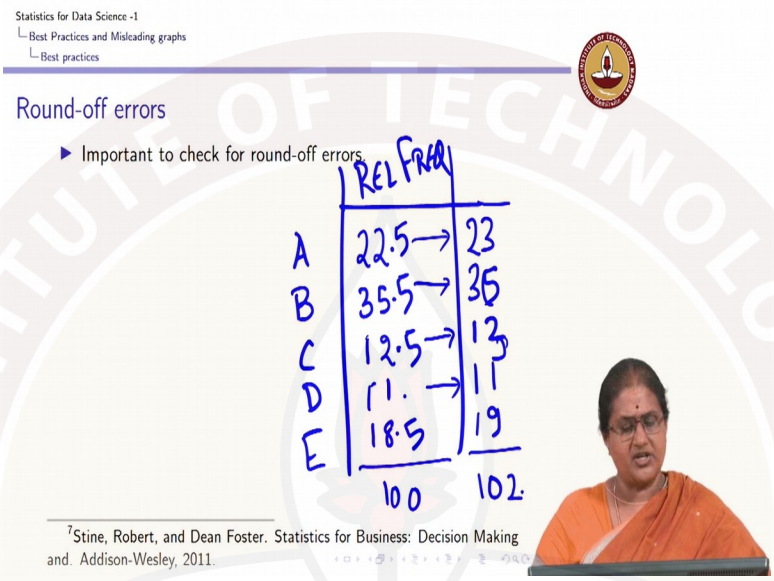
IIT Madras

ONLINE DEGREE

Statistics for Data Science - 1
Professor. Usha Mohan
Department of Management Studies
Indian Institute of Technology, Madras
Lecture No. 2.5
Mode and Median

(Refer Slide Time: 0:14)



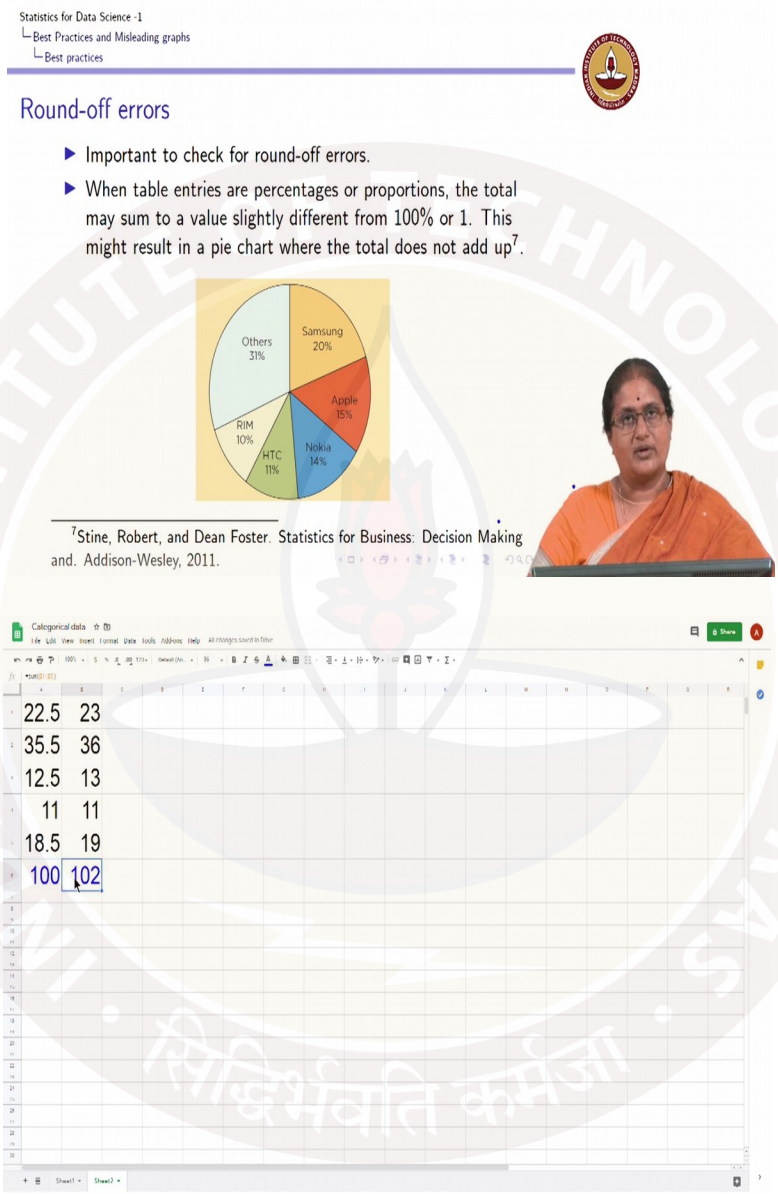
Now the other thing which is very important to check is whether you are introducing round off errors. Now what do I mean by round off errors? Now suppose I have a data set where I am having 22.73 I had 35.74, 11.30 and I have 18.23. Now suppose these are the relative frequencies of 5 categories I am naming the Category A, B, C, D, and E and these are the relative frequencies.

Many a time we are tempted to round them off. So, I just do a 30, 20, 35 dwell, I can look at 11.3 and it is 11 point 11.20, 11.53 and 19. A very quickly we look at this is suppose I have 12.5 and 18.03, 12.5. And I happen to round it off, I round it at 13 and things like that. So, you can see that when you round this data, so I am round in 22.5 to 23, 35.5 to 36, 12.5 to 13, 11 remains at 11 and 18.5 to 19.

What happens in this case is you can quickly see that now, in this case what happened in the earlier cases I had the total 22, 35, 12. So, when I round it up, I get a 23, I get a 36, I get a 13, I

get 11 and I get a 19. And you can see that what you could, what you achieve by rounding up is my total here is 100% whereas my total here is 102.

