Essentials of Data Science with R Software – 1
Professor. Shalabh
Department of Mathematics & Statistics
Indian Institute of Technology Kanpur
Lecture No. 55
Needs for Drawing Statistical Interferences

Hello friends welcome to the course essentials of Data Science with R Software – 1 in which we are going to understand the basic concepts related to the probability theory and statistical inference. So, from this lecture, we are going to start a discussion on a new topic, that is estimation of parameters. Now, the first question comes, what is this estimation of parameters? Now, if I ask you a very simple question, why are you trying to learn this data science? What is the basic objective? What you really want to achieve?

So, the answer is very simple. I want to know something or I want to get an answer of some questions, queries on the basis of a sample of data, but, the main problem is, I am trying to solve this problem on the basis of a sample of data. But what I want? I want my conclusions to be valid or they should remain valid for the entire population. Think about it. This is exactly what are you trying to do?

For example, suppose you want to test the efficacy of a medicine. What we try to do? We try to take some number of people who have got the disease and we try to give them the medicine, and based on their outcome that how many people have responded that they are cured, they are not cured, etc, we try to make a conclusion. Have you ever heard that medicine like paracetamol will control the fever of only Indian people or you have heard that this medicine will not work in say Germany, U.S., African countries or anywhere in this world? No. Wherever and whenever any person in this world has a problem of simple body temperature, the persons are using the paracetamol and they are getting cured.

But do you think that this experiment was conducted on all the people of all over the countries in this world? No, this experiment was conducted as some specified place by choosing some number of people, but the conclusion that this paracetamol is going to be effective for controlling the body temperature is going to be valid for the entire population. So, my question is how to achieve this, how to get it done?

So, essentially, what are we trying to do, there is some characteristic which is for the entire population and we want to know the answer of that question or I want to find out the value that is valid for the population. But this is pretty difficult so we try to do that we try to take a

small sample from the same population and based on the sample, we try to do our statistical calculations. But what we want that whatever statistical calculations we have done on the basis of a small sample they should remain valid for the entire population.

So, now, how to get it done? How to ensure that whatever we are doing on the basis of a small sample will really remain valid over the entire population. Or for example, if I say that you want to know the population mean that means the mean value of all the units in the entire population. But suppose you try to take a sample and try to compute the arithmetic mean. And you assume that this arithmetic mean is going to indicate the value of the mean, which is in the population not in the sample.

So now, there are a couple of questions. Who told you? Who asked you to compute the arithmetic mean? That can be median that can be mode that can be geometric mean that can be harmonic mean. So, the question comes, which of them is a better measure? Which of them is going to give us a better value of the population mean, which is unknown to us? That is not known to us, but we want to know only on the basis of this sample. So, what are the issues related with this type of problem, how to get this value, how to know whether the value is good or bad, etc.

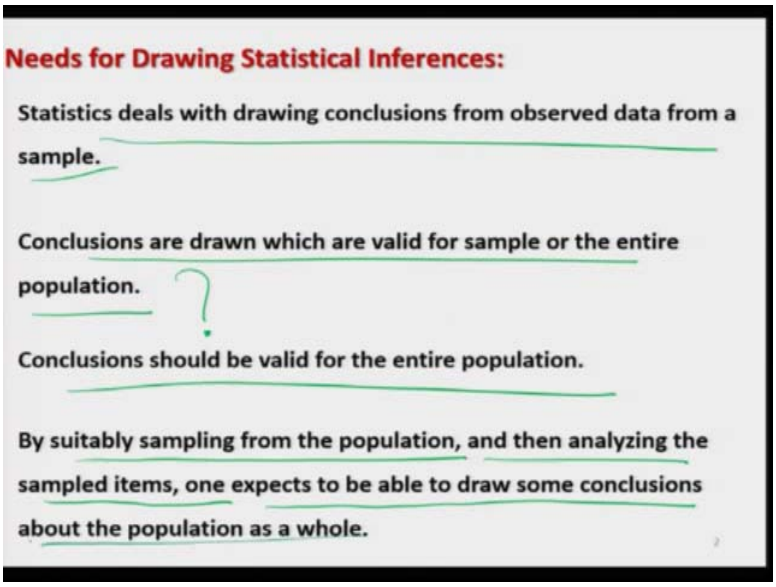
These are all the issues which are handled under the topic of estimation of parameters and that is what we are going to start from today. Well, one thing I would like to make it clear before I come to my lecture formally that in this estimation of parameter part my main interest will be to let you know that how are you going to get good value once you get the value then you can easily compute it in the R software.

So, the role of R software in this chapter may be comparatively less than other lecture. So, here we have to basically concentrate on the basic fundamentals. Once you are clear about the basic fundamentals, then estimating them on the basis of sample of data in the R software is not a difficult thing at all. That is my promise to you. And as soon as I come to the lecture, you will realize the same thing.

So, now, what I'm going to do in this lecture, I will try to explain you the issues the concept etc related to the estimation of parameters. So, basically in this lecture, the amount of mathematics or the amount of algebra is going to be very less, but, I will try to give you the basic ideas. So, let us begin our lecture and try to understand what is this estimation of

parameters and what we really want to do. So, let us begin our lecture. So, now, in this case, first, we are trying to learn about what is the need for drawing statistical inference.

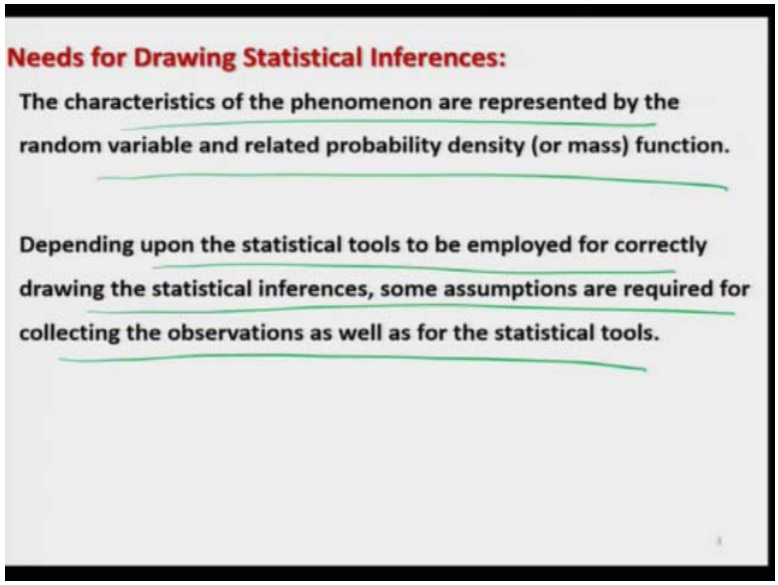
(Refer Slide Time: 07:12)



And then we will try to discuss the smaller topic. So, we know that statistics deals with drawing conclusions from observed data from a sample. And these conclusions are drawn that are valid for sample or the entire population that is my question mark. What do you want? Means the conclusion that you are getting on the basis of sample will they or do you want them to remain satisfied only for that sample, and as soon as the sample changes their conclusions will change or they should remain valid for any of the samples that you can draw from the population? So, the answer is the conclusion should be valid for the entire population.

