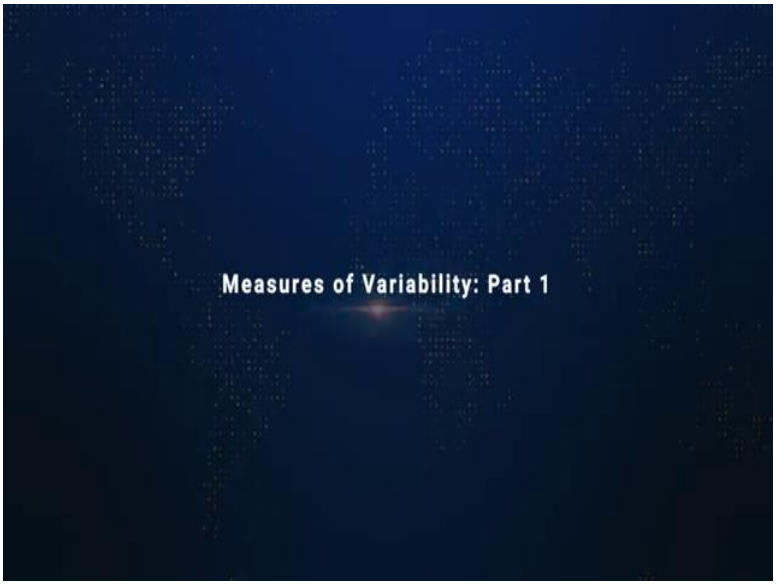
Artificial Intelligence (AI) for Investments
Prof. Abhinava Tripathi
Department of Industrial & Management Engineering
Indian Institute of Technology – Kanpur

Lecture – 16
Measures of Variability: Part 1

(Refer Slide Time: 00:13)



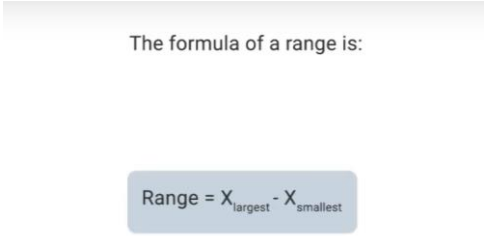
(Video Starts: 00:13) We will discuss about Variability and how to measure it? Measures of central tendency yield information about the centre or middle part of a data set. However, businesses can use another group of analytical tools called measures of variability to describe the spread or the dispersion of a data set. Using measures of variability in conjunction with the measures of central tendency makes possible for a more complete numerical description of the data.

The key measures of variation are range. Range describes the difference between the largest and smallest data point in our data set. Next variance, the variance is the average of square deviation about the arithmetic mean for a set of numbers. Standard deviation, standard deviation measures the dispersion of a data set relative to it is mean. It is defined as the square root of the variance lastly mean absolute deviation.

The mean absolute deviation is the average of the absolute values of the deviations around the mean for a set of numbers. We will start with the measure of range. Range describes the difference between the largest and smallest data point in our data set. The bigger the range, the

more is the split of data and vice versa. A simple measure of variation is the simplest descriptive measure of variation for a numerical variable.

The formula of range is quite simple it is range equal to largest value minus smallest value in the sample.

