

**Six Sigma**  
**Prof. Jitesh J Thakkar**  
**Department of Industrial and Systems Engineering**  
**Indian Institute of Technology, Kharagpur**

**Lecture – 22**  
**Fundamentals of Statistics**

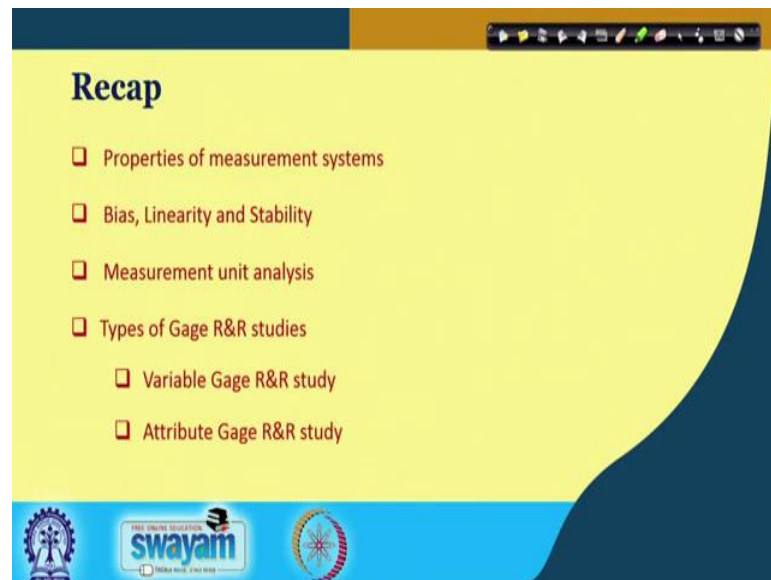
Hello friends, I once again welcome you to the journey of Six Sigma and at present I would like to remind you that we are in the measure phase of DMAIC cycle and discussing various topics and issues specific to measure phase. So, today say as a part of lecture 22, we will see the Fundamentals of Statistics and try to appreciate the importance of statistics and some of the important issues.

(Refer Slide Time: 00:52)



So, let us begin with a very good quote. It is the mark of a truly intelligent person to be moved by statistics and the quote is given by none other than the great scientist researcher George Bernard Shaw. So, this quote says that if you believe in scientific decision making than you cannot avoid the use of statistics and exactly six sigma also emphasizes on the use of scientific tools and techniques.

(Refer Slide Time: 01:30)



So, if you see the recap then we have talked about properties of measurement system, bias linearity and stability, measurement unit analysis, types of gage R & R study, variable type, attribute type. We focused mainly on variable type gage R & R and in that also we had three options R range, gage R & R,  $\bar{x}$  bar, and R mean, and average and ANOVA. So, we have studied in detail mean and average variable gage R & R study.

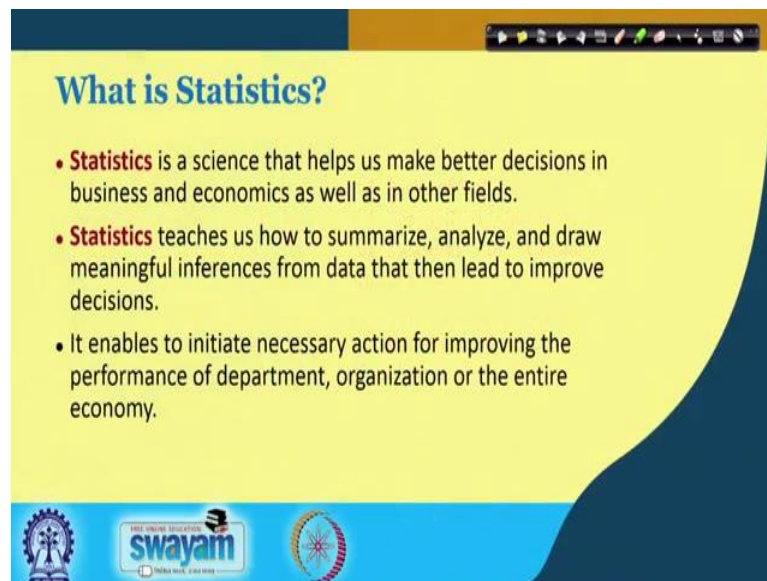
(Refer Slide Time: 02:05)



Now, as a part of this lecture, we would like to see some of the basics of statistics, what is the difference between descriptive and inferential statistics, measures of central

tendency and dispersion, shape of the distribution, numerical descriptive measure for a population. The most important theorem in the statistics please remember is the central limit theorem. So, we will see this central limit theorem and then the random variable; the concept itself is the base for all the statistical analysis. And we will try to appreciate the definition concept of random variable.

(Refer Slide Time: 02:52)



### What is Statistics?

- **Statistics** is a science that helps us make better decisions in business and economics as well as in other fields.
- **Statistics** teaches us how to summarize, analyze, and draw meaningful inferences from data that then lead to improve decisions.
- It enables to initiate necessary action for improving the performance of department, organization or the entire economy.

At the bottom of the slide, there are logos for 'swayam' and other educational institutions.

So, what is statistics? So, statistics typically it is a branch, a science of mathematics and it basically enables the organizations, departments may be entire economy to take the sound data based decisions and helps the people to appreciate, accept the decisions which are based on scientific analysis.

(Refer Slide Time: 03:23)

**Descriptive v/s Inferential Statistics**

• Descriptive Statistics	• Inferential Statistics
✓ Collect	✓ Predict and forecast values of population parameters
✓ Organize	✓ Test hypotheses about values of population parameters
✓ Summarize	✓ Make decisions
✓ Display	
✓ Analyze	

The slide features a yellow background with a dark blue header and footer. The title is in blue. The content is organized into two columns. The footer includes the Swayam logo and the text 'FREE ONLINE EDUCATION swayam'.

Now, if you see the overall domain of statistics then it is divided into descriptive statistics and inferential statistics. If you look at the descriptive statistics you have collect organize, summarize, display, analyze and if I look at the inferential statistics this we will subsequently see then, I would like to predict and forecast values of the population parameter. I would like to taste the hypothesis about my particular assumption of a phenomena and then I would like to draw the inferences based on my statistical analysis.

(Refer Slide Time: 04:07)

**Methods for Descriptive Statistics**

- ✓ **Measures of Central Tendency:** Mean, Median, Mode
- ✓ **Measures of Dispersion:** Range, Quartiles of data set, Variance, Standard Deviation
- ✓ **Shape of the Distribution:** Skewness and Kurtosis
- ✓ **Graphical or Tabular format:** Histograms, Stem-and-leaf display, Contingency tables, scatterplots

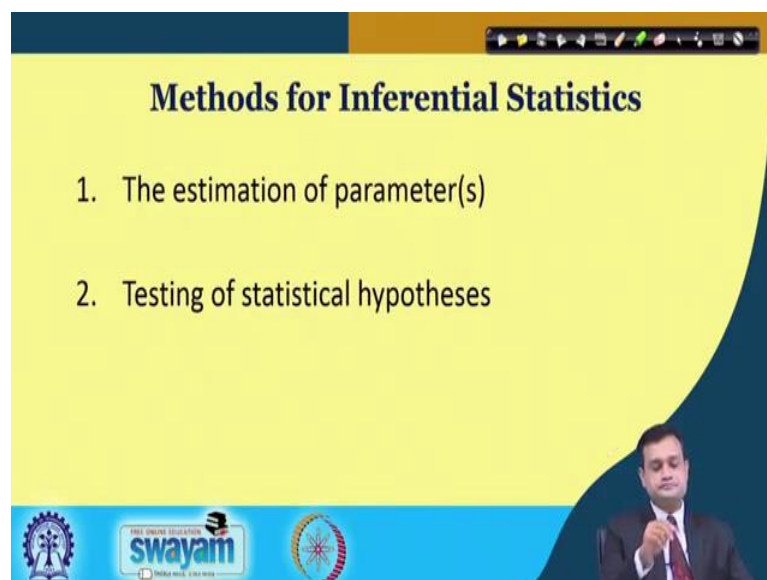
Summarize our group of data using a combination of tabulated description (i.e., tables), graphical description (i.e., graphs and charts) and statistical commentary (i.e., a discussion of the results).

The slide features a yellow background with a dark blue header and footer. The title is in blue. The content is organized into a single column. The footer includes the Swayam logo and the text 'FREE ONLINE EDUCATION swayam'.

So, there are various methods of descriptive statistics typically we classify this into say finding measures of central tendency mean, median and mode. Measures of dispersion you have range, interquartile say range or data set, variance, standard deviation. You have shape of the distribution because usually many of the assumptions in statistical analysis, we try to go with the normal distribution. But, here say it is important to know that my data is coming from which kind of distribution and even if it is normal what is the shape.

So, we try to say check the shape of the distribution and the concepts are Skewness and Kurtosis and you can have graphical or tabular format histograms, stem and leaf diagram or display contingency table, scatter plots and so on. So, basically we have four different ways to describe my data and take the decision based on the descriptive statistics. Here you can see in the box that you can summarize the data using different ways and means you may put it in table or you may have the graphical description or you can have a statistical commentary. And, there are various ways by which the data can be structured and represented in a meaningful way that can help the manager decision maker to see the phenomena in depth and understand the overall perspective of the problem under investigation and take the decision.

(Refer Slide Time: 05:59)



**Methods for Inferential Statistics**

1. The estimation of parameter(s)
2. Testing of statistical hypotheses

swayam  
MOE

So, when we say look at the inferential statistics you have estimation of the parameters and testing of the hypothesis. These are the two broader domains of inferential statistics and this part in continuation we will study in the coming lectures.