(Refer Slide Time: 2:36)

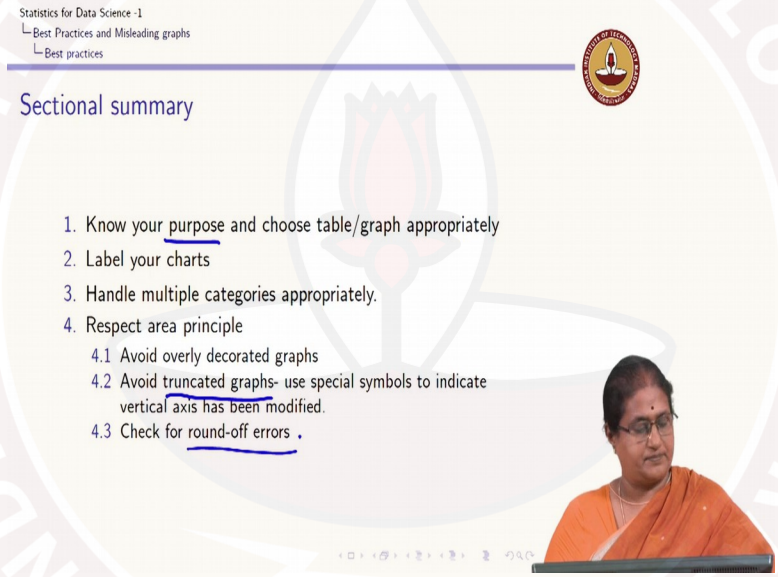


What is the implication of this, when you have actually table entries that are percentages or proportions, as we saw in the earlier case, what was happening is for simple data of this kind, you can see that I actually what is happening in this case is for simple data of this kind by rounding it off, you can see that this total is actually becoming actually the total becomes

changes from 100 to 102. So what would happen to this? So, look at a case where the data was rounded up, you can see 20 plus 15 is 35, 35 plus again, a 25.

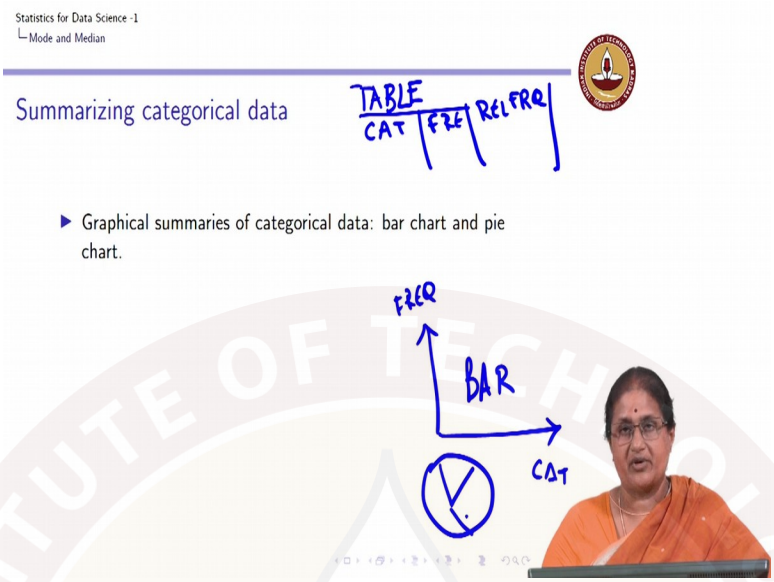
Here I have the 20 I have 15, I have 14, I have 11, I have a 10 and at 31, you can see that this does not add up to 100. And you can see that when it does not add up to 100. This is not forming what I call a pie chart. So, it should be extremely careful when you are actually rounding up. So, you have an error here, where my data is actually adding up to 101 and not to 100. So, what you should be very careful about is when you are actually rounding off you should be careful to see that the round off errors are avoided.

(Refer Slide Time: 4:12)



So, the first thing which we have learned so far is whenever you want to graphically or summarize your data through a table or graph know the purpose would be you want to label your charts or annotate them, handle multiple categories appropriately, even if the count of a particular category is negligible, combine categories as together of all categories, which have a small count, respect area principle, avoid overly decorated graphs. Avoid truncated graphs, because truncated graphs are mostly misleading. And finally check for round off errors.

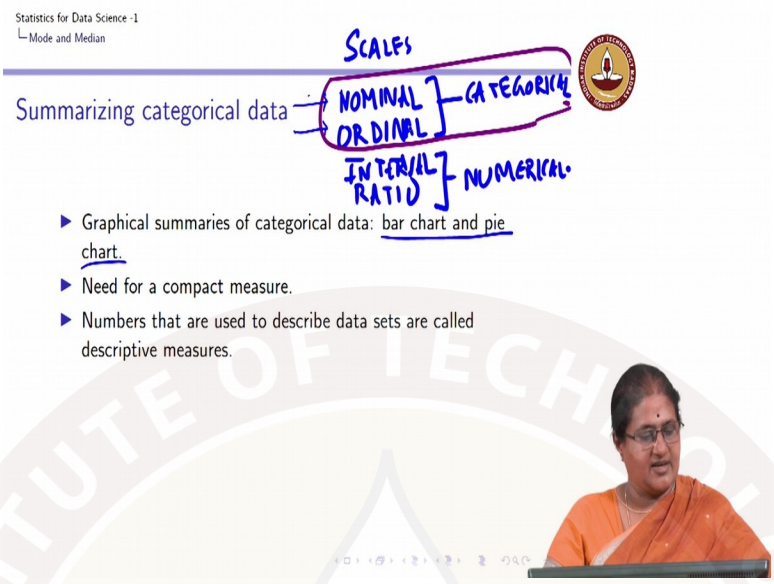
(Refer Slide Time: 5:03)



Now, the next thing which you are going to discuss is, So, far we have looked at graphical summaries and to tabulated data that is we looked at a frequency table, when we looked at a frequency table we talked about a frequency table where I have a category, I have a frequency, I have relate frequency. This is what I call was a frequency table.

Now, when I have when I plot category on my x axis with the frequency or count, I refer to it as a bar chart. And when I actually look at the relative frequency distribution, I call it a pie chart. This is what we have seen so far. And in the last section we saw how to label a pie chart how to label a bar chart and what are the certain things which we need to take into consideration when we are actually graphically summarizing data.

(Refer Slide Time: 6:04)

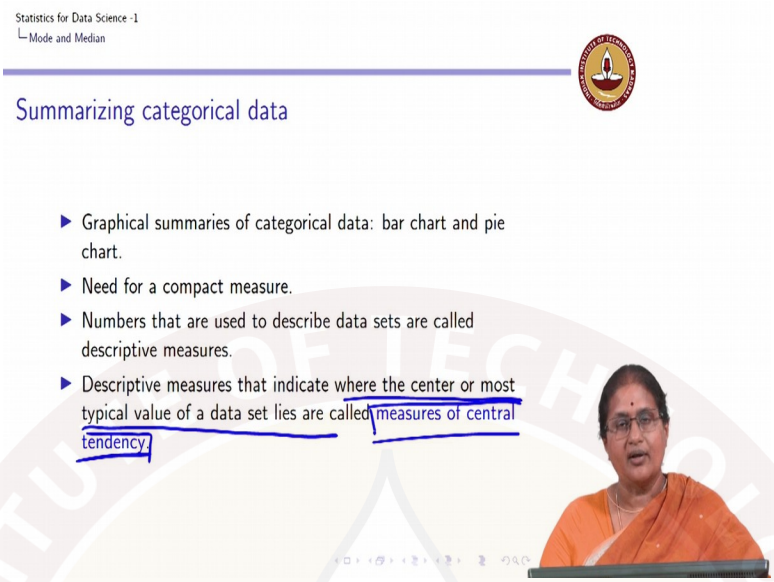


However, we need a compact measure sometimes to describe our data, what do we mean by disc compact measures. So, numbers that are used to describe data sets are called descriptive measures. Now, you may recall that when we looked at different types of variable we said the types of variables were nominal ordinal interval and ratio. We label these as the scales of measurement, where these two are basically when the variable is categorical or qualitative in nature and this is when I have numerical or quantitative variable.

