Statistical Inference

Prof. Somesh Kumar

Department of Mathematics

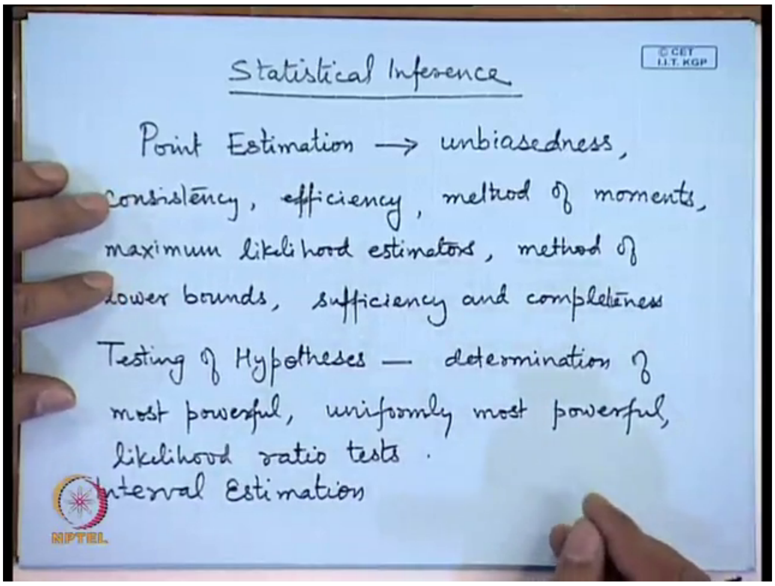
Indian Institute of Technology, Kharagpur

Lecture No. # 01

Introduction & Motivation

Friends, in this course we will cover important topics of statistical influence, 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.

(Refer Slide Time: 00:51)



Statistical Inference

Point Estimation → unbiasedness, consistency, efficiency, method of moments, maximum likelihood estimators, method of lower bounds, sufficiency and completeness

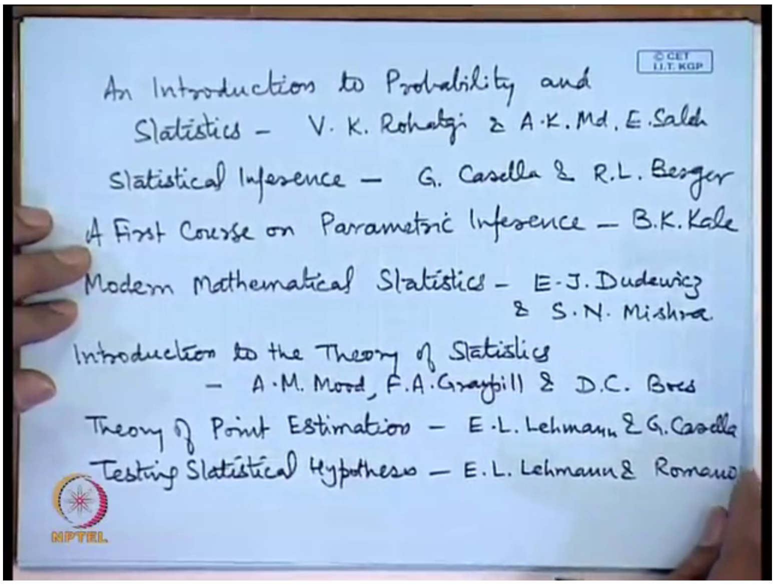
Testing of Hypotheses — determination of most powerful, uniformly most powerful, likelihood ratio tests.

Interval Estimation

We will be mainly covering in this course, point estimation, in the point estimation we will cover the fundamental concepts, such as unbiasedness, consistency, efficiency,aAnd 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 and biased estimation, the method of lower bounds for determining this method of lower bounds, and another concept is that through sufficiency and completeness.

We will cover another major area in 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 tests we will discuss most powerful, uniformly most powerful tests, and then related concepts 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 estimation problems.

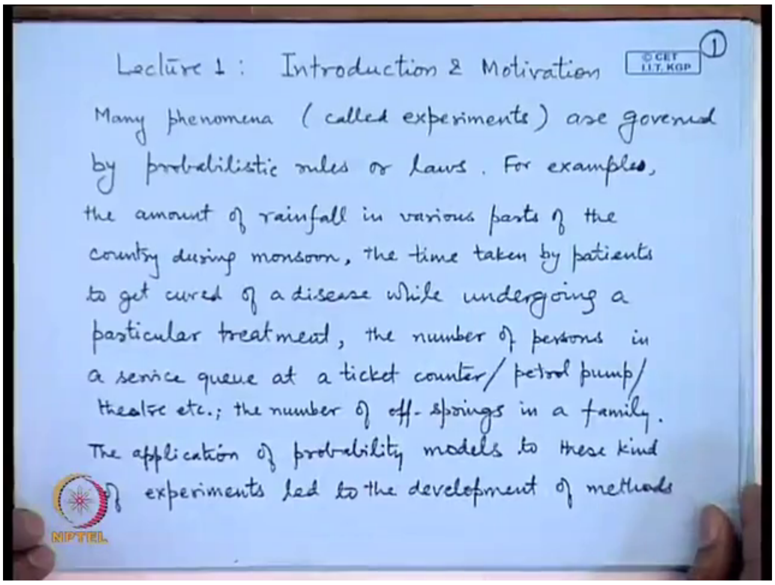
(Refer Slide Time: 03:27)



An Introduction to Probability and
Statistics – V. K. Rohatgi & A.K. Md. E. Saleh
Statistical Inference – G. Casella & R.L. Berger
A First Course on Parametric Inference – B.K. Kale
Modern Mathematical Statistics – E.J. Dudewicz
& S.N. Mishra.
Introduction to the Theory of Statistics
– A.M. Mood, F.A. Graybill & D.C. Boes
Theory of Point Estimation – E.L. Lehmann & G. Casella
Testing Statistical Hypothesis – E.L. Lehmann & Romano

The important texts books, that can be used for these course are, an introduction to probability and statistics by V.K. Rohatgi and A.K Md.E Salah another important book is statistical inference by G Casella and R.L Berger, a first course on parametric inference by B.K Kale. Modern mathematical 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 Cassella and testing 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 introduce, what is the problem of a statistical inference and why should we study it.

(Refer Slide Time: 06:29)



Lecture 1 : Introduction & Motivation

© CET I.I.T. KGP

1

Many phenomena (called experiments) are governed by probabilistic rules or laws. For examples, the amount of rainfall in various parts of the country during monsoon, the time taken by patients to get cured of a disease while undergoing a particular treatment, the number of persons in a service queue at a ticket counter/ petrol pump/ theatre etc.; the number of off-springs in a family. The application of probability models to these kind of experiments led to the development of methods

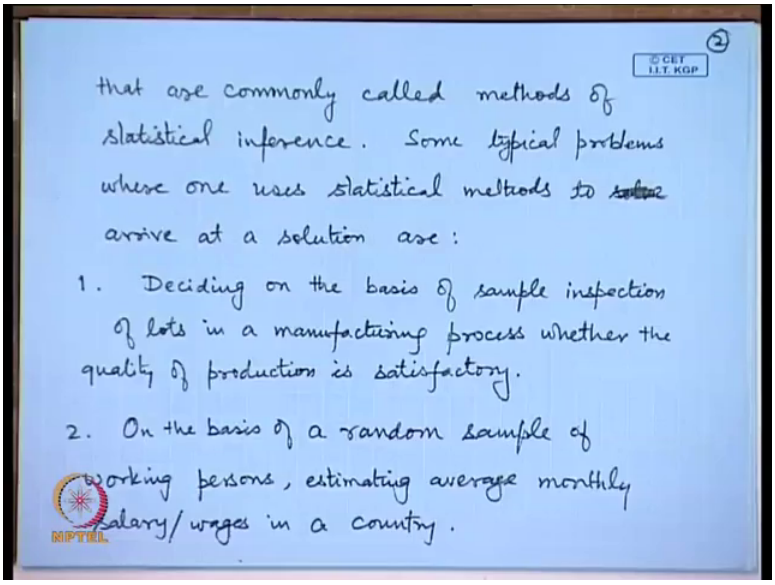
So, let me talk about the introduction and motivation of statistical inference, we noticed that many phenomena, in various signs says they are governed by loss, which are not deterministic in nature. So, we can call then stochastic or probabilistic in nature. For example, if you look at the amount of 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 defer from the previous year, whether over the entire geographical region. In which we are interested in whether there will be uniformed distribution of the rainfall or in some portions, there will be tremendous rainfall and in some other places, there will be conditions of drought.

