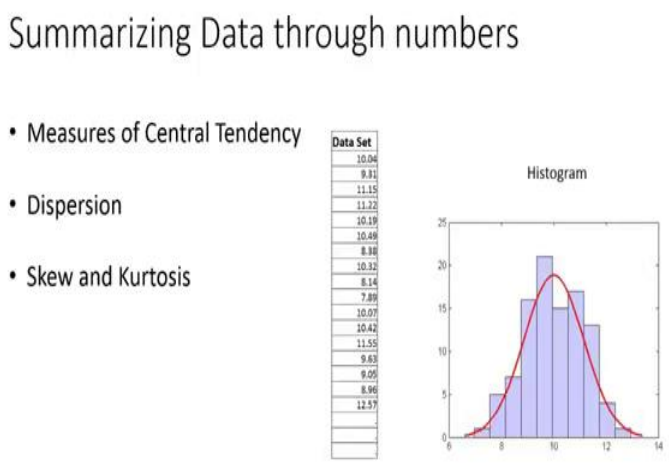
Introduction to Data Analytics
Prof. Nandan Sudarsanam and
Prof. B. Ravindran
Department of Management Studies and
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Module - 01
Lecture - 04
Descriptive Statistics: Summary Statistics: Measures of Central Tendency

Hello and welcome to our next module in the course, Introduction to Data Analytics. In this module, we continue our previous work on Descriptive Statistics and we… In our last session, we spoke about descriptive statistics through the use of various graphical and virtualization techniques. In this module, we start with the use of summary statistics or the idea that you can describe data with numbers, with numbers that summaries the data and most specifically, we are going to be talking about measures of central tendency in this lecture.

(Refer Slide Time: 00:56)



So, just to jog your memories we spoke about the idea that there could be a data set and a data set essentially would be representing, could potentially be representing a particular variable and so, I provided for you a simple example of a data set. And this data set is what is been captured in this histogram. The histogram is a visualization tool that we

spoke about in our last lecture. Now, the histogram essentially is a very rich representation, because it not only captures just some parameters associated with this data set, but it captures various new answers associated with it.

So, just to give you a quick reminder on how this works, essentially the entire x axis breaks down possible values that the data sets could take. So, for instance this bend is the series of values between 10 and from the looks of it 10.66 on this side. Now, depending on the number of data points that you see here, that fall within that range of 10 and 10.66 that would get counted here and out here it looks like that is about 15 points.

So, that is essentially how a histogram is calculated and the idea is that, using a histogram you could then fit something called distribution and the distribution is this red line that is shown on top of the histogram. And, so in some sense the histogram and the distribution that some time follows, tells us the full story associated with the data. And usually this red line is represented through some kind of a formula and we just call that, let say f(x) for now.

But, the basic idea behind summary statistics is that you do not even need to go this deep, I gave you this full picture just to tell you what the richest or the most detail story could be and our next session's, next modules are going to be about distributions. But, now let us take a step back. Is there something simpler that we could do? Is there something simpler without even fitting this distribution or even creating this histogram that could tell us a part of the story?

And the answer is yes and there are these various summary statistics that do exactly that and I am just going to talk about a few of them now. The first and the most common one are these measures of central tendency and what they mean is that, you have this data set and for now, let us just occupy ourselves with the histogram not the distribution that is fit on the top of the histogram. But, essentially with this histogram, what is a fairly central value.

So, it is clear that the values of this distribution go from here to about here, but what is something in the center and how do you define that. So, one idea is often to say, well one measure of central tendency is to see the minimum value going all the way to the maximum value and take something that is in between, so that is one way. Another way

could be to say, at what point in this histogram I am really covering about 50 percent of the area.

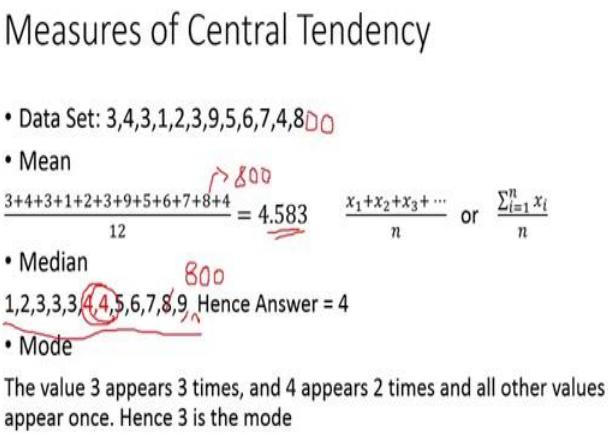
So, this histogram is defined by these blue bars and at what point am I covering about 50 percent of that area, so that could be one way of saying, what is central. There are other innovative ways, one of the most common one is to say think of this as a balance, as a sea saw essentially, this x axis line and all these blue bars are weights on top of it. So, the idea is, where would you want to put a fulcrum, such that this whole thing balances. This does not, one does not tip off the sea saw is essentially in balance.

