

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

**Lecture – 83**  
**Lecture 37A - Goodness of Estimators 1 -2**

What we will do today is we look at actually some examples concerning Fisher's information and then move on to talk about bias variance and mean square error and so on.

Yesterday if you recall we had defined Fisher's information as the variance of the score. So, particularly we discuss the concept of likelihood, and the score is nothing but the derivative of the likelihood and Fisher's information looks at the variance of this score. So, you can remember this score as a sensitivity of the likelihood function with respect to the parameters that we are estimating. And what Fisher's information is actually looking at is how the sensitivity is changing with respect to the outcome.

(Refer Slide Time: 01:12)

Fisher's Information and Properties of Estimators

### 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  $\theta$  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 26, 2016

So, expression for the Fisher's information is the expectation of double phi to the whole square the expectation of that double theta. And we talked about regularity of pdf's under the regularity conditions you can show first of all that the average of this score across the outcome space is 0, and you can also show that this expectation that appears in

the Fisher's information definition can be rewritten as the negative expectation of the second derivative of the likelihood.

(Refer Slide Time: 01:44)

Fisher's Information and Properties of Estimators

### 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

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

Anu K. Tongolo Applied TSA October 28, 2018

And we did mention that the second derivative of their likelihood is actually measure of whether you have reached the maximum or the minimum. So, you can think of this Fisher's information also as what is it average across the outcome space of y, whether there is a peak that whether there is a maximum or not in the likelihood function.

(Refer Slide Time: 02:06)

Fisher's Information and Properties of Estimators

### Example 1: Information about mean and variance

Consider the case of estimating mean  $\mu$  and variance  $\sigma^2$  of a random signal.

**Mean and variance**

Given that a stationary signal  $y[k] \sim \mathcal{N}(\mu, \sigma^2)$ , determine (i)  $I(\mu)$  and (ii)  $I(\sigma^2)$  in a single observation.

1. The log-likelihood function (assuming  $\sigma^2$  is known) is

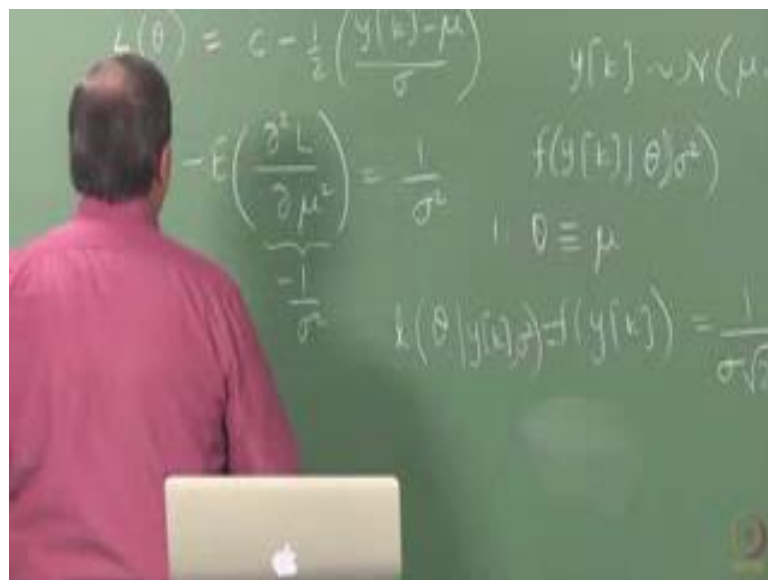
$$L(\mu; Y) = \ln f(y|\mu) = -\frac{1}{2} \ln(2\pi\sigma^2) - \frac{1}{2} \frac{(y - \mu)^2}{\sigma^2} \quad (10)$$

Anu K. Tongolo Applied TSA October 28, 2018

So, best understood with an example; let us look at a very very basic example here. I have a stationary signal as usual, it is more or less the same example that we have looked at before in optimization. And here we are interested in estimating the mean. And we assumed that sigma square is known.

So, the first question that we are going to ask is, if I pick an observation at random the goal here is to estimate the mean one of the crudest estimates of mean is just the observation itself. Suppose, I pick an observation and I want to ask how much information does the signal observation have with respect to mu assuming sigma square is known. So, the procedure always in Fisher's information is to construct the likelihood function, and from where you construct the log likelihood function.

(Refer Slide Time: 03:02)



In this case since you are given that  $y_k$  falls out of Gaussian distribution with mean  $\mu$  and variance  $\sigma^2$  we know straight away that; so what are we given here? We had given only one observation unlike the usual case where I am given  $n$  observations. So, the first step is to construct the likelihood of the given observations. Since I am given only one observation and that likelihood is nothing but the pdf.  $\theta$  for us here in the first case;  $\theta$  is just the  $\mu$ .