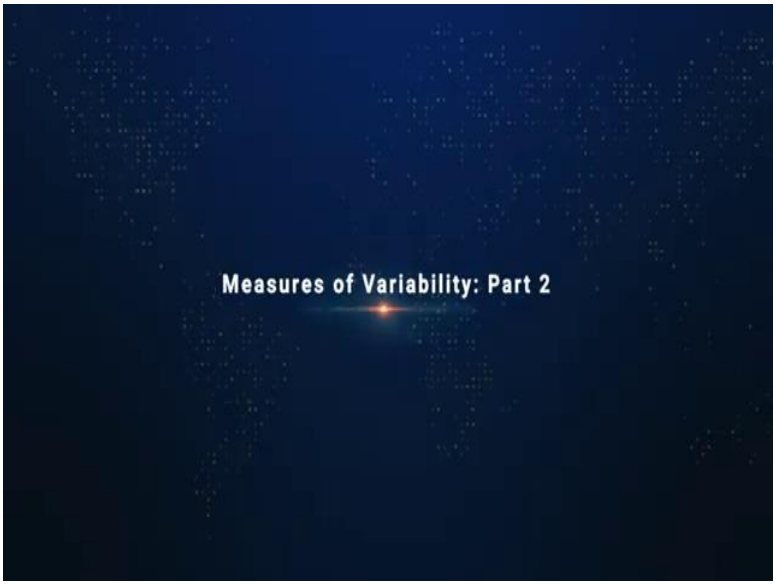


Calculating the range would further analyze the sample of 10 get ready times. We have already seen this data set. To compute the range of this data set first, we will rank the data from smallest to largest, as we have shown in the table here. As per the formula discussed earlier, the maximum value which is 52 minus the minimum value which is 29.

The range is 23 minutes. The range of 23 minutes indicates that the largest difference between any 2 days in time to get ready in the morning is 23 minutes. The range measures the total spread in the set of data. However, the range does not take into account how the values are distributed between the smallest and largest values. In other words, the range does not indicate whether the values are evenly distributed, clustered near the middle or clustered near one or both extremes.

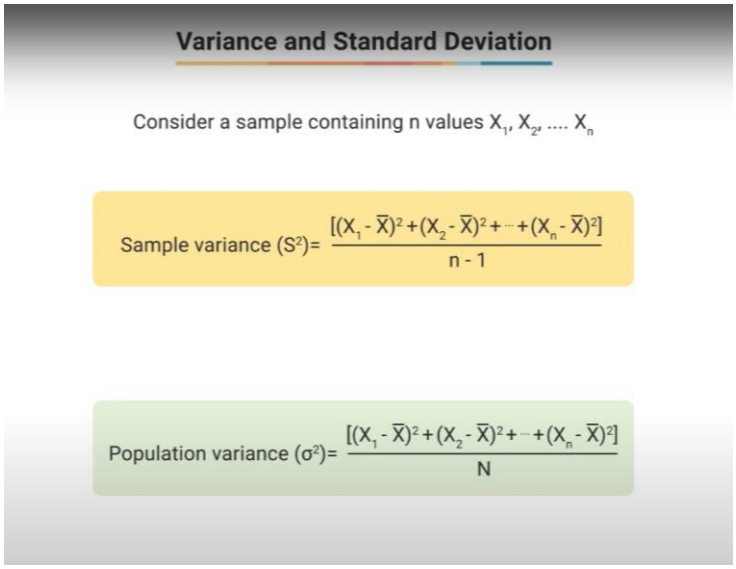
Thus, using the range as a measure of variation when at least one value is an extreme value is misleading. **(Video Ends: 02:35)**

(Refer Slide Time: 02:36)



Now, we will discuss the measure of variance and standard deviation. **(Video Starts: 02:40)** Being a simple measure of variance the range does not consider how the values are distributed or clustered between the extremes. Two commonly used measures of variation that account for how all the values are distributed are the variance and standard deviation. These statistics measure the average scatter around the mean.

How larger values fluctuate above it and how smaller values fluctuate below it. A simple calculation of variance around the mean might take the difference between each value and the mean and then sum these differences. However, some of these differences would always be 0 because the mean is the balance point for every numerical value. Thus, the calculation of variance requires to take square of the difference between each value with its mean and then sum those square differences this sum of the square difference is called the sum of squares forms the basis for calculating the variance and standard deviation as being discussed shortly. For example, if a sample has a sample variance S square then that square is the sum of squares divided by the sample size n – 1. When we are working with population, the denominator n is the population size.



However, we often work with samples and therefore we use sample size – 1 as denominator. The sample standard deviation S is the square root of sample variance S square because the sum of squares can never be negative. The variance and standard deviation will always be non-negative values. And in virtually all cases, the variance and standard deviation will be greater than 0.

