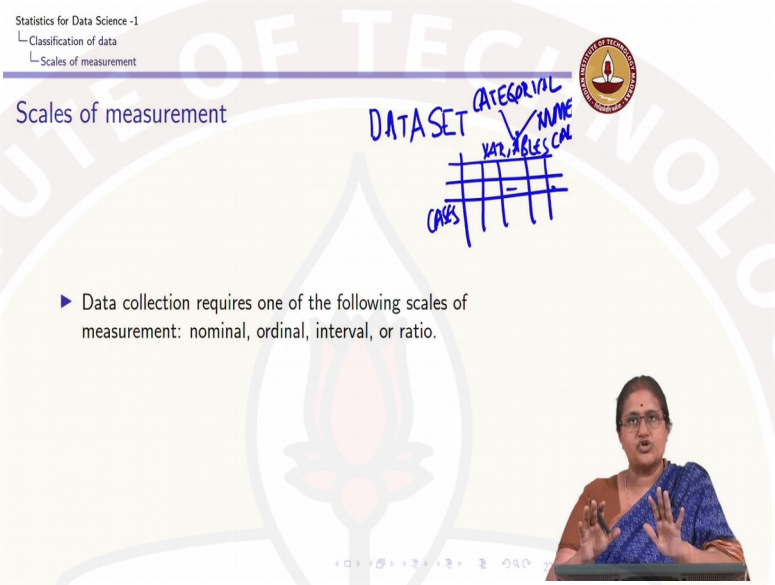
IIT Madras

ONLINE DEGREE

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

Lecture - 06
Introduction and Types of data Part – 4

(Refer Slide Time: 00:16)



So, when you have come here I assume at this point of time you know what is data, you know you have a good data set. By meaning you have a tidy data set what I mean is you have actually a data set which is organized into variables and you have cases. Each case is recorded for each variable and if I do not have a data, I will not record that data. So, this is a tidy data set.

Further, from this data set you are able to look at the variables, you are able to broadly classify these variables as categorical variables and you are able to classify them as numerical variables. So, at this point of time this is what you should be able to do with your data set.

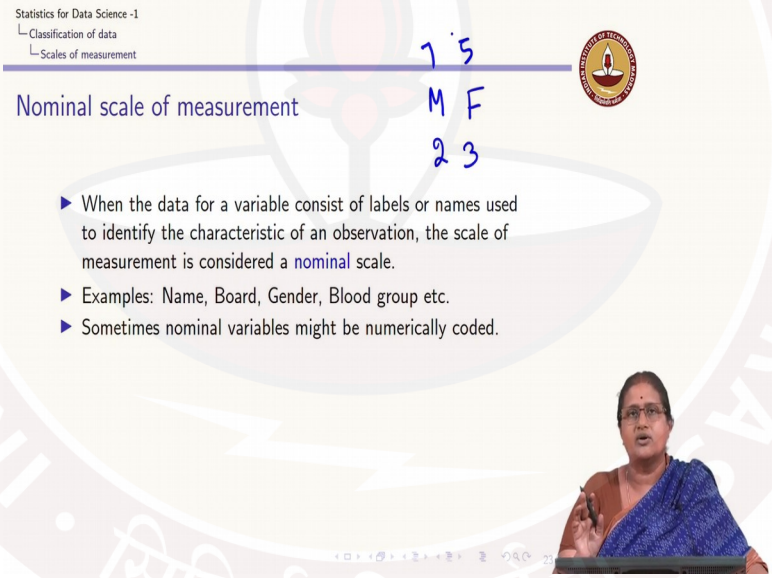
Now, what do we want to do next? Now, we again examine each of our variables in greater detail because if you look at the definition of the statistics which we gave earlier, it said that you want to learn from data. Now, when I want to learn from data the immediate thing is I want to see whether I can summarize this data. Again, when I say I want to summarize this data, the question I ask is can I come up with some graphical

summaries and can I come up with numerical summaries. The minute I say numerical summaries, I need to know whether I can do arithmetic operations on the data, ok.

