

**Introduction to Probability and Statistics**  
**Prof. G. Srinivasan**  
**Department of Management Studies**  
**Indian Institute of Technology, Madras**

**Lecture - 03**  
**Categorical Data**

In this lecture we continue the discussion on data. In the previous lecture we describe data. We also categorized data; we categorize them as qualitative and quantitative. We also categorize them as categorical data and numerical data. And within the framework of categorical and numerical data, we further classified them into nominal ordinal interval and ratio. Now we will take some examples to understand these types of data in more detail. It is also important that in statistics the more examples we look at the more variety of situations we look at, our understanding of the concepts get better.

So, first we will learn the concepts, and then we also try to apply them to some situations, and we will have some kind of a tutorial session. On each of this topic and the first part of this lecture would act as such a tutorial where we try to apply and try to solve, some simple problems to understand what we learnt in the earlier lecture. So, with this let us begin this.


(Refer Slide Time: 01:40)

**Cases and variables**

In a data table, rows are called cases or observations while columns are variables.

Example

Name	Gender	Age	CAT score	Work experience
Akhil	M	27	95.27	5



The slide contains a diagram illustrating the relationship between cases and variables in a data table. The title is "Cases and variables". Below the title, it states: "In a data table, rows are called cases or observations while columns are variables." An example data table is provided with the following structure:

Name	Gender	Age	CAT score	Work experience
Akhil	M	27	95.27	5

Annotations on the table: A blue arrow labeled "cases" points to the first row (Akhil, M, 27, 95.27, 5). A blue arrow labeled "variable" points to the "Work experience" column.

So, what we saw in the last lecture is that when we create data tables, we have rows and columns the columns are called variables, we saw the same example in the last lecture.


So, variable names such as name gender age, cat score or score in a competitive exam, and work experience could act as variables; while cases or observations are specific to individuals for example, in this table. And if there is at a student or a candidate by name Akhil, then you can have a case where the name is Akhil the gender is male the age is 27. The score is 95.27 and work experience is 5. So, we also saw that cases are observations are rows and variable names are columns, and in general we will have more of cases and observations in a given table than the number of variables.

(Refer Slide Time: 02:43)

### Exercise

For the following variables, give a name and indicate the type of variable (categorical, ordinal, numerical)

1. Car owned by ten friends
2. Income of 20 employees
3. Size of clothes as S, M, L, XL
4. Number of students absent for class
5. Education of people as High School, Graduate, PG, PhD




So, let us do a small exercise, and let us say for the following variables give a name and indicate the type of variable whether they are categorical ordinal numerical as the case may be. So, a simple example could be car owned by 10 people. Another example would be income of 20 employees. A third situation could be size of clothes if you go to a garment shop and start looking at shirts, you could find small medium large extra-large and so on.

You could think of number of students who are present in a class or who are absent in a class. You could think of education of people such as studied up to high school, graduate or a postgraduate or a PhD and so on. Each of these we try to give a name and indicate the type of variable which this name belongs to.

(Refer Slide Time: 03:39)

No.	Description	Variable Name	Variable type
1	Car owned by ten friends	Model (or Brand)	categorical
2	Income of 20 employees	Salary	numerical
3	Size of clothes as S, M, L, XL	Cloth size	ordinal
4	Number of students absent for class	Absentees	ratio
5	Education of people	Education level	Categorical



So, if we take these cars owned by 10 friends, the variable name could be model or brand. For example, the car could be a ford car or a Hyundai car or a Maruti as the case may be. So, that is the variable name which is the model or brand, and that data is categorical. So, you would say either this or that or the other. Now, if we look at income of 20 employees, the variable name that we can give could be salary or could be income and the data type is numerical.

Now we can understand why it becomes a numerical data, because we already have learnt that numerical data we can add and subtract we can also multiply and divide. So, if the data is such that we add and subtract which means the difference is interpretable or we can interpret the difference. Then it becomes interval type data. While even a multiplication can be explained then it becomes a ratio level datum.

