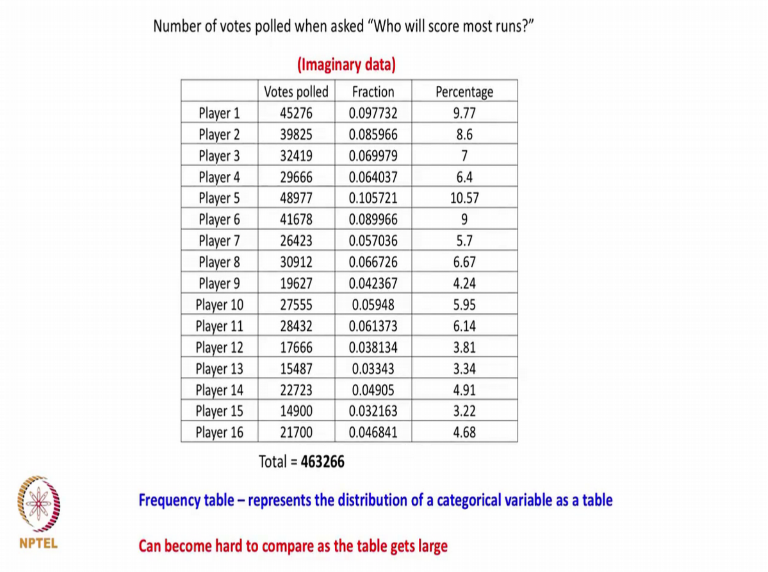
Introduction to Probability and Statistics
Prof. G. Srinivasan
Department of Management Studies
Indian Institute of Technology, Madras

Lecture - 04
Describing Categorical Data

In this lecture, we continue the discussion on Categorical Data. Towards the end of the earlier lecture, we started introducing the bar chart. So, we continue with the same example and which talks about the number of votes polled.

(Refer Slide Time: 00:32)



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**

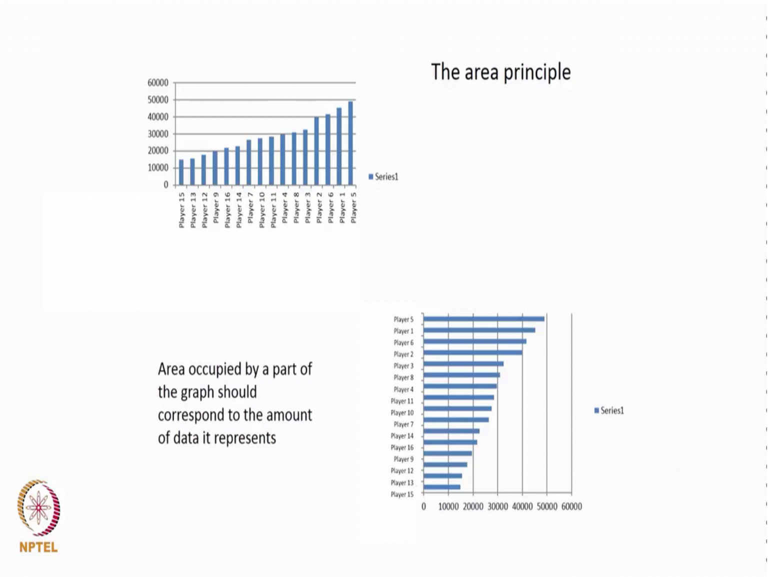
Frequency table – represents the distribution of a categorical variable as a table

Can become hard to compare as the table gets large

Let us say a in a cricket website, when the question who would score most runs was asked, let me say that, this would be imaginary data; then the list of players given. And then, this table, which is the frequency table, now has 4 columns. The first column is the name of the player, the second is the number of votes polled, the third would be the fraction of votes polled and the fourth is the fraction represented as a percentage. And let us assume that, these are the only players who are considered. Therefore, the percentages add up to 100 and the fraction adds up to 1.

