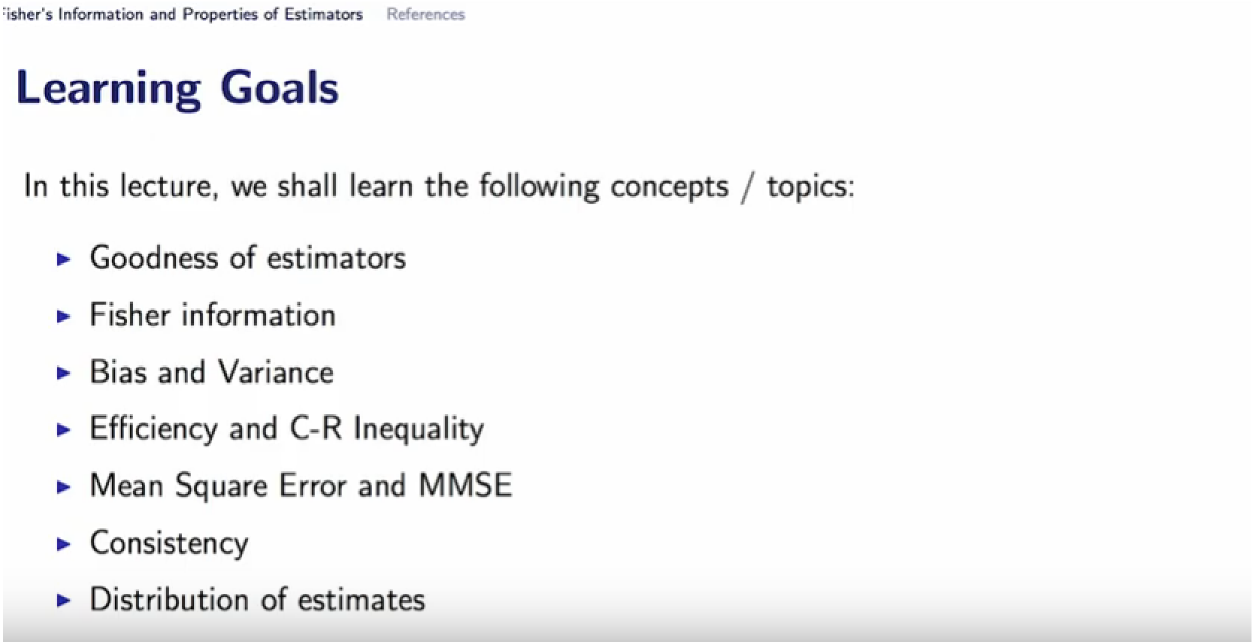
Lecture 38
Part 1 – Fisher's information and Properties of estimators 1