So, you can say your given  $\theta$  and  $\sigma^2$ . When  $\mu$  and  $\sigma^2$  are unknown then  $\theta$  becomes both  $\mu$  and  $\sigma^2$ , we will look at that case as well shortly. So,  $\theta$  is the mean, now since we had given that  $y_k$  is a Gaussian falls out of a

Gaussian distribution we know from simple probability theory that  $f(y_k)$ ; I am avoiding the  $\theta$  here is  $\frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(y_k - \mu)^2\right)$ .

So, this is your pdf which is nothing but the likelihood itself. Since we work with log likelihood you take the logarithm of this. Once you take the logarithm the exponential vanishes, and what you have this is your likelihood but when you write likelihood it is better to explicitly stated as a function of  $\theta$  given  $y_k$  and  $\sigma^2$ . That is your likelihood; write it at the top here for the log likelihood.

So, the log likelihood of  $\theta$  I am just going to write of  $\theta$  just for the sake of convenience I am avoiding given  $y_k$  and  $\sigma^2$  it should be understood; you would obtain logarithm of this factor since  $\sigma^2$  is given is just some constant does not change with  $\mu$ . So,  $c - \frac{1}{2}$  you have  $(y_k - \mu)^2 / \sigma^2$ . This is your log likelihood; that is it.

So, now all one has to do once you have constructed the log likelihood that is usually the most demanding step, it is not easy in many situations to come up with this likelihood function even though you are given the pdf. So now that you have you have the log likelihood the next step is to differentiate twice because we know that the Fisher's information can be computed as expectation average of the second derivative of the likelihood.

So, what is the second derivative of the likelihood, a log likelihood here; when I say likelihood more or less you should understands log likelihood. Would it be minus 1 by  $\sigma^2$ ? Right, is it true? Sure?

Student: (Refer Time: 06:27).

Why? Why it is a negative sign coming? Good; anyway, so you have not answered where is negative sign comes in it comes in because you have a  $(y_k - \mu)^2$ . There is a  $y_k$  that is missing in this equation it reads as  $y$  that is the small correction you should note down. Since it is a single observation I have just written this  $y$ . So, the second derivative of the likelihood is minus 1 over  $\sigma^2$  and what we are interested in is the negative expectation of the second derivative. Since, this itself is minus 1 over  $\sigma^2$

square the Fisher's information is simply  $1/\sigma^2$  and that is the answer for you.

So, what does it tell us? You just arriving at the answer is not enough, what it tells us is the information contained in a single observation about  $\mu$  is inversely proportional to the variance of the process itself; which means more the randomness in the signal lesser is the information contained in  $y_k$ . It does not tell us how good the estimate is going to be remember that, Fisher's information at the moment it does not tell us how good the estimate is or how you should estimate there is nothing like that; there is no reference to how you are going to estimate it.

All it is saying is this is the information contained in the single observation. Now it is the problem of still estimating mean remains. Later on we will learn very fundamental result known as a Cramer Rao's inequality which will tell us regardless how you estimate the precision or you can say the variance of the estimate is very closely related to this information, in fact the bound the least error or the least variability that you can achieve is the inverse of this information.

What this tells us is; this example tells us is as  $\sigma^2$  increases the information decreases, and what I just now said is the variance of any estimate of mean for this process it does not matter whatever method you use that estimate will cannot have a variance lower than  $\sigma^2$ . That is the Cramer Rao's inequality result. We will visit when we talk about it formally we will revisit this example. But what is important for you to take from this example is that a single observation has information of  $1/\sigma^2$  and you cannot do better than this.

Now suppose I am interested in estimating  $\sigma^2$ , let us say I am given mean so we are now turning our attention to the variance. It does not make sense to look at a single observation and try to estimate  $\sigma^2$ , because notionally if you look at estimate of  $\sigma^2$  we would expect to have at least two observations.

(Refer Slide Time: 09:41)

Fisher's Information and Properties of Estimators

**Example 1** **... contd.**

2. Now,  $\theta = \sigma^2$ . The information contained in a single observation is

$$I(\sigma^2) = -E\left(\frac{\partial^2 L}{\partial \theta^2}\right) = -E\left(\frac{1}{2\sigma^4} - \frac{(y - \mu)^2}{\sigma^6}\right) = \frac{1}{2\sigma^4} \quad (12)$$

Arav K. Tongolo Applied TSA October 28, 2018

But let us say- by some wild method I am going to use a single observation to estimate sigma square. Again here there is no reference to how I am going to estimate it, so do not confuse the estimation method with Fisher's information. Remember we are not talking of the estimator here; we are only talking about the data.

