



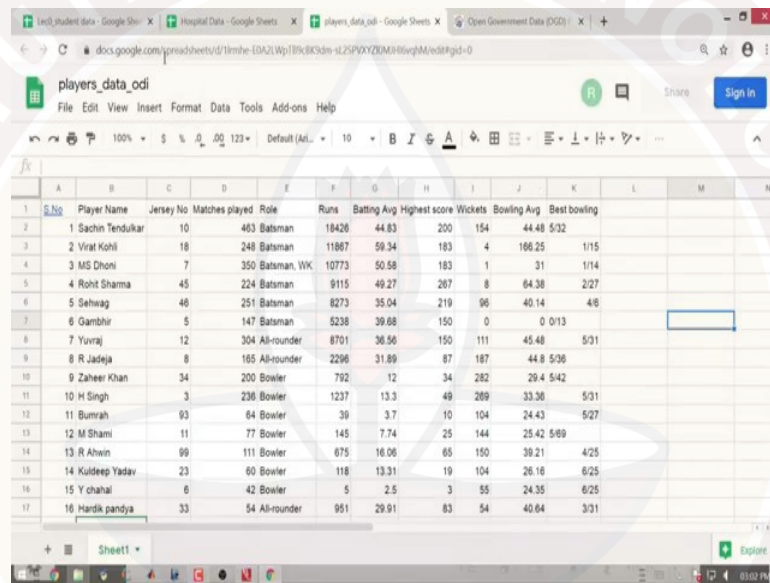
IIT Madras
ONLINE DEGREE

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

Lecture - 05
Introduction and Types of data Part – 3

If you are come up to this module what would I expect you to know is you know what is data and you know how the data is organized as a data table.

(Refer Slide Time: 00:26)



The screenshot shows a Google Sheets spreadsheet titled 'players_data_odi'. The spreadsheet contains a table with 11 columns: S.No, Player Name, Jersey No, Matches played, Role, Runs, Batting Avg, Highest score, Wickets, Bowling Avg, and Best bowling. The data is organized into rows, with each row representing a player's statistics. The table is displayed in a standard Google Sheets interface with a menu bar at the top and a toolbar below it.

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	289	33.36	5/31
11	Bumrah	93	84	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 Ashwin	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

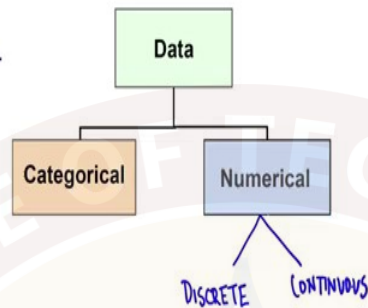
And when I say it is organized as a data table I mean that you understand that this is the data where I have columns representing variables and rows representing the cases or observations.

(Refer Slide Time: 00:39)

Statistics for Data Science -1
└ Classification of data
└ Categorical and numerical



