Applied Statistics and Econometrics
Professor Deep Mukherjee
Department of Economic Sciences
Indian Institute of Technology Kanpur
Lecture 01
Introduction to Statistics

Hello friends. Welcome to the first lecture of the course titled Applied Statistics and Econometrics. So, let us start with a brief definition of the field statistics. What is statistics? Now, as I have told you in the intro video that data is silent. Data cannot speak. So, you have to make the data speak. So, you have to provide some language to the data through which the data will be able to speak and it will be able to communicate the information that it contains.

And statistics is basically the language that actually data is given by the statistical theorists so that they can draw important information from the data. And the data will now tell through the statistical measures that what information it contains. So, statistics can be defined as the science of data collection, data management, data analysis and interpretation of various measures that are derived from data.

(Refer Slide Time: 01:38)

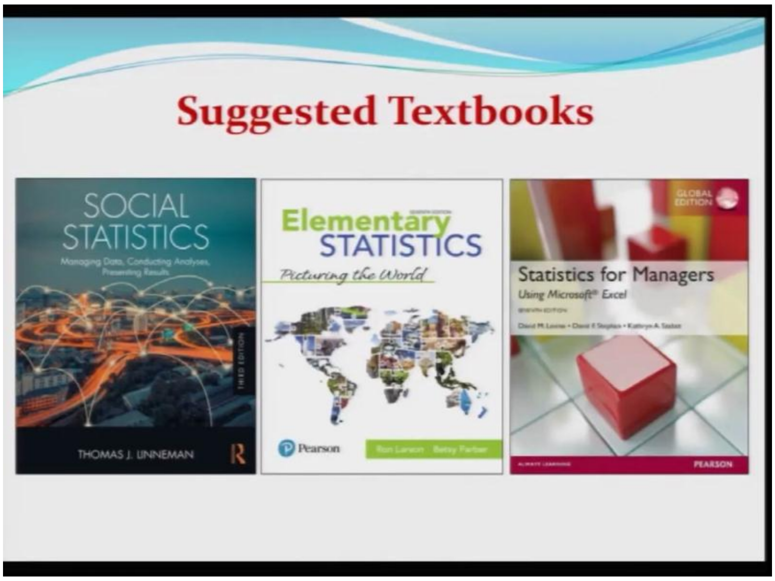
Today's Agenda:

- What is Statistics?
- Stages of Empirical Social Research
- Types of statistics
- Types of variables
- Key elements of statistical analysis

So, before we start the formal discussion on the topics in the course, let me first set the agenda of today's lecture. So, first we are going to discuss what is statistics? Then we are going to talk about the 6 stages of Empirical Social Research using data and statistical tools. Then, we are going to discuss different types of statistics that are being used in Empirical Social Research. Then, we are going to talk about different types of variables. And finally, we will end this lecture with describing the key elements of statistical data analysis.

Okay, so I borrowed a lot of material from some statistical textbooks for the first part of the course. And I am going to now show you the textbooks which are very useful. And I suggest that in addition to listening to the lectures, you also get a copy of any of these textbooks and read.

(Refer Slide Time: 02:34)

