Dynamic Data Assimilation

Prof. S Lakshmivarahan
School of Computer Science
Indian Institute of Technology, Madras

Lecture – 24
Statistical Estimation

So, far we have been talking about Data Assimilation in Deterministic model; deterministic static; deterministic dynamics with respect to deterministic models, what is the basis? The models perfect, the model can be static or dynamic; to do data assimilation, I need observations; when we did the static deterministic model even though we recognize that observations in general are corrupted by noise to make things simple; we assume as if the observations did not have any noise.

But when they came to the dynamic data assimilation, deterministic perfect model assumption; we assumed the observations are noisy. And we knew the observational covariance; so with this we have pretty much completed data assimilation into static deterministic and dynamic deterministic models.

The next topic is a model can be stochastic; the model can be static and dynamic model can also be stochastic. So, what does it mean? I am going to now start with static stochastic model; dynamic stochastic model stochasticity randomness; where does the randomness comes into being randomness? Essentially comes from one observation noise.

Another one randomness comes from model may have a random forcing function; what is the reason? Why we consider model forcing? Some time models are approximations of reality. The model captures pretty much good aspect of the physics, but still there are some leftover processes; that I have accounted for the left over unaccounted terms is called the model errors; it is not if we know what the error we are committing; we would have always taken that into account, it is not acceptable for one to know that I have committed the error and not correct the error.

So, when somebody says this is the model that essentially gives the complete understanding of the model at that time, but whatever be the model, you may want to account for unaccounted terms that is called the model error.

One way to simplify the incorporation of model errors is; to assume the model errors are random. So, addition of a randomized version of model errors makes the model stochastic, considering the observation noise makes the observation also random or stochastic. So, we are going to move into a newer realm; where the stochasticity in the observation, as well as stochasticity in the model errors both could be part of our analysis.

So when we were going from deterministic stochastic, the principles of data assimilation has to depend on statistical probabilistic ideas. I am assuming the readers are familiar with basic fundamental concepts of from probability theory. So, under that assumption; I am now going to build some of the basic tools from statistical estimation theory that one needs to be able to perform data assimilation into static and dynamic stochastic models.

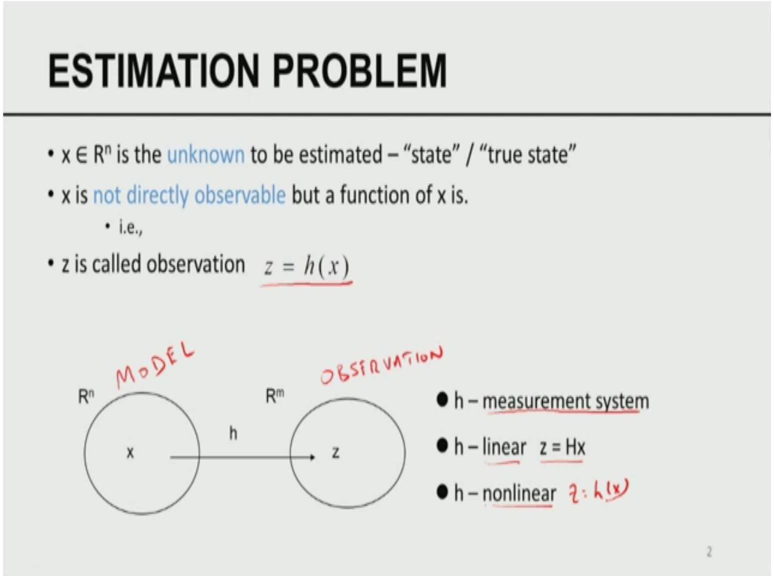
So, that gives raise to the mathematical background that underlie statistical estimation theory. Please remember from the first lecture; data assimilation can be thought of as a regression data assimilation, can be thought of as estimation. So, estimation within the deterministic context is what we finished talking about estimation within the stochastic context is what we are moving into.

So, the first topic in module 6 is called principles of statistical estimation; it is the preparatory work that we need to do to gain, an understanding of the fundamental principles involved in statistical theory. So, this is the part of the mathematical requirement; please go back, we have already talked about final and electro space matrix theory; multi-variated calculus optimization theory, matrix methods optimization algorithms.

Now I am going to be talking about statistical estimation algorithm. So, you may see this course is heavy in mathematics; why this is heavy in mathematics? Because that is what the definition is all about, if you do not understand the mathematics; we may not be able to get like crux of the algorithm; that underlie the data assimilation process itself. It is very easy to be able to use the algorithms there somebody developed, but in addition to be able to use the algorithm that somebody develops.

If you want to be able to venture into the new world of being able to develop newer methods, you need to understand the models the algorithms and the models, the data and the process of bringing; the model to data which is essentially an engineering process in my view and this engineering process involves lots of mathematical preliminaries and that is why our approach is quite mathematical.

(Refer Slide Time: 06:17)

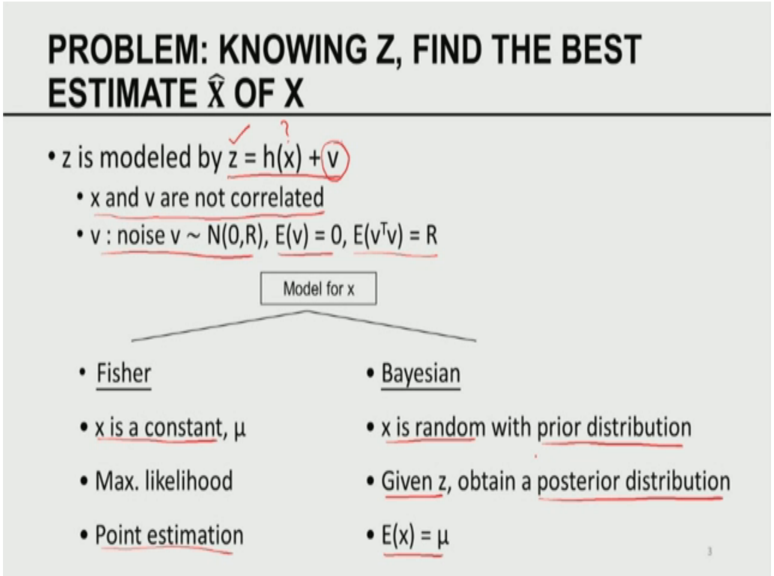


So, with that preamble I would like to be able to describe the fundamental principles of statistical estimation. So, I am going to pose the estimation problem; let x be an unknown vector to be estimated; x is called that state or the true state. I want to know the temperature in the city of Bangalore this afternoon at 3 o'clock, I want to be able to estimate that the unknown the true temperature; not known, I would like to be able to estimate this oftentimes; x is not directly observable; the state of system may or may not be directly observable.

But function of the state may be directly observable; therefore, z is called the observation. The observation is related to the true state by a function z is equal h of x; we already know that we have utilize this term it again. So, R n is a model space; x is the model state, z is an observation vector. We have the observation space which is R; R m h is the map from the model space the observation space; h essentially refers to the measurement system if h is linear, z is equal to h of x.

If h is non-linear; you simply have z is equal to h of x that is what we have. These are very familiar territory for us because we have used this several times over.

(Refer Slide Time: 07:50)



The problem knowing z; I want to be able to best, I want to get a best estimate x hat of x. So, z we know z is a related to h, but h z is related to x; x is not directly observable, a function of x is observable. So, knowingly z; I would like to be able to estimate x where does a stochastic, stochasticity comes into play.

There is an additive observation noise; we are going to assume this noise is mean 0 Gaussians, the known covariance. I will also like to be able to generalize what we have been doing; so far we have been thinking about x as a state deterministic, but the state itself could be random. Therefore, we are having an observation z; z is random in 2 ways because the unknown itself is random, if the unknown is fixed random is fixed that is called static model.

If the unknown x is fixed, but it is random; what I mean by fixed? By random it is a random variable, its value can be different based on a particular distribution. We just do not know what is the distribution based on which the values of x is selected. The distribution is the distribution that mother nature chooses. So, x the unknown is random; the state is random, the noise is random; so z is random. So, given a random observation I would like to be able to recover the x.

So, x hat is the estimate of the unknown x because x and v are random; I am going to simplify maters, assume x and v are uncorrelated. So, x are random variable that represents the natural variation for example, this year in some parts of the world; the temperature is warmer than usual in the winter. In some parts of the world, it is better than normal and these variations are related to a phenomenon called El Nino and El Nino occurs; with some rhythm over time.

So, many of the weather variables around the world have a natural variation associated with them. So, that is what the distribution of x is all about, so x could be the temperature in a specific region of the world that I want to estimate. And x is the random variable is subjected to certain natural variability controlled my other events that happens around the world; z's are observations in addition to the underlying natural variability, there is also an observational error.

So, given the observation z; I would like to be able to have a realization in estimate of the realization of x; that is x hat. So, v is the noise is normally distributed here v is 0; the covariance of v is r. So, x is the unknown I assume x in general could be random, so if you look into the statistical literature, this stochastic estimation problem; there are 2 competing schools of thoughts, one is the Fisher school; another is the Bayesian school.

Within the Fisher's school x; is assumed to be a deterministic constant and it is mu, but unknown Fisher developed a method called maximum likelihood estimation technique. To estimate mu and Fisher's technique lies; Fisher's technique; one can call it as a point estimation because mu is a point. In a vector space of dimension n; because n is the dimension of the state vector x. So, I am interested in estimating an unknown value; the vector mu and mu is a deterministic constant. So, Fisher formulated this problem as a point estimation problem and he developed a method of; what is called maximum likelihood estimate.

