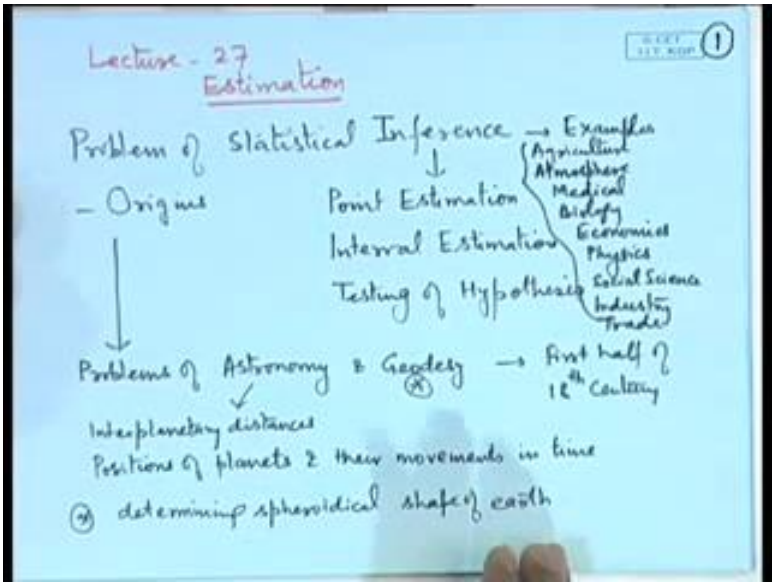
Probability and Statistics
Prof. Somesh Kumar
Department of Mathematics
Indian Institute of Technology, Kharagpur

Lecture - 53
Introduction to Estimation

Today I will introduce the problem of a statistical inference. So, far we have concentrated on discussing the concepts of probability, the concepts of statistical distributions, various kind of discrete and continuous distributions, multi variate random variables and their distributions and we also looked at the concept of sampling distributions.

(Refer Slide Time: 00:47)



Further I described something called descriptive statistics; that means, when a data is given, then how do we plan to analyze it; that means, how to present that data graphically or we draw certain basic say basic characteristics such as measures of central tendency, measures of dispersion or variability from that data; however, all of this is actually to be utilized for drawing inference on populations.

So, what is the problem of inference? So, for example, a government is interested that how much will be the average wheat production in the coming year, how much will be say the production of sugar in the country, how much will be the production of a particular commodity, how much will be the production of say cotton, how many farmers

or what percentage of land is utilized for farming of a fruits. In atmospheric sciences scientists are worried about what is the average temperature likely to be in the month of January or in the year 2010 is it going to be more than the year 2009.

In medical sciences we are interested about the occurrences of diseases. So, what is a estimated number of people who will be affected by a certain kind of disease, and what will be the effect on the longevity of the people by that disease in biology, in economics, in physics, in social sciences, in industry, trade and commerce; in almost every area of human activity we come across such situations or such problems.

Now, one may question that why do we have to use these statistical methods here. For example if I am looking at say occurrences of disease or say agricultural production then where is the statistical thing coming into picture or suppose I am measuring the say diameter of a star in universe in a faraway galaxy, then where is the statistics coming into picture. The statistics comes into picture that although we may feel that the diameter of a star is deterministic and one should be able to get an exact figure of it, but we do not have methods of getting that value exactly.

So, certain formula will be used and in that is formula certain ingredients will be there which will be measured by certain instruments repeatedly, now that measurements, process of taking measurements including say certain errors which we assume are random or a statistical in nature and therefore, when we draw any inference based on those measurements, the inference becomes a statistical inference. And therefore, this entire topic or you can say the entire subject then needs that we use correct methodology of a statistical inference, so that the conclusions drawn from that data are correct. So that brings onto the focus the problem of a statistical inference.

Primarily speaking the problem of inference can be divided into 2 portions: one is called the problem of estimation and another is the problem of testing of hypothesis. For example, if we want to actually get a value that what is a average of longevity of the people of India or people of a particular country, then we actually do not know the value of what we want and therefore, we actually get a value based on a sample. So, this is called the problem of estimation; that means, to get the value.