Since now we are focused on what we are looking at now is the nominal or the categorical variable and you are looking at how to summarize the categorical variable. And what we looked at here so far is we looked at the graphical summaries when we looked at graphical summaries, we are looking at bar chart and pie chart. Now the question is, do I have any descriptive measure to describe datasets where my variable is categorical in nature? The answer is yes.

Now again you go back to your nominal and ordinal variables, we said that we cannot have any arithmetic operations described on this except for counting the number of observations in a particular category. The difference between nominal and ordinal is there is an order in the categorical variable for example, sizes of a T shirt from small, medium large there is an order good, excellent superlative. There is an order good bad, ugly, there is an order if ABCD represents great A is better than B is better than C is better than D again the represents an order.

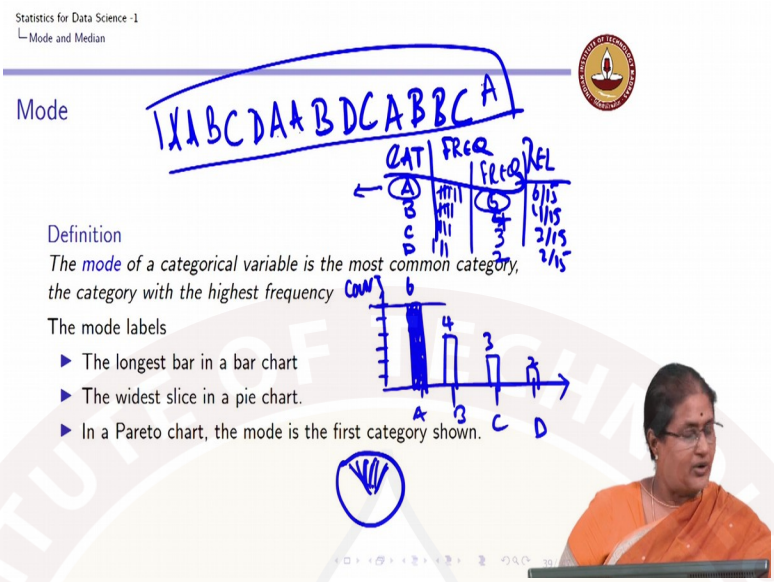
(Refer Slide Time: 8:32)



So, the question is, is there a number that I can use to describe these data sets? The answer is yes. When I describe the data set, the most typical value of a data set where the centre or most typical value of a data set lie is generally referred to as a measure of central tendency. That is where the centre or the typical value of a data set lies is called a measure of central tendency.

So, when we talk about a categorical variable, what is the measure of central tendency that you are referring to? I know that I cannot do any arithmetic operation on it.

(Refer Slide Time: 9:17)



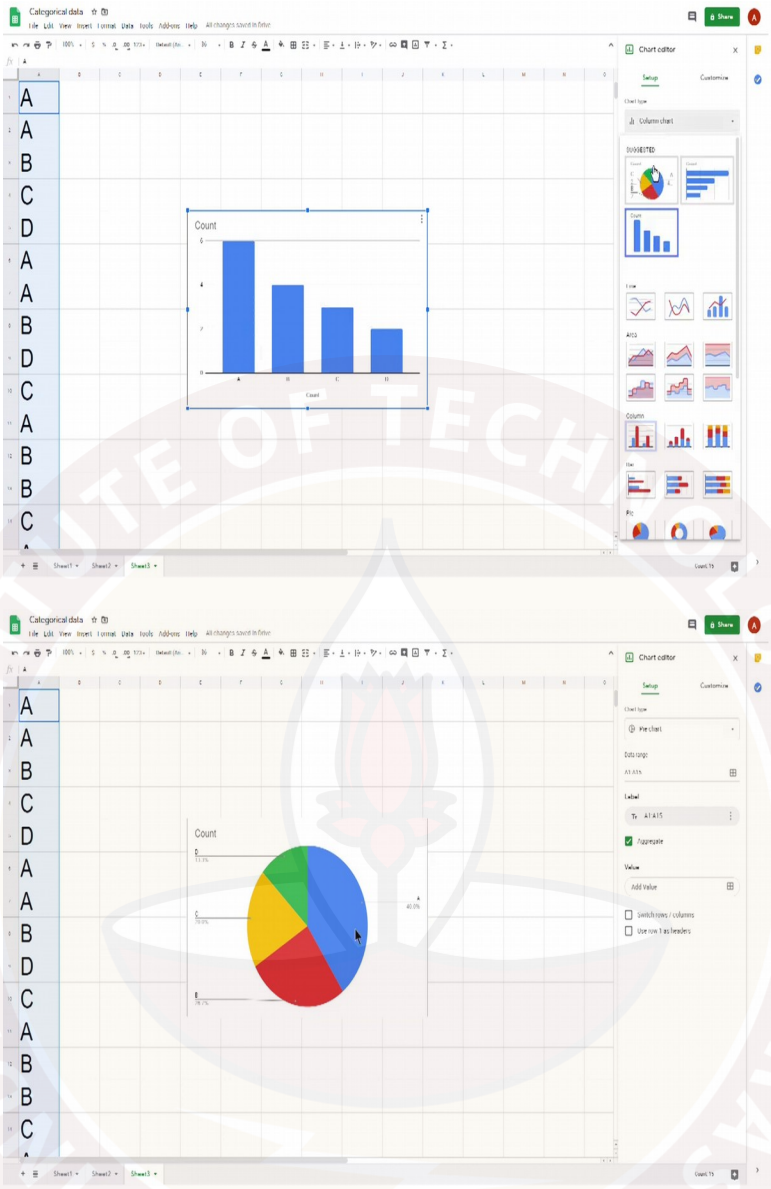
So, the first measure of central tendency is to count what is that category which has the highest frequency or highest count? So, if I have a data which is A, A, B, C, D, A, A, B, D, C, A, A, B, B, C, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14 and I have a 15 again categories.

