Introduction to Probability and Statistics

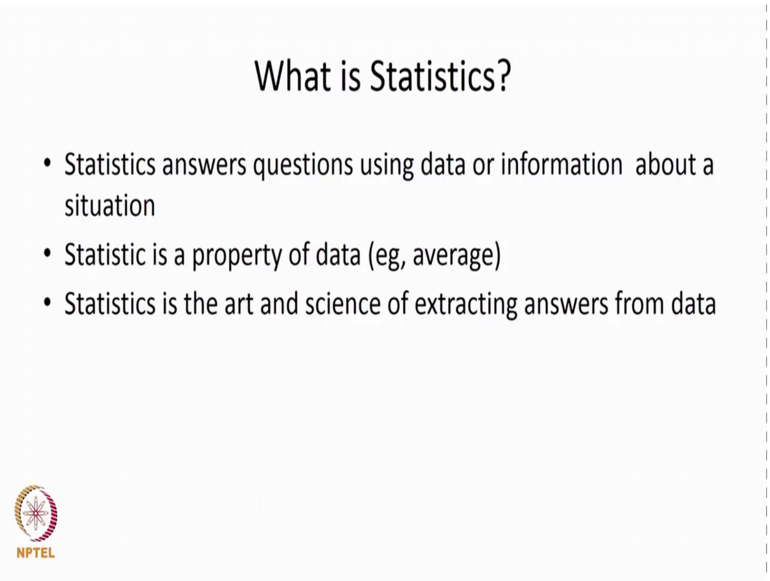
Prof. G. Srinivasan
Department of Management Studies
Indian Institute of Technology, Madras

Lecture - 01
Probability and Statistics

Welcome to this course on Probability and Statistics. This is an NPTEL MOOC's course, and as I had explained in the introductory video, this course is a very basic elementary course which can be treated as an introductory or a pre-term course which will lead you to probability and statistics. As mentioned this is a 4 week course with about 8 hours of content, maybe going to about 10 hours of content.

We will initially look at topics in statistics, and concentrate lot more on descriptive statistics and later in the course move towards understanding probability. So, in this first lecture, we only define what is statistics, where it is going to be used and how it is going to be used.

(Refer Slide Time: 01:28)

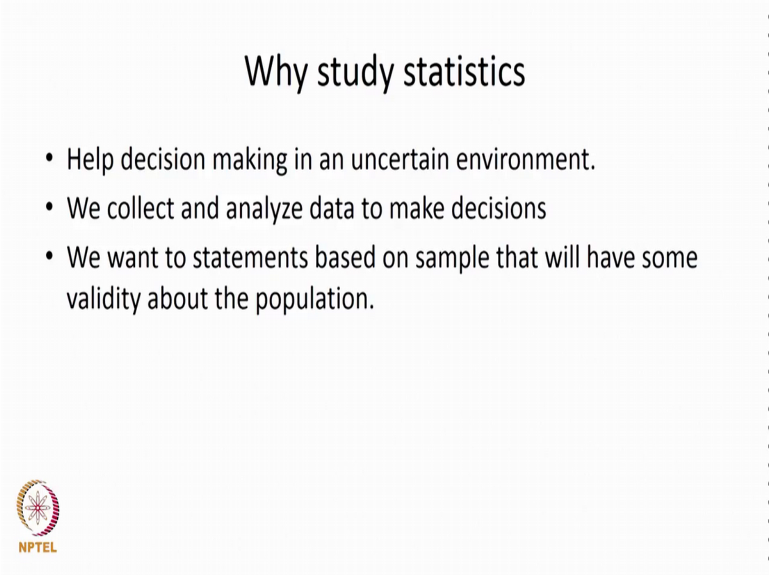


What is Statistics?

- Statistics answers questions using data or information about a situation
- Statistic is a property of data (eg, average)
- Statistics is the art and science of extracting answers from data

So, we asked the first question what is statistics. The answers are statistics answers questions using data or information about a situation. There is also this word called statistic, and you can observe that the s is missing. Statistic is also commonly used and a statistic is a property of data. Example, simple average is a statistic, median is a statistic. Statistic is a property of data or a parameter that represents data in some form. Statistics which is the field that we are talking about is the art and science of extracting answers from data. Therefore, we understand that data is very, very important to learn and understand statistics.