Now, that estimation itself can be split into 2 parts: one is to get an actual value suppose I said a value of every longevity is average age is 65 years then we are assigning a single

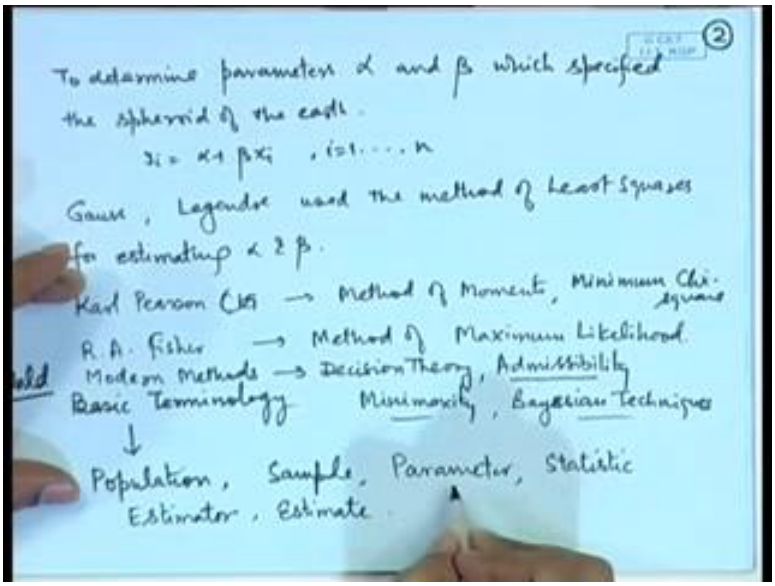
value for the characteristic to be estimated, this is called the problem of point estimation. On the other hand we may not give the exact value, but we may give an interval of the values and say that with a certain confidence or certain probability the given value lies in that interval.

For example we may say that the average age of a person in India is from 62 to 68 years with 95 percent of confidence, this is called the problem of interval estimation or confidence intervals. On the other hand sometimes we would like to test a fact for example; a new drug has been introduced in the market for treating a certain disease, now the manufacturing company which has introduced the drug will certainly like to know that whether the new medicine is more effective than the previous one.

So, if I say p 1 is the proportion of people which were treated earlier and p 2 is a proportion of the persons which are treated now successfully, then whether p 2 is bigger than p 1 this type of judgment. That means, to tell on the basis of the sample whether p 2 is bigger than p 1 or p 2 is less than p 1 etcetera this is called the problem of testing of hypothesis. So, probably we divide the problem of a statistical inference into 3 parts: 1 is the problem of point estimation, another one is the problem of interval estimation and another is the problem of testing of hypothesis. There are various other facets of a statistical inference like prediction, sequential inference and other things, but they can be considered to be follow up from here. So, these are the basic you can say facts of the, or basic parts of the statistical inference.

Let me give some historical facts about how the problem of a statistical inference was initially studied. So, it seems to have been have origins in the problems of astronomy, and geodesy, in the first of half of 18th century then many scientists were finding out like distances between the stars. That is interplanetary distance the positions of the stars, their shapes; how do they move with the time; that means, for example, mercury takes this much time to rotate around the sun how it takes this much time to rotate on it is axis and all those kind of statements; that means, the problems in astronomy are in geodesy

(Refer Slide Time: 08:11)



So, for example, some of the earliest measurements were made on to check whether the spherical shape of earth to determine that thing and so it turned out that the data is of the form that we have observations x 1, x 2, x n and y 1, y 2, y n and they are related with the equation y i is equal to alpha plus beta x i, which today we know as a equation of a simple linear regression model. So, these are earliest occurrences of this model. So, the famous mathematician's Gauss and Legendre they used the method of least squares for finding out the value of alpha and beta.

So, you can say that the method of least squares is probably one of the oldest methods for finding out the estimates of parameters. Towards the end of 19th century Karl Pearson introduced the method of moments and minimum chi square for estimating parameters. In the beginning of the 20th century R.A Fisher he introduced the method of maximum likelihood. In fact, as I already mentioned he is credited to be you can say the initiator of most of the methods of modern statistical inference which we use today. So, he was the 1 who introduced the concept of maximum likelihood estimation. In the mid-20th century Abraham Wald introduced the some decision theoretic methods and methodology such as admissibility, Minimaxity and Bayesian techniques in a statistical inference.

Now let me introduce the basic terminology to be used in a statistical inference: the first term is a population, so a statistical population is a collection of measurements in which

we are interested. For example, we are interested in estimating the average per capita income of persons in a state, then there may be a household survey or there may be a survey of people in different organizations and the incomes of individuals are noted.

