

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

Lecture – 82
Lecture 36B - Goodness of Estimators 1 -1

Let us now move on to studying the properties of estimators. And this is going to be somewhat long interesting some time tedious, but very very essential journey. This is what is missed out in many many estimation courses; I mean simple estimation courses that you take that you must have gone through already.

So, we will predominately talk of is goodness of estimators, but as I said yesterday it is not appropriate to blame the estimator if it produces a poor quality.

(Refer Slide Time: 00:40)

The slide is titled "Fisher's Information and Properties of Estimators" and "References". The main heading is "Learning Goals". Below it, it states "In this lecture, we shall learn the following concepts / topics:" followed by a bulleted list of topics: Goodness of estimators, Fisher information, Bias and Variance, Efficiency and C-R Inequality, Mean Square Error and MMSE, Consistency, and Distribution of estimates. At the bottom, it includes the name "Arun K. Tangirala", "Applied TSA", the date "October 19, 2016", and the NPTEL logo.

Suppose I go to a builder and I ask for a cost estimate of constructing a house. The final estimate to call goodness of the estimator that the builder gives me depends on two things: the information that I have given to him and the ability of the builder to take that and give process that and give you a good estimate.

So, it may not be appropriate to always blame the builder, there is some responsibility and our part as well. Likewise here, it may not be appropriate and it is not appropriate to blame the estimator all the time hold it responsible for delivering a good estimate, there

is also huge responsibility on the part of the experimentalist to generate what is known as informative data. And therefore, early on the Fisher was; I do not know how many of you he was actually an agriculturist. There are many people who have contributed to maths and statistics who are not mathematicians and statisticians; you dread to imagine what if they were. Without being in that field they have actually contributed enormously; (Refer Time: 01:57) for example.

So, Fisher came along and he said where let me quantify this so called information, because somebody has to do that and I have to find out way of knowing up front given data how informative it is with respect to the parameter I am estimating. Remember, information content is the relative, you cannot ask this absolute question is the data informative; it is an imposed question. You have to ask is; yes the data is informative but it depends on what you want. So, the complete statement is the data informative with respect to a parameter or some parameters that you are estimating. And Fisher's information is a metric that gives you that quantifies the information. Later on more complicated versions came along like Kullback-Leibler divergence and so on, but Fisher's information was first on the fore front.

We have list the other properties that I have listed here there is something that I have talked about, but what I am trying to I tell you here is we will formally in define bias variance efficiency and so on as we go along.

(Refer Slide Time: 03:09)

Fisher's Information and Properties of Estimators References

Fisher information

Fisher introduced the notion of information in a data through a series of works by and some existing results. Intuitively, larger the information index is, the "better" the estimator is.

The Fisher information (FI) (Fisher, 1922, 1950) is based on the **likelihood function** of the given data.

The likelihood function stems from the notion of conditional probability, i.e., the probability of observing an event within the vicinity of given data.

Arun K. Tangirala Applied TSA October 19, 2016 NPTEL 4

So, let us move on to Fisher's information. As I said in early 1920s and so on Fisher's seminal papers on the concept of likelihood, of course he coined this term much later on initially he coined the term inverse probability and so on for two reasons: his goal was to obtain optimal estimates of parameters of distribution of a pdf. So, simple things suppose I give you I generate data from some random number generator and I give you data and I ask you to estimate the parameters of the pdf. There are two things that you would require to do: first you have to guess the pdf, you have to guess from which distribution I have sampled the data and then you have to sit down to estimate the parameters of the pdf. What we mean by parameters is, if you take a Gaussian distribution how many parameters do we have? 2.

If you take again uniform distribution then again you have 2. If you take chi square distribution what are the number of parameters that you have? 1 degrees of freedom, so every distribution or density depending on the case has some parameters and you want to identify that. You can simplify the problem and Fisher decided that the problem has to be simplified further by assuming a pdf. Suppose I give you the pdf also I tell you that I have generated data from Gaussian pdf, if I have done that then the only goal that you have objective that you have is to estimate the parameters of the pdf in an optimal manner.