So, as I mentioned in the earlier lecture, this table adequately summarizes what we want to see. But, as the number of cases and observations increases, one would get a feeling that this table, it would look a little cluttered and would perhaps look a little difficult to understand the data. So, the next question that we look at is, can we represent this in the form of a picture.

(Refer Slide Time: 01:49)



The area principle

Area occupied by a part of the graph should correspond to the amount of data it represents

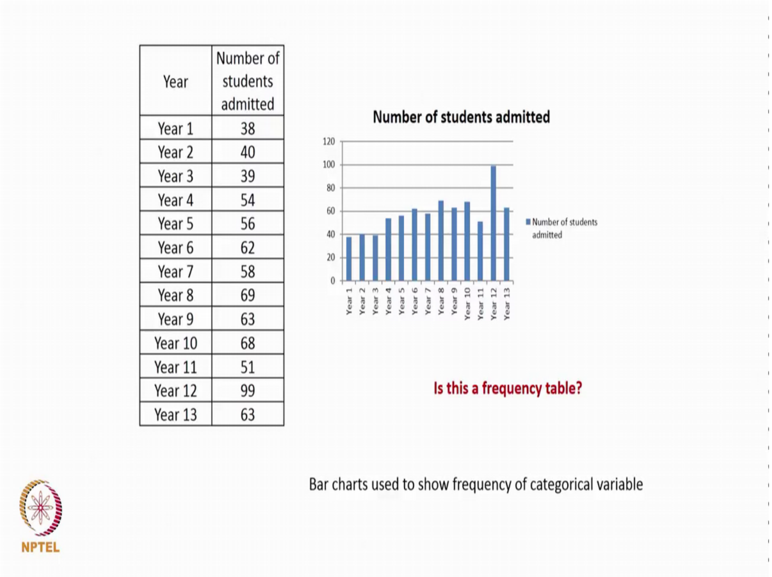
So, this picture was what we saw in the last lecture and this picture is called a bar chart which is a horizontal bar chart, which now shows the number of votes polled by different players. Now, we these bars actually tell us a few things. Now, we looking at the longest bar, one can conclude that this player has scored the maximum number and something between 40000 and 50000 and perhaps closer to 48 or 49000. Though it is very difficult at this point, by merely looking at the bar to find out the actual number which we could get from this table which spoke about as 48977.

In spite of that, the bar chart is a very convenient way to represent data and is widely used as a way to represent categorical data now, the same information shown in two different forms. Now, what we do is, in both these, now this is the horizontal bar chart, the one that is shown here is a vertical bar chart and what we have tried to do here is that, if you observe carefully, in this vertical bar chart the players are already sorted from the smallest number of votes polled to the largest number of votes polled.

And similarly, the there is also sorting from the largest number of votes polled to the smallest number of votes polled in this. Now, such a chart is called Pareto chart. In a Pareto chart, we the bars are arranged in a manner that the one with the largest frequency comes first and progressively, it reduces to the smallest frequency.

So, both these are bar charts. At the same time, both these are also Pareto charts. Now, let us look at another set of data to understand the bar charts.

(Refer Slide Time: 03:53)



Year	Number of students admitted
Year 1	38
Year 2	40
Year 3	39
Year 4	54
Year 5	56
Year 6	62
Year 7	58
Year 8	69
Year 9	63
Year 10	68
Year 11	51
Year 12	99
Year 13	63

Is this a frequency table?

