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

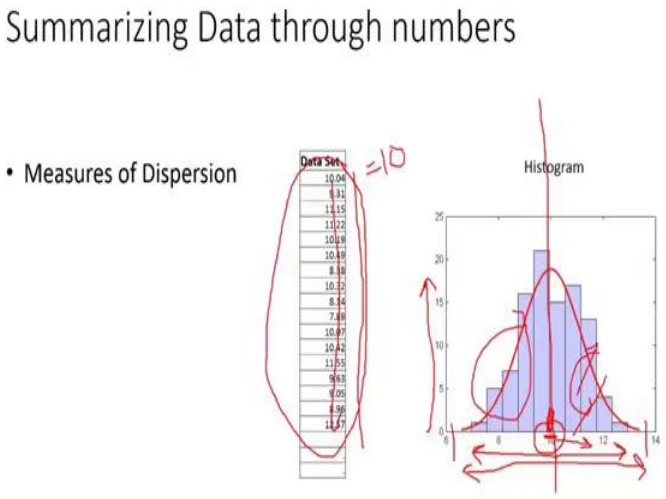
Module - 01

Lecture - 05

Descriptive Statistics: Summary Statistics: Measures of Dispersion

Hello and welcome to our course Introduction to Data Analytics. In this lecture, we continue our work in Descriptive Statistics to give you a timeline of where we are. We have discussed within descriptive statistics, the various graphical and visualization techniques and the second half of the descriptive statistics deals with Summary Statistics, The use of numbers to describe and summarize data. Within summary statistics, in our previous lecture we spoke about Measures of Central Tendency. In this lecture, we will be concluding the use of Summary Statistics with discussion on Measures of Dispersion.

(Refer Slide Time: 01:03)



So, we should be at this point fairly familiar with the use of this data set, essentially the data set is just a sample, which talks about different data points over a certain range and the histogram that you see to the right hand side of this is a histogram that is generated

from this data set. So, we covered histograms during our lecture on graphical techniques and we use them extensively in our discussions of measures of central tendency.

Again to be very clear, measures of central tendency and measures of dispersion and the specific matrix that we would discussed in them. So, for instance in measures of central tendency we spoke about mean, median mode and today we going to be speaking about some matrix associated with dispersion. All of these matrices drawn in any way shape or form need this histogram. They directly operate on this data set and in some sense, you might say why even talk about the histogram.

And the idea is that you are absolutely right, we do not need this histogram, but it really helps to explain the concepts and it is also healthy way to start thinking about these matrices and start thinking about these distributions. It will help us also in the long run, when we cover concepts and probability distributions. So, now, having said that let us talk about what measures of dispersion seek to capture.

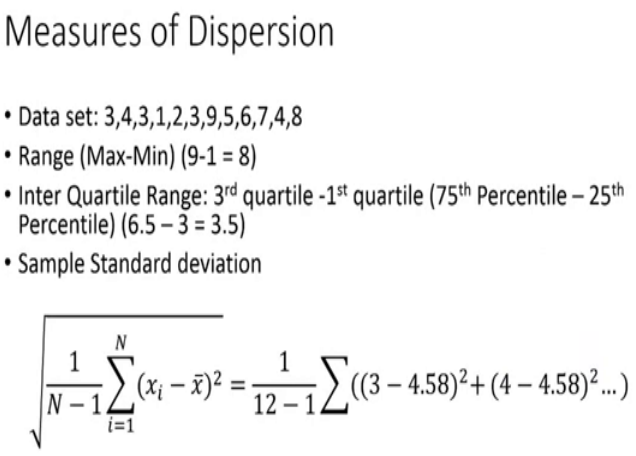
We spoke about in the last lecture, how measures of central tendency try to capture in some sense a central value; some central value within this range of values that this data set takes up. So, the range of values of this data set takes up is shown here and in some sense, the histogram captures their likelihood in this axis. So, the range is here in the x axis and the y axis is in some sense, the likelihood of seeing that data and we spoke about, how measures of central tendency try to captures, what appears to be like a central value and we spoke about different matrix that do that.