And for this purpose Fisher introduced the notion of what was called likelihood later on. And I will first explain what is likelihood, then I will explain what is Fisher's information; because Fisher's information and maximum likelihood approach both rely on this concept of likelihood. So, what is this concept of likelihood? What we did just say? We said that I am going to be given data and I am going to be also given the form of the pdf and the objective is to estimate the parameters.

Suppose, I gave you the pdf; so now the problem statement is as follows that is to understand the likelihood it is a very simple concept. Suppose I gave you the pdf, generally what we use the pdf for? Pdf meaning I give you the parameters also, what would you use the pdf or computing probabilities? One of the common uses of pdf is to compute probability.

Now if aware to ask, what is the probability of obtaining data within the vicinity of what I have observed? Remember I have observed some data, I have a record of N

observations and I want to ask this question what is the probability of obtaining data within whatever I have observed; I mean it is not a yield pose question that I have that is why I am saying within the vicinity, I cannot ask exact question. I cannot ask what is the probability of obtaining the data that I have obtaining; because that would be 0 by probability measure.

So we will ask a more theoretically correct question; what is the probability of obtaining a data within an infinitesimal neighbourhood of whatever I have observed.

(Refer Slide Time: 07:02)

Fisher's Information and Properties of Estimators References

Likelihood function

The probability of obtaining data within the vicinity of y_N is given by (with some abuse of notation)

$$\Pr(y_N < Y < y_N + dy_N) = f(y_N|\theta)dy_N \propto f(y_N|\theta) \quad (1)$$

For a given y_N , the probability is solely a function of θ . Fisher's argument (and the likes of it) rests on the **maximum likelihood** premise that

Among all possible values of θ , the one that maximizes the probability, i.e., the one that renders the event most likely is the winner!

Arun K. Tangirala Applied TSA October 19, 2016 NPTEL 5

That would approximately f of y given θ , N here is a vector of observations went to us times $d y$. Of course, you just have to imagine this in a N dimensional space do not think of it in a single dimensional space. It would be unidimensional if you are looking at a signal observation. But, now since the observation is fixed you can say this probability is proportional essentially, because that infinitesimal quantity $d y$ is kind of fixed; the only thing that determines the probability is your f of y . So, you can say the probability of obtaining the data per within the vicinity of the data is proportional to the pdf at that point.

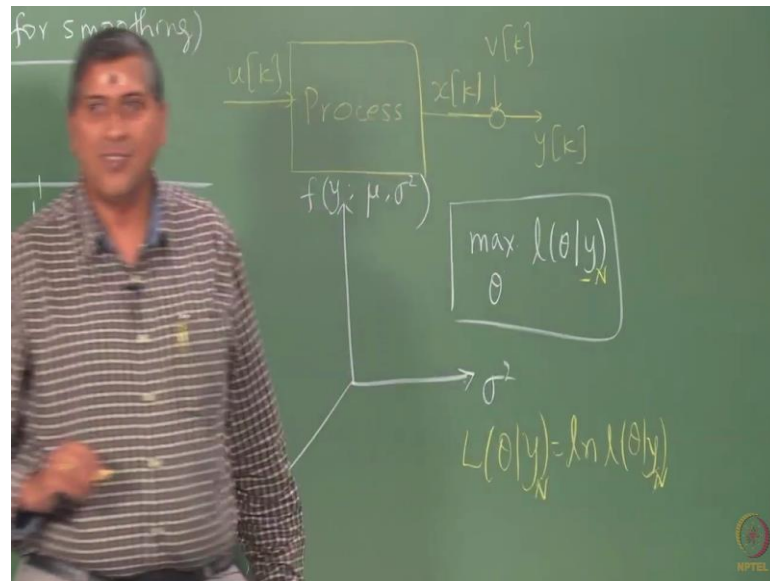
We are not saying is equal to not it is proportional to it. Why are we looking at this? Because Fisher's premise and it also make sense, whatever the premise that Fisher has given which is that of maximum likelihood is; suppose now I do not know θ which is the practical problem and I have the data within then there are many periods that could

have generated the data. When I am given only the data there are many periods meaning I may fix Gaussian, but there are among the Gaussian there are many values of parameters. So, that is what we mean by many pdf. So, several pdf's are candidate pdf, and I want to pick among all the candidate pdf's one pdf that would have produce this data. And Fisher's argument was that it is that pdf which produces this data with maximum probability which is the winner.