(Refer Slide Time: 06:22)

**Statistical Measures**

Measures of Central Tendency	Measures of Variability
✓ Median	✓ Range
✓ Mode	✓ Interquartile range
✓ Mean	✓ Variance
	✓ Standard Deviation

**Shape of the Distribution**

- ✓ Skewness
- ✓ Kurtosis

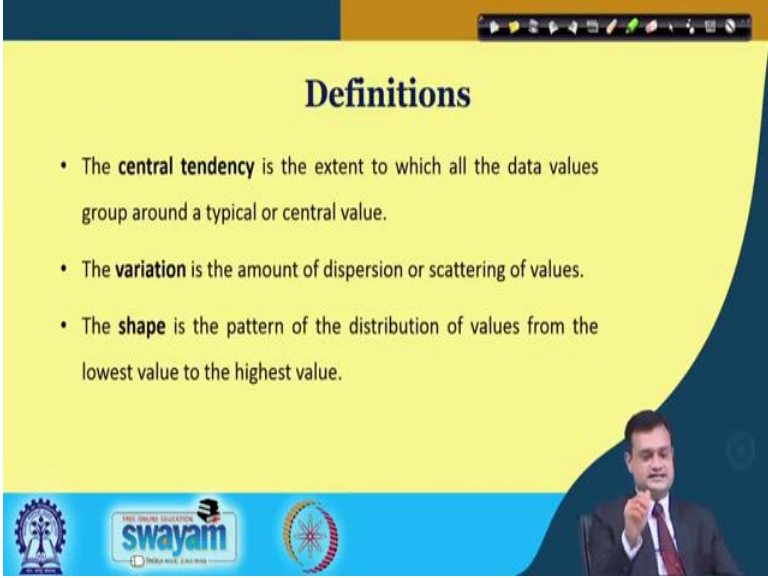
swayam

So, now say various statistical measures as I mentioned, we can just summarize like this there is measures of central tendency which basically targets to say check the mean central tendency of the process. And, this you can do with median; mean; median; mode and you have measures of variability. So, the bear minimum or very say preliminary measure is the range finding the difference between maximum and minimum. You can extend it to interquartile range that will divide your data into quartiles.

You can have variance and when you take the square root of variance you have the standard deviation, you can use Skewness and Kurtosis as the shape of the distribution. So, we will try to basically focus on descriptive statistics as a part of this lecture. And, study all these different measures for presenting my data in a more meaning full manner and analyzing it with the help of some of the measures used in descriptive statistics.







(Refer Slide Time: 07:32)



### Definitions

- The **central tendency** is the extent to which all the data values group around a typical or central value.
- The **variation** is the amount of dispersion or scattering of values.
- The **shape** is the pattern of the distribution of values from the lowest value to the highest value.



So, let us see the definition. Central tendency when I say is the extent to which all the data value they are grouped around a typical central value. So, suppose I have I am measuring the diameter of the shaft and suppose let us say my mean value central value is 10. Then I am taking number of readings; obviously, all reading will not be 10 because there would be certain process variation. So, let us say I am getting 10.1, 10.3, 10.4, 10.2, 10.25 and so on; I would like to see that to what extent my data is close to the central value of interest that is 10 in this case.

So, the variation second term is the amount of dispersion or scattering of the values. So, same example you can see that you have minimum value may be 9.8 and you have maximum value may be 10.4. So, your data range is typically between 9.8 to 10.4 and you would be interested to see that what is the variation I am getting because of this process and this is specific to variability. Then the shape, typically it is pattern of the distribution of values from the lowest value to the highest value. So, we will see all this with the example.

(Refer Slide Time: 09:05)

### Arithmetic Mean or Average

The **mean** of a set of observations is their average - the sum of the observed values divided by the number of observations.

Population Mean	Sample Mean
$\mu = \frac{\sum_{i=1}^N x_i}{N}$	$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$

Logos: UGC, swayam, and a circular emblem. A small video inset of a man in a suit is in the bottom right corner.

So, you have arithmetic mean or average and here we have already gone through the concept of population and sample. You draw the sample from the population to make certain conclusions, inferences based on say measures of central tendency, dispersion or inferential statistics. As it is not possible for me to focus on entire population, I go by the sample and related statistics.


So, if you have a mean usual practice is that you represent your mean as  $\mu$  and it is basically the summation of  $\sum x_i$ ,  $i$  is equal to 1 to  $N$ . And, you divide it by the total number that is capital  $N$  and if you have sample mean then you have  $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$ , small  $n$  is your sample size.



(Refer Slide Time: 10:08)

### Example: Mean

Sales
10
7
12
11
13
15
16
14
14
16
17
16
24
21
22
18
19
18
20
17
320

$$\bar{x} = \frac{\sum x}{n} = \frac{320}{20} = 16$$


Now, just see the example very simple; I have just put some sales data and I want to say I want to see that what is the average sales mean sale that is taking place in the last couple of months. So, I have just summed up my sales data. The sum is 320, I divide 320 by 20 that is the total number of data in a sample and I get average  $\bar{x}$  that is 16.


(Refer Slide Time: 10:36)

### Example: Median

Sales	Sorted Sales
9	7
7	9
12	10
10	12
13	13
15	14
16	14
14	15
14	16
16	16
17	16
16	17
27	17
21	18
23	18
18	19
19	20
18	21
20	23
17	27

← Median

- The **median** is the middle value of data sorted in order of magnitude.
- It is the 50<sup>th</sup> percentile.

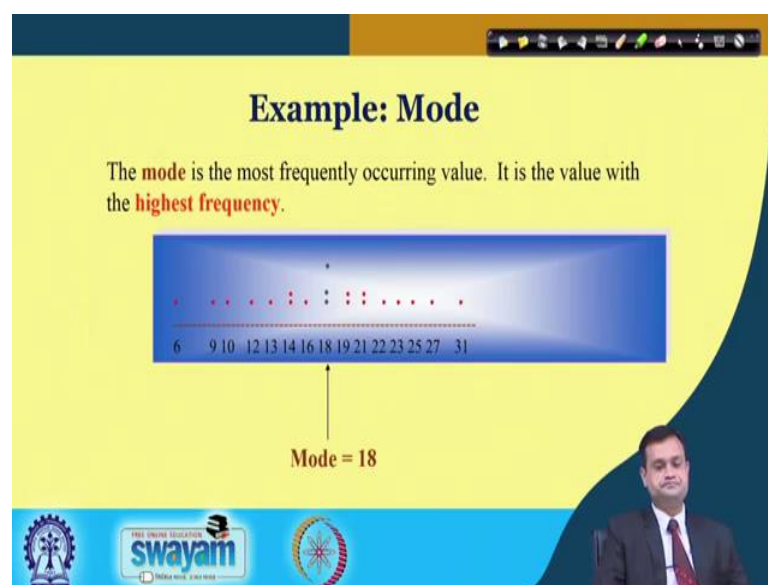


Now, the second important say measure of central tendency is median. I am using again the sales data and I am just sorting it out in the ascending order. So, in the initial first

column you have the data which is not sorted. In the second column I am just putting it in the ascending order lower value first and then the higher values.

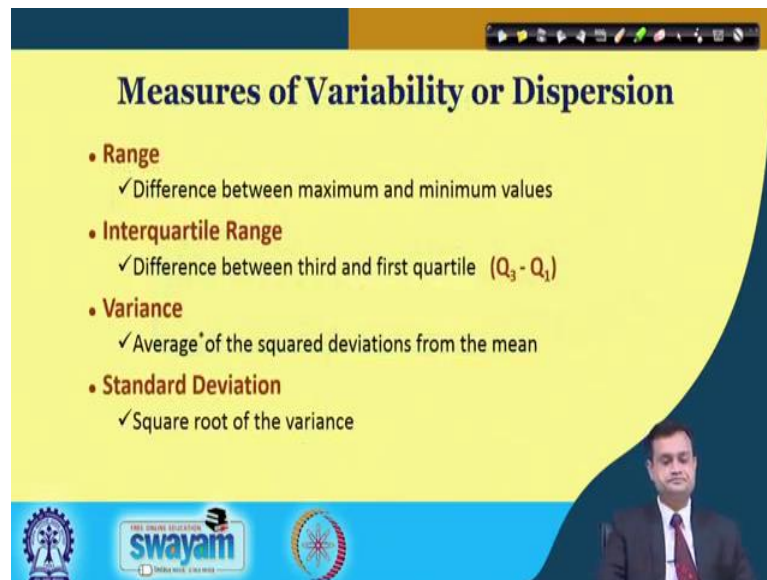
So, now median is basically the middle value of the data sorted in order of magnitude. So, higher magnitude I have sorted and basically it is the 50th percentile. So, here if you see then you have 20 data point exactly 10 10 can be divided; I am not getting exactly one say middle value. So, I can take the average of two middle values. So, in this case it is 16 plus 16 by 2. So, I will have median equal to 16.

(Refer Slide Time: 11:39)



Now, the third measure of central tendency is mode and mode is basically the highest frequency value. So, the particular value which is occurring maximum number of time highest frequency in the data set I will consider that as the mode of my data. So, if you see the example here then I have just put my data in the ascending order ranging from 6 to 31 and, I am just putting the frequency in the form of a dot that how many times particular number you can you can say that these numbers are the sales or may be the demand. So, I am counting the number how many times it is occurring in my data set and I can see that 18 is occurring maximum number of time. So, mode of this data is 18.