(Refer Slide Time: 02:32)



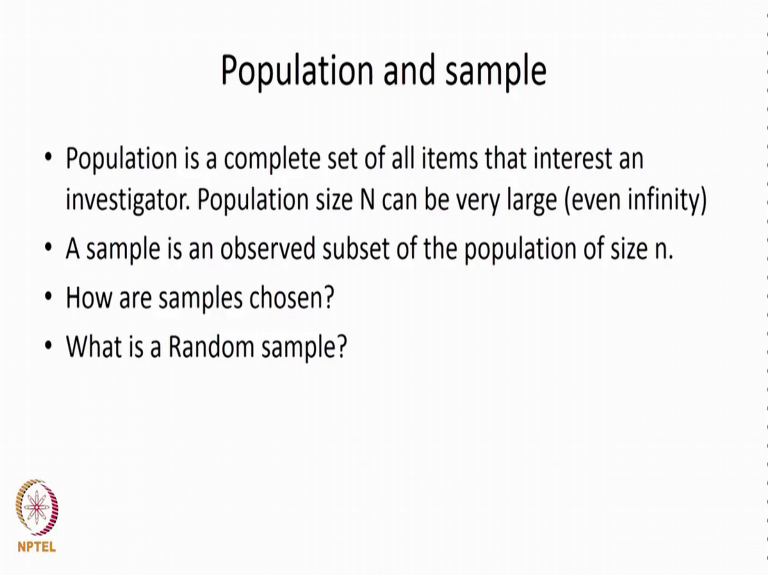
Why study statistics

- Help decision making in an uncertain environment.
- We collect and analyze data to make decisions
- We want to statements based on sample that will have some validity about the population.

So, why do we study statistics? Statistics helps in decision making in an uncertain environment. There are times it also helps decision making in certainty. Primarily the purpose is to make good decisions, and to make good decisions with data. Because decisions made using data are important, and can be consistent compared to decisions that are made through opinions. Therefore, we need to make decisions using models involving data, and statistics as a field of study provides us with models methods that help us make good decisions using data. Therefore, we collect and analyze data to make the decisions.

At times we collect data and we will not be able to cover the entire population as it is called. So, we collect data from samples we collect data from subsets of the population. So, we collect data from samples, and then try to understand something about the population by analyzing the data that is collected from or collected using samples.

(Refer Slide Time: 04:01)



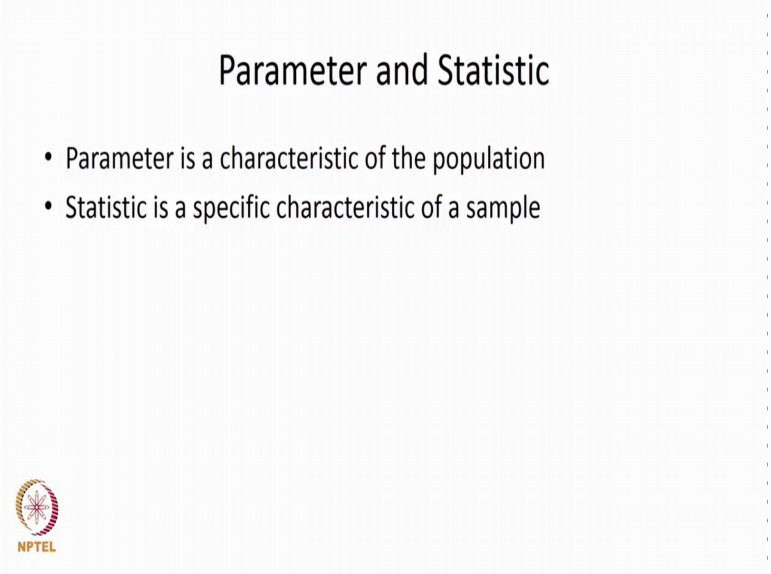
Population and sample

- Population is a complete set of all items that interest an investigator. Population size N can be very large (even infinity)
- A sample is an observed subset of the population of size n.
- How are samples chosen?
- What is a Random sample?

So, what are population and sample? Population is a complete set of all the items that interest and investigator. Population size generally denoted by capital N or uppercase N, can be very large and at times even infinity.

