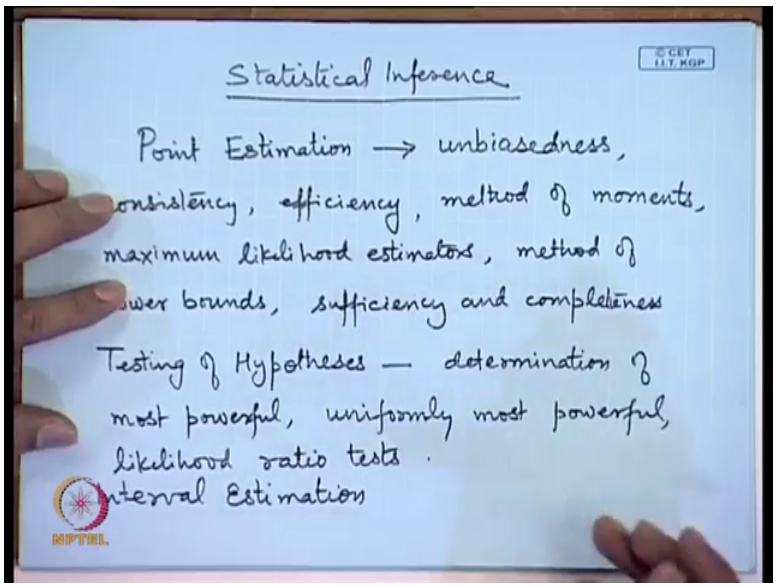
Statistical Inference
Prof. Somesh Kumar
Department of Mathematics
Indian Institute of Technology, Kharagpur

Lecture - 01
Introduction and Motivation –I

Friends, in this course we will cover important topics of a Statistical Inference. This is a second major course in the subject of probability and statistics. We have one basic course on probability theory where we talk about the concepts of probability and distributions. And, in this course we broadly discuss the methods of statistics as applied to day to day problems. We will be mainly covering in this course point estimation.

(Refer Slide Time: 00:53)

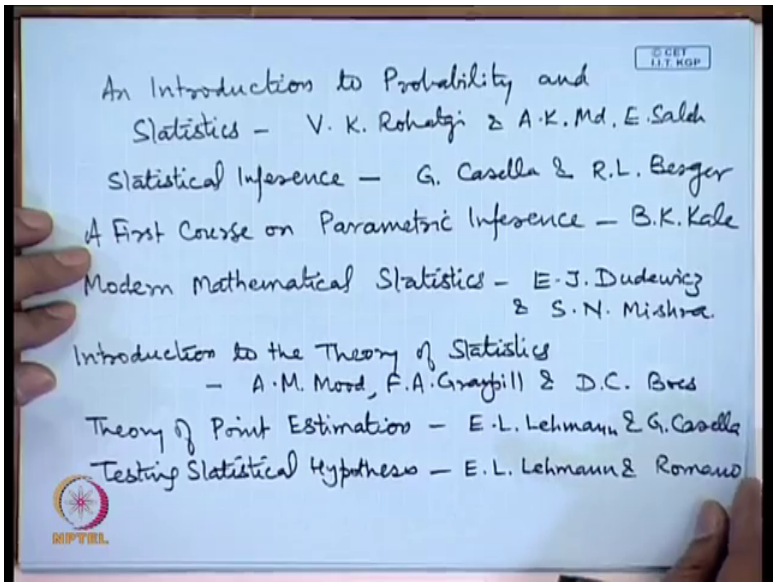


In the point estimation we will cover the fundamental concepts such as unbiasedness, consistency, efficiency. And, then we will discuss the methods of finding out the estimators such as method of moments, maximum likelihood estimator estimators, then we will discuss the uniformly minimum variance unbiased estimation.

The method of lower bounds for determining these, method of lower bounds and another concept is that of through sufficiency and completeness. We will cover another major area and inference that is testing of hypothesis. We will discuss how to find out various kinds of tests, what are the types of errors, the fundamental concepts of testing of hypothesis and determination of the test.

So, in the determination of test we will discuss most powerful, uniformly most powerful tests and then a related concept that is of likelihood ratio tests. We will also discuss the problem of interval estimation. In this we will discuss the methods of finding out confidence intervals for usual one sample and two sample normal distribution problems.