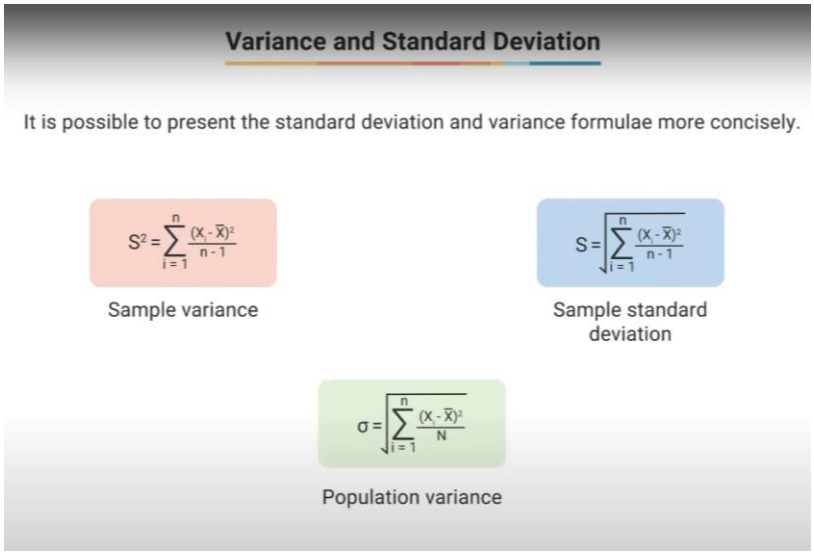
Both the variance standard deviation will be 0, meaning no variation only for the special case in which every value in the sample is same. Consider a sample containing n values, X_1, X_2, X_3 and so on up to X_n. The sample variance S square is simply defined as the formula shown here.

$$s^2 = \frac{[(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2]}{n-1}$$

Please notice that the formula for population variance is different from sample variances for a population containing n values X_1, X_2, X_3, X_n. The population variance sigma square is $\sigma^2 = \frac{[(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2]}{N}$. Where N is the population size. Notice, the difference between sample and population variances.

In the population variance, we have capital N in denominator as compared to n – 1 which is sample size in the context of sample variance. While the mathematical derivation of this is not required here. Please observe that the difference between these two, the divisors that is an n – 1, becomes smaller as the sample size increases and converges to large population size n. The

formula of standard deviation and variance as discussed here can be further provided in a more compact manner.



We can say this $s^2 = \sum_{i=1}^{n} \frac{(x-\bar{x})^2}{n-1}$ or $s = \sqrt{\sum_{i=1}^{n} \frac{(x-\bar{x})^2}{n-1}}$. This is a more compact form of the same formula provided in summation notation. And the formula for population standard deviation that is sigma can also be replicated in a similar manner. Again, here $\bar{X}$ is sample mean, n = sample size, X_i is the ith value of the variable X, $X_i - \bar{X_i}$ is the summation of the square differences.

Note that in both equations, the sum of squares is divided by the sample size – 1 for sample and N for population. For now and our understanding that is the sample estimate converges to population estimate as the sample size increases and converges to n. Let us illustrate the computation of sample for our 10 get ready times data. There are following steps that we will discuss one by one.

Variance and Standard Deviation

Illustrating the computation of the ten data samples on the time taken to get ready

Step 1: Calculate the difference between each value and its mean

Time (X)	Step 1: $(X_i - \bar{X})$	Step 2: $(X_i - \bar{X})^2$
39	-0.60	0.36
29	-10.60	112.36
43	3.40	11.56
52	12.40	153.76
39	-0.60	0.36
44	4.40	19.36
40	0.40	0.16
31	-8.60	73.96
44	4.40	19.36
35	-4.60	21.16
Mean = 40		Sum = 412.40
		Sum Divide by (n - 1) = 45.82

Variance and Standard Deviation

Illustrating the computation of the ten data samples on the time taken to get ready

Step 2: Compute the square differences

Time (X)	Step 1: $(X_i - \bar{X})$	Step 2: $(X_i - \bar{X})^2$
39	-0.60	0.36
29	-10.60	112.36
43	3.40	11.56
52	12.40	153.76
39	-0.60	0.36
44	4.40	19.36
40	0.40	0.16
31	-8.60	73.96
44	4.40	19.36
35	-4.60	21.16
Mean = 40		Sum = 412.40
		Sum Divide by (n - 1) = 45.82

Variance and Standard Deviation

Illustrating the computation of the ten data samples on the time taken to get ready

Step 3: Compute the sum of the square differences

Time (X)	Step 1: $(X_i - \bar{X})$	Step 2: $(X_i - \bar{X})^2$
39	-0.60	0.36
29	-10.60	112.36
43	3.40	11.56
52	12.40	153.76
39	-0.60	0.36
44	4.40	19.36
40	0.40	0.16
31	-8.60	73.96
44	4.40	19.36
35	-4.60	21.16
Mean = 40		Sum = 412.40
		Sum Divide by (n - 1) = 45.82