So, to know whether I can do arithmetic operations on data, I need to understand what are the scales of measurement I use for my data. Now, when we look at scales of measurement, I have 4 scales of measurement and they are called the nominal, ordinal, interval and ratio scale. We are going to understand about each of these scales of measurement in great detail.

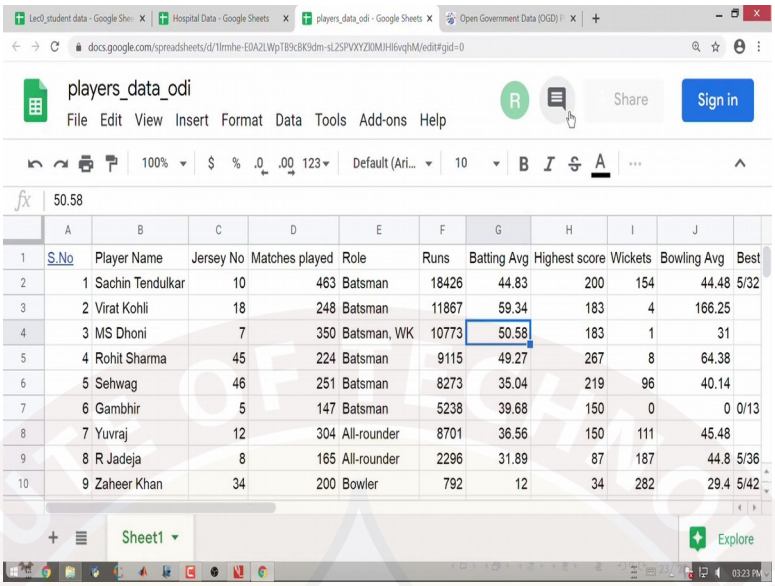
Why is it important? It is extremely important for us to know what is the scale of measurement for each of the variables I have in my data set to eventually come up with what is the kind of summary I can do for that variable. Hence, it is extremely important for us to know what is the scale of measurement for each variable.

(Refer Slide Time: 03:01)



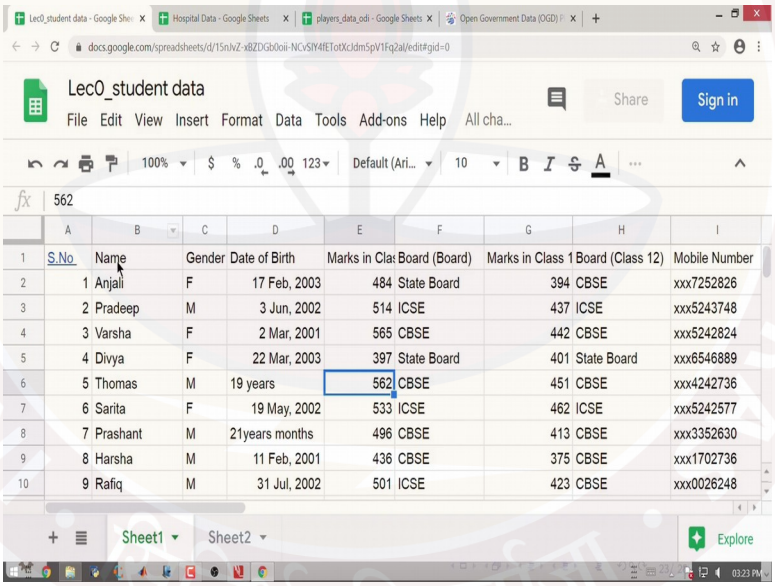
We start with the nominal scale. When the data consists of labels or names, the scale of measurement is considered a nominal scale.