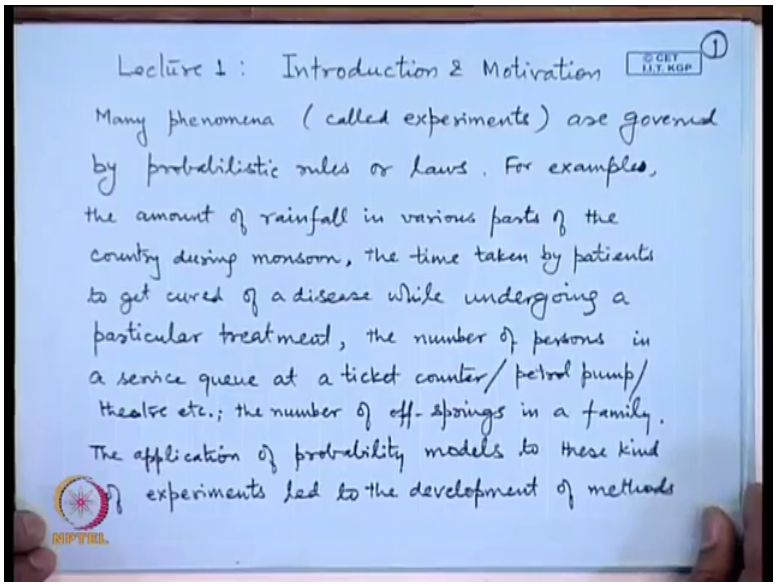
(Refer Slide Time: 03:29)



The important text books that can be used for this course are: An Introduction to Probability and Statistics by V. K. Rohatgi and A. K. Md. E. Saleh. Another important book is a Statistical Inference by G. Casella and R. L. Berger. A First Course on Parametric Inference by B. K. Kale; Modern Mathematically Statistics by E. J. Dudewicz and S. N. Mishra; Introduction to the Theory of Statistics by A. M. Mood, F. A. Graybill and D. C. Boes. Those who are interested to get advance knowledge on statistical inference may further look at the books: Theory of Point Estimation by E. L. Lehmann and G. Caslla; and Testing a Statistical Hypothesis by E. L. Lehmann and Romano.

These books cover almost all the topics that will be taught in this particular course that, I am going to start today. So, let me firstly, introduced what is the problem of statistical inference and why should we study it.

(Refer Slide Time: 06:31)



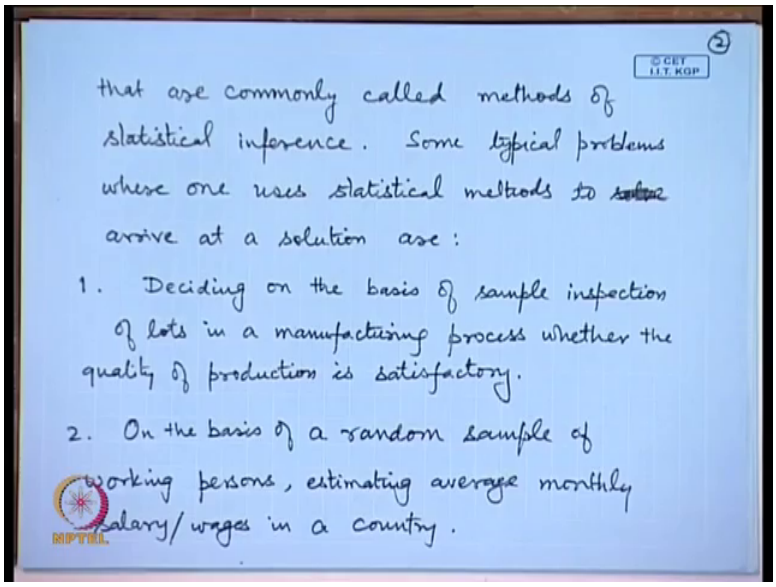
So, let me talk about the introduction and motivation of a statistical inference. We notice that many phenomena in various sciences they are governed by loss which are not deterministic in nature. So, we can call them stochastic or probabilistic in nature. For example, if you look at the amount of rainfall in various parts of the country during a monsoon season, then it is not sure that how much rainfall is going to be there in the next year. How much it will differ from the previous year whether over the entire geographical region in which we are interested in, whether there will be a uniform distribution of the rainfall. Or, in some portions there will be tremendous rainfall and in some other places there will be a condition of drought.