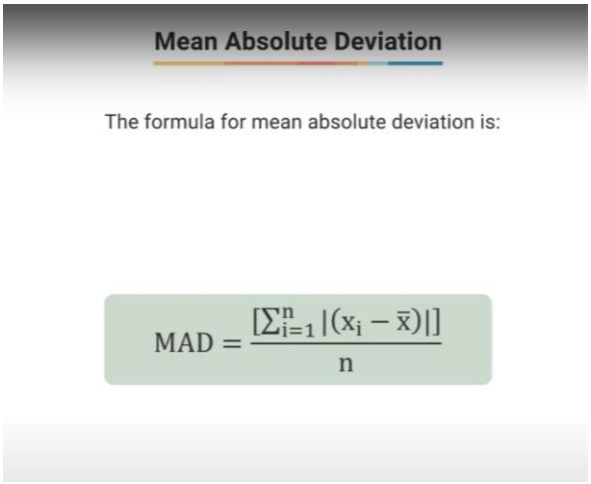
And they are also shown in the slide for our data that is get ready times. As a step 1 we will calculate the difference between each value and it is mean. The mean is computed as 40 and the differences with each value are shown in second column as step 1. In the next step, step 2 will compute the square differences that are shown in the third column as a step 2. And we compute the square differences for each observation. In the third step, we compute the sum of the square differences the value is 412.4.

And then we divide that by n – 1 that is 45.82 becomes our sample variance. The square root of this number which is 6.77 that is the standard deviation of the sample. The same steps can be replicated as shown in this slide to compute the variance and standard deviation of the population. Notice that everything remains the same, except that the denominator that is being used here. The denominator is 10 instead of 9. **(Video Ends: 07:37)**

(Refer Slide Time: 07:38)

Now, we will discuss the another measure of variability. That is mean absolute deviation. **(Video Starts: 07:42)** The mean absolute deviation of a data set is the average distance between each data point and the mean. It gives us an idea about the variability in a data set. The steps to calculate the mean absolute deviation are provided here. First step is calculate the mean.

Second step is calculate how far each mean data point is from the mean using positive distances. These are called absolute deviations. Then these deviations are added in the third step. And in the fourth step, the sum is divided by the number of data points. This is also summarized in the form of formula, $MAD = \frac{\left[\sum_{i=1}^{n} |x_i - \bar{x}|\right]}{n}$. And notice that we are taking the absolute value of these differences.



Let us try to compute these values for our get ready times data set with 10 observations. The MAD measure for this data can be simply computed in those steps. In the first step, we will compute the mean that is for T that is S1. Next step will compute the mean differences of each observations that is $X_i - \bar{X_i}$ this is our second step S2. And the third step will sum up all these values. The sum is equal to 50.

And then in the fourth and last step we will divide this sum by the number of observations that is done. So, we will get our MAD value of 5. Till now we have discussed the measures of variability that is range variance, standard deviation and mean absolute deviation, MAD. In the next set of videos, we will discuss different measures of shape. **(Video Ends: 09:09)**

(Refer Slide Time: 09:10)

We will start with excel implementation of measures of variability. **(Video Starts: 09:14)** First, we will discuss the measure of range which is the difference between maximum and minimum value in a given sample. So, the computation is quite easy. First, we will compute the maximum value which is 200. Then minimum value which is 80. Now, the difference between these two maximum and minimum values is equal to the range of the data.

In this case it is 120. Similarly, if you look at another data which is slightly modified version of this calorie data. Where the extreme value is replaced by 1000 instead of 200, we can compute these measures and let us have a look at these measures. Now, the maximum value is 1000 and therefore the range has been extended considerably to 920. And therefore, if there are extreme outliers in the data, the range measure can increase a lot.

Next, we will discuss the excel implementation of standard deviation and variance measure. We start by looking at the standard deviation and variance of a population. And then we will compute the standard deviation and variance of the sample. So, as a first step, let us compute the standard deviation and variance considering this calorie data as population. So, for that we need to have the mean of this data first.