(Refer Slide Time: 03:16)



	S.No	Player Name	Jersey No	Matches played	Role	Runs	Batting Avg	Highest score	Wickets	Bowling Avg	Best
2	1	Sachin Tendulkar	10	463	Batsman	18426	44.83	200	154	44.48	5/32
3	2	Virat Kohli	18	248	Batsman	11867	59.34	183	4	166.25	
4	3	MS Dhoni	7	350	Batsman, WK	10773	50.58	183	1	31	
5	4	Rohit Sharma	45	224	Batsman	9115	49.27	267	8	64.38	
6	5	Sehwag	46	251	Batsman	8273	35.04	219	96	40.14	
7	6	Gambhir	5	147	Batsman	5238	39.68	150	0	0	0/13
8	7	Yuvraj	12	304	All-rounder	8701	36.56	150	111	45.48	
9	8	R Jadeja	8	165	All-rounder	2296	31.89	87	187	44.8	5/36
10	9	Zaheer Khan	34	200	Bowler	792	12	34	282	29.4	5/42

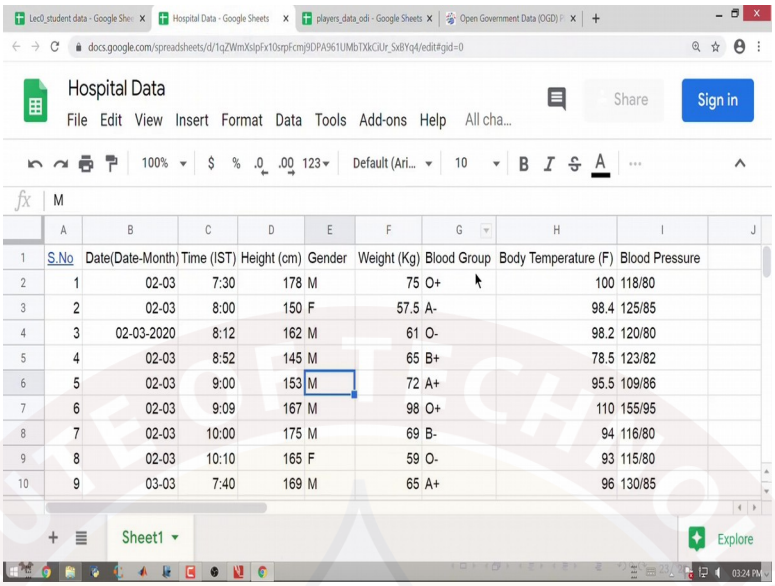
(Refer Slide Time: 03:19)



	S.No	Name	Gender	Date of Birth	Marks in Cla	Board (Board)	Marks in Class 1	Board (Class 12)	Mobile Number
2	1	Anjali	F	17 Feb, 2003	484	State Board	394	CBSE	xxx7252826
3	2	Pradeep	M	3 Jun, 2002	514	ICSE	437	ICSE	xxx5243748
4	3	Varsha	F	2 Mar, 2001	565	CBSE	442	CBSE	xxx5242824
5	4	Divya	F	22 Mar, 2003	397	State Board	401	State Board	xxx6546889
6	5	Thomas	M	19 years	562	CBSE	451	CBSE	xxx4242736
7	6	Sarita	F	19 May, 2002	533	ICSE	462	ICSE	xxx5242577
8	7	Prashant	M	21years months	496	CBSE	413	CBSE	xxx3352630
9	8	Harsha	M	11 Feb, 2001	436	CBSE	375	CBSE	xxx1702736
10	9	Rafiq	M	31 Jul, 2002	501	ICSE	423	CBSE	xxx0026248

Again, let us go back to our example. You can see in the first example I have name which is evidently a nominal scale because it is only names, ok. Again board you can see that it is labels in some sense I have a state board I have a ICSE, I have a CBSE, ok. Whereas gender, again it is a label, I am labelling this category with a female and a male.

(Refer Slide Time: 03:57)