No matter, what physical theory we develop or what atmospheric scientist are able to develop 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, quiet 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 two 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 quiet subjective in nature, they depend upon various conditions therefore, these are stochastic in nature. Similar examples are the number of persons in a service q at a ticket counter or at a petrol pump. So, for example, if we go to a railway counter at a given time of a day, we find that at that time, there are a large number of people standing at 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 a 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 are the number of persons standing in the queue is much less.

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

(Refer Slide Time: 09:45)



that are commonly called methods of statistical inference. Some typical problems where one uses statistical methods to ~~solve~~ arrive at a solution are:

1. Deciding on the basis of sample inspection of lots in a manufacturing process whether the quality of production is satisfactory.
2. On the basis of a random sample of working persons, estimating average monthly salary/wages in a country.

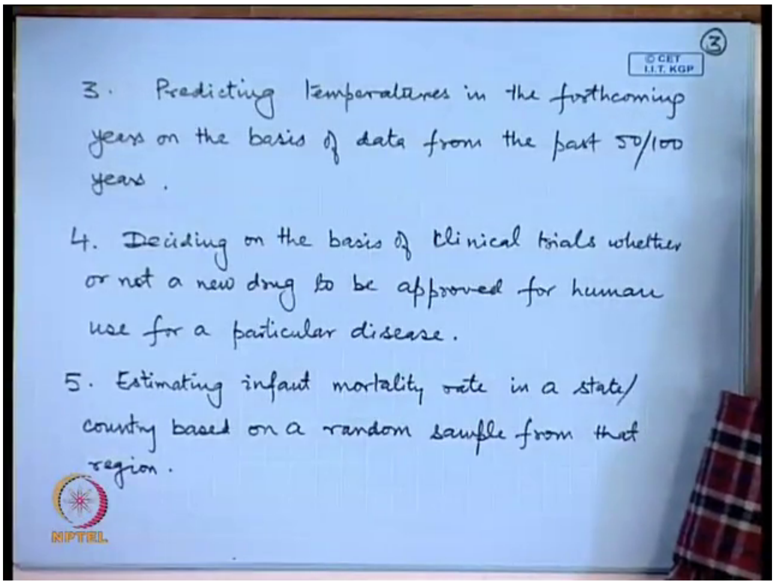
Some typical problems, where one uses is statistical methods to arrive at a solution or deciding on the basis of sample inspection of lots in a manufacturing process, whether the quality of product is satisfactory. So, you consider a factory, where certain kinds of nuts and bolts are being produce. Now, 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 it will be sent to the market at a appropriate price. And 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. Now, how do we goes about it, he looks at the batches of the product for example, maybe in a 1000 or in a 100. And he inspects 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 observes that every lot of 10, it does not produce more than one defective product. Then he may conclude that production of the bolts is satisfactory or up to the giving marks.

Another problem of inference could be 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 inference is important for determining economic policies of the government, it is important for the say consumer goods producing companies, 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 than high caste goods, may not be able to be sold in the market in those places. 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 past 100 years. So, on the basis of these we will be able to estimate, how much will be the average temperature 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.

(Refer Slide Time: 13:13)



3. Predicting temperatures in the forthcoming years on the basis of data from the past 50/100 years.

4. Deciding on the basis of clinical trials whether or not a new drug to be approved for human use for a particular disease.

5. Estimating infant mortality rate in a state/country based on a random sample from that region.

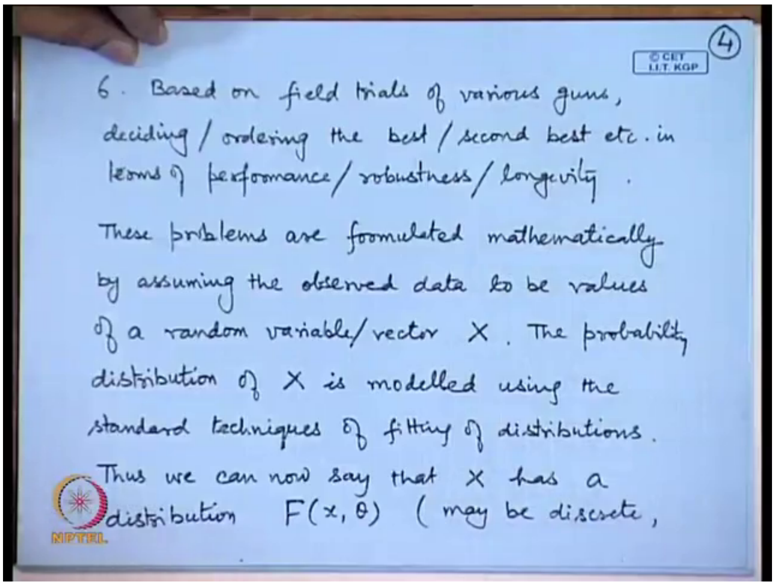
Another typical example of statistical inference is deciding on the basis of clinical trials, whether or not a new drug is to approved for human use for a particular diseases. So, there is a diseases for which certain drug may be or may not be available. Now, doctors are the persons, who are involved in the development of the medicines, they come up with certain substances, certain bio chemical substances 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 the a certain medicine is produced using that, using certain amounts of that clinical, that biochemical substances.

And now once it is decided or the amount is decided or amount is determine that, this is actually going to be 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. Even, if the medicines are not taken, if some people are getting cured then this medicines should be better than that. So, it is the job of the statistician on the basis of a random sample he will decide, how to take a decision in arriving at the 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 G D P 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, 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 notice, find out what is the infant mortality rate in that country, if the infant mortality rate despite having high G D P or high average salaries, even if it is even then, if it is high then that means, it relates to certain other kind of traditions, certain other kind of conditions, which may exists there which are not, which despite high G D P growth are not going to improve infant mortality rate etcetera. So, a social scientist and the planners of the country are interested in knowing, the what is a estimation or what is a estimate of the infant mortality rate.

