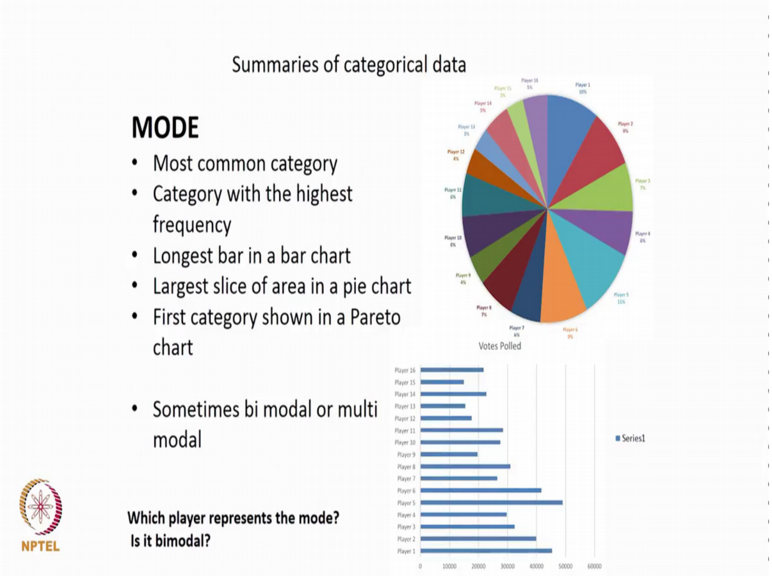
Introduction to Probability and Statistics

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

Lecture - 05
Describing Categorical Data (continued)

In this lecture we look at measures of summarizing categorical data. In the previous lecture we looked at presenting categorical data in the form of bar charts and pie charts towards the end of that lecture, we introduced the mode and we will now go in detail understanding the mode of the categorical variable.

(Refer Slide Time: 00:42)



Summaries of categorical data

MODE

- Most common category
- Category with the highest frequency
- Longest bar in a bar chart
- Largest slice of area in a pie chart
- First category shown in a Pareto chart

- Sometimes bi modal or multi modal

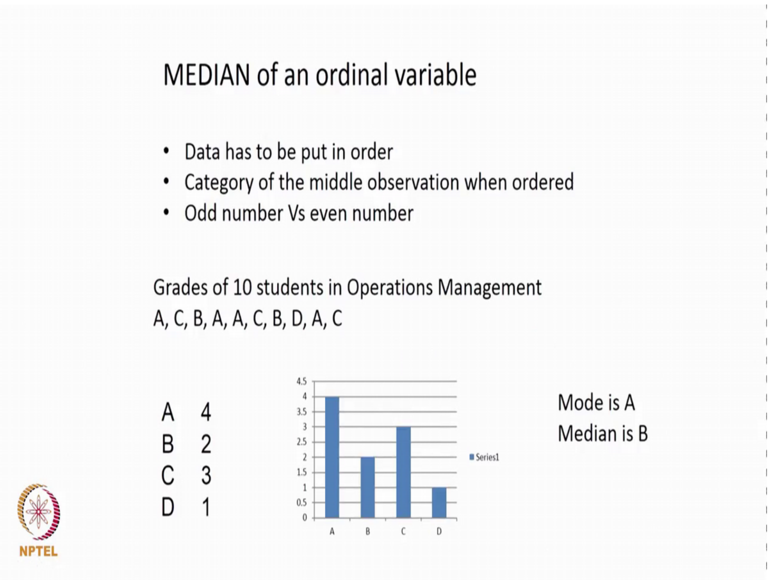
Which player represents the mode?
Is it bimodal?

So, we want to summarize categorical data the most commonly used measure is called the mode. So, mode represents the most common category or the category with the highest frequency.

So, if we look at the table we look at the frequency corresponding to all these cases and the 1 that has the highest frequency is the more, if we look at a horizontal bar chart or a vertical bar chart the mode is the longest bar in the bar chart. If we look at the same data represented as a pie chart the mode is the largest size of slice of area in the pie chart, the 1 that occupies the largest area in the pie chart. And if we represent the data in the form of a Pareto chart as shown here, then it is the first category that is shown in the pareto chart.

So, mode is the most common way of representing categorical data and it is the category with the highest frequency, it is very easy to identify the mode in the categorical data. So, in this case which player represents the mode the player corresponding to this bar represents the mode. Sometimes we could have bi modal which means you could have more than one having the same mode. So, is this bi modal from this picture it is not bi modal because, we do not find 2 bars of equal length therefore it is not bi modal.