Now, when we compare income or salaries, it is also possible to say that salary of person x is more than a certain quantity compared to salary of person y. And it is also fair to say that person x gets 20 percent more than the salary of person y, or gets one and a half times the salary of the other person and so on. And therefore, income of 20 employees comes under the numerical type of data.

Size of clothes as an example, you go to a garment shop you could find at least 4 different sizes which are small medium large and extra-large. So, the variable name can be cloth size, and it becomes ordinal if the sense we can rank them. And we can

generally conclude that extra-large is greater than large, while large size is bigger than medium size, and medium size is bigger than small size. It is very difficult to say that if we look at those measurements, and then compare and then say that medium is bigger than small by a certain quantity. To do that we need more data we need more measurements.

Just seeing the classification, a small, medium, large and extra-large we can categorize them as ordinal data. And conclude that small size is smaller than medium size, which in turn is smaller than large which in turn is smaller than extra-large. But the extent to which one is smaller or larger is not explained and therefore, it becomes ordinal data number of students absent for a class could be absentees and quickly falls into ratio data.

Because if 10 peoples were absent yesterday and 5 people were absent today, it is not only possible to say that today's absentees was 5 less than yesterdays, it is also possible to say that yesterday twice the number of people were absent. Therefore, both addition subtraction as well as multiplication division is possible and therefore, we can call this as ratio type data.

Look at education of people. There are distinct education levels that were given. So, the variable name could be education level, and it would come under categorical people, and the example given in the previous slide could be studied up to high school, did graduation did post-graduation and pursued a PhD we realized that a particular person could fall into any one of these categories. So, it comes as a categorical variable and within that it could be a nominal variable.

(Refer Slide Time: 07:53)

### True or false

1. Pin codes are examples of numerical data
2. Cases represent columns in a data table
3. Frequency of time series is the time spacing between data
4. Likert scale represents numerical data
5. Aggregation of data adds more cases



Now, let us look at some more examples to understand. Pin codes are examples of numerical data is it true or false, the answer is false and pin codes are example of categorical data. Even though pin codes are numbers one might immediately think that it would represent numerical data, actually do not does not represent numerical data, because we can neither add or subtract nor can we multiply divide and make meaningful conclusions out of that. And therefore, pin codes are examples of categorical data.

Second one would be cases represents column in a data table. So, if we go back and quickly see what we learned, the variables are the columns in the data table and cases are observations or the rows in the data table. And therefore, cases represent columns in a data table is false. Frequency of time series is the time spacing between the data. So, to answer this question we also need to understand what is time series data.

So, time series data is essentially data measured across time. For example, if we are looking at let us say an MBA class, and then we could go back and say in the year 2018 we have 70 students in the class. The year 2017 we had 65 students in the class. Year 2016 we could have 73 in the class and so on. So, we measure something over a period of time.

Example number of students in a class. Example sale in 12 months of the year. Example stock prices in the last so many weeks. Example the price of petrol or fuel in 30 days of a month and so on. So, one can give several examples for time series data. We will also see

some situations in this course where we look at time series. And with this information let us come back to the question frequency of time series is the time spacing between the data.

So, time spacing between the data is the frequency in a time series. Likert scale represents numerical data. So, Likert scale is a scale where we say whether or we like something it moves from a very strong like to a dislike. And Likert scale does not represent numerical data, Likert scale represents ordinal data and therefore, categorical data. It is kind of ranks it at the same time it is very difficult to say that if I accept it or I strongly agree versus agree.

In this scale we we say that we start with strongly agree to agree, and then it goes to strongly disagree, and a person takes one of them given a situation. So, while we can say that strongly agree is a more stronger agreement than taking agree. It is difficult to say how strong or measure the difference between the two things. And therefore, it does not represent numerical data, it represents categorical data.