The category I have here is A, B, C, D, the frequency the way looked at is A11, B, C, D, A, A, B, D C, A, B, B, C, B. So, I can see that the frequency 5, 5, 3, 2 now instead of B here, if I had a A, this would have been a 6, this would have been a 4 I would have had 15 observations with this as my frequency table. Now, what is the mode of this data set? The mode of this data set is this value A because it is has the highest frequency.

So, you can see that the mode of this data set is this category A. Now, when we are plotting charts, again I know that I can plot a chart here, this is the Category A, B, C, D, A 1, 2, 3, 4, 5, 6 this is the count or the frequency A is 6, B is 4 6 4 C is 3, D is 2. So, this A is the mode because it has the highest frequency and this is the longest or the tallest bar in the bar chart.

Now, similarly, if I am to come up with a pie chart, again I look at the relative frequencies we have seen in the earlier classes how I find the relative frequency 6 by 15 four by 15, 15 and 2 by 15 I find out the angles and then you can also check and you can find out that the relative frequency.

(Refer Slide Time: 12:10)



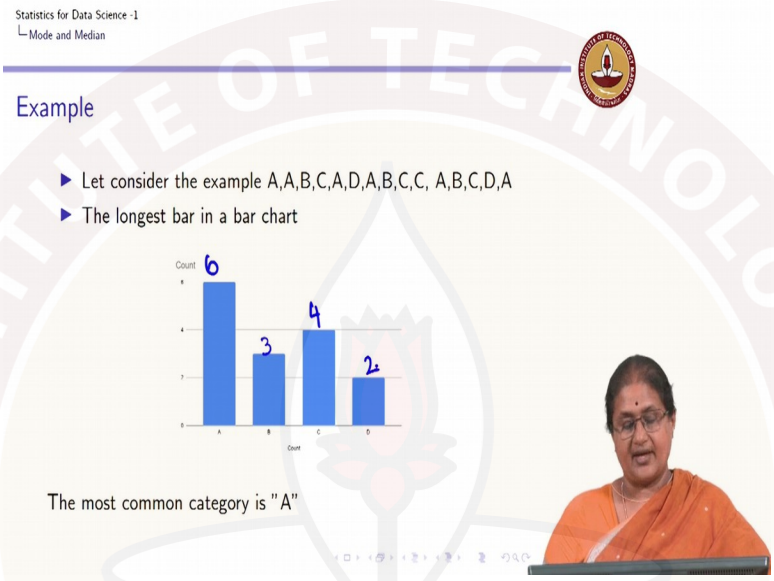
So, if I am going to plot this data, so, I look at this data, so, it is a A, A, B, C, D. So, I go to this I have my data A, A, B, C, D, A, A, B, C, D, A, A, A, B, D, C. So, if you look at this data here, so, this is the data which we just had, I look at this data so, this is the data for this data insert a pie chart. You can see that A which is here is the largest pie, go to a bar chart. A again has a count of 6 B has a count of 4, C has a count of 3 and D has a count of 2, the length of the longest bar and the largest by both of them are the mode.

Now again, if you look at the bar chart for this case, here the A, B C, D are actually this is 6, B, C, D, this is a Pareto chart, but in case I had something of this kind, this is not a Pareto chart because the Category B is appearing first and then a then C and D. This is just a bar chart which

is actually listing these categories B, A, C, D, but when I have a Pareto chart, the Pareto chart for it actually the mode is the first category shown.

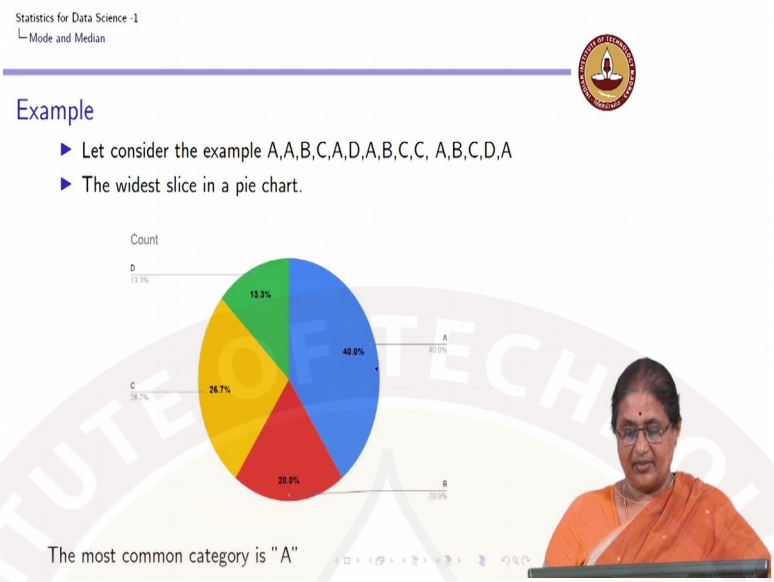
This is a Pareto chart the mode this is the first category that is shown and it is a Pareto chart. So, this is about what is a mode, the mode is the most common category with the highest frequency.

(Refer Slide Time: 14:09)



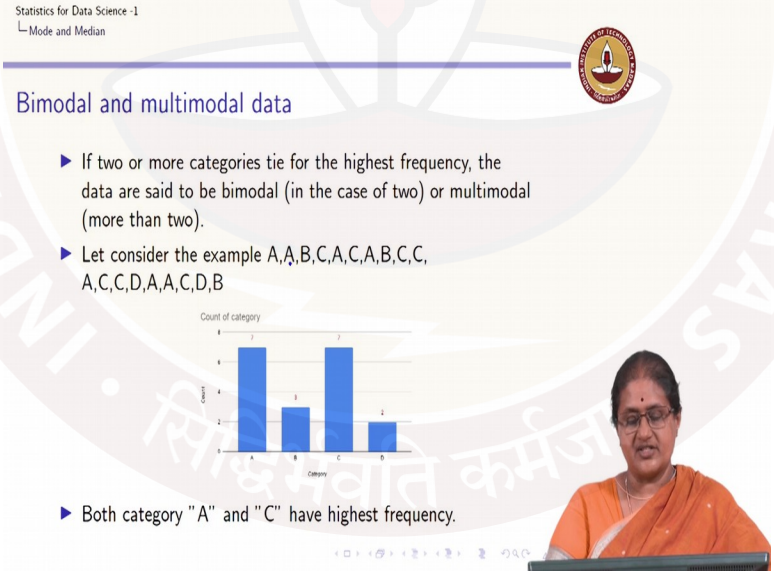
Now, this is the data which we are talking about, the most common category here is A something that you can notice in this data is the most common category is A, I have C, so, A, I have 6, I have B, which is 3, I have C which is 4, I have D which is 2 and A is my mode.

