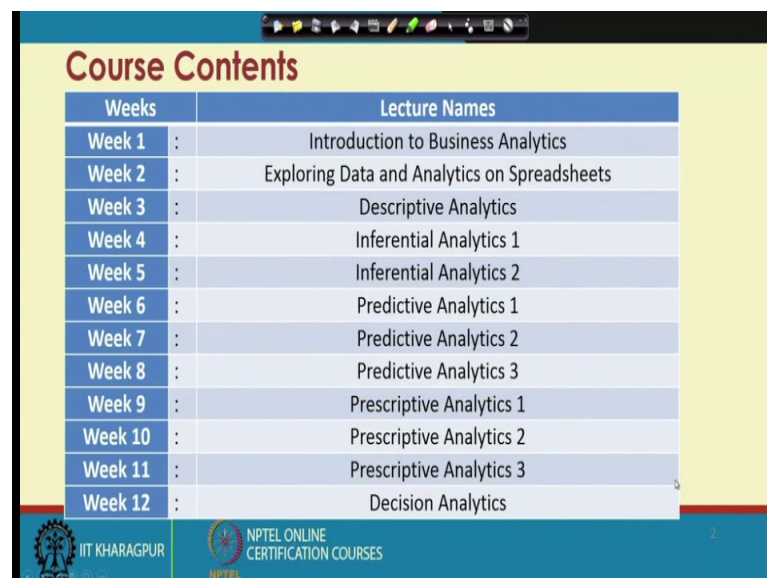


**Business Analytics for Management Decision**  
**Prof. Rudra P Pradhan**  
**Vinod Gupta School of Management**  
**Indian Institute of Technology, Kharagpur**

**Lecture - 11**  
**Descriptive Analytics**

Hello everybody. This is Rudra Pradhan here, and welcome you to BMDA course. So, today the topic of discussion is; descriptive analytics, that to we are here to highlight the unit 3 lecture.

(Refer Slide Time: 00:25)



Weeks	Lecture Names
Week 1	Introduction to Business Analytics
Week 2	Exploring Data and Analytics on Spreadsheets
Week 3	Descriptive Analytics
Week 4	Inferential Analytics 1
Week 5	Inferential Analytics 2
Week 6	Predictive Analytics 1
Week 7	Predictive Analytics 2
Week 8	Predictive Analytics 3
Week 9	Prescriptive Analytics 1
Week 10	Prescriptive Analytics 2
Week 11	Prescriptive Analytics 3
Week 12	Decision Analytics

The slide is titled 'Course Contents' and features a table with two columns: 'Weeks' and 'Lecture Names'. The table lists 12 weeks of the course, starting with 'Introduction to Business Analytics' in Week 1 and ending with 'Decision Analytics' in Week 12. The slide also includes logos for IIT Kharagpur and NPTEL Online Certification Courses at the bottom.

And in the previous 2 lectures, we have discussed about the introduction to business analytics, and exploring data and analytics on spreadsheet. So, technically we have discussed details about the business analytics requirement the tools, and the importance the applications. So, accordingly as per the requirement of you know analyzing any business problem through business analytic tools.

So, the first-hand requirement is the data. And that is how in the unit 2 we have discussed the data in various angles. So, like you know understanding the data, data view, data visualizations and we have discussed various tools graphical tools, and quantity tools to understand the data to visualize the data, and then make the data in a particular structure so that you know we can analyze in a better way as per the problem requirement. And

then then solve a business problem. So, here we will go a little bit more and that to the particular requirement of you know analytics tools.

(Refer Slide Time: 01:47)

**Objectives of Business Analytics**

- ▶ **Descriptive analytics**
  - What happened in the past?
  - Many organizations use DA as part of business intelligence.
- ▶ **Predictive analytics**
  - What will happen in the future
  - Many organizations use predictive analytics.
- ▶ **Prescriptive analytics**
  - What is the best action.
  - Small proportion of organizations use the prescriptive analytics.