So, the information contained in a single observation with respect to sigma square; again is obtained in a similar way; the likelihood function does not change, what changes here? Theta changes that is all. So, now, your given mu is known and sigma square changes, so instead of evaluating the second derivative with respect to mu.

(Refer Slide Time: 10:23)



Now you would estimate, you would calculate the second derivative with respect to sigma square. So, you are going to actually; in fact there is again a slight mistake there we should read as negative expectation of sigma square 1 by dou sigma square square, because sigma square is a parameter with which you are differentiating. And when you do that this is standard calculus it turns out that the information contained is 1 over 2 sigma to the power of 4.

Now the information here is even lower then what you can expect to see with respect to mu. And that make sense because, if I say a single observation gives me an estimate of mean somewhat fair enough but if I say now the single observation is going to give me an estimate of sigma square then it is going to be even worse. Therefore, the information is even lower. And also remember that the information contained with respect to sigma square about sigma square is different in general from the information contained about sigma.

Suppose I were to ask you what is the information contained in the single observation about sigma, what would be your answer; instead of sigma square. So, you want to know how much information is contained in the single observation about a standard deviation instead of variance. What would you do? The likelihood function does not change, but now the parameter is sigma. So, you would actually evaluate this like I have shown you

the expression that I have given you on the screen, bit confusing. So, what would be the answer, can we work it out? What is the answer that you get?

Student: (Refer Time: 12:44).

Sorry.

Student: 4 by (Refer Time: 12:47).

4 by?

Student: Sigma square.

Sigma square; so we will write down the answers here.

(Refer Slide Time: 12:54)

The image shows a green chalkboard with the following handwritten content:

$$y[k] \sim \mathcal{N}(\mu, \sigma^2)$$
$$f(y[k] | \theta) \propto \frac{1}{\sigma^2} \exp\left(-\frac{1}{2\sigma^2}(y[k] - \mu)^2\right)$$
$$\theta = \mu$$
$$f(y[k]) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{y[k] - \mu}{\sigma}\right)^2\right)$$
$$I(\mu) = \frac{1}{\sigma^2}$$
$$I(\sigma) = \frac{1}{2\sigma^4}$$
$$I(\sigma) = \frac{4}{\sigma^2}$$

The first thing that we observed is information contained about mean is 1 over sigma square, information contained for sigma square we have just derived is 1 over 2 sigma to the power of 4, and information contained about sigma is 4 over sigma square; is it correct? Does anyone else get this answer 4 over sigma square? 2 over sigma square.

Student: 2.

Not 4 over sigma square, now you need a tie breaker; sorry.

Student: 3.



3, that seems to be the average of the answers.

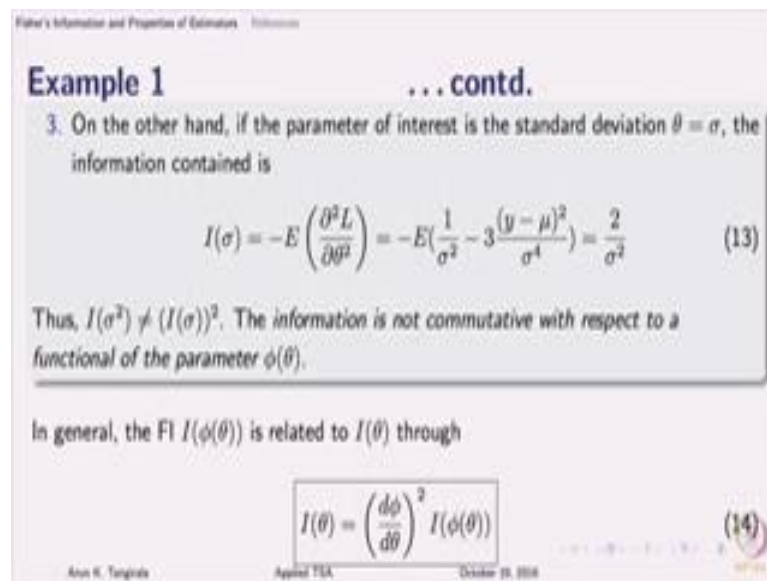
Student: (Refer Time: 13:35).

3 is impossible.

Student: 2 over sigma square.

2 over sigma square; now it turns out there is a way of checking this. You can actually check if this answer is correct or not without having to calculate this. If you know this you can actually this, or if you no information contained about sigma you can actually calculate information contained about sigma square. And that is got to do with this relation.