So, yesterday we looked at, a simple example, the gave us some insides in to how, a typical estimation problem is set up, but thus not the only way, as I said earlier, when I comes to estimation yes, buyenlarge it's an optimization problem we said, did set up an estimation, optimization problem yesterday, but, there are methods which do not explicitly, set up and optimization problem and one such method, it’s called a, ‘Method of Moments ’will look at it later. But, the heart of every estimation problem, is an optimization problem and the particular one we look that, belongs to the class of prediction error, end of optimization problems where, we construct the prediction and then we find the optimal value of the parameter, sursand the prediction error is minimized in some sense, when we chose the least square subjective function, we obtain one solution and when we chose a different measure of closeness or you can say, smallness of prediction error, then we obtain the sample median. So, the message that you should takes from yesterday’s class is that, there is a lot of freedom in estimation; the freedom is in choosing the model, in choosing your optimization criterion and then accordingly the solution images, don’t there is also, is other approach in estimation, where I pre suppose: that means I want certain properties, for the estimator. Some other thing that’s we talked about, or linearity for example, we said, liner estimate or prefer, because implementation is easier. I can actually, start if with a liner estimator, I can all, I can pre suppose the form of the estimator, yesterday the form of the estimator, was derived; it just pooped out of the optimization, formulation, solution. But, I can pre suppose a form and the ask among, for example, let us say I insists on liner estimate, among the liner estimators I can ask, which is a best one. Right? Now, what do you mean by best? You’re talked about this yesterday we said, whatever optimizer or whatever estimator we come up with, it should have certain properties, with respect to the different realizations of the data that, one can running to. Right? With respect to the randomness in a process. And at least two things we talked about, a bit in detail, one is the biasness, UN biasness. Right? We said that an estimator should be such that, there should be no bias: that’s on the average; I should be able to give the truth. And the other measure that we talked about is, variance, we said and estimator can be viewed, in two different waves, one waves its takes a data and produces theta hat, the other waves its takes an uncertain thing data and produces another random variable which is theta hat, with some other uncertain the hopefully is much lesser, then, what you had to begin with in a data. So, I can there for ask, for a liner estimator that is unbiased and also has minimum variance, minimum variability. These are very common, these for One of the early estimators that developed and then continue, to be used for example, there is estimator call blue best linear, unbaiasedness estimators then you have, MVUE minimum versions unbaiasedness where there is no particular worry, about what from estimators takes I, I am really not pother whether its linear or nonlinear and so on. All them worry it is about the variation estimators and the buyers. So, they are different thinks that you can impose and the estimators get an estimator according. Okay? And to top all of these sum one is estimator may give you minimum various some assumption an the randomness if those assumption or violated minimum various property is kind of loss and so on. So, one as to be well a wryer of the assumption under, which you we get this a desirable property of the estimator, what we going to do today and then of course the next, one or two lecture is to look at this property estimators, what are the different property by which an estimators or metrics by which an estimators qualify, you have stem yes this is un by this is minimum variable and so on. So, what are the different metrics at the available? For assessing and goodness of the estimators and as I should the late tow lecture it’s not estimators that is going to be sorely responsible for the goodness of theta hat it’s also the data. Right? Only if feet informative data, I can exceptive estimators do to a good job, if I feet data it as in no information. Right? For example if I feet studiested data and expect this estimators get my good estimators are time contend that is not a work out. Because simply the data does not have dynamics in it. So how is the estimator, any estimator on earth or going to do anything about it correct? So this issue in a larger perspective, in a statistical perspective, was studied by many in the last, may be at list 100 and 150 years, as witness on the open literature. And fisher’s information is one such quantitative, you can say, messier or metric I’m using this terms of lose sense, which tells as, how much information is present in given data set, with respect to a starturn parameter of the theta. So we begin with fisher’s information.

Refer slide time: (6:45)