So, in this particular case the statistical population is the measurement corresponding to the individual incomes. If you are interested in the average longevity of persons, then suppose we are considering a particular state or a particular country then the total life span of each person of that country or that state will constitute the statistical population. If we are interested to study the yield of wheat in the state of Punjab, then corresponding to each plot of land where the wheat is grown, if we look at the total output or yield of the wheat from each of the plot, then those values will be considered the statistical population for this purpose.

So, a statistical population is a collection of measurements with respect to certain characteristic which we are interested to study. Here one thing I would like to mention, but it is not necessary that all the time we will have to look at only numerical values, sometimes it may be in the form of yes no or some answers which we can call attribute data.

For example if we are looking at preferences of people for a certain opinion whether they have a positive opinion about certain issue. So, they may answer in yes or no. So, corresponding to each person you will be noting down the data yes or no and you may put it as value say 0 or 1. You may record the persons who are say possessing a certain characteristic say an IQ greater than 100 or below 100, persons who was average incomes are above say a particular level or below a particular level or we may classify them in according to 4 different levels very poor, lower middle class, upper middle class and say higher income group.

So, we may assign for each person or each household according to the level of income that the person is having the value say 1, 2, 3, 4 or 0, 1, 2, 3 etcetera. So, this is qualitative data or attributes data and in a statistical population one also studies such data. Next is sample; so what is a sample? The exact definition of sample is that sample is a subset of population.

So, in general when we want to study any characteristic about the population, it is requiring the entire measurement; that means, complete enumeration of the population,

which is not visible. So, for example, if we are studying say per household expenditure on say medical expenses in a particular town then it will require to go to each household, and get the monthly expenditure on the medical; however, this may not be feasible. So, the best solution for various such enumeration problems is to take a representative sample from the population and draw the inferences based on that.

So, the concept of sampling techniques or sampling methodology is widely developed in statistics. So, here we assume that the sample has already been selected and we will draw the inferences based on that. So, sample critically speaking is a subset of the population and we will assume that it has been randomly selected.

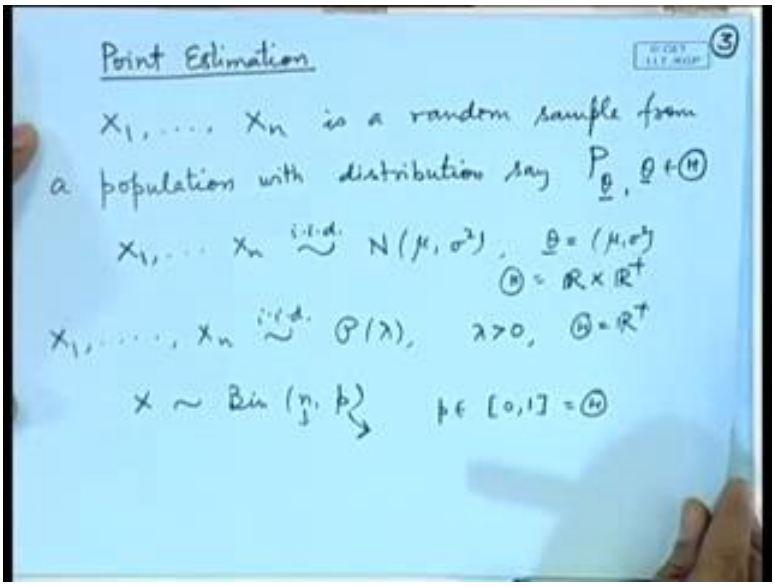
A parameter of a population is the characteristic in which we may be interested in. So, for example when we talk about the population of say incomes, then we may be interested to know the range for example, what is the difference between the maximum salaried employee and the lowest salaried employee. If we are interested in the say yields of different states for say wheat then per hectare wheat production maybe in a particular state is much higher corresponding as compared with the other one.

So, we may be looking at the averages, the maximum value, the minimum value, the variability, the medium value. So, these characteristics of the population they are termed as parameters. So, since we are interest to know about the characteristics of the populations that means parameters, the statistical inference problem relates to either finding out an estimate or you can say point estimator or an interval estimator for the parameter or to test about those parameter values.