For example, if we want to look at what is the average height of people in the world, then we realize that the population is large very, very large. Whereas, a sample is an observed subset of the population. It is also important to note the notation that we have used. So, a sample has a small n or a lowercase n; whereas, population has a big N or capital N or uppercase N. So, that leads us to simple things like how our samples chosen are they chosen randomly, sometimes are they chosen systematically and many other ways or what is a random sample.

(Refer Slide Time: 05:13)



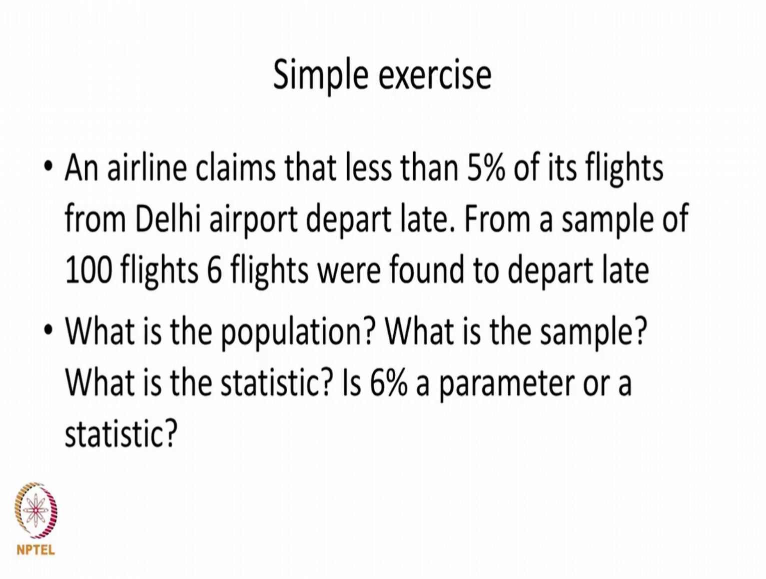
Parameter and Statistic

- Parameter is a characteristic of the population
- Statistic is a specific characteristic of a sample

We also need to understand two more words commonly used a parameter and a statistic. We already used or saw the word statistic, we saw the meaning of the word statistic. Now we go on parameter is a characteristic of the population. Statistic is a specific characteristic of a sample. For example, if average is what we are looking at the population average will be a parameter. And sample average will be a statistic.

Population average we use the notation mu; whereas, sample average we would use the notation x bar. So, we would be dealing with statistics like sample averages and so on. And there are models where we can try and estimate parameters of the population using data collected from samples, or using statistics from sample. So, we just start with a simple exercise to understand a couple of things. Let us say an airline claims that less than 5 percent of it is flights from Delhi airport depart late.

(Refer Slide Time: 06:24)



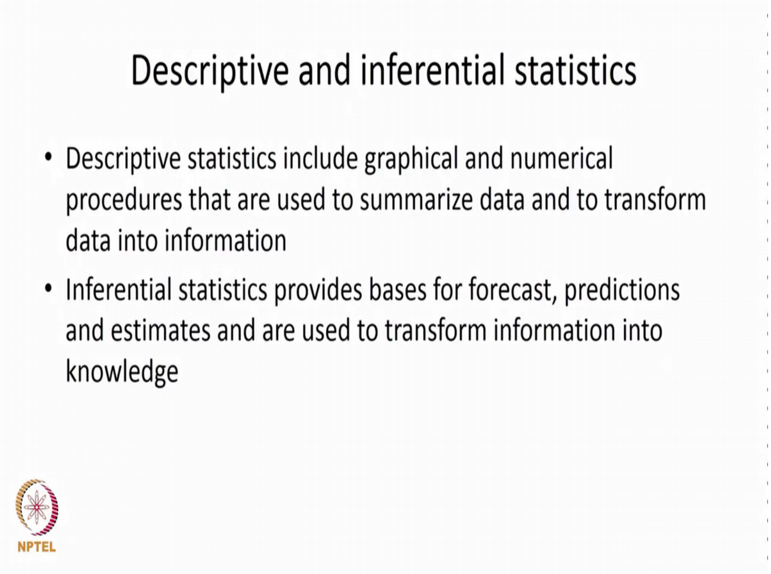
Simple exercise

- An airline claims that less than 5% of its flights from Delhi airport depart late. From a sample of 100 flights 6 flights were found to depart late
- What is the population? What is the sample? What is the statistic? Is 6% a parameter or a statistic?

