Applied Time-Series Analysis
Prof. Arun K. Tangirala
Department of Chemical Engineering
Indian Institute of Technology, Madras

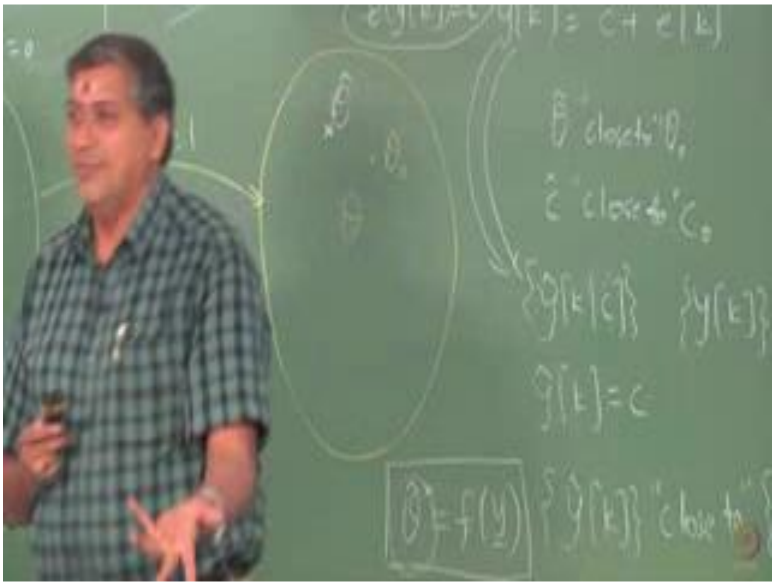
Lecture - 80
Lecture 35C - Introduction to Estimation Theory 3

Now we move to post estimation analysis. I have not talked about sigma square e estimate, we will come to that later on it is not necessary at this moment. So now that I have estimated I would like to ask, even let I say I chose sample mean as the winner I say I know that there are no out layers I would like to work with the simple one I would like to now ask how good this estimate is. So, you may wonder why am I revisiting that question. Already I have made sure that the y hat of k is very close to y and I have got the optimal solution what do you mean by good I have only made sure y hat is driven very close to y. I have to now guarantee that by doing so I have actually managed to drive c hat close to c naught, I have to prove that.

This is the point that is missed out, I mean at least in terms of explaining in many many text books. It is understood I mean kind of understood that you understand all of that, but it is not so. That is this is the key point you have to remember. We have only derived an optimal estimate and this optimality is in the sense of y hat of k being driven very close to y k. I have to be now, guarantee that c hat has been driven very close to c naught, and how do I come up with that kind of an analysis, how do I quantify the closeness and so on.

Anyway, before we plunge into this different matrix the first point that you should keep in mind and forever is that.

(Refer Slide Time: 02:02)



The estimate that you are constructing whether it is star or naught is a function of your observations which have randomness in that. So, the DNA the randomness DNA in y will propagate to theta hat. In other words your theta hat is also a random variable. What does it mean if I change the data record? I will get another value of theta hat. That is what we mean by randomness. Therefore, theta hat is random variable in its own right; it has its own mean, it has its own variance, it has its own pdf, everything. For all practical purposes now theta hat is a random variable, now you can recall our discussion that we had when we talked about covariance matrix; variance-covariance matrix at that time we had said that we will end up looking at sigma theta hat. Now is the time to talk about that.

So, now with this observation, with this fact that any estimate is random variable we will come up with different matrix of closeness of c hat to c naught. Why is it important? Because when I started off I did not realize c hat would be a random variable, but now the problem is more complicated than what I thought c naught is it random or fixed.

Student: Fixed.

C hat is random; now I have a problem, how can I talk of closeness of a random variable to a fixed one? It keeps changing c hat keeps changing with the data set. So now I have to come up with ways of determining the proximity of c hat to c naught taking into account the fact that c hat is random variable see there is. So, much to estimation in high school when we were thought how to fit a straight line none of this would have been even touched the periphery of our imagination, but you see how involved and how much people have actually thought through that is you can also say in those days they did not have any internet, Whatsapp and so on even if you take that away from me I will also think of all of these.