No matter what physical theory we develop or what atmospheric scientist are able to developed the theory, they can never be sure of the exact amount, the timing of the rainfall etcetera in different geographical regions. So, we can say that these are governed by probabilistic laws. Similarly, suppose I consider the time taken by patients to get cured by a disease while undergoing a particular treatment. So, quite often we observe there are patients who are given a certain treatment for a certain disease. We observe that a patient a gets cured within 2 days whereas, patient b takes 10 days to get cured. And, there may be a patient say c who may not get cured by that particular medicine and he may have to be given another type of medicine.

So, the effect of the medicine on different patients are quite subjective in nature, they will depend upon various conditions. Therefore, these are a stochastic in nature. Similar, examples are the number of persons in a service queue at a ticket counter or at a petrol pump. So for example, if we go to a railway counter at a given time of day, we find that at that time there are large number of people standing in a queue. So, for the next time of the when we need the booking we go to the counter at another time thinking that at this time there will be less number of people and we find that that is not true. At another time when we go we find that the number of customers or the number of persons a standing in the queue is much less.

So, the number of persons this is the need of the persons to book in the different trains etcetera is also not deterministic in nature, the number of children in each family. So, these kinds of problems when we look at, this kind of phenomena they are all a stochastic in nature. So, these are nicely modeled using the probability distributions. So, the application of probability models to these kinds of experiments this has led to the development of methods which are commonly called the methods of a statistical inference.

(Refer Slide Time: 09:47)



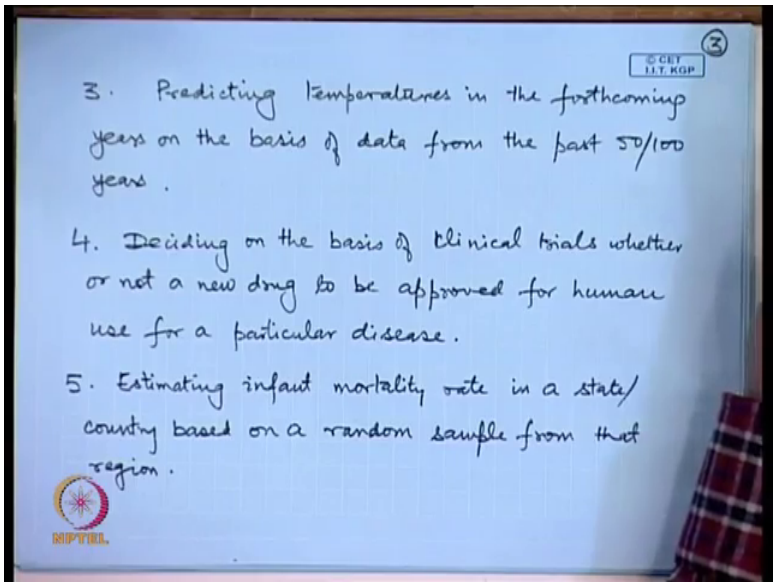
Some typical problems we are one use as the statistical methods to arrive at a solution or deciding on the basis of sample inspection of lots in a manufacturing process whether the quality of product is satisfactory. So, you consider a factory where certain kinds of nuts

and bolts are being produced. Now, we are the factory owner or the manager of the production he is interested to know the quality of the production. If the quality is alright it will be sent to the market at a appropriate price, on the other hand if the quality of the product is not good then it will not be sold at high price, on the other hand it may even be returned.

So, the quality of the production is very important. So now, how do he goes about it? He looks at the batches of the product for example, maybe in a 1000 or in a 100 and he inspect randomly say 10 out of each lot of 100. And, he makes a decision based on that whether the there are more defectives or less number of defectives. For example, he observe that every lot of 10 it does not produce more than 1 defective product. Then he may conclude that the production of the bolts is satisfactory or up to the given marks. Another problem of inference could be that on the basis of a random sample of persons who are in certain employment we want to estimate the average salary or wages in a country.