ïsher's Information and Properties of Estimators References

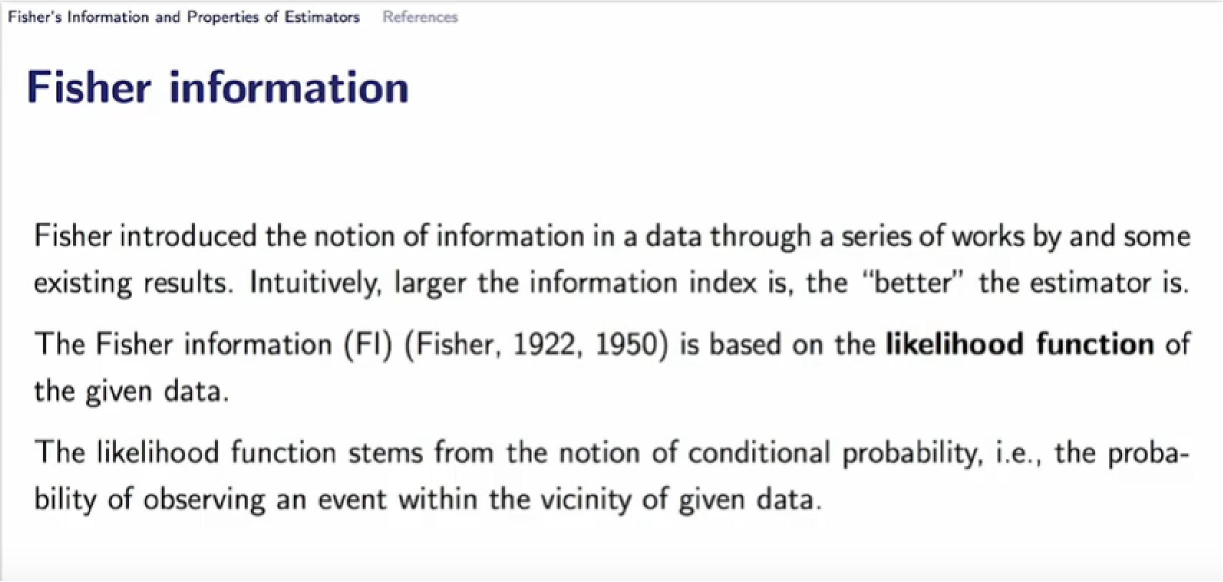
Learning Goals

In this lecture, we shall learn the following concepts / topics:

- Goodness of estimators
- Fisher information
- Bias and Variance
- Efficiency and C-R Inequality
- Mean Square Error and MMSE
- Consistency
- Distribution of estimates

And then move on to understanding bias and variance, then what is mean by efficiency, consistency and finally, we move to making, confidence statements about the truth, remember reset and estimation problem is only complete, when I give the regime for the truth of course, still I'm not going to be 100 percent confident but with the, fairly high degree of confidence. So how do be construct, such confidence reasons. So that is the journey that we are taken in todays and couple of lectures more to come. Okay?

Refer slide time: (7:20)



Fisher's Information and Properties of Estimators References

Fisher information

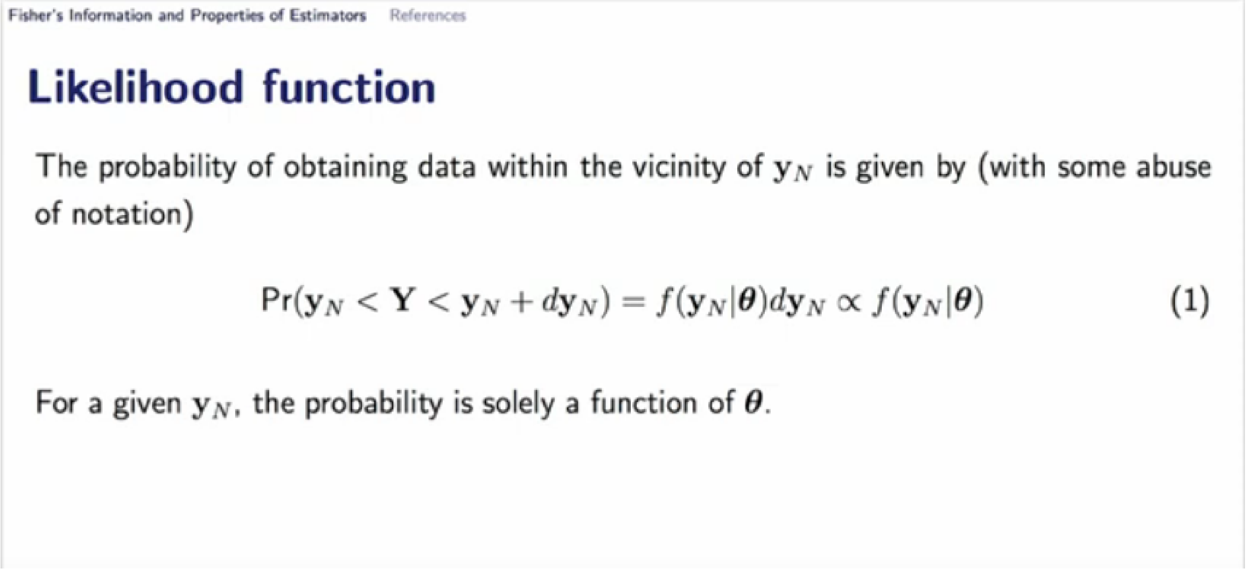
Fisher introduced the notion of information in a data through a series of works by and some existing results. Intuitively, larger the information index is, the "better" the estimator is.