So, what we try to do that we suitably try to draw a sample of observation from the population or that is called sampling from the population and then we try to analyze the sampled items or sampled units, and respect that we can draw the conclusions about the population as a whole they are not remain valid, they will not remain valid only for the sample, but they will remain valid for the entire population.

(Refer Slide Time: 08:20)

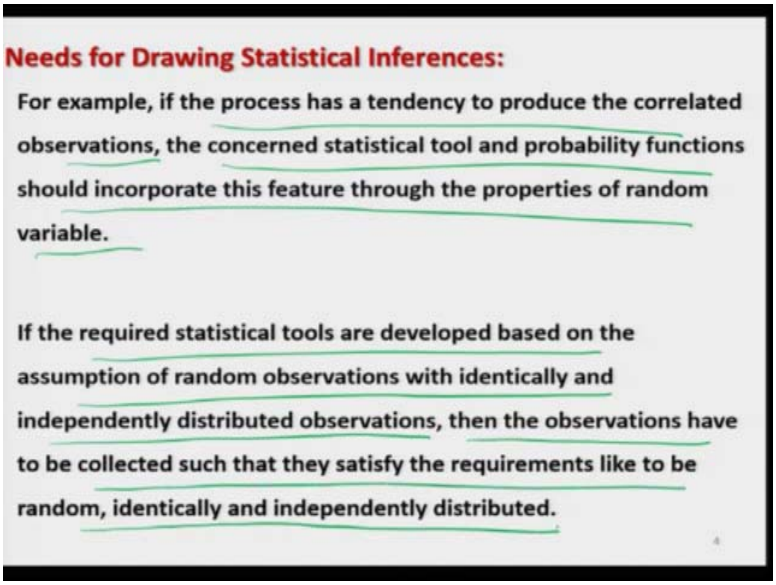


So, under this type of objective the first question comes whatever is your question whatever you want to know that has to be transformed or that has to be defined through a random variable. And whatever are the characteristics of that phenomena or the process from where you want to draw the inference or you want to know the answer of your question, the characteristic of that phenomenon can be prescribed by a probability function. That can be a probability density function or a probability mass function depending on the nature of the random variable.

And then we have to think and decide that which of the statistical tool has to be used, so, that you get a correct statistical outcome. And for that, depending on the tool, we need to have some basic assumptions, which are to be satisfied while we are trying to collect the observation, and when we are trying to use this statistical tool over the collected observations. So, for that, we will always have to make some basic assumptions.

And this is a two-way process, that first you try to fix that what type of statistical tool is going to give you the correct outcome and then you have to look into the need and requirement of the tool which are to be satisfied while collecting the data and vice versa also that while collecting the data, you have certain characteristics. You try to see that the characteristics can be satisfied by which of the probability function or what statistical assumptions you need. So, that these assumptions are or these characteristics are satisfied in the probability function?

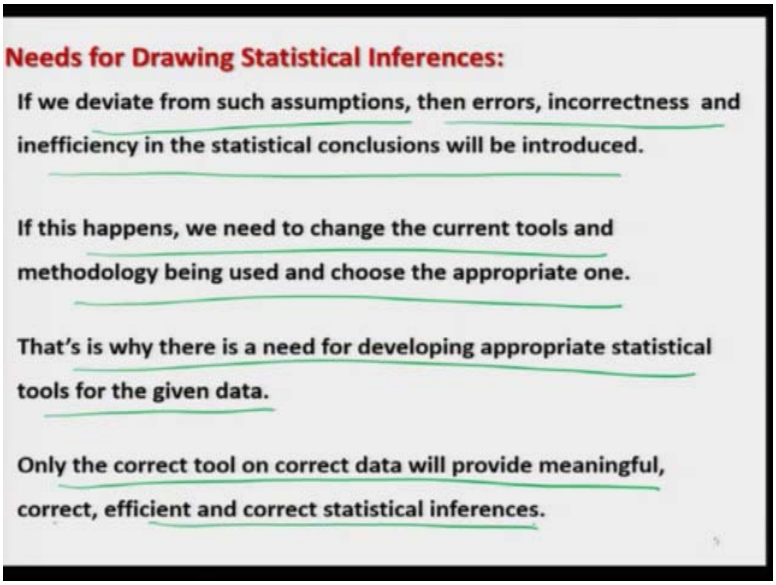
(Refer Slide Time: 10:12)



Now, for example of the process has a tendency to produce the correlated observation that means the concerned statistical tool and probability function should incorporate this feature through the properties of random variable. For example, up to now, you assume in most of the cases that let $X_1, X_2,\ldots, X_n$ be identically and independently distributed random variables. Now, you can choose the proper assumption that either they are only identical, only independent or say none.

And maybe the required statistical tools are to be developed based on the assumption of the random observation with IID that means Identically and Independently Distributed observation, then the observation should also be collected such that they satisfy the requirement like to be random, identically and independently distributed. So, assumptions should be satisfied in the observations, and observation should also satisfy the assumptions.

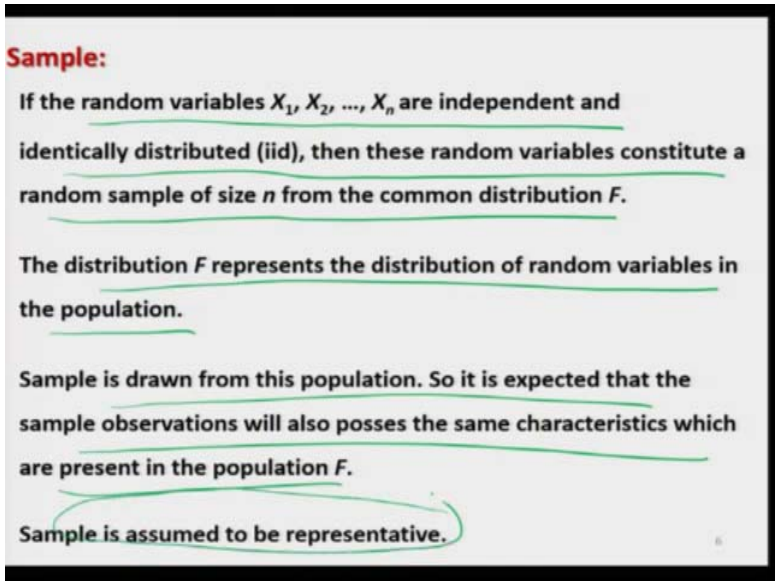
(Refer Slide Time: 11:08)



And in case if you try to deviate from any such assumption, there is going to be some violation. And as soon as there is any violation, then some errors incorrectness or inefficiency in the statistical conclusion will be introduced. Something will happen. What will happen? It will be depending on what type of assumption is getting violated under what type of circumstances, but in case, if this happened, then we need to in the tool that we are trying to use, and the methodology that we are trying to employ, and we have to choose the correct tool and correct methodology.

