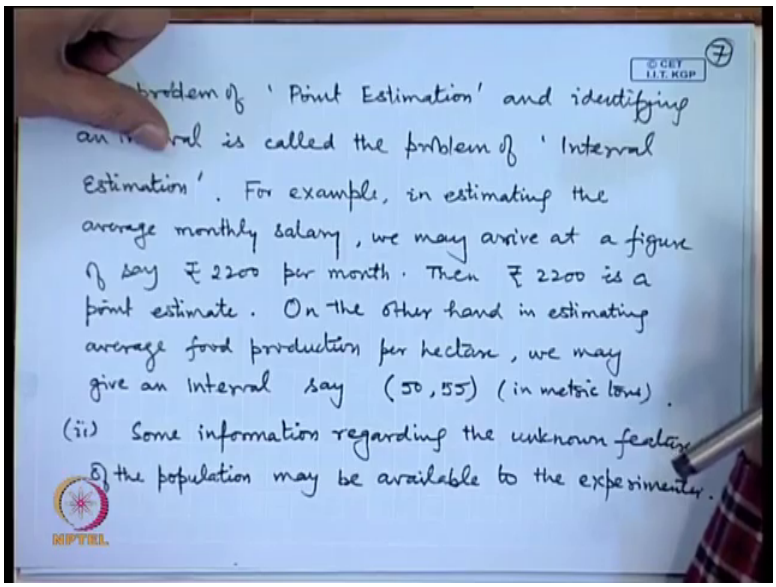
Statistical Inference
Prof. Somesh Kumar
Department of Mathematics
Indian Institute of Technology, Kharagpur

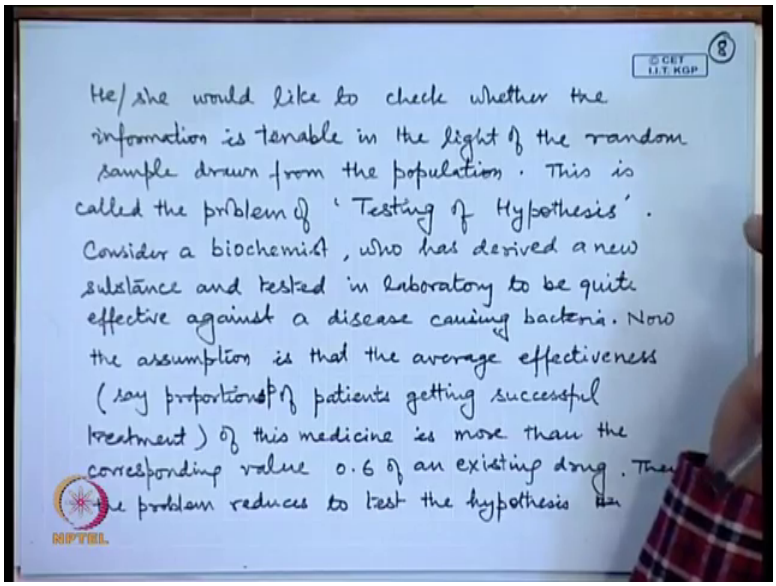
Lecture - 02
Introduction and Motivation-II

(Refer Slide Time: 00:21)



The second major area or broad categorization of statistical inferential problems is that we may have some information regarding the unknown feature of the population which is available to the experimenter.

(Refer Slide Time: 00:35)



Now, the experimenter would like to check whether the information is appropriate or it can be sustained in the light of the random sample which is drawn from the population. So, this is called the problem of testing of hypothesis. So, let us go back to the example of a new medicine getting developed. So, a biochemist, he has derived a new substance and tested in the laboratory that it is quite effective against a disease causing bacteria.

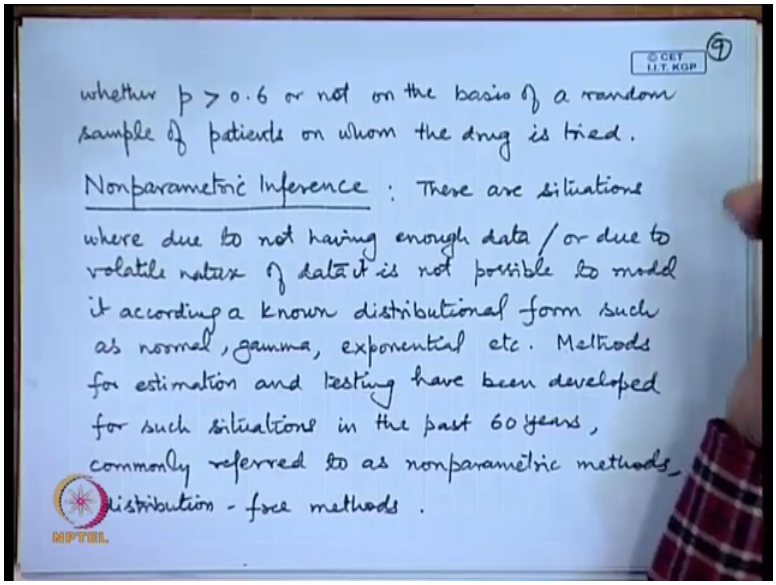
Now, the assumption is that the average effectiveness of the medicine by which is prepared using this new substance will be more than the corresponding value or you can say. So, now when you are testing this effectiveness you have to identify in what terms you are measuring the effectiveness. Is it the proportion of the patients getting treated successfully, or is it the length of the treatment, or is it the survival rate, etcetera.