(Refer Slide Time: 12:37)



### Measures of Variability or Dispersion

- **Range**
  - ✓ Difference between maximum and minimum values
- **Interquartile Range**
  - ✓ Difference between third and first quartile ( $Q_3 - Q_1$ )
- **Variance**
  - ✓ Average of the squared deviations from the mean
- **Standard Deviation**
  - ✓ Square root of the variance

swayam

Digital India

Now, the other way to look at the descriptive statistics is to check for the variability. And variability measures basically include range, that is the difference between maximum and minimum value. Interquartile range so, you divide your data set into quartiles and take  $Q_3$  minus  $Q_1$ ; we will see the example that would show your interquartile range.

Variance; average of the standard deviation from the mean. So, how much my data values are varying with respect to mean that is my variance and simply when I take the square root of the variance it becomes my standard deviation.

(Refer Slide Time: 13:24)

### Range

Sales	Sorted Sales	Rank
9	7	1
7	9	2
12	10	3
10	12	4
13	13	5
15	14	6
16	14	7
14	15	8
14	16	9
16	16	10
17	16	11
16	17	12
28	17	13
21	18	14
22	18	15
18	19	16
19	20	17
18	21	18
20	22	19
17	28	20

Minimum

Maximum

**Maximum - Minimum = 28 - 7 = 21**




Just look at the data set and once again I am going with the first column sales, second column sorted data. The data is put in the ascending order and you can say that 7 is the minimum and 28 is the maximum. So, the difference is 28 minus 7 that is 21.

(Refer Slide Time: 13:45)

### Interquartile range (IQR)

- IQR is a measure of variability, based on dividing a data set into quartiles.
- Quartiles divide a rank-ordered data set into four equal parts.
- The values that divide each part are called the first, second, and third quartiles; and they are denoted by Q1, Q2, and Q3, respectively.

- ☐ Q1 is the "middle" value in the first half of the rank-ordered data set.
- ☐ Q2 is the median value in the set.
- ☐ Q3 is the "middle" value in the *second* half of the rank-ordered data set.
- ☐ The interquartile range is equal to Q3 minus Q1.



Interquartile range here what you need to do. So, IQR typically is a measure of variability again based on the dividing a dataset into quartiles. So, typically you consider 25 percent, 25 percent, 25 percent and 25 percent. So, I can divide my entire data set into

four quartiles four parts and typically what you say that each part is called first second and third quartile.

So, typically I will denote this as  $Q_1$ ,  $Q_2$  and  $Q_3$  and  $Q_1$  is the middle value in the first half of the rank order data set. We will see the example so, it would be better clear.  $Q_2$  is the median value in the set and  $Q_3$  is the middle value in the second half of the ranked data set. And when you take  $Q_3$  minus  $Q_1$  you get the interquartile range.

(Refer Slide Time: 14:48)

**Example: IQR**

- Consider the following numbers: 1, 3, 4, 5, 5, 6, 7, 11.
- $Q_1$  is the middle value in the first half of the data set. Since there are an even number of data points in the first half of the data set, the middle value is the average of the two middle values; that is,  $Q_1 = (3 + 4)/2$  or  **$Q_1 = 3.5$**
- $Q_3$  is the middle value in the second half of the data set. Again, since the second half of the data set has an even number of observations, the middle value is the average of the two middle values; that is,  **$Q_3 = (6 + 7)/2$  or  $Q_3 = 6.5$**
- The interquartile range is  $Q_3$  minus  $Q_1$ , so  **$IQR = 6.5 - 3.5 = 3$**

So, just see this example I have the numbers 1, 3, 4, 5, 5, 6, 7 and 11. And what I can do I can first find the  $Q_1$  which is the middle value; in the first half of the data set. So, there are total 1, 2, 3, 4, 5, 6, 7, 8. So, you have 8 data in this particular data set and I would try to find  $Q_1$  which is the middle value. So, if I just partition it into four data four data I am not getting the exact middle value. If it would have been the odd number 5, 5 then I will get 3 as the middle value. But, here let us say 1, 3, 4, 5 is one set of data and 3 and 4 they both become middle value. So, I take the average of this 3 plus 4 by 2 so,  $Q_1$  is 3.5.

Similar way  $Q_3$  is the middle value in the second half the data set. So, my second half the data set is 5, 6, 7 and 11. So, similar way I will take the average of 6 and 7. So, 6 plus 7 by 2 my  $Q_3$  is 6.5 and your interquartile range is basically 6.5 minus 3.5, 3. So, this will give you an idea that what is the overall variability in your data when you look at the  $Q_3$  that is the last 25 percent of the data and the first. So, what is the difference and how the data is basically spreaded

(Refer Slide Time: 16:28)

**Variance and Standard Deviation**

<p style="text-align: center;"><b>Population Variance</b></p> $\sigma^2 = \frac{\sum_{i=1}^N (x - \mu)^2}{N}$ $= \frac{\sum_{i=1}^N x^2 - \frac{\left(\sum_{i=1}^N x\right)^2}{N}}{N}$ $\sigma = \sqrt{\sigma^2}$	<p style="text-align: center;"><b>Sample Variance</b></p> $s^2 = \frac{\sum_{i=1}^n (x - \bar{x})^2}{(n - 1)}$ $= \frac{\sum_{i=1}^n x^2 - \frac{\left(\sum_{i=1}^n x\right)^2}{n}}{(n - 1)}$ $s = \sqrt{s^2}$
---	--

Now, if you see the range interquartile range and then third measure which is the variance in standard deviation. So, here typically each particular value of the data set is compared with the mean value. So,  $x - \mu$  or  $\frac{x_i - \mu}{N}$ , when I use this expression  $\frac{\sum_{i=1}^N (x - \mu)^2}{N}$ ; I have used  $\mu$ . And this is my standard this is my variance for the population when I take the square root it becomes standard deviation.

Similar, way you can do it for sample I am just putting here  $n - 1$  instead of  $N$ ; we will see the logic later on. But, in order to say go by or comply with my central limit theorem and also because I am dealing with the sample overall variability captured by the sample; obviously, would be on the higher side compared to the population I am dividing this by

$n - 1$ . So,  $\frac{\sum_{i=1}^n x^2 - \frac{(\sum_{i=1}^n x)^2}{n}}{(n-1)}$  say is the denominator and  $s = \sqrt{s^2}$  that is the standard deviation.



(Refer Slide Time: 17:56)

Calculation of Sample Variance			
$x$	$x - \bar{x}$	$(x - \bar{x})^2$	$x^2$
6	-9.85	97.0225	36
9	-6.85	46.9225	81
10	-5.85	34.2225	100
12	-3.85	14.8225	144
13	-2.85	8.1225	169
14	-1.85	3.4225	196
14	-1.85	3.4225	196
15	-0.85	0.7225	225
16	0.15	0.0225	256
16	0.15	0.0225	256
16	0.15	0.0225	256
17	1.15	1.3225	289
17	1.15	1.3225	289
18	2.15	4.6225	324
18	2.15	4.6225	324
19	3.15	9.9225	361
20	4.15	17.2225	400
21	5.15	26.5225	441
22	6.15	37.8225	484
24	8.15	66.4225	576
317	0	378.5500	5403

$$s^2 = \frac{\sum (x - \bar{x})^2}{(n - 1)} = \frac{378.55}{(20 - 1)}$$

$$= \frac{378.55}{19} = 19.923684$$

$$s^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{(n - 1)}$$

$$= \frac{5403 - \frac{317^2}{20}}{(20 - 1)} = \frac{5403 - \frac{100489}{20}}{19}$$

$$= \frac{5403 - 5024.45}{19} = \frac{378.55}{19} = 19.923684$$

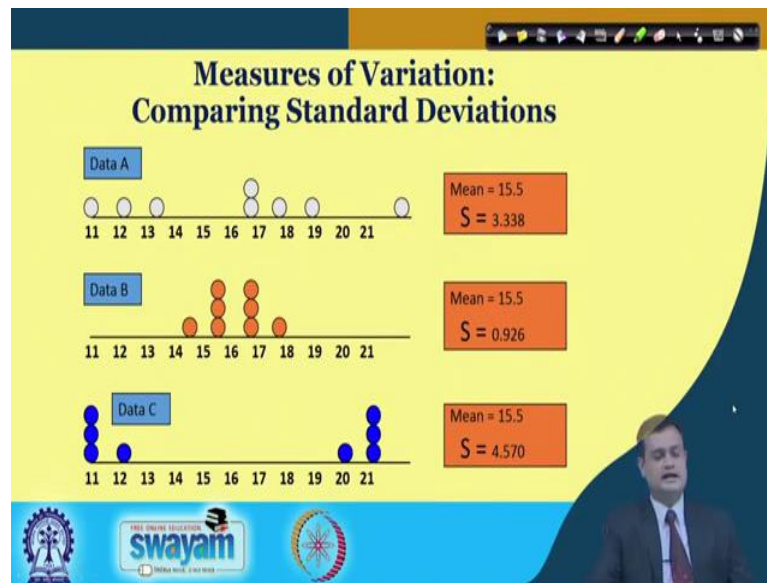
$$s = \sqrt{s^2} = \sqrt{19.923684} = 4.46$$

So, I have a data set and  $x$  is 6, 9, 10, 12, 13 and so on. I am taking  $x - \bar{x}$  or  $x_i - \bar{x}$  for various data and then I am taking the square of it. So, you will have all the positive values. I am taking the  $x^2$  and then I am just trying to plugging the values in the expression of  $s^2$ . So, you will see that my  $s^2$  comes out to be 19.923 and further numbers.

And you put your values in this expression so, you will get 19.923684. So, you have  $s = \sqrt{s^2}$  that is 4.46. So, this is my typically say standard deviation of the process. So, how much variability is there in my process is basically depicted by the standard deviation and the variance.