Measures of dispersion talk really about, how the data is dispersed around this value, how does the data deviate from this value. For instance, if every single data point and let us for now assume that 10 is our measure of central tendency. One measure of central tendency is the mean, so for now let us just say we are using the mean. So, 10 is the mean, then if every single data point in this data set was equal to 10 and it is not, what is here, but it is every value is equal to 10.

Then, there would be no deviation of data from 10, you would just see a single tall line in this histogram and none of the sides, none of these would exists all together, but that is not the case. Typically, most data sets, the values are going to be different and you might have some measure of central tendency, but there is going to be some amount of deviation of the data points on either side to the central value. Now, measures of

dispersion try to capture that, do the values deviate a lot from the center or do they deviate very little from this center and that is what measures of dispersion seek to capture.

(Refer Slide Time: 05:11)



To understand the different measures of dispersion, let us go back to the data set that we were using, when we were speaking about measures of central tendency. So, I have used the same data set out here and the simplest measure of dispersion is range and the range is quite simply nothing but, the highest value minus the lowest value. So, in this particular case the highest value is 9, the lowest value is 1, so quite simply 9 minus 1 is 8 and that should be intuited for you.

The given that the mean of this data set is about 4.5, 4.6, a measure of dispersion is just the max minus min. Now, if for instance if there was very low dispersion, then the highest value would be close to 4.6 and the lowest value would be close to 4.6 and so that, range between max minus min could have been smaller. At the same time, if this dispersion is very high, on the high side and on the low side your max and min value is going to deviate a lot from the 4.6 and so, you would have high dispersion. So, that is the simple one.

This second one is the Inter Quartile Range and the idea here is highly related to the concept of median, where you would arrange the data points and you would kind of take a central data point. Another way of saying that, we discussed that procedure of median

during measures of central tendency, but another way of thinking of it is that, you are taking the 50th percentile point. With the Inter Quartile Range, what you are doing is you are taking the 75th percentile point or the 3rd quartile and subtracting from at the 25^{th} percentile point.

The idea being that within this data set if there is a high level of dispersion, then that range between the 75th percentile point to the 25th percentile point also be high and if the dispersion is low, then this range would be low and it is really noteworthy that this is the concept that gets captured in box plots, which we discussed in a graphical techniques. We spoke about, how in the box plots the upper line of the box plot and the lower line of the box plot, correspond usually to the third quartile and the first quartile of your data set. So, and this is also known as the Inter Quartile Range. So, it might be abbreviated to IQR in some text books, but this is also a measure of dispersion.

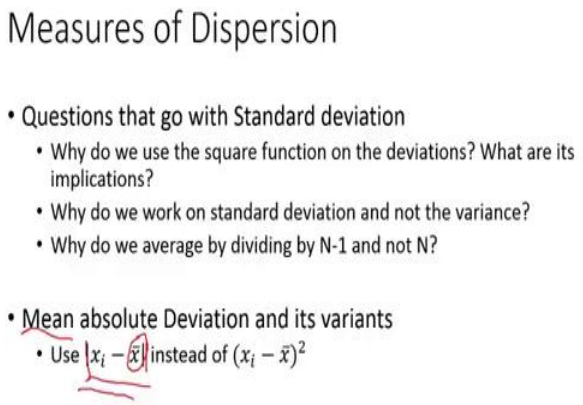
We then come to, what is a fairly popular measure of dispersion and the idea behind this is to essentially look it, how much each data point deviates from the mean that you just calculated. So, x i represents each data point, because i goes from the first value to the n th value, when in a particular example n is 12. And, so we wanted to take each data point, see how much it deviates from the mean. In a particular case for this data set, the mean is 4.58.

So, we will take the first data point which is 3, so the data point 3 and we subtract them from 4.58 and square that value. We would take the second data point to the same thing and we would keep adding up these squares and once you add up these squares, you take something that kind of looks like an average and it is not an exact average, because you have this minus 1 and we will talk about that in a minute. But, in concept you essentially are trying to get an average of the square deviation and you will ultimately take a square root of this.