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

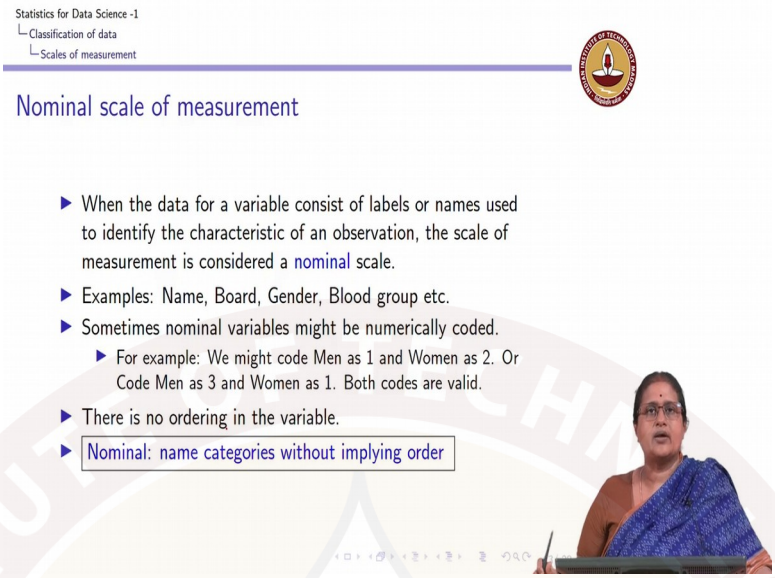
So, you can see that names, board, gender. We go back here, we have the blood data. In the blood group data you see again I can label it as O positive, A minus, O negative. One thing we need to notice here is I just have labels, I have only names; there is no particular order of these names.

For example, the data would not have made any difference that is whether I am having a female male or a male female. It is absolutely it does I am just having this as a label to identify the characteristic. So, this is the called a nominal scale of measurement. Sometimes we can see that nominal variables may be numerically coded. What do I mean by a numerically coded nominal variable?

For example, we have gender. Gender takes two labels which is male and female. I might code a male a 0 and a female a 1, I might code a male a 1 and a female is 0. So, this numerically coded is again equivalent to just labelling this variable, this numeric; no sanctity about having a 0 or a 1.

I could label a man or 2 and woman a 3, or a woman a 5 and man a 7. All it says is this label has the same understanding; these numbers have no meaning when you are coding the nominal variables. Both the codes are valid that is what we mean.

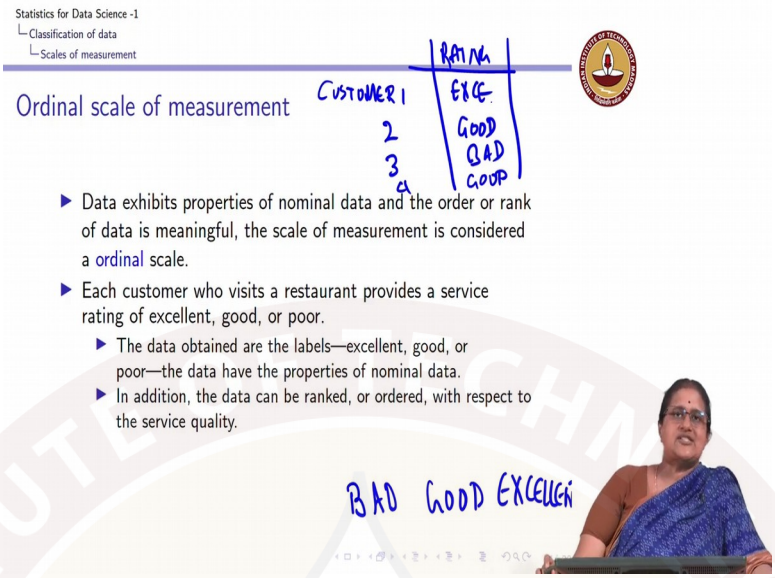
(Refer Slide Time: 05:40)



When no nominal variables are coded whether 1 or 2 or 3 and 1, it has both codes are valid. There is no ordering. This is extremely important that we understand when I talk about nominal variables; there is no ordering in the variable. So, a nominal variable is just name categories without implying order.

Going back to a data sets in the student data set, name, gender, board are nominal variables. In a blood bank data set, we have gender and blood group which are nominal variables. In the cricketing data set jersey number is a nominal variable. The role of batsman is a nominal variable, that is and the player name is a nominal variable. So, nominal scale of measurement is used when I have name categories without implying any order.