(Refer Slide Time: 19:17)

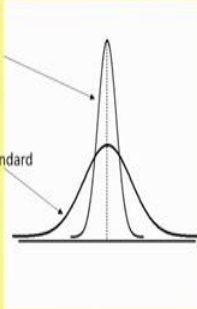


So now, let us move ahead and we have the display. We want to appreciate that actually when I say S is some number; what does it mean and how should I really compare. So, you just see the data set A, data set B and data set C. The data set A has a mean value 15.5 and S is 3.338. Data set B has mean value 15.5, S is equal to 0.926 and data set C has a mean value 15.5 and S value 4.570.

Now, there is one thing to observe that all the data set they have same mean value that is 15.5, but your standard deviation or variance if you take the square associated with each particular dataset is different. So, your dataset A if you observe and I have just put the frequency that how many times a particular number is occurring. Then it has very large variability associated with the data and that is 3.338. But, data set B has same mean value if I look at the S 0.926, so, your data is very much centric towards your mean value and hence the variability is less. But, if you see dataset C some data you have on one extreme, some data you have on other extreme and this is exactly reflected in  $S = 4.570$ . So, this particular display would help you to understand, that if I am analyzing different processes in six sigma then my process may have various processes may have same mean, but different variability. And, the process which has much higher variability needs to be corrected and controlled; otherwise at any point of time it will start producing the defective items.

(Refer Slide Time: 21:27)

### Measures of Variation: Comparing Standard Deviations



**Key Insights**

- The more the data are spread out, the greater the range, variance, and standard deviation.
- The more the data are concentrated, the smaller the range, variance, and standard deviation.
- If the values are all the same (no variation), all these measures will be zero.
- None of these measures are ever negative.

swayam

So, this is the important part about variability and you can further see here and we are talking about the six sigma. So, just see the first particular say your distribution typically bell shaped normal distribution. And, you will say that it has a larger standard deviation. When I see the second one which is more centric towards the central value; you will say it is a smaller standard deviation. And, this particular process which has a smaller standard deviation can help you to achieve the better sigma level five sigma, Six Sigma and this is exactly what we are trying to achieve through DMAIC cycle.

(Refer Slide Time: 22:20)

### Measures of Variation

**The Coefficient of Variation**

- Measures relative variation
- Always in percentage (%)
- Shows variation relative to mean
- Can be used to compare the variability of two or more sets of data measured in different units

$$CV = \left( \frac{S}{\bar{X}} \right) \cdot 100\%$$

swayam

Now, I would just like to focus on couple of more measures which are just the derivative of some of the fundamental measures we have studied that is mean and variability measures. So, there is an important measure called measure of variation. So, this measure of variation is explained as coefficient of variation; typically CV and it measures the relative variation.

So, we have seen that the process may have various processes may have same mean, but the variability is different and the process with very high variability is considered a poor process. So, here just same thing is reflected in terms of ratio that  $CV = \left( \frac{S}{X} \right) \cdot 100\%$ ; I want to check that what is the percentage variability with respect to my mean value.

(Refer Slide Time: 23:14)

**Measures of Variation:  
Comparing Coefficients of Variation**

- Stock A:
  - Average price last year = \$50
  - Standard deviation = \$5
$$CV_A = \left( \frac{S}{X} \right) \cdot 100\% = \frac{\$5}{\$50} \cdot 100\% = 10\%$$
- Stock B:
  - Average price last year = \$100
  - Standard deviation = \$5
$$CV_B = \left( \frac{S}{X} \right) \cdot 100\% = \frac{\$5}{\$100} \cdot 100\% = 5\%$$

Both stocks have the same standard deviation, but stock B is less variable relative to its price

swayam

So, just see the example here the example is very simple; suppose I am analyzing two different or three different or four different stocks. And let us say there is stock A which has average price last year dollar 50 or some rupees, standard deviation 5 and I am just trying to find out what is the percentage variability with respect to mean. So,  $\frac{5}{50} \cdot 100 = 10\%$ . Same way you see the stock B which has average price last year 100 and standard deviation I am keeping same 5; so, it is 5 percent.

(Refer Slide Time: 23:57)

**Measures of Variation:  
Comparing Coefficients of Variation**  
(continued)

- Stock A:
  - Average price last year = \$50
  - Standard deviation = \$5

$$CV_A = \left( \frac{S}{\bar{X}} \right) \cdot 100\% = \frac{\$5}{\$50} \cdot 100\% = 10\%$$

- Stock C:
  - Average price last year = \$8
  - Standard deviation = \$2

$$CV_C = \left( \frac{S}{\bar{X}} \right) \cdot 100\% = \frac{\$2}{\$8} \cdot 100\% = 25\%$$

Stock C has a much smaller standard deviation but a much higher coefficient of variation

swayam

And, then if I see the stock A and C then stock A has 10 percent C with high say average and higher lower standard deviation it has 25 percent. So, what you can reveal in the first case you can say that both stocks have the same standard deviation, but the stock B is less variable relative to its price. In this case what you can say that stock C has a much smaller standard deviation, but much higher coefficient of variance.

So, please try to understand many a times when we do not check it with respect to mean and this ratio we may get misguided; if we just look at the mean value or if we just look at the variance. Here you can very well see that stock C has a much higher coefficient of variation with some value of mean and your standard deviation.

So, now the issue is that how can I establish the relationship between mean and the standard deviation.

(Refer Slide Time: 25:09)

### Relations between the Mean and Standard Deviation

- **Chebyshev's Theorem**
  - ✓ Applies to **any** distribution, regardless of shape
  - ✓ Places lower limits on the percentages of observations within a given number of standard deviations from the mean
- **Empirical Rule**
  - ✓ Applies only to roughly **mound-shaped** and **symmetric** distributions
  - ✓ Specifies approximate percentages of observations within a given number of standard deviations from the mean

Logos at the bottom: Swamyam, a circular emblem, and a small video inset of a man in a suit.

So, there are some simple procedures and rules to follow in order to set the relationship between my mean and the standard deviation. So, number 1 is Chebyshev's theorem and number 2 is Empirical rule. So, Chebyshev's theorem typically it applies to any distribution and regardless of the shape. So, I am not very much sticky about the shape, it could be applied to any kind of distribution.

Now, in Chebyshev's theorem what it does it basically places the lower limits on the percentage of observations within a given number of standard deviation from the mean. And when you look at the empirical rule then typically it is applied to mound shaped or symmetric distribution like normal and specifically approximates the percentage of the observation within a given number of standard deviation. So, plus or minus  $k$  sigma with respect to mean I am trying to find out.

(Refer Slide Time: 26:15)

### Chebyshev's Theorem

- Regardless of how the data are distributed, at least  $(1 - 1/k^2) \times 100\%$  of the values will fall within  $k$  standard deviations of the mean (for  $k > 1$ )
- Examples:

At least	within
$(1 - 1/2^2) \times 100\% = 75\%$	$k=2 \quad (\mu \pm 2\sigma)$
$(1 - 1/3^2) \times 100\% = 89\%$	$k=3 \quad (\mu \pm 3\sigma)$

So, just see this Chebyshev's theorem you have the rule like this that regardless how the data is distributed let it not be normal, not symmetrical  $(1 - 1/k^2) \times 100\%$  of the value fall within  $k$  standard deviation of the mean for; obviously,  $k > 1$ . So, you can just see the numerical example  $(1 - 1/2^2) \times 100\% = 75$  so, my  $k$  value is 2. So, I would say that 75 percent of the data set falls within  $\mu \pm 2\sigma$ . Similar, way if I assume  $k$  is equal to 3 then 89 percent. So, my 89 percent of the data set typically they fall in the range of this.

Now, you would say what is the need of finding this. You just think about demand data, you just think about product specification, you will find that you would be interested to find that what is that number may be consumer may be demand data that would fall in a particular range. And, you would definitely like to focus on that for developing a product, for developing a marketing strategy, for satisfying the customer and finding this range is very important and, hence the establishing relationship between  $\mu$  and  $\sigma$ .




(Refer Slide Time: 27:44)

### Chebyshev's Theorem

- At least  $\left(1 - \frac{1}{k^2}\right)$  of the elements of **any** distribution lie within **k** standard deviations of the mean

	$1 - \frac{1}{k^2}$		Standard deviations of the mean
At least	$1 - \frac{1}{2^2} = 1 - \frac{1}{4} = \frac{3}{4} = 75\%$	2	
	$1 - \frac{1}{3^2} = 1 - \frac{1}{9} = \frac{8}{9} = 89\%$	3	
	$1 - \frac{1}{4^2} = 1 - \frac{1}{16} = \frac{15}{16} = 94\%$	4	

Lie within




So, this another example can also further help you. So, if I take  $k = 2$ ; 75 percent. If I take  $k = 3$ ; 89 percent, 4 it is 94; percent and this is 2 times standard deviation of the mean 3 times standard deviation of the mean and 4 times standard deviation of the mean.

