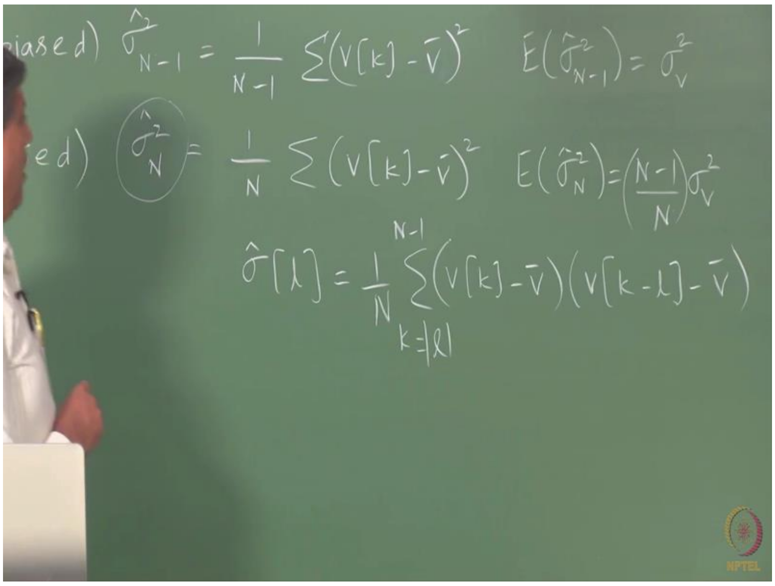
Applied Time-Series Analysis
Prof. Arun. K. Tangirala
Department of Chemical Engineering
Indian Institute of Technology, Madras

Lecture – 86
Lecture 38A - Goodness of Estimators 2 -1A

So let us continue with the discussion that have been having on estimators; until the last class we talked about Fishers information the bias and variance and so on. If you consider any estimation method, there are this important properties that we have talked about bias, variance, efficiency, consistency and so on.

We have given formal definition for bias and variance, while bias is concerned with the accuracy of the estimator, variance is concerned with the precision that is how much the estimate is going to vary across experiments and one of the things that we should remember is as much as unbiasedness is desirable, it is ok to have a biased estimator. But what is more serious is the precision; if there are 2 estimators one being biased one and having a lower variability that is higher precision. And another estimator which is unbiased, but having a larger accuracy larger variability then one would prefer perhaps the more precise estimator at the cost of bias; that does not mean that always one has to sacrifice bias for getting lower precision, but if it has to be sacrificed then it is ok.

(Refer Slide Time: 01:48)



iased) $\hat{\sigma}^2_{N-1} = \frac{1}{N-1}\sum(v[k]-\bar{v})^2$ $E(\hat{\sigma}^2_{N-1}) = \sigma^2_v$
ed) $\hat{\sigma}^2_N = \frac{1}{N}\sum(v[k]-\bar{v})^2$ $E(\hat{\sigma}^2_N) = \left(\frac{N-1}{N}\right)\sigma^2_v$
$\hat{\sigma}[l] = \frac{1}{N}\sum_{k=|l|}^{N-1}(v[k]-\bar{v})(v[k-l]-\bar{v})$

So, as a simple example suppose you take the estimator of variance, we know that there are two different forms of estimator; we have discussed this. The estimator of variance is given by unbiased estimator has a 1 over n minus 1, v k minus v bar square the sum of that whereas, the biased one which is given out by MLE has this expression to it; of these two this is the unbiased one, whereas, obviously if this is unbiased estimator of the variance then this is the biased one; what we mean by unbiased is the expected value of sigma square hat n minus 1 is sigma square n; sigma square v itself. Whereas, here the expected value of sigma square hat n is going to be n minus 1 by n, sigma square v which means; obviously there is a bias in the second estimator which is given out by the MLE.