(Refer Slide Time: 16:20)



4

© CET I.I.T. KGP

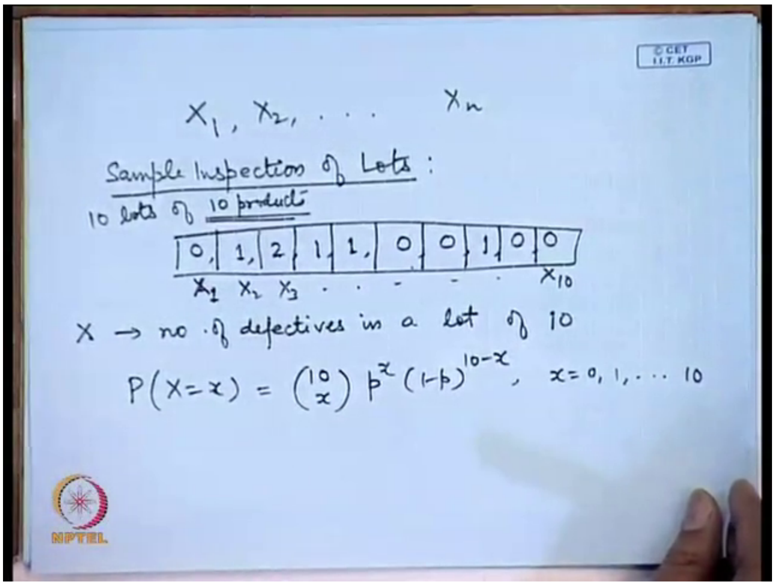
6. Based on field trials of various guns, deciding / ordering the best / second best etc. in terms of performance / robustness / longevity.

These problems are formulated mathematically by assuming the observed data to be values of a random variable/vector X. The probability distribution of X is modelled using the standard techniques of fitting of distributions. Thus we can now say that X has a distribution $F(x, \theta)$ (may be discrete,

A frequently encountered problem is that a there are various kinds of guns, which can be used in the by the army now, now army wants to buy new guns to replenish it is stock of the arms. So, various arms manufactures, they give the samples from their factories. And then the field trials of those guns are conducted. So, the job of the statistician to determine that, which is the best gun among these, which is the second best etcetera in terms of its performance. Performance could be in terms of accuracy that or the range that, that the gun can cover in hitting the it is targets, it is robustness in the sense that depending upon the different terrain, whether it is hilly region or whether it is a plain 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 them in some preference. So, this 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, who is observed values are given to the statistician by the user, by the end user.

(Refer Slide Time: 18:40)



$X_1, X_2, \ldots \quad X_n$

Sample Inspection of Lots :

10 lots of 10 products

0,	1,	2,	1,	1,	0,	0,	1,	0	0
X_1	X_2	X_3	.	.	-	-	-	.	X_{10}

$X \rightarrow$ no. of defectives in a lot of 10

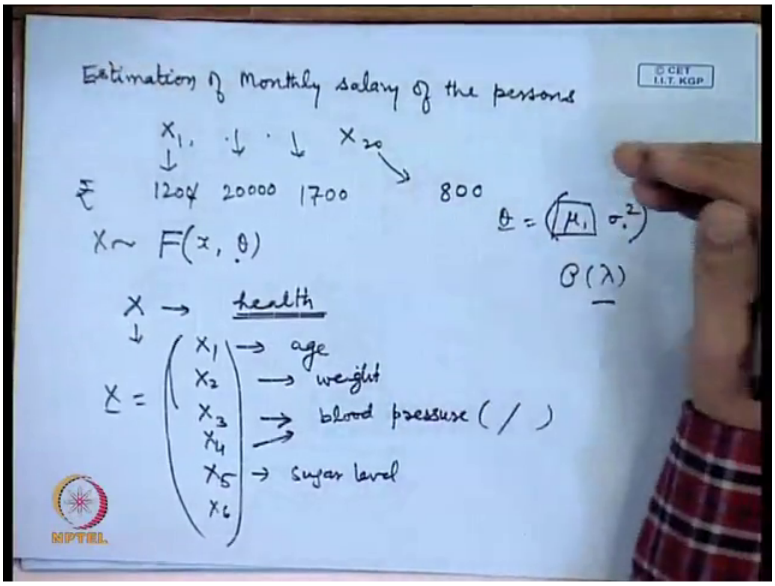
$$P(X=x) = \binom{10}{x} p^{x} (1-p)^{10-x}, \quad x = 0, 1, \ldots 10$$

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 bolts 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, we which is the quality control terminology used for describing the goodness of that product. So, we may use that now the data could go 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 have one defective, lot number 2 have two defective, lot number 4 had one defective, lot number 5 had one defective, lot number 6 and 7 had no defective, lot number 8 had one 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 statistician 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 tha,t 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 equals 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: 22:09)



Estimation of Monthly salary of the persons

$X_1, \ldots, X_{20}$

₹ 1204 20000 1700 800

$X \sim F(x, \underset{\sim}{\theta})$

$\underset{\sim}{\theta} = (\boxed{\mu_1}, \sigma_1^2)$

$\theta(\underline{\lambda})$

$X \rightarrow$ health

$$\underset{\sim}{X} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ X_6 \end{pmatrix}$$

$X_1 \rightarrow$ age

$X_2 \rightarrow$ weight

$X_3, X_4 \rightarrow$ blood pressure (/)

$X_5 \rightarrow$ sugar level

© CET I.I.T. KGP
NPTEL

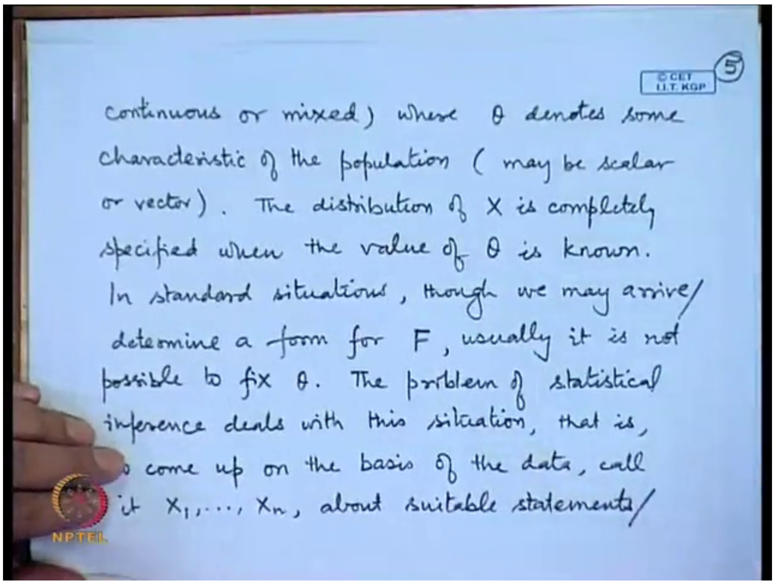
So, in this particular case, if we are looking at estimation of monthly salary of the persons, say employed persons say the values here would be calculated as X 1; X 1 say X 20. Suppose, 20 persons have been considered here then the values could be in terms of rupees say. So, you may have say for a person it could be 1200 rupees for another one it could be may be 20,000 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 a now we may use this data to have a model. So, for example, the model may be given by certain distributions say F it could be normal distribution, it could be Gama distribution etcetera, depending upon the actual values that have been obtained there.

So, the probability distribution of X is modeled using standard techniques of hitting of the distribution. And therefore, we can say that X has a distribution say F x theta, which of course, may be discreet or it could be continuous or it could be mixed. And theta denotes some characteristics of the population, which could be scalar or the vector, here X itself can be 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 discreet random variable, however there maybe 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 consists of certain components say X 1 which may relate to his age, X 2 may relate to his weight that is his body weight, X 3 may denote his blood pressure. Now, blood pressure may consist of two values. So, you have two 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 a monthly salary, if we 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 we are considering say the number of persons arriving in a q 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 read as the observed values of a random variable X and X will is assume, to have a known distribution effects theta. And therefore, the distribution consists of certain parameter, which is called characteristic or parameter of the distribution.