And that is why there is a need for developing the appropriate statistical tool for the given data. As we have discussed in the lectures in the beginning that only the correct tool on correct data will provide the meaningful correct efficient and correct statistical inferences that we always have to keep in mind.

(Refer Slide Time: 12:01)

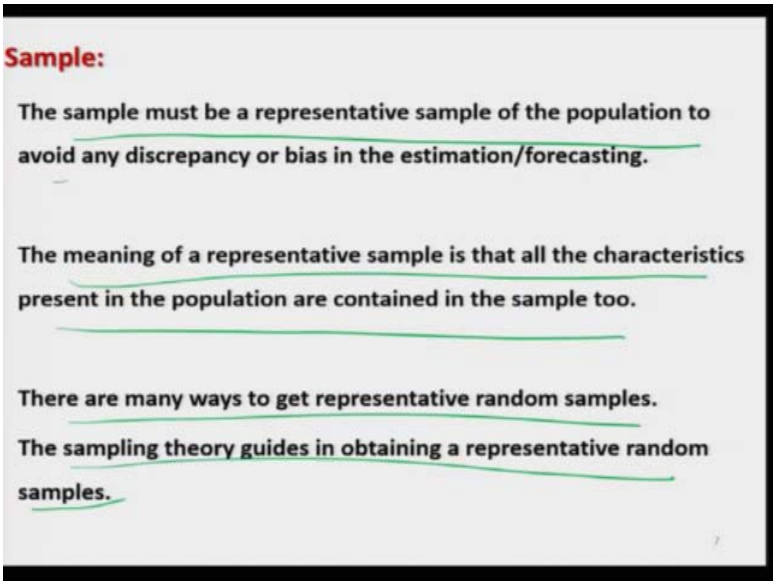


So, whenever we are trying to learn something about the population that has to be done on the basis of a sample that is the basic objective of statistics. So, of the random variable X_1, X_2,…, X_n are independent and identically distributed, then the random variables constitute a random sample of size n from the common distribution F. Because, means you have done many such example, let X_1, X_2,…, X_n be a random sample from normal population or normal distribution with mean μ and variance σ^2 or let X_1, X_2,…, X_n be from exponential distribution with parameter λ.

So, we try to assume that these random variables are going to constitute a random sample of size n from a common distribution F. Means, if you try to say that the first observation is coming from normal, second is coming from Poisson, third is coming from binomial, fourth is coming from exponential then I think the statistical process will become too complicated. So, we are trying to assume here that all the observations are coming from a common distribution that can be a probability density function or that can be a probability mass function.

So, this distribution is going to represent the distribution of the random variables in the population. And when we are trying to draw a sample from this population, then we expect that the sample observation will also have the same characteristic which are present in the population capital F. And this goes without saying you will not find anywhere written, but it is assumed by default that the sample is always representative. As soon as you say this is a sample that goes without saying that the sample is a representative.

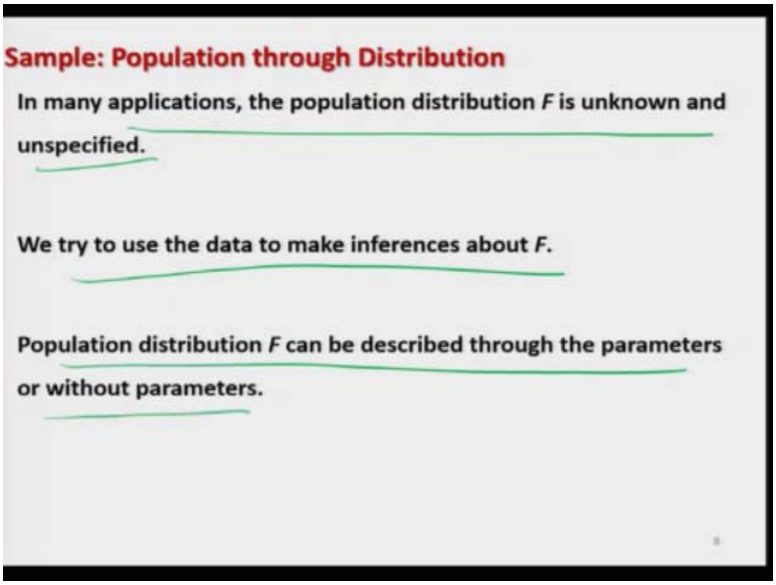
(Refer Slide Time: 14:00)



And representative in the sense that it is going to present all the characteristics, which are present in the population. So, that meaning of the representative sample is that all the correct juicy present in the population are also contained in the sample. Now, that is a different aspect that how are you going to collect such observation. And there are several ways to get such representative samples.

In fact, in statistics, we have an area sampling theory and then we try to learn about different types of sampling schemes and they try to help us in taking a decision that how should we draw the random sample, so, that it is representative. And in case if the sample is not representative, then the statistical inferences may have some type of error, bias, discrepancy, etc. So, that is why we always assume that the sample μst be represented to sample of the population and this is going to avoid any type of discrepancy or say bias in the estimation or forecasting of the parameters.

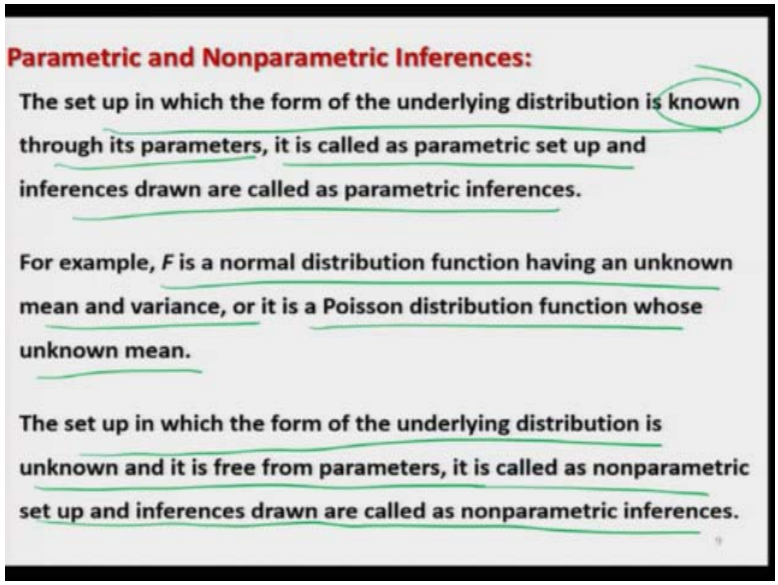
(Refer Slide Time: 15:10)



Now, in many, many applications, you will find that the population distribution function F is unknown to us and that is unspecified. We do not know what is the form like as we do not know whether it is normal or say Poisson, we do not know what are the values of the parameter like as μ, σ^2 in case of normal distribution or say λ in the case of Poisson distribution.