Now, what is happening here is we have lost some accuracy as we move from this estimator to this estimator. However, you can show that the second estimator has lower variability than the first one and this fundamental principle should be remembered always in estimation theory, whether you are estimating parameters of a pdf or parameters of a model; in any estimation exercise there is always a compromise between bias and variance always and we will keep talking about that in fact, if you recall we talked about AIC; Akaike Information Criterion, a while ago when we were going through a case study on building ARMA models from data, there we talked about AIC and AIC measures the tradeoff between bias and variance, but the bias that we are talking about in AIC is the bias in a prediction and variance that we are talking about an AIC is variability in parameter estimates.

So when you have a model, very simple model let us say a first order AR or something very simple model does not have to be in the context of random processes; any for any process if you build a simple model it is likely that the simple model may do a poor job of a prediction, but has very few parameters; maybe one or two parameters and then in a bit to improve the prediction, we start increase in the complexity of the model.

So, we may for example, move from AR 1 to AR 4 which has more parameters to estimate. Clearly more the parameters that you have in the model better is your ability to predict because you have more parameter power, you can say more man power to do the job for you. But then remember the information content in the data is fixed, you are using the same data; you are not going to use any other data set and this has to be the information content in the data is like the food and that has to be distributed among the parameters. When you have more parameters in the model; I am giving you qualitative arguments here, later on when we move to least squares you received quantitative expressions as well.

When you have more parameters in the model, you will see that the variability in the parameters estimates is going to shoot up, but what have you achieved, you have actually reduced the bias in the prediction; you have gotten the predictions of the model closer to the absorbed values. So, that is a standard trade of that you would have and now you can extend this argument further as I have more and more complicated models, lower is going to be the bias in the prediction, but higher is going to be the variability in parameter estimates. So, in any estimation exercise there is going to be a trade of between bias and variance.

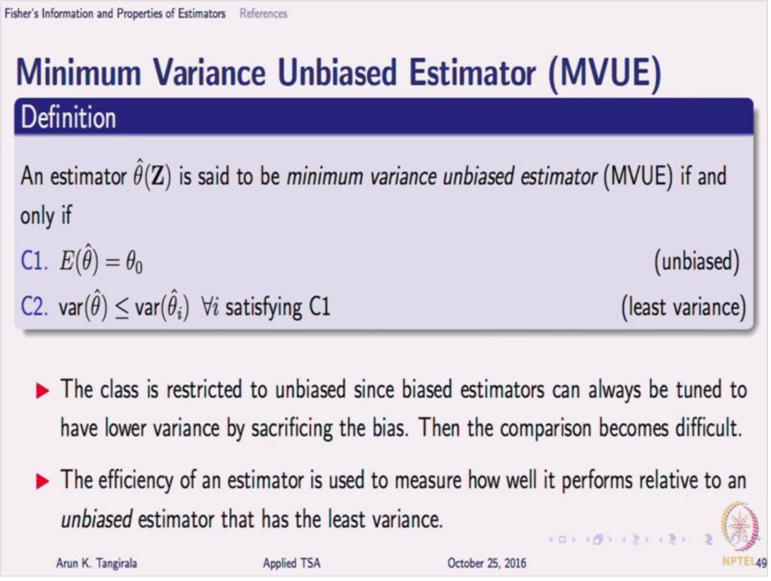
So, in this example normally one prefers to work with this for large samples, when you have large observations you say I will just use this. When you have small observations may be you want to work with this, another example that we will come across very soon is the estimator of auto co-variance function; we have used that; the sample auto covariance function as a factor of 1 over n; if you recall. Let me write that expression for you; if you have forgotten, sigma hat of l is 1 over N sigma v k minus v bar and v k minus l minus v bar and here k runs from l to n minus 1 you can say mod l if you are looking at positive.

So, in this case obviously, you have n minus l terms at any lag l there are n minus l terms in the summation and intuitively if you want an unbiased estimate of the auto covariance function, you should have had a 1 over n minus mod l as a factor here so that you get an unbiased estimate of the auto covariance function. But we still work with 1 over n, very well knowing that this 1 over n is going to give me a bias estimate of the auto covariance function.