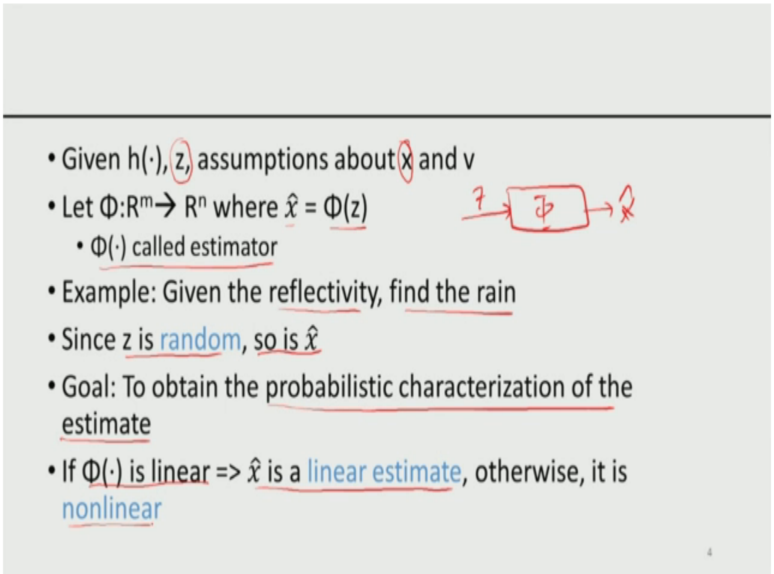
As opposed to Fisher's approach, there is called the Bayesian approach within the context of Bayesian; approach x is considered to be a random, x is sit to have a prior distribution. The prior distribution captures the natural variability of x, so if x denotes the temperature distribution around the world; the temperature distribution around the world is subjected to climatic conditions. The climatic conditions itself vary in some rhythmic fashion therefore, we can predict to some degree of accuracy the natural variability in x.

And this natural variability is captured as prior distribution; now given z, so there is a particular; they also prior distribution relates our belief as to what x would be? X access z is the actual state. So, for example, in this current year we know we are under the grip of El Nino; so we know under El Nino, what kind of temperature variations could take place? Even though, we have a prediction from the base in the prior; which are climatic data, we make an actual measurement z.

So, z contain some new information x; the prior contain some old information, I would like to be able to combine them the prior and the new information to get what is called the posterior distribution, E is random in this case mu is expected value of x expectation is taken with respect to the prior distribution.

So, you can think of x being a constant and x being random; these are 2 complementary points of view as an existing statistical estimation theory.

(Refer Slide Time: 14:45)



So, given a function H and assumptions about x and v as; we had done, I would like to be able to now concoct a function phi, which is R m to R n; what is R m z belongs to? R m. Please remember that z belongs to R m z contains information what is R n? R n x is in R n. So, I would like to be able to transfer the information from z to x z; is not x is not known, I want to transfer information from z to x, this information transferred.

I am going to represent through a function phi that maps from R m to Rn; so phi of z is equal x hat. So, phi is the process by which I by; I analyze the observation the output is x bar. So, you can think of it like this phi is the process into which you give the observation outcomes H hat as an estimate of the unknown.

So, if phi generates x hat which is the estimate of x based on z phi is called an estimator. So, estimator is a map from the observation space into the model space. So, what are the examples? Given the reflectivity from the radar that is z, I would like to be able to find the amount of rain. So, the state of the system is rain but the measurement is a reflectivity which is z in this case z is random.

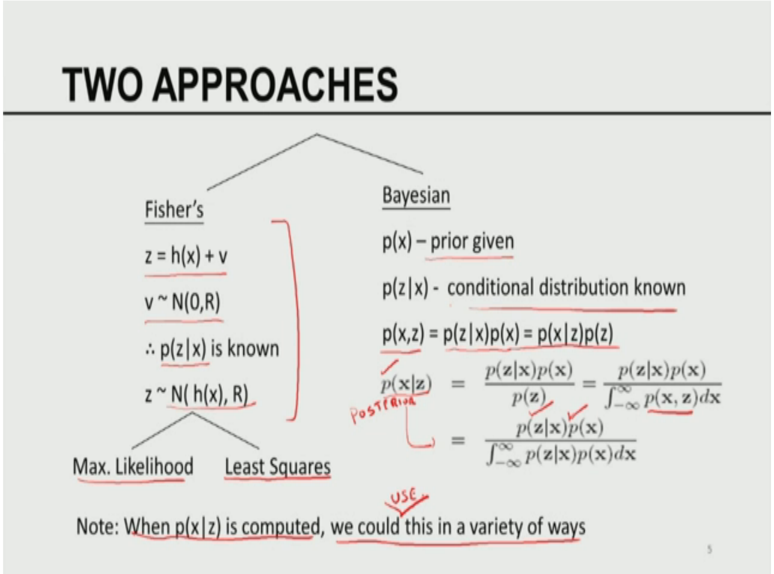
So, phi of z is a function of a random variable; so x hat is random. So, the estimate x hat is the random variable, the goal is to be able to obtain probabilistic characterization of the estimate. What are the probabilistic characterization estimate? So, there are 2 things; first we need to be able to concoct a way to design phi that is output x hat in estimate of the unknown x.

Once it puts out an estimate, we have to talk up an H hat is random; we need to be able to talk about the probabilistic characteristics of x hat, a complete probability characterization involves knowing the entire distribution of x bar. Sometimes is often difficult to get that in lieu of that; sometimes we will be contented with knowing what is the mean? What is the covariance?

So, the problem of characterizing the properties of x hat is the problem that is associated with statistical estimation. If phi is a linear function of z, then x called the linear estimate; otherwise it is non-linear. So, estimate can be either linear or non-linear, so what is the summary of that? I have a unknown x, which could be random; there is a natural variability, there is a prior distribution; I make observations.

Observations are also corrupted by noise; I know the distribution of the observation. I want to be able to combine the distribution prior with the given distribution to be able to get the property; the probabilistic characterization of x hat. So, I want to be able to design an estimator; an estimator could be either a linear estimator or a non-linear estimator.

(Refer Slide Time: 18:38)



So, in Fisher's approach x is fixed constant; so x is equal to H x plus v; v is in randomized v; v is Gaussian random. Therefore, the probability density function of z given, x is again a normal distribution. The mean of z is normal distribution which H x as the mean and R is the covariance.

So, if x is deterministic; the randomness in z comes precisely from the randomness in v and H x is added to be v is 0 mean and a covariance. Or if you add a deterministic quantity to a random quantity, it simply shifts the mean without changing the covariance and that is essentially the analysis that we have given in this discussion. In this case, there are 2 approaches to estimation; one is called the maximum likelihood estimation.

Another is the least square estimation in the Bayesian approach; b x p of x has a; is called the prior distribution; its belief that we had about the unknown the natural variability. I would like you to think of it as a information; that we have an climate that is a prior information; then at when you start taking actual observation, the actual observation as a conditional distribution.

So, given x; given a particular realization of x, the observation has a distribution that is called the conditional distribution; generally that is known prior is given. So, I can now compute the joint distribution p of x of z by simple rule unconditional probability is conditional of z with respect with given x times p of x or it can be written as conditional of x with given z; with respect to p z; by equating these two, we can now see p x of z is given by the product of p z given x times p x by p z; p z is the dense probability; density observation.

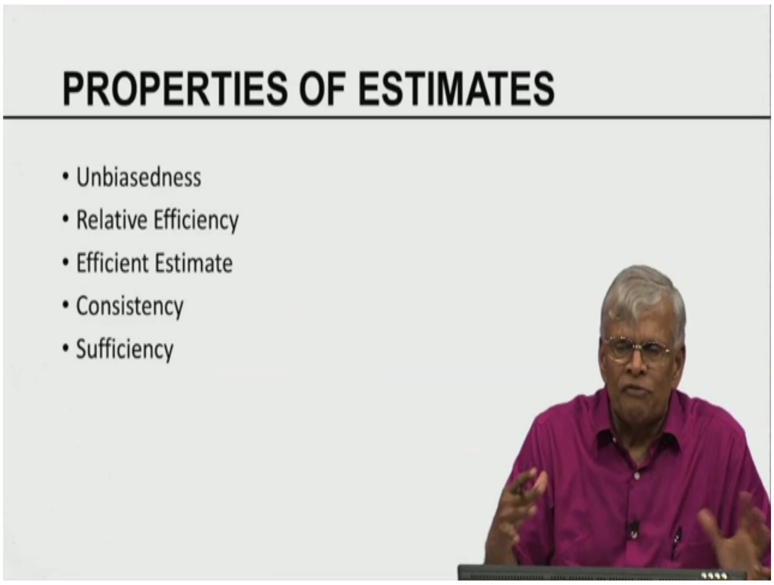
It can be expressed as the integral of the joint density, which is p x z with respect to x. If you integrate the joint density with respect x; you get p z; so this relation has come to be called the Bayes rule. So, what is the Bayes rule? Say if you give me the prior, if you also tell me the conditional distribution of the observation; I can combine them to get this and this is what is called the posterior.

So, what does it mean? I am updating the prior; prior is the belief before I came into the game when I started playing the game, I got observation. The observation gives me some new information; the new information helps me to revise my old belief, so the new belief is called the posterior. So, the old belief changes to a new belief by virtue of getting new information through observations; when pxz, the posterior is computed within the Bayesian setup.