Suppose we fix here our measurement of effectiveness of a medicine by the proportion of the patients which get successfully treated. So, let us call it P. So, now, that means, suppose we give the medicine, the medicine is given or the treatment is given to say 100 patients, out of that how many get cured. So, we look at the proportion. Suppose this proportion is P for the new drug.

Now, there is an existing drug which had 60 percent cure rate; that means, 0.6 is the proportion which was curable using the previous drug. So, now, in order to have or you can say in order to introduced this new medicine in the market, we would like to check whether this P, the proportion of the patients getting cured using the new medicine is

greater than 0.6 or not, this is called the problem of testing of hypothesis. So, this is the outcome of this test will be determined by the statistician using an appropriate a statistical method. So, in this particular case, it will be an appropriate test. So, based on a random sample of certain patients who are given the medicine, one will need to check these things.

(Refer Slide Time: 03:06)



There is another distinction which I would like to make at this point, there are situations where due to not having enough data or due to volatile nature of the data, it is not possible to model the data according to a known distributional form such as normal distribution or a gamma distribution or an exponential distribution. Because many times the data is huge and it may be having lot of variations, therefore appropriate known probability models are not suitable to fit that distribution.

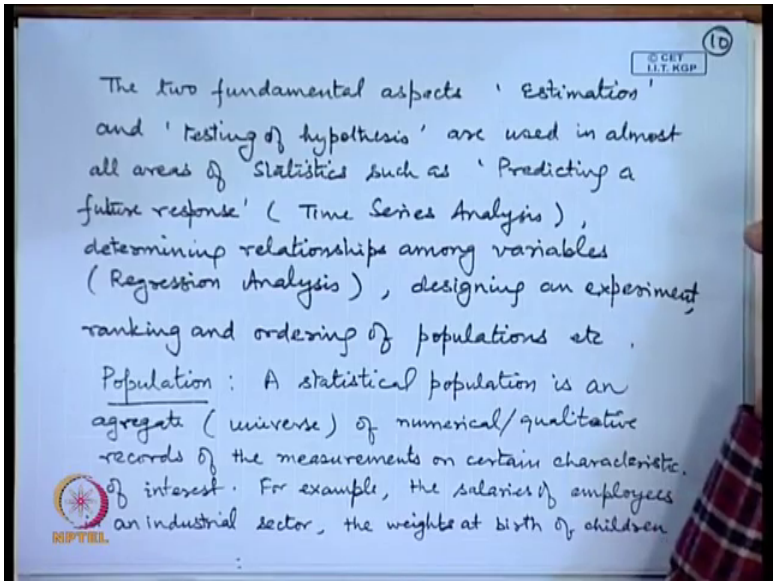
So, such situations are considered by statisticians over the years, and they have developed methods for estimation and testing etcetera, these are called popularly as nonparametric methods, or parameter free methods, or the distribution free methods. And this comes under the topic nonparametric inference. In this particular course, we will be spending almost all our time in discussing parametric inference.

So, by parametric inference then we refer to the problem, when the appropriate probability distribution has been is specified and the problem is now reduced to making inferences about the parameter or a function of the parameter in the form of estimation

which could be point estimation or interval estimation or testing of hypothesis. So, these two fundamental aspects that is estimation and testing of hypothesis, they are used in almost every area a statistical methodology.

For example, we consider predicting a future response I mentioned the problem of predicting the temperature for the forthcoming year, we would like to predict the average food production in the next year, we would like to predict the average industrial growth in the next year. So, these are the problems where the past data and certain other variables are used to predict the future thing. So, here this type of inferential problem is treated under the topic of time series analysis. Similarly, there are areas where we determine the relationship among the variables.