So, that is the core idea behind measures of central tendency, what is a central value. Now, you then have measures of dispersion and the idea behind measures of dispersion are, that so you might have something that is central value. But, how a data point actually dispersed around the central value? Are they are very far away from them or are they very close to it and so on. And these are the two major forms of summary statistics that we will cover in this course and that you are likely to encounter.

But, you might have also heard of the concepts of skew and kurtosis and so, it just briefly what mentioning it, skew concerns the shape of this distribution itself and the like, the fact that sometimes distributions lean to one side versus the other and this is the fairly colloquial way of saying it. But, instead this distribution the red line being the way it is could you have a distribution that looked like that, which means it is leaning, the whole thing is leaning to one side.

Again, I am just giving you conceptual feel for it, it is not a formal definition, we and then, in the same line kurtosis is the idea that how fat are the tails of the distribution and what we mean by that is, you know having a similar distribution that looks like this. So, that tails themselves are fatter than the other and that property gets captured with kurtosis, so great. So, this is, this should I have given you some a brief overview of the different, the over hatching idea of summarizing statistics through mean, describing statistics through summary statistics or numbers as a means of describing distributions.

(Refer Slide Time: 07:43)



We now, go into the major subject of this lecture which is measures of central tendency. So, the best way to do this is through a concrete example and we, that is what we will do with this step. So, there are three major measures of central tendency and they are the mean, median and mode. With the mean, the core idea and many of these you might have encountered, you might have come across before.

So, bear with me if you already heard this, the idea behind mean is just that it is the concept of average. If you have a data set and I am just given you a sample data set out here and, so it is the numbers 3, 4, 3, 1 and I have kept the data sets smalls, so that I can illustrate the concept typically might be dealing with much larger data sets, but the idea remains the same. So, for this data set the mean is nothing but, the sum of all the numbers divided by the total number of a numbers there are.

So, we take each numbers 3, 4, 3 and we add them all up and the 12 that you get out here is the actual number of numbers that there are in this list. So, once you divide it and that is the concept of mean, which can also be represented mathematically in this form and I have just shown that, you use of that, when you see that it is, you not surprised by it. So, great and incidentally the mean is the concept that I was speaking about in this histogram of balancing the seesaw, where would you place the fulcrum; such that this seesaw gets balanced, that is the same concept of mean.

We now move on to the next measure of central tendency, which is called the median. The median is calculated by arranging all the numbers in order. So, you had this data set and it was 3,4,3 that is what you had out of here, but when you bring it to median, you basically takes the smallest number put it first and then, in some order ascending or descending you arrange the numbers. Once you do that, you choose the central number and that is your median.

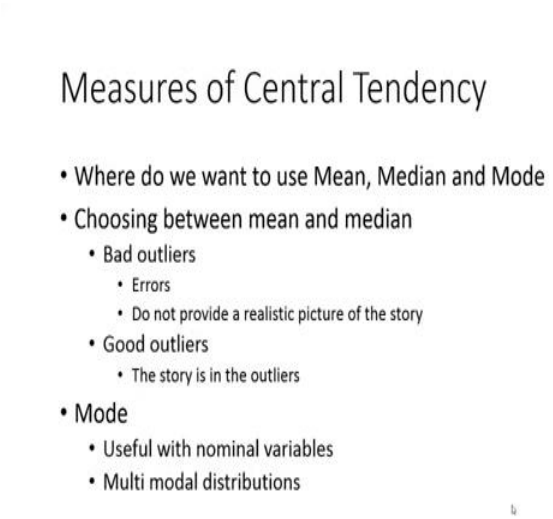
Now, choosing that central number is quite easy when you have an odd number of numbers, so if you had 9 numbers the 5th number would be the central number, which you have 4 numbers before and you have 4 numbers after words and that is your central number, but when you have an even number of numbers… So, in this particular case we have 12 numbers, the central number is really not 1 number it is 2 numbers. So, at here it winds up at being a 6th and the 7th number.