Bar charts used to show frequency of categorical variable

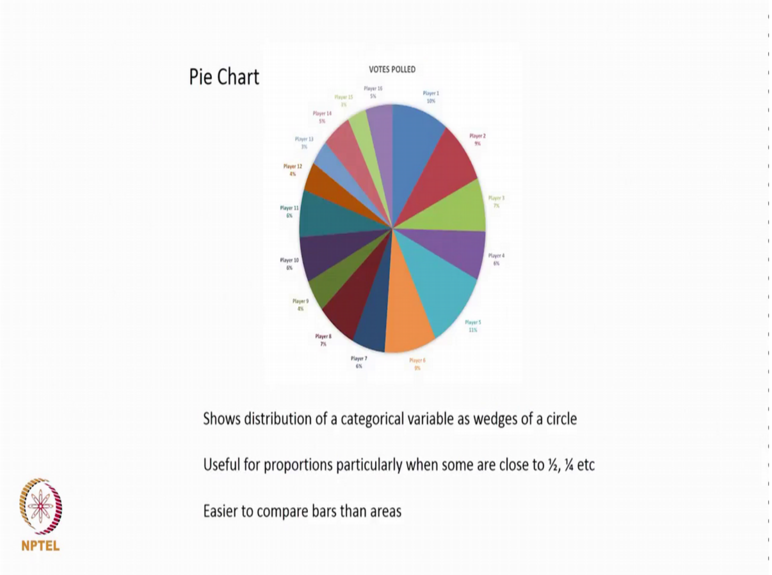
Now, let us say that, we have collected data for about 13 years on the number of students admitted to an MBA class. And the data shows that, in year 1, 38 students were admitted; while in year 13, 63 students were admitted.

Now, this table, here the variables there are the variable that we want to look at is a number of students. It is also a time series data because, we show this from years 1 to 13 and then this when represented as bar chart, would look like this; with the 13 years representing the bars and the number of students shown here. This is an example of a vertical bar chart and one could see the one can also observe from this bar chart that, year 12 had maximum number of students and quite close to 100. If not 100, very close to 100. One can see the same thing here because, year 12 has 99 students.

Now, one can particularly, if we can draw this bar chart and nowadays, we can draw this bar chart very comfortably using available software on the computer with nice excellent color codings and so on. So, the bar chart looks extremely pleasant to the eye and it also takes very little time to actually draw this chart on a computer and presented in any formal presentation.

So, many times, the bar charts replace the table and are convenient ways of presenting categorical data. Bar charts here are this is an example where bar charts are used to show the frequency of the categorical variable which is the number of students admit.

(Refer Slide Time: 05:48)



Pie Chart

Shows distribution of a categorical variable as wedges of a circle

Useful for proportions particularly when some are close to ½, ¼ etc

Easier to compare bars than areas

The next chart that we see is called a pie chart; very very popular and very commonly used chart which is called pie chart. And this is a pie chart corresponding to our imaginary data on the votes polled and so on. Now, we quickly understand from the color coding that different players, this player is perhaps here and so on. Pie chart shows the distribution of a categorical variable as wedges of a circle. The most important thing to understand in a pie chart is, pie chart is used for fractions or proportions. Pie charts are useful for proportions particular when these proportions are closer to half 1 by 4 and so on.

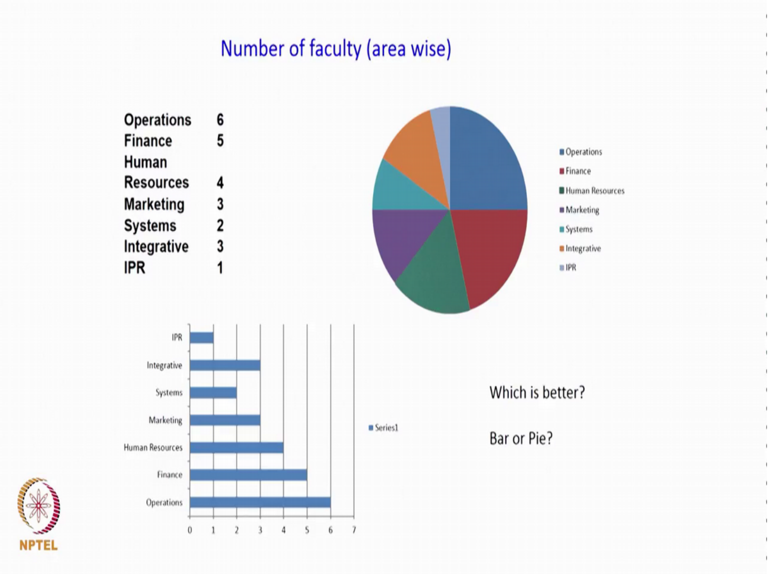
The simple reason pie charts are used for proportions bar charts are actually used for absolute values and number there is a general feeling there that, it is actually easier to compare bars than compare arrays. For example, if we look at this bar, it is quite easy and quick to say that this bar is much taller than the rest of the bars. And therefore, this is the one that has the highest frequency. Now, let us look at year 8 and year 10.