(Refer Slide Time: 14:36)



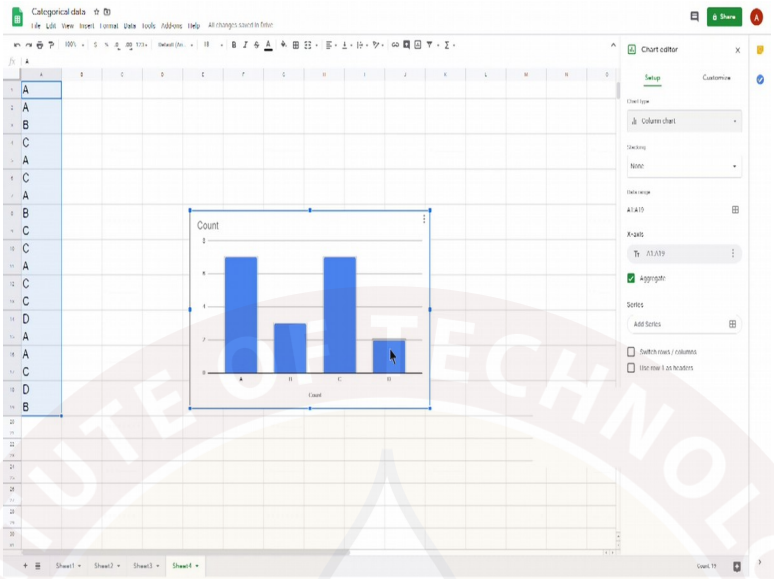
Similarly, for the same data, I have A which is 40% the most common category is A.

(Refer Slide Time: 14:44)



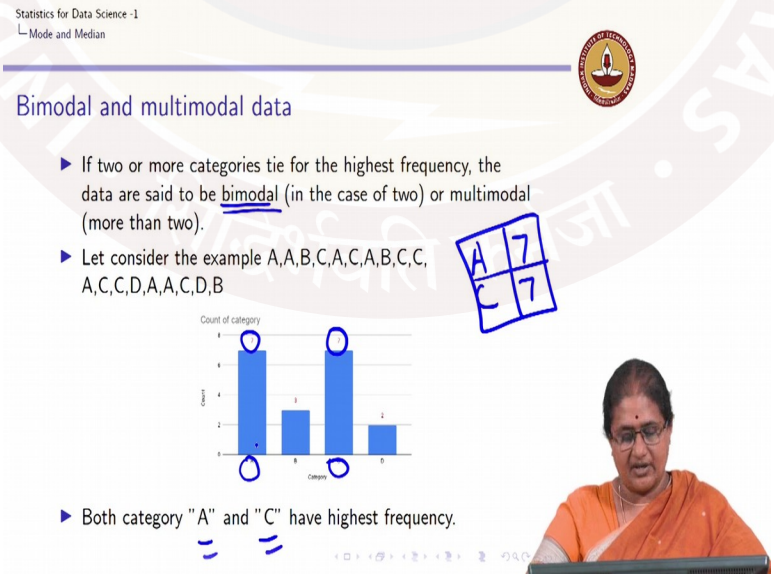
Now suppose, you have this data which is A, A, B, C, A so, I look at this other data where my data is A A B C A C A B C C.

(Refer Slide Time: 15:12)



So, this is my data set now and I just plot a column chart for this. Now what you look at this pie chart as you can see that both A and C are equally distributed or share of the pie is the same, which is 36.8%. Now this comes out much better in a bar chart where this is again 7 this is 7 B and C 2 and 1 respectively. But what is this info, what can we talk about this? So, when I have more than one category, in this case, I have A and C which have a frequency 7 each.

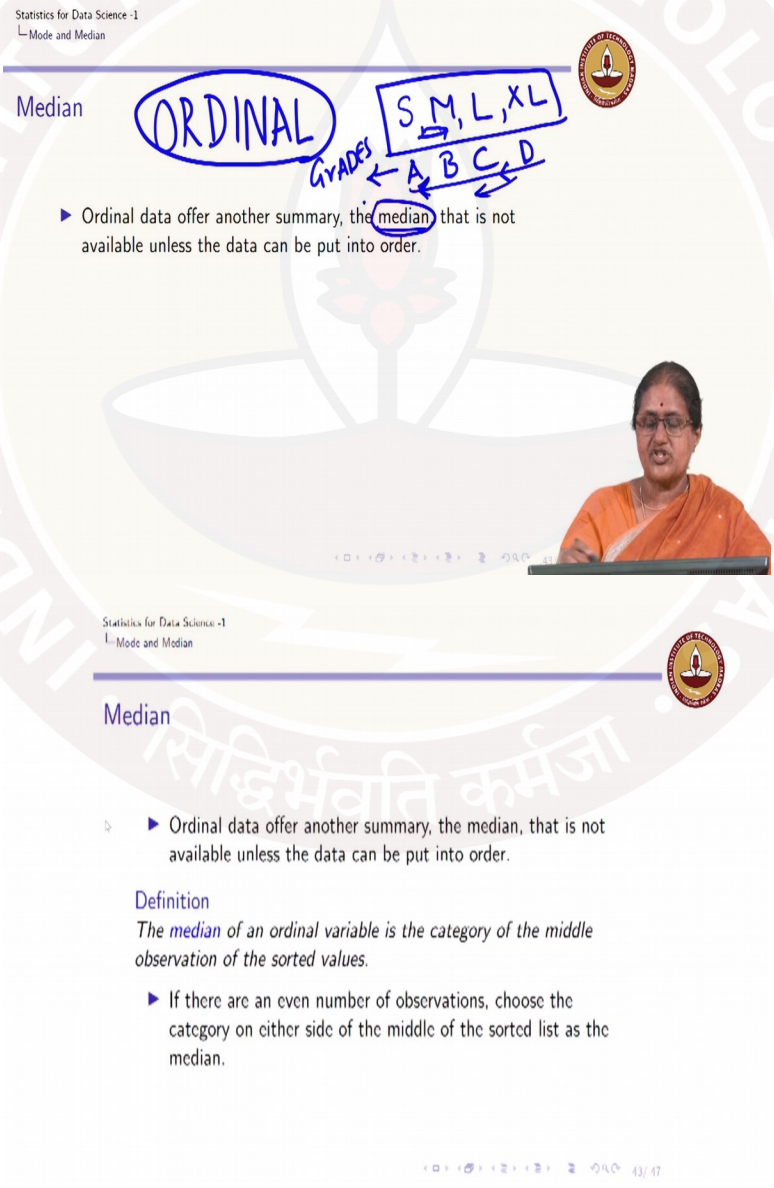
(Refer Slide Time: 15:55)



So, in a sense, both A and C have the highest frequency and I say the data is said to be bimodal. I repeat, if I have two categories or two or more categories that tie for the highest frequency as in this case, both A and C had a frequency of 7 each which is the highest frequency. Hence both A and C are having a tie for the highest frequency. My data is referred to as a by bimodal data.