We could use this in a variety of ways, we could compute the mean, we could compute the covariance. We can make lots of analysis with respect to different properties of x; based on the posterior distribution. So, these are the 2 competing approaches to statistical estimation in one form. The stochasticity arising purely from observation noise that known is fixed, the other one the unknown is also random; the noise also for the corrupts; the observation.

In this case, I have a posterior; I have a conditional distribution the process of combining a posterior with the conditional distribution is the one that gives you the prior and the conditional distribution, when combined gives you the posterior. So, posterior is the new belief posterior is the revised belief; posterior is the one that we should use in our decision process.

(Refer Slide Time: 23:30)



So, now we talked about the need for creating estimates unknown estimates of the known be random are deterministic. There are several properties; the estimate one has to be concerned with one is called unbiasedness. Another is called the relative efficiency of the estimate; we have to understand what is called an efficient estimate? We have to understand what is called consistency in estimation? A consistency of the estimate, we also need to worry about what is called sufficiency the estimate?

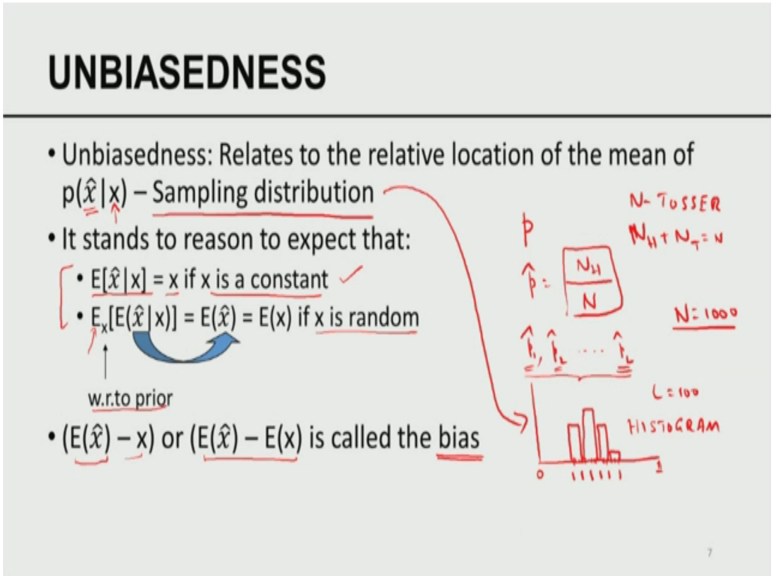
So, these are all the norms against which estimates are evaluated as if we can induce as many of these properties into the estimate as possible. Those estimates are better estimates for example, I would like to be able to have an unbiased estimate, I would like to be able to have the efficient estimate; I would also like to have a consistent estimate. So, while data assimilation is in a inputs; where as an estimation problem within the context of stochastic estimation.

We need to be aware of different properties; the estimates will process and the properties that the estimates process, depends on the design of the estimate or the function phi. So, how do you define; design a function phi the estimator such that the estimators unbiased efficient consistent and so on. So, statistical analysis has been concerned with the development of this theory for well over century, there is a very well established body of literature.

So, if you want to be able to become an expert in the area of stochastic data assimilation problem or randomized stochastic data assimilation; randomized stochastic in data simulation problems. You need to be cognizant to have very many fundamental results from this contemporary statistical literature.

So, you can see how different areas of applied mathematics are going to be involved in trying to make this area of data simulation work.

(Refer Slide Time: 26:03)



So, I am now going to define what is called unbiased? When do I say the estimators unbiased unbiasedness relates to the relative location of the mean; of the sampling distribution. So, let us talk about that now, so there are lots of little things in here sampling distribution.

So, let me first talk about the notion of a sampling distribution; suppose I have a coin, I do not know whether the coin, I do not know the probability with which the coin falls head or tail. Let p be the probability with which the coin falls head; so, 1 over 1 minus p will be the probability; it falls tail I want to be able to estimate this p; what do we do? We conduct experiments in which we do N tosses; I am doing N tosses.

When I am going to compute how many of these N tosses a head turned up. So, N H is the number of tosses where head turned up N T; where the number of process tosses, where the tail turned up and that is equal to N. So, what is the estimate of p? P is essentially given by N H by N; this estimate becomes.

So, let us assume we are picked N is equal to 1000. So, I first conduct an experiment; I get the first estimate p 1, which is given by the estimate p 1; given by the first set up 1000 experiments. Let us conduct a second set of 1000 experiments again N remains the same m H the number of times, it falls head in the first set of 1000 tosses and the second set of 1000 tosses; need not be the same.

So, let me call that as p 2 of hat; let us consider this as p L of hat. So, what is that we are trying to do? I am trying to fixed N as 1000; we are conducting an experiment and estimating the probability that the chosen coin falls head. I am doing L experiments; I will also assume L is 100 in the first set of 1000; I compute the number of heads I get p 1.

In the second set of 1000 experiments, I count the number of heads; I get p 2 it turns out, you can very easily see the number of even though; the total number of tosses remains the same. The number of heads in each chunk of 1000 tosses need not be the same, but there will be slightly different. So, p 1 in general need not be equal to p 2; in general, it need not be equal to p L.

So, if I now plot the value that p 1, p 2, p L; they will take different points in a real line. This is 0, this is 1, p L; p is not in between that; so they will take different values; there will be 100 points. Now we can divide this interval; where this lies into different bins; we can then compute the number of times. The p hat lies in here, we can put a bar, we can put a bar, we can put a bar, we can put a bar. So, the bar refers to the number of times the estimate has fallen into that bin, so this is called the histogram. The histogram essentially refers to the sampling distribution of the estimate p hat.

So, please remember p is that constant; unknown p hat is the estimate of p, p hat is random because p hat depends on the number of tosses; so p hat the estimate. So, this is the estimator; this estimator gives you an estimate, a estimate is a random variable. If I repeated this experiment L times; I get L different values of this estimate, because it is random; they are distributed in a range. I can then bin this range and compute the number of times the values of p falls in each of these. I can compute what is called the histogram; the histograms, gives you the sampling distribution.

The histogram is an approximation for the sampling distribution; so what is the sampling distribution? The distribution of the estimate conditioning of the fact the unknown is x; in this case unknown x is p x hat is p hat. So, even though the probability that the coin falls head p is fixed unknown; its estimate varies. Estimate has a distribution that is what is called the sampling distribution it stands to reason to x expect; that the expected value of this random of this estimate, which is the random variable; the conditional expectation of x hat given x is equal to x.

Then x is a constant the conditional expectation of x; x hat with respect to x. So in the first case x is a constant, in the second case x is random. So, nature picks x from the prior distribution, so this is the expectation with respect to the prior. The second one is the sampling distribution that is related to the randomness arising from sampling.

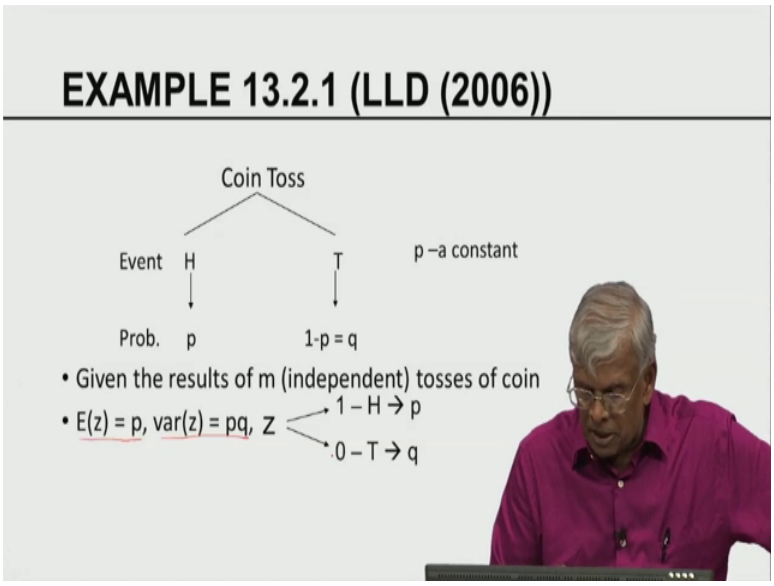
Therefore, we would expect a; my estimator x hat; to be such that the expectation with respect to the prior of the conditional expectation of H hat; with respect to x must be equal to E of x hat must be equal to E of x; what is the E of x? Is the mean of the original random variable x; with respect to the prior? So, these are the 2 conditions for un-biasness; that is there are very natural conditions for un-biasness, if an estimate is not biased is not unbiased; there is a bias, the difference between the expected value of x and the x; that is called the bias or the difference between the expected of x hat and the E of x is called the bias.

We also know biased arises in other ways for example, if you have an altimeter; if you have been using the altimeter for a long time, the properties change. So, if the actual voltage is 15 degrees; it may show, it may always underestimate this. There could be an error of minus 2; so that is called bias, there the bias if the reading of the instruments that can be corrected by calibration. You can calibrate a meter against the standard; we can correct the bias, but in here the bias arises because of the way I estimate bias is the property the estimator.

So, what is the desirable attribute of an estimator? An estimator is a desirable estimator is one; where the output of the estimator, which is an estimate. The estimate must be unbiased since we are considering 2 alternate cases, where x could be deterministic or random. In the case of deterministic x; the conditional expectation of x hat given x must be x, in the case of random; this repeated expectation, the expectation of the prior expectation with respect to prior of the conditional expectation must be equal to the expectation of x that is the mean of the prior.