(Refer Slide Time: 02:21)



MEDIAN of an ordinal variable

- Data has to be put in order
- Category of the middle observation when ordered
- Odd number Vs even number

Grades of 10 students in Operations Management
A, C, B, A, A, C, B, D, A, C

A	4
B	2
C	3
D	1

Mode is A
Median is B

We also have another measure which is called the median, but median is used for a ordinal variable and we remember that he categorized the qualitative variables as categorical and ordinal and the numerical variables were interval and ratio. So, for all categorical variables we had the mode while for ordinal variable we have the median.

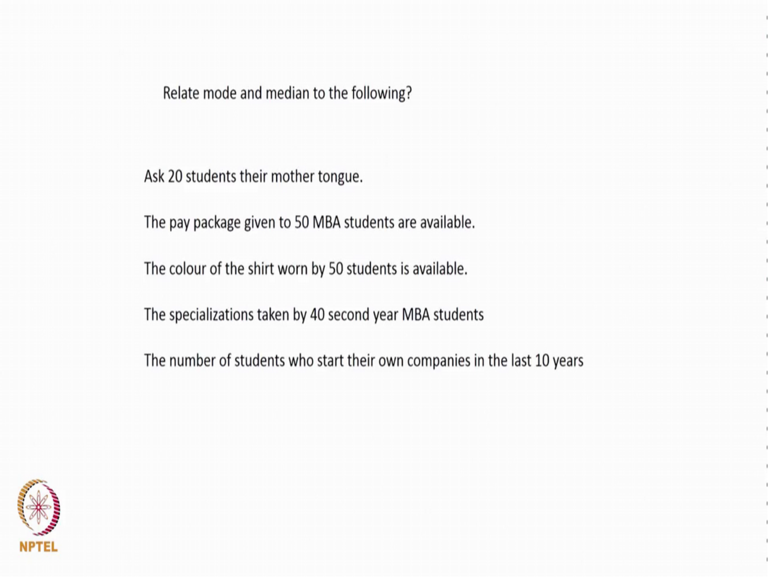
So, the very word ordinal would tell us that there is an order except that in the order we cannot say or we cannot compute the difference between the terms in the order therefore it becomes ordinal and if you are able to understand the difference then it becomes interval. Now, in an ordinal variable now data has to be put in the order and the category of the middle observation when ordered is called the median and we also know that the median computation is slightly different for odd number and for even number. So, just to give an example the grades of 10 students in an operations management course is given, so A, C, B, A, A, C, B, D, A, C so A is considered as the highest grade B is the next highest C D and so on.

So, first thing if we draw a bar chart the bar chart will look like this, so the bar chart has 4 people corresponding to A grade 2 people corresponding to B grade 3 people corresponding to C grade and 1 person corresponding to D grade, first we look at the bar chart and the bar chart is ordered and then drawn in this case; therefore, there are 10 observations or 10 cases.

Now, because it is an even number we look at case number 5 and 6 and then look at what it has. Now, in this case we realize that both 5 and 6 have be great the first 4 have A, once we have sorted them in the order the first 4 have A grade the fifth and the sixth person has B grade and the other 4 3 have C grade and 1 has a degrade. And therefore, the median is the average of the fifth and sixth person in this case they are equal and therefore, the median is a B grade.

For the same data the mode is the longest bar and therefore, A grade is the mode. So, if the data is ordinal we can calculate mode as well as median, while the data as categorical we only calculate the mode which is the same thing is explained here with A being the mode and B being the median.

(Refer Slide Time: 05:09)



Relate mode and median to the following?

Ask 20 students their mother tongue.

The pay package given to 50 MBA students are available.

The colour of the shirt worn by 50 students is available.