(Refer Slide Time: 14:19)



Fisher's Information and Properties of Estimators

**Example 1** ... contd.

3. On the other hand, if the parameter of interest is the standard deviation  $\theta = \sigma$ , the information contained is

$$I(\sigma) = -E\left(\frac{\partial^2 L}{\partial \theta^2}\right) = -E\left(\frac{1}{\sigma^2} - 3\frac{(y - \mu)^2}{\sigma^4}\right) = \frac{2}{\sigma^2} \quad (13)$$

Thus,  $I(\sigma^2) \neq (I(\sigma))^2$ . The information is not commutative with respect to a functional of the parameter  $\phi(\theta)$ .

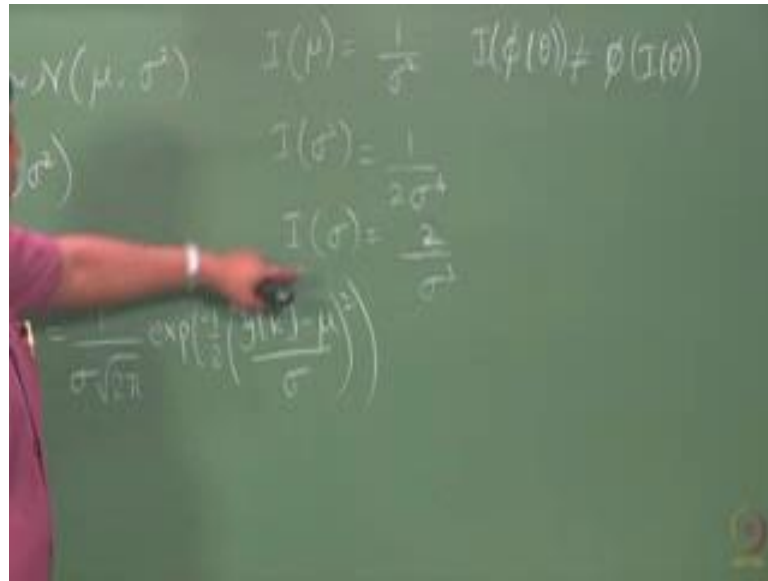
In general, the FI  $I(\phi(\theta))$  is related to  $I(\theta)$  through

$$I(\theta) = \left(\frac{d\phi}{d\theta}\right)^2 I(\phi(\theta)) \quad (14)$$

Anu K. Torgata Applied T&A October 25, 2018

In general if you have information about theta calculated, the information contained about some function of theta let say some phi of theta or whatever that function is; it is related to I of theta through this relation given in equation 14. In fact, I have calculated here, but if you were to cross check let us say you do not want to recalculate your I of sigma all fresh you can use the relation in 14 by either knowing. So, you can use I of sigma square to calculate I of sigma or vice versa and you can cross check.

(Refer Slide Time: 15:00)



So, what this says is the information contained in theta is not simply; so let me put it this way here information contained, let us say in phi of theta some function is not equal to phi of information contained in theta.

So, here suppose theta is sigma then we know already now information contained in a single observation with respect to sigma is 2 over sigma square. Now what this relation; suppose I were to blindly square this, I have to imagine that the information contained the sigma square in simply just square of the information contained in sigma, then obviously that is not correct because the square of this is 4 over sigma to the power of 4, whereas the actual answer is 1 over 2 sigma 2 the power of 4. And you can check using the relation in 14 that this is satisfied for the results that we have derived. And this is in general true it is not a coincidence, you can prove formally that the information contained between theta and phi of theta related in this way.

So, what this tells us also in general is in estimation; suppose I want to estimate alpha or I want to estimate standard deviation let us talk about standard deviation, in general it may not be true that I the optimal estimate of sigma is the square root of the optimal estimate of sigma square. So, that is something that one has to keep in mind. In generally it may not be true; it turns out for MLE it is true, but for a general estimator suppose I want to estimate sigma square I cannot estimate sigma optimally and necessarily claimed

that the square of that optimal estimate is the optimal estimate of sigma square, because the information content is different.

So, this is called lack of invariance property in general but MLE alone has that; that means if you were to estimate the parameters using maximum likelihood approach you can then guaranteed that whatever MLE estimate that you have for sigma through the maximum likelihood approach you can simply square that and you can obtain, it would be the same as the maximum likelihood estimate of sigma square, but in general it is not true.

Therefore, when you want to estimate parameters formulate your estimation problem directly in terms of that parameter. Do not try to do actually do a post processing later on and claim optimality; that is something to keep in mind.