Artificial Intelligence (AI) for Investments
Prof. Abhinava Tripathi
Department of Industrial and Management Engineering
Indian Institute of Technology – Kanpur

Lecture – 15
Descriptive Analytics: Introduction

(Refer Slide Time: 00:13)

In this lesson, we will discuss Descriptive Analytics. **(Video Starts: 00:17)** Businesses use analytics to explore and examine the data and then transform their findings into insights that ultimately help executives, managers and operational employees make better more informed business decisions. One of the key types of analytics used in the business is descriptive analytics.

Descriptive analytics is a commonly used form of data analysis, whereby historical data is collected, organized and then presented in a way that is easily understood. Descriptive analytics is a field of statistics that focuses on gathering and summarizing raw data to be easily interpreted. When we begin to analyse data, it is essential to first explore our data before we spend time building complicated models.

One easy way to do so, is to calculate some descriptive statistics for our data. Descriptive statistics analysis helps to describe the basic features of a data set and generates a short summary of the sample and measures of the data. In descriptive analysis, we describe our data with the help of various representative methods like charts, graphs, tables, excel, files, etcetera.

In the descriptive analysis, we describe our data in some manner and present in a meaningful way to be easily understood.

Most of the time it is performed on small data sets and this analysis helps us a lot to predict some future trends based on the current findings. The descriptive statistic can be categorized into three parts first measures of central tendency, second measures of variation and third measures of shape. **(Video Ends: 01:40)**

(Refer Slide Time: 01:41)

In this video, we will introduce descriptive statistics namely measures of central tendency, measures of variation and measures of shape. **(Video Starts: 01:49)** As a fund manager several prospective clients are asking for more information on which they can base their investment decisions. In particular they would like to compare the results of an individual specific fund to the results of similar funds.

For example, while the earlier work your team did shows how the three year return percentages are distributed? Prospective clients would like to know, how to value for a particular mid-cap growth fund as compared to the three year returns of all mid cap growth funds? They also seek to understand, the variation among returns but all the values relatively similar.

And does any variable have outliers value that are either extremely small or extremely large. While doing a complete search of the funds, data would lead to answers to the preceding questions, you wonder if there are better ways than extensive searching to uncover those

answers. You also wonder if there are other ways of being more descriptive about the sample of funds providing answers to questions not yet raised by prospective clients.

If you can help the provide such answers prospective clients will be able to better evaluate their investment in funds that your firm features. Business use analytics to explore and examine the data and then transform their findings into insights that ultimately help executives, managers and operational employees make better and more informed business decisions.

One of the key types of analytics used in business is descriptive analytics. Descriptive analytics is a commonly used word for data analysis whereby, historical data is collected, organized and then presented in a way that is easily understood. Descriptive analytics is a field of statistics that focuses on gathering and summarizing raw data to be easily interpreted. When we begin to analyse data it is essential to first explore our data before we spend time building complicated models.

One easy way to do so, is to calculate some descriptive statistics for our data. Descriptive statistical analysis helps to describe the basic features of a data and generate short summary of the sample and measures of the data. In descriptive analysis, we describe our data with the help of various representative methods like charts, graphs, tables, excels, files, etcetera. In the descriptive analysis, we describe our data in some manner and present it in a more meaningful way to be easily understood.

Most of the times it is performed on small data sets and this analysis helps us with a lot of information to predict future trends based on current findings. The three key descriptive statistics that can be categorized are measures of central tendency, measure of variation and measures of shape. Let us start with measures of central tendency. Central tendency is the extent to which values of a numerical variable group around a typical or central value.

A measure of central tendency is a summary statistic that represents the centre point or typical value of a data set. In statistics the three most common measures of central tendency are the mean, median and mode. Most variables show a distinct tendency to group around the central value, when people talk about an average value or the middle value or the most frequent value they are talking about mean, median and mode, the three key measures of central tendency.