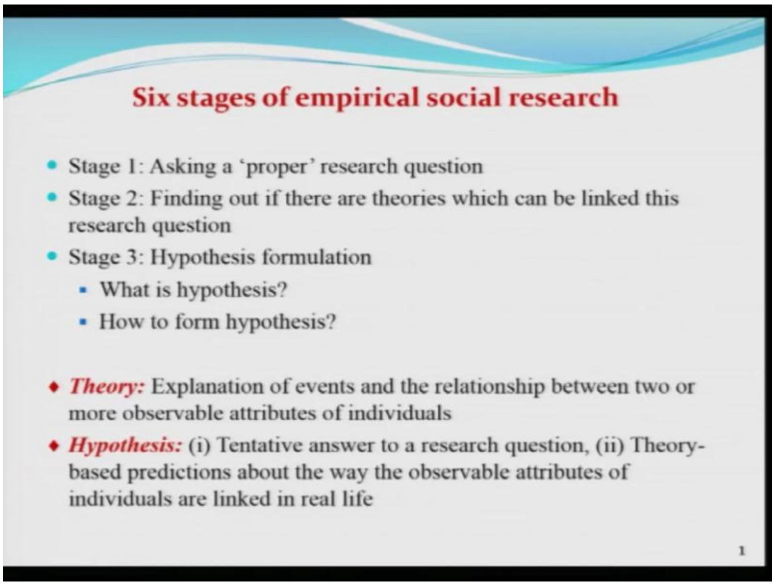


Suggested Textbooks

So, here are these textbooks based on which I, you know, developed my lectures for the first part. And the first book is called or titled, Social Statistics and it is written by Thomas Linneman. And the book actually covers a lot of statistical tools and it shows how to get computations done using the popular software, SPSS. The second book Elementary Statistics by Larson and Farber is also a very good textbook. It is written in very simple language. Now the third textbook which is titled Statistics for Managers by Levine, Stefan and Szabat.

So, this textbook actually shows you how to use Microsoft Excel to do statistical computation and I find this particular textbook very useful for the course because in the course we are also going to see how Microsoft Excel could be used for statistical computations. But of course, you know, there are other textbooks available. And you know, you are free to choose any one of them and read.

(Refer Slide Time: 03:55)



Six stages of empirical social research

- Stage 1: Asking a 'proper' research question
- Stage 2: Finding out if there are theories which can be linked this research question
- Stage 3: Hypothesis formulation
  - What is hypothesis?
  - How to form hypothesis?

- *Theory:* Explanation of events and the relationship between two or more observable attributes of individuals
- *Hypothesis:* (i) Tentative answer to a research question, (ii) Theory-based predictions about the way the observable attributes of individuals are linked in real life

1

So, let me move on to the first slide where we are going to talk about stages of empirical social research. Now, in stage 1 it is very important to ask a proper research question which would be answered using data. In other words, the research question should be quantitatively oriented so that data could be collected and conclusion or an answer could be drawn based on the data.

Now, it is not only sufficient to ask a proper research question but also you have to figure out whether there are theories coming from either the field of Management Science or Economics or Sociology or Psychology which could be linked to the research question you are asking. So, once the research question is formulated and you have linked this research question with the established theory, then it is time to formulate your hypothesis. Now, what is a hypothesis? And how to form a hypothesis?

In this lecture, we are going to talk about the meaning of hypothesis. And as the course progresses, as we move on, you will see how we can form various hypothesis. So, here at this juncture, it is important to understand what do we mean by theory? Because, I have already emphasized the role of theory behind hypothesis formulation. So, these are the stage 2 and stage 3 of empirical social research.

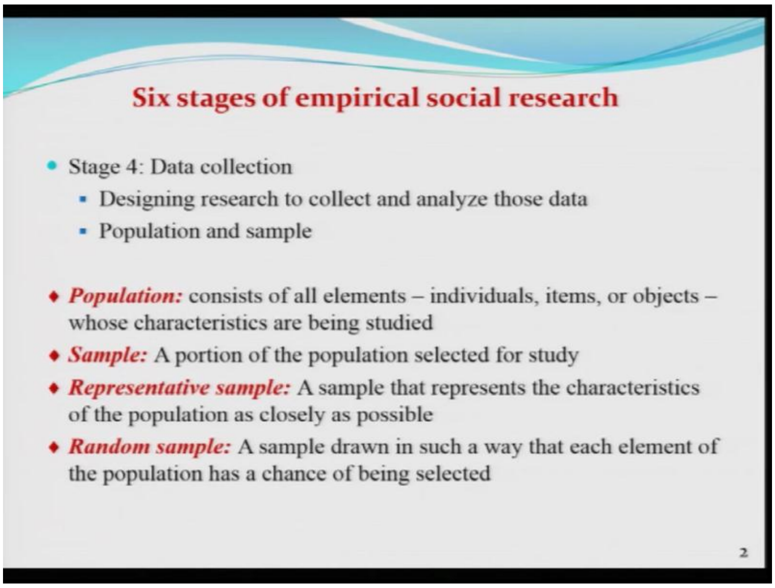
Now, what is theory? The theory is basically an explanation of already happening events and/or the relationship between two or more observable attributes of different individuals or it could be you know entities like nations. Now, what is hypothesis? Now, hypothesis could be seen from two different angles. Now, if I link hypothesis concept to stage 1 of empirical social research then you know one can say that well hypothesis is the tentative answer that a researcher is expecting to a research question asked.

Now, if I want to link the concept of hypothesis to the stage 2 of empirical social research, then I can say that hypothesis is basically the theory-based predictions about the way the observable attributes of individuals are linked in real life. So, now, let me explain the concept of hypothesis using some simple example from economics.