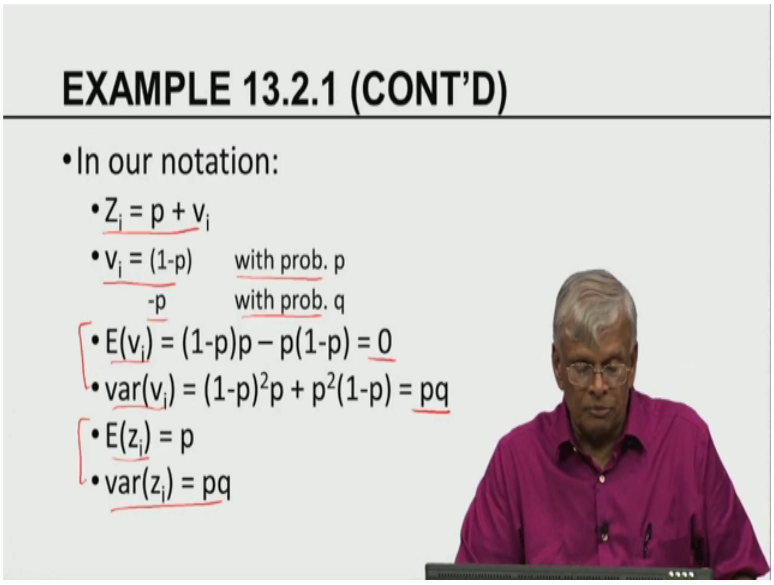
So, that is the condition we should always seek to force un-biasness in the estimates.

(Refer Slide Time: 34:44)



Again a coin toss an experiment; if I want a coin, I am going back to the coin tossing experiment even H R; T p q 1 minus p, given the results of m independent tosses. I assume m is 1; 1000 in my illustration E f z is I am going to E f z is p variance of z is p q; these results you may ask where does it come from? It comes from the standard binomial distribution, z takes the value 1; when it falls head z takes, the value 0; when it falls tail, so one with the probability p 0 the probability q.

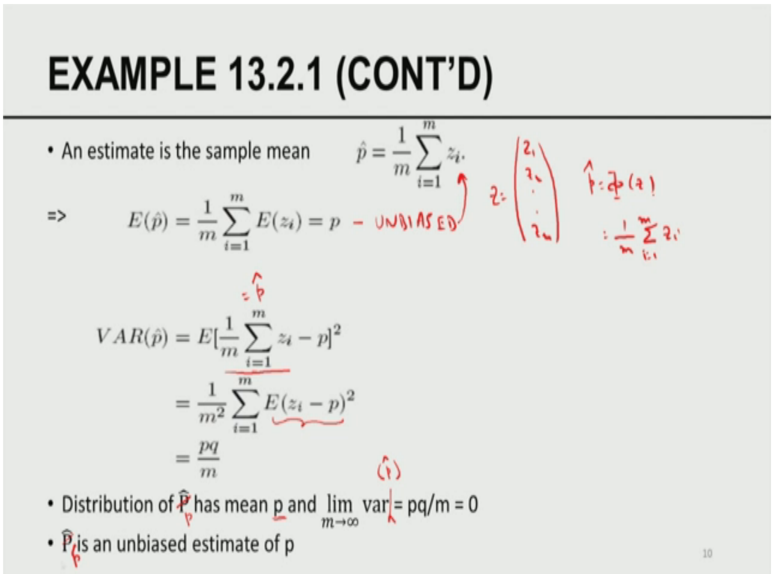
(Refer Slide Time: 35:25)



So, in our rotation z; i is equal to p plus v i; z i is equal to p plus v i; z i is equal to p plus v i; v i is equal to 1 minus p with the probability p; it is equal to minus p with the probability q. Therefore, the expectation value of v i is 0; based on this calculation, the variation of the variance of v i is p q; as it should be and the x value of z i is p; the variance of z i is p q; that is what comes out of this.

So, we calculate the properties of v and then we calculate the properties of z, these are simple calculation that comes from fundamental analysis.

(Refer Slide Time: 36:09)



Now, I am going to talk about estimation of the sample mean. Now that we have seen 2 different formulations of the estimation problem, the Fisher's formulation and the Bayesian formulation and we have also seen the definition of what unbiasness and what is the measure of bias is all about. We are going to illustrate the concept of bias; using a simple coin tossing experiment.

So, example thirteen point 2 point one thirteen point 2 point one is taken from our book Lakshmivarahan Lewis and Dhall Dynamic Data Assimilation published in 2006; consider a coin tossing experiment the events are coin fast falling head or tail the probability of head is p the probability tail is q 1 minus p.

They are assuming p is a constant; so we are following the Fisher's framework given the results of m; independent tosses of a coin. In the previous illustration, I used m is equal to 1000; we would like. So, the estimate we would like to be able to get an estimate of p; the observations are z, z is equal to 1, when it falls head z is equal to 0. When it falls tail, the probability of head is p; probability of tail is q.

Therefore, the expected value of E z expected value of z is p; the variance of z is p q anybody who has done basic probability theory and statistics should be able to recognize that. This is a simple example of a binomial Bernoulli random variable; which is taking 2 values head or tail. So, we are going to rewrite this in our notation observations are z; z is equal to p plus v i; p is the unknown v i is the noise. The unknown p is to be estimated the observations is the sum of the value of the unknown p plus the noise v i.

In order to be able to make sure the z matches with the previous description; we are going to concoct a noise v i is equal to 1 minus p with probability p; v i is equal to minus p, with the probability q. With this; first we are going to compute the expectation and the variance of v i. The expectation and the variance of v i; expectation of v i is 0, the variance of v i is p q.

Once I know the mean and the variance of v i since z i is a sum of a constant plus a random variable adding p 2; v i simply shifts the mean. So, the mean of E i; z i is p, the variance of z i is p q. So, this is the fundamental result that comes from probability theory; if you add a constant to a random variable, the distribution of the sum is the same as the distribution of the original random variable except that the variance remains the same, but the mean is shifted that is the basic idea in here

So, I am now going to talk about estimate problem, I am going to perform m experiments; m could be 1; 1000 z are the results of the tossing coin in the i th toss. Please remember; z i takes values 0 or 1; so, the sum of z i; i running from 1 m is equal to the total number of times the head came the total number of head divided by m is an estimate of the unknown. The estimate is characterized by p hat; this p hat is the random variable; p hat has an underlying distribution, that distribution is called the sampling distribution.

It can now be verified E of p hat is equal to expectation of the sum; is the sum of the expectations. Therefore, E of p hat is equal to the average of the expectations of the i random variable, z i running from 1 to m; the average of each z i is p as it was shown in the previous slide. So, the average of the expected value, of the estimate is equal to the true value; that means, this estimate p hat is unbiased.

Unbiased variance of p hat variance of the sum is equal to sum of the variances is; if the random variables are independent in this particular that is the result, that comes from basic statistics and probability theory. And here, we are concerned with the sum of the results of independent tosses. The tosses are independent therefore; there is no correlation between 2 successive results of the 2 successive tosses.

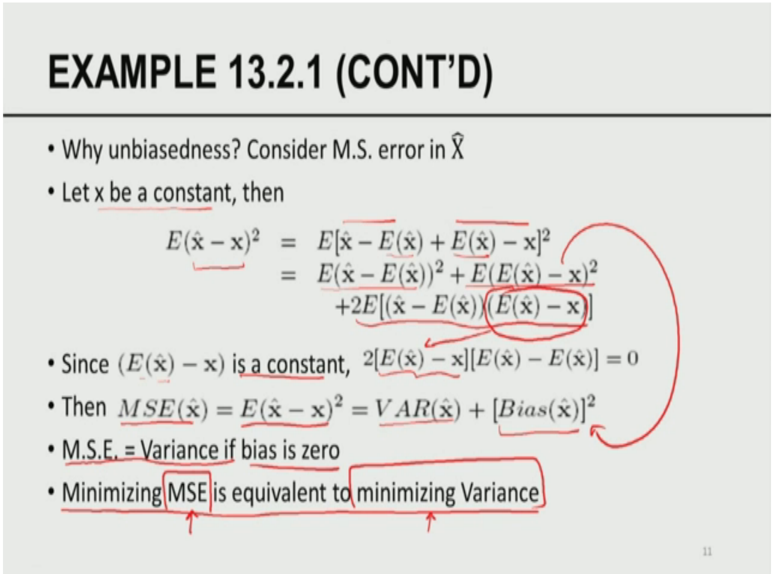
Therefore, the variance of p hat is given by expected value average minus; this is the random variable, which is the average of the; which is the estimate; this is equal to p hat. As you can see p hat minus p whole square expected value of that; this from basic probability theory relating to the properties of expectation, expectation of the sum is sum of the expectations.

So, this reduces to 1 over m square times; the sum of the variances of the individual term. The variance of each individual term is p q; I am adding m times. So, m times p q divided by m square that leads to p q divided by m. Therefore, the distribution of p hat, there should be a smaller p; the distribution of p hat has the mean p; p is the unknown to be estimated and the variance of p hat I think we should put within parenthesis.

The variance of p hat is equal to p q by m, in the limit it goes to 0. So, p hat is called an unbiased estimate and so in this case p hat is an unbiased estimate. As you can readily see the estimate has no bias; therefore, the estimate p hat as given in here is unbiased. So, that is an important attribute of this particular estimate. So, you can think of z as a set of all observations, so I would like to go back to the structure of the estimator. So, what is that we have the observations are z 1, z 2, z m; that is the vector that is given to us; that is z p hat is equal to a function phi the estimator of z.