(Refer Slide Time: 28:05)

### The Empirical Rule

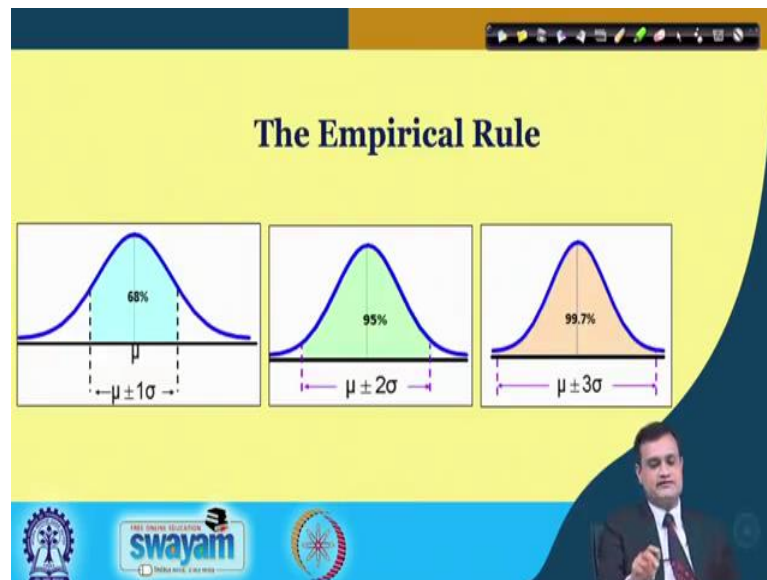
- The empirical rule approximates the variation of data in a bell-shaped distribution
- Approximately 68% of the data in a bell shaped distribution is within 1 standard deviation of the mean or  $\mu \pm \sigma$
- Approximately 95% of the data in a bell-shaped distribution lies within two standard deviations of the mean, or  $\mu \pm 2\sigma$
- Approximately 99.7% of the data in a bell-shaped distribution lies within three standard deviations of the mean, or  $\mu \pm 3\sigma$



If you look at the empirical rule it applies to the symmetric distribution like normal and we say that approximately 68 percent of the data fall in bell shape distribution normal distribution; when I say  $\mu \pm \sigma$ . When I say  $\mu \pm 2\sigma$  it is 95 percent data they fall in this range  $\mu \pm 3\sigma$  which is a widely used range 99.7 percent.



(Refer Slide Time: 28:34)



So, this is where you can see pictorially that what exactly I mean to say in the empirical rule  $\pm 1\sigma$ ,  $2\sigma$ , and  $3\sigma$ .

(Refer Slide Time: 28:45)

### Empirical Rule

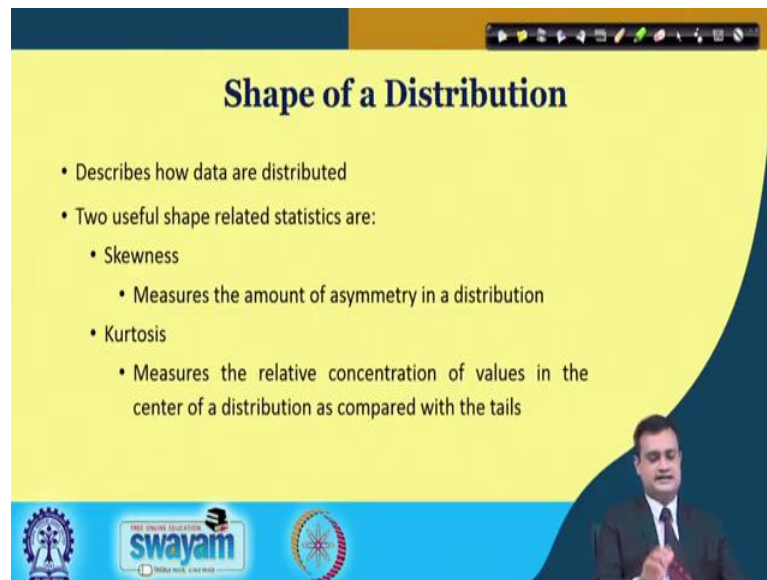
- For roughly **bell (or mound)-shaped** and **symmetric** distributions, approximately:

68%	<b>Lie within</b>	1 standard deviation of the mean
95%		2 standard deviations of the mean
All		3 standard deviations of the mean

The figure shows a table summarizing the Empirical Rule. The table has three rows and three columns. The first column contains the percentages: 68%, 95%, and All. The second column contains the text 'Lie within' centered vertically. The third column contains the corresponding standard deviation ranges: 1 standard deviation of the mean, 2 standard deviations of the mean, and 3 standard deviations of the mean. At the bottom of the slide, there are logos for 'swayam' and 'INDIA RISE, EDUCATION RISE'.

And you have say one standard deviation from the mean just the summary 68 percent, 2 95 percent, 3 almost 99.7 percent in this.

(Refer Slide Time: 28:59)



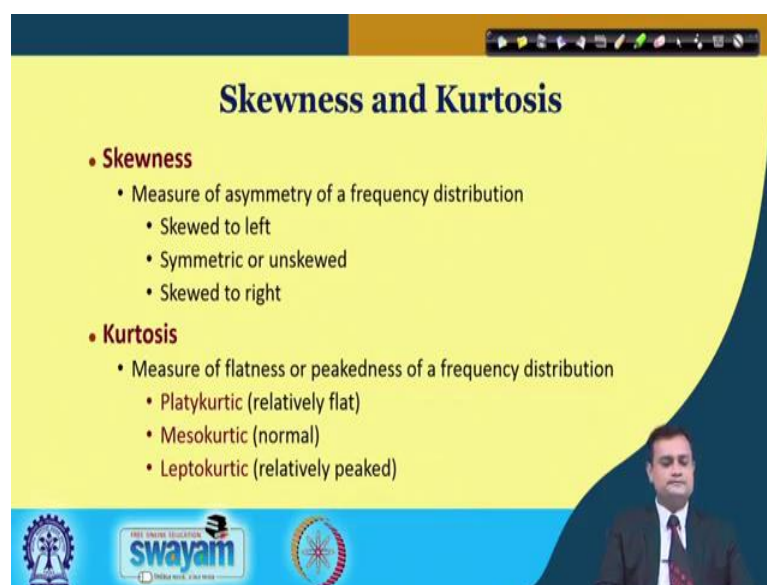
**Shape of a Distribution**

- Describes how data are distributed
- Two useful shape related statistics are:
  - Skewness
    - Measures the amount of asymmetry in a distribution
  - Kurtosis
    - Measures the relative concentration of values in the center of a distribution as compared with the tails

The slide features a yellow background with a dark blue curved border on the right. At the bottom, there are logos for 'swayam' and 'INDIA RITE, A PAFI INITIATIVE'.

Now, another important dimension is the shape and we talk about Skewness and the Kurtosis as a part of the shape. So, typically Skewness measures the amount of asymmetry in a distribution and Kurtosis typically talks about the peakedness.

(Refer Slide Time: 29:16)



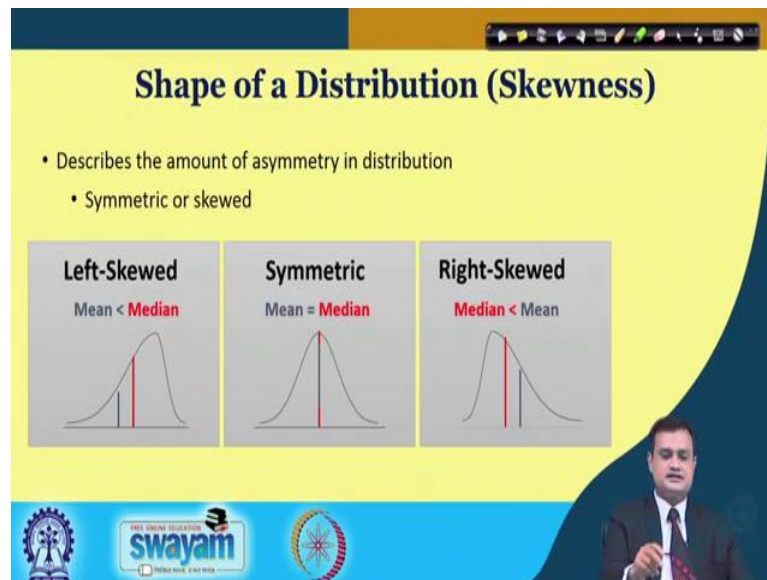
**Skewness and Kurtosis**

- **Skewness**
  - Measure of asymmetry of a frequency distribution
    - Skewed to left
    - Symmetric or unskewed
    - Skewed to right
- **Kurtosis**
  - Measure of flatness or peakedness of a frequency distribution
    - Platykurtic (relatively flat)
    - Mesokurtic (normal)
    - Leptokurtic (relatively peaked)

The slide features a yellow background with a dark blue curved border on the right. At the bottom, there are logos for 'swayam' and 'INDIA RITE, A PAFI INITIATIVE'.

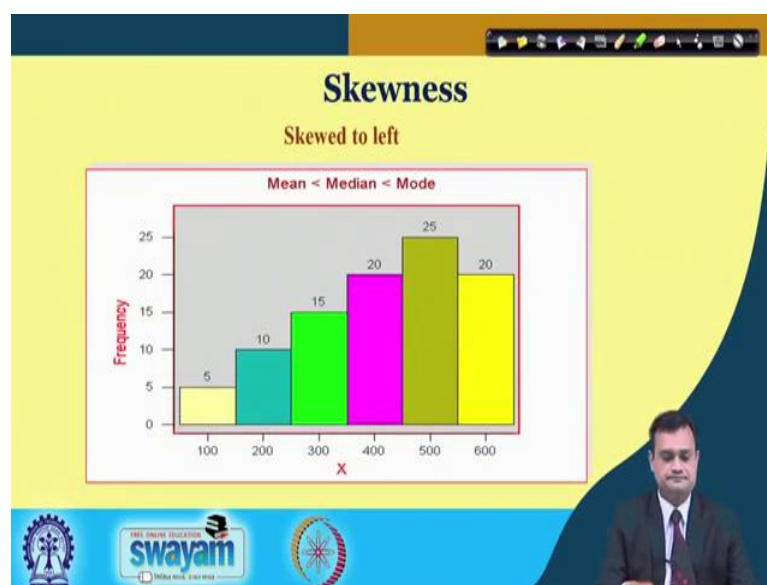
So, we have the different variety when we talk about the shape. So, Skewness it may be skewed to left, unskewed right. Kurtosis you may have relatively flat distribution, you may have normal, you may have relatively peaked.