So, what we have to do we try to use the data to make some inferences about F. And whatever is my population that is going to be described by the distribution F, and distribution F is going to be specified by the parameters. But there are two options that are sometime the F can be prescribed through the parameters and sometimes it can be prescribed without parameters.

(Refer Slide Time: 16:03)

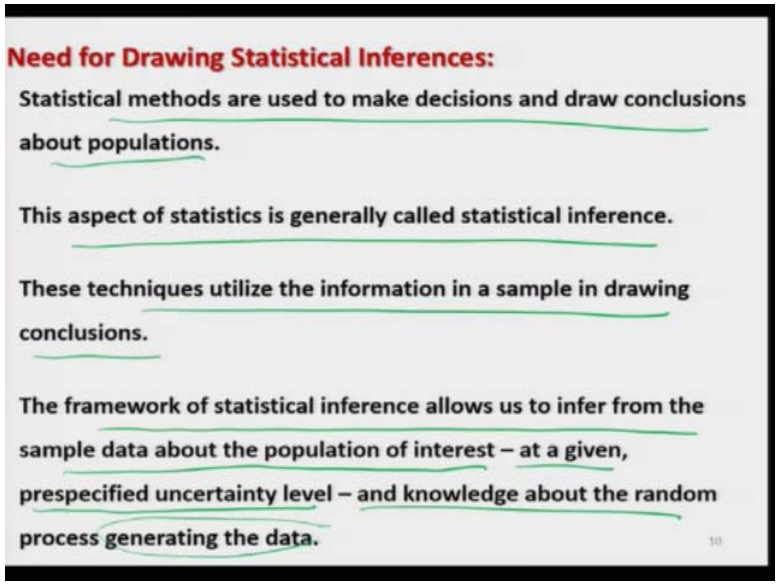


So, based on this criterion we have two types of statistical inferences or we have two types of setups, which should help us in dealing such problems and they are called as parametric and non-parametric inferences. So, the setup in which the form of the underlying distribution is known through its parameters that means we know that it is a normal population and its mean is this variance is this, then it is called a parametric setup. That means, you have information about the form and the parameters.

And whatever statistical inferences we try to draw from such setup they are called as parametric inferences. For example, if I say that F is normal distribution function having an unknown mean and say variance. The variance can be known that can be unknown it depends on different type of conditions or I can also say that F is a Poisson distribution function whose mean is unknown to us that mean λ is unknown to us.

And on the other way, the setup in which the form of the underlying distribution is unknown and it is free from parameters it is called as non-parametric setup, and in such a case the inferences are drawn using this setup and those type of inferences are called as non-parametric inference. And in statistics, you will see that we have an entire area of this parametric inference and nonparametric inference. In this lecture and in this course, we are going to talk only about the parametric inference.

(Refer Slide Time: 17:41)



So, now, the statistical methods are used to make decision and draw conclusions about the population. And this aspect of a statistic this is generally called as statistical inference. There are certain statistical techniques which utilize the information in the sample and they help us that how should we draw the statistical conclusions in the correct way.

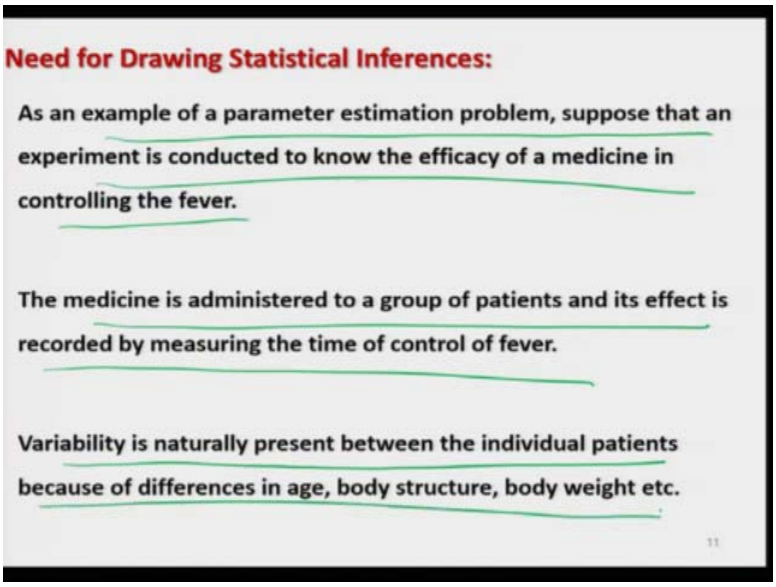
So, what is really going to happen, the framework of statistical inference allows us to infer from the sample data and our conclusions are going to be about the population of interest. And this is obtained at a given pre-specified uncertainty level and knowledge about the random process which is generating the data. Well, you have to understand that you are trying to know about a very big population on the basis of a very small sample of data.

And every time you try to draw the sample from the population, there can be different sample. The probability of getting the same sample is very, very less, which you have seen in many applications, whenever I was trying to generate the random numbers from the R software in a specified probability mass function or probability density functions. You have seen that every time you are getting a different sample and those values are trying to give a different value of the statistic that you are trying to compute to draw statistical inferences. So, certainly there is going to be uncertainty involved industrial scale inferences.

And since, we really do not know with 100 percent confident that what is the exact form of the distribution function F, so we try to assume at an appropriate form and but there may be some discrepancy between the true form and the form that we are going to consider. So, the

uncertainties from such factors will be reflected in the process when we are trying to take conclusions.