(Refer Slide Time: 05:31)



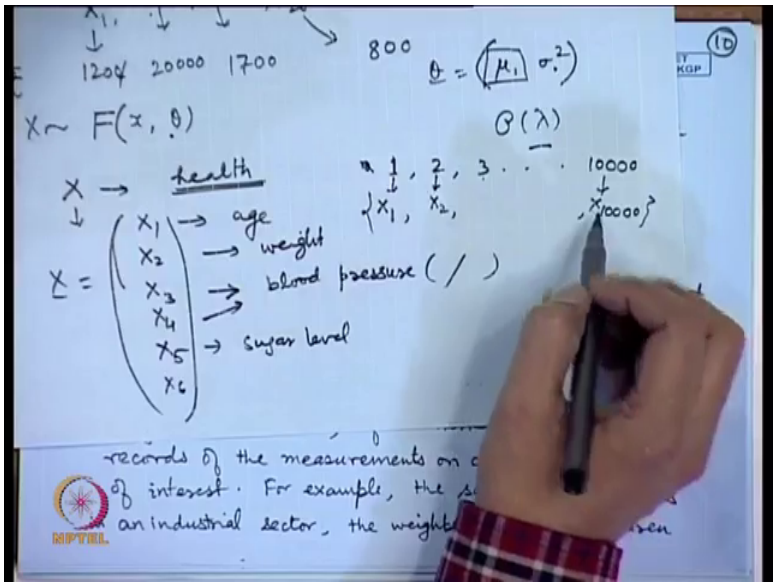
For example the effect of providing say irrigation, say modern equipment, good quality seed and say good quality of insecticides or pesticides etcetera to the farmers, and we look at the response in terms of the increased food production or increased yield of that particular crop. So, here the response variable is y that is the yield, and the variables which are determining this they are called regressor variables here x 1, x 2, etcetera, that could be the amount of irrigation facility, the amount of modern equipment, the modern fertilizers and other kind of things. This is this topic is generally covered under the subject regression analysis.

Designing of the experiments which is again used in the various industrial agricultural medical experiments ranking and ordering of populations etcetera. So, all of these advanced areas of statistical inference they use this fundamental aspects that is the estimation and the testing.

Now, at this stage I will introduce certain terminology and their exact meanings in the context of a statistical inference. The first important terminology is the term population which I have been using till now from the beginning of this lecture. So, a population in a layman terminology refers to a collection of individuals could be human beings, or it could be cattle, or it could be insects. So, generally a population refers to living beings that means the entities themselves. For example, a population; population of a country, population of say sheep in a population of a sheep in a state, population of say rats. So, we say that there are problems because the population of rats is increasing rapidly in a particular city or in a particular a state. However, a statistical population is not the collection of individuals or the units. it is the collection of the measurements or you can say aggregate of numerical or qualitative records of measurements on certain characteristics of interest.

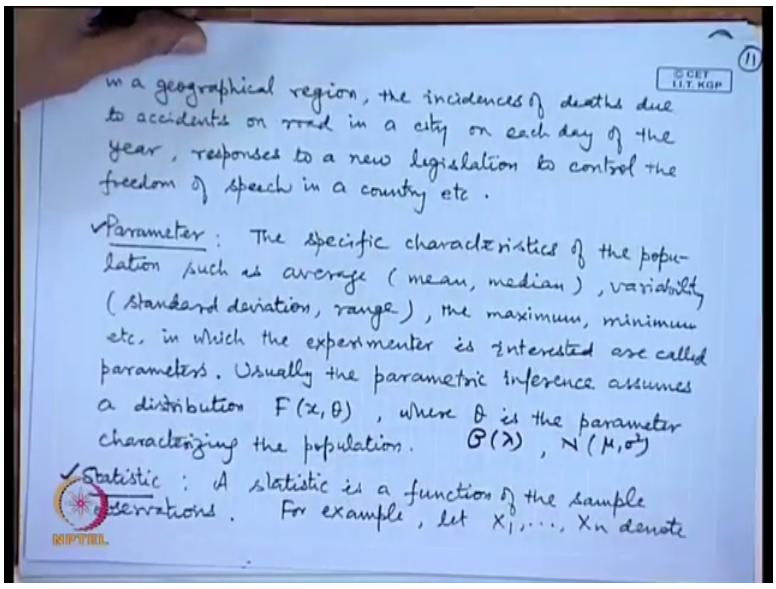
So, we looked at various problems just a while ago. So, we considered one problem of say estimating the average salary of the employees. So, here what would be the population, the population is the records against the salaries of the employees. So, suppose we are looking at an industrial organization. So, we may look at that all the employees which are employed in that particular industrial organization. And the so suppose there are 10,000 employees there, and we have them marked according to their employee code or any other identification code, then the values corresponding to their salaries.