So, let us now get on to the matrix of goodness of estimator. What we mean by here is we want to quantify now come up with matrix that tell us how close c hat is to c naught. One of the matrix is to now demand that the expected value of c hat, we know c hat keeps changing with the data record. Now we can demand different ways of or we can imagine different ways of proximity of c hat to c naught.

(Refer Slide Time: 05:00)

i. $E(\hat{\theta}) = \theta_0$
ii. $E((\hat{\theta} - E(\hat{\theta}))^2)$
$\sigma_{\hat{\theta}}^2$
Σ_z
Z
σ_y^2

The first thing is we can ask if expectation of c hat is equal to c naught. So, what we mean by this is I repeat the experiment many many many times I generate all possible data records and for each such data record I have a c hat. And I take the average of all of that, I do not change the sample size; the sample size remains the same and I take the average of all of that; will that be equal to the truth? See expectation of c hat is a deterministic quantity. So, it is fair enough to compare a expectation of c hat with c naught. Directly I cannot compare c hat with c naught, because c hat is a random variable. If this is the case then we say the estimator is accurate and we say that estimator is unbiased; this is what has got to do with accuracy.

So, unbiased and accuracy are synonymous terms. What else can we think of? The other thing that we can think of is variability how does the variance of c hat, what about the variability? This is a different measure it has got not necessarily with respect to the proximity of the c naught, the first one talks about proximity of c naught but apart from that just because this is satisfied it does not make an estimator very good; just because the average of c naught gets the truth you will never be able to do this in practice but let us say theoretically this is guaranteed. It does not make the estimator necessarily very good. One of the expectations of a requirement of a good estimator is if I change the data record I know c hat will change, but it should not change widely.

Remember the uncertainty in the data here is your z propagates to theta hat as well. So, here is your sigma square, let us say sigma z or if you are looking at single signal or variable sigma square y here you have sigma square theta hat for the single parameter; between the data and theta hat is the estimator. The estimator while giving you the estimate of the parameter is also doing something else behind the scenes, what is it doing? It is letting the uncertainty propagate; uncertainty in data is propagating to theta hat.

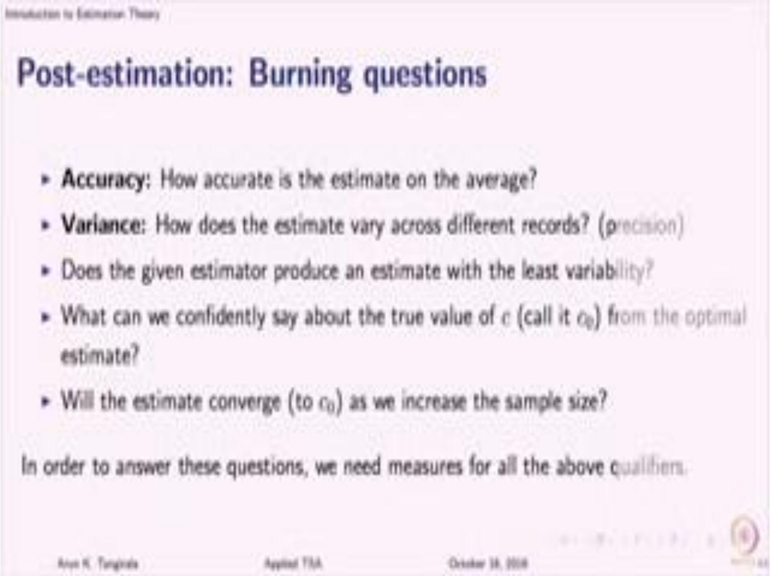
Remember you can recall this in the context of Fourier analysis, we said there is a signal decomposition there is an energy decomposition or power decomposition likewise here the role of the estimator yes is to get you given y it will give you theta hat, good; but behind the scenes what is it doing? It is also propagating the uncertainty. And we want the estimator to shrink that uncertainty; we do not want that uncertainty to propagate just like that. We want it to reduce, yes the data changes from experiment to experiment, theta hat will also change but I do not want it to change the same way as the data changes. I want that variability to be as small as possible; I know I cannot drive it to 0. If I drive it to 0 what happens?

Student: (Refer Time: 08:44).