From a sample of 100 flights, it was observed that 6 flights were found to depart late. So, let us look at this sentence and try to understand, what is the population? What is the sample? What is the statistic, is the is for example, 6 percent a parameter or a statistic. So, let us read the sentence again, an airline claims that less than 5 percent of it is flights from Delhi airport depart late.

So, we can assume that the population in this case, or all the flights that depart from Delhi airport. Out of these 100 flights data on 100 flights were collected, which means the sample in this case is 100. So, small n is 100 capital N is large very large. And then it was observed that 6 flights were found to depart late. So, the 6 flights is actually a statistic that we found from the sample. And so, the answers are population in this case are the all the flights that depart from Delhi airport. Sampled capital a small n equal to 100 are the flights for which data has been taken; 6 percent that is observed assuming that this 6 percent or 6 flights out of 100 were observed then it becomes a statistic.

(Refer Slide Time: 08:04)



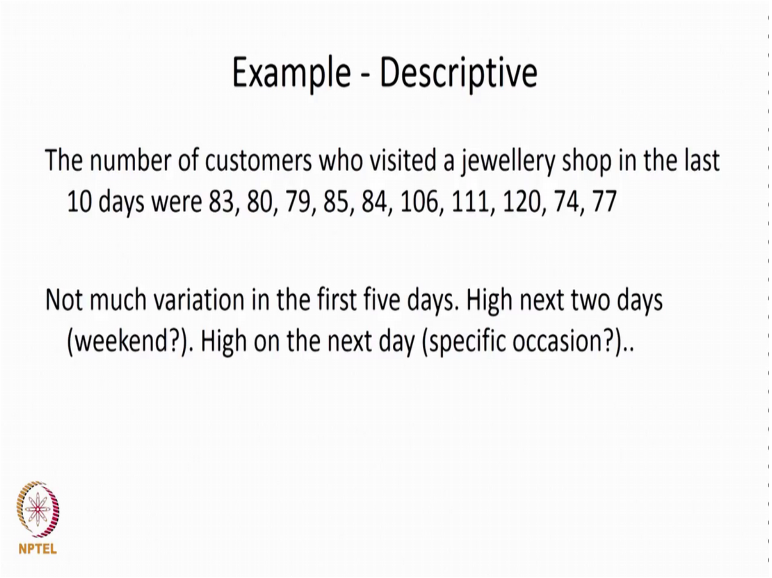
Descriptive and inferential statistics

- Descriptive statistics include graphical and numerical procedures that are used to summarize data and to transform data into information
- Inferential statistics provides bases for forecast, predictions and estimates and are used to transform information into knowledge

Now, going back to the field of statistics, we have 2 broad types of models in statistics, these are called descriptive statistics and inferential statistics. So, descriptive statistics use graphical and numerical procedures to summarize data and to transform data into information.

Inferential statistics provides base for forecast predictions and estimates which are used to transform information into knowledge. So, we will begin with learning descriptive statistics. And in this particular course, whatever statistics we are going to look at are descriptive, and we will not be looking at inferential statistics in this course.

(Refer Slide Time: 08:49)



Example - Descriptive

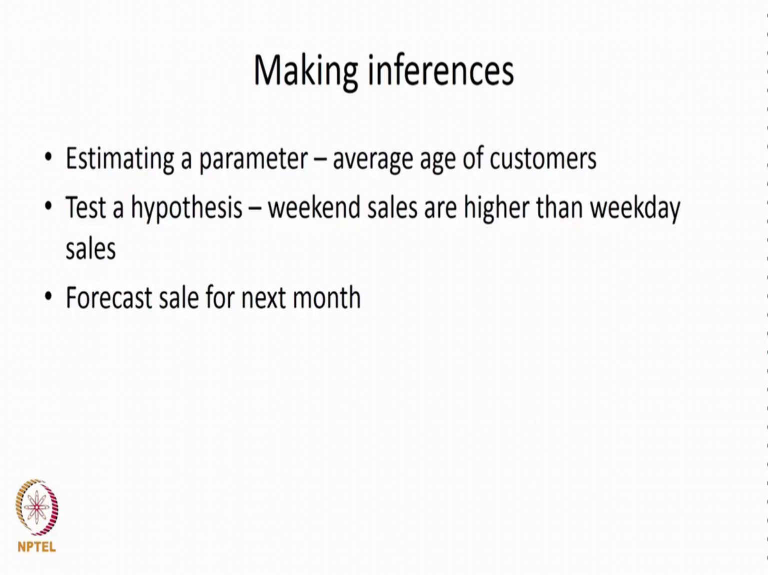
The number of customers who visited a jewellery shop in the last 10 days were 83, 80, 79, 85, 84, 106, 111, 120, 74, 77