If I have more than two categories, which are where there is a tie for the highest frequency, I refer to that data as multimodal data. So, this is about both Category A and C have the highest frequency.

(Refer Slide Time: 17:05)

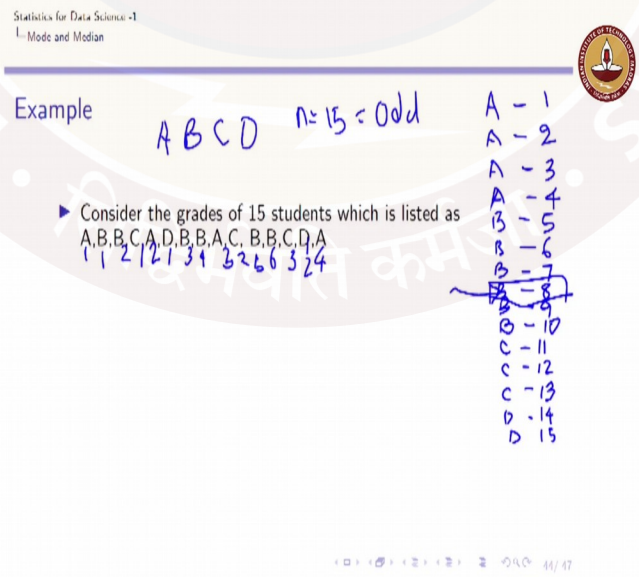


Now, another useful descriptive measure is what we refer to as a median. But a word of caution here is my data has to be ordered or ordinal in nature, I cannot define a median when I have a data, categorical data which is nominal. Again remember, small, medium, large, XL these are sizes of T shirts, this is ordinal data. A, B, C D if they referred to grades, this is again ordinal data because A is better than a B grade which is better than a C grade which is better than a D grade usually.

Now when you have ordinal data, I can offer another summary which is called the median. When we talk about mode for categorical data, we are counting that variable which appears the most number of times or that variable which has the highest frequency and that is what we refer to as the mode. Now, what is a median? Now, in order to compute a median, we first require the data to be ordered or ordinal data. So, unless the data can be put in some sort of an order, we cannot talk about the measure median, what is the measure median?

So, I can define the median of an ordinal variable to be that category of the middle observation of the sorted values. So, if there are even number of observing than the median could be either of the values of the middle, middle there be two middle observations, it could be either of those values, if it is odd then it is exactly the middle observation.

(Refer Slide Time: 19:16)



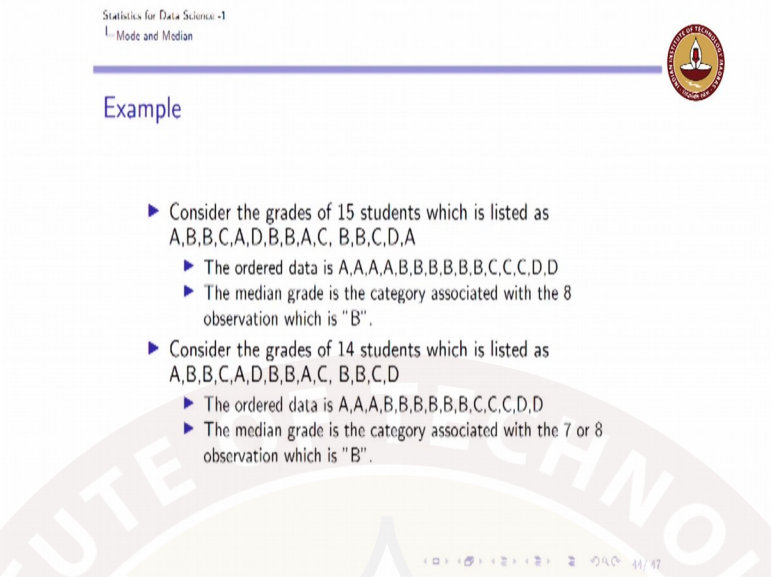
So, how do we compute the median? So, now you consider the grades of 15 students, which is listed as given. So, I have a A. Now, if I am going to actually order this consider A is the highest grade B is the next highest C is the next highest and D is the next highest. I am going to have an order of this kind. I have an A I have how many A's I have 1, 2, 3, 4 I have four A's So, I can order it A, A, A, A. So, my first observation is in A second is again an A, and a third is an A fourth is an A that is a rank order given to these observations, then I have a B 1, 2,3, 4, 5, and 6.

So, I write 6 B's ,1,2,3,4.5 and 6. So this is my fifth, 6,7,8,9, and 10. I have a C 1, 2, 3 a C, C, C, 11, 12, 13, I have a D. 2 D's. So 14, 15. So, what I have done here is we have listed the variables, what are the variables I have here I have A, I have B, I have C and I have D these are the variables. I have here. I have listed them in an order the order is A is better than B is better than C is better than D.

You could have chosen to have it in the other order also. Now once I have listed the variables in this order, you can see that the number of observations n equal to 15 is odd number that is what you can notice. So, if I look at what is that observation which divides this data set into exactly two halves, you can observe that this 8 observation here, has 7 above it 1, 2, 3, 4, 5, 6, 7. 1, 2, 3, 4, 5, 6, 7. So, this 8 observation which is corresponding to the variable B is the middle observation of the data set.

Hence, I can find out the ones I have ordered data, the median grade is that value which is associated with the 8 observation as given here and that is B.