(Refer Slide Time: 26:39)



© CET I.I.T. KGP

5

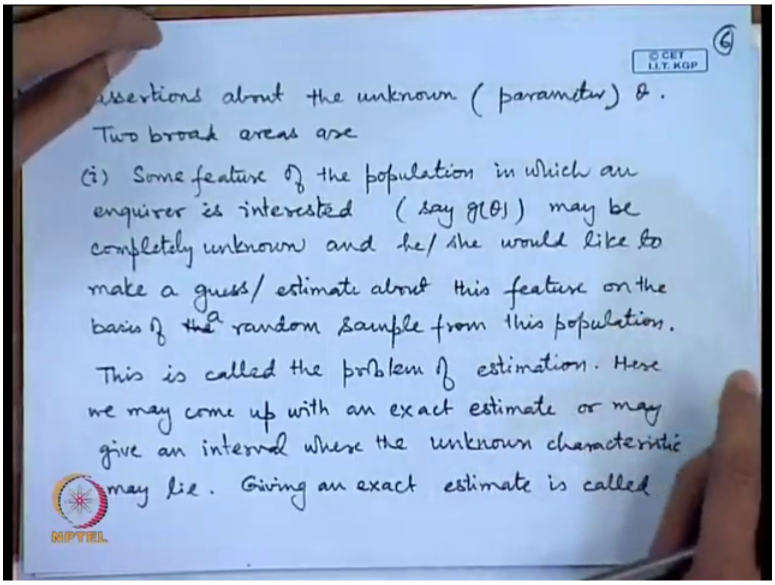
continuous or mixed) where θ denotes some characteristic of the population (may be scalar or vector) . The distribution of X is completely specified when the value of θ is known.

In standard situations, though we may arrive/ determine a form for F, usually it is not possible to fix θ. The problem of statistical inference deals with this situation, that is, o come up on the basis of the data, call it $X_1, \ldots, X_n$, about suitable statements/

So, the distribution of X is completely specified, when the value of theta is known. Now, in standard situations although we may arrive or determinate a 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 during the time period the number of failures. Similarly in some situations, we may arrive at a conclusion that binomial model is more useful or gamma distribution is more useful, but the appropriate parameters of that distribution, one may not know in advance. 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 are the problems 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: 27:51)

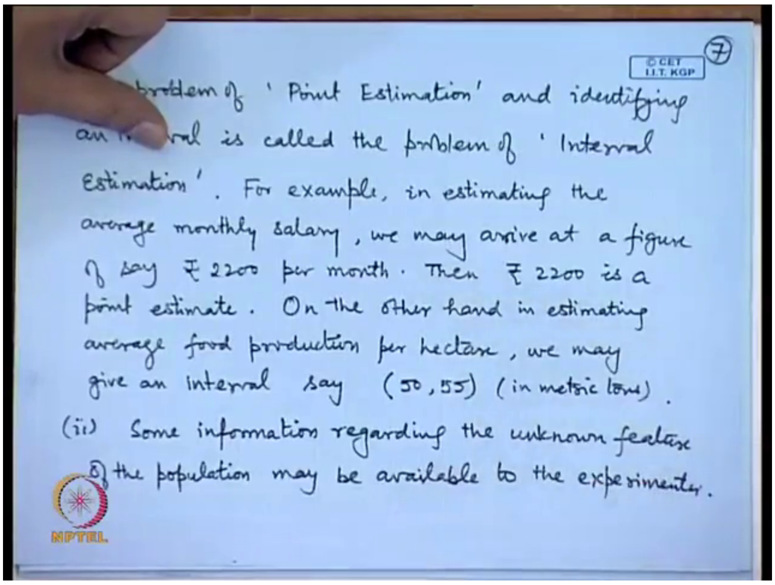


ssertions about the unknown (parameter) θ.
Two broad areas are
(i) Some feature of the population in which an enquirer is interested (say g(θ)) may be completely unknown and he/she would like to make a guess/estimate about this feature on the basis of the a random sample from this population. This is called the problem of estimation. Here we may come up with an exact estimate or may give an interval where the unknown characteristic may lie. Giving an exact estimate is called

So, it is not necessary that we just tell the value. So, the general problem of statistical inference is to make suitable statements or the assertions about the unknown parameter of the population. Here we can break this up in two broad areas, one is some features of the populations in which an experimenter or inquirer is interested. So, let us say 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 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, even 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 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, this is one major area statistical inference.

(Refer Slide Time: 30:13)



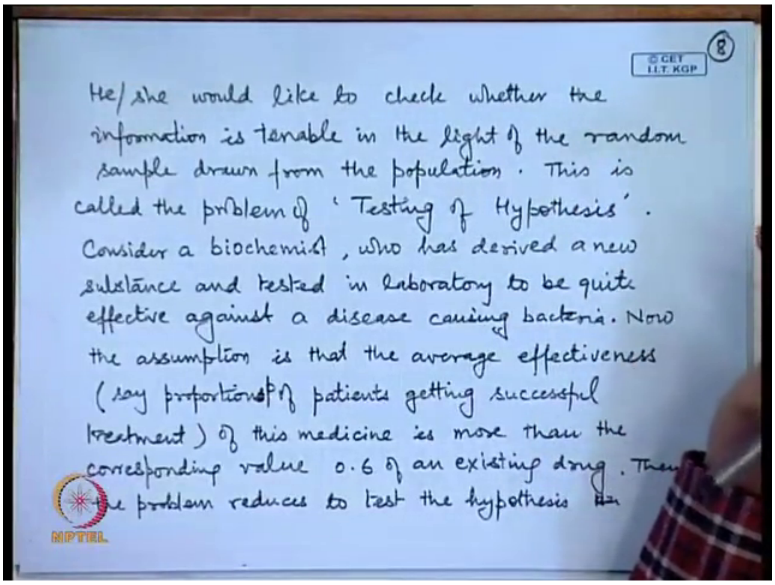
7

problem of 'Point Estimation' and identifying an interval is called the problem of 'Interval Estimation'. For example, in estimating the average monthly salary, we may arrive at a figure of say ₹ 2200 per month. Then ₹ 2200 is a point estimate. On the other hand in estimating average food production per hectare, we may give an interval say (50, 55) (in metric tons).

(ii) Some information regarding the unknown feature of the population may be available to the experimenter.

The second major area are broad categorizations 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: 30:28)



8

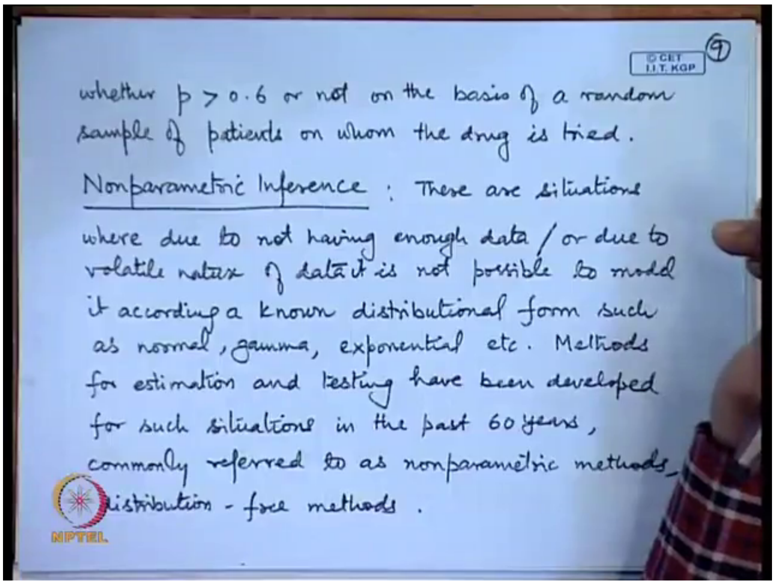
He/she would like to check whether the information is tenable in the light of the random sample drawn from the population. This is called the problem of 'Testing of Hypothesis'. Consider a biochemist, who has derived a new substance and tested in laboratory to be quite effective against a disease causing bacteria. Now the assumption is that the average effectiveness (say proportions of patients getting successful treatment) of this medicine is more than the corresponding value 0.6 of an existing drug. Then the problem reduces to test the hypothesis [illegible]

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 bio chemist he has derived a new substance and tested in the laboratory, that it is quiet effective against a diseases causing a 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. Now, when you are testing this effectiveness, you have to indentify 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 by the proportion of the patients, which gets successfully treated. So, let us call it p 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. Now, in order to have or you can say in order to introduce 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 great then 0.6 or not, this is called the problem of testing of hypothesis.