Now, at our disposal we have a random sample say x 1, x 2, x n. Now whatever we want to draw our inference from x 1, x 2, X n we will be using certain function of that. So, for example, if I say I wanted to use find out average height and from the sample I take the average and I use it as an estimate so; that means, I have used a function of the sample observations, so these sample observations if we make a function out of that that is called a statistic. So, the statistic will have different uses for example, I can use them to make a point estimator, I can use them to make a confidence interval, we can use them to create a test. So, we may use it as a test statistic. So, when I use it as to estimate certain parametric function then it is called an estimator and the realized value of that is known as an estimate.

So, now let me introduce the basic features of estimation.

(Refer Slide Time: 17:29)



So, let me concentrate on the problem of point estimation; now to begin with I mentioned that there are several mathematicians or statisticians who gave some methods of estimation, for examples I mentioned the word least squares estimates, the method of maximum likelihood, the method of movements, the minimum chi square method etcetera.

So, each of these methods is based on certain concept or you can say certain theory that why this is desirable method; now the question is that they may give different values of the estimators or they may give the same values of the estimators, then the question comes that how do you distinguish that which 1 should be used? So, for that purpose we introduce certain criteria of estimation. So, before going to give the actual methods of estimation let me introduce certain criteria.

So, in any statistical inference problem the model is like this that we have X 1, X 2, X n is a random sample from a population with distribution say P theta, theta belonging to say a script theta. Let me explain this normally we will be talking about sentences such as X 1, X 2, X n is a random sample from Poisson lambda distribution; X 1, X 2, X n is a random sample from normal mu sigma square distribution; so, what is the meaning of this?

The meaning of this is that in the inference problem we assume that the determination of their statistical model has already been done; that means, the problem is already specified. For example, if I am saying it is estimation of say average longevity, the estimation of average temperatures etcetera the problem has already been identified by the person who is going to use it, it may be a government agency, it may be a commercial organization etcetera and then their statistician has already determined the parametric model for that. That means, if we are talking about average heights then their statistician has determined that this population follows a normal distribution; that means, if we have a large data set from the our target group and we have taken the heights and then we make a histogram and a frequency curve and we find that it looks like a normally distributed random variable.

Therefore the problem to for inference is now to draw certain inference on the parameters of the population; that means, what could be the value of mu, whether mu is equal to 0 or mu is less than or equal to a certain value, whether sigma square is a known value or unknown value etcetera. That means, we are going to do a testing or confidence interval or point estimation about the parameters of the population; that means, when I am saying point estimation or testing etcetera we are talking about parametric inference.

So, there are 2 types of inferences: we have parametric inference and non parametric inference. So, where the non-parametric inference will arise? When we are unable to determine the model from which the data has come from; that means, we may not be able to say that it is normally distributed. So, this could be in several ways for example, the data is too f hazard or the data is too less or we are not having sufficient experience to determine the data is coming from which population.

Then there are certain methods which we call distribution free methods of non parametric inference, in this particular topic we will be restricting our redemption to parametric inference; that means, we assume that the model is coming from a certain population P theta. So, P we already know that what distribution it could be only thing to be determined is that parameters we may not know. So, by statements such as X 1, X 2, X n is a random sample from normal mu sigma square.

So, we go back to our terminology which we used in the distribution theory that we write that X 1, X 2 X n are independent and are identically distributed random variables with

normal mu sigma square distribution. So, here what is theta? Theta is equal to mu sigma square; now what is this script theta? It is a set of all possible values of the parameter for example, if I am saying normal mu sigma square then mu varies from minus infinity to infinity and sigma square is positive; that means, here theta is your R that is a real line cross R plus that is the positive half of the real line.

We may say X 1, X 2 X n follow Poisson lambda distribution. So, here lambda is a positive parameter therefore, my parameter displace is R plus. If I say X follows a binomial n p distribution, I know what is n because I know in how many trials I am looking it at for the number of successors so the parameter could be p, and we may say that p belongs to the interval say 0 to 1. So, this is the parameter displace in this situation.

So, depending upon the different parametric model the distribution and the parameter displace will be specified. So, in any inference problem we start with this model that we have a random sample from a given population. So, the meaning of that is that we have identically and independently distributed random variables from a given population and our objective is to make certain inference about the parameters of the population in the form of point estimation, interval estimation or confidence interval. So, now, for the time being we restrict attention to the problem of point estimation.