean square error goes to zero as N goes to infinity. And the third one is the strongest form of consistency which has got to do with almost sure convergence. Now, this almost sure convergence is for all practical purposestheta hat as follows,theta hat actually converges to thetanot for all practical purposes. But there are these rare situations where it doesn't. Okay. There are this extremely rare situations that it may not. That's why we're saying almost sure. That means there is a very remote possibility that theta hat N may not go and sit at exactly at theta not. But it will for a lot of times go in and sitat thetanot. That is what we mean we're almost sure.

I have stated in very descriptive and loose terms, the exact meaning of almost sure convergence. I've explained in the lectures you should sit through and understand. The almost sure convergence has got to do with this notion of events that occurred but they have probability 0. It's strange, when you think of probability. How can you say that an event occurs on the probability is 0. Can you give an example? Can you give an example of an event that occurs but the probability is 0. Any examples? It is very easy.

Continues.

Correct. So in the continues valued random variable case, you have that. In fact it's an irony. At every point the probability is 0 but over on interval it isn't, correct. So it is possible that the random variable exactly achieves that value but as but our measure of probability that the probability 0 because probability is a measure. So it is possible that the random variable exactly goes and attains that value but from our measure theory viewpoint probably 0.

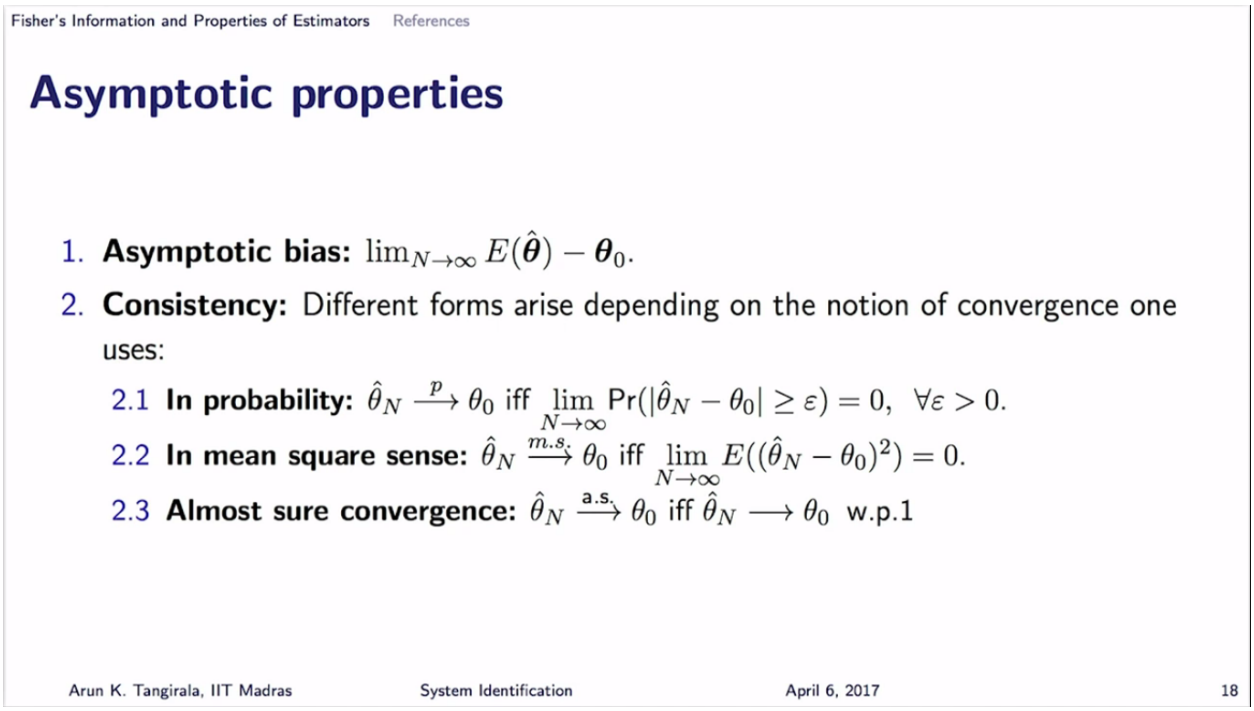
So it's almost sure convergence is related to that. Exactly theta hat N equal the theta not. If you ask what is probably that theta hat N will exactly equal theta not that is 0 for continuous value parameters. So that is why we say that it is almost sure convergence. But leaving aside the technicality, what this means is, if you are able to show that an estimatorsatisfies the, it has the almost sure consistency property. We say that theta hat N reaches theta not with probability 1. That's a very strong statement that you are making. And if you're able to make such a statement for an estimator,if you're able to prove that then the other two follow, the means square consistency and consistency in probability also convergence and probability also holds.