(Refer Slide Time: 19:50)

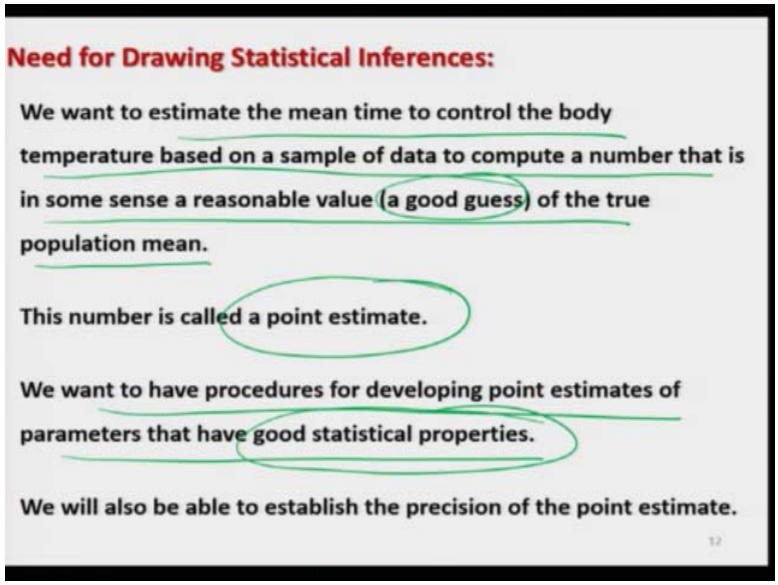


Now, as an example of a parameter estimation problem suppose that an experiment is conducted to know the efficacy of the medicine in controlling the body temperature that is fever. The medicine is administrated to a group or patient and its effect is recorded by measuring the time of control of fever.

Now, in case if you try to take a group of people and they are given the medicine, some persons may have the body temperature control up to say this 5 hours only some people may have the body temperature control for more than 5 hours, somebody may have 6 hours somebody may have 7 hours somebody may have 5.5 hours and so on. So, you will see that these values have certain amount of variability.

So, this variability is naturally going to be present between the individual patients and this variability can enter into the data because of several reasons. Like as the age or patient's body structure body weights, etc. So, that can happen.

(Refer Slide Time: 20:59)



So, now, our objective is to estimate the average time or the mean time to control the body temperature based on a sample of data to compute a number that in some sense a reasonable value or a good guess of the true parameter mean. For example, you have seen that many times we say that whenever a person has a body temperature or fever the doctor says, okay try to take the medicine after every 8 hours.

How does this doctor comes known that the medicine has to be given for 8 hours. Because the experiment was conducted and it was found that once the dose of the medicine is given to a patient on an average, this can control the body temperature up to 8 hours. So, the doctor does not know what is your body structure inside your body, but after having a lot of experience he can say that okay, whatever be the body structure, this medicine is going to work approximately 7 to 8 hours.

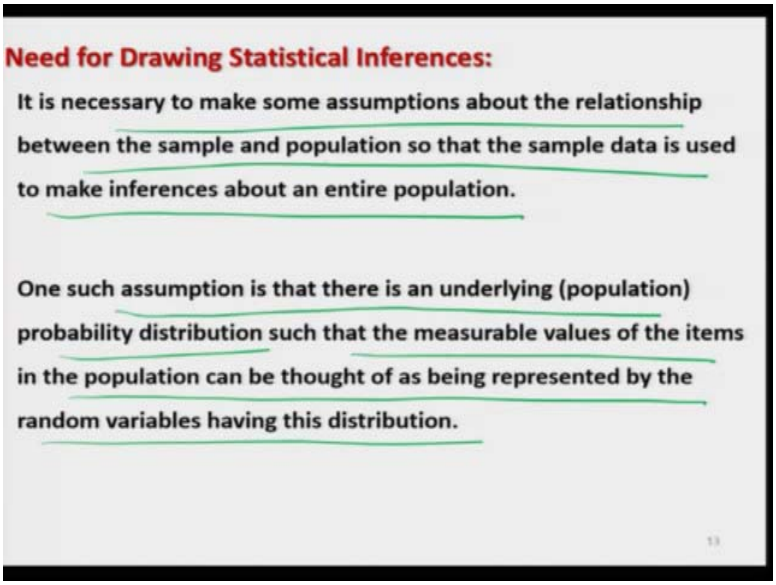
So, sometimes you can see the body temperature remained control for seven and a half hours then 8 hours and 9 hours or seven sometimes more than this. So, that is what we mean by saying that we are trying to estimate only the average value. And the number of what we are finding this is called as a point estimate. Have you ever observed that the doctor says okay, this medicine can control the body temperature for 6 hours or for 8 hours or that time that you are going to take from your home to college is about 20 minutes. So, you are trying to specify a value which is a point.

So, that is why this type of estimation or this type of statistical inferences, they fall under the purview of point estimate. And we want to have the procedure for developing such point

estimates of the parameter that have good statistical property. Good statistical property mean? For example, if the doctor says to the patient that your body temperature will remain in control for say 8 hours and then the patient ask, will this be 7 hour 30 minutes or say 8 hour 30 minutes? And if the doctor says no that can be any duration between 2 hours to 20 hours, do you think that is it a very efficient statement.

So, that means the way it is computed that is not a good way. So, we always want to have a value of the estimate with some good statistical properties. What are those good statistical property that I will start discussing one by one from the next lecture. And we will also be able to establish the precision of the point estimate that how good it is how what is the variability about this value.

(Refer Slide Time: 23:47)

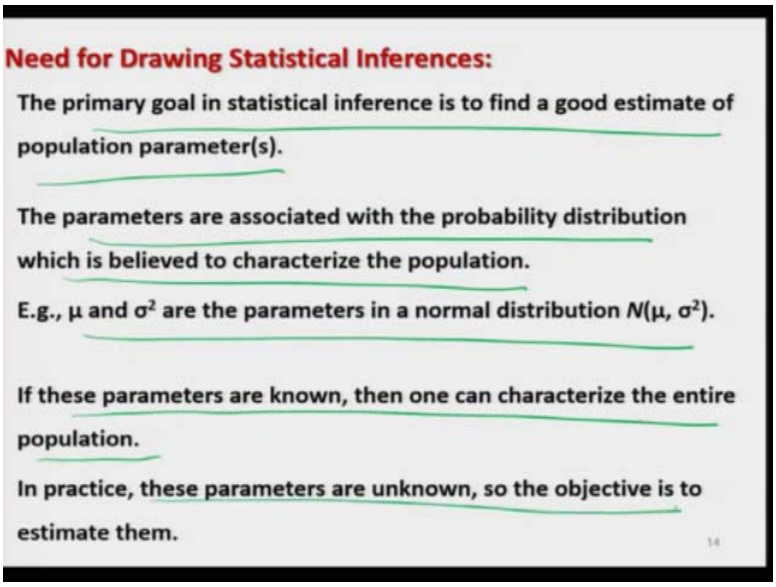


So, in case if you want to draw such a statistical inference, so, it is necessary to make some assumption about the relationship between the sample and population, so that the sample data is used to make inference about the entire population. And one such a basic assumption is that there is an underlying probability distribution that is going to represent the population and this probability distribution is such that the measurable values of the items in the population can be thought of as being represented by the random variable with having this distribution.