(Refer Slide Time: 29:33)



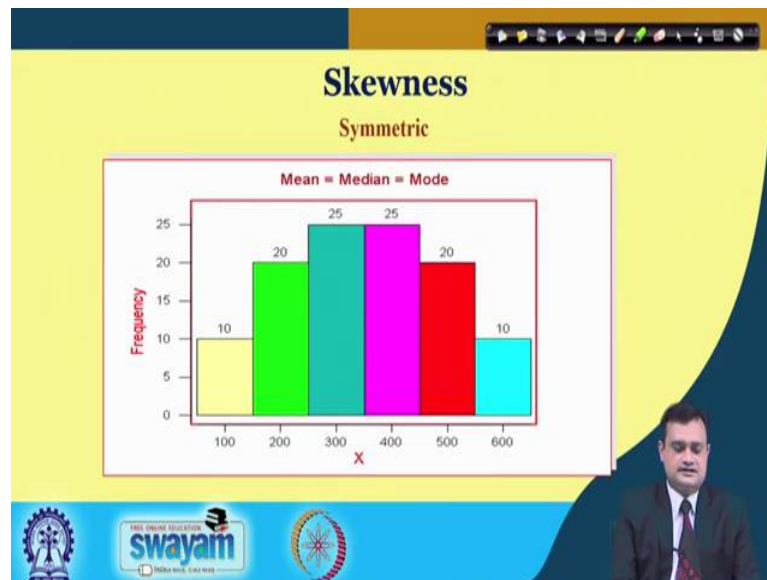
And these are some of the examples that says that left skewed symmetric, right skewed and you have mean is equal to median in the central case, mean is less than median, median is less than mean and so on.

(Refer Slide Time: 29:49)



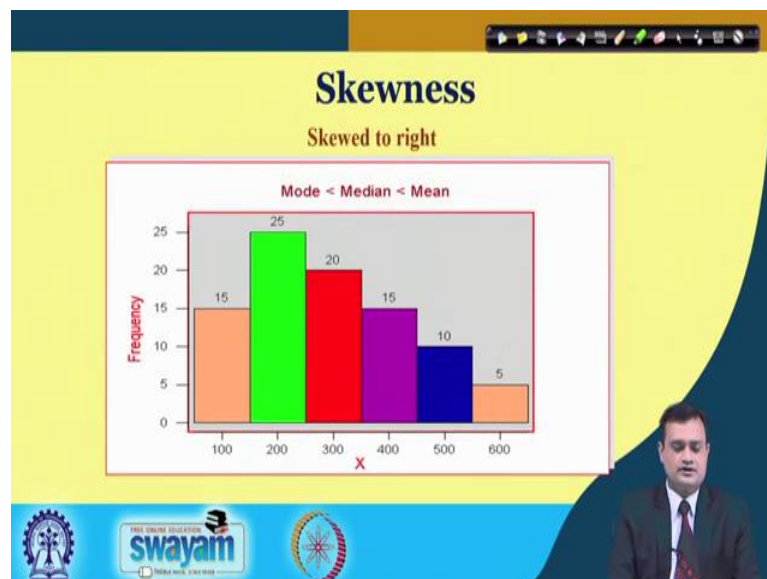
Some of the examples I have putted for your say I have put for your self-study, Skewness. .

(Refer Slide Time: 29:56)



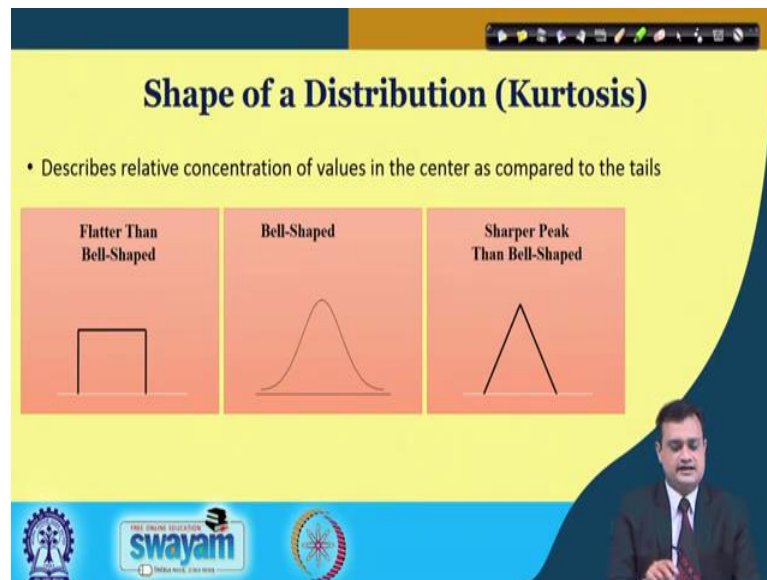
Then another examples Skewness, where all three are equal mean median and mode.

(Refer Slide Time: 30:03)



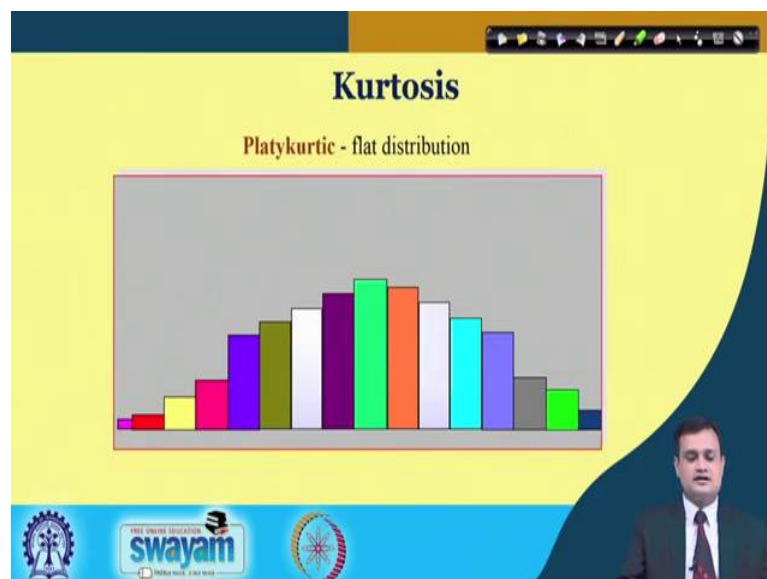
You have mode less than median less than mean that is  $Q_2$  right.

(Refer Slide Time: 30:08)



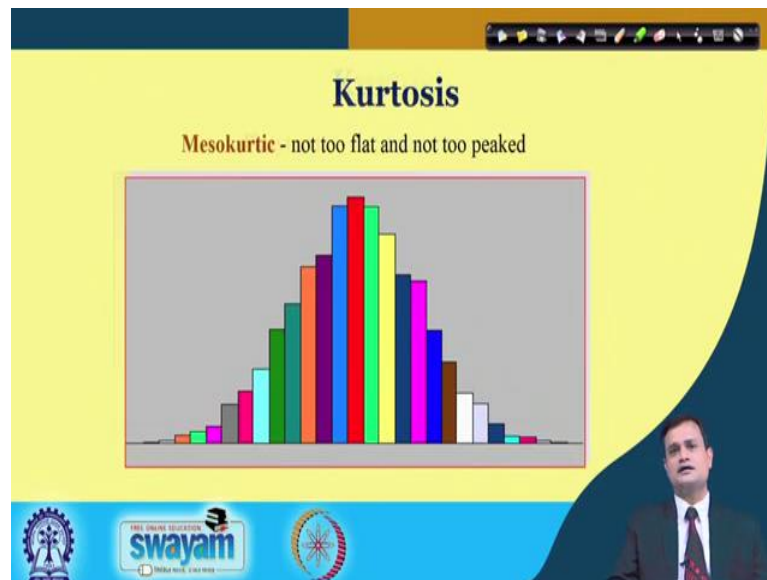
And as a summary you can see that various ways you can present this Skewness ok. So, shape of a distribution typically we talk about the Kurtosis and you can see that you can have a flat shaped, bell shaped which is typically a normal symmetric and sharper peak. So, typically it is a bell shape then the bell shape which has a sharper peak.

(Refer Slide Time: 30:35)



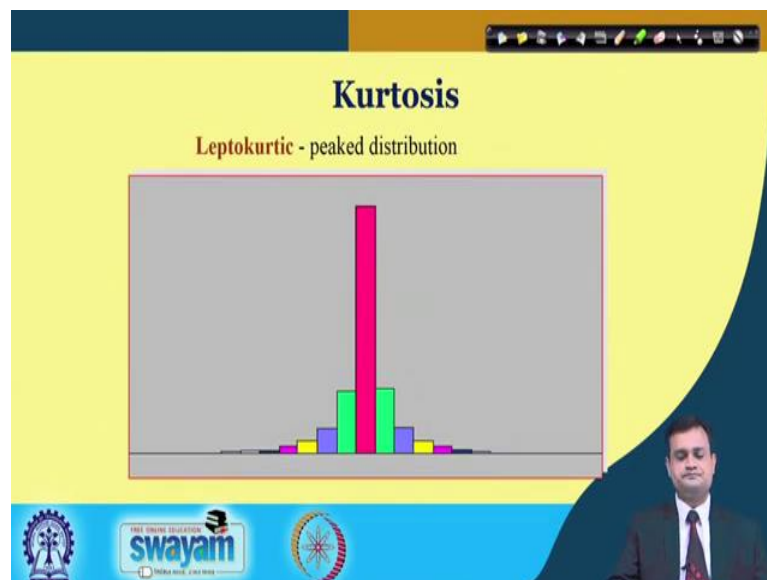
So, Kurtosis I am just presenting in the form of say histogram and you can see that it has more or less flat distribution.