Generally what one tries to prove is mean square consistency. Of course, if you can prove the third one it's great. But if you cannot then you try to prove that mean square consistency and if you cannot prove that at least you show the weak form of consistency. But you have to work with an estimator that is consistent. That's extremely important because if that is not the case, your entire effort of collecting more observations will go futile. It will go in vein, right, because you collect more observations with a firm

belief that as you collect more data your parameter estimate will improve. But if the estimator is such that, how many ever data points you collect. As I said periodogram for example as an estimate of spectral density is an inconsistent estimator. How many ever data points you collect the periodogram never gets you the truespectral density. So do not be under the impression thatalways increasing the number of observations will improve estimate. It depends on how your estimating. Okay.

So generally, the two desirable properties of an estimator in the order is consistency. One is consistency. Right. And the other is efficiency. These are the two different forms properties that one desires.

(Refer Slide Time: 12:53)



And once you show mean square consistency for example, since mean square error is the sum of the bias square and variance asymptotically also the estimator becomes unbiased. So this asymptotic bias is saying that, for finite samples a bias may exist but for large samples the bias can vanish. Right.So for example, in the variance estimators that we wrote on the board yesterday. We wrote two different estimators, one which has a one over N minus 1as a factor and the other that has 1 over N. The one that has N minus 1 in the denominator is an unbiased estimator. As a consequence the other one that has 1 over N in the denominator is a biased estimator. However when N is very large it doesn't matter whether you have 1over N minus 1 or 1over N. Therefore although for finite samples size the sigma square hat N is a biased estimator it is asymptotically unbiased estimator. At least you should say for large samples the bias should vanish.

So those are the desirable properties of an estimator. Any estimator that you take, you should first ask if it is consistent. Then you should ask if it is the most efficient. What is efficiency got to do with?Correct. Exactly.Very good.Minimum variance. That is,it will give the least errors among all the estimators. As long as the estimators these two properties. It is very good. Maximum likelihood estimator have this property and therefore that very popular. All right.Least squares estimators also are consistent and most efficient but only under some restricted conditions. Okay. That we'll talk about briefly when we when we talk of Least square methods and Émilemethods. Okay.