So, this is the outcome of a this test will be determine by the statistician using an appropriate statistical method. So, in this particular case, it will an appropriate test. So, based on a random sample of certain patients, who are given the medicine one will need to check this thing.

(Refer Slide Time: 32:58)



9

whether $p > 0.6$ or not on the basis of a random sample of patients on whom the drug is tried.

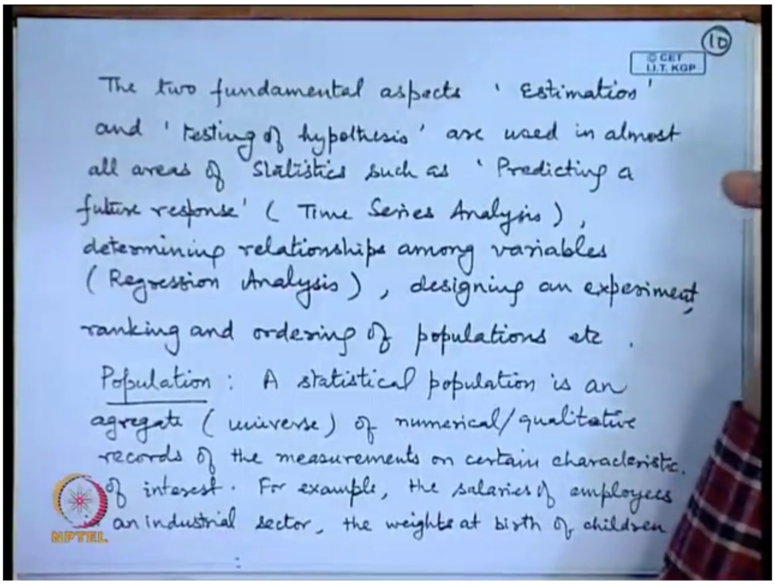
Nonparametric Inference : There are situations where due to not having enough data / or due to volatile nature of data it is not possible to model it according a known distributional form such as normal, gamma, exponential etc. Methods for estimation and testing have been developed for such situations in the past 60 years, commonly referred to as nonparametric methods distribution - free methods .

There is another distinction which I would like to make at this point. There are situations 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 non parametric methods or parameter free methods or the distribution free methods.

And this comes under the topic non parametric inference in this particular course; we will be spending almost all over time in discussing parametric inference. So, by parametric inference then we refer to the problem, when the appropriate probability distribution has been 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, this two fundamental aspects that is estimation and testing of hypothesis, they are used in almost every area statistical methodology for example, we consider predicting a future response. I mention the problem of predicting the temperature for the forth coming 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.

(Refer Slide Time: 35:23)



The two fundamental aspects 'Estimation' and 'testing of hypothesis' are used in almost all areas of Statistics such as 'Predicting a future response' (Time Series Analysis), determining relationships among variables (Regression Analysis), designing an experiment ranking and ordering of populations etc.

Population : A statistical population is an agregate (universe) of numerical/qualitative records of the measurements on certain characteristic of interest. For example, the salaries of employees an industrial sector, the weights at birth of children

:

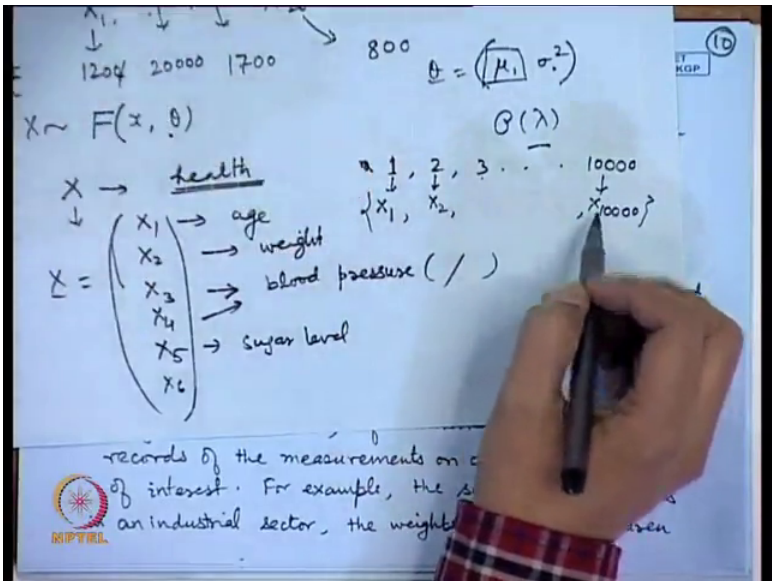
Similarly there are areas, where we determine relationship among the variables for example, the effect of providing say irrigation say modern equipment good quality seed and say good quality of a insecticides or pesticides etcetera to the farmers. And we look at the response in terms of the increased 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 regression 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 this advance area 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 there exact meanings in the context of 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 Lehman 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 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 are increasing rapidly in a particular in a city or in a particular state, however a statistical population is not the collections 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 will consider 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 the mark according to their employee code or any other identification code, then the values corresponding to their salaries.

(Refer Slide Time: 38:58)



1204 20000 1700 800

$\underline{\theta} = (\boxed{\mu_1}\ \sigma^2)$

$X \sim F(x, \underline{\theta})$

$P(\underline{\lambda})$

$X \rightarrow$ health

$1, 2, 3 \ldots 10000$

$\{X_1, X_2, \ldots, X_{10000}\}$

$\underline{X} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ X_6 \end{pmatrix}$

$X_1 \rightarrow$ age

$X_2 \rightarrow$ weight

$X_3, X_4 \rightarrow$ blood pressure (/)

$X_5 \rightarrow$ sugar level

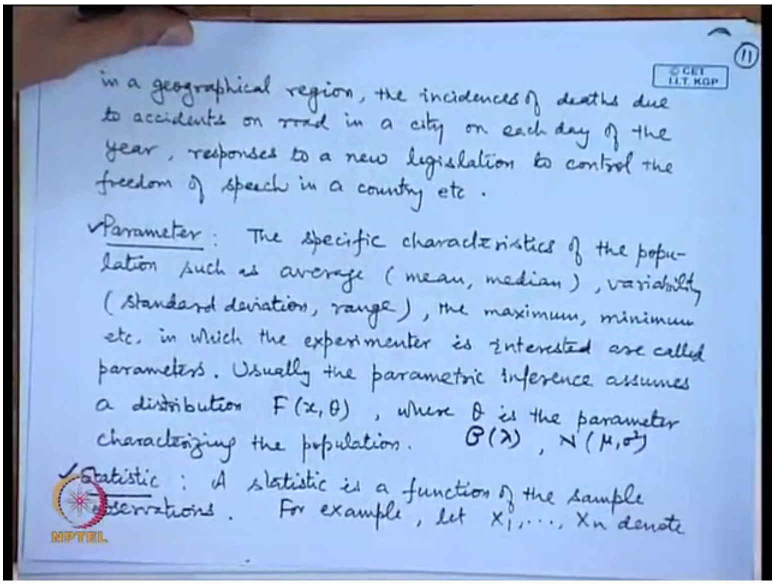
records of the measurements on c

of interest. For example, the s

an industrial sector, the weight

So, for example, I am identifying the employees as 1 2 3 and so on upto 10,000. Now, the salary of the employee number 1, that is X 1 the salary corresponding to the employee number 2, that will be called X 2 and so on X 10,000. So, in this particular case, the population of interest is these 10,000 increase.