(Refer Slide Time: 09:06)



So, for example, I am I identifying implies as 1, 2, 3 and so on up to 10,000. Now, the salary of the employee number 1 that is x 1 the same salary corresponding to the employee number two that will be called x 2 and so on so x 10,000. So, in this particular case, the population of interest is these 10,000 entries.

(Refer Slide Time: 09:41)



If we are looking at the weight at birth of the children in a certain geographical region; then, for all the children which are born during a particular period in a particular geographical region. So, we look at the value of the weights taken in say pounds or in

kilograms or in grams corresponding to all the children born. So, here the population is that aggregate.

If we are looking at the incidence and incidences of deaths due to accidents on the road in a city on each day of the year, then each day we record the number of accidents taking place and then the corresponding deaths in those accidents. So, the population here is the number of the deaths on each day. Responses to a new legislation to control the freedom of speech in a country, so a new legislation is placed in the parliament or it is proposed by say a by the cabinet.

Then say the opinion polls are taken whether it is a popular measure or not. So, the here the responses by the persons will be in the form of say they are whether they favor it or no not. So, it could be answers could be in the form of yes or no. So, the answers which are now here in the form of quality in a qualitative ones and that means, it is in the form of attribute that is also consisting creating or you can say this collection is my population in this particular problem. On the basis of this we may have to make the inference whether it is going to be a popular measure or not.

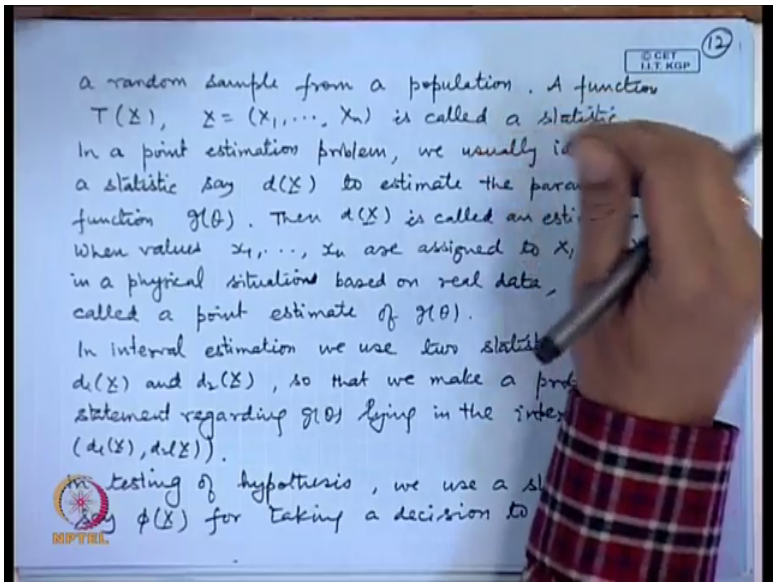
Once we have identified a population of our interest, the next key the key term is parameter. I have been using this term parameter repeatedly beforehand , but however what is the proper meaning of the parameter. So, by specific characteristics of the population such as average for example, it would be mean, median mode, arithmetic mean, harmonic mean, etcetera or a characteristic of this used for which determines the variability such as standard deviation range, suppose it is determining the whether the population is symmetric or not, maximum value minimum value etcetera

So, whatever the characteristics which in which the experimenter is interested in. So, the characteristic which are related to the population, they are called the parameters. So, usually the parametric inference assumes a distribution F x, theta. So, here theta is the parameter which characterizes the population. So, the popular examples like we say Poisson lambda distribution, so lambda is the parameter.

Here if I say normal mu sigma square distribution, so the distributional model is normal and it is characterized by the parameters mu and sigma square etcetera here mu and sigma square are the mean and variance respectively in the Poisson distribution lambda itself is the mean as well as the variance of the distribution.

A statistic so this is the next terminology. A statistic is a function of the sample observations. So, from the population the statistician has at his disposal a random sample on the basis of this which he will make the appropriate inferences. So, the sample is termed as observations x 1, x 2, x n.