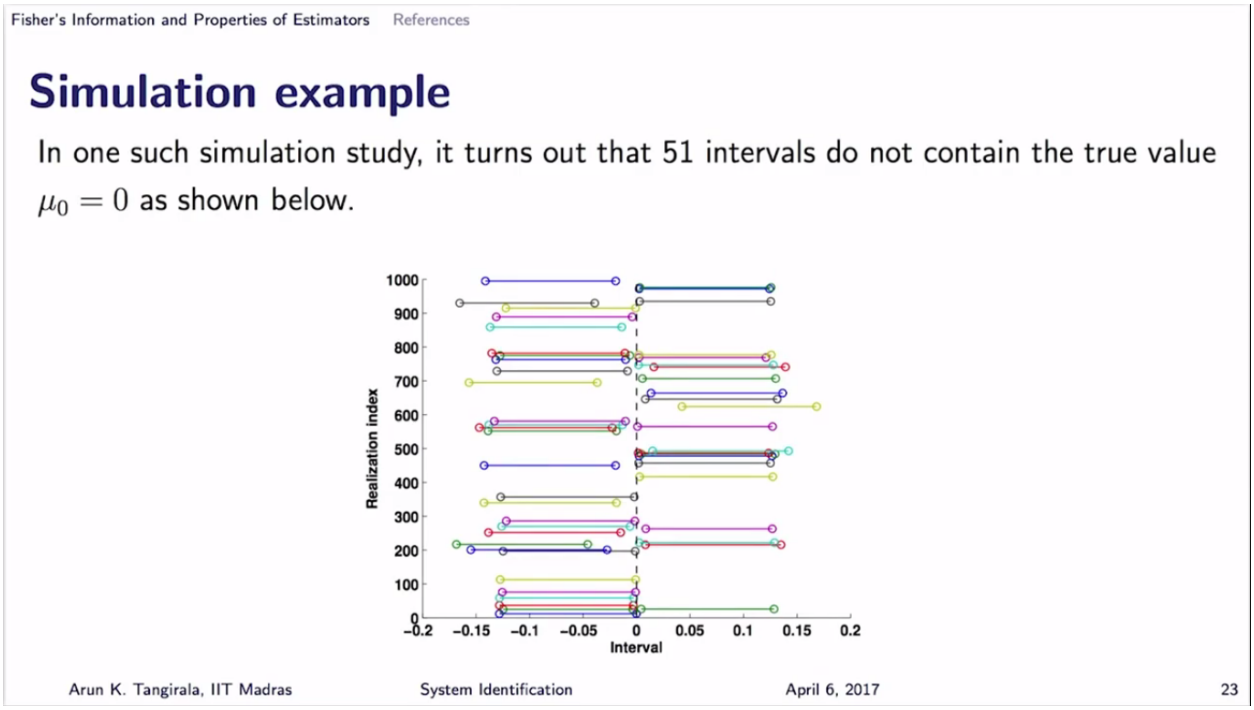
So let's move on and quickly conclude this discussion on properties. The other thing that we are interested in estimation as you know, is making confident statements about the truth which then eventually becomes important in hypothesis testing. Right.Remember if you recall. Remember that when we talked about estimation we said the purpose of estimation is not just to obtain a single value, that's called a point estimate. But the grander purpose is to say something about the truth. That is, give an interval in which you believe the truth is. And that is what we mean by confidence intervals. And the only thing you have to remember about confidence intervals, is that it is not possible to have 100% confidence interval that is finite length, finite width interval because of randomness. I cannot say with 100% confidence that the truth is in some finite region. If I ask you what is a true temperature outside or you know something else you can say what is a true measurement of something else. You can say with 100% confidence that it is somewhere between zero and infinity or minus-- Even for that outside temperature, the 10% confidence interval would be the lowest achievable thermodynamically which is 0 Kelvin. And maybe temperature of the sun or temperature of the earth or whatever 100 degree Celsius. But that for all practical purposes is like your minus infinity to infinity. And that's of very little use. Rather than that you would want a finite interval, finite width confidence interval with a high degree of confidence. And usually these are conflicting.As I try to narrow down the truth my degree of confidence will go down. But what we should strive at is to, for the same, for a given interval, if I say that I want to confine the truth to I want to be able to say something about the truth with as narrow interval as possible. But I want highest degree of confidence, then what kind of an estimator should I work with. What kind of estimators will give me for a given. Let's say, a given the degree of confidence, minimum in confidence region interval.

Efficiency.

Efficient, that's where efficiency comes in the picture , right.Because this confidence region has got to do with errors in your estimate. Larger the standard error in the estimate, larger the-- wider the interval is going to be, a confidence interval is going to be for a given degree of confidence. So if I fixed a degree of confidence like typically to 95% or 99% then the one that gives me the narrowest interval for the truth.

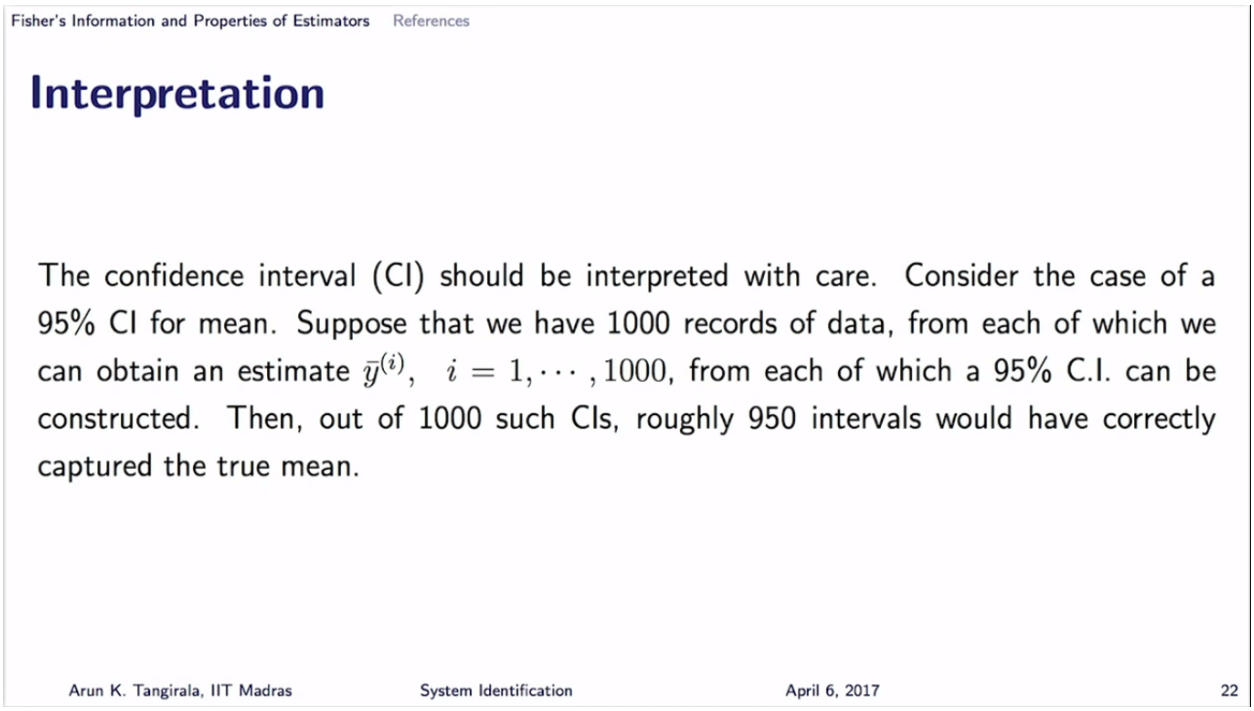
See truth is fixed. When we say-- When we talk about interval it is our estimate. And we may be long and that is what we mean by confidence. Right. So we when we say 95% confidence, what we mean is that there is a, in some sense although we shouldn't say a chance but that truth is not contained in this interval there is a possibility and that possibility you can say is 5%.