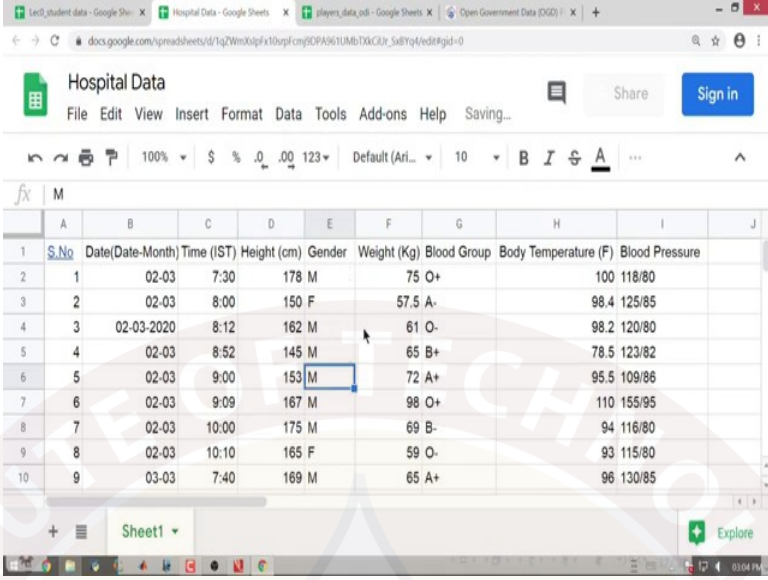
Categorical and numerical



Now, once we go back to this, you can see that again let us go back to the data set. The minute you look at a data set of this kind, you can see that when I look at name it is just Anjali, Pradeep, Varsha, Divya, I have gender. When I look at gender I have two categories female and male. When I have marks, you can see that this marks. Let me put back this mark here, let me put it as say 565 and remove the percentage.

So, when I look at marks, you can see that there is 484, 514, 565 etcetera, but state board again it is again some sort of a category here. I have a State Board, I have ICSE, I have CBSE. Marks in class 12 again 394, 437 again the board etcetera.

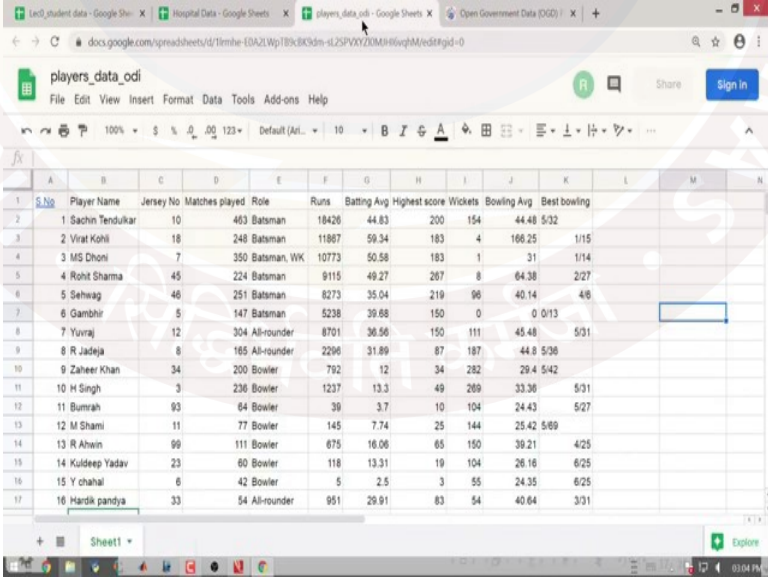
(Refer Slide Time: 01:35)



S.No	Date(Date-Month)	Time (IST)	Height (cm)	Gender	Weight (Kg)	Blood Group	Body Temperature (F)	Blood Pressure
1	02-03	7:30	178	M	75	O+	100	118/80
2	02-03	8:00	150	F	57.5	A-	98.4	125/85
3	02-03-2020	8:12	162	M	61	O-	98.2	120/80
4	02-03	8:52	145	M	65	B+	78.5	123/82
5	02-03	9:00	153	M	72	A+	95.5	109/86
6	02-03	9:09	167	M	98	O+	110	155/95
7	02-03	10:00	175	M	69	B-	94	116/80
8	02-03	10:10	165	F	59	O-	93	115/80
9	03-03	7:40	169	M	65	A+	96	130/85

Let us go to the other hospital data. Height, again I have centimeter. So, let me rephrase this into centimeters I put it as 152, 167 and I have a 175 centimeter here, gender again male female, weight is again you can see 75, 57.5, 65, 98. There is something called blood group, body temperature in degree Fahrenheit and you have blood pressure.