(Refer Slide Time: 13:32)

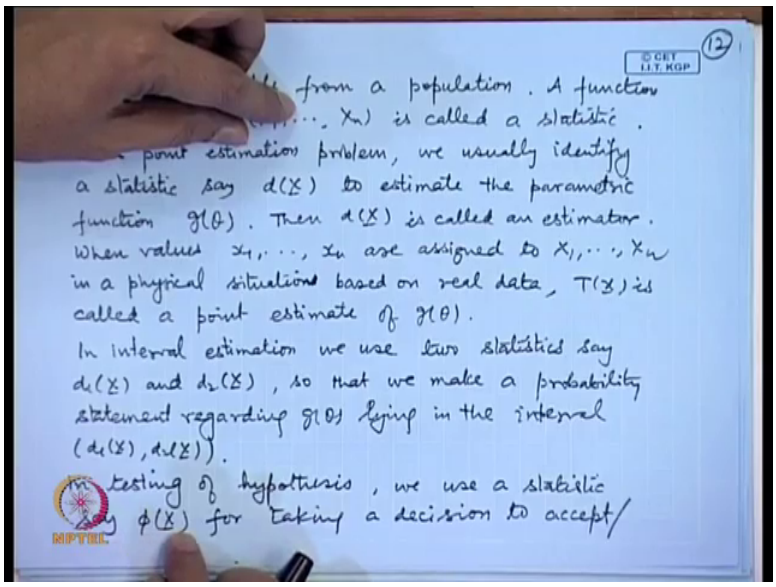


So, any function of these observations let us call a T T x where x is denoting the sample x 1 x 2 x n this is called a statistic. So, in a point estimation problem we usually identify a statistic let us call it say dx, this is called an estimator of the parametric function g theta. So, for example, in the suppose we have a Poisson model, and we are having the rate lambda. And we are interested to estimate say 1 by lambda. So, my parametric function is 1 by lambda.

So, now it is a (Refer Time: 14:09) question whether we can find out an estimate for 1 by lambda or we may be interested to estimate say lambda to the power 3. We may be interested to estimate in a normal distribution say mu. We may be interested to estimate sigma I square we may be interested to estimate sigma and we may be interested to estimate mu plus say p sigma which is denoting a quantile. So, depending upon the interest of the enquirer or the experimenter, one needs to determine which parameter is to be estimated or inference on which parameter is to be made, and the corresponding a statistic has to be frame from the sample which will be useful for the purpose.

So, for example, in the normal distribution, one may use sample mean to estimate mu, one may use sample variance say 1 by n sigma x I minus x bar whole square to estimate sigma square or 1 may use 1 by n sigma x I minus x bar whole square to estimate sigma square. In a interval estimation problem in place of one statistic say in this point estimation we are proposing 1 that is d x, but in interval estimation we need 2 that is endpoints of the interval where my parameter of interest is supposed to lie. So, we need to specify say d 1 x and d 2 x so that we can make a probability statement regarding the parametric function g theta lying in the interval d 1 to d 2.

(Refer Slide Time: 15:37)

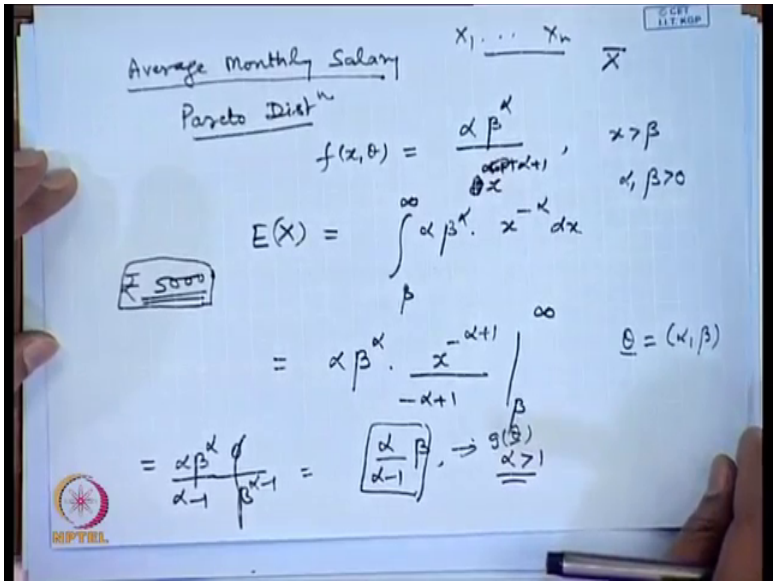


In testing of hypothesis, we use a statistic let us call it is a phi x for a taking a decision to accept or reject a given hypothesis. In this case phi x is termed as the test function or test a statistic. So, these are the basic terminologies which are to be used in statistical inference. We have a population, so that is the first thing that where we are interested what is our interest to a study in the given setup. So, we identify the population. We draw a random sample from the given population.

Now, drawing of a random sample itself is a matter of full investigation; it comes under the topic of methods of sample surveys or methods sampling techniques. And it is another aspect of the a statistical methodology where we discussed various methods of taking of random sample in this particular case we assume that a random sample is already available to us.