(Refer Slide Time: 30:46)



You can see that it is called Mesokurtic and it has not too flat and not too peaked. So, may be very close to normal.

(Refer Slide Time: 30:57)

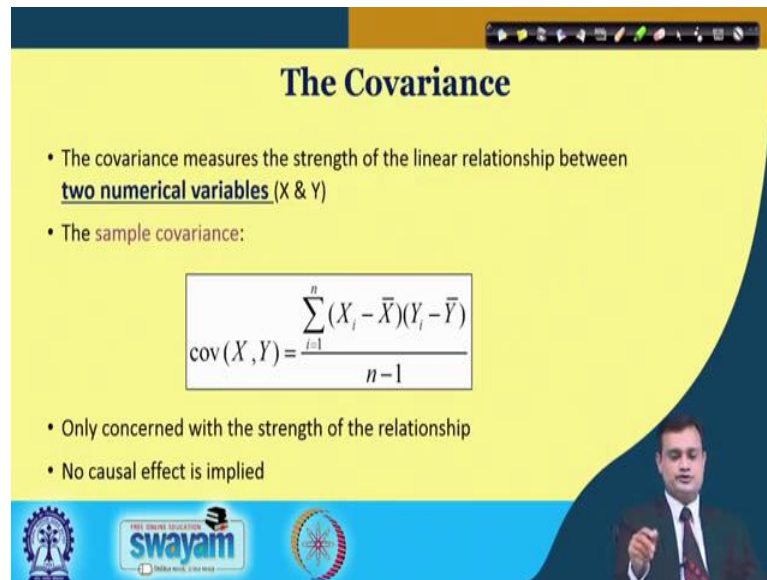


And this you can see that it is a peaked distribution. So, once again you will have the question in mind that what is the need of presenting my data in this form. So, again I would say that basically I want to see that how my data is distributed whether it is symmetric whether it is going high in one value or it is skewed on left side, right side.



So, this kind of information can really help me to understand; suppose I am talking about the case of sales data and if I take the 12 month. And, if I plot my say data it would give me immediate an idea that fine in which month or in what are what are the couple of months in which the sales is maximum and how the data is distributed.

(Refer Slide Time: 31:49)






### The Covariance

- The covariance measures the strength of the linear relationship between two numerical variables (X & Y)
- The sample covariance:

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

- Only concerned with the strength of the relationship
- No causal effect is implied

The another important measure is a covariance whatever we have so, seen so, far is mainly with respect to one data let us say sales data. But, do not you think you would also be interested to check that sales of let us say wafers, how it is changing with respect to the sales of the cold drink like Pepsi. So, may be you can assume another variable that suppose sales with respect to let us say season and you would try to study that how the variables they are associated.

So, here you can say that you have two numerical variables X and Y and covariance is X and Y. So, may be in manufacturing set up you can say that you have set of operators and their skill and then you have the output. So, is there a relationship between these two

variables. So, you can check it by using a very simple expression;  $\sum_{i=1}^n (X_i - \bar{X})$  that is the


first variable and  $\frac{(Y_i - \bar{Y})}{(n-1)}$  because we are using the sample.



(Refer Slide Time: 33:11)

### Interpreting Covariance

- **Covariance** between two variables:
  - $\text{cov}(X,Y) > 0 \rightarrow$  X and Y tend to move in the **same** direction
  - $\text{cov}(X,Y) < 0 \rightarrow$  X and Y tend to move in **opposite** directions
  - $\text{cov}(X,Y) = 0 \rightarrow$  X and Y are independent
- The covariance has a major flaw:
  - It is not possible to determine the relative strength of the relationship from the size of the covariance



So, you can interpret if co variance of X Y greater than 0, X and Y tend to move in the same direction X increases Y increases. If they are less than 1 then X increases Y decreases opposite direction, if 0 there is no relationship they are just independent.


(Refer Slide Time: 33:30)

### Coefficient of Correlation

- Measures the relative strength of the linear relationship between two numerical variables
- Sample coefficient of correlation:

$$r = \frac{\text{cov}(X,Y)}{S_X S_Y}$$

where


$$\text{cov}(X,Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$
$$S_X = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}}$$
$$S_Y = \sqrt{\frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n-1}}$$


There is one another important measure which is called coefficient of correlation and typically either it is described as  $r$  or  $\rho$ . So,  $r = \text{cov}(X,Y) / S_X S_Y$  I am talking about two variable divided by  $S_X S_Y$ ; we have already seen how to find  $S_X$  and  $S_Y$  and when you plug in the values in these you will get  $\text{cov}(X,Y)$ .

(Refer Slide Time: 33:57)

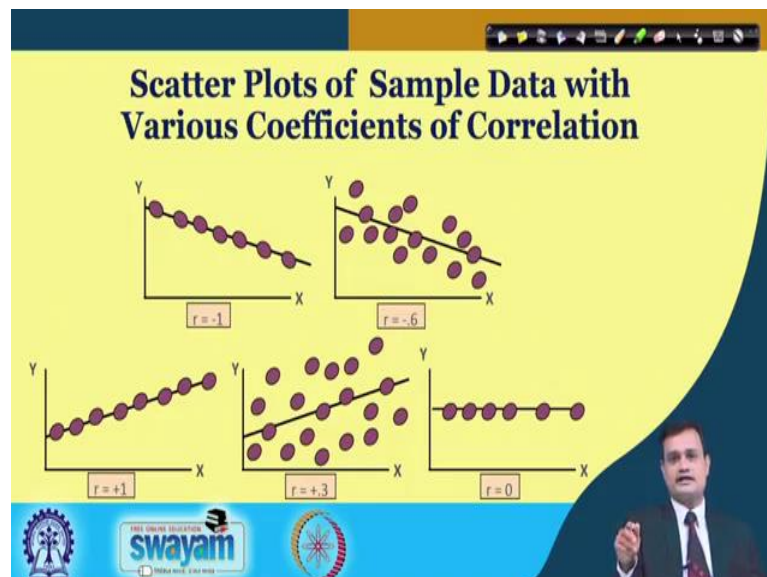
### Features of the Coefficient of Correlation

- The population coefficient of correlation is referred as  $\rho$ .
- The sample coefficient of correlation is referred to as  $r$ .
- Either  $\rho$  or  $r$  have the following features:
  - Unit free
  - Ranges between  $-1$  and  $1$
  - The closer to  $-1$ , the stronger the negative linear relationship
  - The closer to  $1$ , the stronger the positive linear relationship
  - The closer to  $0$ , the weaker the linear relationship



So, typically you can visualize that covariance can vary from minus 1 to 1. And if it is closer to minus 1 then strong negative relationship exist. If it is 0 then very closer to 0 rather very weak linear relationship exist and if it is closer to 1 then a positive linear relationship you can predict between two variables.

(Refer Slide Time: 34:27)

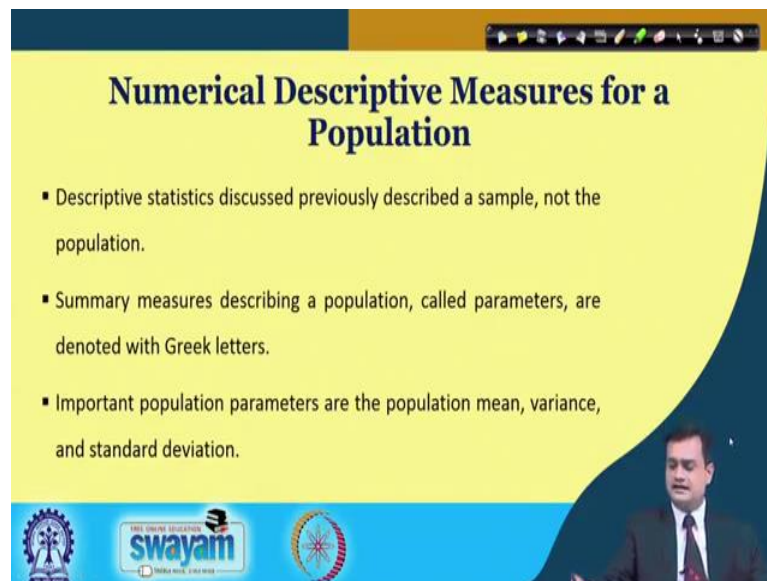


So, this is just the example with various  $r$  values; in first case it is  $r$  is equal to say this case  $r$  is equal to minus 1. So, there is a negative relationship. Here  $r$  is equal to minus 0.6 negative relationship, but it is data is more scattered. So, not a strong relationship.

So, basically it talks about the strength of the relationship  $r$  is equal to 1 positive strong relationship, data is not scattered, your say line is covering all the data.

When you see this  $r$  is equal to 0.3 so, there is a positive relationship, but your data is scattered and your line is only passing through some the data set and  $r$  is equal to 0. So, there is no relationship between  $X$  and  $Y$ .

(Refer Slide Time: 35:27)



**Numerical Descriptive Measures for a Population**

- Descriptive statistics discussed previously described a sample, not the population.
- Summary measures describing a population, called parameters, are denoted with Greek letters.
- Important population parameters are the population mean, variance, and standard deviation.

The slide features a yellow background with a dark blue curved border on the right. At the bottom, there is a blue banner with logos for 'swayam' and other educational institutions. A small video inset in the bottom right corner shows a man in a suit speaking.