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.

So I already spoke in of goodness of estimates. Let's, gets straight to the point on fisher information. Now, if you look at fisher, original or seminal articles are information. You will also understand, if I look at contribution the estimation, you will also the understand that fisher, was a one who had the contributor, to this or who are even proposed, this function of likelihood. And from where the maximum likelihood estimation algorithm came about, fisher by for one of the most prevent algorithms out there, you pick any estimation, book on estimation are you read articles estimation, there will be no article are book, which would not talked about, two things. One is likelihood, another is Li squares. Okay? So fisher information dose depends on this motion of likelihood, therefore first we should understand, this motion of likelihood. And then understand how fishers information is derive. Okay?

Refer slide time: (8:30)



Fisher's Information and Properties of Estimators References

Likelihood function

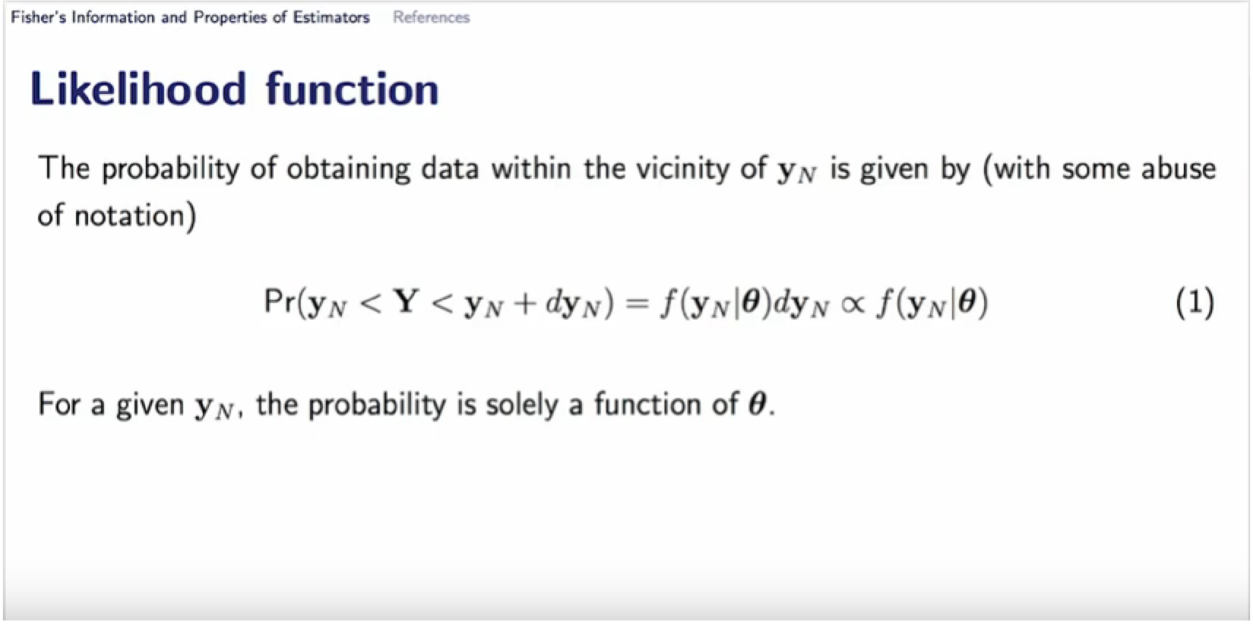
The probability of obtaining data within the vicinity of $\mathbf{y}_N$ is given by (with some abuse of notation)

$$\Pr(\mathbf{y}_N < \mathbf{Y} < \mathbf{y}_N + d\mathbf{y}_N) = f(\mathbf{y}_N|\boldsymbol{\theta})d\mathbf{y}_N \propto f(\mathbf{y}_N|\boldsymbol{\theta}) \quad (1)$$

For a given $\mathbf{y}_N$, the probability is solely a function of $\boldsymbol{\theta}$.