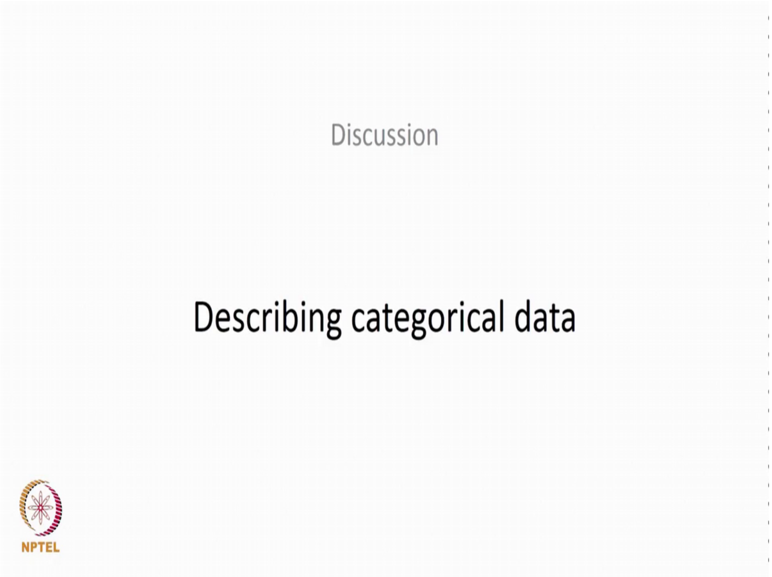
The specializations taken by 40 second year MBA students

The number of students who start their own companies in the last 10 years

Now, relate the mode and median to the following we will do small we will do a small exercise here you just to understand sometimes particularly in cases where both mode and median are possible how they are related. So, first example we asked 20 students what is their mother tongue, so mother tongue of twenty students is a categorical variable. So, we will find only the mode and whichever language has the highest frequency would become the mother tongue, pay package given to 50 students are available. So, in this case we can actually find out is there a particular salary where most number of people have got, then it is a mode times we may even fit these in a range and then look at this and then we can sort these 50 salaries in a descending order and try to find out what is the median.

Color of the shirt owned by 50 students it is clearly a categorical variable and therefore, only the mode can be found out. So, whichever color has more students we do this specializations taken by 40 second year MBA students, now these specializations is categorical. So, it could be finance it could be marketing it could be operations and so on and therefore it is a categorical variable only the mode can be found out. The number of students who start their own companies in the last 10 years is also a categorical variable and over the years it is a time series. So, whichever year has the maximum number and that can represent the mode of that value.

(Refer Slide Time: 06:58)

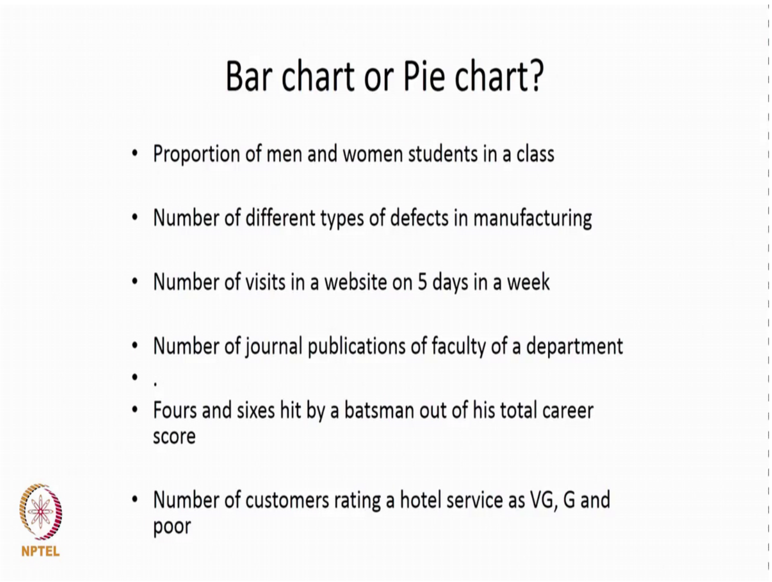


Discussion

Describing categorical data

So, we will continue this discussion on describing categorical data with some more examples.

(Refer Slide Time: 07:02)



Bar chart or Pie chart?

- Proportion of men and women students in a class
- Number of different types of defects in manufacturing
- Number of visits in a website on 5 days in a week
- Number of journal publications of faculty of a department
- .
- Fours and sixes hit by a batsman out of his total career score
- Number of customers rating a hotel service as VG, G and poor