(Refer Slide Time: 18:48)



So that interpretation is extremely important.

(Refer Slide Time: 18:47)



Let me actually do this for you. So what I explain, this is also there in the lectures but I just want to reiterate what confidence interval means because many people have wrong notions of a confidence interval. So what I did here was, I generated a thousand observations of Gaussian white noise process computed at the sample mean. And we know the confidence interval for mean, when you compute, when you estimated the sample mean.So your confidence interval depends on it estimator that you're using. If you sample mean and if I assume that the underlying process is Gaussian White. So there are so many assumptions. One I should work with the sample mean and two that the data is coming from Gaussian White Noise process or at least white noise process. Then we know that the confidence region is given by at least 95% confidence the region for the large sample case is given by this.Right.This is what is it, 95%confidence region for what?For?

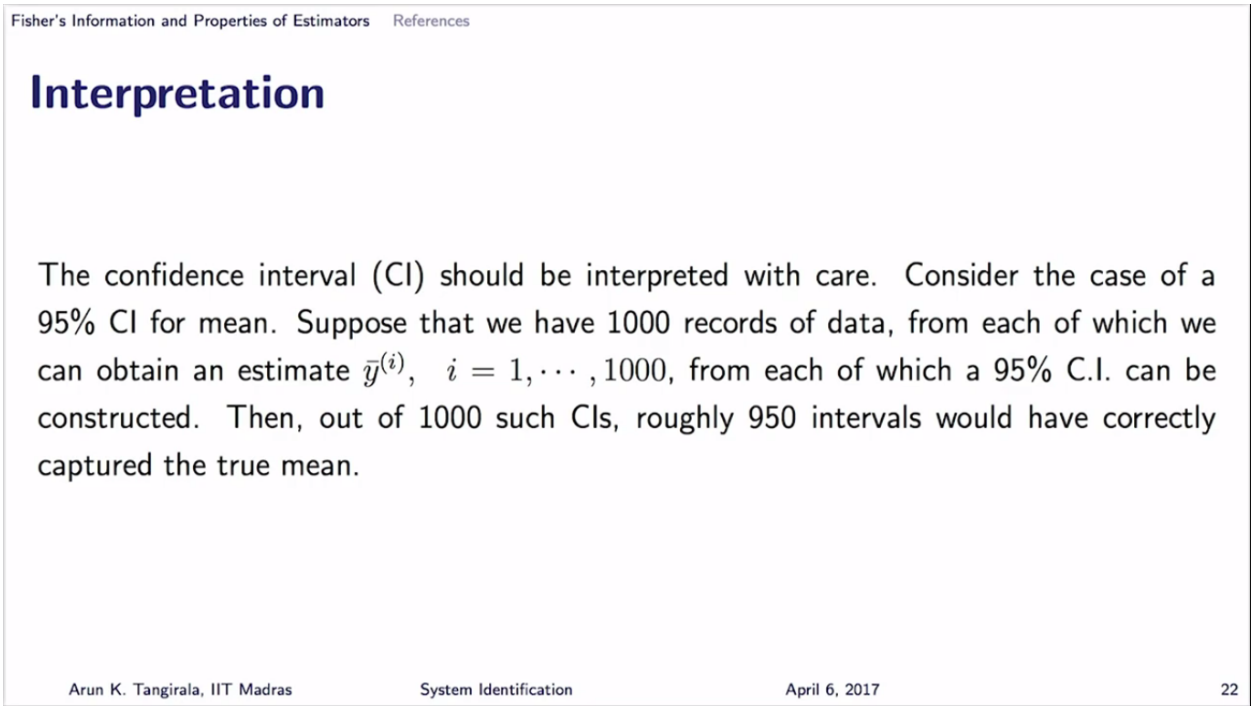
Mean.