If you look at the actual data, year 8 has 69, year 10 is 68. And the bars, if you look at them carefully and we look at them very closely, it is possible to understand that even though both are equally tall, one can carefully look at this and understand that year 8 is slightly taller than year 10; whereas, if these were represented in the pie, it would be extremely difficult to tell which area is actually bigger than the other if you look at this pie.

For example, it is very difficult to say whether this is a larger area or whether this is the larger area. So, this is the general limitation of the pie chart compared to the bar chart, but have been said that pie charts are very good; particularly when you have smaller number of wedges which is at represented by this sentence. And the areas are closer to half 1 by 4 and there is a perceptible difference in the areas. By just looking at the pie, we should be able to say this is bigger than the other.

And if we have those kind of situations, pie charts are used and it is very very important to note that pie charts are used for proportions. And it is also important to note that, it is easier to compare bars than to compare areas. And therefore, wherever relevant bar charts have to be used ahead of pie charts, though at times pie charts are more attractive to the eye than bar charts are.

(Refer Slide Time: 08:54)



Number of faculty (area wise)

Operations	6
Finance	5
Human Resources	4
Marketing	3
Systems	2
Integrative	3
IPR	1

Which is better?

Bar or Pie?

So, let us continue this. Let us go back to this example of the number of faculty. Let us say in an MBA department. So, let us look at an MBA department of a college and then this MBA department has grouped the faculty that they have into groups. So, these groups are called the operations group, the finance group, human resources management group, the marketing group and so on.

So, now we have about 24 people 6 plus 5 11 15 18 20 20 4 people and this is the grouping. So, this is the frequency table which talks about number of people in each group. Now, if we draw bar chart, a horizontal bar chart, one would get a bar chart like this. Now, this bar chart would tell that there are 6 people in the operations group. Well, there is one person in the intellectual property rights group and so on.

And, let us try to represent the same thing in the form of a pie chart. Now, that is shown here. This pie chart is shown here and along with the color coding, one can quickly understand that 6 people given here in the operations group is the larger group and then you get the next group and so on. Now, this is a good representation in a pie chart, but with a small issue, in the sense that, from this one can now if you look at this pie chart carefully, we know see that there are 7 different colors and seven wedges and it friction some of them are closed to 1 by 4.

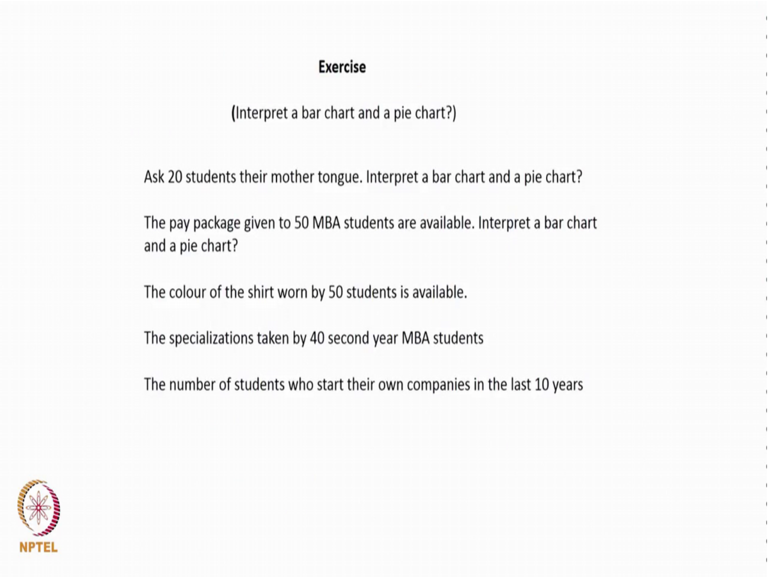
It is also possible quickly understand that this is larger than this which in turn is slightly larger than this and so on. But when we represent this in the form of a pie chart, we are representing proportions. So, we are trying to say that about 25 percent, there are 24 people. About 25 percent belong to this group. About 5 by 24, which is nearly about 20 percent belong to the next group and so on. The absolute numbers do not come in this pie chart. Whereas, the bar chart tells as everything the bar chart does not give us the percentage, but the bar chart gives the actual number of people who are in various groups.