Now, our job is to use this random sample to draw appropriate inferences in the form of point estimation, interval estimation or testing of hypothesis to inform the end user about the appropriate conclusion of for about the population parameters. So, parameters are the characteristic of the population in which we are interested in. The decision is based on the random sample and for that purpose we use a function which is called a statistic. So, in the point estimation problem, we will create a point estimator using the statistic. In an interval estimator, we will create an interval which is in the form of two statistics giving a range. In a testing of hypothesis problem, we will specify a test function or a test statistic using that random sample.

(Refer Slide Time: 18:10)



At this point let me briefly give example here. So, let us consider the problem of say average monthly salary of the employees in an organization. Now, let us assume that the model for this is described by say Pareto distribution. So, a Pareto distribution may be having a is a continuous distribution. The density function is of a given form say alpha beta to the power alpha divided by theta to the power say alpha plus 1 sorry x to the power alpha plus 1, where x is greater than beta. So, in this particular case, we have considered a two parameter model where the parameters are alpha and beta both are of course, positive.

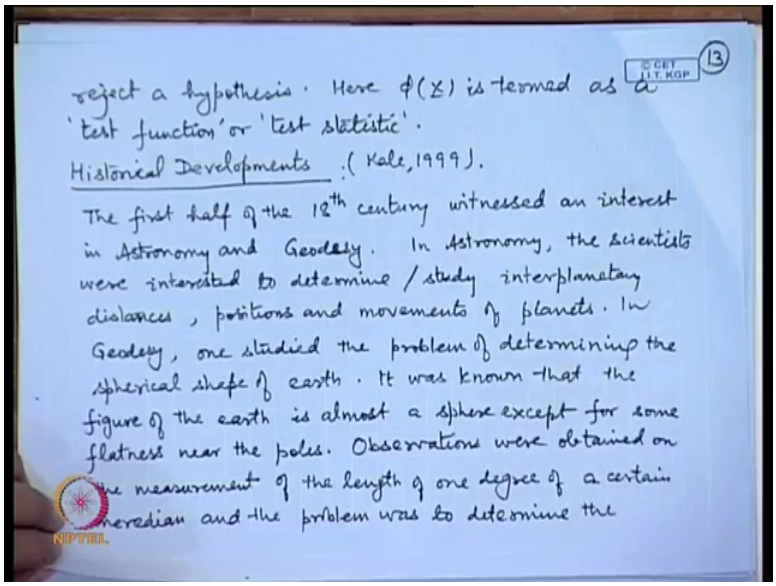
Now, here we may be interested in the average monthly salary. So, average monthly salary denotes expectation of x that means, from this distribution what is the value of the

expectation of x which can be of course, easily calculated. So, this value turns out to be alpha beta to the power alpha x to the power minus alpha plus 1 divided by minus alpha plus 1 from beta to infinity so alpha beta to the power alpha. And then when we substitute the value at infinity this will vanish and at beta this will become. So, we will get beta to the power alpha minus 1. So, the value turns out to be alpha by alpha minus 1 beta where of course, alpha has to be greater than 1 otherwise this expression will not be valid.

So, now in this particular problem, we want to estimate this parametric function. So, this is my g theta here, theta is a vector parameter consisting of two components alpha beta. So, now, to estimate this now there may be different procedures as a layman one may say that take the random sample x 1, x 2, x n and we may use x bar that is the sample mean to estimate this. So, this could be one method. And of course, depending upon the situation one may develop the different methods as we will be seeing during the course of during this course.

On the other hand one may have to do some sort of testing here one we may like to check whether the average income levels are low or high. So, for low or high we may identify a control. We may say that if the average monthly income is more than say 5000 rupees then we may say that they are well off or well paid. And in that case we may device a test statistic based on x 1, x 2, x n, to take a decision whether this hypothesis is tenable or not that means, we want to have a hypothesis whether the average monthly salary is more than 5000 dollar.