Now, when you take the square root, what you get is the standard deviation and when you do not take a square root, you get this measure called variance and variance is also a measure of dispersion. In concept, the only difference between standard deviation and variance is that, 1 is the square root of the other. Again, now that you understand how a standard deviation is calculated. Let us go through some questions that might have come

up, when we discuss standard deviations. Given that the other two methods that we have discussed are of fairly straight forward and clear.

(Refer Slide Time: 09:45)



So, here is some questions that always go with standard deviation. Why do we use this square function on the deviations and what are it is implications? So, what we are referring to here is the fact that, we actually take the square of the deviation. So, why, what is the purpose? If you want to calculate, see if you want to get some measure of average deviation, why not just take the deviation and take the average of it and the answer is fairly straight forward to that.

The answer is that just by definition, because you are looking at the deviation from the mean, there are going to be some points that deviate from the mean on a positive side, there going to be some points that deviates from the mean on the negative side. So, 3 minus 4.58 would lead to a negative number, whereas 9 minus 4.58 would have been a positive number and again by definition, because of how you calculate a mean and the math for this is fairly straight forward.

You will find that if you just took the deviations, some positive numbers and some negative numbers and you added them up, you would always get zero and that is because of, how the mean is calculated, because the mean is nothing but, the sum of all the numbers divided by the total number of such numbers. So, by definition just taking the

deviation would result in some positive numbers and some negative numbers, which should cancel each other out and give you zero.

So, what you really trying to capture is an average deviation, but you do not want the signs. So, what is one great thing you can do is to square it all. So, when you square a number, whether it is negative or positive, you always get a positive number and the other really interesting thing is the, only thing that matters is the magnitude. So, minus 3 square is 9, which is also the same as plus 3 square. So, the idea is that the square function is symmetric on the plus minus side and always gives you a positive number. So, for that reason we use the square function.

Now, are there some implications of that and the answer is, yes there are some implications ((Refer Time: 12:07)). The implications is, the effect that squaring has. So, let us say you had two separate deviations of one unit each, so let us say you had two data points that signified that there was a deviation of one unit. So, given the average is 4.58, let us say you had a 3.58 deviation. So, 3.58 and you had another data point, which was 5.58, so both of these would have a deviation of minus 1 and plus 1 and when you square these two numbers, so you square these two numbers, the answer comes out to be 2. So, that is what happens when you do this entire squaring process.

Now, what happens when in one case? So we had two… We just focused on two data points. Now, what happens in one case when you just… In one case, you are right on the mean. So, you are right on the mean, in the other case you are deviating by two points essentially or whatever the unit you are using, a deviation is two units. So, here the deviation, because you are comparing it to the mean, you are essentially just replacing this 4 point, you are replacing this 3 with this 4.58 and we are looking at what would happen.

In case, because you are doing that your deviation is zero. Here your deviation is 2 and because it gets squared, that becomes 4 and so, your cumulative deviation in some sense is 4, whereas in the previous case your cumulative deviation is only calculated as 2. In both cases, you deviated by two units from your mean across the two data points. In one case, you deviated by one unit in the first data point and one unit in the second data point, but the sum of the squares led you to a number 2.

In the second case, you deviated from the mean by zero data points in the first, by zero units in the first data point and again two units in the second data point. So, in both cases if you just look at the actual deviation from the mean, in both cases you have deviated by only two points, but in the second instance, in this instance you would be recording a square deviation of four units, which is twice as much as the square deviation of the first case, which is two units.

Now, many people like that and there are many contexts, where that makes a lot of sense. There are some contexts, where this justice not make sense, but that is one of the implications of squaring the deviations. So, second question is, why do we work on standard deviation and not the variance? So, the idea is, ((Refer Time: 15:35)) why do we take this square root. Why not just report the variance, why do we report mean, because they both the same function and the answer again is fairly straight forward.

You have a data set and some units, that is 3, 4, 3, 1 could have some units and these units could be things like, simple things like rupees or kilometers per hour, whatever it is that you know. You might have then collecting data own and when you report a deviation from the mean, the units would then be in squared if you are using variance. So, if you use variance you will have to report a value that is in square and, so what is it mean to say rupees square.