Theta hat will remain fixed, match fixing any data record you give me theta hat is the same; that cannot happen. So, estimator should change the value of theta hat, but also should work on the variance of theta hat. So an ideal estimator, how is this variance defined? The same way as the variance is defined for a random variable; expectation of c hat minus expectation of c hat square this is sigma square so let us me replace with theta here, so that we keep things as generic as possible. So, this is your entire thing is your sigma square your theta hat. This is sigma square theta hat not of the data or of the parameter the true parameter is fixed.

So, one metric of goodness of c hat is unbiasness, other metric of the goodness is with respect to its variability which is sigma square c hat very often the term used is precision. You must have heard of high precision instruments, wherever you see you know people advertising manufacturers advertising their instruments on pamphlet us they talk of high precision instruments. What they mean by that is if you use that instrument again and again you will see a slightly different reading but not too much different, it should not be that sometimes your thermometer that you have at home sometime shows 37, next time it shows 40. It is too much right it may show some variability in maybe fifth decimal and so on

(Refer Slide Time: 10:31)



Introduction to Estimation Theory

Post-estimation: Burning questions

- **Accuracy:** How accurate is the estimate on the average?
- **Variance:** How does the estimate vary across different records? (precision)
- Does the given estimator produce an estimate with the least variability?
- What can we confidently say about the true value of c (call it c_0) from the optimal estimate?
- Will the estimate converge (to c_0) as we increase the sample size?

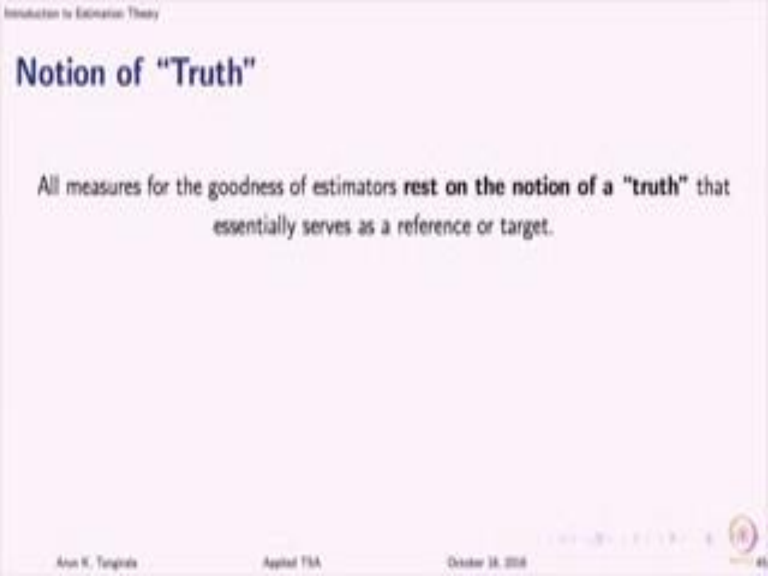
In order to answer these questions, we need measures for all the above qualifiers.

Arun K. Tangirala Applied TSA October 18, 2016

Fine, the general question that we are always interested in is; does the given estimator produce estimates with the least variability? So, among all estimators that are possible the hunt is always for the estimator which has the highest precision that is the lowest variability. That is why the manufacturers are advertising their instruments are high as high precision instruments. There are many instruments to measure the same variable, but this guy claims that my instrument has high precision; another guy is also making a same claim, but then it is your duty to compare the precisions.

So, accuracy has got to do with bias or unbiased or lack of it and variance has got to do with precision. An estimator can be biased, but highly précised. What we mean by that is this may not be satisfied that is, but it may have very low variability. We will talk about that more in detail when we formally define what is bias, what is variance and so on, but it is possible for an estimator to be biased; that means, there can be a systematic error in your estimate, but the variability from experiment experiment is very low.

(Refer Slide Time: 12:07)



Introduction to Estimation Theory

Notion of "Truth"

All measures for the goodness of estimators **rest on the notion of a "truth"** that essentially serves as a reference or target.

Arun K. Tangirala   Applied TSA   October 18, 2016

Then the other thing that we want to ask is the next question that we want to ask is what can we confidently say about the true value. I will come to that let me actually go to the notion of truth before we answer what can we confidently say about the true value requires this brief discussion maybe 2- 3 minute discussion on what we mean by notion of truth in estimation. So, we have looked at two matrix accuracy and variability, but then we said early on in this lecture and previously also ultimately I should end the estimation exercise with some confident statement about the truth. But to do that first I have to define what is meant by truth.