So, we will compute the mean or average of this data. Next, we will compute the difference of each observation from the mean. Now, we need to square these values in the next step, as we have already discussed in the previous videos. Then we will compute sum of squares. Whether we are computing, the standard deviation or variance of population or sample will depend upon this step where we are dividing the sum of squares by the number of values.

So, if I divide the sum of square by number of observations that is count, there are seven observations, this will become population variance. So, this is my population variance. Instead of this if I divide this number, not by the overall count or size but size – 1 then they will become sample variance. Notice because of change in denominator the value has increased. The standard deviation is nothing but the square root of this value.

So, it is easy to calculate we can just write this simple formula sqrt to get the population variance. And in a very similar manner I can get this standard deviation of population. And in a very similar manner I can get the standard deviation of sample by taking the square root of this sample variance. We can observe both of these values. There is also a very simple way to compute these values in excel.

So, if I use this formula VAR dot P that will give me variance of the population but it will be a more direct way and you can see these numbers are matching with those calculated by us. Also, I can compute the variance of sample using this VAR dot S. Now, the sample version of this variance formula will apply notice it is same. And again, for the standard deviation also, I can use the formula for population. Alternatively, I can use the formula of sample.

Generally, in real life we do not have populations, we often work with samples and therefore the formula that is the one with sample is more applicable in real practical situations. Next, we will discuss the excel implementation of another important measure of variance which is mean absolute deviation. So, we will compute mean absolute deviation. The computation is quite easy. First, we need to compute the mean of the data which is average formula.

Now, we will compute the absolute value of mean differences which is ABS deviation. So, these are my absolute deviations from the mean we will compute all these deviations. These will be positive because we are computing, absolute values and then we will sum them up to compute the absolute deviation. This is the summation of all the deviations. And now we need to average it out by dividing it with number of values in the sample. So, we get a value of 37. So, this is mean absolute deviation of the sample. **(Video Ends: 14:33)**

(Refer Slide Time: 14:34)

In this video, we will discuss the R implementation of measures of variability. **(Video Starts: 14:40)** We will continue with our existing data which is calorie data from cereals. A brief description of this data is provided on the console when I run this data with my control enter command notice the data is populated on the console window. It starts from 80 and up till 200. As a first step we need to compute the range of this data.

That is our first variability measure. Now, I can check the range of this data by simply running range command which will give me the lowest and highest values in the data that are 80 and 200. If I want to see the range, I can simply compute the maximum value of this data and subtract the minimum value of data from this to get the range which is 120. So, this was our range computation. This was our range measure.

Next, we will try and compute the variance and standard deviation measures of from the data. As a first step, I will show you how to compute the variance of the data, simply var command data and calories. This will give me the variance of this data which is 2200. But please notice this is sample variance that is sample var. So, this is my sample variance. In order to compute population variance from this data I need to multiply this data from its denominator that is –1 and divide it by the length of the data.

This will result in the population variance. I can assign this value to a new variable as well that is population bar or population variance alternatively and store it in this new variable notice. If I want to extract this value from population var, I need to simply just select this population var

variable and press control enter on this to get this value which is 1885.7. All these values are appearing in my console window.

Similarly, I can compute the standard deviation which is quite easy. I can use this SD command standard deviation and data dollar calories. This is the standard deviation of the data. I can assign this to a new variable as sample SD which will be the sample standard deviation of this value. However, in order to compute the population standard deviation as we did earlier.

I need to multiply this with the square root of the factor which is length of the variable –1 divided by length of the variable which is just for the population part. So, if I run this, I will get the sample standard deviation which is 43.42 which is different from 46. I can assign this to a new variable which is population standard deviation. If I want to extract this variable, I can simply take it out and run ctrl enter to get the value of population standard deviation.

So, this is my variance and standard deviation measures for population and the sample. The last measure that we are going to compute is our mean absolute deviation measure or you can say MAD measure. As part of MAD measure, it is quite easy. First, I will do the manual way or a slightly more complicated. So, what we will do is we will take our data subtract the mean of the data.

Mean of this data is as we remember, it is 130. So, we will subtract the mean from all observations and we will compute the absolute value of these which are populated on the console window. Now, what we need to do is we need to sum up these absolute values and then divide them by the length of the sample that is data dot calories, dollar calories and we will get the MAD value.