(Refer Slide Time: 06:55)

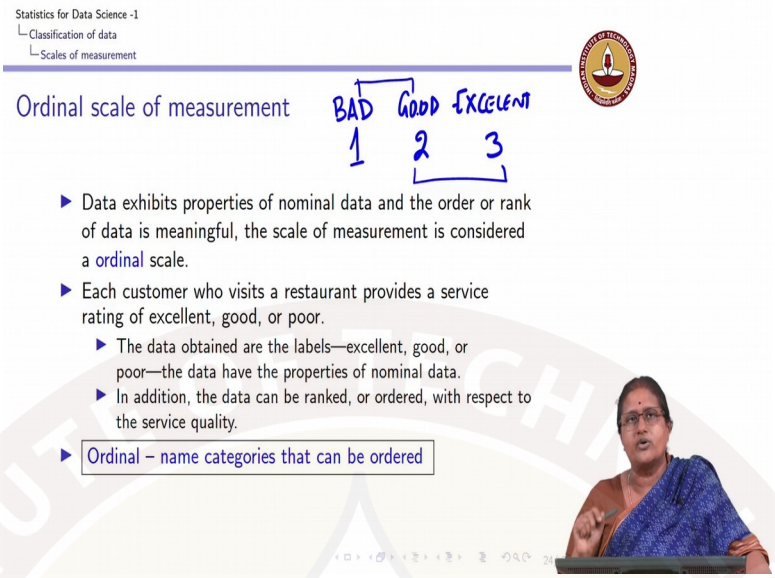


The next scale of measurement is called an ordinal scale of measurement where the data exhibits the same property of nominal data, but here an order or rank is meaningful. What do I mean by this is for example, a customer who visits a restaurant provides a service rating of excellent good and poor. The data are again labels.

The data is again labels. For example, I can have a data where I have customer 1 who gives a rating of excellent, customer 2 good, customer 3 bad, customer 4 again good. So, if you look at the variable, the variable is rating.

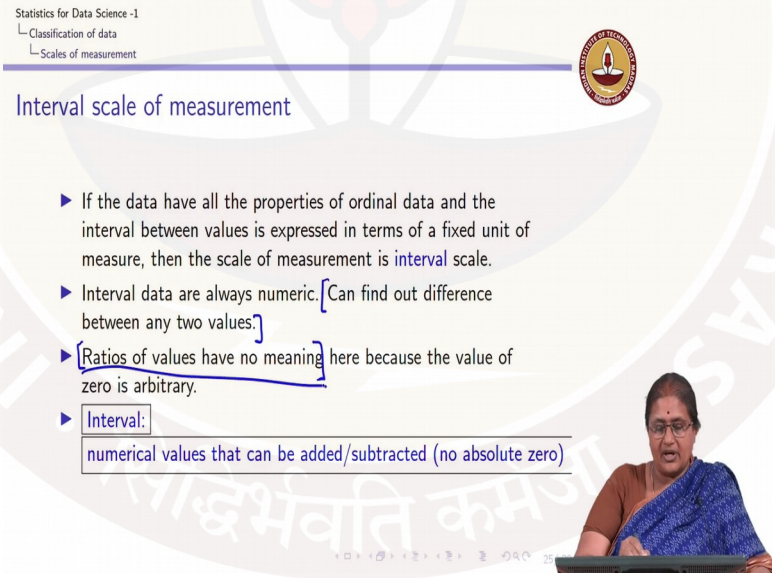
Here you can again see this is taking a nominal value. By nominal it is taking a categorical value where my categorical variable has 3 categories; excellent, good and bad, but within this categorical variable there is an order. You know the order is bad, good and excellent. So, categorical or nominal data which exhibit some rank or an order or rank is meaningful is said to have the measurement, the scale of measurement is said to be a ordinal scale.

(Refer Slide Time: 08:42)



So, ordinal scale of data is name categories that can be ordered.

(Refer Slide Time: 08:53)



The next scale is called an interval scale of measurement. Now, when we talk about ordinal scale, again I can code an ordinal scale of measurement. For example, in our earlier example my bad could have been coded as 1, my good can be coded as 2, and my excellent can be coded as 3. I could code them.