Now in any estimation analysis we have this notion of truth; that means, we say that if this is the truth how will your estimator perform. Then you may ask no the reality is lot more complicated, for example here in this example suppose we say this is how the true process is generating the measurements suppose that is the case then we want to ask how good the estimate is, whether it is accurate, what is the variability and so on. But we know very well that the true process may not be the simple maybe lot more complicated, but we do not worry about it if we want to qualify; that means, if we want to say an estimator is good we will first fix the truth and ask if this is the truth at least you tell me for estimator is good.

So, there is always this notion of truth that is introduced in any estimation exercise you define what is truth and then you qualify the estimator. You say well, suppose you come up with some other estimation method tomorrow. And you want to present to the world you want to convince that it is a very good way of estimating parameters then your estimator will be put to test. Some known problems will be taken and then your estimator will be subject to test, in the sense for this known problem does it recover the truth. In reality whatever estimate you get we do not know what the truth is.

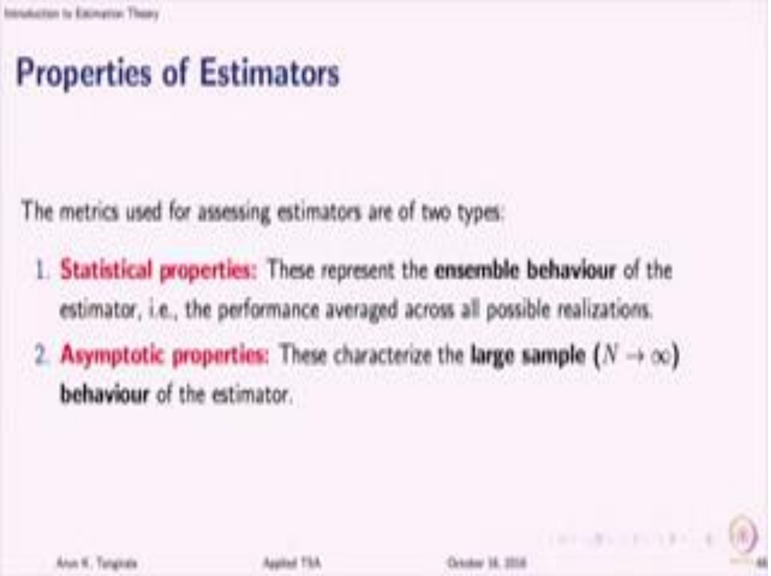
Please remember that this notion of truth is very essential for the analysis the theoretical analysis, but when it comes to practice we do not know what the truth is. Then what do you do? Suppose you have estimated parameters of a model. How do you know they are good, because you are you do not know the truth. What do you do there? Suppose you have estimated some model; that means, you have estimated some parameters. How do you know they are good? You do not know the truth, I also do not know the truth, nobody knows the truth; how do you convince that your estimates are good?

Student: (Refer Time: 14:58).

Now again map it back to the knowns, given the estimates construct y hat and show that there are very close to y you will come back to this. But at the moment what we are saying is suppose I know the true values will your estimator recover that or not. If it cannot then there is no hope even for y hat that is the point here. So, in practice your test for goodness of estimates will be based on y hat, and of course variability also will know how to calculate variability. But the rigorous part of the estimation analysis consists of specifying the truth and requiring that your estimator is able to recover that; that is what is important to keep in mind.

So, when it comes to properties of estimators there are basically two types of properties.

(Refer Slide Time: 16:02)



Introduction to Estimation Theory

Properties of Estimators

The metrics used for assessing estimators are of two types:

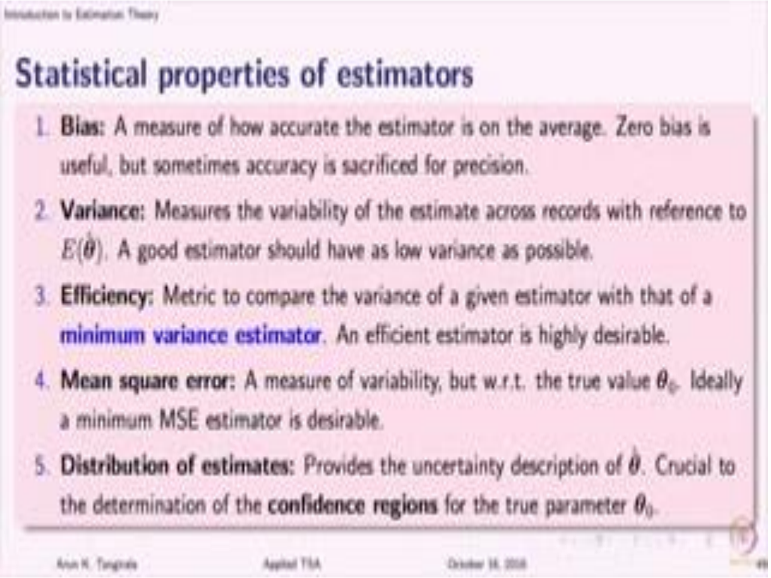
1. Statistical properties: These represent the **ensemble behaviour** of the estimator, i.e., the performance averaged across all possible realizations.
2. Asymptotic properties: These characterize the **large sample** ($N \to \infty$) **behaviour** of the estimator.

Arun K. Tangirala Applied TSA October 16, 2016

One is the statistical properties, other is the asymptotic properties. These two that we have discussed are statistical properties of the estimator, and then there are a few more. Then there is something called asymptotic properties, what is the difference between these two?

Statistical properties of the estimators examine the quality, the response of the estimator to change in data records, when I change the data record how does the estimator behave? Is it going to sure or is it going to do very well. So, it looks at the ensemble direction. Asymptotic properties are examined or quantify the property of the estimator with respect to sample size. How do these statistical properties or the property of the estimate itself how do they respond as I keep increasing the sample size; does it get better and better. I mean the intuitive expectation, the natural expectation is that the estimate improves as I increase the sample size unfortunately there are a few estimators that do not improve at all even if you supply million billion trillion data points still the estimate quality is going to be very bad. So, we want to avoid such estimators. I will just list these properties.

(Refer Slide Time: 17:28)



Introduction to Estimation Theory

Statistical properties of estimators

1. **Bias:** A measure of how accurate the estimator is on the average. Zero bias is useful, but sometimes accuracy is sacrificed for precision.
2. **Variance:** Measures the variability of the estimate across records with reference to $E(\hat{\theta})$. A good estimator should have as low variance as possible.
3. **Efficiency:** Metric to compare the variance of a given estimator with that of a **minimum variance estimator**. An efficient estimator is highly desirable.
4. **Mean square error:** A measure of variability, but w.r.t. the true value θ_0. Ideally a minimum MSE estimator is desirable.
5. **Distribution of estimates:** Provides the uncertainty description of $\hat{\theta}$. Crucial to the determination of the **confidence regions** for the true parameter θ_0.

Arun K. Tangirala Applied TSA October 16, 2016

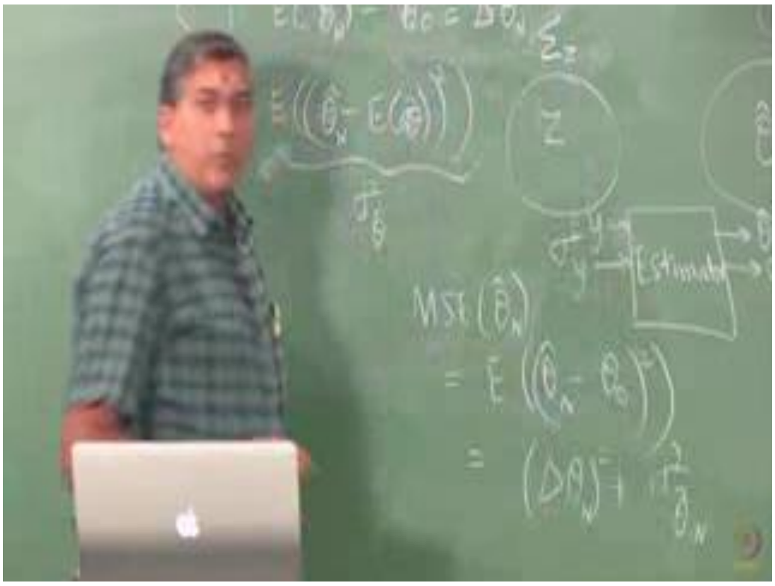
And then we will spend a couple of minutes to close the section. So, I have already talked about bias and variance there are three other statistical properties, there are many more but I have just listed the most important once. There are three other statistical properties that are of interest: one is the efficiency in fact efficiency is a property based on variance. We said we want an estimator that has lowest variability. When we have struck gold; that means, we have found an estimator that has the lowest variability among all estimators we say that is the most efficient estimator.