Mean. You have to--Many people say it's a confidence interval for the sample mean. There's nothing like confidence interval estimate. Where estimate is a single value. So this is the 95% confidence interval for mean. Meaning, I'm 95%confident that the true value is contained in this interval. Large sample case,White Noise process and if you sample mean. If I change any of these assumptions this result doesn't hold good.

For example if I'm looking at a correlated process this expression is not correct. If I'm looking at the small sample case, then this is not correct. If I'm looking at a different estimator,like a sample median, then this

expression is not correct. So there are this three assumptions that are involved. So what I did was,in this each realization, I generated a thousand observations and I computed sample mean and that constructed confidence intervals. Right.

(Refer Slide Time: 21:13)



For each realization I have one confidence region. Will the confidence intervals with the same for every realisation?Why?

[21:33 inaudible]

Y bar is going to change. Right. It's very simple. There's no complication here. You look up on this has a formula. Y bar is going to change from realisation to realisation.

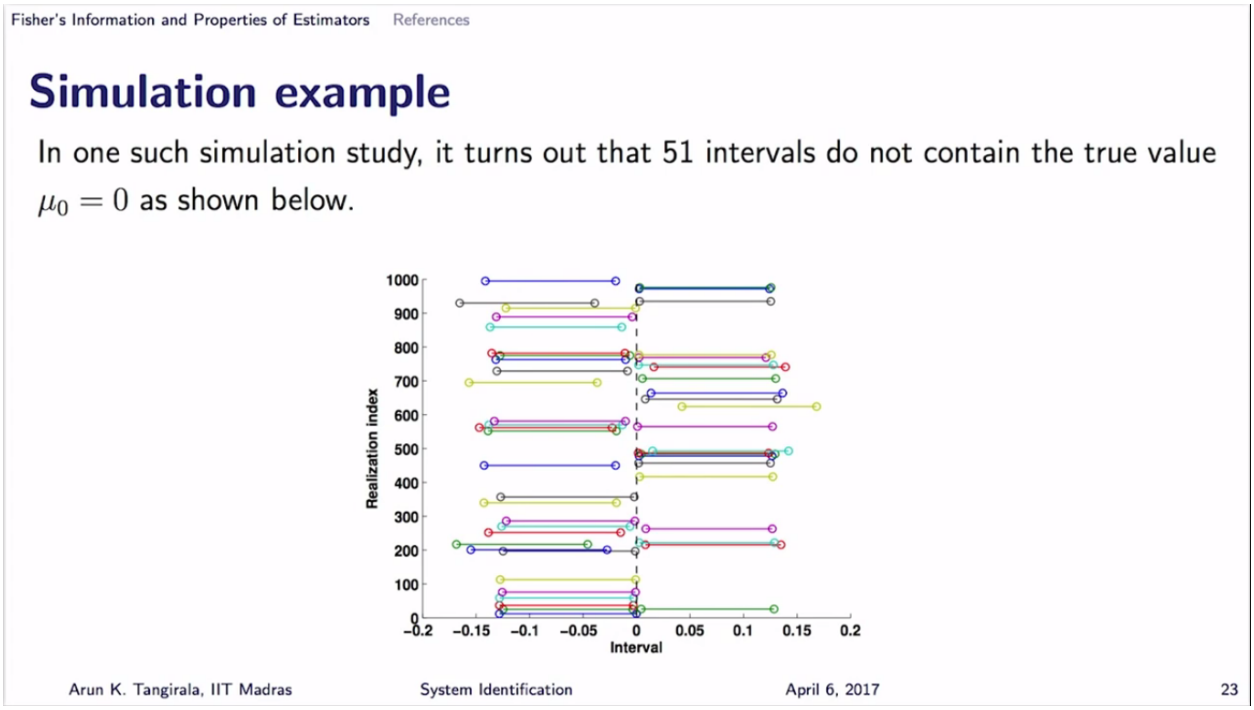
It will be the dame but the position --

No. But the width also can change. Because of course, I assumed here that sigma is known, when sigma is not known then you have to estimate. And for large sample case you can still use this expression. Strictly speaking,if you work with an estimator sigma, then you have to go back to T distribution and change that 1.96 but will as I said assume large sample case.