In this case, the function phi is essentially the average of the components of z i is equal to 1 to m. So, this estimator which is given by the average is an unbiased estimator. So, that is the fundamental concept of un-biasness and is; so un-biasness is one of the properties of this particular estimator.

(Refer Slide Time: 44:41)



Why un-biasness? We are often interested in mean square errors in the estimate x hat of x. Let us go back to x being the unknown, x hat being the estimate of that. So, what if x is a constant? Even estimate the expected value of the square of the difference; the expected value of the square of the difference is called the mean square error; the error in the estimate.

So, now I can add and subtract E of x hat to this expression inside; then I can combine this 2 and combine this 2 that becomes a plus b whole square; that becomes a square plus b square plus 2 E a b. So, you get the result of three terms; again we have used the sum of the expectation; the expectation of the sum is the sum of the expectations.

Now, I am assuming x is a constant to start with. So, x bars are random variable; if I took the expected value of x bar with respect, with the sampling distribution; this also becomes a constant. So, this minus x is a constant, so the mean square error in x hat is given by E of x hat minus x whole square; that is what we have been discussing. Now since E x hat minus x is a constant; if you look at this particular term; in this particular term the second factors are constant.

I can take that second factor out as a common, as a factor out if I took that out; in here we are left with E of x hat minus E x hat. If I distribute that E operator in sign, it becomes E x hat minus E x hat that is 0; therefore, this term with the coefficient 2 in this expression becomes 0. Therefore, the mean square error now is simply the sum of these 2 terms; this term as well as that term. The first term is called the variance of x hat that comes from the fundamental definition of a variance is expected value. The expected value of the random variable minus is expected value whole square.

So, the first term is the variance; the second term as you can readily see from the definition of bias. If E x hat is minus is equal to x is unbiased in this particular case; what is that we have? This is we have and when x is a constant; E x hat minus x is a constant. So, expected value of a constant is itself therefore, we get that term equal to the second term equal to the square of the bias.

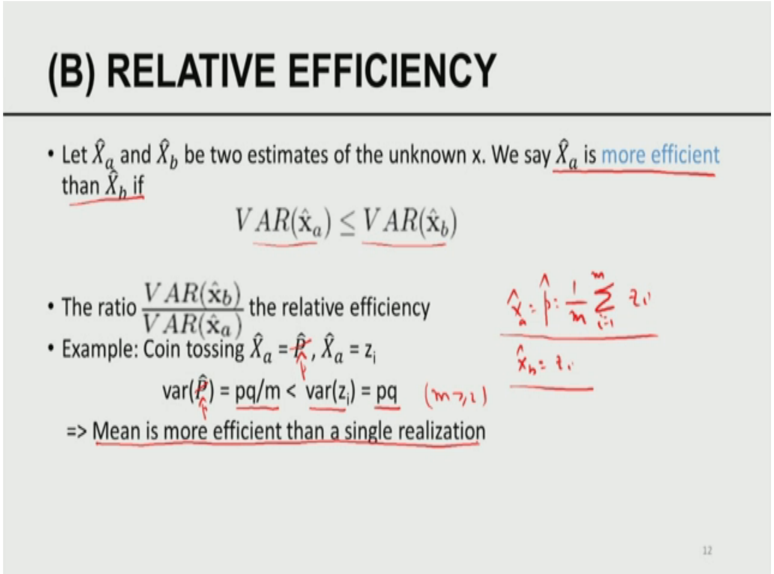
So, now you can see the impact of bias on the mean square error. So, bias is something Bias Square is always positive; so mean square error in the estimate is equal to the variance of the estimate plus the square of the bias. Since bias is always positive, the minimum value of the mean square error happens; when the bias is 0. And the minimum value of the bias possible is equal to the variance of the estimate.

Therefore, when the bias is 0; the minimum squared error is equal to the variance of the estimate. Therefore, minimizing the mean square error is equivalent to minimizing the mean square error. When there is no bias is equivalent to minimizing the variance because mean square error becomes the variance is the biases 0; this is one of the reasons why we are always looking for unbiased estimate minimum variance.

Estimation is one class of estimation that we will deal with and that relates to derivations in Kalman filters means; minimizing the mean square error. There is another criteria that comes; from again estimation theory therefore, mean square error criterion is one thing, minimum variance error criterion another thing; these 2 criteria coincide when the bias is

0 and so these 2 problems become one and the same. If the bias is 0 and that is one of the reasons why we are always motivated to find estimates with bias 0 or unbiased estimates.

(Refer Slide Time: 49:54)



Now, I am going to go to; so far we have talked about the role of bias, un-biasness and some of the reasons for seeking un-biasness and when the bias is 0; we also saw mean square error is equal to the variance.

Now, I am going to go to the next attribute of the estimate called relative efficiency. Let x a hat and x b hat be 2 estimates of the unknown x; we say x; x a hat is more efficient. We say x a hat is more efficient than x b; if the variance of the estimate x a is less than the variance of x x hat b. So, x hat a is one estimate; x hat b is another estimate, suppose somebody gives you 2 estimates of the same unknown; how do we compare?

First we compute the variance of these estimates; please remember these estimates are random variables a random variable has a distribution; hence it has a variance. So, each of the estimates being random has an associated variance, the one estimate with the lesser variance is said to be more efficient than the other. The ratio; the variance of x b to variance of x a; is called the relative efficiency of the 2 estimates in a coin tossing experiment.

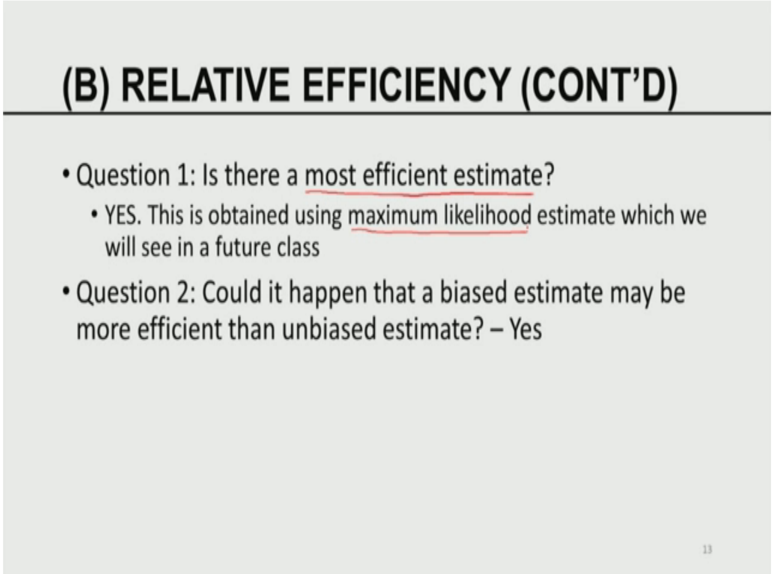
Let us assume; I have one estimate which is p hat, I think it is smaller p hat. So, what is p hat? p hat if you remember is equal to 1 over m times summation z i is equal to 1 pm. So, this is going to be my first estimate x hat a; my second x, my x hat b is the H of that is one observation. So, what is that I am going to do? Now I am picking 2 estimates; one is the average of the observations arising from m tosses. Second one is 1 observation itself; so you can see the differences in the sample size used in this estimator.

It is a simple exercise to show that the variance of p hat is p q by m, but the variance of z i is essentially p q. So, for every m greater than 2 for every m greater than or equal to 2; this inequality variance of p hat is less than variance of z i any one observation; that means, the mean is more efficient than the single observation; I think this is a very fundamental result.

So, if you are trying to estimate more the merrier; you have more observations, you take the mean of a large number of observations. If the number of observations becomes large, there is a theorem called central limit theorem. Even though the average is a random variable as the number of samples becomes larger and larger; the sampling distribution becomes a delta distribution centered around the known p.

And that is the very well known result and that result is borne by at least the essence of the result is borne by this example. Therefore, they are always whenever there are different possible choices for designing estimators, we are going to be looking for estimators that gives you estimate; which are unbiased and more efficient. More efficient means the variance of the estimate is small; if the variance of the estimate is small, the confidence of the estimate is larger; that is why relative efficiency matters.

(Refer Slide Time: 53:59)



So, the question is this; if one is more efficient than the other, it behaves us to ask a question; is there a most efficient estimate? I want to think about this. Now, if there is a possibility of improving the variance of the estimate, there is a fundamental interest in asking a question; is there a most efficient estimator? Or is there most efficient estimator? The answer is yes.

One of the theoretical ways in which one can establish this most efficient estimator estimate is by resorting to a technique called maximum likelihood. Estimate this maximum likelihood estimation technique was essentially introduced by Fisher. So, Fisher assumed; I have an unknown x, which is a deterministic constant. I have observations; the observations are going to give you estimate and I would like to have an estimate; which is unbiased, which is relatively more efficient.

In fact, I want to have an estimate which is most efficient; that means, there is nothing else which is more; which is nothing else, which is more efficient than the one that is given by maximum likelihood estimate. So, that is the theory developed by Fisher well over 100 years; well over several decades ago.