Mean, it is the sum of observations divided by total number of observations. Median, it is the middle value of the data set, it splits the data into two halves. Mode, it is a value that has the highest frequency in a given data set. Quartiles, quartiles are a measure of central tendency that divide a group of data into four subgroups or parts. In the next set of videos, we will talk about these measures of central tendency in greater detail with more examples. **(Video Ends: 05:29) (Refer Slide Time: 05:29)**

In this video, we will start with an important measure of central tendency that is mean. **(Video Starts: 05:34)** Mean is defined as an average which is the sum divided by count. The arithmetic mean that is everyday usage the mean is the most common measure of central tendency. To calculate a mean, sum the values in a set of data and then divide that sum by the number of values in the data set.

The mean can suggest a typical or central value and serves as a balance point in a set of data. Similar to a fulcrum on a seesaw, the mean is the only common measure in which all the values play an equal role and that is why it is very important. The symbol X here in the formula read as X bar or mean of X or average of X represents the mean of the sample. The sample mean is the sum of the values in a sample divided by the number of values in the sample.

For a sample containing n values, the equation for the mean of the sample is X bar is equal to sum of n values divided by n. The same formula can be described using the series of values that is X 1, X 2 and so on up to X n to represent the set of n values. This equation becomes

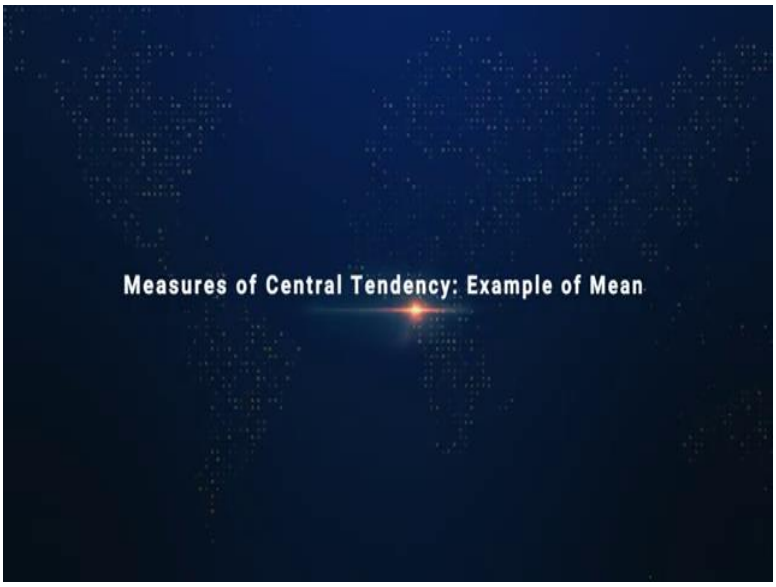
$$\bar{X} = \frac{(X_1 + X_2 + X_3 + \cdots + X_n)}{n}$$

Using summation notation to replace the numerator that is X 1 + X 2 and so on up to Xn with the term sigma Xi which is the summation or sum of all the X i values starting from the first X 1 to the last value that is X n provides the formal definition of the mean

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Here X = sample mean, N = sample size, X = i th value of the variable X_i and sigma X_i is the sum of all the X_i values in the sample. Because all the values play an equal role, mean is greatly affected by any value that is very different from others. **(Video Ends: 07:23)**

(Refer Slide Time: 07:25)



In this video, we will go through the examples of the mean. **(Video Starts: 07:28)** When a set of data contains extreme values, avoid using the mean as a measure of central tendency. As an example of using a sample mean, consider that knowing the typical time to get ready in the morning might help people to better manage their weekday schedules. The collected values for time to get ready in the morning is provided here in the form of table.

The following observations are collected as shown in the table here, the mean of this data is simply computed as shown here

$$\bar{X} = \frac{(X_1 + X_2 + X_3 + \cdots + X_n)}{n}$$

In this case it is 396 divided by 10 which is 39.6 even though no individual day in the sample data has a value of 39.6 minutes, allotting this amount of time to get ready in the morning would be a reasonable decision to make.