So for example, if you want to measure the height of some students, so you say that my random variable is going to be height, and you have to specify that, the distribution of the

heights in the student population is for example, going to follow a normal distribution with the with some mean and some variance. So, this type of assumption is needed.

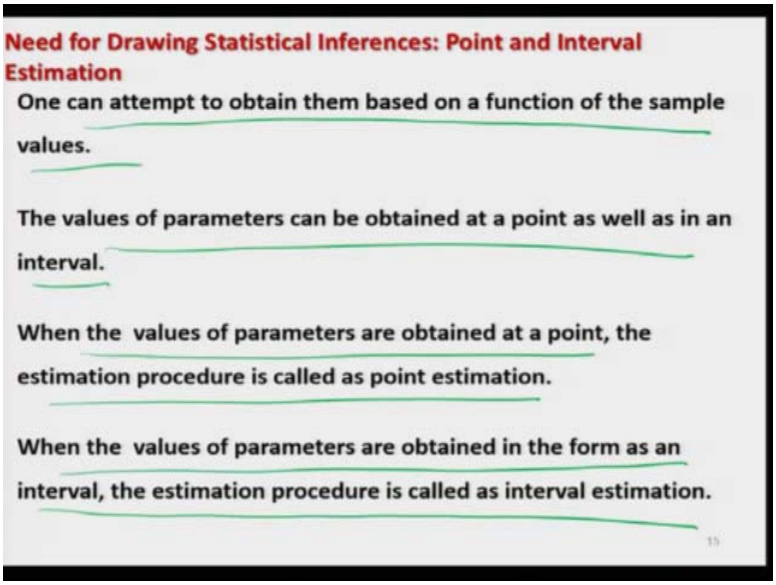
(Refer Slide Time: 24:47)



So, the primary goal in a statistical inference is to find a good estimate of the population parameter. And if there are more than one parameters, we have to take care of them also. And these parameters are associated with a probability distribution, which is believed to characterize the population. For example, if I say, we are assuming that my population is normal, then there are going to be two parameters μ and σ^2, which are going to characterize the normal distribution.

And if these parameters are known to us, then one can characterize the entire population. For example, you know that if you know the value of μ and σ^2 and the normal μ, σ^2 distribution then you can find out its mean, variance, skewness, kurtosis, occur, etc. whatever you want, but, what is really happening in practice? That these parameters are unknown to us. So, one of the very important objectives is to first estimate them.

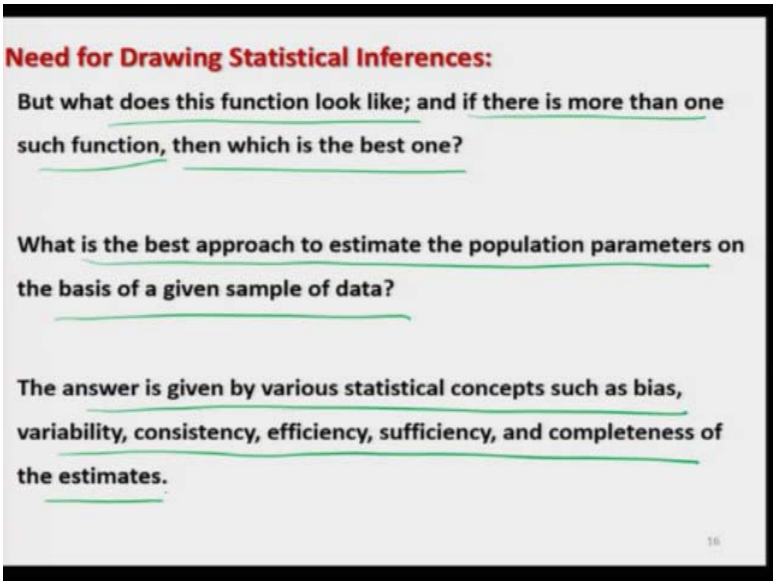
(Refer Slide Time: 25:50)



And one can attempt to obtain these estimates as a function of the sample values. That the value of the parameter has to be known on the basis of observation that you have collected and the values of these parameters can be obtained at a point as well as in the form of an interval. So, when the values of the parameters are obtained at a point, the estimation procedure is called as point estimation, and when the values of parameters are operating in the form of an interval estimation procedure is called as an interval estimation.

For example, means, if I asked you a very simple question that how μch time do you take from your home to college. Now, you have two options. You tell me that I take 20 minutes that means, you are trying to tell me the value of the time at a point, and this is going to be the point estimate. But you can also tell me, I will take say between 15 minutes to 25 minutes. So that means, you are trying to specify the time in the form of an interval 15 to 25. So, this is called as an interval estimation, and when you are trying to give me only here a value say, 20 minutes then it is called as a point estimation.

(Refer Slide Time: 27:06)

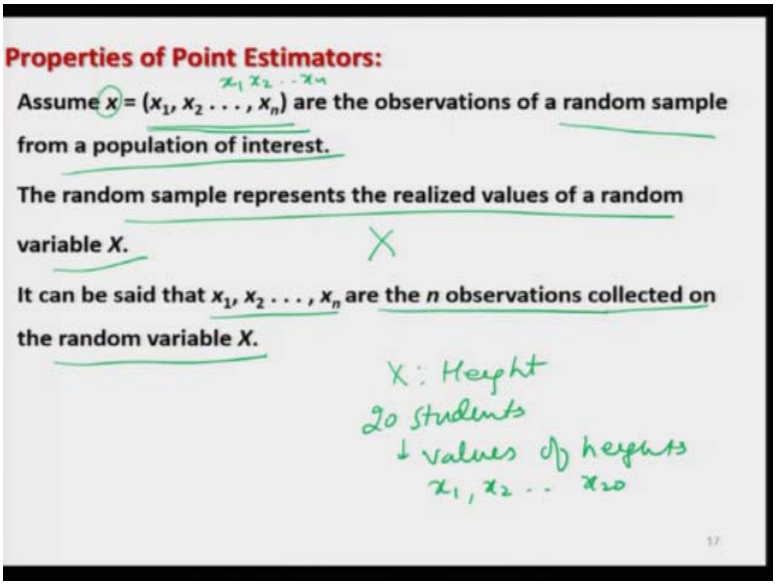


But the question is this when you are trying to find out these values 20 minutes or 15 minutes or 25 minutes, how are you going to obtain them? They have to be obtained on the basis of given sample of data. But where you have to employ the data? The question is what you really have to compute? What you really have to calculate so that you get these values which are representing the truth?