Not much variation in the first five days. High next two days (weekend?). High on the next day (specific occasion?)..

So, example of a descriptive statistics, example number of customers so, visited a jewelry shop in the last 10 days are given 83 80 79 85 and so on. So, we can describe this data in many forms we can try to understand something from this date. For example, one could say that looking at this data, the first 5 days we did not find too much of a variation.

Whereas the next 3 days we found a lot of variation. Among the 3 days there was little variation, but compared to the first 5 there is an increase and then there is a reduction. So, some things that we could observe are the first first 5 days could be Monday through Friday. The next 2 could be a Saturday Sunday. The third perhaps could be a holiday therefore, the number of people increased, and then it went back to working days and so on. So, we can try and describe something from the data that we actually have. How do we do it? Little more formally we will see as we move along in this course.

(Refer Slide Time: 10:07)



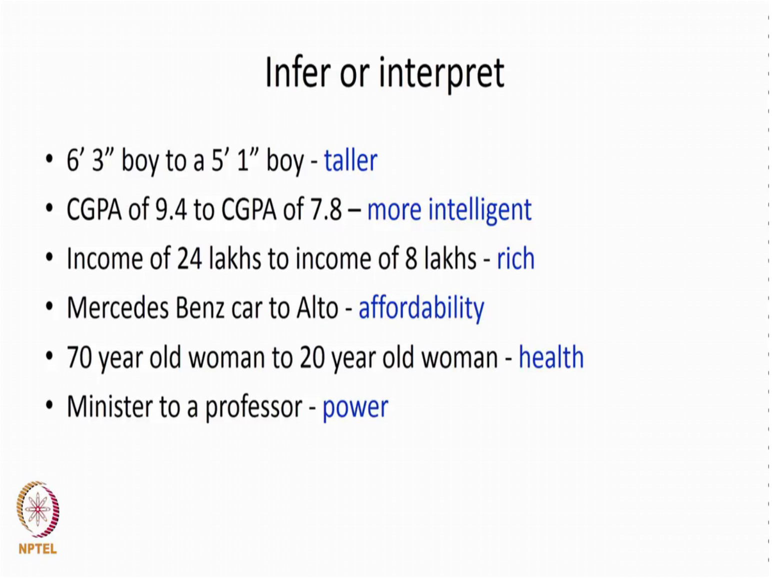
Making inferences

- Estimating a parameter – average age of customers
- Test a hypothesis – weekend sales are higher than weekday sales
- Forecast sale for next month

So, some simple things about making inferences. Estimating a parameter average age of customers. Testing a hypothesis for example, is it true that weekend sales are higher than big day sales a number of people who visit the shop during the weekends and holidays are much higher than those who visit during the normal days.

Another inference could be how do I make a forecast of the sale for the next month next month using some past data or old data. So, these are some examples of making inferences. And as I pointed out we would not be looking at models to do this in this course. Whereas, we would be looking at models that would describe the data for example, maybe the first part finding the average and so on.

(Refer Slide Time: 10:59)



Infer or interpret

- 6′ 3″ boy to a 5′ 1″ boy - taller
- CGPA of 9.4 to CGPA of 7.8 – more intelligent
- Income of 24 lakhs to income of 8 lakhs - rich
- Mercedes Benz car to Alto - affordability
- 70 year old woman to 20 year old woman - health
- Minister to a professor - power

Now, what more can we do with data? First thing that data does or we do with data is to compare. Example, we can compare a 6 feet 3 boy to a 5 feet one boy, and say that this boy is taller considerably taller. We could compare a student with the CGPA of 9.4 with another student with the CGPA of 7.8. And perhaps come to a conclusion or come to a decision, that the student with the CGPA of 9.4 has performed academically better than the student with the CGPA of 7.8.