So, this kind of situation or this kind of inference is important for determining economic policies of the government. It is important for the say consumer goods producing companies they, if they find that the average salaries are high or on the higher side in particular place then they may like to sell the appropriate items to that zone. Because they may get more customers, on the other hand if they find that the average salaries are much lower then high cost goods may not be able to be sold in the market at those places.

(Refer Slide Time: 12:15)



Another problem could be that for example, nowadays there is a lot of talk about climate change, global warming etcetera. So, the scientists are interested in knowing how much increase of the temperature will be there in the global climate during next say 10 years or next 5 years on the basis of the data which is available to us from the past 50 years or from past 100 years. So, on the basis of this we will be able to estimate how much will be the average temperatures in different places or overall global temperature. This will this is going to be useful to determine the policies of the various organizations, various governments that what should they do to reduce the global warming or the effect of the global warming.

Another typical example of a statistical inference is deciding on the basis of clinical trials whether or not a new drug is to be approved for human use for a particular disease. So, there is a disease for which certain drug maybe or may not be available. Now, doctors are the person who are who are involved in the development of the medicines. They come up with certain substances certain biochemical substance which they find to be effective against the disease causing bacteria or virus. Now how do they go about it?

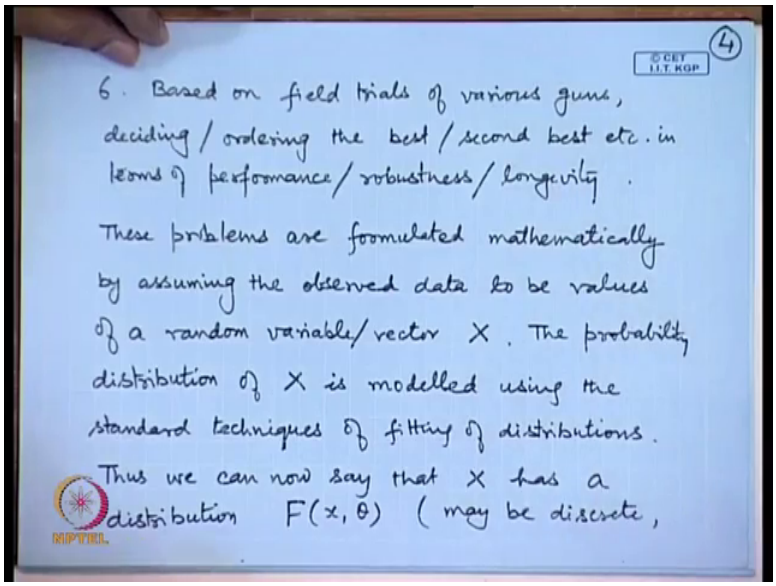
So, they design an experiment where a certain medicine is produced using that using certain amounts of that clinical and that biochemical substance. And, now once it is decided the amount is decided or amount is determined that this is going to be actually useful in curing this disease then it is the question of introducing in the market. Now, the

medicine can be introduced in the market only if it is found that after taking this medicine there are no side effects. As well as it has a high efficacy as compared to the previously used medicine.

Or, if there was no medicine then it should be better than the control; that means, even if the medicines are not taken if some people are getting cured then this medicine should be better than that. So, it is the job of the a statistician on the basis of a random sample he will decide how to take a decision in arriving at a conclusion whether this new medicine is going to be effective or not. Estimating infant mortality rate in a state or a country based on a random sample from that region. So, people talk about the development. So, we say that a country has a high GDP or Gross Domestic Product and therefore, the country is on the path of a growth; but, whether from the human development index point of view whether there is an overall development. So, we look at other parameters such as infant mortality rate, the literacy levels and other factors.

So, we want to note a find out what is the infant mortality rate in that country. If the infant mortality rate despite having high GDP or high average salaries even if it is even then if it is high than. That means, it relates to certain other kind of traditions certain other kind of conditions which may exist there which are not which despite high GDP growth are not going to improve the infant mortality rate etcetera. So, a social scientist and the planners of the country are interested in knowing what is the estimation or what is the estimate or the infant mortality rate.