So, the question now, here is what does this function look like? And if there is suppose more than one such function then what to do? For example, if you want to find out the average value and suppose you have two options, you can compute the arithmetic mean or you can compute the say this median, mode, harmonic mean, geometric mean then which of them has to be utilized and which of them is going to give you the correct value or more precise value? That is the question.

So, the next question is what is the best approach to estimate the population parameter on the basis of a given sample of data? And the answer is given by various statistical concepts such as bias, variability, consistency, efficiency, sufficiency and completeness of estimates.

(Refer Slide Time: 28:18)

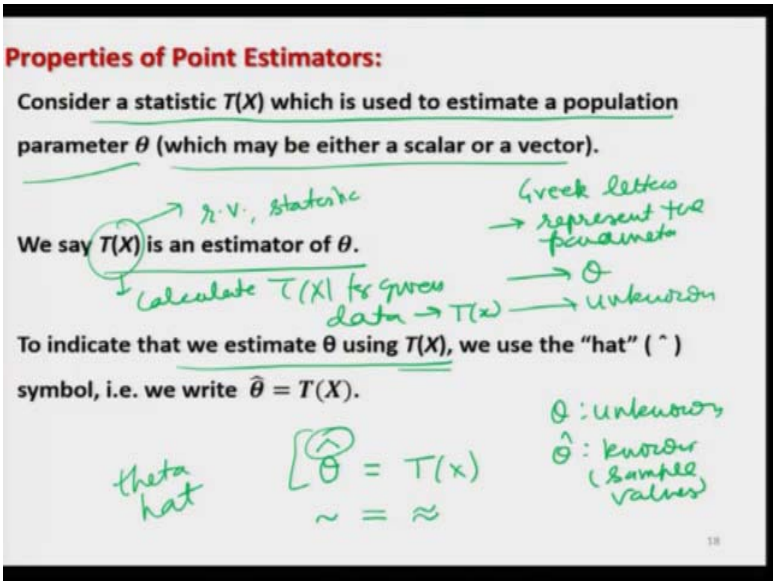


And that is what we are now we are going to study in this chapter in the forthcoming lectures. So, what we are going to assume that we are going to assume that there is a sample of observations and this sample is a random sample, which has been drawn from the population of interest, and the sample values are going to be indicated by say x_1, x_2,…, x_n where x_1, x_2,…, x_n are written in the lowercase alphabet that is a small x like this one is small x_1, small x_2,…, small x_n.

And sometimes just in order to make the representation clear, we will simply write here as a x. So, my random variable here is going to be capital X and these values are going to represent the realized values of this random variable X. So, it can be said that this x_1, x_2,…, x_n, they are the small number of observations which are collected on the random variable X. For example, if X is your here height, so, what you try to do? That you try to suppose take a sample of 20 students from your class out of say this 100 students and you try to find out their values of heights.

So, there are going to be 20 values of height. So, this can be written as x_1, x_2,…, up to here x_{20}. So, this is what we mean when we are trying to say that to let a small x_1, small x_2,…, small x_n are the say this an observation which are collected on this random variable x. And this capital X is going to follow a certain distribution.

(Refer Slide Time: 29:54)



Now, definitely when you want to estimate a population parameter then you need to do all the calculations only on the basis of the random values or the random sample that you have obtained. So, in statistics, there is a general sat this understanding or there just is an understanding that with Greek letter, Greek alphabets, we try to represent the parameters. So, usually you will see that μ, σ^2, λ they are representing the parameters.

So, in general we will assume that now we have a parameter θ is unknown to us and θ can be a scalar quantity the quantity or this can be a vector quantity, and this θ is unknown to us. So, we want to know its value on the basis of the random sample or that read the value of what we what we have obtained. So, obviously, means, we are looking forward to find out the form of a function that can give us then appropriate value of the θ.

So, definitely this function is going to be a function of the random variable, so the function of the random variable is called as statistics. So, now, we can assume that we are going to consider a statistics T(X) which is used to estimate the population parameter θ and θ can be a scalar as well as a vector. And once we are trying to do it, we say in a statistical language that T(X) is an estimator of θ. So, T(X) is always going to be random variable, that is a statistic which is a function of a random variable.

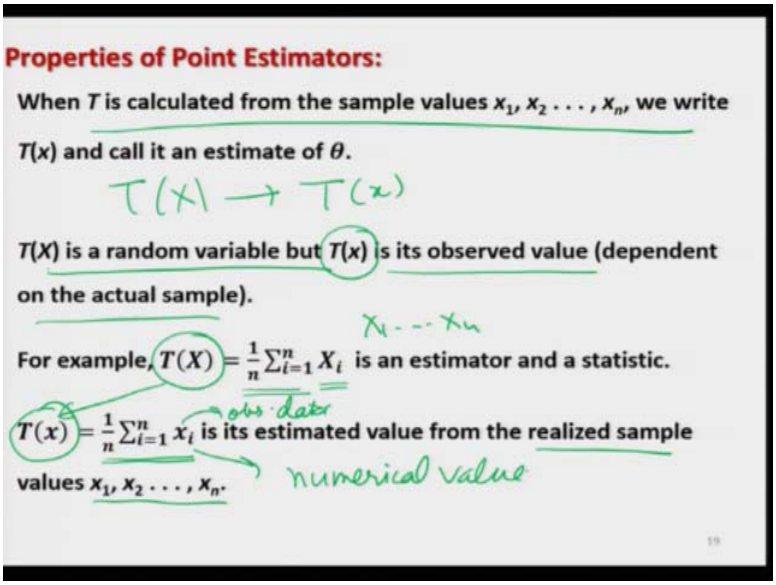
And now, what will happen now, you will try to calculate the value of T(X), calculate T(X) for given data. So, now, what will happen, this will give us a value here like is here T say small x. So, that will be the value of that capital T(X). So, now, in case if you want to

indicate that this unknown parameter θ is going to be estimated by the statistics T(X), then we try to write it like a like this we will write down the parameter θ that is equal to here T(X) but definitely because this is a wrong statement that θ is equal to T(X), because θ is an unknown value which is based on the population and T(X) is the value that is based on the sample.

So, what we try to do, we try to put here a hat like this and this is called here as a $\hat{\theta}$. So, as soon as you try to put a hat on the parameter that is going to indicate that this value has to be obtained on the basis of given sample of data. So, θ is unknown, but $\hat{\theta}$ is going to be known, and this is going to be known based on sample values, this is what you have to keep in mind. And means I am taking here the symbol hat, but many times we can also use other symbols also like tilda or double bar or double tilda and so on.

So, well I am simply trying to say that, we try to put a symbol on the upper side of the parameter to indicate that this is going to be an estimator.