There is an order in 1, 2, 3. But then one thing which I need to understand here is the distance between bad to good need not be the same as the distance between good and

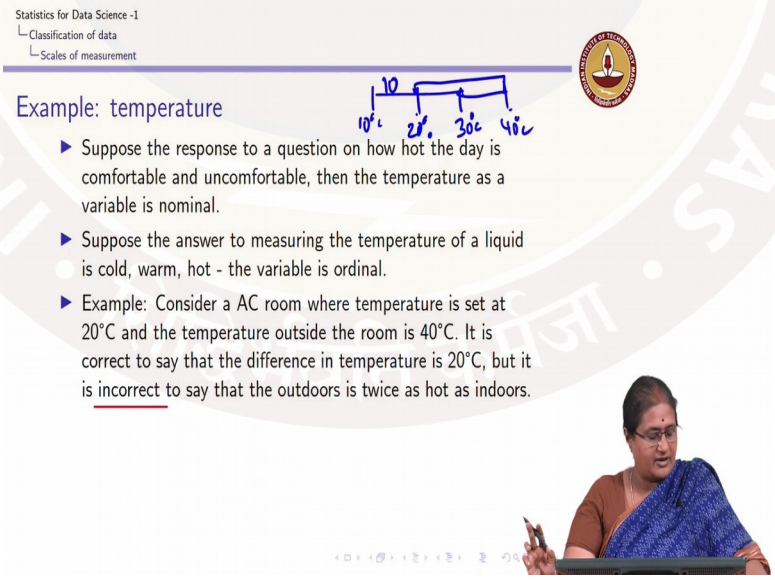
excellent. It is just an order, I know excellent is better than good, but I cannot say that excellent the difference between good and excellent is the same as the distance between good and bad. I have an order, but at this point of time I am not able to comment anything more about this order.

So, when I go to interval scale of data, interval scale of data has all the properties of interval scale of data, but the interval between the values is expresses a fixed unit of measure. Remember, when I said bad, good and excellent, I said the difference between good and bad need not be the same as the difference between excellent and good.

Whereas, when I have internal data I have an ordering, but in this case whenever I am ordering my data the interval between the values is expressed in units of a fixed unit of measure. When the differences expressed in a fixed unit of measure; so, if I have for example, it can I can find out the difference between any two values, ok.

Now, here ratios do not have any meaning because the value of 0 is arbitrary. Let us explain this through an example. Interval data and numerical values that can be added or subtracted, it has no absolute 0.

(Refer Slide Time: 11:05)



Let us look at temperature as an example. Suppose, the reference or response to a question is how hot the day is, and you respond as just comfortable uncomfortable or just good, bad. I am just giving a label to my feeling. I really I am not grading or ordering

this feeling I have, whether I am just telling it is comfortable it is categorized into comfortable and uncomfortable or it could be just leisurely or the something. But in a sense you might feel that there is an order, but at this point of time I am not having any order between comfortable and uncomfortable. Temperature is a variable of interest and here you can see temperature is a nominal variable.

Now, suppose temperature is again the variable of interest, but here I am interested in knowing how hot a liquid is, whether it is cold, warm, or hot; you see that there is a order in this variable. I know cold has warm is warmer than cold or hot is warmer than warm, the variable is ordinal. However, here I do not know whether the difference between a warm and a cold beverage is the same as a hot and a warm beverage.

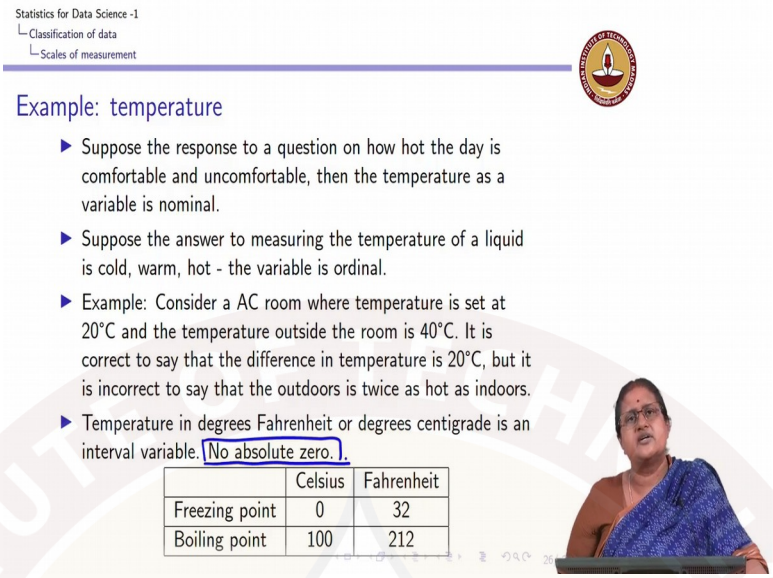
But now suppose I am measuring the temperature, consider an AC room which is set at 20 degree centigrade and temperature out of their outside the room is 40 degree centigrade it is correct to say that the difference in the temperature is 20 degree centigrade, absolutely fine.