Why do we do that? One of the reasons is being that it has lower variance then the unbiased one, there is another reason that we will learn later on, but one of the reasons is that the variance is lower. These does not mean as I said earlier that if you have to get more précised estimates, you have to sacrifice the bias; it does not mean that, but if you have to do it then it is ok; that is what it means. Now let us talk about again come back to the world of unbiased estimators, what I have essentially tried to convey is precision is a far more serious property than accuracy; desirable property than accuracy.

There is nothing like having both accurate précised estimator and that is what we are seeking here; minimum variance unbiased estimator, but can happen that for a process you may not be able to find a minimum variance unbiased estimator, which means you may not be able to find an accurate; both an accurate and précised estimator then you may have to sacrifice one of the things and typically we would like to sacrifice accuracy for precision that is the message.

(Refer Slide Time: 08:58)



Fisher's Information and Properties of Estimators   References

Minimum Variance Unbiased Estimator (MVUE)

Definition

An estimator $\hat{\theta}(\mathbf{Z})$ is said to be *minimum variance unbiased estimator* (MVUE) if and only if

C1. $E(\hat{\theta}) = \theta_0$ (unbiased)

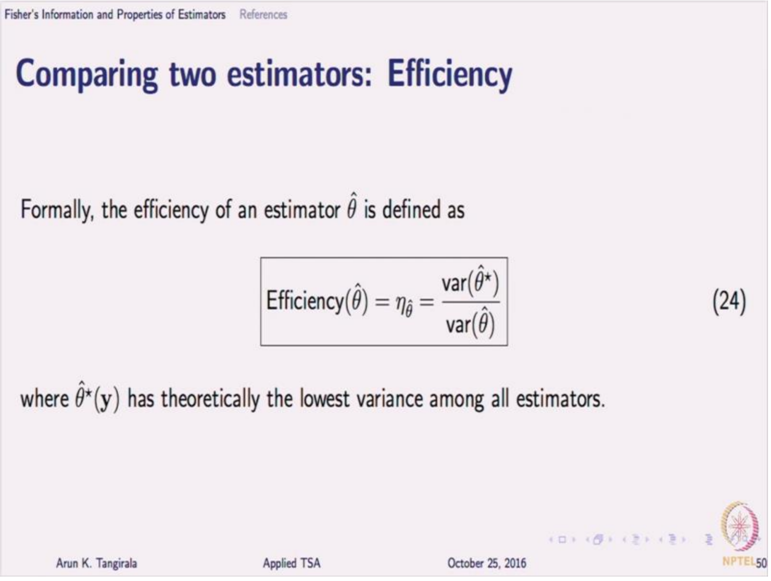
C2. $\text{var}(\hat{\theta}) \leq \text{var}(\hat{\theta}_i) \ \forall i$ satisfying C1 (least variance)

- The class is restricted to unbiased since biased estimators can always be tuned to have lower variance by sacrificing the bias. Then the comparison becomes difficult.
- The efficiency of an estimator is used to measure how well it performs relative to an *unbiased* estimator that has the least variance.

Arun K. Tangirala   Applied TSA   October 25, 2016   NPTEL 49

So, what is the definition of MVUE or minimum variance unbiased estimator, well it is straight forward; first of all it should be unbiased and among this world of class of unbiased estimators, it will have the least variability. So, we are being fair we are not comparing this estimator with the biased one; if I start comparing the variance of an unbiased estimator with a biased one then the biased one will beat it and there is no limit, I can always sacrifice the bias highly to get more précised estimates. So, it is not fair to compare the variance of an unbiased estimator with that of a biased one, therefore we restrict ourselves the class of biased estimators. Now the notion of efficiency has actually stems from this concept of minimum variance estimators; these are all extremely fundamental concepts to estimation theory there is no escape to this.

(Refer Slide Time: 09:49)



Fisher's Information and Properties of Estimators References

Comparing two estimators: Efficiency

Formally, the efficiency of an estimator $\hat{\theta}$ is defined as

$$\boxed{\text{Efficiency}(\hat{\theta}) = \eta_{\hat{\theta}} = \frac{\text{var}(\hat{\theta}^{\star})}{\text{var}(\hat{\theta})}} \qquad (24)$$

where $\hat{\theta}^{\star}(\mathbf{y})$ has theoretically the lowest variance among all estimators.