The mean is a good measure of central tendency in this case because the data set does not contain any exceptionally small or large values. To show how mean can be greatly affected by any value that is very different from the others, imagine that on day 3 a set of unusual circumstances delayed the person getting ready by an extra. So that the time for that day was 1 or 3 minutes this extreme value causes the mean to rise by 45.6 minutes as shown in the calculations here.

The one extreme value that has increased the mean by 6 minutes this extreme value has also moved the position of the mean related to all the other values. For example, the original mean of 39.6 minutes had the middle or central position among the data values, five of the times were less than that mean and five were greater than that mean. In contrast the mean using the extreme value is greater than 9 of the 10 times making the new mean a very poor measure of central tendency. **(Video Ends: 09:09)**

(Refer Slide Time: 09:09)

In this video, let us look at the measures of central tendency that are mean, median, mode, quartiles and percentiles. **(Video Starts: 09:16)** Median is the middle value in an ordered array of data that has been ranked from smallest to largest. Half the values are smaller than or equal to the median and half of the values are larger than or equal to the median. Extreme values do not affect the median, making median a good alternative to the mean when such values exist in data.

To calculate the median for a set of data first rank the values from smallest to largest and then use the equation shown here to calculate the rank of the value, that is, median. Median here is $\frac{n+1}{2} th\ ranked\ value$

There are two simple rules to compute median. Rule number one if the data set contains an odd number of values the median is the measurement associated with the middle rank value.

Rule number two if the data set contains an even number of values the median is the measurement associated with the average of the two middle rank values. To further analyse, let us consider a sample of ten time to get ready values as we have seen earlier, we will try to compute the median for this data. To do so, we need to first rank these as per their magnitude starting from 1 to 10.

Because the result of dividing n + 1 by 2 for this case is 10 + 1 divided by 2 which is equal to 5.5. One must use the rule number two and average the measurements associated with the fifth and sixth rank values that is the average of 39 and 40 which is 39.5 and therefore in this case median is 39.5. The median of 39.5 means that for half of the day time to get ready is less than or equal to 39.5 and for other half of the days the time to get ready is greater than or equal to 39.5.

In this case the median time to get ready is 39.5 minutes that is very close to the mean time to get ready for 39.69. In the previous section we noted that substituting 103 minutes for the time of 43 minutes increase the mean by 6 minutes. Doing the same substitution does not affect the value of median which would remain at 39.5 this discussion illustrates that the median is not affected by extreme values.

Our next measure of central tendency is mode, mode is the value that appears most frequently like the median and unlike the mean extreme values do not affect the mode. For a particular variable there can be several modes or no mode at all. For example, consider again that values of time to get ready those ten values to get ready in morning, the values are shown in the table provided here.

These values are ranked according to their magnitude, notice that there are two modes that is 39 and 44 because each of these values occur twice in the data. Consider another example

where a systems manager is in charge of a company's network and keeps track of the number of server failures that occur in a day. The failure data is provided here for past two weeks which represent the number of server failures per day for the past two weeks.

The observations are shown and they are also ranked according to their magnitude in the table. The ordered array of these data suggests that because 3 occurs five times more time than any other value the mode is 3. Thus, the systems managers can say that the most common occurrence is having three server failures in a day for this data set the median is also equal to 3 and the mean is equal to 4.5.

The value 26 is an extreme value for this data, the median and the mode are better measures of central tendency than the mean. Our next measure of central tendency is quartiles. Quartiles are measures of central tendency that divide a group of data into four subgroups or parts. The three quartiles are noted as Q1, Q2 and Q 3 the first quartile Q 1 separates the first or lowest one-fourth of the data from the upper three-fourths and equals the 25th percentile.

The second quartile Q2 separates the second quarter of the data from the third quarter, Q 2 is located at the fiftieth percentile and equals the median of the data. The third quartile, Q3 divides the first three quarters of the data from the last quarter and is equal to the value of 75th percentile. The computation of quartile requires that the values have been first ranked from the smallest to the largest.