Suppose, I had set it at 10 degree, 14 degree centigrade and the temperature outside was 28 degrees, it is perfectly right for me to tell that there is a difference in the temperature of 14 degree centigrade. But it is incorrect; it is absolutely incorrect for me to say that outdoors is twice as hot as indoors because 40 degree centigrade is not twice as hot as 20 degree centigrade.

So, when I am talking about an interval scale I know in an interval scale, I know the difference between 10 degree centigrade and 20 degree centigrade is 10, which is same as a difference between 20 degree centigrade and 30 degree centigrade, which is same as a difference between 30 degree centigrade and 40 degree centigrade.

But I cannot make a statement that 40 degree centigrade is twice as hot as 20 degree centigrade. It is incorrect. So, when I am able to; so, that tells me that I can talk about the difference between any two values, but here ratios have no meaning.

(Refer Slide Time: 14:13)

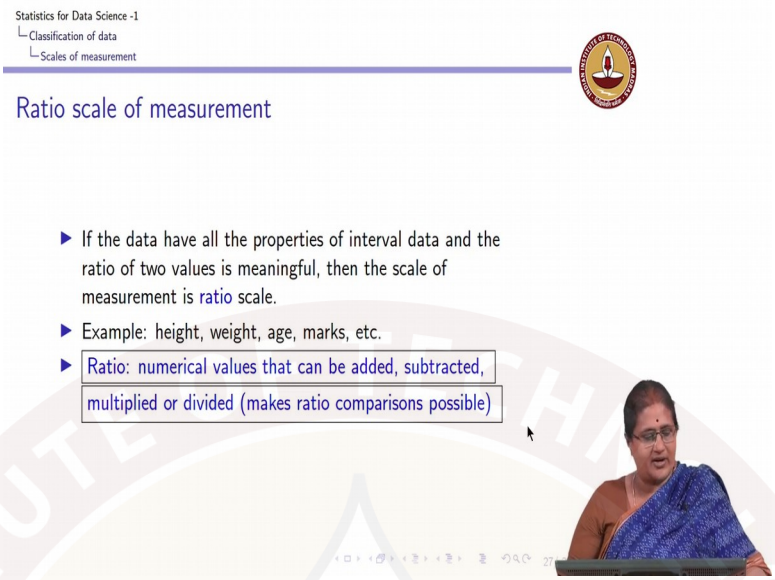


Again, we understand from temperature, at least when we talk about Celsius and Fahrenheit scales there, there is no absolute 0, in the Celsius 0 and 100 are set to be as the freezing point and the boiling point whereas, in Fahrenheit it is 32 and 212. Only in the Kelvin you have a 0 degree, where 0 means absolutely no temperature. But when you are talking about Celsius and Fahrenheit we understand that there is no absolute 0.

So, when you talk about an interval scale, it is extremely important for us to understand there is no absolute 0. However, the difference between an interval scale and an ordinal scale of measurement is in an interval scale the difference between the values is fixed unit of measure whereas, for a ordinal scale that need not be a fixed unit of measure that is good to bad need not be the same difference as excellent to good. This is the key difference.

The last scale of measurement is what we refer to as the ratio scale of measurement.