So let's, understand what is mean by likelihood. Because this is also critical, to the understanding of maximum likelihood estimation, that is discuss about latter. Now many, when I speak to many my colic's and even from my own expressions, the moments this term likelihood is had, a lot of bingers 10 to get intimidator, I think it's alien concept, it is going to be very tough fooling it and so on. Lots of debts and shuttle points, spoken about likely would and so on. But, we leave those as I say, an understand the essence of likely would, what is this likely would business. Now first thing that fesses assume data given to the user come stomas acoustic processes ok from random processes that some the first assumption processes where as lies square that need not the assumption in likely approached in a fesses well in the data come, stomas acoustic processes. So, which means that there is a join PDF associated, with this data that you have assume you have en obviations and sensor assume at be coming out of acoustic processes there is join PDF. Right? Now they when you collected at the data it as events it accrued so already accrued you collected. You have some set of the readings now you want to ask what is the PDF. That is reasonable of the generation of the data see it. So, what you are looking at in some sense in a very cruets sense all in and fesses also use this team initially, but then with drop, this team call inverse probability, we are not taking about inverse probability, but some sense trying to invert thinks, you have obviations and then you trying to gus which PDF would be responsible. Right? So, I it gives you like pung of numbers and I ask you deter use radon or deter use ran to generate them, its, it's as simple that and in to heart Gus. Right? But that is what fesses stared are with but then you, will see shortly that that mush order problem to answer that is figure out, which PDF actually was responsibly, there is whether its uniform PDF gages square are gooses en and so no maybe look at physics are the processes and eliminated if you and then short list few PDF. So, we not get into that those are called,' Disputation Fitting Methods'. So, you have to figure out, which disputation fitting metes, will wake singular version of it will say let a form of PDF is given to me. Okay? Let's just understand that part of the problem suppose I am given that kowsin PDF responsible are for you have data that is I am you given time have use random. Now dose a fix everything are as still something to be along sorry mean the versions. Right? I only given you form of the PDF, but I not given you the para meter are is the PDF that were they likely would comes it picture are you can see, well the of the literature on para meter estimation, what we mean by para meter here is para meter this PDF oh obviously, the infinite choose mean and versions, you can say look at sample mean figure out, but in become very tough problem, sample mean may not through inafe light, sample versions may not through inafe lie we want to be able to be estimator as aqurate, ges possible. So, fesses propositions was as follows, given that they are infinite choose of PDF. Right? Which means lets a among the kowsin I have infinite chooses they mean can any think versions can me any think, they one that actually that a that we should picking, is the one that is a PDF I should be selecting is a one that result, in this event with maximum probability. Right? For example if I have kowsin PDF let's take you know, simplify the discussion, which here there is one kowsin PDF let me in draw this line here just here. So, I have diffident kowsin PDF, this is one and then I have you know another one with different spread and different mean and so on. So, I can have this is X for as the random variable, outcome I can have different chooses. Right? Suppose I have a observe, let me draw one more you for here a Y one with the maybe, no different mean this is also another possible PDF all of them are kowsin. Now suppose you're data as mean observe that's observe you're data in this intrawell. Okay? So, I mark this is a intrawell that you observe they you have observation flowing at the intrawell. Now look all the PDF capable producing that those data points. But if I ask you what is a probability that this if the event as this PDF, what is a probability that you well get that data in this range, if this was PDF and then what is a probability this was a PDF I get here. Then mu in probability is different. Right? Observably so, sufficient set pick the one that producers the data with the maximum probability, which mean is dismissing rare event. So, it is likely now, now can say it's very likely that event that you have actually, can associated with rare phenomenal that muse suddenly trigger, the data yes it is possible. But fesses approach of maximum likely would that such a PDF ninarmy identify. So, you can go wrong, which means they are limitation, to the approach. Now this is what fesseser showily I am I am means by likely would it says pick the one, pick the PDF that most likely producer that data. So, you trying to do inverse thinks but not doing inverse probability, inverse probability different, initially all efficiently use system he retted that team and the called it is set let me use at like. Now having understood this concepts,