So, we have seen 2 types of representation for this and then we ask this question which is better, a bar chart or a pie chart. And at the moment, the answer for this is, the bar chart is a better representation. If we want to represent that there are 6 people in this group out of 24 people, then the bar chart is the representation. If you want to say that 25 percent of the people, they belong to a particular group, then pie chart is the representation and between the 2, in this example bar is a better representation than the pie chart.

This is something which we need to learn and we when we start doing this kind of thing and start representing different types of data, then we will know exactly which is a better way to represented and we have to keep in mind a few things which we just now saw which is pie charts are used for fractions and proportions bar charts are used for the actual numbers.

It is always easier to understand from the bars to find out which one is bigger rather than from the areas and 3rd; particularly, when the number of wedges is small and these wedges are distinct and close to say 20 percent 25 percent, then pie chart becomes meaning full. So, we need to understand these 3 things and then try to work out several situations to finally conclude whether we use a bar chart or whether we use a pie chart. The most important thing is that, when we want to represent fractions or precautions, we use pie and when we want represent the numbers, we use bar.

(Refer Slide Time: 13:11)



Exercise

(Interpret a bar chart and a pie chart?)

Ask 20 students their mother tongue. Interpret a bar chart and a pie chart?

The pay package given to 50 MBA students are available. Interpret a bar chart and a pie chart?

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, let us try to do this. We ask 20 students there. Now, the exercise is, should I draw a bar chart or should I draw a pie chart. In this ask, 20 students, their mother tongue interpret a bar chart and a pie chart. So, we can do that. We could draw bar chart; we could draw pie chart. And in this case, a bar chart would be preferred to a pie chart and let us we want to generalize saying that, so many percentage of people belong to a certain mother tongue, but we restrict ourselves to the 20 students.

Then, the bar chart is a better representation than the pie chart pay packages given to 50 students are available. Interpret a bar chart and a pie chart. Once again, I would at this point, say that bar chart is a better example; however, if these 50 are going to represent a generic sample from a large population and so on.

And if we generally want to conclude from this saying that, 10 percent of them would be having a pay package of more than 20 lacks and so on, then a pie can be used. Otherwise, one would use a bar chart color of shirt worn by 50 students is available once again bar chart unless we want a generalized specializations taken by MBA students bar chart, but then we would look at a pie chart only if we want to generalize.

And finally, say that out of MBA students 20 percent take this specialization, 15 percent take another specialization and so on. The number of students who start their own companies in the last 10 years is clearly a bar chart it is a time series data. So, we use a bar chart in this. So, like this we should try a different examples and situation to actually understand whether we would be using a bar chart or we would be using a pie chart. This is very simple idea called the area principle which is important in a bar chart.

So, whenever we draw a bar chart, we need to observe two things. First and foremost,, if you look at this vertical bar chart and look at this horizontal bar chart, you will observe that all the bars of the same color. So, in this case, we have not shown the bar corresponding to different players using different colors and this horizontal bar also tells us that, while there are more bars than the number of players and some player names are not written here, but that is alright for the discussion.

