



IIT Madras

ONLINE DEGREE

Statistics for Data Science - 1
Prof. Usha Mohan
Department of Management Studies
Indian Institute of Technology, Madras

Lecture - 03
Introduction and Types of data – Part 1

In this week 1, the learning objectives are you understand first, why you are learning this course.

(Refer Slide Time: 00:21)

Statistics for Data Science -1



Learning objectives

1. What is statistics?
 - ▶ Descriptive statistics, inferential statistics.
 - ▶ Distinguish between a sample and a population.
2. Understand how data are collected.
 - ▶ Identify variables and cases (observations) in a data set
3. Types of data-
 - ▶ classify data as categorical(qualitative) or numerical(quantitative) data.
 - ▶ Understand cross-sectional versus time-series data.
 - ▶ Measurement scales
4. Creating data sets; Downloading and manipulating data sets; working on subsets of data.
5. Framing questions that can be answered from data.



What is statistics? We are just going to tell briefly about the two main branches of statistics which will be relevant at this point of time to you people will tell what you mean what is understood by descriptive statistics and inferential statistics.

The minute I talk about inferential statistics, I need to introduce what is the notion of a sample and a population. So, that is what I am going to introduce. Then we move on to understand why we need data; we will understand a bit about how data is collected and we will talk about how to organize data in form of what we call a data set.

Once we have a data set, we will understand to more about data by classifying data in terms of categorical and numerical or cross-sectional and time-series, and we will talk a bit we will discuss a bit about measurement scales.

Finally, I think any statistical analysis, the key is to understand your data and frame questions based on data. So, we will focus some time to try and understand and train ourselves to frame questions based on data. So, these are the learning objectives for the week 1.

(Refer Slide Time: 01:36)

Statistics for Data Science -1

- Introduction
- Basic definitions

What is Statistics?

Definition

Statistics¹ is the art of learning from data. It is concerned with the collection of data, their subsequent description, and their analysis, which often leads to the drawing of conclusions.

¹Ross, Sheldon M. Introductory statistics. Academic Press, 2017.

What is statistics? If you go through the definitions of statistics over the years, you can see that there has been a transformation and that has been changing over the period of time. What started as just summarizing data, then afterwards gradually improve to inference from data and then afterwards now with lot of data available, statistics is being redefined as the art of learning from data.

Now, the minute I say learning from data, it includes that you want to seek some information from data. So, Sheldon Ross defined statistics as the art of learning from data, you are concerned with collection of data, subsequent description and their analysis which often leads to drawing of conclusion. So, the main idea of statistics and statistical analysis is to actually draw conclusions based on data.

(Refer Slide Time: 02:44)

Statistics for Data Science -1
└ Introduction
└ Basic definitions

Major branches of statistics


1. Description

Definition
*The part of statistics concerned with the description and summarization of data is called **descriptive statistics**.*

2. Inference

Definition
*The part of statistics concerned with the drawing of conclusions from data is called **inferential statistics**.*

- ▶ To be able to draw a conclusion from the data, we must take into account the possibility of chance- introduction to **probability**.



So, if you look at the classification of statistics, even though there are newer branches of statistics and new titles given, you may broadly classify the branches of statistics or you might broadly look at the main branches of statistics to be two: one way you are describing data that is a part of statistics which is concerned to description and summarization of data more popularly referred to as the descriptive statistics branch.


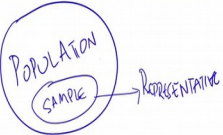
The part of statistics which is concerned with drawing conclusions from data is called the inferential statistics branch that is you want to infer from data. Now, when you want to infer from data, there is one very important thing which is the possibility of chance because when you are inferring from data there is an element of chance you do not have exactly what you are having what you know.

And, hence we are preparing in this course in this foundation course with an introduction to probability, to help you understand or help you prepare for the next league or the next course where will you where you will be learning about inferential statistics.

(Refer Slide Time: 04:11)


Statistics for Data Science -1
└ Introduction
└ Population and sample

Population and sample



Suppose we are interested in knowing

- ▶ The percentage of all students in India who have passed their Class 12 exams and study engineering.
- ▶ The prices of all houses in Tamil Nadu.
- ▶ The total sales of all cars in India in the year 2019.
- ▶ The age distribution of people who visit a city Mall in a particular month.



So, primarily when you talk about inferential statistics, we are trying to talk about drawing of conclusions from data. Now, a branch of inference as inferential statistics, one important thing is many a time you are interested perhaps in knowing about the percentage of all students in India who have passed their Class 12 exams and study engineering; the prices of all households in Tamil Nadu; the total sales of all cars in India in the year 2019; the age distribution of people who visit a city Mall in a particular month.

So, one way of answering all these questions is one is through a complete enumeration – you go and collect data on everybody or everything you are interested. For example, in this question you are interested in knowing about the percentage of all students in India, but very quickly you understand that getting this kind of data might not be very easy.