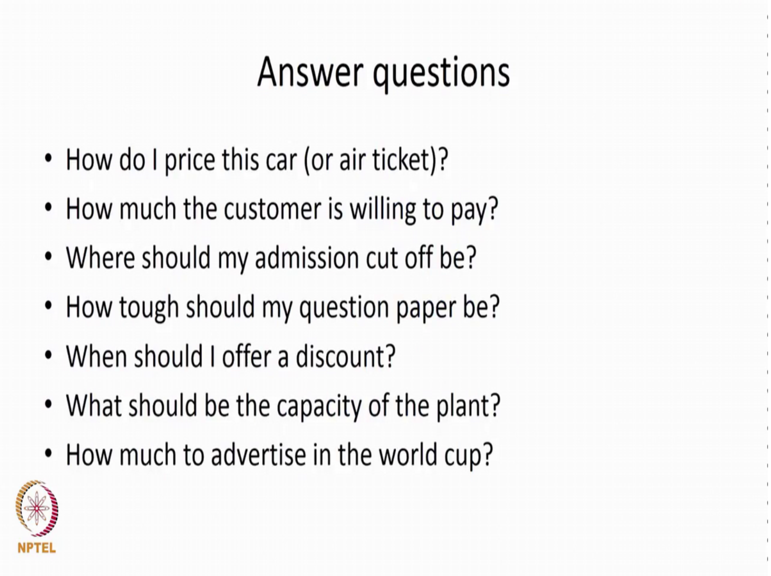
We can compare 2 people, one would one having an income of 24 lakhs per year. To another who has an income of 8 lakhs per year, and then concluded that the first person is earning more than the second person. We could compare different types of cars, and then form a certain judgment saying that, this person has a costlier car car that is costlier than the other car.

We could compare a 70-year-old woman to a 20-year-old woman and compare that this person is older than the other. We could compare 2 people, one could be a minister the other could be a professor and say that they are in different professions. Or they enjoy certain privil each enjoys a certain privilege in society and so on. Or each takes part in certain types of decisions which would benefit the society and so on.

So, data helps us to compare, and that is we have given you examples of how you can use different data to compare. Data also helps us to infer or interpret. Going back to the same example, the 6 feet 3-inch boy is taller than the other. The CGPA of 9.4 can be taken as more intelligent than the other. Though one could say has performed better than the other.

The 24 lakh person can be said as richer than the person whose income is 8 lakhs. The the person who drives a better car, one can say that the affordability is higher, and one could compare the health of a 70-year-old person to a 20-year-old person, and one could compare the power that a minister has with respect to what a professor would have. Therefore, data helps us to infer or interpret it.

(Refer Slide Time: 13:37)

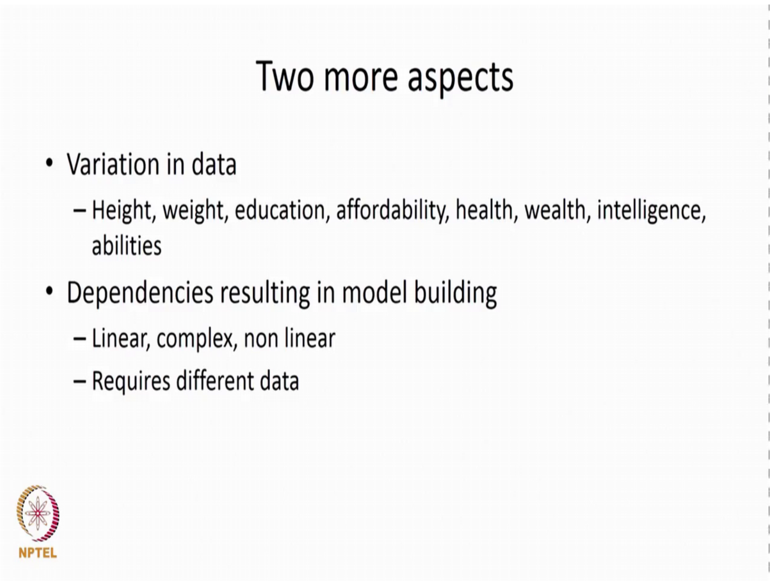


Answer questions

- How do I price this car (or air ticket)?
- How much the customer is willing to pay?
- Where should my admission cut off be?
- How tough should my question paper be?
- When should I offer a discount?
- What should be the capacity of the plant?
- How much to advertise in the world cup?