Even if you are not a student of economics, you probably know that the quantity demanded of any commodity is negatively related with price of that commodity. So, basically there is a law of demand which states that as price of a commodity increases, quantity demanded of that commodity shall decrease. So, this is coming from economic theory.

And if you want to form a hypothesis regarding that theory, if you want to empirically test whether this law of demand actually takes place in market data or not; whether it is evident from the market data or not; what kind of hypothesis you can set? So, think about a simple straight line demand function negatively sloped. And you know, if you do a statistical analysis, you may want to test the hypothesis whether the demand function from obtained from the real-life data is indeed negatively sloped or not. Now, let me move on to the other stages of the empirical social research.

(Refer Slide Time: 07:42)



Six stages of empirical social research

- Stage 4: Data collection
  - Designing research to collect and analyze those data
  - Population and sample

- ***Population:*** consists of all elements – individuals, items, or objects – whose characteristics are being studied
- ***Sample:*** A portion of the population selected for study
- ***Representative sample:*** A sample that represents the characteristics of the population as closely as possible
- ***Random sample:*** A sample drawn in such a way that each element of the population has a chance of being selected

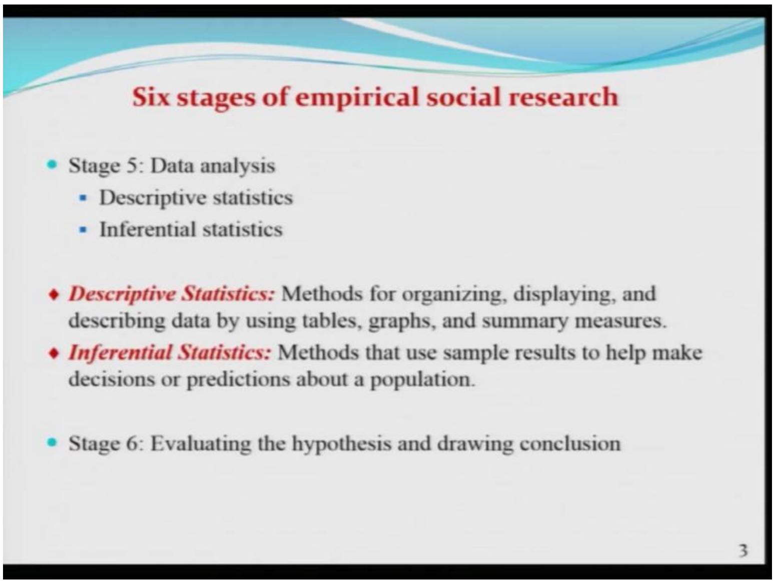
2

So, here we are going to focus on data collection which is stage 4. Once your research question is formed and hypothesis is framed, then it is time to collect data so that you can empirically validate your theory and you want to find an empirical evidence-based answer for your research question that you had set. Now, while you are in stage 4 it is important to understand two major concepts of statistical data analysis and they are population and sample.

So, let us look at them one by one. So, what is a population? So, population consists of all elements. It could be individuals; it could be items or objects; it could be nations whose characteristics are being studied in an empirical social research project. What is a sample? Sample, it is a proportion of the population that you have selected for your study.

Now, in this regard, it is important to understand two more related concepts namely representative sample and random sample. What is a representative sample? Representative sample is a sample that represents the characteristics of the population as closely as possible. And what is a random sample? A random sample drawn in such a way that each element of the population has a chance of being selected. Now, these concepts will become even more clearer as we move on with this course.

(Refer Slide Time: 09:20)



Six stages of empirical social research

- Stage 5: Data analysis
  - Descriptive statistics
  - Inferential statistics

- *Descriptive Statistics:* Methods for organizing, displaying, and describing data by using tables, graphs, and summary measures.
- *Inferential Statistics:* Methods that use sample results to help make decisions or predictions about a population.