(Refer Slide Time: 16:23)



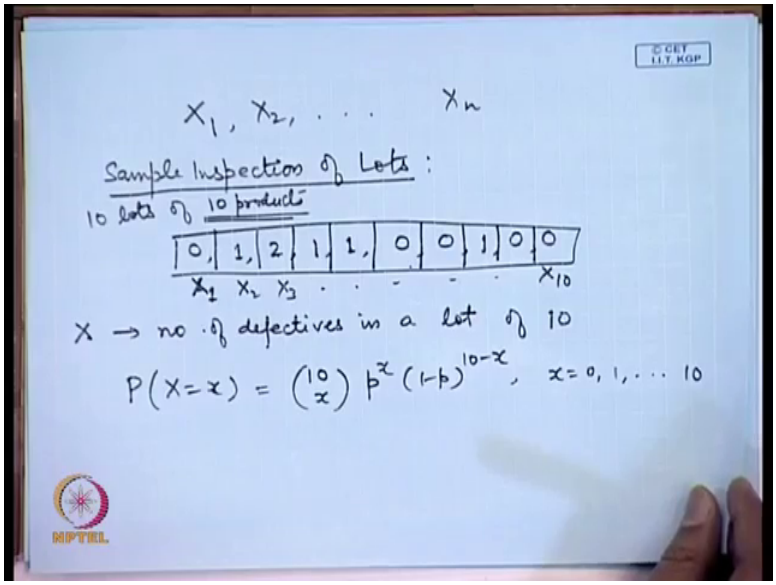
A frequently encountered problem is that there are its various kind of guns which can be used in the by the army. Now, now army wants to buy a new guns to replenish its stock of the arms. So, now various arms manufacturers they give the samples from their factories and then the field trials of those guns are conducted. So, the job of the statistician is to determine that which is the best gun among this, which is the second best etcetera.

In terms of its performance, performance could be in terms of accuracy that or the range that the gun can cover in hitting its targets. Its robustness in the sense that depending upon the different area in whether it is hilly region or whether it is whether it is a plane region or whether it is a desert region; whether the gun can be equally effective in different temperatures, in different timings of the day etcetera and also the longevity of the gun that is also important.

So, the job of the one needs a statistical methods to determine which is the best which is the second best and so on; that means, we have to order then in some preference. So, these kind of problems are usually formulated using a mathematical model or you can say a statistical model. So, on the hand you can say the field person gives some data to the statistician. So, the data is in the form of certain numerical values or some observed values. So, the statistician treats these values as the values of a random variable X. So,

we use a notation say capital X which denotes the random variable whose observed values are given to the statistician by the user by the end user.

(Refer Slide Time: 18:41)



So, they call it the values as X 1 X 2 and X n; as an example we looked at the problems. So, we are looking at say sample inspection of lots. In the example of sample inspection of lots what will be X 1 X 2 X n. So, suppose 10 lots are inspected of say 10 products each; that means, each lot of the bolt contains 10 bolts. And, we are looking at in terms of say the diameter of the top or the length of the bolt etcetera. Anything which is the quality control terminology used for describing the goodness of that product so we may use that.

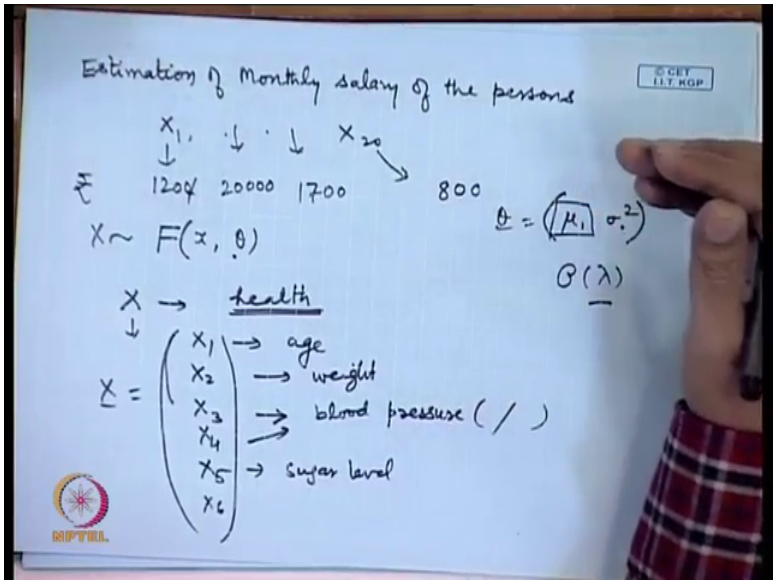
Now, the data could be in the form of say 0 1 2 1 1 0 0 1 0 0. So, these are the 10 values; that means the lot number 1 it had no defective, lot number 1 had 1 defective, lot number 2 had 2 defective. Lot number 4 had 1 defective, lot number 5 had 1 defective, lot number 6 and 7 had no defective. Lot number 8 had 1 defective lot number 9 and 10 had again no defective bolts according to the criteria that has been fixed by the quality control manager for that particular product. So, for a statisticians these values will represent the values of X 1 X 2 X 3 up to X 10.