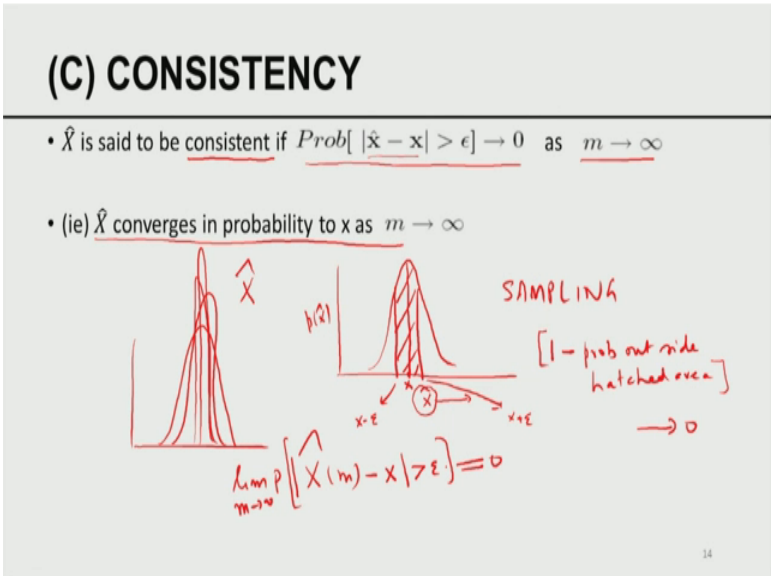
Question 2; could be happen that the biased estimate may be more efficient than the unbiased estimate. Again the answer is yes; now look at this now bias is one attribute of the estimate efficiency is another attribute of the estimate, these are 2 different attributes.

So, we need to ask ourselves when we tried to design estimates to estimate the unknowns in a data simulation problem; we need to be aware of the following question.

What are the underlying properties of the estimate? So, generate is an unbiased because there exists another estimate, which is more efficient than this. What is it take to be able to generate the most efficient estimate? Is the most efficient estimate is always a linear estimate, is it a non-linear estimate is that; is there a possibility that a biased and a biased estimate will be more efficient than an unbiased estimate?

So, these are all the class of question that statisticians have worked around and undeveloped a beautiful theory. I am trying to provide a snapshot of some of the fundamental underpinnings of this theory, because of its intrinsic interest in intrinsic relation between estimation theory and data assimilation theory.

(Refer Slide Time: 57:06)



Now, we come to the next property; which is called consistency. Consistency is again another fundamental attribute; please recall that we have seen x star; I am sorry x hat is a random variable. If x hat is a random variable, there is going to be a probability distribution which is called the sampling distribution of x hat. Let us assume the sampling distribution comes like this; so this is a sampling distribution we talked about the method of generating sampling distribution in the context of coin tossing in the previous slides. So, what is that we are looking for?

Let this be the x; that is unknown x hat is an estimate, the x the estimate is random; we would like to ask what is the following question; so, this is x I would like to consider an epsilon strip plus or minus epsilon, this is x minus epsilon; this point is x plus epsilon. If you consider this trip; if you integrate the probability density from x minus epsilon to x plus epsilon, that is going to be a total probability mass under this curve.

So, the probability that the absolute value of the difference between x hat and x; that is the probability that I have pictured here and well I should not say; this is epsilon may be, I think there is a bad notation. I will change my notation; a little bit please forgive me, this is not epsilon because let us assume this is x minus delta and x plus delta you know; I will go back to epsilon; sorry you know I am that is right; sorry I will go back to epsilon that is right.

So, x hat could be x could lie in between x plus epsilon and x minus epsilon; that is correct my original statements are right, that is correct this probability. So, the probability within the hatched area is 1 minus the probability outside the hatched area; probability outside the hatched area. Now hatched area outside the hatched idea; so I would like to ask in the following question; what is the probability that my estimate x hat will lie outside of an epsilon band around x a?

That is the question; that probability is given by 1 minus; the probability that it will lie inside. If this probability where to tend to 0, that will happen only when the probability of the hatched area is closer to 1. If the probability of the hatched area is closer to; is becoming closer to 1 means what the probability distribution becomes more and more peaked, it was originally like this; then becomes like this, then it becomes like this, then it becomes like this.

So, we are looking for a thin narrow region around the unknown within which the entire sampling distribution the probability mass resides. So, that outside of this thin strip; the probability mass is 0; that is what exactly this, the relation tells you. As m tends to infinity; what is the m? M is the number of samples. As I increase the number of samples, my estimate x hat as a random variable; finds itself in an epsilon strip around the x with probability 1.

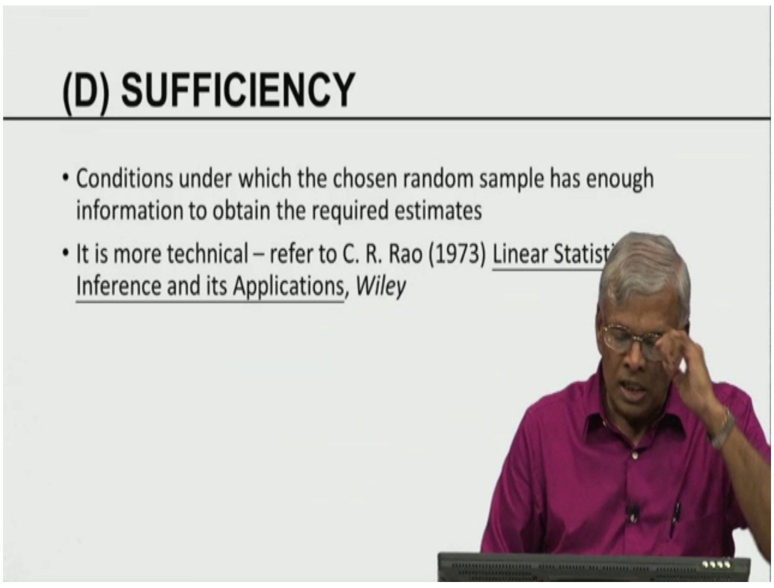
What does it mean? The probability of my estimate lying outside the epsilon band; that means, x hat minus x is greater than epsilon, it goes to 0. If the sampling distribution satisfies this property; that is what is called consistent estimate; that means, as the number of samples increases my estimate becomes closer and closer and closer to the truth and the probability of it being not equal to the truth, goes to 0 continuously. As the number of samples goes to infinity; that means, my estimator becomes more and more closer to the truth.

In the probabilistic language, this is the very special connotation or a special name, this is called convergence in probability. So, what does it mean? The estimate x hat; which is a function of the number of samples in the limit as m goes to infinity; as m goes to infinity lies in a region which is very small; that means, it lies this the probability of this goes to 0.

The limit; the probability is 0, if my estimate I had; is satisfies this property it is called consistent. So, consistent estimators are very natural choices convey; so this is called convergence. In probability of x hat; x in the probabilistic language; so consistency of our estimate is another fundamental attribute. So, we have seen three attributes biasness or un-biasness relative efficiency. Efficiency most efficient estimate and then the third one is called consistency.

So, what are we looking for? We are looking for consistent unbiased most efficient estimate is what we are looking for; that is the ultimate goal from a statistical perspective.

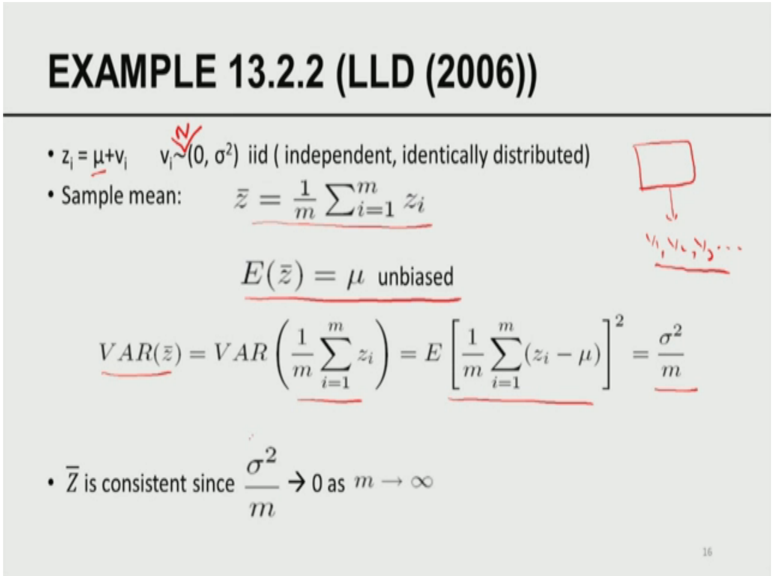
(Refer Slide Time: 63:59)



Sufficiency is another criterion we are not going to go about too much into the discussion. Sufficiency is a little bit more technical conditions, under which your chosen random sample has enough information to obtain. They are quite estimate; that is the relate to the sufficiency; in other words that is the chosen sample that is used to estimate has sufficient information to provide you good estimate.

Under what condition such sufficiency can be guaranteed is a very technical condition. I am not going to go into the details, one of the most thorough discussion of all these attributes, biasness, un-biasness, relative efficiency, maximum efficiency, consistency, sufficiency all these properties are discussed in great detail in one of the classic books on statistical analysis by Professor C.R Rao; published in 1973; is a classic book Linear Statistical Inference and Applications and in my view anybody who wants to do data simulation; especially in the statistical arena should have a copy of this book in their personal library. It is a bible with respect to most of the fundamental statistical principles and their applications.

(Refer Slide Time: 65:18)



Now, I am going to continue the example; let mu be the unknown, but constant we are simply concerned with the coin tossing experiment. Again mu is there known z i is equal to mu plus v i; in this particular case we are assuming v i is normal; I think it should be normal v i has a normal distribution v i's are independent, identically distributed; what does it mean?