(Refer Slide Time: 39:34)



11

© CET I.I.T. KGP

in a geographical region, the incidences of deaths due to accidents on road in a city on each day of the year, responses to a new legislation to control the freedom of speech in a country etc.

✓<u>Parameter</u>: The specific characteristics of the population such as average (mean, median), variability (standard deviation, range), the maximum, minimum etc. in which the experimenter is interested are called parameters. Usually the parametric inference assumes a distribution $F(x, \theta)$, where θ is the parameter characterizing the population. $\mathcal{P}(\lambda)$, $N(\mu, \sigma^2)$

✓<u>Statistic</u>: A statistic is a function of the sample observations. For example, let $X_1, \ldots, X_n$ denote

If you 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 you are looking at the incidents of deaths due to accidents in a 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.

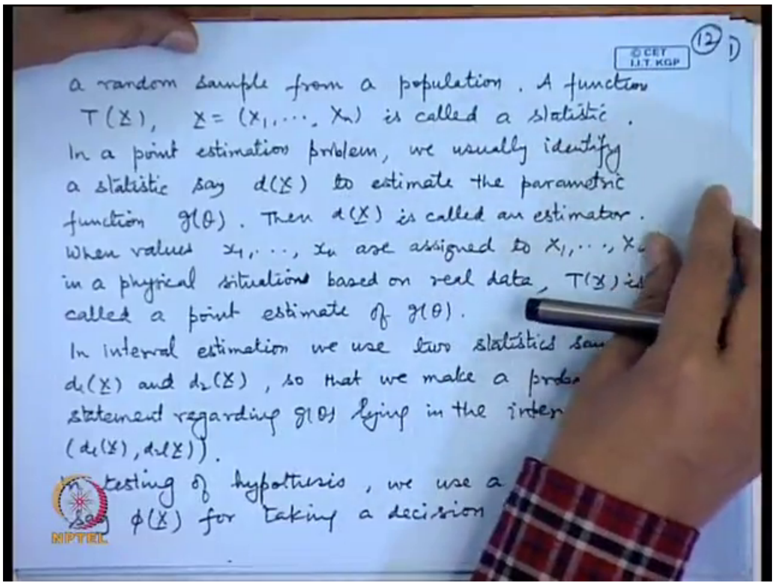
So, the population here is the number of the deaths on each day, responses to a new legislation to control the freedom of pitch in a country. So, a new legislation is placed in the parliament or it is proposed by say a by the cabinet, then say the opinion polls are taken whether it is a popular measure or not. So, 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 answers, 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 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 beforehand, but however, what is the proper meaning of the parameter. So, the specific characteristics of the population such as average for example, it could be mean, median, mode, arithmetic mean, harmonic mean etcetera or a characteristic of this use for which determines, the variability such as standard deviation range. Suppose, it is determining the weather, 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 characteristics, which are related to the population they are called the parameters. So, usually the parametric inference assumes a distribution effects theta. So, here theta is the parameter which characterized 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 here. So, the distribution model is not normal and it is characterized by the parameters mu and sigma square etcetera.

Here mu and sigma squares are the mean and variance respectively, in the poison distribution lambda itself is the mean and as well as the variance of the distribution, a statistic. So, this is a next terminology a statistics is the function of the sample observations. So, from the populations the statistician has at his disposal a random sample on the basis of this he will make the appropriate inferences. So, the sample is termed as observations X 1 X 2 X n.

(Refer Slide Time: 43:25)



12 1)

a random sample from a population. A function $T(\underset{\sim}{X})$, $\underset{\sim}{X} = (X_1, \dots, X_n)$ is called a statistic.

In a point estimation problem, we usually identify a statistic say $d(\underset{\sim}{X})$ to estimate the parametric function $g(\theta)$. Then $d(\underset{\sim}{X})$ is called an estimator.

When values $x_1, \dots, x_n$ are assigned to $X_1, \dots, X$ in a physical situation based on real data, $T(\underset{\sim}{x})$ is called a point estimate of $g(\theta)$.

In interval estimation we use two statistics sa $d_1(\underset{\sim}{X})$ and $d_2(\underset{\sim}{X})$, so that we make a prob statement regarding $g(\theta)$ lying in the inter $(d_1(\underset{\sim}{X}), d_2(\underset{\sim}{X})))$.

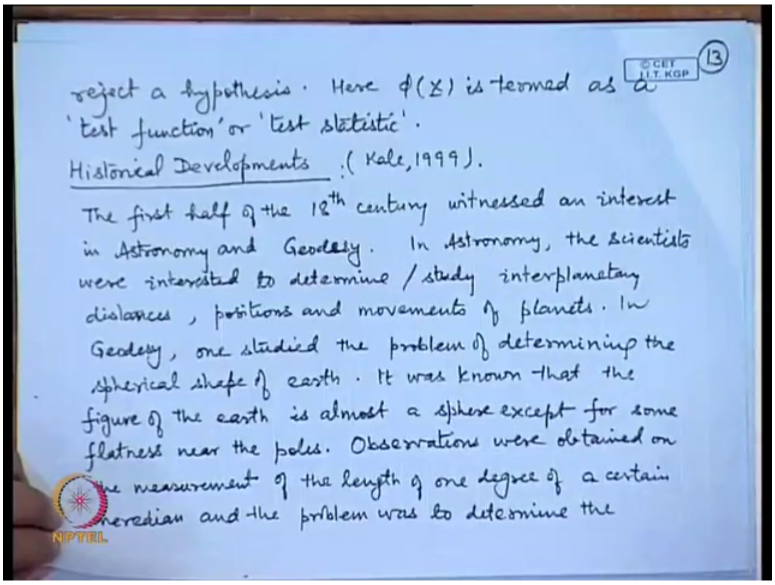
In testing of hypothesis, we use a say $\phi(\underset{\sim}{X})$ for taking a decision

So, any functions of these observations, let us called T X where X is denoting the sample X 1, X 2, X n this is called a statistics. So, in a point estimation problem, we usually indentify a statistics that is called 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 one by lambda. So, my parametric function is one by lambda. So, now it is a good question whether, we can find out and estimate for one by lambda or we may be interested to estimate say lambda to the power three, we may be estimate in a normal distribution say mu, we will be interested to estimate sigma 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 on 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 statistics has to be framed 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 X square r 1, may use one by n sigma X I minus by X square to estimate sigma X square.

In a interval estimation problem in place of one statistics say in this point estimation, we are proposing one that is dX, but in interval estimation, we need two that is end points of the interval where by parameter of interest is suppose to lie. So, we need to specify say d 1 X 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. In testing of hypothesis, we use a statistics

(Refer Slide Time: 45:49)



13

reject a hypothesis. Here $\phi(X)$ is termed as a 'test function' or 'test statistic'.