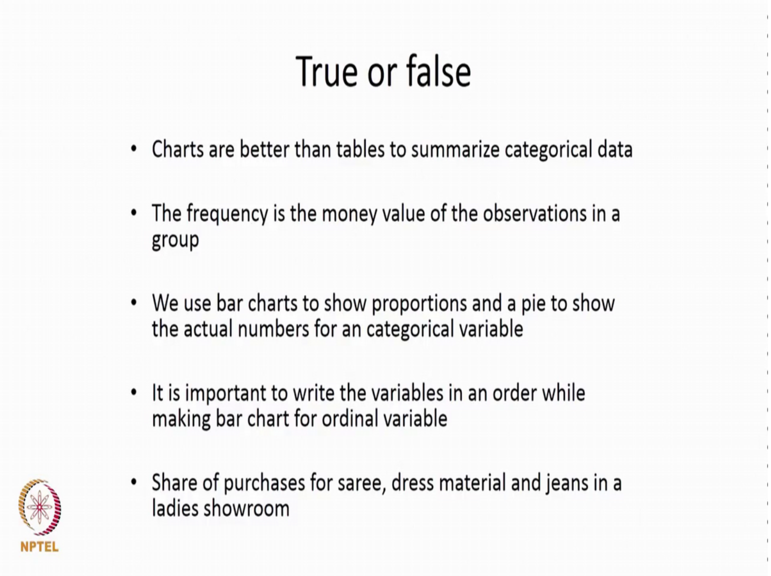
So, we will now get back to the pie chart and bar chart examples and look at situations and try to answer questions whether something can be told as a bar chart or told as a pie chart proportion of men and women students in class, number of different types of defects in manufacturing number of visits in a website 5 days in a week, number of publications of faculty 4 and 6 hit by a batsman out of this total career score, number of customers rating a hotel service proportion of men and women the answer is obvious we talked about proportions fractions and therefore pie chart.

The number of different types of defects in manufacturing, so first we have to find out what are the different types of defects and within each defect type we can find out a certain number, we could have a bar or a pie depending on how we generalize it. So, at the end if we want to generalize saying that 10 percent of the defects are of this type and so on then we could use a pie in this case. Number of visits in a website on 5 days in a week could be clearly a bar chart and we unless we want to say that 20 percent of the people visited on day 1, 40 percent visited on day 5 and so on.

Number of general publications of the faculty if represented as a number it is a bar chart and represented as a percentage, particularly when we want to compare the time series data saying that out of the last publications in the last 5 years 40 percent happened in the fourth year and so on, then it is a pie otherwise it is a bar chart 4 and 6 hit by a batsman out of his total career once again starting with the bar chart.

But if you want to generalize it as a percentage then it becomes a pie, number of customers rating a Hoteliers very good, good and poor would generally become a pie; considering that we want to generalize it rather than saying 25 people said excellent or good we want to know what percentage of people said very good, so that we could generalize it. So it becomes a pie chart.

(Refer Slide Time: 09:21)



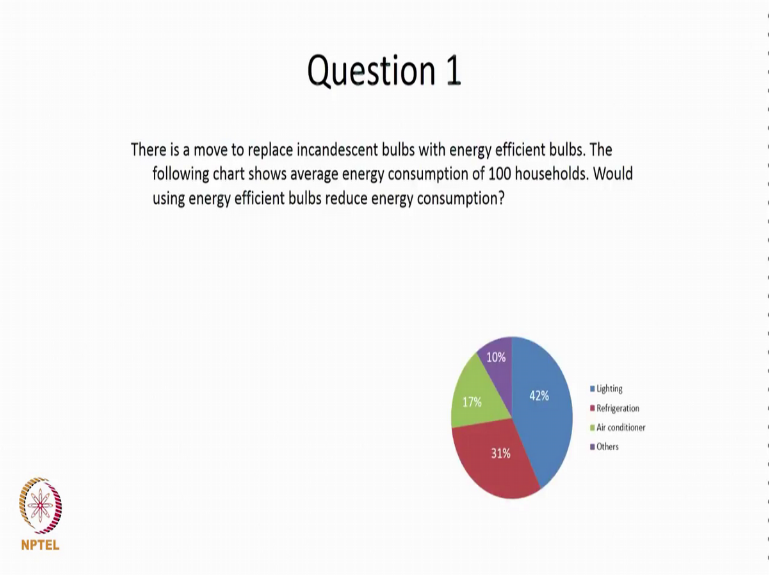
True or false

- Charts are better than tables to summarize categorical data
- The frequency is the money value of the observations in a group
- We use bar charts to show proportions and a pie to show the actual numbers for an categorical variable
- It is important to write the variables in an order while making bar chart for ordinal variable
- Share of purchases for saree, dress material and jeans in a ladies showroom

Some true or false questions charts are better than tables to summarize categorical data true because, charts are usually used to represent and summarize the data. Frequency is the money value of observation in a group not necessarily frequency is usually a number which represents the variable that is used. We use bar charts to show proportions and pie charts to show numbers for a categorical variable.