So, many a time what we are interested in knowing is the percentage of all students in India. Now, if I just want to construct a database and I would want the actual data of all the students who have passed class 12, but if my intention is just to know an overall feel of what are the kind of people who finally, end up taking engineering then one thing I would want to know is work with a smaller subset of all the students in India. All the set of all students in India is what we refer to as a population. A smaller subset of this is referred to as a sample. It is a subset, so, I am putting it as a sample.

Now, many a time you might be wanting to know about the prices of all houses. Again, you need not go and find out about all the houses that have been sold in a particular year; you might want to know about a smaller subset of the entire population. One thing you want about the sample is you want it to be as representative as possible you want the sample to be as representative as possible.

(Refer Slide Time: 06:42)

Statistics for Data Science -1

- Introduction
- Population and sample

Population and sample

Definition
The total collection of all the elements that we are interested in is called a *population*.

Now, what do we mean by representative sample? For example, let me define the population is a collection of all elements that we are interested in. If this is the population so, let me draw different colours here. What is the tool I use?.

So, suppose this is a population and I take another subset here. Suppose I take a subset, this is a subset. The smaller set is actually a subset of the larger set, but we very quickly notice that the smaller set does not have any yellow elements in it. So, I cannot say this smaller set is actually a good representative sample of the larger set.


(Refer Slide Time: 07:39)

Statistics for Data Science -1
└ Introduction
└ Population and sample

Population and sample

Definition
*The total collection of all the elements that we are interested in is called a **population**.*

Definition
*A subgroup of the population that will be studied in detail is called a **sample**.*



So, a sample is basically a subgroup of the population that will be studied in detail.

Now, we need the idea of a population and sample and you will be introduced to this concept of population and sample in greater detail when you do your inferential statistics course. But, nevertheless why do we need the concept of a population and sample in this course is, eventually when we are going to come up with summary statistics, we always need to understand whether the summary statistics is for a population or a sample and this is something which we will know in due course.


(Refer Slide Time: 08:15)

Statistics for Data Science -1
└ Introduction
└ Population and sample

Purpose of statistical analysis

- ▶ If the purpose of the analysis² is to examine and explore information for its own intrinsic interest only, the study is descriptive.
- ▶ If the information is obtained from a sample of a population and the purpose of the study is to use that information to draw conclusions about the population, the study is inferential.
- ▶ A descriptive study may be performed either on a sample or on a population.
- ▶ When an inference is made about the population, based on information obtained from the sample, does the study become inferential.

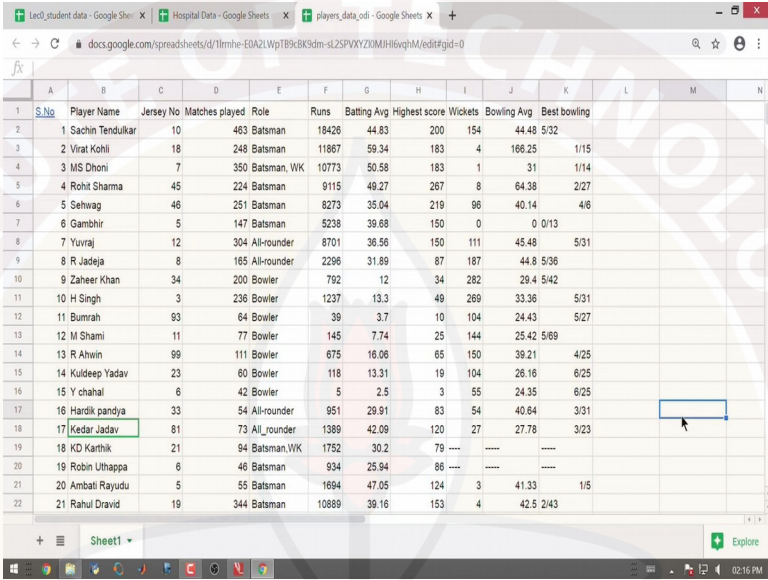
²Weiss, Neil A. Introductory Statistics: Pearson New International Edition. Pearson Education Limited, 2014.



So, what is the purpose of statistical analysis? Now, when would you use a descriptive statistics? When would you use inferential statistics? Now, if the purpose of your analysis is just to examine and explore information for its own intrinsic interest only, this study is descriptive.

Now, what do we mean by that? Let me demonstrate it to you through a data set ok.