Now, this was slightly lengthier and more complicated way. A more easier way is to u see matrix package. So first, if you have not installed, you need to run this command matrix. Since, I have already installed matrix package in my current working library for installation of library and installing in their current working environment. That is already discussed the previous videos. So, we will not explain that in detail here.

So, we will simply put the library matrix in our current working environment. That is quite easy. Once I have put it now, I can use the functionality. Then a very important functionality

is this mae command. If you want to know more about this mae command, you can simply type ?mae. And notice this mae gives you the absolute difference between vectors. So, what I can do is I can simply write this mae.

Then I will use my vector of data which is data dollar calories. And then I can subtract it from the mean value which I have already computed. And this will be my mean of absolute mean absolute error or mean absolute deviation here in this particular case which is just check here small error. So, I am just running this instead of minus I need to put comma so now, I will get the same value. So, this is my mean absolute deviation.

So, this is how you compute the mean absolute deviation directly using the R functionality from matrix package. **(Video Ends: 19:54)**

(Refer Slide Time: 19:55)

In this video, we will discuss about measures of shape. **(Video Starts: 19:58)** The measures of shape is tool that can be used to describe the shape of a distribution of data. In this video we will examine two measures of shape that is skewness and kurtosis. Skewness refers to a distortion or asymmetry that deviates from the symmetrical nature of data around it is mean. While kurtosis measures the peakness of the curvature of data around the distribution.

The distribution of data in which the right half is the mirror image of the left half is said to be symmetrical. One example of symmetrical distribution is the normal distribution or bell curve. Here, skewness is a property where distribution is asymmetric or lacks symmetry around its

central measure like mean median or something like that. Skewness measures, the extent to which the data values are not symmetrical around the mean.

There are three possibilities either mean is less than median. This is called negative or left skewed distribution, as shown in panel A. Then you have mean equal to median which is symmetrical distribution 0 skewness, as shown in panel B. And then you have mean greater than median, where you have positive skewness that is right skew distribution, as shown in panel C.

In a symmetrical distribution like panel B, the values below the mean are distributed in exactly the same way as the values above the mean and the skewness is 0. In skewed distribution there is an imbalance of data values below and above the mean. And the skewness is a non-zero value that is less than 0 for left skewed distribution or greater than 0 for a right skewed distribution. Panel A displays a less skewed distribution.

In a left skewed distribution most of the values are in the upper portion of the distribution. Some extreme small values cause the long tail and distortion to the left and cause the mean to be less than the medial. Because this skewness statistic for such a distribution will be less than 0 some use the term negative skewed to describe this distribution. Panel B displays a symmetrical distribution.

In a symmetrical distribution values are equally distributed in the upper and lower portions of the distribution. This equality causes the portion of the curve below the mean to be the mirror image of the portion of the curve that is above the mean and makes the mean equal to the median. Panel C displays a right skew distribution. In a right skewed distribution most of the values are in the lower portion of the distribution.

Some extremely large values cause the long tail and distortion to the right and cause the mean to be greater than the median. Because the skewness statistics for such a distribution will be greater than 0 some use the term positive skew to describe this distribution. Next, we move to the kurtosis measure. Kurtosis measure the peakness of the curve of the distribution. That is how sharply the curve rises approaching the centre of the distribution.

Kurtosis compare the shape of the peak of a bell-shaped normal distribution that is bell shaped normal distribution which by definition has a kurtosis of 0. A distribution that has a sharper rising centre peak than the peak of a normal distribution has a positive kurtosis. A kurtosis value that is greater than 0 and is called leptokurtic. A distribution that has a slower rising that is flatter centre peak than the peak of a normal distribution has a negative kurtosis.

A kurtosis value that is less than 0 and is called platykurtic. A leptokurtic distribution has a higher concentration of values near the mean of the distribution compared to a normal distribution. While a platykurtic distribution has a lower concentration compared to a normal distribution. In affecting the shape of the central peak, the relative concentration of values, near the mean also affect the ends or tails of the curve of distribution.

A leptokurtic distribution has fatter tails many more values in tails than a normal distribution has, when analysis mistakenly assumes that a set of data forms a normal distribution. That analysis will underestimate the occurrence of extreme values if the data actually forms a leptokurtic distribution. **(Video Ends: 24:04)** Some suggest that such a mistake can explain the unanticipated, reverses and collapses that financial markets have experienced in the recent past.