Arun K. Tangirala Applied TSA October 25, 2016 50

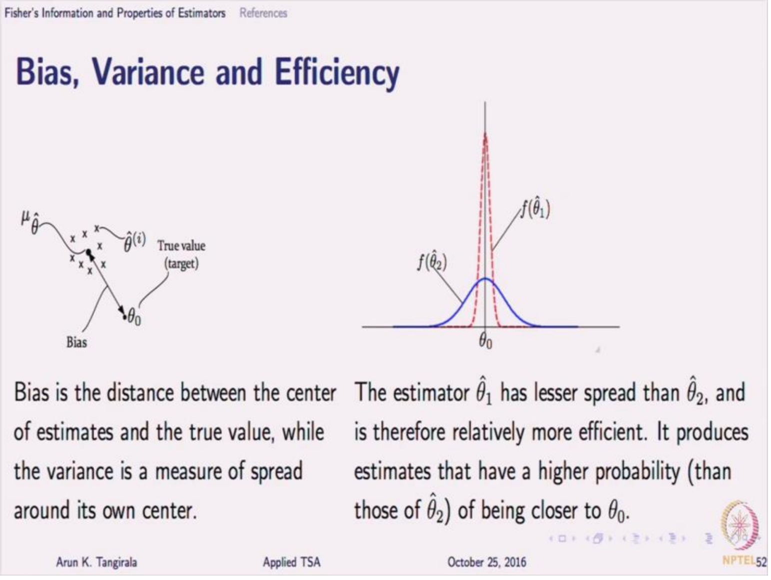
Efficiency is very straight forward, we have talked about this earlier efficiency is the ratio of the variability of the best estimator that is the minimum variance estimator to the variance of that estimator under study. So the theta hat star is the minimum variance unbiased estimator, it goes without saying I do not have to keep saying unbiased when I say minimum variance unless otherwise stated you should assume it unbiased also. So, theta hat star is a minimum variance estimator whereas, theta hat is a estimator under study.

So, the efficiency is the ratio of the variability of the theta hat star to the variance of theta hat because theta hat star is a the minimum variance estimator, what does it mean? The variance of theta hat star; the numerator is always going to be less than maximum or to the equal to the denominator, if theta hat is the minimum variance itself then they are going to be equals.

So, the maximum efficiency that an estimator can achieve is 100 percent and this is called efficiency for reasons I had explained earlier. What you are pumping in to the estimator is data fine that is what you are feeding in; outcomes a estimate, but behind seen what your feeding in is the uncertainty in the data and outcomes the uncertainty in the estimate. So, efficiency of an estimator is a measure of the ability of the estimator to reduce or shrink that uncertainty to produce a more certain number; that is how you can interpret efficiency as and there is also relative efficiency suppose I have two estimators;

two methods of estimating then which is more efficient then they you just divide the ratio of the variances accordingly.

(Refer Slide Time: 11:42)



Fisher's Information and Properties of Estimators   References

Bias, Variance and Efficiency

Bias is the distance between the center of estimates and the true value, while the variance is a measure of spread around its own center.

The estimator $\hat{\theta}_1$ has lesser spread than $\hat{\theta}_2$, and is therefore relatively more efficient. It produces estimates that have a higher probability (than those of $\hat{\theta}_2$) of being closer to θ_0.

Arun K. Tangirala   Applied TSA   October 25, 2016   52

Now, this is a schematic which kind of gives you an idea of; what is bias, what is variance and efficiency. So if you look at the left hand side plot here, what I have here is just some estimates indicated by this crosier marks and then there is that is indicated by theta hat of i; corresponding to the ith realization of data that you have; theta naught is the true value and then you have mu theta hat which is a average; it is a average of the estimates that you have across all the realizations.