Initially I was puzzled to see this term efficiency why would someone use the term efficiency for defining an estimator that achieves the lowest variability, but you give it a thermo dynamic prospective or you give it a general process prospective what you are putting in is sigma square y. What is coming out, of course what you are putting in is y to get theta hat that is what you see, but behind the scenes what you are pumping in is sigma square y. And what estimate is pumping out is sigma square theta hat. Obviously, efficiency is its ability to convert that uncertainty to or to shrink that uncertainty to as low value as possible; and that is why the name efficiency seems to be justified.

So, it is very efficient it is doing a lot of hard work. And the most efficient estimator is the one that achieves the lowest variability. And the hunt is always for the so called fully efficient or minimum variance estimate. In fact, the ideal estimators that we want is minimum variance unbiased estimator, you will see this everywhere in estimation literature MVUE; minimum variance unbiased estimator there is always hunt for it.

Then the forth property that is of interest is mean square error. This mean square error is also a measure of variability, but the only difference with the mean square error is.

(Refer Slide Time: 19:49)



It is called the MSE of theta hat is that the variance is calculated with respect to the truth. What is the difference between this and the variance? The reference point here is its own average which may be truth or may not be the truth depending on whether it is biased or unbiased, whereas MSE specifically looks at the statistical distance of theta hat from the truth.

Now again you have to wonder if I do not know the truth how am I go to calculate MSE? Which is true, I do not know the truth. But there are ways to calculate this, and even though I cannot calculate it I can guarantee certain estimates achieve minimum mean square error. In fact, earlier I said the hunt is for minimum variance unbiased estimator, but the hunt is even more for an estimator that achieves minimum mean square error.

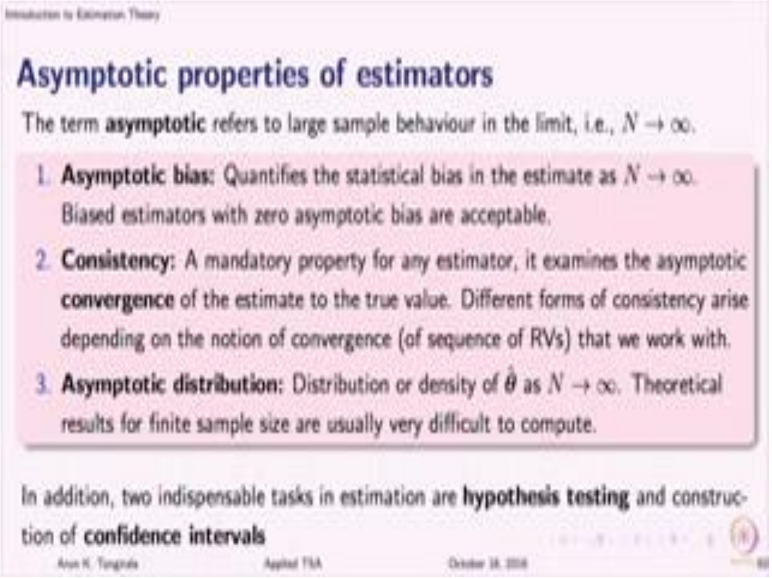
Obviously, I want an estimator that is as close as possible to the truth; that is a period I mean there is no other statement that I need to make. I do not have to say minimum variance unbiased nothing, I will combine that into a single statement; I want an estimator that achieves minimum mean square error and it turns out Bayesian estimators can do that. In fact you can show just with the couple of steps that this has got to do with your delta theta square plus sigma square theta hat; delta theta is this difference if you were to define this has delta theta or even other ways theta naught minus expectation of theta does not matter essentially the bias.