The answer is not true it is the other way bar chart shows numbers pie chart shows proportions. It is important to write the variables in an order while making a bar chart for an ordinal variable, not necessarily but if you want to make a pareto chart then we need to put them in order otherwise we could have bar charts; but it is normal practice to put them in order and do it for ordinal variables share of purchases for Saree, dress material etc in a showroom becomes a pie chart because it represents a proportion.

(Refer Slide Time: 10:24)



Question 1

There is a move to replace incandescent bulbs with energy efficient bulbs. The following chart shows average energy consumption of 100 households. Would using energy efficient bulbs reduce energy consumption?

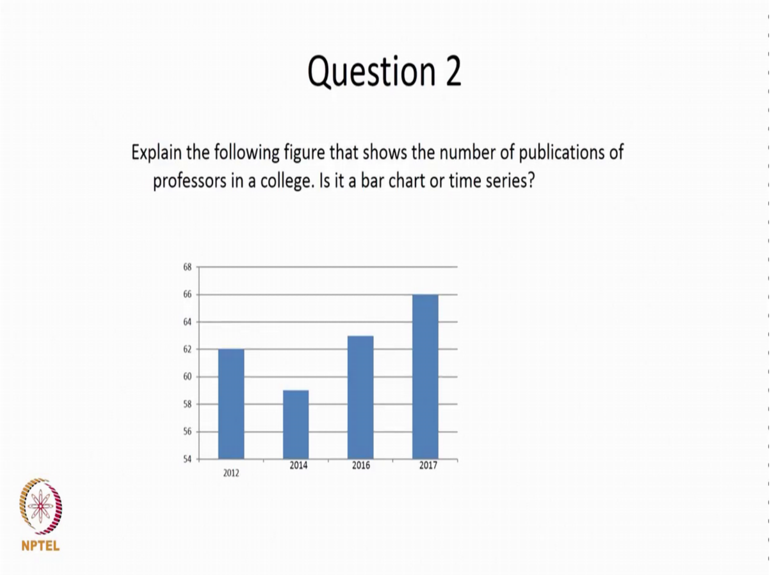
So, we look at a few questions, now so the first question would be there is a move to replace incandescent bulbs with energy efficient bulbs. The chart shows the average energy consumption of 100 households would using energy efficient bulbs reduce the energy consumption, the obvious answer to the question is yes using energy efficient bulbs would reduce energy consumption.

But then we use the data that is provided and then we try to make a decision based on the data that we have, now the pie chart shown in this slide tells us the average energy consumption from 100 households. So, from this we observe that 42 percent of the energy consumed by these households goes in lighting, therefore the decision to replace incandescent bulbs with energy efficient bulbs is going to affect 42 percent of the consumption which by itself is a large percentage and therefore, for this question we say that since lighting occupies a significantly large percentage of energy consumption, replacing incandescent bulbs with energy efficient bulbs would have a good effect on the reduction in energy consumption.

We also note that this is a pie chart that is given here which means we are looking at proportions, proportions taken by these 4 things which are lighting refrigerators air conditioners and others and this is data that has been collected by surveying 100 households. Therefore, the average values that we get based on the survey of hundred households is now converted to proportions and used in this pie chart.

So, this is an example of using a pie chart and is also an example where data collected from a sample is now generalized to represent the broad ways by which energy is consumed in households and so on. So, we moved to another question.

(Refer Slide Time: 12:57)