And by large we mean anywhere after 40, 50 and so on. Here in thisestimation process. So sigma hat is also likely to change from realization to realization. Y bar is going to change as a consequence the

interval is going to change. For every realisation I have an interval. What 95% confidence interval means is, if I were to work with thousand realizations. How many intervals will I have? Thousand intervals. Correct. Out of this thousand confidence intervals roughly about 50 will miss the truth. Right.And I show this. That's what I show in this graph. This 50, in fact in this case it turns out to be 51.

(Refer Slide Time: 23:02)



Ideally if you have maybe 10,000 or maybe more then you will have better and better, you know, correct number of confidence intervals that do not capture the truth. So the 51 confidence intervals that miss the truth are what I'm showing you. The centre line is a true 0. I have generated data from a 0 mean white noise process. And what do you notice here. I have 51 confidence intervals that have missed out the truth. So there is a chance that you would run into one of these realizations. Suppose, I had told thousand students in my class, I generated thousand realizations, I gave one realisation to each student, fifty one students will come back and tell me that this is a confidence interval. That means, if I were to conduct the hypothesis test that the mean is 0. So if I want, this is a null hypothesis that I am testing which is called a significant test. In a class of 1000,949 students will they reject or not reject the null hypothesis.

Will they fail to reject the null hypothesis or they'll reject the null hypothesis. 949 students.

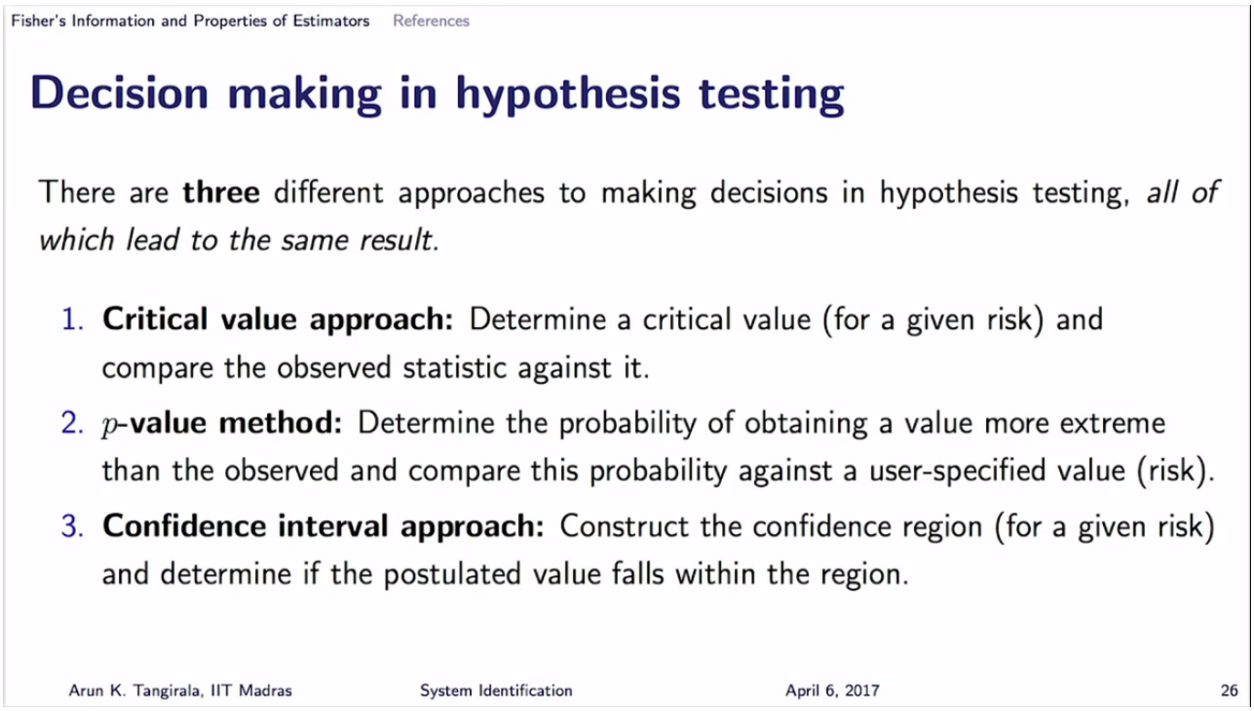
Failed to reject.