We use the same color. We use the same color because; the variable that is represented is the same. So, unless we represent different variables in the same bar chart, we do not use different colors. Sometimes, when we think when we present the data using different colors, would actually give a more pleasing appearance to the eye. But, what is important to understand is that, as long as we are representing the same variable, it is important to use the same color. That is the first principle.

Then comes the area principle. The area principal talks about the width of the bar being the same in all the bars. Now, the area principle was all the more important; when we created these bars by hand using a pencil or using a ruler and so on. In which case when we did it by hand, we have to ensure that the thickness of the bars are actually the same while the length of the bars are different and correspond to the variable that is being measured; Today with increasing use of software and increasing use of computers to do this.

The area principle is actually understood the way it is done. But then, to the user, it is important to understand the area principle which talks about the width of the bars being the same. So, that the area generally represents what is the variable that is being measured in somewhere, the area here would be comparable to the area in the pie chart when we make then, we present the same data using a pie. It is also important to have this spacing between the bars same.

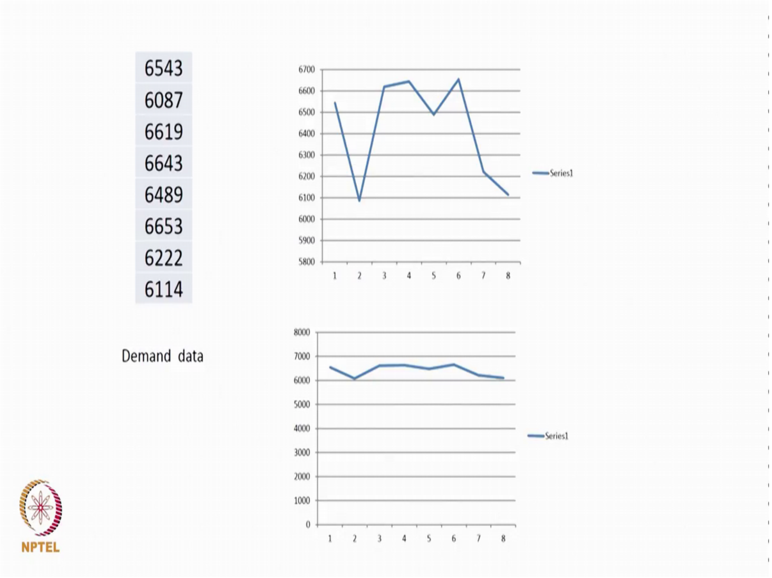
So, that it is pleasing to the eye, but what is important is area occupied by the part of the graph should correspond to the amount of data it represents, which means the width of the bars have to be the same. Now, this is a vertical bar chart, this is a horizontal bar chart. We also often have this question; should I represent something as a horizontal bar chart or should I represented as a vertical bar chart.

Now, there are only two issues and we explain that using the same bar. Now, this is a vertical bar chart and we when you represented as a vertical bar chart, it becomes extremely difficult to write the name in the horizontal manner which is comfortable to the eye and here we end up writing the name in the in the y direction or in the vertical direction and reading it becomes a little difficult. So, that is the first disadvantage when we do this. Otherwise, we need to put 1 2 3 4 and then give a legend here saying 1 represent something else.

That way, horizontal gives us a very comfortable way to represent it here. The more important thing is that; it is possible to quickly understand the difference in a horizontal bar chart than in a vertical bar chart. For example, if we see these 2 names, you can see the slight difference, but then one has to come closer to the bar to quickly come and see which one is taller; whereas, the same two things.

Let us say, are represented here as these 2 bars, then we realize quickly that this bar is long. So, wherever possible a horizontal bar is a more comfortable representation than a vertical bar for two reasons. One is a ability to write the name in a manner; that is easy to read, but the more important reason is, the ability of the eye to quickly understand which one is longer or bigger. When you we draw it horizontally rather than doing it vertically, the other thing that we need to understand is the y axis and the scaling particularly when we plot data. So, this example tells us.