So, the random variable X denotes here the number of defectives in a lot of 10. So, in a lot of 10 there are X number of defectives in fact, in this particular case one can find out the corresponding probability distribution. For example, if I say that there are x number

of defectives in a lot of 10. So, one may use a binomial model or one may use a hyper geometric model depending upon the conditions that have been imposed on this kind of situation. Suppose there is a constant probability p of being defective then this probability will be 10 c x p to the power x 1 minus p to the power 10 minus x, for x is equal to 0 1 to 10; that means, you may have a binomial model to describe the distribution of x.

Let us consider another problem; suppose we are looking at the random sample of working persons and we are looking at estimating average monthly salary in the country.

(Refer Slide Time: 22:11)



So, in this particular case if we are looking at estimation of monthly salary of the persons say employed persons. So, the values here would be calculated as X 1 X 2 say X 20; suppose 20 persons have been considered here then the values could be in terms of rupees say. So, you may have say for a person it could be 1200 rupees, for another person it could be maybe 20000 rupees, for third person it could be say 1700 rupees, for one person it could be 800 rupees etcetera. So, here X 1 X 2 X 20 denote these values and now we may use this data to have a model.

So for example, the model may be given by certain distribution say F. It could be normal distribution; it could be gamma distribution etcetera depending upon the actual values that have been obtained there. So, so the probability distribution of X is modeled using a standard techniques of fitting of the distribution. And therefore, we can say that X has a

distribution say F x theta which of course, maybe discrete or it could be continuous or it could be mixed. And, theta denote some characteristic of the population which could be a scalar or the vector.

Here X itself can be a scalar or a vector depending upon the type of the things we are having. For example, the situations that have been described now here in all these cases X is a discrete random variable. However, there may be some other situations. For example, I am looking at X as the observations are taken on a person regarding his health. He goes to a medical practitioner and he wants to have an estimate of his average health.

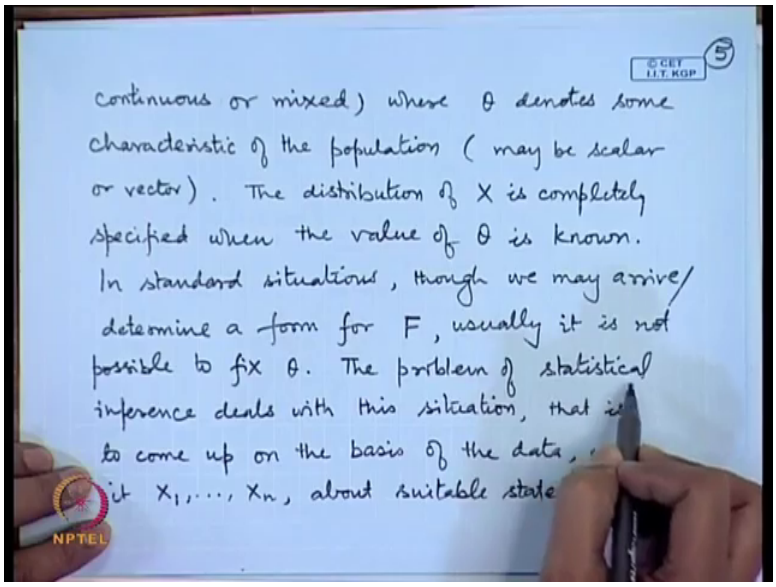
So, the values that X may take here may consist of certain components; say X 1 which may relates to his age. X 2 may relates to his weight that is his body weight, X 3 may denote his blood pressure. Now, blood pressure may consists of 2 values so, you have 2 values here. So, you may consider them as X 3 and X 4 then you may have his sugar level and so on. Say his pulse level, in this case X is a random vector. Similarly, the parameter of the distribution F for the x; so the parameter theta itself may be a vector or it may be a scalar.