So, typically there you choose the 6th and the 7th number and take the average. In a particular case that is not a problem, because it happens to be the same number and quite easily we say that the answer is 4. The mode, which is a third measure of central tendency and there might be a few others, but these are the three most common ones that you will encounter, the mode essentially says what is the most common value. So, if you look at this data set, the number 3 appears 3 times, the number 4 appears twice and then, all the other numbers just appear once.

And, so out here it is fairly clear that the number 3 is the most common one and hence 3 is the answer if you, if the question is, what is the mode. It is the most common number, but and that kind of make sense, if you have a data set, where there are only few numbers that are recurring, but the concept is again generalizable ((Refer Time: 11:49)). So, even if you had a data set that look like this, where numbers you know, it does not make sense to ask the question, which is the most common number in fact, no number might repeat itself.

But, out here the more essentially is based on the range itself. So, you would say this is the most common range and so this is the mode and so the mode out here would have been something like if 9.5 to 9.6, so great. So, we now understand how mean, median and mode are calculated.

(Refer Slide Time: 12:33)



Now, let us take a step and see, where do we want to use which measure of central tendency. So, how do we choose between mean, median and mode? The main concept really comes between mean and median. So, let us talk about that for a minute and becomes kind of obvious, where mode is more useful, because it has a very different property associated with this central tendency, so great.

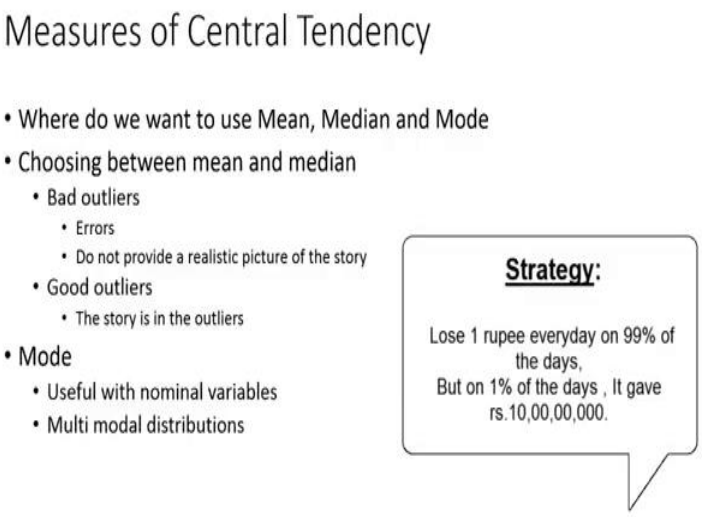
If you have to choose between mean and median, much of the debate usually comes down to outliers. The idea of the outliers is that, it is a number or a value that is not really within that set of most of the other numbers that you see. Now, that can be, because of quite of few reasons. When this come about because of an error in the data, so the data set itself could have an error, then it is easy to say that is around you, so this is a bad outlier.

But, sometimes this state can be that outlier is very much not an error and it tells an important part of the story. In really simple words, the median is not influenced much by the outlier, whereas the mean is greatly influenced by the outlier. For that reason, the median is often kind of expressed as been a more robust metric to outliers. But, that we need to take a step back, just a second about saying you know outliers can either be good or they need not be good and so, it really depends on what we think about the outliers, ask to whether we choose to go with the mean or median.

Obviously, when we think the outlier is a bad think, it should not be there or it is not contributing towards a story that we want to tell. Then, we call it a bad outlier; we prefer the median in that case. So, to actually give you some insight as to, how the outlier affects the mean and not the median. Let us just go back to the previous example Let us say that in our data set instead of 8 we had 800. So, that is clearly a mistake. Someone by mistake type two 0's next to 8 and, so it is 800 and let us assume it is a mistake, it is not obvious.

Now, in the case of the median, this would have a huge impact instead of 8 being here you would put an 800 and that would greatly change this number. So, your mean is largely affected, it probably send the number into the 100's. Now, it will have no impact on the median, this 8 gets listed, it becomes 800, which means that it is not in it is place, it comes after 9. So, make some space here, so the 800 comes here, but the central two numbers still remain these two 4s, I mean these two 4s, so your answer really did not change.

(Refer Slide Time: 16:10)



So, in many ways the outlier has like this huge impact on the mean, it has no impact on the median. Now, clearly if what you have faced with this kind of an error, you like using the median, because the mean is susceptible to this problem. There might be other situations, where you want to use the median and again it pertains outliers, but here we

are not as much scared about errors, but you are scared that there is this one a typical case, which is just skewing a story that I want to tell.