(Refer Slide Time: 20:04)



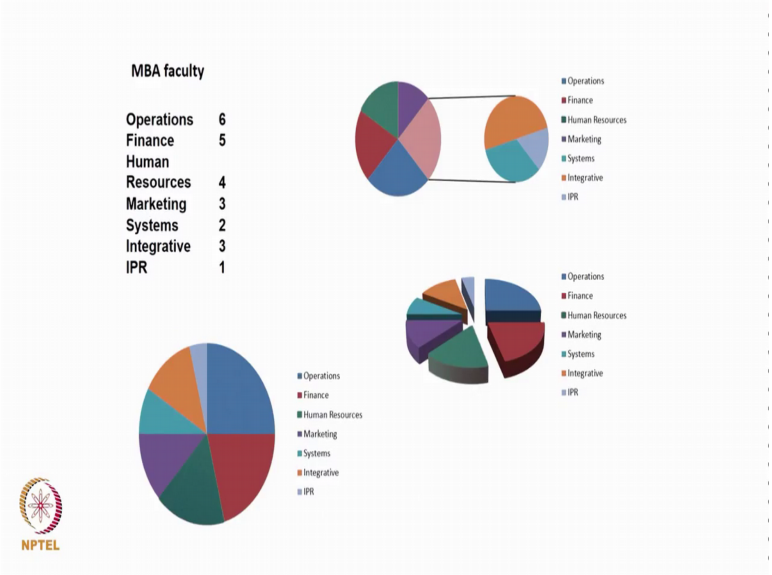
6543
6087
6619
6643
6489
6653
6222
6114

Demand data

Now, suppose we look at, let us say, a these numbers represent the demand of a particular item in about 8 months. So, we have shown actually 2 graphs where we have plotted this and both these are the same data. They are not different though the shapes of the curves look different. Both are actually the same data. The reason is, you have to look carefully at the y axis. In the first part, the y axis is between 5800 and 6700.

And now, we are able to see the spread or dispersion in the data comfortably and clearly here whereas, the same date as presented, but with a y axis between 0 and 8000. We now realize that, we are not able to understand that dispersion in this. So, one also not only before we form an impression from the picture, particularly a graph it is very very important to understand look at the y axis and then form an impression of the picture. Particularly, when we try to understand from only the graph and not by looking at the base data from which the graph was actually drawn. So, this is another aspect when it comes to presenting data.

(Refer Slide Time: 21:35)



MBA faculty

Operations	6
Finance	5
Human Resources	4
Marketing	3
Systems	2
Integrative	3
IPR	1

Now, let us look at the same example. There are multiple ways of presenting this. This particular pie chart we have already seen. So, let us assume, we are going to use a pie chart to represent this data. Though I had said the bar would be more meaningful than a pie, I am going to use this example to show different types of pie charts, two more examples of pie charts here. Now, first we will look at this pie chart.

This is also a very common way of representing with increasing use of software and computers, there are times miss be observe people present the same data this way like the pies being broken and pie is being part of disc and so on which also has a small 3 dimensional effect in this. And now, you will quickly realize that this is actually the same as this which as coming here.

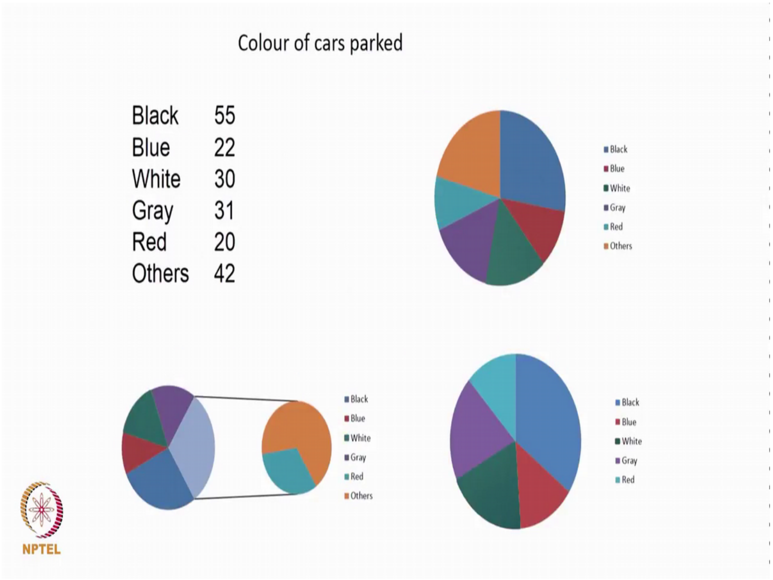
Now, this is in my impression, this is a little more difficult to comprehend compared to this. For the simple reason, that when we look at, I am very clear that one fourth of this circle is actually occupied by this whereas, when I represent the same data in this manner, I find it very difficult to quickly understand that this is actually one fourth of it.