For example, to calculate the quartiles for the sample of ten observations get ready times, first rank the data from smallest to largest. The rules for calculating the quartiles from this data set of rank values are discussed here. Rule number one if the rank value is a whole number the quartile is equal to the measurement that corresponds to the rank value. For example, if your sample size is 7, the first quartile Q1 is equal to the measurement associated with 7 + 1 that is 8 divided by 4 which is second ranked value.

Rule number two if the rank value is a fractional half that is like 2.5 or 4.5, the quartile is equal to the measurement that corresponds to the average of the measurements because of the two rank values involved. For example, if the sample size is n = 9, the first quartile Q1 = 9 + 1 upon 4 that is 2.5 rank value, halfway between second rank and third ranked value. Rule number three if we rank the value that is neither a whole number nor a fractional half, round the result

to nearest integer and select the measurement corresponding to the rank value. For example, if the sample size n = 10, the first quartile Q1 = 10 + 1 divided by 4 that is 2.75 rank value. Now, we will round off this 2.7523 and use the third ranked value that is closest integer to 2.75 which is 3. Consider the following example which involves our earlier time to get ready values, these values are ranked according to their magnitude from 1 to 10.

Notice that the first quartile is $\frac{n+1}{4} = \frac{10+1}{4} = 2.75$. So, using the rule three will round up to the third rank value that is 3, the third rank value for the get ready times data is 35 minutes. Interpret the first quartile of 35 to mean that on 25% of the days the time to get ready is less than or equal to 35 minutes, while that on 75 percent of the days the time to get ready value is greater than or equal to 35 minutes.

The third quartile is 3(n + 1)/ 4 , that is, 3(10+1) /4 = 8.25 ranked value. Again, using the rule number three for quartiles round this down to the eighth rank value which is the closest integer. The eighth rank value here is 44 minutes and therefore we interpret the third quartile to mean that on 75 percent of days the time to get ready is less than or equal to 44 minutes and on 25 percent of the days the time to get ready is greater than or equal to 44 minutes.

Another interesting measure of central tendency is percentiles, percentile is related to quartiles and split the variable into 100 equal parts. By this definition, the first quartile is equivalent to 25th percentile, the second quartile is equal to 50th percentile and the third quartile is equal to 75th percentile and so on. **(Video Ends: 16:09)**

(Refer Slide Time: 16:10)

Our next topic is measures of central tendency that is IQR interquartile range. **(Video Starts: 16:15)** Our final and a very important measure of central tendency is the interquartile range also called as mid spread. It measures the difference in the centre of distribution between the third and first quartiles. Interquartile range is the difference between the third quartile and the first quartile.

And the formula for interquartile range is equal to $Q3 - Q1$, where Q3 represents the third quartile and Q1 represents the first quartile. The interquartile range measures the spread in the middle 55 percent of the values and is not influenced by extreme values, the interquartile range can be used to determine whether to classify an extreme value as an outlier. If a value is either more than 1.5 times the interquartile range below the first quartile or more than 1.5 times the interquartile range above the third quartile, that value can be classified as an outlier as well. Therefore, this interquartile range measure is very important. Calculating the interquartile range can be further analysed with the sample of ten get ready time that we have been using for now. One can calculate the interquartile range for these times. First we need to order the data as per their magnitude. Then using the equation for interquartile range that is Q 3 – Q 1, we can see that Q1 = 35, Q3 = 44, we have already calculated this and therefore the interquartile range is 44 – 35 = 9 minutes. Therefore, for our get ready times data the interquartile range for this ten get ready times is 9 minutes with the interval from 35 to 44 which is often referred to as middle 50. **(Video Ends: 17:56)**

(Refer Slide Time: 17:56)

In this video, we will discuss the excel implementation of measures of central tendency starting with computation of mean. **(Video Starts: 18:04)** We will take an example of different serial brand and their calorific values. The corresponding values are provided here these are different serial brands and their calorie values are provided. We will start by computing their mean values, a very simple way to compute this would be use of average formula.