So, you can express the mini mean square error like the Pythagoras theorem as sum square of bias plus the standard deviation. So, if you are going to achieve minimum mean square error you do not care whether it is bias unbiased and so on; it is staying very close to the truth. But, it is a difficult task and Bayesian estimators do that.

Finally, after having said all of these remember, theta hat is a random variable I need to know the pdf. Why do I need to know the pdf of theta hat? Again on the face of it, it is not so clear why I am interested is it only some academic joy that I get by constructing f of theta hat or is there a practical purpose to it. And the practical purpose is that we will be able to construct confidence intervals for theta naught, it may not be clear at this moment when we actually talk about it will become very clear later on.

So, the purpose of constructing f of theta hat is to make some confidence statements about theta naught. Remember that is the final step in the estimation. We construct theta hat, we ask how good it is and then if it is good then we proceed further and say that this is the interval in which theta naught has. And I just want to conclude the lecture with these asymptotic properties.

(Refer Slide Time: 23:10)



Introduction to Estimation Theory

Asymptotic properties of estimators

The term **asymptotic** refers to large sample behaviour in the limit, i.e., $N \to \infty$.

1. **Asymptotic bias:** Quantifies the statistical bias in the estimate as $N \to \infty$. Biased estimators with zero asymptotic bias are acceptable.
2. **Consistency:** A mandatory property for any estimator, it examines the asymptotic **convergence** of the estimate to the true value. Different forms of consistency arise depending on the notion of convergence (of sequence of RVs) that we work with.
3. **Asymptotic distribution:** Distribution or density of $\hat{\theta}$ as $N \to \infty$. Theoretical results for finite sample size are usually very difficult to compute.

In addition, two indispensable tasks in estimation are **hypothesis testing** and construction of **confidence intervals**

Arun K. Tangirala Applied TSA October 18, 2016

These are the three different asymptotic properties that one would be interested in it. One is asymptotic bias; what is the difference between statistical and asymptotic properties? Asymptotic properties look at how the estimates change as I change the sample size or how the statistical properties themselves change as I change the sample size. Remember when we I am calculating this all of this is based on fixed sample size, finite sample size. These are all finite sample size properties, I am only varying the data records; but now I am saying I would like to also know how this estimate changes as I change the sample size. And earlier I said not all estimators are going to be unbiased, there can be a biased.

(Refer Slide Time: 24:07)

For example, sample mean; what is the sample mean? 1 over N sigma y k. Is this an unbiased estimator or biased estimator? How do you find out? This is mu hat right, mu y hat I want to find out if expectation of mu y hat is the same as mu y. So, simply apply the expectation operator. What do you get? It is unbiased. But suppose for some crazy reason I am going to use N minus 5, it is biased; but what can you say as about the bias as n goes to infinity? Will it become unbiased or not? Bias is equal to 0 or not?

Student: (Refer Time: 25:02).

So, we say this estimator is biased statistically, but asymptotically unbiased. There are many such estimators; on you are calculators you will have s n s subscript n minus 1, what are they doing? They are estimating what? Have you looked at the calculator (Refer Time: 25:22)? S n is s n and s n minus 1 or sigma n sigma n minus 1, what are they estimating? Standard deviations, right; whether it uses a 1 over N minus 1 or 1 over N. It turns out that if you use a 1 over N minus on in the calculation of standard deviation estimating standard deviation it gives you unbiased estimates, whereas you use 1 over N it gives a biased estimate. Obviously, one is unbiased the other has to be biased. But this 1 over N is asymptotically going to give you unbiased estimates. As n goes becomes very large it does not matter whether you have 1 over N minus 1 or 1 over N. So, such is the case with many estimators.

What we are saying, what we require is it is to have biased estimators; that means, for finite sample size, but when n goes to infinity the bias should vanish; that is one requirement. The other requirement is consistency; this is like the golden property that you are seeking in an estimator. And this golden property is that as n goes to infinity; that means, as you have large number of samples theta hat should converge to theta naught. You recall we talked of convergence of random variables in some context, do you recall the context? When? At some point we brought up the notions of convergence of random variables.

Student: (Refer Time: 26:54).

No, one is ergodicity; the other context?

Student: (Refer Time: 27:00).