Failed to reject.Right.Becausethe way you use your confidence intervals for hypothesis testing is your construct the confidence interval and then you ask if the postulated truth is in the region. The postulated truth is 0. In a class of 1000, 949 students will come back and tell me that there is a confidence interval that contains 0. Therefore they will fail to reject the null hypothesis. That means 0 is also possibility. The remaining 51 will reject the null hypothesis. And you may be one of those lucky ones. If you end up with such a realisation for which the confidence interval does not capture the truth then you will end up rejecting the hypothesis, null hypothesis. That is what we mean by 5% significant level in hypothesis testing, right. That 5% significance level has a very close connection with your degree of confidence.

You can look at look up on that significance level orconfidence, doesn't matter. That means, what is the truth. Truth is that null hypothesis holds. But there is an error. There is a chance of rejecting the null hypothesis and that chance is 5%. If you construct a 95% confidence interval. If you work with a 99% confidence interval 10 students or 11 students will come back and tell me that the null hypothesis has to be rejected. Right. That is what we mean by Type 1 error in hypothesis testing. The null hypothesis is true and there is a probability that I rejected. That probability is closely related to your degree of confidence.They are all one and the same.

So in general hypothesis testing can be conducted in three different ways. And you should, you know, at least if you are not familiar, I've already told you,you should sit through the lectures on statistical hypothesis testing or even the lectures here. I will tell you that basically there are three different ways of testing.

(Refer Slide Time: 26:57)



One is a critical value approach, where you compare your observed statistic in this case sample mean with some critical value that you determine based on the significance level or you use a P value approach or do you use a confidence region approach. As far as this course is concerned, in general also you can use a confidence interval approach. Each has its own charm but all the three will give the same result.

You cannot come back and say my critical value approach failed to reject the null hypothesis. Whereas confidence interval approach rejected it. No. If you maintain the same alpha for all the three, all these three approaches to hypothesis testing will give you the same result. The question is which is the most convenient. And I find the confidence reason approach a very convenient and elegant. But there are statisticians who would argue in favour of the P value approach or the critical value approach. Yes, I agree. I mean in some, in the design of experiment, from a design of experiments viewpoint and so on, the p value approach can be useful. But as far as testing is concerned confidence interval is very good.

So for all practical purposes in this course we will adopt the confidence interval approach for hypothesis testing. But the question is, in what kind of situations do we run into this hypothesis test? Well when you build models what are you going to do, you're going to estimate parameters at least parametric models, you're going to estimate parameters. And we want to know whether that's true parameter is 0 that means have a inadvertently included a parameter that was not supposed to be there in the model. Then you have to conduct these kinds of tests which are called significant steps.