Aggregation of data adds more cases; aggregation of data actually reduces the number of cases. Because as aggregation means addition, and as we add we only reduce the number of cases or observations, and therefore, it is necessary to understand that aggregation does not add more cases it reduces the case. So, if we really want to present data in a more precise or a shorter form then we resort to aggregating the data.

So, these examples kind of made us understand given different situations; whether the data falls under categorical or numerical and within that sub categories such as ordinal interval and so on.

(Refer Slide Time: 12:12)

## Cross sectional or time series?

1. Company has data on number of employees who are in PF scheme and the amount in PF
2. 1000 people are asked if India would win the cricket world cup
3. The number of people who shopped for more than Rs 5000 on five days of the week
4. 100 customers of a hotel give feedback. 60 ticked excellent, 30 ticked average while 10 said poor
5. Number of sedans and small cars parked in front of a supermarket on 7 days of a week



We look at another aspect of this. And we want to check whether given situations the data is a cross sectional data or is it a time series data. We already saw what time series is, time series is basically data measured across different points in time and category sectional data essentially means looking at the data at a certain instance in time. So, that is the difference between cross sectional and time series data.

We will now look at these 5 examples to understand whether they are cross sectional or time series. So, first situation would be a company has data on the number of employees who are in a PF scheme, and the amount that they have in their provident fund. Now this is cross sectional data, because this data is taken at a certain point, and it is not taken at different points for comparison.

So, the first example is an example of what is called a cross sectional data. Situation to about thousand people were asked if India could win the cricket world cup. Again this is an example of cross sectional data, because at a certain point in time we ask a certain number of people whether something would happen or not happen; Situation number 3, number of people who shopped for more than 5,000 on 5 days of a week. So, this is an example of time series data, because the data is measured according to a certain frequency, which is a day and on 5 consecutive days or 5 days of the week we measure the number of people who shopped for more than 5,000.

Situation 4, 100 customers have given feedback, 60 have said excellent, 30 have said average while 10 have said poor. Again example of cross sectional data, because the statement does not explicitly say that the feedback was collected over different points in time at regular frequencies and so on. So, we could take this as cross sectional data. Number of cars big cars small cars park in front of a supermarket on 7 days of a week.

So, once again it is similar to what we saw in item 3, where this data is collected at different points in time and therefore, it comes under time series data. Times it is necessary for us to understand this classification. Because certain analysis specific to time series we would be studying later in statistics, maybe not in this course and therefore, we introduce this idea that once we look at data we also need to understand whether it is cross sectional, which means it is data that is taken at a certain point in time, or it is time series where it is data that has been collected over a period of time.

Now, we move to some more aspects of data, and we know try to describe categorical data. So, in the earlier in the last lecture we introduced the term called categorical data, and then we classified them further into nominal and ordinal. Now, we try to describe and see how we present categorical data to the user.

(Refer Slide Time: 15:45)

Number of votes polled when asked "Who will score most runs?"

(Imaginary data)

	Votes polled	Fraction	Percentage
Player 1	45276	0.097732	9.77
Player 2	39825	0.085966	8.6
Player 3	32419	0.069979	7
Player 4	29666	0.064037	6.4
Player 5	48977	0.105721	10.57
Player 6	41678	0.089966	9
Player 7	26423	0.057036	5.7
Player 8	30912	0.066726	6.67
Player 9	19627	0.042367	4.24
Player 10	27555	0.05948	5.95
Player 11	28432	0.061373	6.14
Player 12	17666	0.038134	3.81
Player 13	15487	0.03343	3.34
Player 14	22723	0.04905	4.91
Player 15	14900	0.032163	3.22
Player 16	21700	0.046841	4.68

Total = 463266



Frequency table – represents the distribution of a categorical variable as a table

Can become hard to compare as the table gets large

So, I have just given a an example from let us say from cricket, and we have picked up some numbers the data is an imaginary data, it does not represent the live data, and let us



assume that this question was asked in a in a cricket website as to who would score more runs in let us say a popular 20-20 tournament.