(Refer Slide Time: 08:45)



S.No	Player Name	Jersey No	Matches played	Role	Runs	Batting Avg	Highest score	Wickets	Bowling Avg	Best bowling
1	Sachin Tendulkar	10	463	Batsman	18426	44.83	200	154	44.48	5/32
2	Virat Kohli	18	248	Batsman	11867	59.34	183	4	166.25	1/15
3	MS Dhoni	7	350	Batsman, WK	10773	50.58	183	1	31	1/14
4	Rohit Sharma	45	224	Batsman	9115	49.27	267	8	64.38	2/27
5	Sehwag	46	251	Batsman	8273	35.04	219	96	40.14	4/6
6	Gambhir	5	147	Batsman	5238	39.68	150	0	0	0/13
7	Yuvraj	12	304	All-rounder	8701	36.56	150	111	45.48	5/31
8	R Jadeja	8	165	All-rounder	2296	31.89	87	187	44.8	5/36
9	Zaheer Khan	34	200	Bowler	792	12	34	282	29.4	5/42
10	H Singh	3	236	Bowler	1237	13.3	49	269	33.36	5/31
11	Bumrah	93	64	Bowler	39	3.7	10	104	24.43	5/27
12	M Shami	11	77	Bowler	145	7.74	25	144	25.42	5/69
13	R Ahwin	99	111	Bowler	675	16.06	65	150	39.21	4/25
14	Kuldeep Yadav	23	60	Bowler	118	13.31	19	104	26.16	6/25
15	Y chahal	6	42	Bowler	5	2.5	3	55	24.35	6/25
16	Hardik pandya	33	54	All-rounder	951	29.91	83	54	40.64	3/31
17	Kedar Jadav	81	73	All_rounder	1389	42.09	120	27	27.78	3/23
18	KD Karthik	21	94	Batsman, WK	1752	30.2	79	---	---	---
19	Robin Uthappa	6	46	Batsman	934	25.94	86	---	---	---
20	Ambati Rayudu	5	55	Batsman	1694	47.05	124	3	41.33	1/5
21	Rahul Dravid	19	344	Batsman	10889	39.16	153	4	42.5	2/43

This is again another hypothetical data set which is just showing the names of the cricket players. All of us are very well aware of these cricket players – Tendulkar, Kohli, Dhoni. The matches they have played, in what role, what are the total runs, the batting average, the highest score, wickets, bowling average and best bowling.

Now, suppose a purpose is just to understand what are the total runs scored, what is the batting average, what is the who has the highest batting average, who has the highest run scored, who have played the most number of matches, if these who has taken the highest number of wickets – if these are the questions of interest then all these questions of interest which I have just posed now, can directly be just got from the data set.

I might also want to order the number of runs of a batsman has scored; I might want to also know what is among the batsman how have the people scored runs and all of this I can just describe this data. I do not have to do anything more about this data. So, in this case the question I am asking is basically, the purpose I have here the purpose I have

here is to just examine and explore the information that is given. So, the study is just descriptive. I am not asking anything more. I just want to describe the data set that is given here and this study is descriptive.

But, suppose I am using this and one thing which we notice again in this data is the following. If you look at this data this data is not the entire cricketing data about all the cricketers available from all the countries.

It is a sample from an entire population of data. It is just a small sample. I can say it is at best a representative sample of the Indian cricketing data over the last 5 or 10 years or perhaps about this could be about for the in the last decade. It is a sample of definitely, it is a sample of the Indian cricketing data.

But, it is again not the entire population which includes over all batsmen and overall cricketers, but however, if I am just interested in summarizing this data if my inherent interest is just about summarizing this data, then I would be interested in only a descriptive nature of studies for which descriptive statistics is sufficient.

But, now if I am going to use this to draw a conclusions further conclusions; for example, if I want to know about the role a batsman plays with a batting average, I would need more information and I want to pick up a team for the future. For example, you we all know about the IPL auctions and how people are chosen. So, there is a further role. I am just not interested in describing this data.

The bigger role for me or the bigger interest for me is to use this data to gather or infer some information which I am going to use in my decision making process. For that I would I am going to have an element of chance and there I am going to have what I need is I am going to have an inferential study in that case. So, very often we see that a descriptive study we need to understand whether our nature of a study is only going to be descriptive or whether we want to do an inferential study.

When we come to inferential study a descriptive study sorry when we come to for a descriptive study it might be either performed on a sample or on a population. Since in the classes to come we will be talking about descriptive statistics in detail. We need to understand whether a descriptive study is performed on a sample or on an entire

population that is the reason why we introduced the notion of a sample and a population at this stage.

However, if our inference is to be made about a population based on the sample, then the study becomes inferential. Inferential statistics is not the scope of this course, but however, you will be introduced to the concept of probability which will help you develop the methodology towards inferential statistics.

(Refer Slide Time: 13:49)

Statistics for Data Science - I

- Introduction
- Population and sample

Summary

- ▶ Descriptive statistics
- ▶ Inferential statistics
- ▶ Population and sample

So, in summary, you should know the two main branches are descriptive statistics, inferential statistics. You are going to do a descriptive study or inferential study based on what is your purpose of study.

If your intrinsic purpose is just to summarize your data, you would go for an descriptive statistic. But if your purpose of study is to infer into the future or infer about a larger population using a smaller subset, you would go for inferential statistic. To do understand inferential statistic, you need to understand what is the concept of a population and sample.