There is a box random number generator; out of that box, I can continuously keep asking and it will deliver arena number. This sequence is v 1, v 2, v 3 and so on; these are independent, the same sense that if I am trying to toss a coin, the results of the tossing a coin are also independent. So, that is what is I did up; what is iid first, i refers to independent; the samples are being independent. The second i refers to the fact all these samples are drawn from the same distribution, the distribution does not change from one drawing to another drawing.

So, iid in independent identically distributed is one of the standard assumptions used in estimation theory to start with. So, this is very similar to the coin tossing experiment, but not quite the same because in the coin tossing experiment, the evens are 1 or 0; head or tail, but in here I am assuming there is an unknown mu; I can observe the unknown through z i; z i is equal to mu plus v i, v i is not a discrete now is does not take 2 values.

This v i takes; v i has a continuous distribution Gaussian 0 mean and sigma square as the variance. So, if I have a bunch of m observations; what is the estimator? Estimator is in this case, we call it z bar; z bar is the average of all the z i. So, this is the formula for the estimator and the estimate is z bar, by taking the expectation of z bar using the fundamental principle, the sum of the expectations is equal to expectation of the sum; one can readily verify E of z bar is mu, hence this estimate is unbiased.

So, this is an unbiased estimate much like the average leads unbiased estimator. In the case of coin tossing, as well in this case; we can also compute the variance of the estimate. Please remember, the estimator is a random variable; this is equal to the variance of the average. The variance of the average is given by this formula by invoking to the standard definitions of variance.

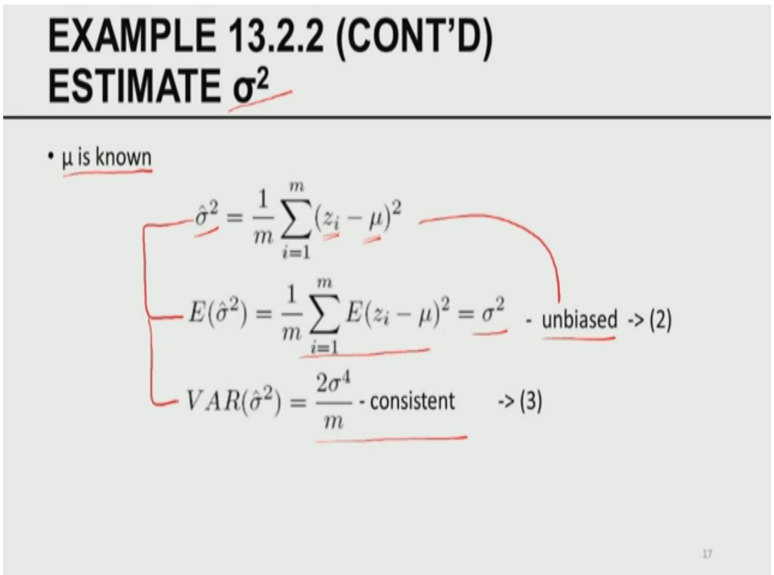
From basic probability theory, a little calculation will reveal; this variance is given by sigma square by m. Please remember; in the case of coin tossing experiment its p q by m; in this case sigma square by m very similar. So, you can readily see the variance of the estimate z bar, which is the average of all the; z i is 1 over m times, the variance of a single random variable which is sigma square.

So, as m goes to infinity; sigma square over m goes to 0; that means, the variance of the estimate becomes closer and closer to 0; that means, the sampling distribution becomes peaked if the sampling distribution is peaked that is called consistency. So, this estimate the average of all the observations is simultaneously unbiased; it is also consistent. So, this is the reason why we say well; if you want to estimate something infinite it is asymptotically it is going to converge to the exact value.

But you may not be able to have the; resources need to do unbounded number of experiment. So, if you have a large sample; using large sample if you take the average, the average arising at a large sample is unbiased and also reasonably good efficiency, it is also consistent depending on the number of depending. The efficiency relates to the number of samples you have, so you can see why average is a estimate.

We have also seen in our static inverse deterministic inverse problem; average is the best least square estimate for example, you may remember the following experiment. Suppose I want to estimate my weight, I make m measurements in m different scales; are m measurements in the same scale at different parts of the day. So, I have m measurements which all the different; I would like to be able to have a best estimate of my weight of a least square theory tells you the average gives you the best least square estimate of your weight; given that you have m observations. The main dependent observations of you weight same thing in here; the average of the observation gives you the estimate; which is simultaneously unbiased and it is sampling, variance goes to 0.

(Refer Slide Time: 70:58)



So it is asymptotically very efficient, it is consistent is unbiased example 3.22; continued in the previous example exercise; we assumed mu to be unknown. So, we estimated the mu now I am going to consider the other part of the story.

Let us pretend mu is known, I do not know sigma square. So, let us go back to the problem; please in here z i is equal to mu plus v i; mu is the unknown constant, v i is the noise with the 0 mean and sigma square is the variance. So, I can formulate several different estimation problem knowings is assumed. I know sigma square estimate mu that is what we finished.

Now what are we going to do? We assume that I know mu, but I want to be able to estimate sigma square; that means, I have I know an unknown, I know the observation is unknown plus some additive noise; that additive noise has an inherent variance. I do not know what the variance of additive noise? My goal is to be able to estimate sigma square; what is sigma square? Sigma square is the variance of the noise in the measurements. I would like to be able to estimate sigma square; how does assumption we use not?

There is one more version of the problem; mu is not known, sigma square is not known. So, you can see there are three kinds of problem; sigma square is known, mu is not known, mu is known; sigma square is not known. As estimate sigma square; mu is not known, sigma square is not known, estimate both of them simultaneously. This is a very classic example; every student in statistics generally go through this.

The aim of this exercise to be is to acquaint ourselves with the fundamental principles of properties relating to estimators and estimates; namely un-biasness relative efficiency, asymptotic efficiency, consistency and so on. So, let us concoct an estimator for sigma sigma square; if I do not know the variance, I am going to have an estimator which is sigma square hat. So, z i these are the observations I know mu from basic probability theories, the variance must be expected values of the square of that.

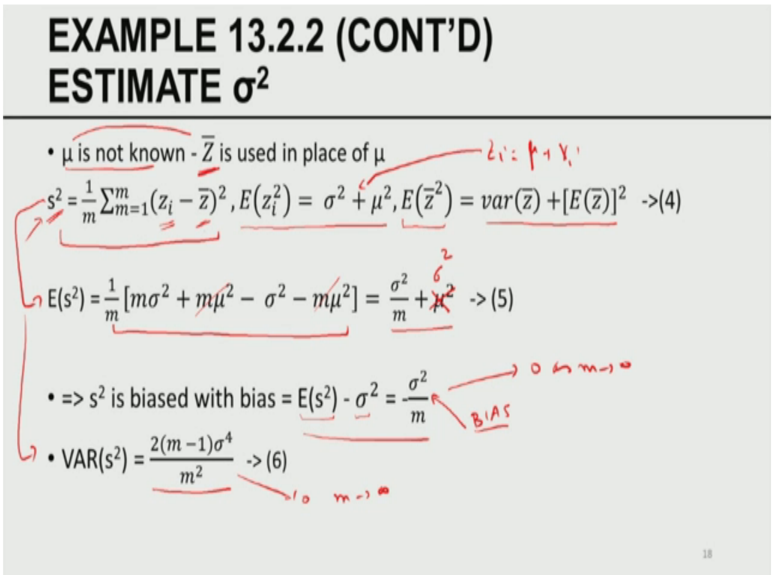
So, what am I going to do? I am going to take the average of the sum of the difference between z i. So, z i minus mu is the error; if the sum of squared errors is the average of the sum of the square errors you can see the least square principle; comes in here also in the underpinnings of list square. You can see here also, but sigma square is a random variable because z i is a random.

So, expected value of the estimate sigma square hat is the sum of the expectations. So, by applying that simple rule; it can be verified that the expectation of the estimate is equal to the true value. Therefore, this estimate is unbiased; I am not going to prove this it can also; one can also compute the variance of this sigma square, it can be shown that the variance is 2 times sigma to the power 4 by m.

As m goes to infinity; this variance of the estimate goes to 0, that is consistent here. There are lots of homework problem here, I would very strongly encourage you to use simple principles of basic statistics and probability theory to be able to compute the variance of this. So, this is the random variable; it is the mean, it is the variance please compute the variance and verify; I am heating under all the major conclusions. Some of the derivations, I am going to leave it as an assignment for you to be able to do. I think it is a worthwhile assignment to be able to check, whether you understand some of the fundamental principles involved in calculating these quantities.

Especially sample moments and properties of and analyzing the properties of sampling distributions. So, in the previous case; what is that we have seen? if mu is known; sigma square is not known, I can estimate sigma square; I have an estimator which are unbiased and consistent much like the estimate for me.

(Refer Slide Time: 75:35)



Now, we are coming to the harder case; I do not know mu I want to estimate mu, I also want to estimate sigma square. So, then mu is known sigma square is not known; the estimator for sigma square is called sigma hat square. When mu is not known, sigma square is not known; I am going to call it sigma bar is the estimate of mu; I am going to call S square as the estimator for sigma square.

So, this is the f square is the estimator for sigma square c bar is the estimator for mu that. So, z bar is simply the average; so how do I estimate? The variance this is the sample value; if mu had been known, I would have used mu here, but mu is not known; I am going to use the sample mean in here. I am going to compute the difference square; some of the square difference average value.

So, this is an estimate of the variance; when the mean is not known. This is the estimate of the mean given a particular samples, so I do not have any truth; I simply have to rely everywhere, but estimates whatever I have. So, from basic principles of the definition of variance it can be real it can be verified that E of z i square is equal to sigma square plus mu square. Because you already know z i is equal to mu plus v i; I am sorry mu plus that is v i.