We will use the average of all these calorie values, the value is 130. Another simple example or way to do that would be to compute the sum of these values, like this ,and then divide it by the number of values and we are expecting the same result. Also please notice, as we discussed earlier this mean or average value is considerably affected by the distribution of values around the mean and any extreme value can considerably change the results.

For example, notice if I modify one of the values to extreme outlier for example, I modify 200 to 1000 notice now, the mean value is substantially high and it is almost more than the second largest value which is 190. So, it is even higher than 190 and it becomes 244 and therefore it does not reflect the mean distribution or the central tendency of the data. This is happening because one of the values in the data is extremely high which is the highest value, we can consider it as extreme outlier on the higher side

In this video, we will discuss the excel implementation of median computation. We will carry on with our example of serial brands and their calories. A very simple way to compute the median for a data is to use this median formula provided in excel and the value is 110, also we can use the discussion that we had in the previous videos to compute this median.

First, we need to count the number of observations in this data and that is 7. Now, we will use that (n + 1)/ 2 which is 4 and therefore since it is an even number 8, we get a value of 4 which is 110. As we can observe in the data the fourth value if ranked as per magnitude that is 110 this one. And therefore, again we get the same consistent result that is 110 had it been an odd value like 9 or 11.

We would have to choose as the average between those two values. For example, if it was 9 then the value would have been 4.5 and we would have taken the average of fourth and fifth value. But in this case, we have a value of fourth position which is 110, the interpretation of this value is quite simple. There are 50 percent values which are either less than or equal to this value or 50 percent value which are more than or equal to this value.

Notice what happens if we increase one of the values to very high levels as we did previously. So, even if one of the value is 1000, the median is not affected it is still 110. And therefore, median is not affected by the presence of extreme outliers unlike the mean value. In this video, we will discuss the excel implementation of another important measure of central tendency that is mode.

Mode is a measure that shows the value that is most frequent in a given data set or sample. A very simple way to compute mode for a given data set or sample is to use mode function and then choose the set of values. In this case since as we can see 100 is occurring two times which is the maximum number of times, the mode of the data is 100 that means 100 is the most occurring value.

Please notice, even if you increase or decrease or there are some extreme outliers present in the data, mode is not affected. For example, look at this in the second set of data we have an extreme value as we have seen earlier but the mode of the data will not change, it will still remain 100 which is the most frequent value. Another way to compute this mode is to see how frequent a value is appearing.

So, we will compute the frequency of a value in this data set and the simple function to perform that is countif. In countif function, we need to select the data, we will fix this range and then we need to specify the value. Now, we will just drag it and notice the value 100 is appearing two times and therefore it remains the mode of the data.

In this video we will see the excel implementation of range computation.The measure of range for a data is the difference between maximum and minimum value in that data. It can be easily computed using this max function to obtain the maximum value in the data, minus min function to obtain the minimum value in the data. In this data the maximum value is 200 and minimum value is 80 so, we will get the range as 120. Notice the other data set where a maximum value is 1000 and minimum value is 80.

The range is expected to be quite different and as we can see it is 920 which was expected and therefore the presence of extreme outliers have increased the range of this data. In this video, we will discuss the excel implementation of quartiles, as we already know that quartiles distribute the data into four intervals or four groups. To start the computation of quartile we need to see how many observations are there?

For example, in our calorie serial data we have total seven observations so, we will apply our n + 1 by four formula. So, in order to get the value of Q1, we need to compute n + 1 upon 4, since this is a whole number or integer, we will select the second value which is 100 in this case. So, the answer in this part is quite simple Q1 = 100 which is the second position then we need Q2, Q2 is essentially the median of the data itself which is quite easy to compute.

We will compute the median which is 110. In order to compute Q3, we will use that formula which is provided in the rule number which is three times n which is 7 here + 1 divided by 4 so, we get a value of 6. And therefore, the sixth position is of held by 190 value so, we will use this 190. The interpretation of this 190 value is that there are 75 percent observations that are equal to or less than this 190 and there are 25 percent observations that are equal to or more than 190.