(Refer Slide Time: 22:10)



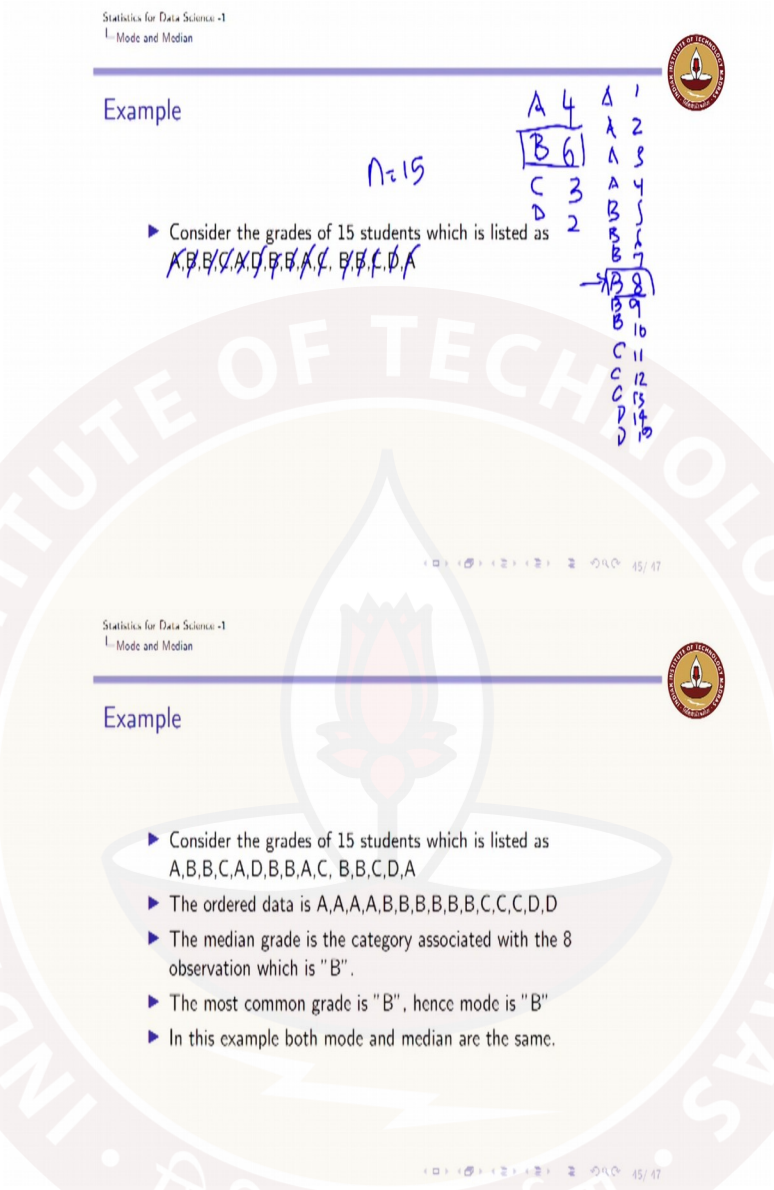
Now suppose I consider the grades of 14 students, when I have a grade of 14 students may n equal to 14. I can repeat the same exercise, and I can write it as I have 1. 2, 3 A's, so, A, A, A have 1, 2, 3, 4, 5, 6 B's, B, B, B. I have 1, 2, 3 C's and I have two D's. I again write an order. In this case, you can see the following is that I have n which is equal to 14. So, there is not n observation which can exactly have equal number of observations above it and below it.

So in this case, if I look at the seventh and eighth observations together, you can see it has 6 above it and 6 below it. So again, my median is again B, which is either the seventh observation, or the eighth observation, which in this case is B. Now, in case instead of the eight observation instead of a B I had a C, the median would have be either B, or C. So, this is how you compute a median of a categorical data that is ordinal. Why would you compute a median for categorical data?

The mode in a sense gives you the length of the longest bar in a bar chart or the largest pie in the pie chart because it gives you the that category or that variable which has the highest frequency. The case of a median when you have ordered data in a sense, for example, if this ABCD with sizes, then you would have said that this particular size particularly divides the data into two halves.

So, median in some sense gives you a measure of central tendency. We will look at this measure in greater detail when we look into measures for numerical data.

(Refer Slide Time: 24:45)

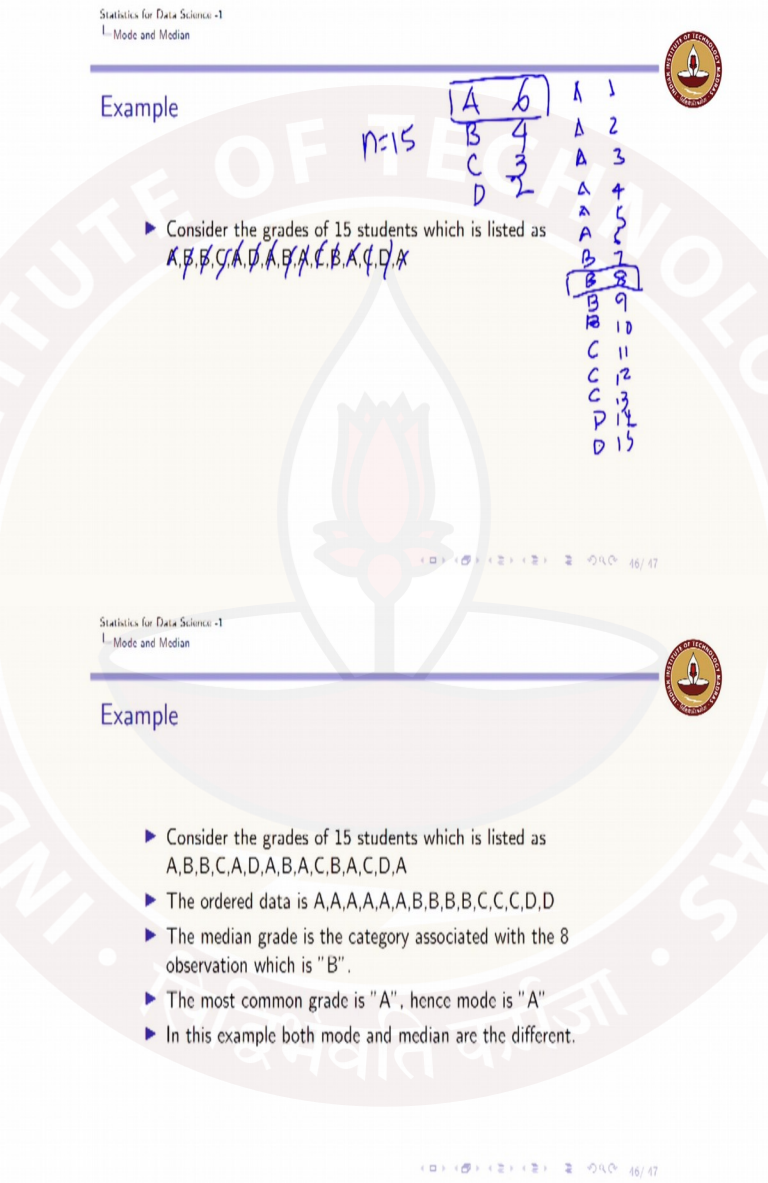


So, again, if you consider the grades of these 15 students again, I go back to the example here how again listed the grades of the 15 students. So, again, what do we have I have n equal to 15 I have listed the grades. So let us again count how many A's do I have I have 1, I have a 2, I have a 3, I have a 4.