(Refer Slide Time: 02:12)



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	289	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 Ashwin	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

So, the minute you look at the data of this kind. Let us look at the cricket data, you again have a bowling average which is 44.48, 166.25, 64.38 whereas, matches played you have 463, 248, 350. Now, here you have a jersey number which is 10, 18, 7 etcetera. So, the

minute the you look at a data set of this kind two things you observe. The first thing you observe is you have numbers, you have text. By text I mean that you have names, you have role of a batsman, all rounder, go back to your blood group you have O plus A minus, you have gender which is again captured as male and female, you have board again which is captured as CBSE state board.

So, you immediately see that all data is not of the same kind. You see that there is a basic difference in the way the data is presented. So, the next thing is we seek whether I can classify this data into broadly two categories.

Immediately I notice that some data is numerical in nature or quantitative in nature. For example, I see that the marks and the bowling averages the height, weight, the bowling average, the highest score, the batting average, the matches played, etcetera can be clubbed into some kind of variable whereas the name the gender, the board, the blood group and here the role. These represent certain kind of variables.

Interestingly if you look at this jersey number, it appears to be numbers; but all of us know that these numbers have no meaning. So, there are kinds of variables which could take numerical values as in this thing they could be numbers, but they might not.


So, there is definitely a difference between the variable, a jersey number and matches played even though both are numbers. So, it is very important for us to understand how to classify data and to what category or what type of data my variable belongs to extremely important. So, when we look at data, data is broadly classified into two categories; categorical data and numerical data.

(Refer Slide Time: 05:17)

Statistics for Data Science -1
└ Classification of data
└ Categorical and numerical

Categorical and numerical variables

- ▶ Categorical data
 - ▶ Also called qualitative variables.
 - ▶ Identify group membership
- ▶ Numerical data
 - ▶ Also called quantitative variables.
 - ▶ Describe numerical properties of cases
 - ▶ Have measurement units
- ▶ Measurement units: Scale that defines the meaning of numerical data, such as weights measured in kilograms, prices in rupees, heights in centimeters, etc.
 - ▶ The data that make up a numerical variable in a data table must share a common unit.



So, when we look at categorical data, these are also called as qualitative variables. Now, it identifies group membership. What do we mean by group membership? Again we go back to our student data, let us look at gender. Gender is a categorical variable. I have two categories here. I can classify any observation into one of these two categories.

So, it is a group membership. Similarly, when I look at board I have a category which is a State Board, I have ICSE, I have CBSE. So, again you can see that this categorical variable has three categories and any observation can be categorized into one of these three groups.

So, when we go back you can see that you in a sense, I am giving membership of an observation to a particular group in that particular variable. So, this category has groups. Let us go to the hospital data. You see that blood group every patient is either an O positive or an O negative or a B positive or a A positive or A negative.

So, you can see that there are many blood groups I again this is a categorical variable; gender is a categorical variable. What kind of variable is mobile number? I leave it as an exercise. I want you people to come up with an answer; mobile numbers are again numbers. What kind of a variable do you think is a mobile number?

Similarly, what kind of a variable do you think is the jersey number? Even though jersey number is 10, 18, 7, what kind of a variable is it? Is it a categorical variable or is it a numerical variable?

So, now, the first thing which we need to understand is I have categorical data I also have what are numerical data. When we have numerical data, numerical data is also called quantitative variables. Here I can talk about numerical properties of data.

Now, go back here. Marks obtained both in class 10 and class 12 are numerical data and you talk about marks this is 484 marks, this is 514 marks, this is 565 marks. Come to the hospital data I have 178 centimeters, 150 centimeters; weight 75 kilograms, 57.5 kilograms; body temperature in terms of Fahrenheit degree Fahrenheit 100 degree Fahrenheit, 98.4 degree Fahrenheit. When I come to cricketing data matches played, I have batting averages, I have wickets taken 154 run score 200 runs.