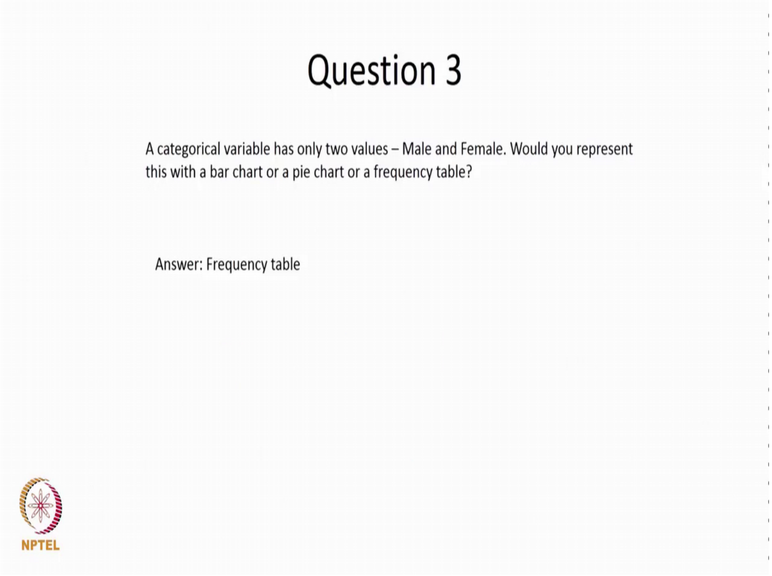


Question 2

Explain the following figure that shows the number of publications of professors in a college. Is it a bar chart or time series?

Explain the following figure that shows the publications of professors in a college, is it a bar chart or it is a time series. Now, here we show data for 4 years 2012, 14, 16 and 17 and the y axis shows the number of publications. Now, this is time series data because the x axis represents time and data is collected over 4 distinct time periods, it is more a time series chart than a bar chart though these bars represent the data over a period of time.

(Refer Slide Time: 13:36)



Question 3

A categorical variable has only two values – Male and Female. Would you represent this with a bar chart or a pie chart or a frequency table?

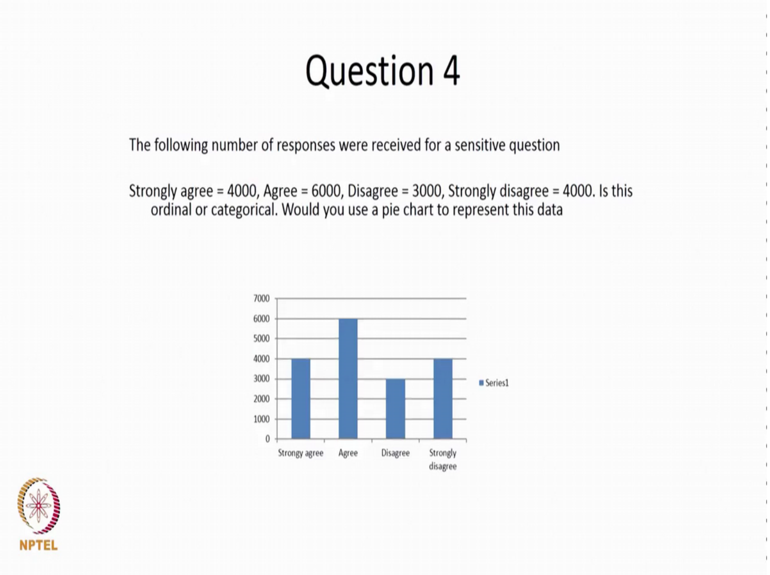
Answer: Frequency table

We look at another question a categorical variable has only 2 values which is male and female, would you represent this as a bar chart or a pie chart or a frequency table. The answer is also given that the answer is frequency table, the reason why the answer is frequency table is something that we will discuss. Now, we can use a bar chart we can also use a pie chart if the data represents proportion of people who are male and proportion who are female.

When we started our discussion in trying to understand pictorial representation of data we started with a frequency table and then we said that as the number of cases or observations increases, it becomes difficult to interpret or understand from the frequency table and that led us to charts or pictorial representation of data. And then, we learned the bar chart and the pie chart and then we also said that if we are looking at proportions or fractions or percentages we look at pie chart, otherwise we represented with a bar chart.

Now in this case there are only 2 cases which means male and female are the only 2 types of observations or cases; therefore, the frequency table would be something like the number of male and the number of female. Since the number of distinct cases is small a frequency table is an adequate representation in this case, though bar charts are not entirely wrong. From the frequency table itself we will be able to get the exact numbers of male and female, we also understood through other examples that at times by looking at the bar chart it might be difficult to read the exact number that is implied in the bar chart, pie chart also shows only percentages and therefore in this case frequency table is a good way to represent this type of data.