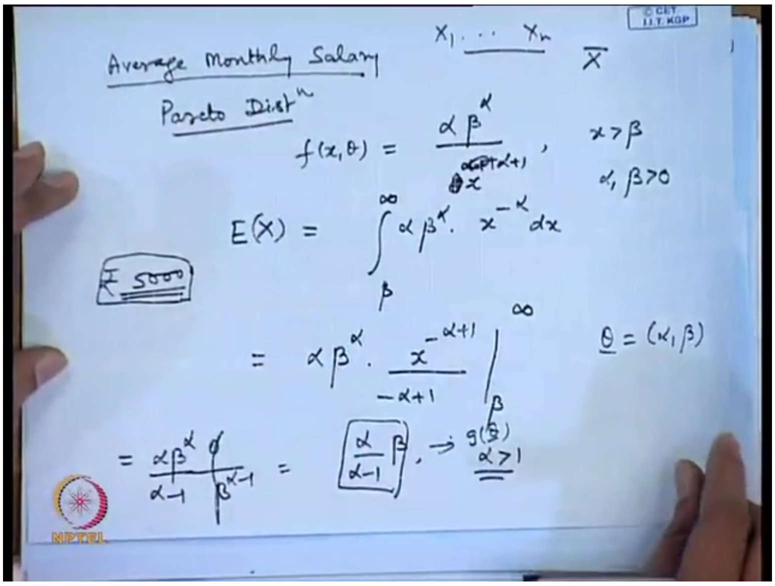
Historical Developments :(Kale, 1999).

The first half of the 18th century witnessed an interest in Astronomy and Geodesy. In Astronomy, the scientists were interested to determine / study interplanetary distances, positions and movements of planets. In Geodesy, one studied the problem of determining the spherical shape of earth. It was known that the figure of the earth is almost a sphere except for some flatness near the poles. Observations were obtained on the measurement of the length of one degree of a certain meredian and the problem was to determine the

Let us call it say phi X for a taking a decision to accept or reject a given hypothesis in this case, y x is termed as test function or test statistics. So, these are the basic terminologies, which are to be used in statistical inference, we have a population. So, that is the first thing, where we are interested what is our interest to study in the given setup, we identify the population, we dry an random sample from the given population. Now, drying of a random sample itself is a matter of full investigation, it comes under the topic of methods of samples are waste or method or sampling techniques. And it is another aspect of the statistical methodology, where we discuss various methods of taking of random sample in this particular case; we assume that 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 estimations, interval estimations or testing of hypothesis to inform the end user about the appropriate conclusion of the for about the population parameters. So, parameters are the characteristics of the population, 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 statistics. So, in the point estimation problem we will create, a point estimator using a statistics 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 to statistics using that test random sample, at this point let me briefly give example here.

(Refer Slide Time: 48:02)



© CET I.I.T. KGP

Average Monthly Salary $X_1 \ldots X_n$ $\overline{X}$

Pareto Dist^n

$$f(x,\theta) = \frac{\alpha\beta^{\alpha}}{x^{\alpha+1}}, \quad x>\beta \quad \alpha,\beta>0$$

$$E(X) = \int_{\beta}^{\infty} \alpha\beta^{\alpha} . x^{-\alpha} dx$$

$\boxed{₹ 5000}$

$$= \alpha\beta^{\alpha} . \left. \frac{x^{-\alpha+1}}{-\alpha+1} \right|_{\beta}^{\infty}$$

$\underline{\theta} = (\alpha,\beta)$

$$= \frac{\alpha\beta^{\alpha}}{\alpha-1} \frac{1}{\beta^{\alpha-1}} = \boxed{\frac{\alpha}{\alpha-1}\beta} . \rightarrow g(\underline{\theta})$$

$\underline{\underline{\alpha>1}}$

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 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 plus 1, where x is greater than beta. So, in this particular case we have consider a two parameter model, where the parameters are alpha and beta both are of course, positive.

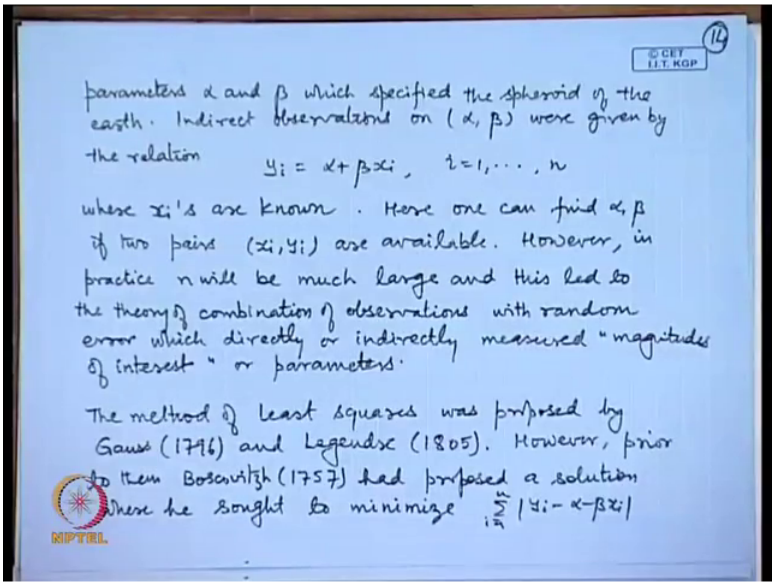
Now, here we may get interested in the average monthly salary. So, average monthly salary denotes expectation of X that means from this distribution, what is the value of the expectation of X, which can be of course, easily calculated. So, these value turns out to alpha beta to the power alpha x to the power minus alpha plus 1 divided by minus alpha plus 1 from beta to infinity. So, alpha beta to the power alpha and when we substitute the value at infinity this will vanish 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 is not be valid. So, now in this particular problem we want to estimate these parametric function so these is my g theta here theta is a vector parameter consisting of two components alpha beta now to estimate this. Now, there may be different procedures as a Lehman one may say that take the random sample X 1, X 2, X n and we may use X bar that is a sample mean to estimate this. So, this could be one method 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 of far well paid. And in that case we may device a test statistics this known as 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 or not.

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 eighteenth century and mostly in the problems of astronomy and g o d c so in astronomy, the interest was to find out the interplanetary distances. The positions of the various planets or stars and their movements in g o d c, we even 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, the standard technique is to take observations not one, but several measurements are taken. For example, they are taken about the length of 1 degree of certain meridian.

(Refer Slide Time: 52:42)



14

© CET I.I.T. KGP

parameters α and β which specified the spheroid of the earth. Indirect observations on (α, β) were given by the relation

$$y_i = \alpha + \beta x_i, \quad i = 1, \cdots, n$$

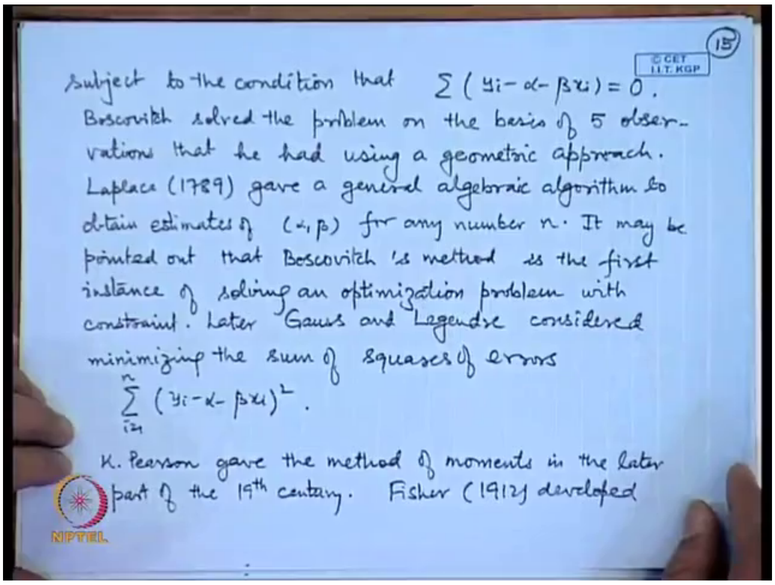
where x_i's are known. Here one can find α, β if two pairs (x_i, y_i) are available. However, in practice n will be much large and this led to the theory of combination of observations with random error which directly or indirectly measured "magnitudes of interest" or parameters.