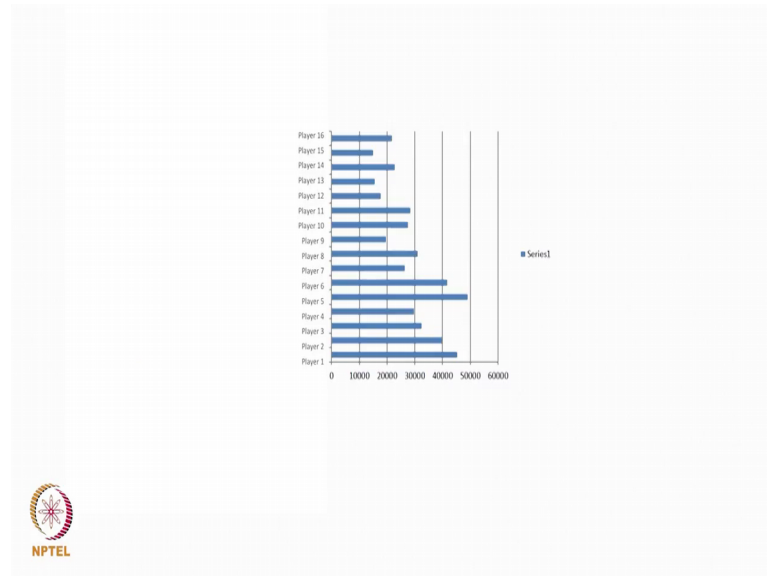
And the users could pick up a certain number, and let us assume that these are the the names of these players who were actually voted by people. And let us assume these are the number of votes that were polled by each of these players. For example, player number one let us say polls 45,276 votes or 45276 people believe that player number one listed here would score the maximum runs.

So, this is a data table where we have 4 columns, and the first column would be the name of the player, the second column is the number of votes polled, the third column would simply be the fraction of the total votes polled and the 4th column is the fraction represented as a percentage. So, because this represented as a percentage, the percentages let us say would add 200 and the ratios would add to one.

So now this is a data table or a frequency table which represents the distribution of categorical variable as a table. And the categorical variable is a number of votes polled. Now this is one way of presenting categorical data, saying these are the cases and this is the data. And the advantage of this frequency table is that we are able to present all the data that we wish to present.

But the disadvantage is this table can become large as the number of cases become large. For example, we already have names here and it would though we present all the information that we wish to present, one gets a feeling that if this runs to a second page or if there are more cases and observations, it becomes difficult to handle this kind of a data.

(Refer Slide Time: 18:24)



So, one way is to look at a table to present it, while the other is to look at pictures to present this type of data. So, this is a picture that presents the same data in a pictorial form. And this picture now shows the names which are here, and it also shows a bar representing the number of words that this person has polled. You can see now the person who polled, the maximum is close to about 50 thousand, polls or votes and here is somebody who has about 15,000.

Now, this is called a bar chart so, a bar chart is a very convenient way of presenting a categorical variable. There are 2 types of bar charts, and this bar chart is called a horizontal bar chart, and the other one which we will see later is called a vertical bar chart. Now in this these bars represent the number that we wish to present and this number is the number corresponding to the categorical variable.

If we take this particular player, then this bar represents the number of votes that this person has got. Now one can get a feeling that this bar chart presents the the data in in perhaps a slightly nice of form where we are able to have these bars representing what we actually want to represent. Perhaps a a slight disadvantage of this representation is that by looking at this bar it is it is slightly difficult to say what is the exact number of votes or polls this person has got.

One can only say there it is between 40 to 50000 and much closer to 50000, one might get a feeling that this is anything; between 48 to 49000. So, in spite of this the bar chart

is accepted as a very as a very convenient and nice way of presenting a categorical variable. Now in the next lecture, we would continue the discussion on presenting this categorical data, and we will see further examples from bar charts and pie charts to present categorical data.