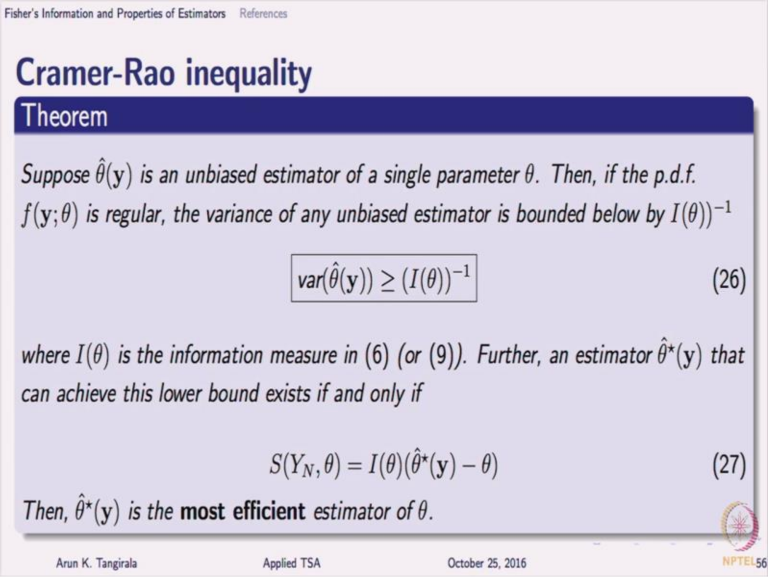
Bias is a distance between the average of the estimates and the truth; if the estimate is unbiased they will coincide. On the right hand side you have two different estimators theta 1 hat and theta 2 hat and I am just showing you the pdf of it. That is the distribution of the estimates obtained from this two different estimators across the realization space and both are unbiased as you can see the both are centered around theta naught, but the one in red has lower variability, lower spread then the one in blue which has a larger spread. So, we say theta 1 hat is more efficient than theta 2 hat.

So, the other way of imagining this or interpreting this is that theta 1 hat produces estimates that have a higher probability of being closer to theta naught then theta 2 does; that is another way of looking at it, but you can say essentially theta 1 hat is more precised than theta 2 hat. So, hopefully this schematic kind of gives you a better picture of this terminology.

Obviously now the hunt is for the most efficient estimator; I would like the most precised estimator and of course, an unbiased one and this was a question that was asked long ago by statisticians and the answer came out in the form of the celebrated Cramer-Rao inequality, which as I will show you is very closely related to Fishers information, the bound itself is the inverse of the Fishers information. So, what we want to ask is for a given estimation problem, what is the minimum variance that I should expect to see; that is question number 1 and the second question; question number 2 is, what is that estimator which will give me that minimum variance?

Now, as I said earlier, you may be able to find the bound on the minimum variance, but you may not be able to necessarily find device an estimator that will get you that bound; in that case the minimum variance only remains an imagination and ideal one that you cannot realize physically. So, there are many such idealities in many different fields, so that MVUE also will become that you know that imagination, something that you wanted to achieve in life and it will remain as an imagination because you could not realize the dream. It is also possible that in estimation you will have the same story. So, what is the Cramer-Rao's inequality? It essentially says if there is an unbiased estimator and of some single parameter right now we will just focus on a scalar case.

(Refer Slide Time: 15:20)



Fisher's Information and Properties of Estimators References

Cramer-Rao inequality

Theorem

Suppose $\hat{\theta}(\mathbf{y})$ is an unbiased estimator of a single parameter θ. Then, if the p.d.f. $f(\mathbf{y};\theta)$ is regular, the variance of any unbiased estimator is bounded below by $I(\theta))^{-1}$

$$\boxed{var(\hat{\theta}(\mathbf{y})) \geq (I(\theta))^{-1}} \qquad (26)$$

where $I(\theta)$ is the information measure in (6) (or (9)). Further, an estimator $\hat{\theta}^{\star}(\mathbf{y})$ that can achieve this lower bound exists if and only if

$$S(Y_N,\theta) = I(\theta)(\hat{\theta}^{\star}(\mathbf{y}) - \theta) \qquad (27)$$

Then, $\hat{\theta}^{\star}(\mathbf{y})$ is the **most efficient** *estimator of θ.*