So, from here we get this formula; we also can compute the expected value of z square. The square of the average; it can be verified that is given by this, that is given by the formula 4. Again, it very simple calculation from basic probability; there in statistics if you take a good course in probability theory, a good course in basic statistics where you will do all these computations in detail; I am assuming many of you have taken courses of this type; if not this is a motivation for you to be able to learn some of the fundamental principles of estimation theory. I think you can use this as an excuse to learn something you probably have not had an occasion to learn.

So, I am doing all the basic ingredients now z bar S square E of z i square; E of z bar square all given in 4. Now, I am going to ask myself what is going to be the mean of the estimate of the variance. And that is given by this formula, again it will take for 5 to 10 minutes for somebody to derive this. But I would like you to go over the detail; use the expressions and 4 to do this.

If you simplify this; it becomes this therefore, the estimate of E square E S square is equal to sigma square by m plus mu square. Look at this now; the actual value of expected value must be sigma square. Therefore, S square is a biased estimate and the bias is given by minus sigma square by m. I can also compute the variance of E of f square that is given by this; the variance is given by 2 times there minus 1; by m square times sigma to the power 4.

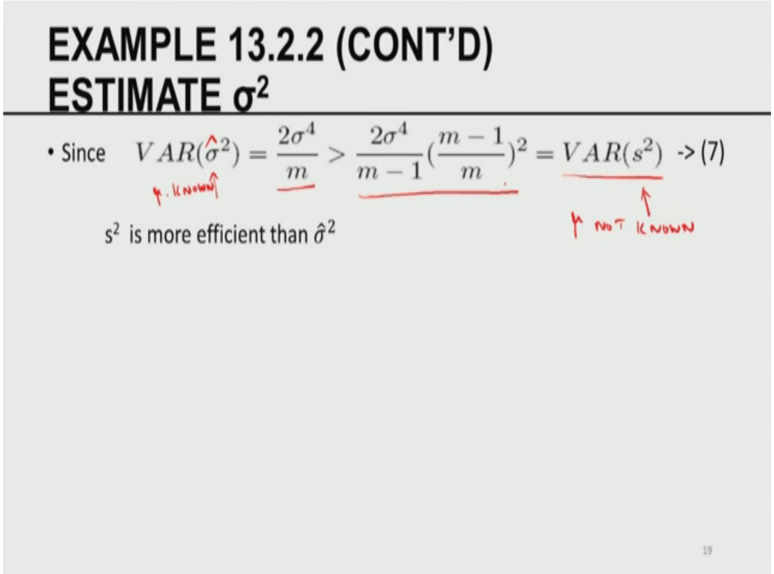
So, you have I think that is an error here. So, this will get cancelled with this; that is correct, so this must be sigma square sorry this must be sigma square that is correct that is right sorry for the error there is sigma square. Therefore if you consider E of S square minus sigma square, that is equal to minus sigma square m. And please remember that that is the bias. So, now you can really see I have an estimate which makes sense, but there are estimates the bias estimate.

So, far we have seen unbiased estimate for the first time, we are seeing an estimate which is a very natural estimate. But it turns out to be biased; it is a variance, the variance is given by this expression. Now, if you let m go to infinity the bias tends to 0 the bias tends to 0 as m goes to infinity.

The variance also goes to 0; as m goes to infinity. So, what does it mean? This estimate is asymptotically unbiased, but final sample is biased, but this is asymptotically; it becomes consistent. So, consistency and biasness un-biasness with respect to finite samples infinite samples. So, what happens? Asymptotically may not happen for a finite sample.

So, you in statistics; there are always 2 types of theories, finite sample statistics and asymptotic analysis. The asymptotic analysis are rather easy than finite statistics, we generally derive conclusions for the finite sample statistics by looking at the asymptotic analog of the finite sample statistics; to be able to judge the impact of not having infinite number of samples.

(Refer Slide Time: 81:46)



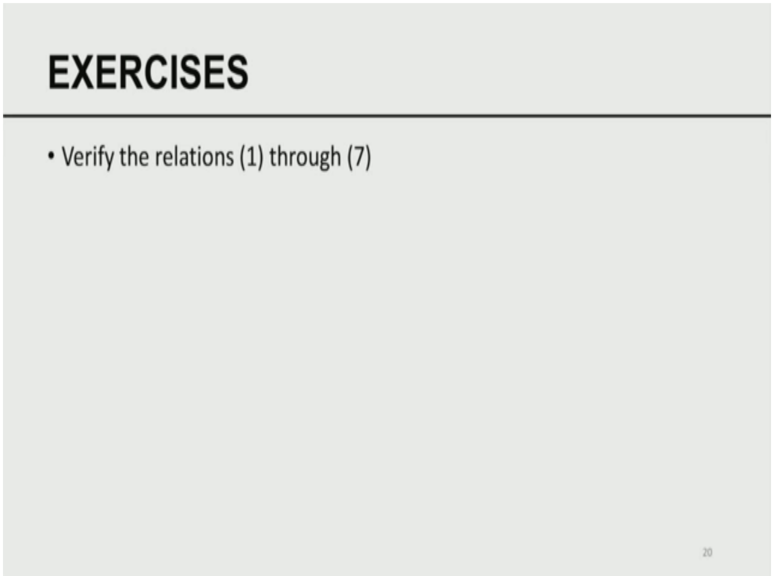
So, that is very clearly bought out by this example here. Again here again I have 2 kinds of estimates, now the variance of x square from the previous page is given by this. If the variance; let me go back the variance of sigma square is 2 sigma to the power 4 by m sigma to the power 4 m. This is variance of hat, I believe I am sorry this is hat; so this is one estimate of the variance, this is another estimate of the variance. This estimate of the variance assumes the mu is known; this estimate assumes the mu not known. We already know, when mu is not known; this estimate is biased, we know this estimate where mu is known; the estimate is unbiased. So, the variance of the estimate is given by this; the variance of the estimate is given by this.

This is something extraordinary the unbiased estimate has a larger variance than the bias estimate wow that is a very nice interesting property. So, what does it mean the unbiased estimate S square is an m; the sigma square hat is the unbiased estimate, this is the biased estimate? So, the biased estimate is more efficient than; then there is estimate; when it comes to the question of variance.

So, here you have to see the choice of estimator; what are the given conditions, under which you design the estimator, what are the knowns? What are the unknowns? All these things matter in the design of your estimator; which spits out the value of the estimate. So, the properties of the estimate very much related to what is known? What is not known? And how the estimator is designed? And we have to deal with the properties of the estimate from many of the different dimensions biasness efficiency, relative efficiency, consistency relative efficiency tells me which one is more efficient than the other.

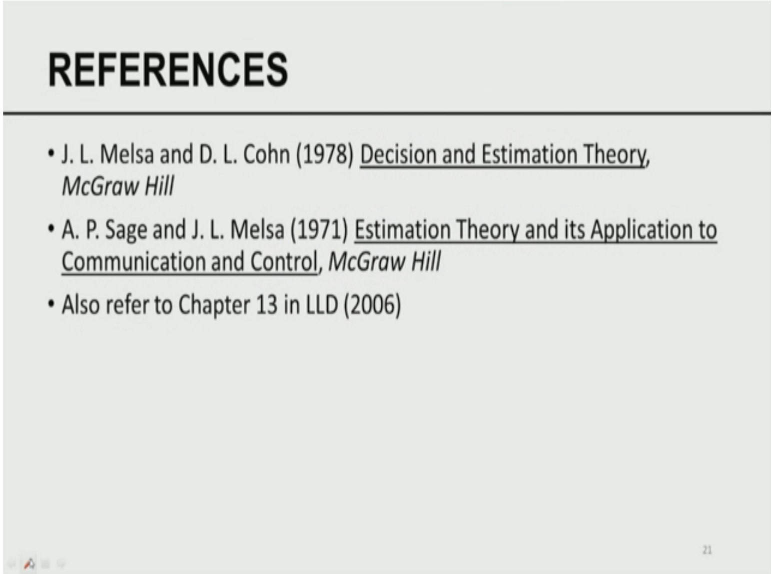
So, you simply cannot say unbiased estimates are the only thing that is of interest. We have already seen; if an estimate is unbiased, the mean square error is equal to variance. So, minimizing mean square error is equal to minimum variance that is an advantage of unbiased estimate. But if you are interested in the overall efficiency, the estimation you cannot rule out the possibility of introducing a small bias. In the estimate to be able to get more efficiency; so it all depends on what you are ultimate go is when you want to be able to estimate the unknown.

(Refer Slide Time: 84:57)



With this we come to the end of the first discussion on the design and other properties of statistical estimation. I would like to ask the reader to verify all the relation the variance expressions I had given. And I would like to very strongly encourage you to be able to derive these from the basic probability theory and statistical experience you may have had.

(Refer Slide Time: 85:27)



And the next slide provides you couple of very good references. These are some of my favorite Melsa and Cohn; 1978 Decision Estimation Theory is a small book published by McGraw Hill is an excellent book largely tailored to Engineering audience; especially Electrical Engineering audience, within the context of communication theory estimation and so on.

The book by Sage and Melsa is another wonderful book Estimation Theory and its Application to Communication and Control. It is tailored to electrical engineers and communication engineers; I coming from an engineering background, I particularly like these 2 of course, the book by C. R. Rao is the ultimate Bible; when it comes to question of statistical principles and techniques. With this, we conclude our discussion of the properties of estimates.

Thank you.