The method of least squares was proposed by Gauss (1796) and Legendre (1805). However, prior to them Boscovitzh (1757) had proposed a solution where he sought to minimize $\sum_{i=1}^{n} |y_i - \alpha - \beta x_i|$

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 relations y i equals to alpha plus beta x i so here x i is given to us y is i given to us. So, alpha and beta are to be estimated. Nowadays, we understand this is a problem of linear, simple linear regression however this problem was studies as early as in eighteenth 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, where he sought to minimize summation of modulus y i minus alpha minus beta x i, so in place of the square initially you consider the mean absolute error actually.

(Refer Slide Time: 53:44)



15

© CET I.I.T. KGP

Subject to the condition that $\sum (y_i - \alpha - \beta x_i) = 0$. Boscovitch solved the problem on the basis of 5 observations that he had using a geometric approach. Laplace (1789) gave a general algebraic algorithm to obtain estimates of (α, β) for any number n. It may be pointed out that Boscovitch 's method is the first instance of solving an optimization problem with constraint. Later Gauss and Legendre considered minimizing the sum of squares of errors

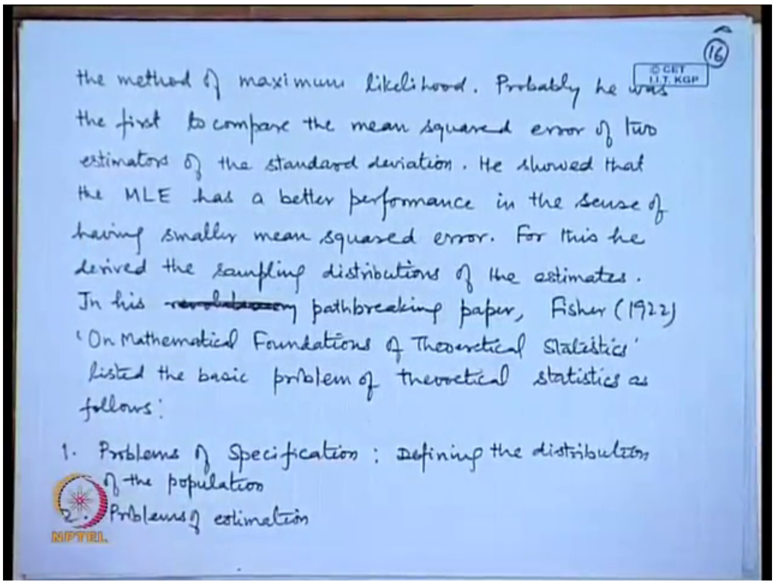
$$\sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 .$$

K. Pearson gave the method of moments in the later part of the 19th century. Fisher (1912) developed

Subject to the condition, that some of the errors must be 0. And he solved this problem using geometry, geometrical methods based on 5 observations. Later on Laplace has given an 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 sum of the squares and that is why came to be known as a method of least squares. So, you can consider the problems of statistical inference or you can say the modern statistical inference started as early as in the eighteenth century.

Further developments or you can say further techniques started to get developed towards later half of the 19 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 be 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 toward normality of the heights. And the first studies you can say the first model of the simple regression were made in this thing. Later on Karl Pearson develop the methods of moments in the later part of 19 century. The modern methods of statistics as we know today and probably were first started by fisher in 1912.

(Refer Slide Time: 55:43)



(16)

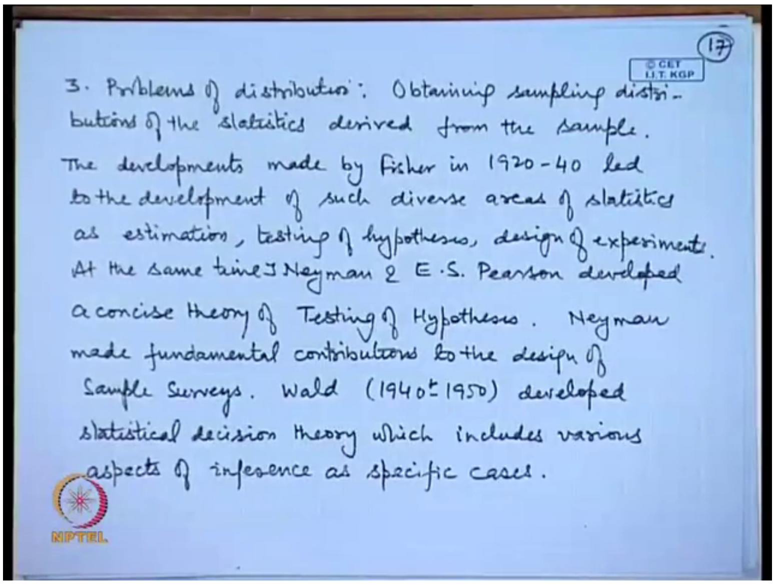
© CET I.I.T. KGP

the method of maximum likelihood. Probably he was the first to compare the mean squared error of two estimators of the standard deviation. He showed that the MLE has a better performance in the sense of having smaller mean squared error. For this he derived the sampling distributions of the estimates. In his ~~revolutionary~~ pathbreaking paper, Fisher (1922) 'On Mathematical Foundations of Theoretical Statistics' listed the basic problem of theoretical statistics as follows:

1. Problems of Specification : Defining the distribution of the population
2. Problems of estimation

Where he developed the method of maximum likelihood, he is probably the first one when he realize the importance of comparing two different methods of estimation. So, he consider two estimates of like standard deviation, he found out the sampling distribution of that and therefore, the mean squared errors. 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, where he listed the basic problem of theoretical statistics, as firstly the problem of specification that is defining the distribution of the population, second is the problem of estimation.

(Refer Slide Time: 56:29)



17

3. Problems of distribution: Obtaining sampling distributions of the statistics derived from the sample. The developments made by Fisher in 1920-40 led to the development of such diverse areas of statistics as estimation, testing of hypotheses, design of experiments. At the same time J Neyman & E.S. Pearson developed a concise theory of Testing of Hypotheses. Neyman made fundamental contributions to the design of Sample Surveys. Wald (1940-1950) developed statistical decision theory which includes various aspects of inference as specific cases.

And 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 statistics which are being used. So, these developments made by fisher in 1920 - 40. And the had these had the effects in the various areas of statistics, such as estimation testing of hypothesis designs of experiments. At the same time Neyman and E.S Pearson simultaneously developed a theory of testing of hypothesis. Lehman also developed a theory of sample surveys later on in 1940s Abraham wall developed a topic called statistical decision theory.

And these 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 its 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, it is main components in this particular case, we focus on the problem of estimation and testing of hypothesis. So, from the next class onward I will start discussion on the problem of point estimation.