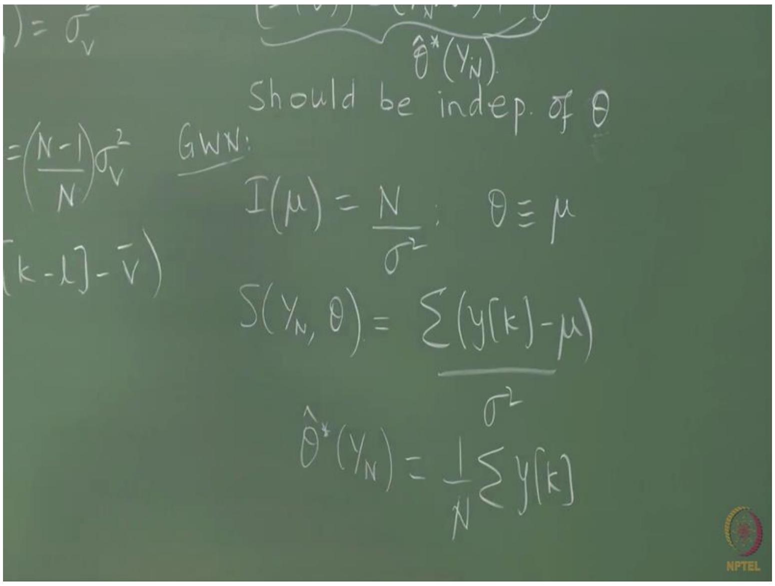
Arun K. Tangirala Applied TSA October 25, 2016 NPTEL 56

Suppose I have an unbiased estimator of a single parameter theta and if the pdf is regular, so there is a condition pdf of what? Of the data is regular then the variance of any unbiased estimator is bounded below by the inverse of the Fishers information, which means the minimum variance that is achievable for any estimator is the inverse of the Fishers information which now helps us appreciate Fishers information lot more; what this result says is as the information content in the data about this parameter theta grows; the minimum variance that you can achieve goes down which is good; that means, you can dream of getting more and more precised estimates which is good.

So, this quantifies the expectations that people may have on how precise an estimator I can construct for a theta; that is the first part of the Cramer-Rao's inequality. The second part talks about its existence whether an estimator exist that will achieve the lower bound and it exist only if you can express this relation that is given in equation 27 here; which relates the score which is nothing, but the derivative of the likely hood to the Fishers information and the theta star minus theta.

So, what is theta here the parameter that you have estimating theta hat star is the efficient estimator that you are searching and Fishers information we already know, the score we already know.

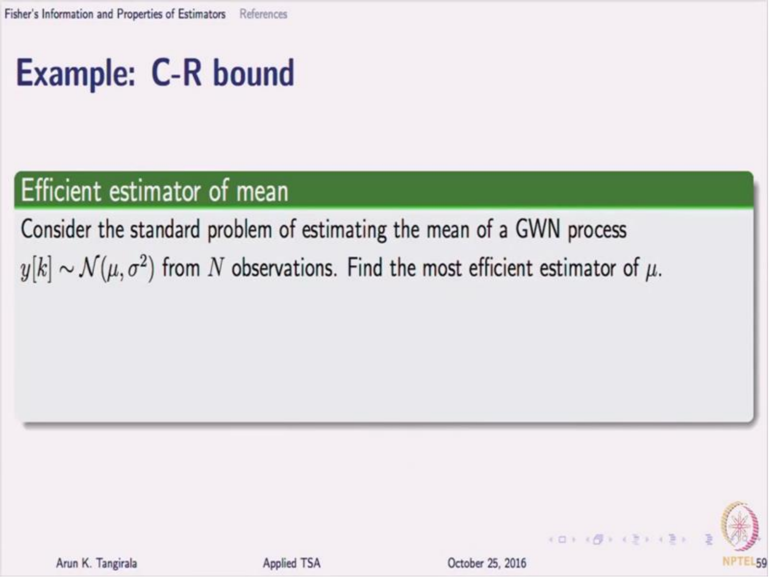
(Refer Slide Time: 17:18)