(Refer Slide Time: 15:25)



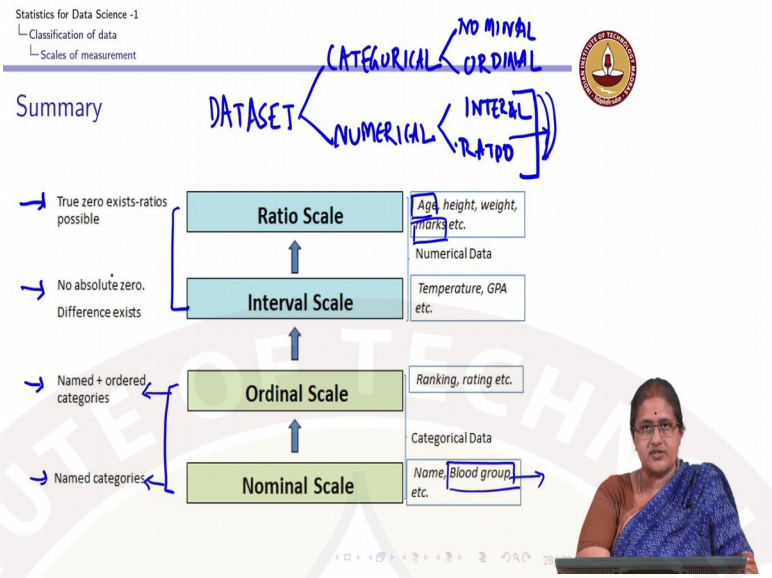
In a ratio scale of measurement it has all properties of interval data and the ratio is a very meaningful measure. The scale is a ratio scale. So, the example, height, weight, marks. Like I know a person who has scored 300 has scored twice as well as a person who has scored 150 marks.

I know a person with a score of 200 is highest score of 200 is twice as good as a person who has scored 100 runs. I know a person who has taken 100 wickets is twice as good as a person who has taken 50 wickets.

So, I can have a notion of a ratio which I can define here, the variables, height, weight, marks, runs, wickets all of them are examples of ratio scale of measurement. So, ratio when you have a variable which is measured in the ratio scale; you can do all the mathematical or arithmetic operations on it.

You can add, you can subtract, you can multiply or divide. Whereas, when you talk about interval scale you can only talk about difference or you can add and subtract. You have no absolute 0, so ratios do not have any meaning.

(Refer Slide Time: 17:04)



Whenever you are presented with the data set after you identify the variables as categorical or numerical, it is extremely important for us to understand that when I have categorical data, I have the nominal and ordinal scale. Within the nominal scale, I have nominal which is only named category, ordinal is a name with an order. Here the difference between order is not a fixed measure.

Again, example for categorical data name, blood group or nominal scale; ordinal scale ranking, rating; there is a order, but then there need not be a fixed order in the rating. Absolute 0 does not exist for an interval scale. This is for numerical data or quantitative data. Absolute 0 does not exist, but proportional difference exists.

I can have the mathematical operation of difference done here. And in the ratio scale, ratios are possible. All variables of age, height, weight, marks etcetera are variables which can be measured on a ratio scale. Temperature, grade point, average everything is a 4 GPA is not twice as good as a 2 GPA, they can be measured on a in interval scale

So, whenever we are given data why are we interested about the scales of measurement? Here you can see no arithmetic operations possible. Here I can have some sense of an order. Here I can do addition, subtraction. Here I can do all arithmetic operations. So, the minute I identify the variable the type of questions, I asked makes sense.

For example, when I have a variable which is a blood group I would not be asking the question of what is the average blood group because I cannot define any arithmetic operation here. Similarly I need to, but when I am talking about a age or a mark, I can ask about some numerical summaries depending on what is this scale of measurement.

So, with this we stop this module. At the end of this module you should be able to look at a data set which is very well organized, identify variables as categorical or numerical. And once you are able to identify these variables further look at what are the scales of measurement whether it is a nominal, whether it is an ordinal scale of measurement or whether it is an interval scale of measurement or ratio scale of measurement.

Most of the textbooks and books written in statistics clubbed; both of these scales together and mentioned it either as an interval or ratio scale. But nevertheless there is a difference, and the critical difference is in an interval there is no absolute 0, whereas in a ratio scale an absolute or a true 0 exist.