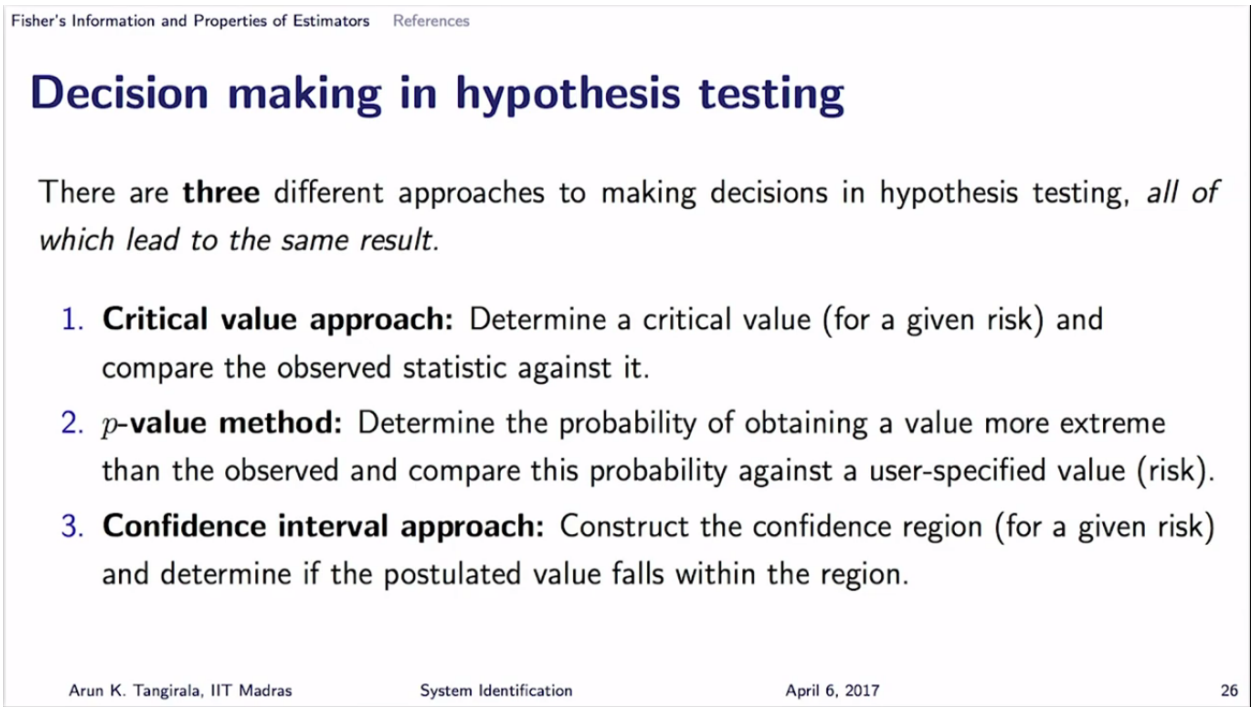
In general, when you replace this Mu with theta, right.Then any test of this form is called significance test. That means you want to look at theta hat and -- theta hat looks very small. Should I ignore it or not? That is what we mean by.That means is theta hat statistically significant or not. Associated to this question is this hypothesis test. Theta is equal to 0 and theta not equals 0.

So what is a procedure there, you construct a confidence region for theta and search for 0, if 0 is contained in the interval then you fail to reject the null hypothesis. So what do we do? Suppose I discovered that a parameter, for a parameter I have failed to reject this null hypothesis that means, it doesn't pass a significant test. So what is the conclusion?What is the consequence of that? What do I do? Do I go and post this live on Facebook. What is it.? Well, many things are posted on Facebook these days including the saddest parts also. So what do we do with this.?

Remove that parameter from the model. And re-estimate the model. Please, don't include some1000 parameters in the model and say , sir, 995 of them are actually turned out to be insignificant. Please take the remaining 5 and go home and be happy. You can say that. It's your duty to throw away those 995 parameters re-estimate those 5parameters. Whyshould I re-estimate those 5 parameters?

I've spent some food in feeding those 995 guests.995 people who did not do anything for me.Right. They didn't do any work for me. So why should they actually do that. So I rather re-invite those 5 people, feed them well and make sure that I get the best work out of them. Right. So I should re-estimate the model with the significant parameters and then report that model. You can't leave things halfway through. This is something that I expect even in the final exam.When I give you data and you have done an over fitting and it's possible. All of us do [31:21 inaudible] with the over- fitting. You recognize that you're working with an over-fitted model then you should remove those redundant parameters and re-estimate and report the most compact from. Okay. So this concludes the discussion on the properties of estimators. Of course, one thing that I have not spoken about but I have elaborated in the lectures is that.

(Refer Slide Time: 31:48)



How to construct a confidence region. There is a procedure and that procedure involves determining the distribution of theta hat. That is why we need distribution of theta hat. And the only thing that I can tell you is, for linear estimators it's easy to come up with an analytical expression for the distribution of theta hat. Why because I have a central limit theorem. For non-linear estimators, it's quite difficult. And generally in in the modern era, one uses this Monte Carlo method or Bootstrapping methods and so on to figure out F of theta hat.Once you have F of theta hat from there you can conduct your hypothesis test or you can construct confidence regions. In some cases, you have asymptotic results. In fact a lot of times you only have asymptotic results, that means distributional properties and under large sample cases. And that is what we use by and large in in this course. We use a large sample distribution properties or asymptotic distribution properties. Generally it turns out to be that theta hat is Gaussian. In all our analysis, except depending on whether you're dealing with power spectrum or power spectral density, variance and so on. Even in such cases the large sample case, we know, I square tends to Gaussian. So you should expect in some sense asymptotic Gaussian distribution.