So, here the pdf form is fixed, the parameters are free to vary. In the parameter space which parameter do I pick? I plot f of y and θ that is f of y given θ for different values of θ . And I pick 1 it has the maximum value, why? Because I am assuming this is only an assumption I may be wrong, but it works for many many situations. I am assuming that among all the candidate pdf that run to produce this observation the winner was the one that had that resulted in maximum probability. That does not mean that I am correct, why? I mean even low probability pdf could I have also produce this, but when I do not which pdf has produced I need some criterion and Fisher proposed this criterion.

So, imagine that you have a Gaussian pdf producing the data, there are two parameters to be estimated which are μ and σ^2 , I plot f of y ; that is I plug in the observations and also now plug in different possible values of μ and σ^2 I would get a three dimensional plot. With μ and σ^2 on the x and y axis I am sorry, and on the z axis you the f of y .

(Refer Slide Time: 10:40)



So, this is how you would plot. So here is mu, here would be sigma square, and this would be f of y given mu and sigma square; that is for different values. And the maximum likelihood principle is based on the premise that the one parameter that is optimal is a one that maximizes the value of this pdf. Why is that? Because if the pdf is maximized the probability is also maximized. Remember, d is independent of theta. That is a point you have to note. The probabilities f of y times d y, but d y is does not depend on theta only f depends on theta. So, if f is maximized, probability is maximized, so we are assuming that among all the pdf's that produce the data within it is vicinity the one that produces with maximum probability is the most likely value of theta. And that is why Fisher called this initially as inverse probably, because this is an inverse probability problem the event as occurred and you are trying to figure out what us what theta could have produced it. Whereas, the general probability question is your given theta and you are trying to compute probability, parameter estimation problem is always considered as an inversion problem in statistics. The event as occurred and you are trying to figure out what could have produced the event.

So, with this idea Fisher proposed the likelihood function. The likelihood function is now nothing else, but the pdf itself, but there is a fundamental difference between these two. The pdf, when you talk of pdf's thetas are given and it is only a function of y where you are I observation space. Whereas, when it comes to likelihood, what is fixed? Because you have given data and searching in space big difference, mathematically they are the

same. So, that is why different names had to be attached there purposes are different; f of y given θ is used for computing probabilities, whereas l of θ given y is used for estimating parameters.

So, that is why different names are given for all practical purposes in fact not only practical even theoretical likelihood function is the pdf itself. There may be proportionality constant, but that does not change the nature of the optimal values of θ . This is essentially the maximum likelihood principle for you, that is all. Beyond this it is all about optimization. So, what is the maximum likelihood principle? Find θ such that this is your MLE.

Very simple and always simple statements have profound solutions. It turns out that this likelihood is not as friendly as the log likelihood. That is when it comes to tractability in optimization when you want to find solutions, when we say tractable it is ability to find the solution in an easy way. Since, pdf's are always non negative valued or in fact you can say positive valued we can offer to take logarithm and maximizing the likelihood is the same as maximizing log likelihood.

So, Fisher therefore proposed to maximize log likelihood instead of likelihood. So, we introduce this big L which is nothing but logarithm of the likelihood. So, you would always maximize this log likelihood; that see convention that is used in literature the big L is for log likelihood, the small l is for likelihood. So, hopefully now you understand the concept of likelihood, now it is all about if you want to obtain let us say MLE of mean. So, if you go back to yesterday's problem you can take it as a simple homework problem where yesterday we said the estimation of c is nothing but estimating mean, mean of y .