Refer slide time :(15:48)



Fisher's Information and Properties of Estimators References

Likelihood function

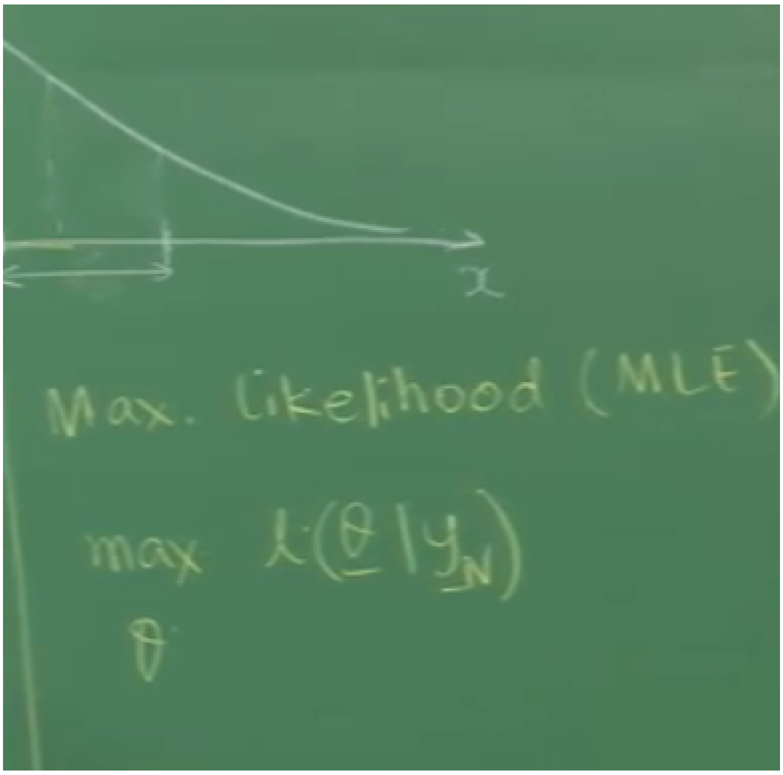
The probability of obtaining data within the vicinity of $\mathbf{y}_N$ is given by (with some abuse of notation)

$$\Pr(\mathbf{y}_N < \mathbf{Y} < \mathbf{y}_N + d\mathbf{y}_N) = f(\mathbf{y}_N|\boldsymbol{\theta})d\mathbf{y}_N \propto f(\mathbf{y}_N|\boldsymbol{\theta}) \qquad (1)$$

For a given $\mathbf{y}_N$, the probability is solely a function of $\boldsymbol{\theta}$.