In the case of monthly salary, if you are having a distribution such as a normal distribution then the normal distribution is characterized by two parameters say mu and sigma square. In this case my theta is consisting of two parts. If you are considering say the number of persons arriving in a queue in a given time period then we may model it by using say Poisson distribution which is having a parameter lambda which is the rate of arrival here. So, for different problems we will use different probability models to describe the setup. So, we are having X 1 X 2 X n as the random sample which we treat as the observed values of a random variable X.

And, X will is assume to have a known distribution F x theta and therefore, the distribution consists of certain parameter which is called characteristic or parameter of the distribution. So, the distribution of X is completely specified when the value of theta is known. Now, in a standard situations although we may arrive or determinate form of F, but in most of the practical situations it is not possible to fix the value of theta in advance. We may find out the distribution is normal like or it may be Poisson, Poisson model is more appropriate to describe the number of arrivals you know during the time

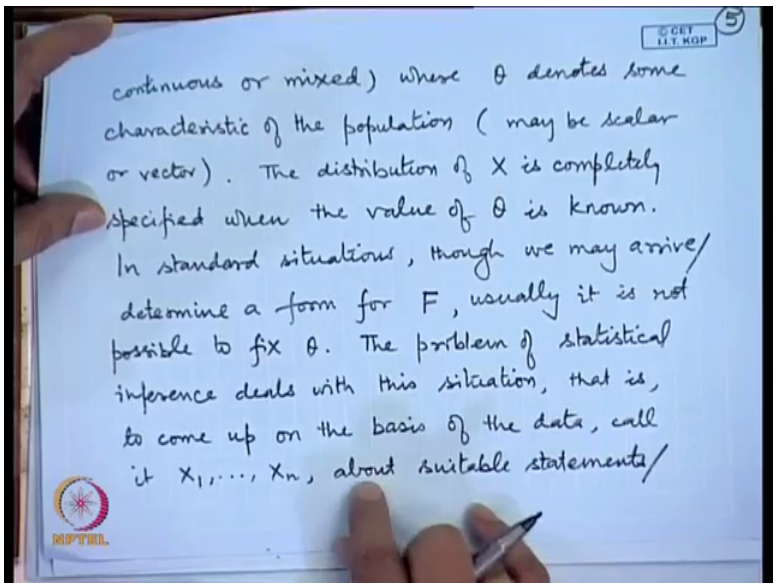
period, the number of failures. Similarly, in some situations we may arrived at conclusion that binomial model is more useful or gamma distribution is more useful. But, the appropriate parameters of the distribution one may not know in advance.

(Refer Slide Time: 27:37)



So, the distribution is completely specified if the parameter is known, but in most of the practical situations it will not be known. So, the definition of the statistical inference or the problem of statistical inference is to determine on the basis of the given data that what would be the value of this unknown parameter.

(Refer Slide Time: 27:53)



So, it is not necessarily that we just tell the value. So, the general problem of statistical inferences is to make suitable statements or the assertions about the unknown parameter of the population. Here we can break this up into two broad areas: one is that some feature of the population in which an experimenter or enquirer is interested. So, let us a g theta this may be completely unknown and the experimenter would like to make a guess or estimate about this feature on the basis of a random sample from this population. So, this is called the problem of estimation.

So, here he may come up with a single value for as an estimate. So, for example, when I say average salaries and he may come up with a figure say 2200 rupees per month. If we give a single value or a unique value for the unknown parameter of the population this is called point estimation, because we are giving a single value that is a point. On the other hand we may specify a range. For example, when we talk about the expected temperature in the coming year then we may say that the expected average temperature during the month of June is likely to be between 42 to 44 degree Celsius in a particular region of the country.

So, here we are not telling a single value like saying average value is 43 degree Celsius rather we are giving a range. Now, this range has to be qualified using certain probability statement. This is called the problem of interval estimation and this is another part of a statistical inference. So, in estimation we may specify a single value or we may specify a range of values. This is called the problem of estimation.

So, this is one major area a statistical inference.