Now, if I ask you to obtain maximum likelihood estimate of c , how would you? Remember this pdf here is the pdf of the joint pdf of the N observations, but that goes without saying. It is not of the single observation; whatever observations you have you are constructing the joint pdf. So, my question to is if I were to ask you to construct an MLE of c what would be step 1. What do I need? I need to construct the likelihood function. To construct the likelihood function what do I need? I need to know the pdf of the N observations.

How would you write the joint pdf of N observations there in yesterday's example? Now you have to assume that e_k is Gaussian white noise. Yesterday when we were solving

we did not have to assume Gaussianity, we just needed to assume white noise; but now I have to assume that e_k has some distribution and we will assume for simplicity Gaussian. If e_k is Gaussian white noise what would be the joint pdf of y ? Remember uncorrelated Gaussian white noise also independent, therefore the joint pdf of y is simply the product of the marginal pdf's. And we know that when vector of random variables are individual are also they are Gaussian.

So, all I need to do is write the joint pdf of y and that becomes my likelihood, then you take the logarithm of that work it out and find out what the solution is without referring to any resource. Even if you have referred I would likely to work it out, it is a good practice, it is a good exercise to go through because then that will make you comfortable with likelihood, because that is likelihood is also required to Fisher's information.

So, let us move on to now Fisher's information quickly.

(Refer Slide Time: 17:33)


Fisher's Information and Properties of Estimators References

Fisher information ... contd.

Fisher's information quantifies "how informative" a vector of observations is about a parameter θ (or θ). It rests on the following quantities (assume **single parameter**):

$l(\theta, \mathbf{y}) = f(\mathbf{y}; \theta)$ (or $f(\mathbf{y} \theta)$)	(likelihood function)	(3)
$L(\theta, \mathbf{y}) = \ln l(\theta, \mathbf{y})$	(log-likelihood function)	(4)
$S(\theta; \mathbf{y}) = \frac{\partial}{\partial \theta} \ln f(\mathbf{y}; \theta) = \frac{\partial}{\partial \theta} L(\theta, \mathbf{y})$	(score function)	(5)

where \mathbf{y} is the set of observations and θ is the parameter to be estimated.



Arun K. Tangirala Applied TSA October 19, 2016 NPTTEL 8

Fisher's information as I said rests on the likelihood concept. Why does it rest on this likelihood concept? What is Fisher's information all about? Very simple, now if you look back it is very simple. The Fisher's information is trying to quantify how much information the data has with respect to a parameter theta.

What do we mean by information? First of all, if I move in the theta space that is if I change theta the log likelihood function should also change. Why? Because then only we

say that the objective function is sensitive to theta. What is the objective function in MLE log? The log likelihood that is your objective function, in any optimization problem if you believe that there is something in that objective function to give you the optimal estimate what is the first step that you would do. You would actually take the derivative of that objective function set it to 0. If that derivative itself is 0 for any value of theta then there is no use in solving that optimization problem.

So, the first requirement is that the objective function should be sensitive to theta. If it is not sensitive to theta there is no question of finding an optimal solution. So, Fisher introduced therefore this concept called score which is nothing as you can see derivative of log likelihood with respect to theta. If there is nothing complicated about this, how would you find the optimal value of theta? You would set up the log likelihood and then differentiate with respect to theta and then set that to 0, the standard KKT conditions we will have to apply in optimization. Set the first derivative to 0, find the optimality and then check if it is indeed maximizing or minimizing standard stuff. So, the first step is to define the sensitivity function which is called the score in those days Fisher use to call this as a score.

Now, remember this likelihood function is conditioned on the data that you have. So, the derivative of the likelihood function with respect to this theta, depends on what? Depends on the data that is given to you; for a given value of theta it depends on the data if I change the data record the sensitivity is also going to change. And Fisher essentially argued and proved that the variability of this sensitive of this core across the outcome space of y that is all possible data records is what is Fisher's information; that is the idea.

(Refer Slide Time: 20:34)

Fisher's Information and Properties of Estimators References

Fisher information ... contd.

FI measures the variability in sensitivity of likelihood, i.e., the score function, across the outcome space (of y).