We will discuss the implication of these shape measures and kurtosis skewness with R and excel.

(Refer Slide Time: 24:18)

We will discuss the R and excel implementation of our measures of shape. **(Video Starts: 24:21)** First, let us discuss the implementation of measures of shape that is skewness and kurtosis in excel using our get ready data. Where we have times observable for getting ready in the morning. To compute this skewness, we will simply use this skew function, functionality in R.

And the skewness of this data is 0.085 which seems that it is possibly a right skew data. For a symmetric distribution this skewness a should be equal to 0. We can also generate the kurtosis of this data. And the kurtosis of this data is 0.1. Please note that this is a relative kurtosis measure. That means excel functionality, computes the kurtosis related to normal distribution.

Normal distribution has a kurtosis of 3, so there we have an excess, positive kurtosis with this time to get ready data that is 0.13. So, absolute kurtosis would be 3.13 that means it has a positive kurtosis excess positive kurtosis. Now, we will discuss the implementation of shape measures with R will carry on with R calorie data set with cereals. So, this was the data set. We can see that data set in our console where I have populated it.

First, we need to install some of the packages relevant packages. These are tseries and moments. Since, I have already installed these packages, I will not run these commands I will simply put them into my current working environment which is first for tseries. And then for moments now, I have installed for both of these libraries both of these packages tseries and moments in my current working library.

So, I will start by computing the skewness of my data which is calorie data. Notice, the skewness, the skewness is turning to be 0.65 which is a positive skew as compared to normal distribution. A symmetric distribution like a normal distribution which has a 0 skew value. Similarly, I can compute the kurtosis for this data. Notice, the kurtosis for this data is 1.807 which is less than three that is normal distribution or a distribution like normal distribution has a kurtosis of three.

So, it seems it has a peak which is lower than a normal distribution. Now, if you want to know exactly how it works for normal distribution will generate a normal distribution like this. Generally, we use our rnorm function if you want to know more about this rnorm function. You can simply put question mark and see all the there is to know about this rnorm,

it will be populated in the help window.

We can check that what are the syntax and parameters to be fed. So, first, let us generate this normal distribution. And that is rnorm will quite probably 10,000 values and the mean is 0 and standard division is 1. So, rnorm distribution is there. Now, we will compute the skewness of this distribution. Notice that this value of distribution is not exactly 0 but it is very close to 0.

It is linked to the number of values we provide. For example, if I decrease the sample size, I put it to 1000 you notice the skewness is slightly more in magnitude. Although, it is positive but it is more in magnitude. If I increase the number of observations in the sample, the skewness will go down in terms of magnitude. Sign does not matter much here because the magnitude is quite low.

Similarly, if I am talking about kurtosis and I am using reasonably large sample then the kurtosis will be closer to a normal distribution which is close to 3. If I decrease the sample size notice, the kurtosis value will also not be as close slightly different from 3. And as I increase the observations may be very large number of observation. Then in fact, you notice that very close to.

So, the kurtosis of a normal distribution is 3. And of course, the kurtosis and skewness of our data which is calorie data is not same as normal distribution. That means it is not exactly symmetric. Because of skewness measure and not exactly same peakness as normal distribution. **(Video Ends: 28:31)**

(Refer Slide Time: 28:32)

(Video Starts: 28:33) In this lesson, we discuss the measures of central tendency, shape and variation with the help of examples. Measures of central tendency include mean median mode, quartiles and interquartile range. Here the mean is the sum of observations divided by total number of observations Median divides the data set into two halves. Mode has the highest frequency in the given data set. Quartiles is divided data set into four parts.

Measures of variation include range standard deviation, variance and mean absolute deviation. That is MAD measure. Range describes the difference between the largest and smallest data point in the data set. Variance is the average of square deviations from the mean. Standard deviation is the square root of the variance. MAD measure is the average of absolute deviations from the mean.

Measures of shape included skewness and kurtosis. Skewness indicates the degree of symmetry of the data around its centre point. Kurtosis represents the peak and tail behavior of the data. Each of these measures contributed to understanding of the data set in different, yet important ways. Before beginning with any kind of advanced data analysis one should carefully examine the data through descriptive measures. **(Video Ends: 29:38)**