- Stage 6: Evaluating the hypothesis and drawing conclusion

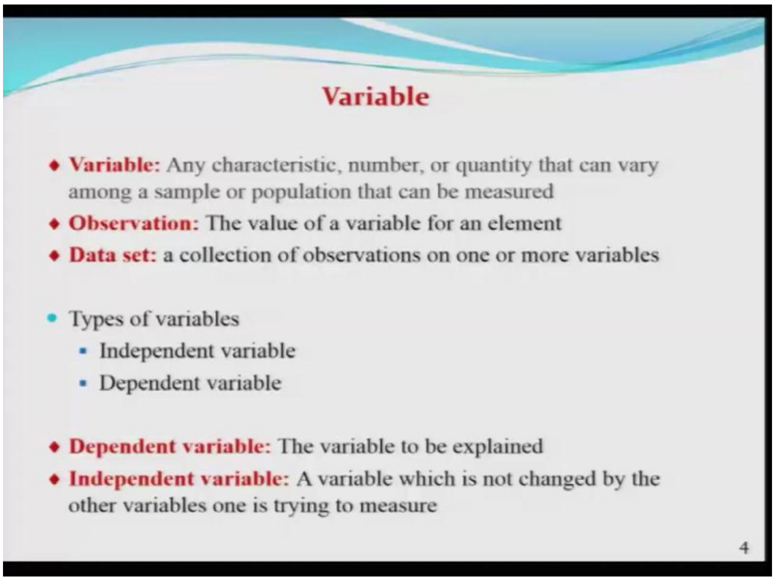
3

So, now, let us look at the stage 5 of empirical social research strategies. And that is basically the stage of data analysis. So, in this stage, we focus on two types of data analysis. One is called descriptive statistics and the other one is called inferential statistics. Now, what is descriptive statistics? So, it is a method for organizing, displaying and describing data by using tables, graphs and summary measures. Now, how we are going to make tables, graphs and what are these summary measures, these things will be cleared to you as we move on in this course.

At this moment, it is important to understand the difference between descriptive statistics and inferential statistics very well because inferential statistics plays a big role in applied statistical data analysis and applied economic research. What is inferential statistics? Methods that use sample results to help make decisions or predictions about a population is called inferential statistics.

Now, after the stage 5 there is only one stage left and that is stage 6. That is basically evaluating the hypothesis and drawing a conclusion. Now, these stages of empirical social research may look a bit vague you know, as I described here. But you know in the coming lectures, these stages will be much more clear to you when I will start again coming back to these stages with some illustration or example from real life.

(Refer Slide Time: 11:05)



Variable

- **Variable:** Any characteristic, number, or quantity that can vary among a sample or population that can be measured
- **Observation:** The value of a variable for an element
- **Data set:** a collection of observations on one or more variables

- Types of variables
  - Independent variable
  - Dependent variable

- **Dependent variable:** The variable to be explained
- **Independent variable:** A variable which is not changed by the other variables one is trying to measure

4

Now, we are going to talk about another interesting concept and you know this is the bread and butter of statistical data analysis and the notion of variable is very important to understand at this moment. So, what is a variable? A variable is basically a characteristic, a number, or quantity that can vary among a sample or the population elements and that can be measured.

So, here you know it is very important to note that we are talking about measurement and we are talking about a particular characteristic or number which is varying over the entities or the elements of either population or sample. So, it is not a constant. Now, what do we mean by an observation? So, an observation is the value of a variable for an element either in sample or in population.

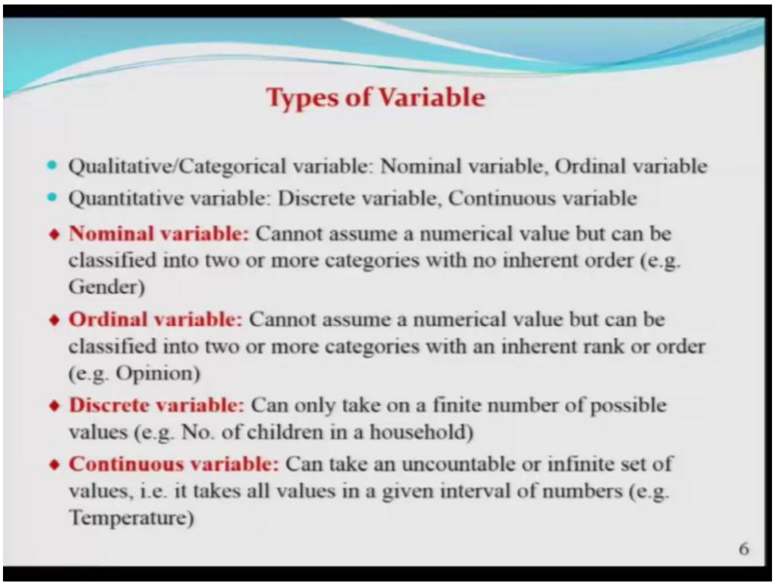
Now, what is a data set? We are going to analyse different data sets in this course. So, it is important to understand what does a data set mean. So, a data set is basically a collection of observations on one or more variables. So, at this point it is important to understand what are the different types of variables that we deal with in applied statistical or econometric research. So, we are going to here talk about two different variables. One is independent variable and the other one is dependent variable.