We would also times this data helps us to answer questions. And some of these questions could be how do I price this car or how do I price an air ticket. How much the customer is willing to pay for something? Where should my admission cutoff be if I am doing an admission for a course. How tough should my question paper be, if I am a course instructor, when should I offer a discount, if I own a shop and I sell things. What should be the capacity of the manufacturing plant? And how much to advertise in when and I have an event like a world cup or whatever.

(Refer Slide Time: 14:28)



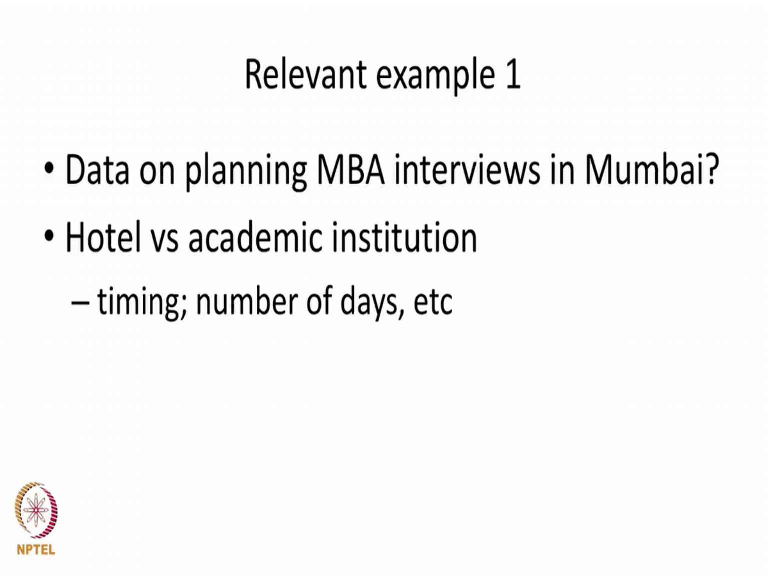
Two more aspects

- Variation in data
  - Height, weight, education, affordability, health, wealth, intelligence, abilities
- Dependencies resulting in model building
  - Linear, complex, non linear
  - Requires different data

So, all these questions are also answered using data, and therefore, we have given you a sample of these kind of questions. There are a couple of more things that we need to look at one should understand that there is a lot of variation in the data. And all the examples that we saw where we looked at data and then did some simple inferences, they also the comparison essentially price to capture the variation in the data. So, variation in height variation in weight, variation education level affordability health wealth intelligence and so on.

The other aspect that we have to look at are some dependencies resulting in model building do these parameters, have a linear behavior or non-linear behavior or do different models require different types of data. So, we need to understand all these aspects as we move along.

(Refer Slide Time: 15:13)



Relevant example 1

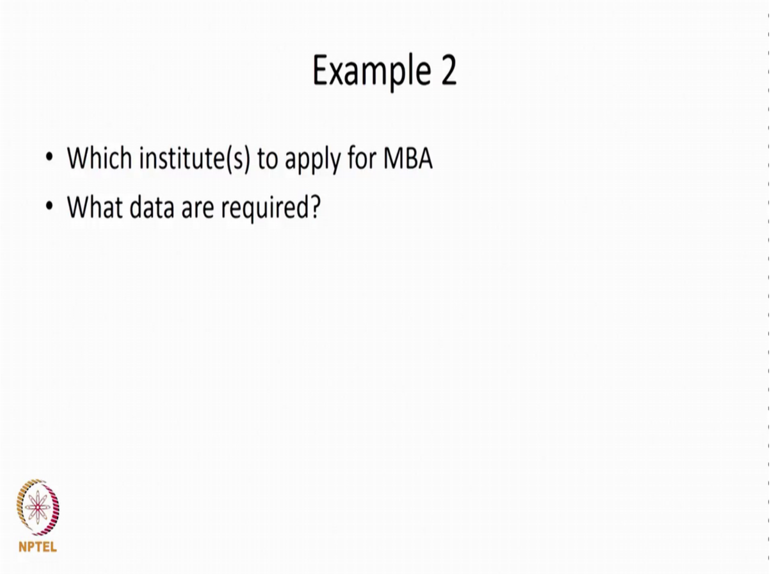
- Data on planning MBA interviews in Mumbai?
- Hotel vs academic institution
  - timing; number of days, etc