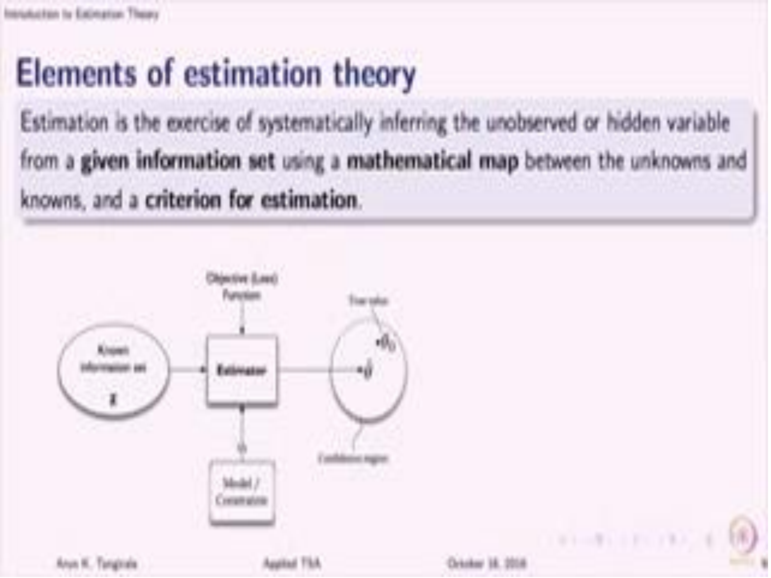
Linear time series models we said that the linear representations sigma h of n e k minus n should converge should produce a random variable, it should converge. And we talked off different forms of convergence: almost sure convergence, weak probabilistic convergence, mean square convergence, and so on. All of that we will talk in detail when we talk of consistency. Consistency is about the convergence of theta hat n to theta naught.

And finally, asymptotic distributions; just now I said we need to construct f of theta at m that is the distribution of this estimator. What happens is unfortunately in many situations it is very hard to theoretically figure out what the distribution is analytically for finite sample size; it becomes easier if you evaluate let n goes to infinity. Again here we are talking of convergence we are asking now of convergence in distribution. I talked about that as well. So, there are four different forms of convergence and all of that are relevant here now. So, today you have methods like goods lapping, Monte Carlo simulations and (Refer Time: 28:19) data methods and so on to figure out the distributions. But when it comes to theory only asymptotic distributions are easy to compute.

So, these are the matrix that we will come across. And these are the matrix that qualifies the goodness of an estimator. You see we have to worry about all of this before we choose an estimator. And we will ask when we study least square we will ask least squares efficient, consistent, unbiased, what is its variability under what conditions and so on; likewise for MLE and so on.

But I want to conclude today's lecture with one very important point. We have been talking of estimator all the time, but let us now I just as a foot for thought when you leave and this is going to be the main topic of tomorrows lecture. Let me take you back to the schematic that I had upfront in today's lecture.

(Refer Slide Time: 29:17)



Introduction to Estimation Theory

Elements of estimation theory

Estimation is the exercise of systematically inferring the unobserved or hidden variable from a **given information set** using a **mathematical map** between the unknowns and knowns, and a **criterion for estimation**.

Arun K. Tangirala Applied TSA October 18, 2016

So, we have been talking of this estimator all the time; how good estimator is and so on, it is like asking how good my digestive system is? What if the food that I eat is very bad? We should also ask that right, if somebody falls ill due to food poisoning or something or ill and let us say its related to food thing there are two possibilities: either that the food that the person ate was poisoned or you know had some problems or the food was but digestive system has gone for a toss.

Until now we have focused only on the digester only on the estimator, but what if the data that was presented itself was of poor quality. Suppose, it did not have any information, let us say I give you some cricketers scores over years and I ask you to credit the grade of some random student; does that data have any information? May be surprising we do not know. But depends on how close the person did not following the cricketer.

But to think of it, do you think that data will have any information? That is what we mean by poor data, that data does not have any information. And Fisher among many was pioneer in coming up with this concept. And today Fishers information is one of the central concepts in information theory, at least in statistical inferencing which quantifies how much information data has with respect to a parameter. And we will learn that and we will then learn another milestone result in estimation theory known as (Refer Time: 31:00) which is based on Fishers information.