Now, what is a dependent variable? It is a variable that needs to be explained through data analysis. And what is an independent variable? This is a variable which is not changed by the other variables one is trying to measure. Or you can say that independent variables actually help to explain the variation in dependent variable.

Now, please note that you know as I am defining dependent and independent variable here, one may think that "So, is there a causality between dependent variable and independent variable?" "Can I say that the movements in independent variable cause the variation in dependent variable?" The answer is well, it may it may; it may not. You know sometimes causality is very difficult to establish through statistical or econometric data analysis. So, even if there is some relationship between dependent variable and independent variable, it may not be causal all the times.

Note that the data for statistical or econometric analysis comes in form of variable and the variables could be of different types depending on the data that you see you know underneath. So, based on that feature now we are going to talk about different types of data. And the major two types are: qualitative variable and quantitative variable. Now, we are going to discuss about the subdivisions of qualitative and quantitative variables with examples.

(Refer Slide Time: 14:12)



Types of Variable

- Qualitative/Categorical variable: Nominal variable, Ordinal variable
- Quantitative variable: Discrete variable, Continuous variable
- **Nominal variable:** Cannot assume a numerical value but can be classified into two or more categories with no inherent order (e.g. Gender)
- **Ordinal variable:** Cannot assume a numerical value but can be classified into two or more categories with an inherent rank or order (e.g. Opinion)
- **Discrete variable:** Can only take on a finite number of possible values (e.g. No. of children in a household)
- **Continuous variable:** Can take an uncountable or infinite set of values, i.e. it takes all values in a given interval of numbers (e.g. Temperature)

6

So, qualitative or categorical variables are basically of two types: either it is nominal variable or it is ordinal variable. Now, the quantitative variable is also of two types: discrete variable or continuous variable. Now, let us look at these 4 types of variable one by one with example. First, we are going to discuss what is a nominal variable. A nominal variable cannot assume a numerical value but it can be classified into two or more categories with no inherent order or rank in it.

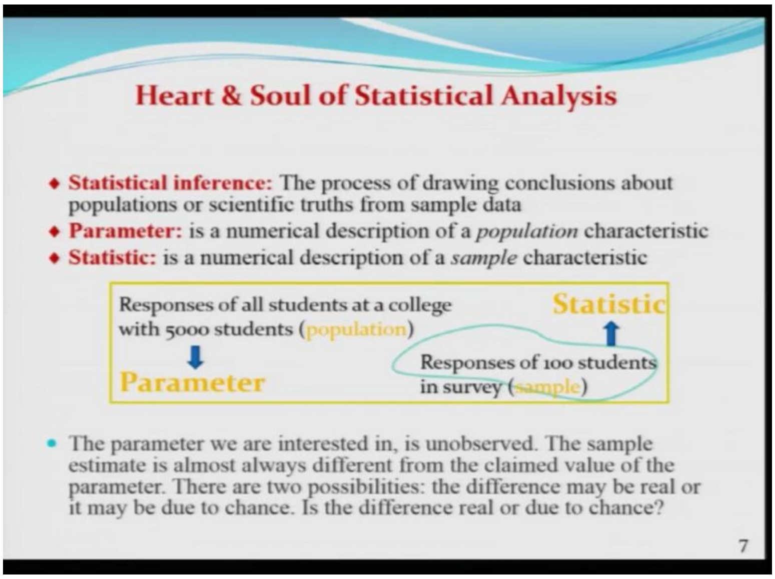
So, for example, we can consider gender. Now note that gender is an attribute or a qualitative feature of you know a person. Now, it cannot be measured. There could not be numerical values against male or female. But of course, we can classify these gender variables into two different categories namely male and female.

Next, we move on to ordinal variable. Ordinal variable is also another type of categorical variable. But this one is some far different from nominal variable in the sense that when we define nominal variable we talked about no inherent order in the categories where we are classifying you know one nominal categorical variable. But in this ordinal variable case actually we can classify these variables or this type of variable into two or more categories with an inherent rank or order.