$) = \sigma_v^2$
$\hat{\theta}^*(Y_N)$
Should be indep. of θ
$= \left(\frac{N-1}{N}\right)\sigma_v^2$
GWN:
$I(\mu) = \frac{N}{\sigma^2}$; $\theta \equiv \mu$
$[k-l] - \bar{v})$
$S(Y_N, \theta) = \frac{\sum(y[k]-\mu)}{\sigma^2}$
$\hat{\theta}^*(Y_N) = \frac{1}{N}\sum y[k]$

In other words, what it says is if I am given the data and likelihood and so on then essentially what you have from these relations. So, let us write for theta hat star essentially I of theta inverse; times score plus your theta should be independent of y, should be independent of theta in fact, it is a scalar I am going to remove that sorry about that.

So, what it says is that if you were to rewrite this expression in a different way which is what I have done then the right hand side should get me theta hat star which is purely a function of y. I should not need anything else to construct the efficient estimator, it can turn out that in many cases that this expression can in turn be a function of true parameter, in which case you cannot use it. This is nothing but according to the Cramer-Rao inequality this is nothing, but your theta hat star. What does it say theta hat star is only a function of the observations of n of nothing else, so let us go through an example and will appreciate it much better.

(Refer Slide Time: 18:52)



Fisher's Information and Properties of Estimators References

Example: C-R bound

Efficient estimator of mean

Consider the standard problem of estimating the mean of a GWN process $y[k] \sim \mathcal{N}(\mu, \sigma^2)$ from N observations. Find the most efficient estimator of μ.

Arun K. Tangirala Applied TSA October 25, 2016 59

So, let me actually go through this example here where we are again returning to the standard problem of estimating mean. Earlier we use this example to illustrate the concept of estimation, how the objective function can change the nature of the estimate then we use this example to compute Fishers information. Now we are using the same example to illustrate the Cramer-Rao inequality, so in this example what we want to know is given n observations of a Gaussian white noise process, what is the most efficient estimator of mean that is it. We are not asking for linear estimators, non-linear; we are not imposing any form the only requirement that we are asking is of course, unbiased as well it is understood because Cramer-Rao's inequality focuses on unbiased estimator, we are asking what is the most efficient estimator of mean.

Now we can easily work this out; for n observations what was the Fishers information that we had for mean? What was I of mu?

Student: (Refer Time: 20:08).

Sorry.

Student: (Refer Time: 20:11).

N by sigma square very good; n by sigma square and what is the score function that we had the derivative of the log likelihood, you recall sorry just 1 by sigma square no way. So, for the Gaussian white noise process summation.

Student: (Refer Time: 20:46).

Sorry.

Student: (Refer Time: 20:48).

Very good y k minus mu divided by sigma square good; correct. Now can you put together these two pieces of information and of course, theta of interest is mu. So, put all of this in to this expression and get me theta hat star and tell me then if it is independent of the parameter that you are estimate; what you get?

Student: (Refer Time: 21:32).

Good, let us see if there is any other person who is able to see that it is very simple algebra nothing much to do, does the anybody else get the sample mean. So, you should get the sample mean as the answer and is it independent of the parameter that I am estimating, yes which means I have hit the jackpot. First of all what I know from Fisher’s information the inverse of this is the least variance that I should expect for estimating mean among all unbiased estimator; it does not matter whom you ask, what is it? As long as it is unbiased no estimator can achieve a lower bound than this right which is inverse of I, which is sigma square by n and we already proved that the variance of sample mean is sigma square by n from that result itself we should get that sample mean achieve this bound.