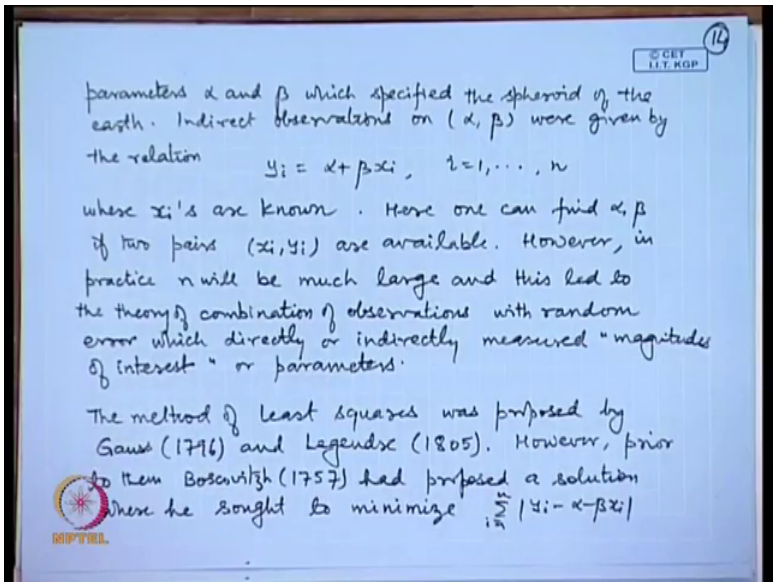
(Refer Slide Time: 21:57)



I will spend a few minutes on the historical developments of the subject. So, the historical development of the subject of statistical inference we can attribute towards the first half of the 18th century. And mostly in the problems of astronomy and geodesy, so in astronomy the interest was to find out the interplanetary difference distances the positions of the various planets or stars and their movements, in geodesy one wanted to find out the spherical shape of the earth. So, it was known that actually the earth shape is spherical, but it is flat on the near the poles.

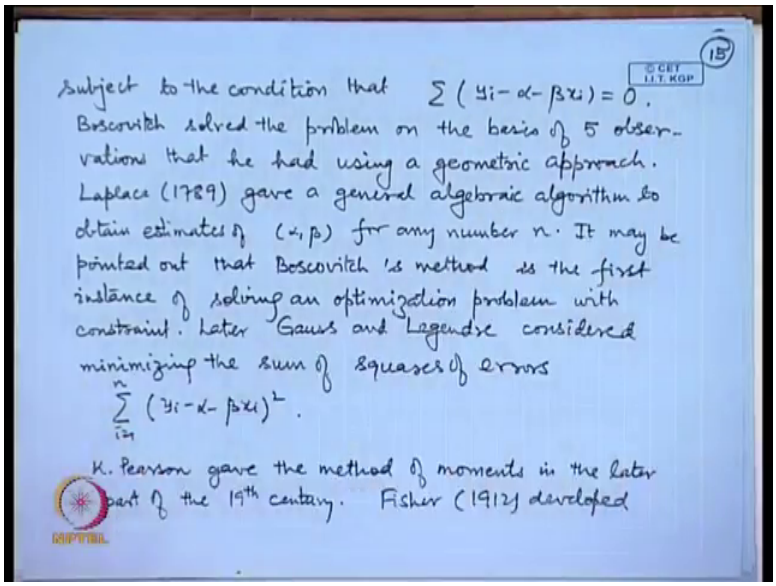
So, a standard technique is to take observations not one, but several measurements are taken. For example, they are taken about the length of one degree of a certain meridian.

(Refer Slide Time: 22:50)



And the problem is to determine parameters alpha and beta which is specified the spheroid of the earth. So, indirect observations on alpha beta are given by the relation y i is equal to alpha plus beta x i. So, here x i's are given to us y i's given to us. So, alpha and beta are to be estimated. So, nowadays we understand this as a problem of linear simple linear regression. However, this problem was is studied as early as in 18th century by Gauss and Legendre who came up with the method of least squares to solve this problem. Even before Gauss and Legendre about 50 years before that Boscovitzh in 1757, he proposed a solution for this problem very sought to minimize summation of modulus y i minus alpha minus beta x i.

(Refer Slide Time: 23:52)