So, typically you have numerical measures for a population as well as sample. And, when I talk about the population I have to consider the population standard deviation and mean and variance.

(Refer Slide Time: 35:42)

### Numerical Descriptive Measures For A Population


- The **population mean** is the sum of the values in the population divided by the population size, N

$$\mu = \frac{\sum_{i=1}^N X_i}{N} = \frac{X_1 + X_2 + \dots + X_N}{N}$$

- The **population standard deviation** is the average of squared deviations of values from the mean

$$\sigma^2 = \frac{\sum_{i=1}^N (X_i - \mu)^2}{N}$$

$\mu$  = population mean  
 $N$  = population size  
 $X_i$  =  $i^{\text{th}}$  value of the variable X




And this is the way we try to write using the mean and the capital N, that is the population size to find the  $\mu$  value and the  $\sigma^2$ . So, many a times I would mention that we are not aware of the population variance and mean and then we have to go with certain assumptions that we will see later on. So, here we are going by the assumption that you can find out the population mean and the variance.

(Refer Slide Time: 36:17)

### Sample statistics versus population parameters

Measure	Population Parameter	Sample Statistic
Mean	$\mu$	$\bar{X}$
Variance	$\sigma^2$	$S^2$
Standard Deviation	$\sigma$	$S$



So, typically your notations are like this if it is for population  $\mu$ , sample  $\bar{X}$ , variance  $\sigma^2$ ,  $S^2$ .

(Refer Slide Time: 36:28)

### The Central Limit Theorem!

If all possible random samples, each of size  $n$ , are taken from any population with a mean  $\mu$  and a standard deviation  $\sigma$ , the sampling distribution of the sample means (averages) will:

1. have mean:  $\mu_{\bar{x}} = \mu$
2. have standard deviation:  $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
3. be approximately normally distributed regardless of the shape of the parent population (normality improves with larger  $n$ ).

swayam  
THINKING NOTE, LEARNING

And central limit theorem which is the most important theorem in statistics I would just like to give you the details of this. So, if all possible random sample you are taking the different samples, each of size  $n$ . Suppose you are manufacturing a bearing and you are taking 10 samples each sample consist let us say 50 bearings for the measurement. Now, each sample having sample size  $n$  is taken from any population with a mean  $\mu$  and standard deviation  $\sigma$  that is about population; sampling distribution of the sample means. So, you take the mean of each particular sample having 50 bearings, you are measuring the outer diameter then you will have  $\mu_{\bar{x}} = \mu$ . So, the mean of your sample mean would be very much close to the mean of your population. And, now I would like to remind you that here standard deviation  $\sigma_{\bar{x}}$  will be  $\frac{\sigma}{\sqrt{n}}$ . So, when I want to find the standard deviation of my mean sample then  $\sigma$  is the population standard deviation should be divided by  $\sqrt{n}$ . And, be approximately normally distributed regardless of the shape of the population; your population may have triangular distribution, it may have exponential distribution. We will see couple of distribution in the next class, but your sample will have normal distribution typically when your sample size  $n$  is larger.



(Refer Slide Time: 38:13)

### The Central Limit Theorem!

- When independent random variables are added, their properly normalized sum tends toward a normal distribution (informally a "bell curve") even if the original variables themselves are not normally distributed.
- **The theorem is a key ("central") concept in probability theory** because it implies that probabilistic and statistical methods that work for normal distributions can be applicable to many problems involving other types of distributions.
- **Example:** If one flips a coin many times the probability of getting a given number of heads in a series of flips will approach a normal curve, with mean equal to half the total number of flips in each series.

Logos: IIT Bombay, Swamyam, and a circular emblem.

So, this is the crux of central limit theorem and it is the central concept in the domain of statistical and probability theory. And typically it helps because, many of the distributions can be approximated to normal and normal distribution is the most convenient, easy to deal with and this is where the central limit theorem plays an important role.

(Refer Slide Time: 38:40)

### Random Variable

- It is a variable whose possible values are numerical outcomes of a random phenomenon.
- For example, the value of the first roll of a die. For example, the sum of a roll of two dice.

**Two types of random variables: Discrete and Continuous**

- **Discrete random variables:** where the possible events are countable. For example, the roll of a dice, or the outcome of a horse race, or whether the firm will default or not.
- **Continuous random variables:** where the possible events are not countable. For example, the number of white hair on my head, or how much dividend I&T will announce next year, or the price of HDFC Bank stock.

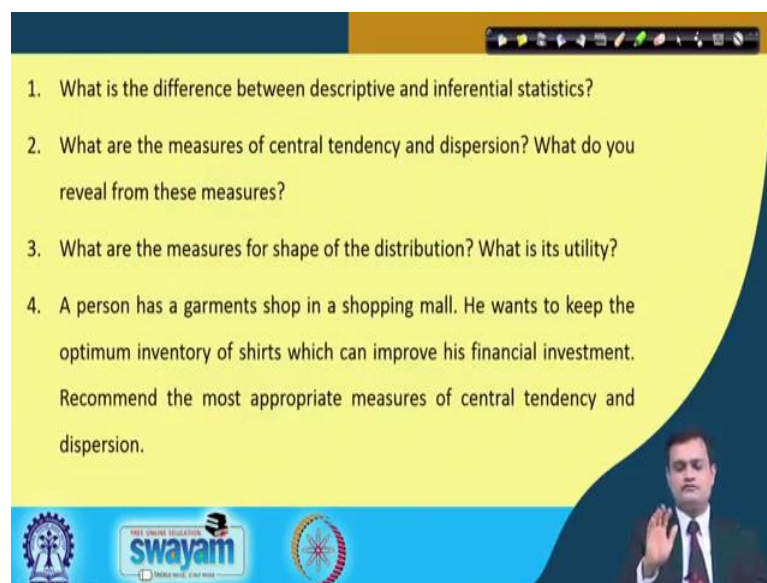
Logos: IIT Bombay, Swamyam, and a circular emblem.

Before, we complete and conclude the important term on which the entire statistics probability theory is based I would like to share random variable. And, typically it is a

variable whose possible values are numerical outcomes of a random phenomena. Suppose you toss a coin or you roll a die you do not know there could be a chance of getting head and tail, there could be a chance of getting 1, 2, 3, 4 whatever.

So, broadly you have random variable which is an outcome of random phenomena and you have discrete means specific value or you have continuous it can fall in a particular range. May be 10 to 11, 10.1, 10.4, 10.7 and these two are the different say types of random variables.

(Refer Slide Time: 39:39)



1. What is the difference between descriptive and inferential statistics?

2. What are the measures of central tendency and dispersion? What do you reveal from these measures?

3. What are the measures for shape of the distribution? What is its utility?

4. A person has a garments shop in a shopping mall. He wants to keep the optimum inventory of shirts which can improve his financial investment. Recommend the most appropriate measures of central tendency and dispersion.

So, before we finish as a part of our practice I want to float couple of questions for your introspection and recap. So, what is the difference between descriptive and inferential statistics? What are the measures of central tendency and the dispersion? And what do you reveal from these, what is its importance? What are the measures of shape and distribution? What is its utility?

And, suppose I give you a case that a person running a garment shop suppose in a shopping mall he wants to keep the optimum inventory of shirts which can improve his financial investment. Suppose you are not able to sale the shirts then this inventory is lost or it will become absolute. Now, in this case what do you recommend as the appropriate measures of central tendency and dispersion, that you should use in order to have the better decision making on inventory keeping.

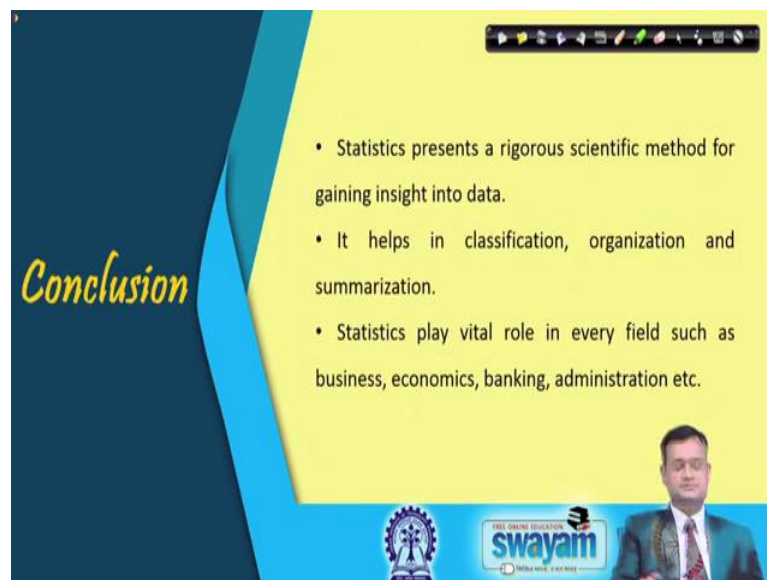


(Refer Slide Time: 40:38)



So, these are the references you can use Aczel, Levine and Kubiak, T. M. Kubiak, Michael Kutner. So, this will help you to appreciate the concepts in detail.

(Refer Slide Time: 40:50)



So, as a summary conclusion statistics presents a rigorous scientific method for going insights into dig data. Helps in classifying organizing and summarizing. And, it plays a vital role in every field such as business, banking, administration, manufacturing and all.

So, thank you very much for your patience in learning the concept and I hope this preliminary understanding will definitely create a base for understanding the other

statistical concepts in the subsequent lectures and keep revising all the concepts. We are discussing the major phase of Six Sigma. So, be with me enjoy.