Similarly, the interpretation of Q1 here is that there are 25 percent value that are equal to or less than 100 or 75 percent value that are equal to or more than 100. Similarly for Q2 which is 110 the interpretation is that there are 50 percent values that are equal to or less than 110 and similarly 50 percent value that are equal to or more than 110. Our last and very important measure of central tendency is interquartile range.

And as we remember from our discussion in the previous videos interquartile range is the value that is the difference between Q3 quartile – Q1 quartile. In this case that is 190 – 100. Q3 is

190, Q1 is 100 and therefore our interquartile range that is Q 3 – Q 1 = 90. So, with this we have discussed the excel implementation of key measures of central tendency including mean, median, mode, interquartile ranges and quartiles. **(Video Ends: 25:39)**

(Refer Slide Time: 25:40)

In this video, we will discuss the R implementation of measures of central tendency. **(Video Starts: 25:44)** First, I will load the data that I have copied, I have already copied the data so, I will use this command read.delim, I have copied the data from excel so, the data is available on my clipboard. I will simply use this clipboard functionality and since there is a head in data I will put header is equal to 2 and now we can check the data, what is inside it?

So, this data is cereals data where calorie values corresponding to serials are provided. As a best practice, I would recommend that we set our working directory properly. So, we can set the working directory, choose an appropriate directory and click on ok. You would notice a command has appeared on console window, you can copy paste this command on your script for future purposes.

So, when you are working on the same code line and same data in future you have all the details readily available. For example, if I save this data let us say I save this data with saveRDS command, I save the data variable. I give it a name let us see Data.rds this data will be directly saved into this location and next time if I want to read this I can simply run this command data is equal to readRDS and I will give it the name.

I need not specify the complete address because I have already set the working directory properly. So, my data file is available with me, let us look at this data file so, it is a small data file. So, I can see it carries the calorific values of all the cereals that were available all the seven observations. And now, we will start with our descriptive exercise as a first step, we want to see the mean of this data.

In R, I can compute this very simply by using this mean of data, I will use the dollar operator which will connect me to the variable inside data which is calories and I can compute the mean of calorific values. I can also compute the median using with very simple command Data$Calories, I can do that as well. Next, I want to compute the mode of this data and for that I need to install a package functionality which is using this command install.packages and I will install a package statip.

I have already installed it if you have not you can run this command and it will install the package statip for you. Once the package is installed, you need to add this functionality to your current working library using this library statip, once you run this command, the statip functionality is added to your current working library. This functionality will help us in computing the mode of the data using this mfv most frequent value function.

And I will write Data$Calories so, I can check that 100 which appears the most that is two times in the data is considered as the mode of the data set. Next, I want to compute the quartiles of this value, in R there is a functionality called quantile where you can generate any percentile of the data. So, for example, if I want to generate Q 1 which is 25th percentile, I can simply use Q1 equal to Data$Calories and I want to specify the percentile as 0.25.

So, it is generating the Q 1 first quartile for me so, I can check the value Q 1 which is 100. Similarly, I can generate Q 2 although Q 2 is same as median but using this I can also generate Q 1 with at 50 percentile and I can see the value of Q 1. Similarly, I can generate the value of Q 3 also which is quite easy, I need to only put 75 percentile and I will get the value of Q 3.

Now, in case you are interested in looking at the interquartile range that is quite easy, what you need to do is simply Q 3 – Q 1 and you will get the value of interquartile range. Also, a very important functionality that is provided by R is computation of percentiles. We can easily

compute the percentile corresponding to any data by using this quantile function, I can write Data$Calories and I can choose any percentile.

Let us say I want 10 percentile, I can compute this that is 92, similarly any set of percentile I can compute using this formula. Notice that I am simply changing the values of percentiles and it is providing me the value for that data. **(Video Ends: 30:08)**