The **Fisher information** of a parameter θ in y is defined as

$$I(\theta) = \text{var}(S) = E \left(\left(\frac{\partial L}{\partial \theta} \right)^2 \right) \quad (6)$$

Under the regularity assumption, it can be shown that

$$\mu_S = E(S|\theta) = 0, \quad \text{var}(S|\theta) = E(S^2) = E \left(\left(\frac{\partial L(y, \theta)}{\partial \theta} \right)^2 \right) \quad (7)$$

Arun K. Tangirala Applied TSA October 19, 2016 NPTEL 10

So, first is likelihood, next is of course log likelihood, then you say that if there is information in the data with respect to theta it should be sensitive to changes in theta, otherwise so it is like this as I said you take a cricketer scores and then you look at student grades. Suppose I am trying to estimate the student grade I change the student grade in my guess will a cricketer scores change, there is no impact so you say there is no information about the student's grade in the cricketer score.

So, the first is sensitivity of the objective function with respect to theta, but this sensitivity depends on the data record. It is a function of y it is conditioned on y . Now I have to walk across the entire outcomes space of y and see what is the variability? That is I want to see how the sensitivity changes with realizations of data. If it does not change then there is a problem again, it should change.

And later on we will come across fundamental result in fact tomorrow we will talk about this fundamental result known as Cramer-Rao's inequality which tells us how the precision of the estimate theta is related to Fisher's information. In fact, we will see that the precision of theta hat is inversely proportional to the not precisions, sorry is directly proportional to the Fisher's information or you can say the variance of theta hat is inversely proportional to Fisher's information. More the information lower the variability in theta hat.

So, to summarize Fisher's information is nothing but variability of the score, variance of the score. And you have to understand this expectation is being evaluated in which space.

Student: (Refer Time: 22:29).

No, theta is not random.

Student: (Refer Time: 22:35).

Theta is fixed; in the white space across all possible outcomes you are calculating the average in not average, but the square essentially the variance how does the sensitivity change when I change the data record that is what is Fisher's information. So, we will conclude with a simple example, in fact you can show there is a very important assumption by the way and that is that assumption is central to the use of MLE which is that the pdf is regular.

What do we mean by regular pdf's? Regular pdf's are those they satisfy two conditions: one that and we will talk about that in MLE later on, but one of the main conditions of a pdf that is regular is it is parameters are not dependent on the range of values that the wise concrete. For example, if I take a uniform distribution what are the parameters a and b . A and b are themselves the range of values that why an why can take, so uniform distribution is not a regular one, whereas I take a Gaussian pdf the parameters are μ and σ^2 they do not determine the range of possible values, they only at measures but they are not they down. So, the Gaussian pdf is fortunately regular.

When that condition is satisfied you can show that the expectation, so remember here score is the first derivative of likelihood with respect to theta and information is expectation of the square by dou l by dou theta. And you can show that the expectation this variance can be calculated in a simpler way; simpler way meaning, you take the second derivative of the likelihood and then simply take the negative expectation of that.

(Refer Slide Time: 24:30)

Fisher's Information and Properties of Estimators References

Fisher information ... contd.

Since

$$E\left(\left(\frac{\partial L}{\partial \theta}\right)^2\right) = -E\left(\frac{\partial^2 L}{\partial \theta^2}\right) \quad (8)$$

the information can also be computed as

$$I(\theta) = -E\left(\frac{\partial^2 L}{\partial \theta^2}\right) = -E\left(\frac{\partial S}{\partial \theta}\right) \quad (9)$$

Arun K. Tangirala Applied TSA October 19, 2016 NPTEL 11

The actual definition is take the first derivative of the log likelihood square it up and then take the expectation with respect to y . But there is a simpler way out, you take the second derivative of \ln log likelihood with respect to θ and then take the negative expectation.

By the way what is the second derivative of log likelihood with respect to θ tell you? Whether the obtained optimal θ are maximum or minimum, the sign of it. So, essentially what Fisher's information is doing is it is looking at that and it is now looking at across all possible values of y . That is all that is another interpretation. So, let us look at a very simple example and then adjourn. We have done? Ok sorry. So, we will continue tomorrow.

Thank you.