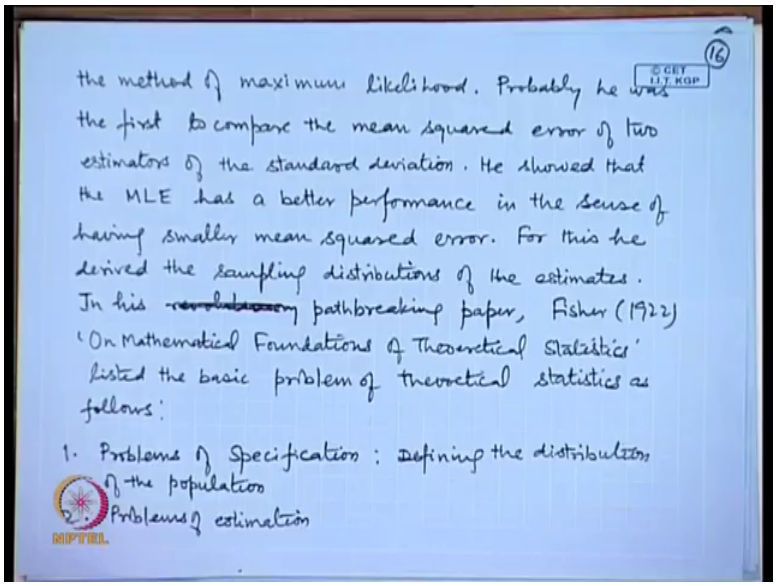


So, in place of this square, he initially considered the mean absolute error actually subject to the condition that some of the errors must be 0. And he solved this problem using geometry geometrical methods based on five observations. Later on Laplace has given a general algebraic solution to this problem. This can be considered as the first you can say attempt to solve an optimization problem under constraints. Later on Gauss and Legendre considered the minimization of the sum of a squares and that is why it came to be known as the method of least squares.

So, you can consider the problem of statistical inference or you can say the modern statistical inference is started as early years in the 18th century. Further developments or you can say the further techniques started to get developed towards latter half of the 19th century for example, Francis Galton, he started to study something called the relationship between the variables and he called it regression. So, he wanted to he sort of predicted that the tall parents have tall children, but less tall than the parents; and shorter parents have short children in the height, but taller than the parents. So, this was called regression towards normality of the heights. And the first studies are you can say first model of the simple in regression we are made in this thing.

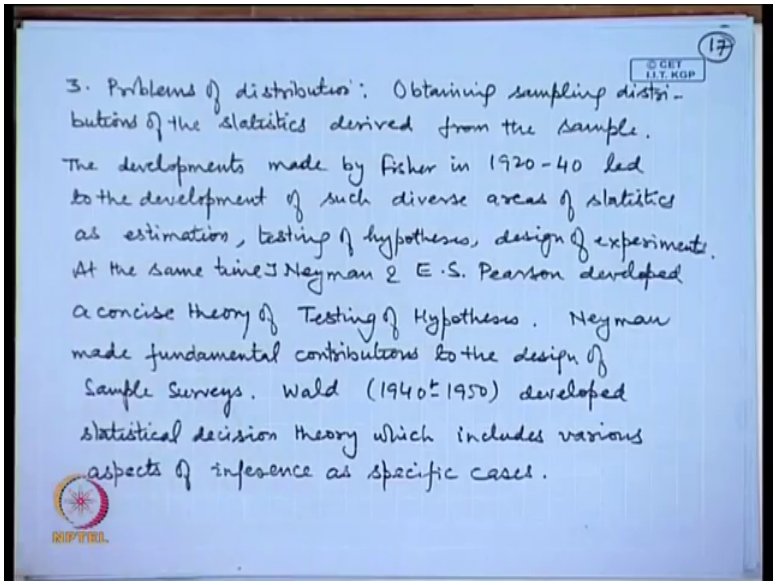
Later on Karl Pearson developed the method of moments in the latter part of 19th century. The modern methods of statistics as we know today and probably they were first started by Fisher in 1912, where he developed the method of maximum likelihood.

(Refer Slide Time: 25:51)



He is probably the first one when he realized the importance of comparing two different methods of estimation. So, he considered two estimates of the standard deviation. He found out the sampling distributions of that and therefore the mean squared and he showed that one of the estimators has a smaller mean squared error than the other. So, probably this is the fundamental or you can say path breaking paper in 1922 that is called the mathematical foundations of theoretical statistics. Here he listed the basic problem of theoretical statistics as, firstly the problem of a specification that is defining the distribution of the population; second is the problem of estimation.

(Refer Slide Time: 26:57)



And the third is the problems of distribution that means, how to judge the goodness of the or you can say evaluate the performance of the estimators we need the sampling distributions of the sampling distributions of the statistic which are being used. So, these developments made by fisher in 1920 to 40. And had these had the effects in the various areas of statistics such as estimation testing of hypothesis, designs of experiments at the same time Jerzy Neyman and E S Pearson simultaneously developed a theory of Testing of Hypothesis. Neyman also developed the theory of Sample Surveys.

Later on in 1940s Abraham Wald developed a topic called Statistical Decision Theory, and this includes various aspects of inference as special cases. He showed in fact that estimation testing ranking and selection procedures they are all part of the general problem of decision theory which is actually having it is origin in the theory of games which was developed in 1930s or 1920s by John von-Neumann among others.

So, friends, today we have discussed the basic problem of a statistical inference its main components. In this particular case, we will focus on the problem of estimation and testing of hypothesis. So, from the next class onwards, I will start discussion on the problem of point estimation.