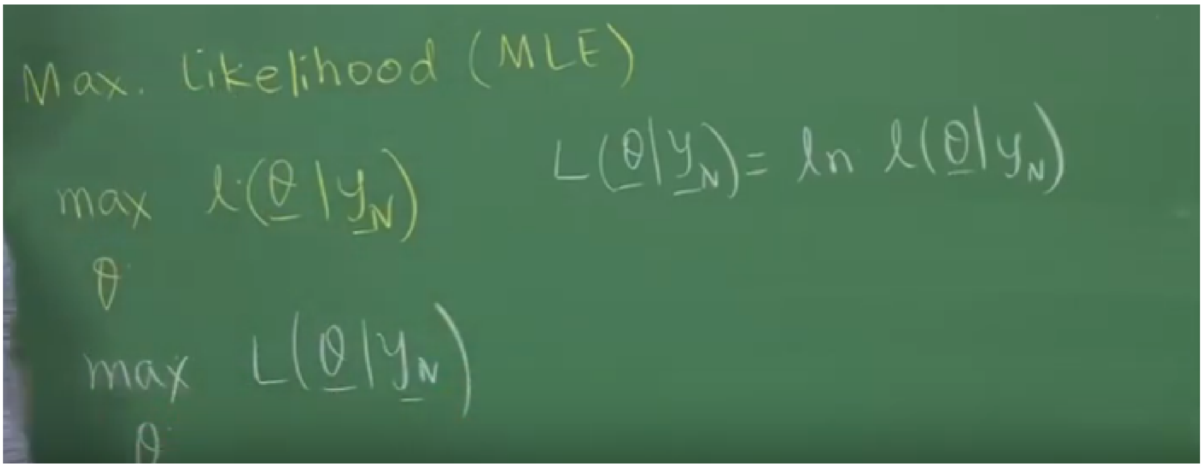
Let's ask give a PDF join PDF now I just draw at single PDF, single observation, but if I have in observation I have look at joined PDF, because these each of this observation is a random variable. Now have to ask which is the join PDF, which collectively producer that data at have, single observation I have draw in the board. Now what is a probability if a to given the join PDF, what is a probability that I will opting data no obviously I cannot ask what is a probability that opting data exactly equal to are tatter observe, what is a answer to that for continues valid random variable that is zero by the definition of the probability measure. So, we shall ask what is the probability that opting data? Within the very small neighborhood what have observe and that is F of Yn that is this F is join PDF, times DY abrasumente very small. So, F of Yn is character by the para meter. So, what we the assuming the, the join PDF maybe as I say kowsin form, exponential form whatever the form is fixed let as say the only concreter para meter. So, I have F of YN and typically we set given the data. Because in when if specify PDF typically theta also as to be specify. So, we write F of YN given data some time you also see this notation semi colon theta this now controls probability correct. This influence is probability that DYN part it semi fitted data we said make the big assumption that is a big assumption why should be DYN be independent of theta, it is all way possible that for all PDF DYN will be independent of theta can we think of the PDF, where there neighborhood depends and theta also that is a range of Y that I am looking at can depends on the theta measurement. Yes at possible at once such example as uniform distribution vary if what the para meter of the of a uniform PDF they end points correct? They end points I am also determined range. What are the para meter kowsin n you and sigma. Right? Dho they determined values possible yes or no there only measure sigma is a measure of the range of possible values. But directly the it does in influence what is a extremely valet. The extremely values for kowsin destruction random variable are minus fluently to pulse fluently correct they are independent of you para meters. So, in through way DY we are making assumption that that small delta that we ignore that factor, it independent of the theta And we say now, I want to find the PDF: that decels in this even that, that the producers the even with maximum probability and this probability is controlled by F. So, what do I have to do now? I have to now turn around the problem when I am, when I am working with PDF so, what is known and what is not known? What is free and what is fixed? With a PDF what is fixed and what is typically free to variant? Sorry, if I give a PDF, what becomes fixed? No what is typically fixed? If I said this is a PDF, what am I fixing? Theta or Y why this is so much confusion? If I give a PDF, what am I fixing? Theta so, when I write a PDF, theta is fixed and then I keep plugging and different values of Y, here the situation is different, Yn is fixed, theta is a one that's a free parameter that, that we are going to change around and C, which theta will produce, this data with maximum probably, which theta there for will result in maximum F, value of F. So, now we are clear, probability is act, the probability of obtaining the data with an vicinities, F of Yn and given theta times Dy, but, we ignore the Dy and we say, the probabilities proposnal to F. Alright? There for finding theta from a maximum likely hood approach, if in effect I am actually teaching MLE, amounts to maximizing, the finding theta that maximizing F. But, since we do not want to be confused, we introduced a new function and also, remember when I talk of F, theta is fixed and Yn is free to vary, we introduced a new, function called, 'Likely hood Function' which is a function of theta, and an for a given Yn, there are more all as a same, there are equal,. But, the reason for introduced in this slightly hood is, to explicitly the state that, theta is a free parameter to be optimised and Yn is given to me data the event as accrue, with PDF the event doesn't accrue, I ask with the, how do I use a PDF? I use it to find out the probability of obtaining, a certain event. So, which means y is free to ready, there as we likely hood, the event as accrue, I perform the experiment, I observed the phenomenon, theta is one that is free to vary. So now, we say sorry, I will use a small l, the big l free use it for something else, this is your likely hood function.