Now, an example of ordinal variable could be public opinion. Suppose I ask you a particular question, “Do you support that we should adopt more renewable energy in place of fossil fuel energy even if it raises our electricity bill?” So, next we move on to the case of discrete variables. So, a discrete variable is a quantitative variable. Now, here a discrete variable can take only countably finite number of possible values. So, an example could be number of children in a household.

The last one in the list is continuous variable. So, it is a kind of quantitative variable that can take uncountable or infinite set of values. It implies that the variable can take all values in a given interval of numbers. Example could be temperature, or body weight, etc. So, after defining the basic concepts that are required for statistical data analysis, now let us look at the heart and soul of statistical data analysis. So, we have already defined concepts like population and sample. We have also introduced the concepts of descriptive statistics and inferential statistics. Let us see what one can do with these concepts.

(Refer Slide Time: 17:21)



Heart & Soul of Statistical Analysis

- **Statistical inference:** The process of drawing conclusions about populations or scientific truths from sample data
- **Parameter:** is a numerical description of a *population* characteristic
- **Statistic:** is a numerical description of a *sample* characteristic

Responses of all students at a college with 5000 students (population)
Parameter
Statistic
Responses of 100 students in survey (sample)

- The parameter we are interested in, is unobserved. The sample estimate is almost always different from the claimed value of the parameter. There are two possibilities: the difference may be real or it may be due to chance. Is the difference real or due to chance?

7

So here in the last slide for today's lecture we are going to introduce the concept called or terminology called statistical inference. So, it is a data analysis process of drawing conclusions about populations or scientific truths from the sample data. Now, here to conduct a statistical inference analysis, you need to know two more concepts which are parameter and statistic.

What is a parameter? It is a numerical description of a population characteristic. And what is a statistic? It is a numerical description of a sample characteristic. Now, let us look at an example which will probably help you to understand these concepts or the problem of statistical inference in detail. It will help you in understanding these complicated ideas of statistical inference analysis.

Suppose you are interested to collect some information on the vital statistics of the students that are studying in your college. So, now the college can have you know a huge population of 5000 students. But it is very costly both in terms of money and time to go to each and every one of these 5000 students and conduct a survey, collect the information and then derive some statistical measure to draw some inference.

So, to avoid that costly route what can you do? You can actually draw a random sample from this population of 5000 students. And then based on that sample you can try to draw some meaningful conclusion. And that is basically the process of statistical inference. Now, look at this chart that I have developed here. So here this rectangle that I am showing, that is basically your entire population of 5000 students who are studying in a particular college and that is my population.

Now, 5000 is a large number. So, you want to just you know derive small sample, you want to draw a random sample. Because random samples have good statistical properties. We will discuss these things later. But you know, you want to draw a random sample of 100 students for a survey and you collect responses from these 100 students. So, these you know 100 students, you know, are shown by this curvature inside this rectangle and then that is my sample.

Now, note that if you have all the time and money to go to these 5000 students and interview them to collect data, then the summary statistic that you are going to get from 5000 surveys is called a population parameter. So, it could be body weight; it could be height; it could be their marks in mathematics; it could be anything that is measurable.

Now, when you have a small sample of only 100 students from the survey and you want to calculate some descriptive statistic from the collected data, then you know you are calling it statistic. Now, note that this parameter that we are interested in is unobserved because we do not have time and money to go to each one of them and collect the data and draw the statistical conclusion.

So, the sample estimate is almost going to be different from the claimed value of the parameter. Suppose I say that the mean weight of the students in your college is around 55 kgs. So, that is basically a claim and that is the value of the population parameter. But note that this is unobserved. Nobody knows what is the actual mean weight of the 5000 students, you know, who are studying in your college.

Now, you get these 100 surveys done and you collect this data and compute the mean weight of these 100 students and you know highly likely that it is not going to be 55. So, if it is 53 or if it is you know 57, then why you know this is differing from the population parameter? So, note that there are two possibilities here. One is that this difference that you get from the hypothesis that you made about the population parameter and the sample statistic that you have calculated that difference is actually real or it could be just due to chance.

So, this is basically in a nutshell statistical data analysis is all about. And I mean, of course, this is not the entire picture of statistics; you know one can do more. But in the next lecture, we are going to continue the discussion and in the next lecture I am going to introduce this field called econometrics. Thank you.