(Refer Slide Time: 33:19)

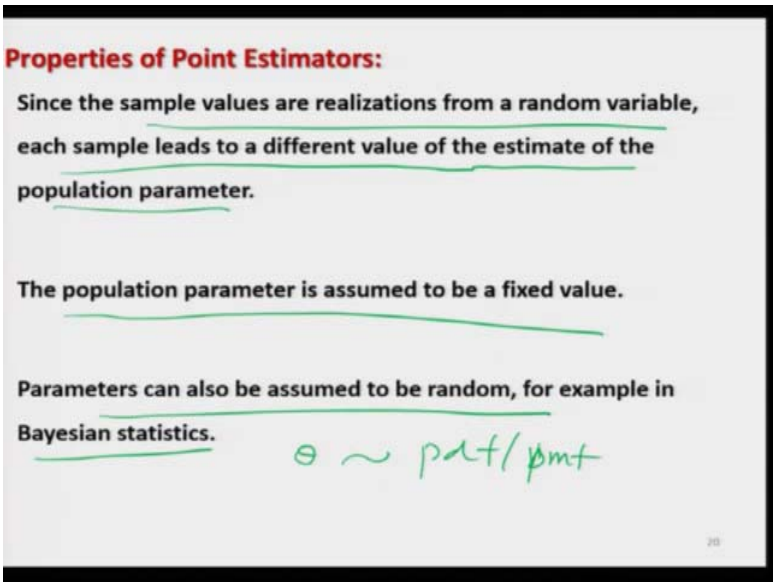


And when we try to calculate the value of T on the basis of given sample of data, we try to write this T(X) now, so, here T small x, because small x is the observed data that is indicating this one. So, this T(X) is a random variable, but this capital T with a small x this is observed value, and this is dependent on the actual sample that whatever sample you are trying to draw, this value is going to be changed. And if you try to change the sample this value is going to be changed.

For example, if I say that capital T(X) is equal to one upon n summation it goes from 1 to N capital Xi which is based on a sample see here $X_1, X_2, \ldots, X_n$. So, this is going to be a statistic this is going to be a random value and this is an estimator and of some parameter. But when we are trying to observe the values, we of this random variable that means, we are trying to observe a sample of data then we try to compute this a capital T capital X that is the statistics, and the value of this capital T(X) here is obtained here as a T small x which is 1 upon n summation i goes from 1 to n is small xi. So, small x_i is the observed data.

So, this $X_1, X_2, \ldots, X_n$ they are the realized sample. And whatever is the value that is going to be a numerical value that is what you have to keep in mind. Because estimator is a random variable, an estimate is its numerical value, which is based on the given sample of data. And in case if you try to change the random sample these numerical values are expected to be changed and this value of the statistics will also change.

(Refer Slide Time: 35:03)

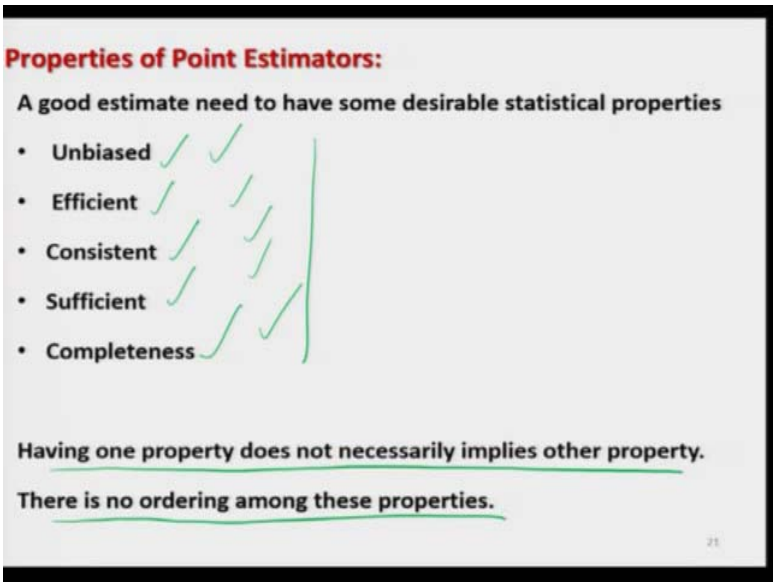


So, change the sample values are realization from a random variable. So, these sample will give us a different value of the estimate of the population parameter. Now, the question is this, whether this population parameter is going to be a fixed value or a random variable. So, we are going to assume in this course that population parameters is a fixed value that is our assumption, but it does not mean that parameters always has to be only fixed. The parameters can also be assumed to be random.

For example, when we are trying to deal with vision statistics where we assume the parameter θ this has got some PDF or say probability mass function. So, both the options are possible,

but once again remember one thing that in this course we are assuming that the parameters are fixed.

(Refer Slide Time: 35:57)



Now, when we are trying to obtain the value of the estimator then several question comes up. That the first question is, how are you going to obtain a form of the statistics so that you come to know that you need to find out $\sum X_i$ or say $\sum X_i^2$? So, that is the approach and these are -- and there are some estimation techniques which will help us. Once you obtain the values from those estimation methodologies then the question comes how to judge whether the obtained values are good or bad.

So, now, from the next lecture, I will try to handle all these issues one by one. So, there are several properties of estimators, which we expect that if those properties are there, then we can say that the estimator is good. So, these properties are unbiased, estimator, efficient estimator, consistency, sufficiency, completeness etc beside a couple of others.

So, we are going to talk about these properties in the from the next lecture. But one thing you have to keep in mind. Now, suppose I have to teach you all these properties, I cannot teach all of them at the same time. I have to take up one by one. So, in most of the books you will see that they handled that first they try to explain about the property of unbiasness then efficiency then consistency then sufficiency and then completeness.

So, sometimes people try to take it that these are the order of importance. So, here I would like to mention very clearly that this is not the order of importance, these are different

characteristics. And having one property does not necessarily imply that other properties are also satisfied. There is no ordering among these properties.

So, now, we come to an end to this lecture. And well that was only a story telling type lecture, but it was very important for you to understand that what are we going to do now. What is our objective? And many times, I have seen that the terminologies like estimator, estimate etc., what is here, hat, that creates a big confusion among the students. So, you please try to revise this lecture. And I have intentionally not given you any mathematical theories or details in this lecture.

Now, you have understood what are we really going to do. Once you know this thing after that the life becomes simpler and you can understand that topics that one by one they are going to come and then you have to learn, but when you are trying to imply them, you have to use all of them together. For example, when you are going to opt a statistic to estimate a parameter you have to test whether this is unbiased, whether this is consistent, whether it is having the property of sufficiency, completeness etc.

So, you try to revise this lecture and then I will see you in the next lecture, and I will try to handle one property at a time, but without any ordering. So, you practice and I will see you in the next lecture. Till then, goodbye.