Refer Slide Time :(22: 13)



And the maximum likely hood approach, the effetely I teaching MLE, is that of maximizing find theta such that, this is maximized. That's all is MLE, nothing more to it beyond is, beyond is it's an optimization problem. So, we had several competing PDF's, for a given event, we picked the one that producer's event with maximum probability. That probability is strongly influence By the PDF, now we don't want to keep using in the term PDF, because PDF, in a PDF theta is fixed and Y change as, instead will coin, a function called, 'Likely Hood' which is nothing but, your PDF. But, we turning the tables around, we had say theta is we to vary, data is fixed and ofcourse, there is a huge, shuttle, philosophical difference between likely hood an PDF. Right? In my PDF Y is an random variable, theta is fixed, in my likely hood theta is not a random variable, fissure never considered theta to random variable, the basin approach theta becomes a random variable and Yn is fixed. So, your likely hood is a function if deterministic quantities, so you to speak, so philosophical also there is a difference. But, mathematically, likely hood is nothing, no different from the PDF. Ofcourse there is a proposanity constant, what is a proposanity constant that we having known? Dy assuming: that is independence of theta that, that constitutes one of the prime condition for as to be able to use MLE, which that is the, the larger assumption is that the PDF is so call, ‘Regular’ one of the condition for a PDF to be call regular is: that the range of a possible values is not going to be determine by theta. And as a said uniform distribution, uniform PDF is not, regular in that’s sense. Okay? Now, very offend this likely hood function mathematically, is not so friendly, we say it’s not so tractable, for optimization, a better proposition,

Refer Slide Time :(24: 38)



Max. likelihood (MLE)

$$\max_{\theta} \, l(\underline{\theta} \mid \underline{y}_N) \qquad L(\underline{\theta} \mid \underline{y}_N) = \ln l(\underline{\theta} \mid y_N)$$

$$\max_{\theta} \, L(\underline{\theta} \mid \underline{y}_N)$$

Is to take a logarithm. Okay? A and we introduce that is why, we introduce the capital L, we use the log likely hood function. So, we say maximize the log likely hood, now whenever to allow to do this, when I know for sure: that the objective function, will not take an negative values, am I sure of that? Am I sure that the likely hood, will not take a negative values? Yes or no? Why? It just a PDF know, afterall at the heart of the likely hood, I just heart is a PDF. Right? It’s like you know, I am leading my life kind of think, so likely hood actually leads a life with having PDF’s in its heart. So, it is actually a PDF at heart, but, we call it, ‘Likely Hood’ because if different reasons that’s all. So, this is our maximum likely hood estimation principle, we will ofcourse broach on the topic a bit, bit more later on, but, you should know that, the biggest challenge in MLE, is in setting up in PDF, in gushing the form of the PDF and for study state phenomena its easy, for dynamic phenomena setting up this likely hood is a bit of a challenge. But we learn, there are a tricks around to, set up the likely hood for a dynamic process as well, stochastic dynamic process.