So, what is it mean to say a dispersion is a 500 rupees square and you know, rupees squared is not something that we can understand ((Refer Time: 16:40)), it is far more intuitive. When it, truly it is form a meaningful to say a deviation is 23 rupees from the mean, you can make decisions based off of that and you can gather some insights based off of that. So, third question and often a very interesting question is why do we average by dividing by n minus 1 and not n.

So, the idea here is that the sum of the deviations is always zero and so the last deviation, because you are essentially ((Refer Time: 17:11)) doing a series of deviations. Now, the last deviation you can be found, once we know the other n minus 1 deviations. So, we are not really averaging n unrelated numbers you are really averaging only n minus 1, a squared deviations. In some sense, it is almost like only the n minus 1 square deviations can vary freely and we average by dividing the total, essentially by n minus 1.

This is also the concept of degrees of freedom, which is how many of the values can actually move freely with, can move freely and still maintain the final statistic and in this case, the final statistic is the mean, because you subtracting each number from the sample mean. Now, the important thing is, this mean which is 4.58 in our case is something that was calculated from this data. So, from the same data, which we are using to calculate the standard deviation, you calculated the mean and that is the reason essentially that you are using the n minus 1.

If instead you are not using this mean, but someone came and told you, what the true mean of this data was. Someone said, here is the data set and by the way, the mean of this data set is 5. So, they just told you the data set or you knew the data set from past experience or you are able to compute that, then you would not have to do the n minus 1 and you would do the n, but also out here you would not substitute 4.58, you would be substituting 5.

So, in each of these places you would be substituting 5, which is the true mean. We call that the true mean and we call 4.58 the sample mean, because 4.58 you calculated from this data, whereas 5 is something that you knew on principle or you are able to use some other source to know, what the true mean was. Now, another way the people like to describe this is also to say for instance that, if this 3, 4, 3, 1 this data is ultimately a sample from some other population, then you need to essentially do what we just discussed, now which is to use this n minus 1 and take the sample mean.

So, again we are talking about the case, where nobody comes and tells you what the true mean is. So, your only hope is to calculate a mean from the data and you calculated 4.58 and because this data set is a sample from something else that generating this data, the right way to do it is the way the standard deviation formula right now is shown. But, if in some sense this data is the population, it is not a sample from some universe, but it is the real deal.

Then, again the idea would be to use n and not n minus 1, because this is the true mean and again out here, you would be substituting the 4.58, but then this should be called a population standard deviation. So, it is POP, population standard deviation. But, more often the not in terms of the more realistic situation that you will encounter in life, I think

it is fairly safe to say that, if you are taking the sample and you are calculating the mean, use n minus 1.

If you given a sample data set, but you already know the mean, that is you are not calculating it from this data set you already know the true mean. In that case, you can just go ahead and use n instead of n minus 1 and that would be the right standard deviation. So, that is as far as standard deviation goes, but before we conclude on measures of dispersion, it is worth mentioning that there are some other measures of dispersion out there and these are called mean absolute deviation and there are many variance to it.

But, the core idea is that, with mean absolute deviations you replace what you use in standard deviation, which is the deviation of each point from it is mean and squaring it, you replace that with an actual deviation. So, the deviation and this sign which is the two vertical lines on either side, what the essentially mean is that the negative symbol just goes away. So, a 3 minus 3.58 minus 4.58 which would result in minus 1 would just be written down as a 1 and so would a 5.58 minus of 4.58.

So, negative signs are just taken off and then you do and the other operation of the same. The good thing with mean absolute deviation is that, it has lot of variance, so it is like what is the average deviation from the mean, that is the typical case and that is what I have written down here, but you can also replace this x bar with the median of the x's. So, the mean absolute deviation from the median is another case and you also have cases like, what is the median absolute deviation from the mean, the median absolute deviation from the median.

Obviously, our previous lecture on understanding the pros and cons of means and medians would play an important role in making such a selection. So, that should conclude a lecture on measures of dispersion. In the next lecture, we will continue with descriptive statistics, but focusing more on distributions.

Thank you.