So, a classic example of where medians are used is, when looking at salaries of people, where the idea is that salaries having some sense of exponential. Many people earn a consistent salary and then, there is these few peoples who just earn these catastrophically high amounts and so something like mean, where and here the idea is the catastrophically high amounts are these outliers. And here talking about a mean will not give you the typical salary that a person earns, because of these one or two people who earn very high salaries.

So, it is not just errors there might be other situations, where you have outliers and you feel like the presence of these outliers is moving you away from talking really about, what is a typical value, now having said that there are many situations again, where you are dealing with outliers, but these outliers are of very important part of the story. So, let me give you an example of this. So, let us say you have this data set, where you were looking at a particular financial strategy.

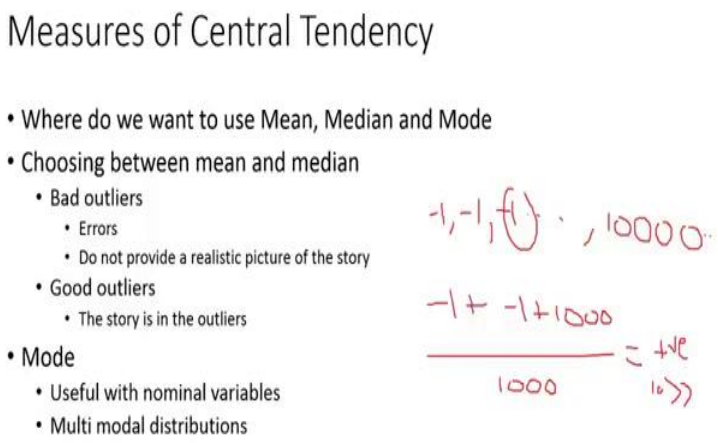
And in this financial strategy you are looking at how much money you made on a daily basis and, so you have taken some historic data and you want to see you want to see, what is a typical scenario of the strategy you want to evaluate the strategy based on based on this data set. So, let us see this financial strategy actually made you lose 1 rupee every day on 99 percent of the days, but on 1 percent of the days this strategy gave you 10 crores, so large enough number.

So, this strategy made you lose money on 99 percent of the days and on 1 percent of the days gave you 10 crores is this strategy you would like to take the very straight forward answer is if you like making money you really like this strategy, because despite the fact that you lose just 1 rupee on 99 percent on the days as long you as you can play this game or you can trade on this strategy in a stock market for long enough period of time here bound to in the long run make good amount of money.

Because, 10 crores more than compensates for the 99 days or during, which you lost the 1 rupee. Now, let us see how mean and median would have represented this data right if you got a sufficiently large enough data set of having actually play the strategy. So, let us say you go and you actually collect your data set of the strategy over a 1000 days or less

than 10, 1000 days, what would be the median of this strategy the answer is said the median would have been the -1 that you that the 1 rupee that you lost. So, quite simply, how is that on that data set for this would look something like this.

(Refer Slide Time: 20:04)



It would look like this -1, 1 1, -1,…., with the lot of minus 1's 99 percent of them are minus 1's and then, the odd time you are going to find this are really large number I am not even going to talk about how many 0's lots of 0's dot, dot, dot for the 0's. So, you put this in ascending order and you choose the middle value that is going to be a minus 1, so the median gives you a minus 1.

However, you put these numbers you add all of these numbers up together include put your 10 crores in there with lots of 0's and then, divided by the total number of data points, which is you know 1000 or 10000 or something whatever the number of data point you have and you going to get a very large positive number positive and large great science. So, here is a story where yes you're not you have an outlier you got this huge outlier which is 10 crores.

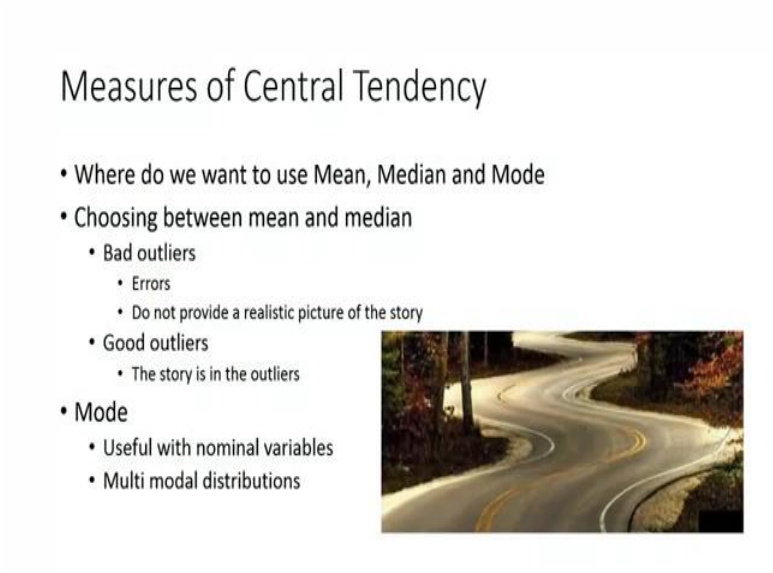
But, the story was in the outlier that is as much real money that you made or lost as the 1 rupee is that you lost. So, here is the case where the mean is probably a great measure to go by if you had to choose whether to play this strategy or not. So, I have that gives you some idea between mean and median, now let us talk a little bit about mode is an