So, I have a count of 4 for A, I have a count of 1, 2, 3, 4, 5, 6 a count of B, I have 6, I have C's 1, 2, 3 and I have a D, which is 2. So, this is again the same example I have a A, A ,A, A I have a B, B, B, B, B, B. I have a C, C, C and a D, D rank them 1, 2,3 ,4, 5, 6, 7, 8, 9 10, 11, 12, 13, 14

and 15 and I see the median is B. The mode is also B because B is that category which appears the largest frequency. So in this example, the median is B, the mode is also B. So the natural question to ask is, will the median and mode in any data set be the same. The answer is no.

(Refer Slide Time: 26:18)

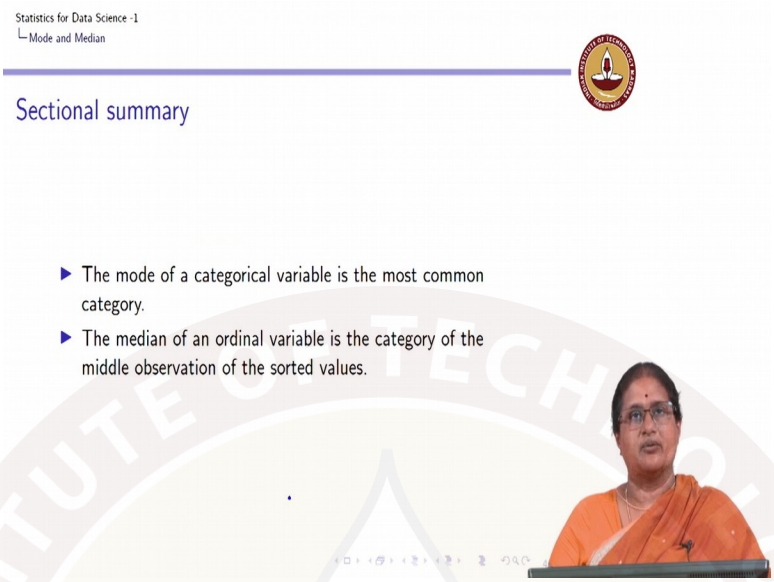


For example, consider the given data set. Again I repeat the exercise, I have again 4 categories A, B, C, D, the total number of observations is again 15. Again, let us count A, A 1, 2, 3, 4, 5, 6 A is 6, B is 1, 2, 3, 4, C is 1, 2, 3. And D is 2. Now let us look at the median I have A, A, A, A, A, A, A, B, B, B, B, C, C, C, D, D. So, I have ranking 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11,12, 13,14, 15

again the median for this data set is B, whereas the mode is A because it appears with the highest frequency and the highest frequency is given by 6.

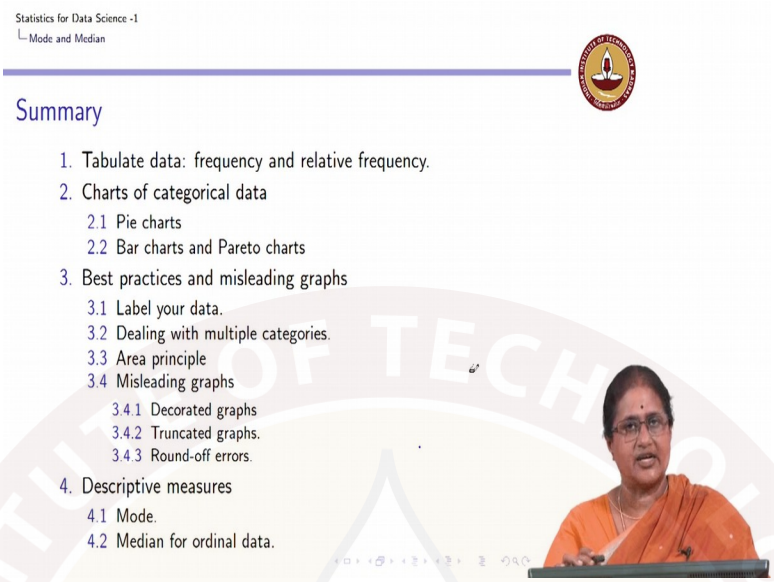
So this is an example of a data set where the median grade is B, whereas the mode grade is A. Sometimes these summaries are helpful in summarizing the performance of a class by relating or by giving a numerical descriptive measure instead of graphical measure in terms of mode and median.

(Refer Slide Time: 27:47)



So, what we have learned so far is when we come to numerical descriptive measures, in case of categorical data, there are two important measures. The first measure as what we refer to as a mode, which just gives you which is the most common category. The second measure is what we refer to as the median, which is the middle observation of sorted values, the mode works for nominal data, but to calculate the median, you need the data to be ordinal. So, with this we come to an end to with the module on describing categorical variables.

(Refer Slide Time: 28:31)



So, what should we have learned here is we started with tabulating data. What do we understand by tabulating data? We first identify what are the number of categories of my particular variable. If there how

many categories I state all these categories A, B, C, D, How many of our categories I have a list down all the categories. When I look at a tabulating or a frequency table, I have categories, I have the frequency.

Frequency is just the count of each category, I define what is the relative frequency, the relative frequency is the frequency by total count or total number of observations. So, this is frequency by total number of observations. If I had to summarize it using a pie chart where the question or purpose is to look at what is it as a composition of the whole, I plot the relative frequencies in terms of a pie chart, I can also plot it through a frequency which is called a bar chart.

A Pareto chart is where my categories are actually arranged in either decreasing or increasing order. When we have either a pie chart or a bar chart, labeling the data or annotating it, having appropriate A titles labeling your access all of them are extremely important. Do not ignore categories when you deal with multiple categories that is many categories. Instead club all the categories which have very few counts into one because they convey information we saw an

example where the multiple categories which had very low counts actually conveyed a lot of information.

Decorated graphs avoid if you can, do not use truncated grass which truncate the baseline and use An artificial baseline it could be very misleading, be aware of round off errors because round off errors could also not add up to your totals. Finally, we looked at descriptive measures. We looked at mode, mode is only for nominal when you have nominal data. The only possible descriptive measure is that for mode, when we have ordinal data we can have both mode than median.

We looked at an example where mod was equal to the median and we also looked at an example where the mode was actually lesser than the median. So, what where do we stand now with respect to the course with respect to the course we have now we understand what is our type of data we know how to categorize these data.

So given a variable, first given a data I can identify what are my variables I can identify what are my cases. I know how to tabulate my data, now given my data I can actually classify them into categorical or numerical. And given a categorical variable I know how to tabulate my categorical variable and summarize my categorical variable using a pie chart or by bar chart and to also give a descriptive measure in terms of a mode and median.