(Refer Slide Time: 15:46)



Question 4

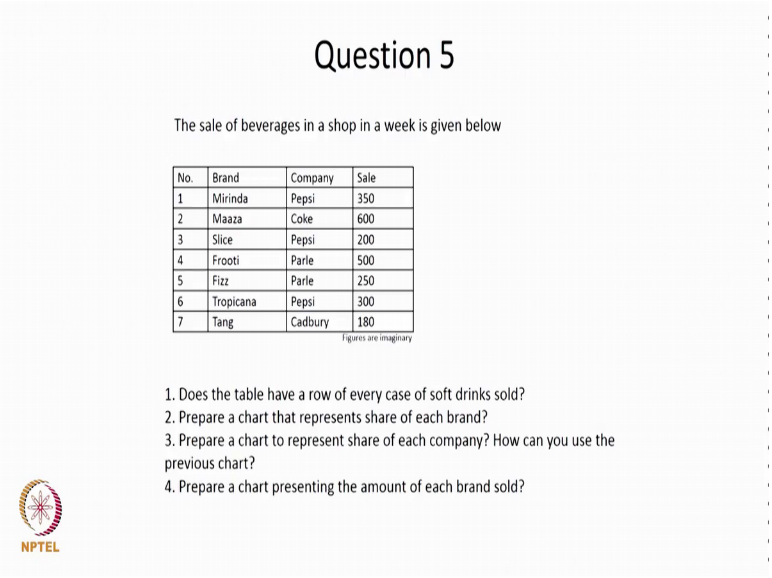
The following number of responses were received for a sensitive question

Strongly agree = 4000, Agree = 6000, Disagree = 3000, Strongly disagree = 4000. Is this ordinal or categorical. Would you use a pie chart to represent this data

We look at another question the following number of responses were received for a sensitive question strongly agree 4000 agree 6000 disagree 3000 and strongly disagree 4000; is it ordinal or categorical would you use a pie chart to represent this data is the question the bar chart is given. Now because, we have these 4 ways of ticking or expressing the opinion which starts from strongly agree to strongly disagree this data is ordinal data, where strongly agree is seen at a level higher or more than agree which implies that a little more than disagree and then comes strongly disagree.

But then as we said this would not be interval data because, the difference between strongly agree and strongly disagree versus difference between agree and disagree are not comparable and measurable; therefore, this represents ordinal data and therefore, it is not advisable to use a pie chart, pie charts are used to represent categorical data. So, bar chart in this case would be a more appropriate representation of the data that we wish to represent.

(Refer Slide Time: 17:08)



Question 5

The sale of beverages in a shop in a week is given below

No.	Brand	Company	Sale
1	Mirinda	Pepsi	350
2	Maaza	Coke	600
3	Slice	Pepsi	200
4	Frooti	Parle	500
5	Fizz	Parle	250
6	Tropicana	Pepsi	300
7	Tang	Cadbury	180

Figures are imaginary

1. Does the table have a row of every case of soft drinks sold?
2. Prepare a chart that represents share of each brand?
3. Prepare a chart to represent share of each company? How can you use the previous chart?
4. Prepare a chart presenting the amount of each brand sold?

NPTEL

Now, we look at another question which talks about sale of beverages in a shop in a week is given, so 7 names or brands of beverages are given and some companies from which these brands are produced are also given and sales figures are given and these sales figures can be assumed to be imaginary figures. There are some questions does the table have a row of every case of soft drinks sold, the answer is maybe not the shop could sell other brands and those have not been represented here.

If we have to prepare a chart that represents the share of each brand which is given here, then it will be good to do a pie chart first we find out the total sale. And then, we find out the fraction or percentage of each sale for each of these brands and then we could draw a pie chart which would represent the word share in the question is indicator of proportions and therefore a pie chart is to be used.

Prepare a chart to represent the share of each company, how can you use the previous chart. Once again you find the word share in the question and therefore we could look at a pie chart to represent even though there are 7 brands that are listed here. There are fewer than 7 companies therefore, we have to do the sale company wise and then find the proportions and then draw a pie chart for this.

How do we use the previous chart we can aggregate the values from the previous chart and then we can use that. Prepare a chart representing the amount of each brand sold, so again there are 7 brands and then the in this case the chart would be a bar chart, where we actually represent the sale figures there are 7 figures that are given in this table. So, this way we could go on and answer questions and we have in this lecture and part of the previous lecture.

See in some questions and some instances where the concepts that we learned in the previous lectures, were used to solve different types of problems. Now, in this lecture and in the earlier lecture we looked at representing categorical data as well as measuring them and representing them through the mode as well as the median, in the next lecture we would look at ways of capturing numerical date.