So, we could think of at this point, what kind of data would be required for planning events. A simple example could be one could think of if an educational institution wishes to have interviews to select MBA students in Mumbai.

So, what kind of data do we require? Would be a good exercise to understand the number of things that we have seen till now. It could for example, begin with I have just given some examples, it could for example, begin with the timing it could begin with of days and so on. The number of students who are going to be called for interview number of days the interview that possible places, if for example, an IIT is doing it would we do it in in another IIT. Or we do it in some other place that is available some other institutions where some space is available.

It could also depend on as I said the number of students or candidates are going to be called for the interview, the timings, the location. So, all these would result in different kinds of data that is required to carry out an exercise.

(Refer Slide Time: 16:30)



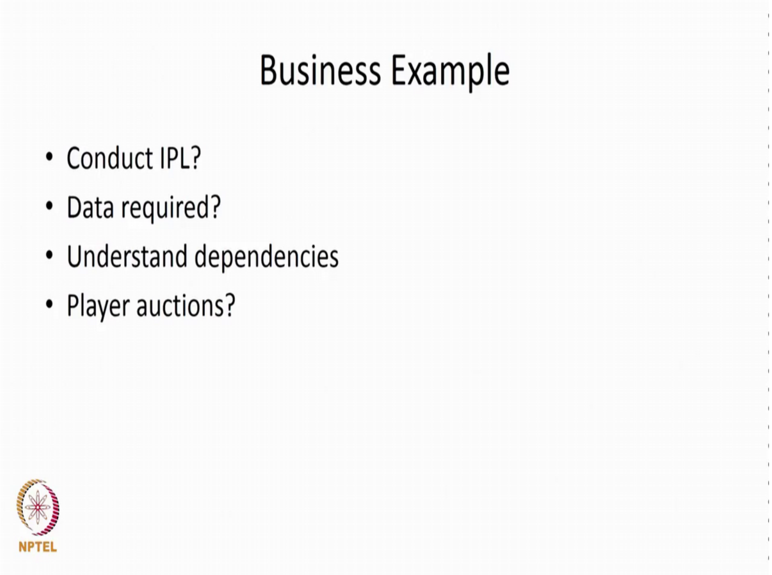
Example 2

- Which institute(s) to apply for MBA
- What data are required?

Another example could be a student aspiring to study MBA, and might want to ask a question which are the institutes to apply for MBA. So, what kind of data are required there? So, the list of colleges that offer an MBA, the qualifying examinations for each one of them, are there multiple exams or do all of them go through the same entrance exam.

The fees that these institutions would charge, the number of seats that are available in each one; The importance given to various aspects such as work experience and so on. So, good exercise at this point is to sit and write about 10 pieces of data that is required for any situation. And I have just described 2 situations right now. So, we could think of several business examples for which we could do this exercise for example, if you looking at conducting big events such as the IPL.

(Refer Slide Time: 17:30)



Business Example

- Conduct IPL?
- Data required?
- Understand dependencies
- Player auctions?

So, one could go back and write about 20 30 different types of data. That could be required to make any decision on this. So, what could be the data required? One has to understand the dependencies on the data, and one could also look at even player auctions as a separate; Example, where we could think about the data that is needed to do this exercise.

So, with this we come to the first end of the first lecture. So, a very quick summary in this lecture we started by defining what is statistics. We will be coming to probability much later in this course. And then we understood the importance of data. We also understood that data helps us to compare and infer. And we also saw some examples of how what type of data as needed, and how this data could help in effective decision making. In the next lecture we would talk about data in more detail, and try to classify data and understand the various types of data, and situations where these types of data could be used.