So, you can see immediately when I talk about numerical data, I have associated with them either measurement units or I have something which are called the bowling average and batting average. Now, you also see that when I talk about matches, it is a whole number it is 463 whereas, when I talk about batting averages you can see that it can take any value. It could be fractions also or it is any value. So, this again tells us that when we talk about categorical and numerical data; within numerical data, I could have discrete data and I could have continuous data. I could further look at data that is discrete and I could look at data that is continuous right.

So, once we understand what is this categorical and numerical data, we need to understand the measurement units that are used for numerical data. Again let us go back to our data, you can see that here height is measured in centimeter, weights is measured in kilograms, the body temperature in degree Fahrenheit. Again we have marks again this is 484 marks. Again when you come to players data you have matches played, highest score; it is in runs wickets taken again in wickets ok.

So, the idea is we need to understand what is the scale that defines the numerical data. Again we have already emphasized on the point that when you have numerical data which take units. We need to ensure that the variable is measured across all observations and shares a common unit. This is something which we need to ensure.

(Refer Slide Time: 11:15)



Cross-sectional and time-series data

- ▶ Time series - data recorded over time
- ▶ Timeplot – graph of a time series showing values in chronological order
- ▶ Cross-sectional - data observed at the same time



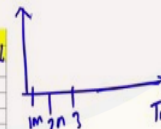
Apart from categorical data and numerical data, we also have data which are referred to as time series data.

(Refer Slide Time: 11:27)



Time-series data- Example

Date	Potato		
	Qty(kg)	cost (Rs.)	Selling price(Rs.)
01-Mar	0	21	24
02-Mar	1300	20.05	24
03-Mar	675	20.5	24
04-Mar	0	NA	NA
05-Mar	675	20.8	24
06-Mar	675	21.25	24
08-Mar	20	20.5	24
09-Mar	900	20.5	24
10-Mar	900	20.5	24
11-Mar	0	NA	NA
12-Mar	900	20.3	24
13-Mar	1125	19.4	22
15-Mar	1125	18.8	22
16-Mar	1125	19.4	22
17-Mar	1125	19.25	22
18-Mar	1125	20.3	24
19-Mar	1125	19.8	24
20-Mar	675	21.25	24
22-Mar	675	20.5	24
23-Mar	0	NA	NA
24-Mar	0	NA	NA
25-Mar	675	19.6	24
26-Mar	675	19.7	24
27-Mar	1125	19.3	24
29-Mar	540	20.6	26
30-Mar	0		28



Let us look at this data. Now, when we look at this data, you can see that this data has a variable which is called date. And this data is actually the data which tracks at a retail outlet, what is the quantity that is procured every day from 1st March to 30th March of a month, the cost of procurement and the price at which it was being sold.

If you look at it the variable is just one thing which is potato and you for all the days you are tracking what is the quantity that is being sold. So, in other words I have a date; I have the first march, I have second march, I have third march. I have a time and over that time, I can actually find out what is the quantity that is being procured every day.

So, this is what we refer to as a time series data where the data on a particular variable; this could be the quantity procured on potato is obtained the variable is the same. What is the variable? It is quantity of potatoes that is the variable, but I am tracking this variable over a period of time which is from 1st March to 30th March. So, this kind of data is what we refer to as a time series data whereas, cross sectional data is the data which is observed at the same time.

(Refer Slide Time: 13:24)

Statistics for Data Science -1

- Classification of data
 - Cross-sectional versus time-series data

Summary

- ▶ Classify data as categorical or numerical.
- ▶ For numerical data, find out unit of measurement.
- ▶ Check whether data is collected at a point of time (cross-sectional data) or over time (time-series data).

So, we will to broadly classify, we should know that given data we classify them broadly as categorical or numerical. So, whenever we are presented with a data set, we should be able to classify all the variables in the data set as a categorical variable or a numerical variable. If it is a numerical variable, find out what is the unit of measurement. Again check if the unit of measurement is consistent across all the observations. The third thing is check whether it is collected at a point of time whether it is cross sectional data or whether it is time series data. So, now, given a data set you should be able to do this.