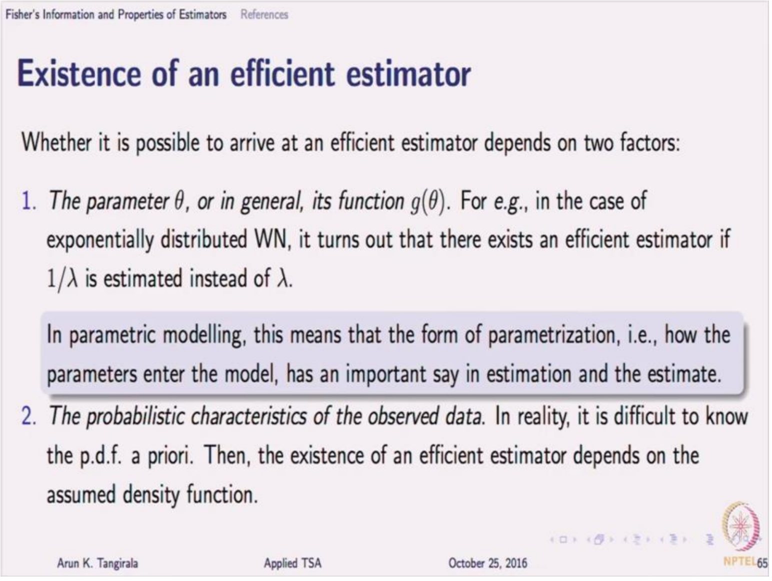
Now, we have also shown that indeed that efficient estimator is a sample mean, look at how beautiful Cramer-Rao inequality is we did not have to solve any optimization problem, nothing we just asked for the most efficient estimator and it came out and said use the sample mean which is what we use daily. But of course, it tells you a lot of things it says sample mean is unbiased of course, is the most efficient estimator only at least when from this result we know that this is true only for Gaussian white noise process.

You should ask yourself if you were to change the pdf instead of Gaussian white noise suppose the pdf, suppose the distribution was laplacian or pause on some other distribution; Kai-square, would you expect to see sample mean coming out as the most efficient estimator, would you expect to see the inverse of this as the bound? Intuitively no, you would see something else, but that is the beauty. This entire result not only tells you what is the most efficient estimator, but also tells you that sample mean which is something that you use routinely is most efficient only for at least from this example for Gaussian white noise process; so that is the beauty of this Cramer-Rao's inequality. Now, the other thing that of course, we see is that it is a linear estimator which is also good news, so that is it. So, we have I am just going to go pass this which we have discussed.

Now let us just briefly talk about existence of an efficient estimator; what does it depend on of course, it depends on the parameter that you are estimating. The first factor that is going to affect the existence of an efficient estimator is the parameter; to be more precise how the parameter enters the model? Where your parameter is sitting in the model is it hiding somewhere behind a corner or is it making itself fairly obvious so that you can estimate it very well; you can think of it this way. And that is example that I have given and in parametric modeling this has got to do with how you parameterize your model alright whether you are actually parameterizing in such a way that you get an auto regressive model or a moving average model or some other model and so on that will determine your ability to find the most efficient estimator.

The second factor of course, as you can see and we have discussed that is going to affect is the pdf itself; whether the pdf is regular or not in the statement itself it says if the pdf is regular or not. If the pdf is not regular then the Cramer-Rao inequality itself does not apply.

(Refer Slide Time: 25:47)



Fisher's Information and Properties of Estimators References

Existence of an efficient estimator

Whether it is possible to arrive at an efficient estimator depends on two factors:

1. *The parameter θ, or in general, its function $g(\theta)$.* For *e.g.*, in the case of exponentially distributed WN, it turns out that there exists an efficient estimator if $1/\lambda$ is estimated instead of λ.

   In parametric modelling, this means that the form of parametrization, i.e., how the parameters enter the model, has an important say in estimation and the estimate.

2. *The probabilistic characteristics of the observed data.* In reality, it is difficult to know the p.d.f. a priori. Then, the existence of an efficient estimator depends on the assumed density function.

Arun K. Tangirala Applied TSA October 25, 2016 65

So, that kind of you know concludes our discussion on Cramer-Rao's inequality and Cramer-Rao's bound, but it has far far reaching implication Cramer-Rao's bound is used as a gold standard for determining the most efficient estimator and tomorrow if you come up with an estimation method, you will also have to show for that parameter that you are estimate; any problem that you have take, your method may give you some variance. If it is unbiased, it cannot give in its variance lower than what the Cramer-Rao's inequality is telling you. So, that is a test that you have to do to show the whether you have achieved that efficiency or how far you are away from the efficiency. Any questions on this before we move on to mean square error.