interesting one, because it just blank it says I am going to take, the value, which is the most popular and that works fine for you know distributions, which are fairly symmetric.

But, in many cases that that people do not find that too meaningful the one big advantage; however, that the mode has is that you can even use a nominal variables. So, you might have a situation, where you are just counting the number of you are coming up with the count associated with a categorical variable an example could be in the number of reds the number of greens the number of yellows and all the mode is going to do is say lets pick the one which has the most number of it and it can also be fairly useful in multi modal distributions.

Let me give you an example of where a multi modal distribution and multi moral just means that there are many peeks to the distributions. So, if you go back to this slide and let me just erase that for you. So, the red line is the distribution and we going by the we going talk a lot more about distributions multi modal distribution is one that might look like this, so there are like two peeks to this distribution. So, let me give you an example a real life example of where, you could have a multi modal distribution and, where you might want to use the mode.

(Refer Slide Time: 23:13)

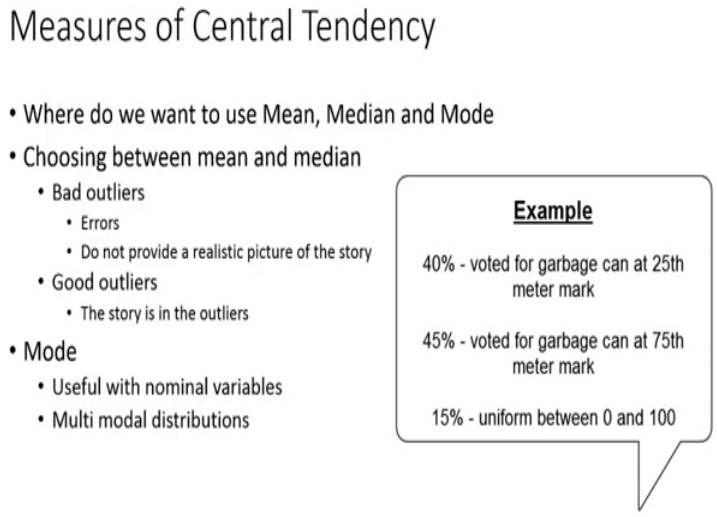


So, let us say you we looked on this street and this street was a 100 meters long. So, one end of this street is 0 meters and then, there are markers on this street. So, this 1 meter, 2 meter, 3 meter and the street goes all the way to 100 meters. So, if someone said this 75

meter you immediately knew, which point of the street or road we are talking about. So, all the residence of this street need to make a decision, on where to place a garbage can a trash can, which lets say for whatever reason people do not people have strong opinions of that.

So, people are all going to go or we going to take we are going to take a survey of all the residence and each ones going to come up with the number. So, person one says I want the garbage can in the 25 meter mark another person says I have want it in the 50 meter mark so on and so forth.

(Refer Slide Time: 24:18)



Now, let us say we collected all these data and we found that 40 percent of the residence said they want the garbage can in the 25th meter mark. Let us say another 40 percent or let us say 45 percent said they wanted the trash can on the 75th meter mark. So, just to recap 40 percent of the people say they want the trash can on the 25th meter mark you know 45 percent of the residency they wanted in the 75th meter mark and the remaining 15 percent they just its all over between its like uniform somewhere between 0 to a 100.

Now, the problem is both mean and median might windup saying the average preference is to keep the trash can somewhere in the 51 52 meter mark, because that is bound to be a central value. Now, that might be something that nobody wanted, where as something like a mode would just categorically say keep it in the 75th meter, because that is the most populist preference.

So, in case is, where kind of taking two extremes and averaging them out and in some sense median also does that as long as there are enough data point does not work and in those cases the mode could be fairly useful application. I have just gives you an idea of the difference measures of central tendency.

In the next lecture we will take up measures of dispersion.