Now, when I look at this and when I look at the other one, now in this circular pie chart, I know that the area occupied by this portion is actually less than the area occupied by this portion. But, when I start presenting the same data in this form, I am slightly confused here because, both of them look alike.

So, one has to be very careful when we present this type of a pie chart. Particularly, when 2 or 3 areas are equal and slightly different about equal, then it becomes hard to distinguish from this. Sometimes we have seen people represented this way. Now, what do we do here? Now, the same data is presented here except that, you will realize that the last 3 are combined into one small pie or a wedge here which is further expanded to this at times, when we want to group data, aggregate data to make the pie look nice.

Now, you can see that this pie looks very nice compared to this pie. It looks very nice. We also realized the same 25 percent is actually here, here the same 25 percent that was here is actually here. And now, you see the last 3 simpler ones smaller ones. Now, aggregated into one which is further expanded and shown that within this within the last 6 50 percent is here the other ones are here. So, these are different ways of representing the pie chart.

(Refer Slide Time: 24:30)

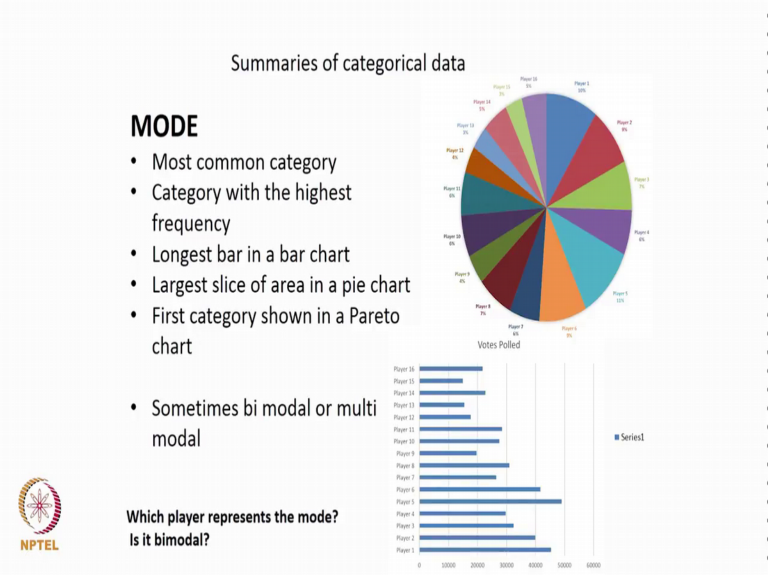


Colour of cars parked

Black	55
Blue	22
White	30
Gray	31
Red	20
Others	42

Now, let us look at some other example of colors of cars parked in a parking area very similar. Now, you can see that, this is one particular pie chart and here you can see that, it is expanded and it is shown here. Now, it includes the others. There is a category called others. Here in this pie the category called others is removed; the others as shown separately and so on. So, different form of representation of pie charts.

(Refer Slide Time: 25:00)



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?

Now, we look at how to summarize a categorical data and one important way of summarizing the categorical data is by what is called the mode. So, we will look at mode and other summaries of categorical data and numerical data in the subsequent lecture.