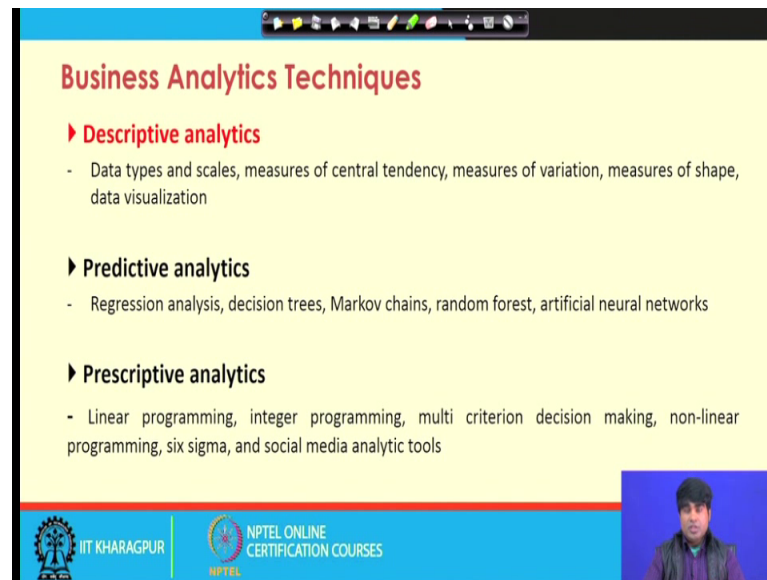
IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | 13

So, the discussion is like this. So, in the last class or in the first 2 units we have drastically highlighted the classification of you know, business analytics tools.

So, business analytics tools basically divided into 3 groups. So far as you know, so far as the discussion is concerned to take some kind of you know management decision. So, the first one is in descriptive analytics. Second one is the predictive analytics. Third one is the prescriptive analytics. So, these are all discussed earlier. So, just I am highlighting. So, that we can connect with you know the todays discussion. In the descriptive analytics the basic idea is to discuss what happened in the past. And knowing the past kind of you know trend or the structure. So, the predictive analytics really analyze the particular requirement, and then like to answer the questions what will happen in the future.

Then on the basis of descriptive analytics and predictive analytics, prescriptive analytics is the structure where we have to find out the best you know possible actions; that means, the optimum values of the decision variables. So now, here we are in the process of you know discussing the descriptive analytics.

(Refer Slide Time: 03:04)



**Business Analytics Techniques**

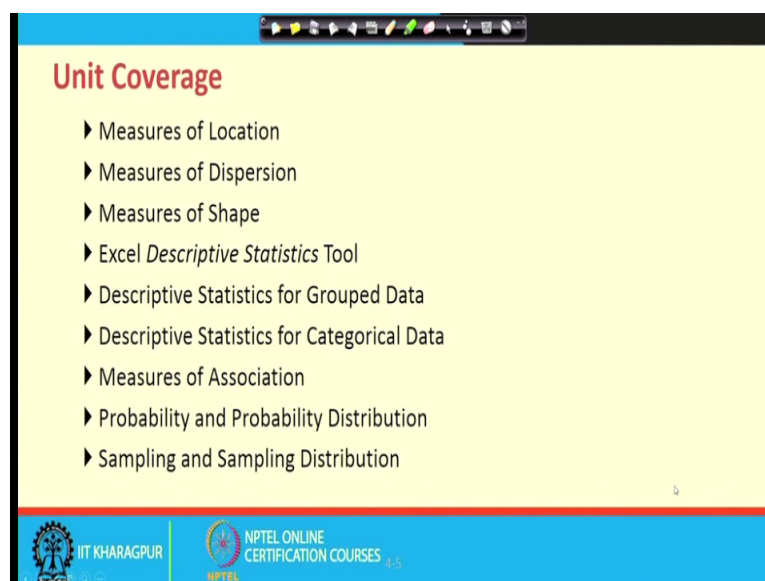
- ▶ **Descriptive analytics**
  - Data types and scales, measures of central tendency, measures of variation, measures of shape, data visualization
- ▶ **Predictive analytics**
  - Regression analysis, decision trees, Markov chains, random forest, artificial neural networks
- ▶ **Prescriptive analytics**
  - Linear programming, integer programming, multi criterion decision making, non-linear programming, six sigma, and social media analytic tools

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

And as per the earlier kind of you know discussion. So, these are the various tools available under descriptive analytics, predictive analytics and prescriptive analytics. And in the case of descriptive analytics. So, we have a data types and scales, measurement of central tendency, measures of variations, measures of safe and data visualization. And in fact, we have already discussed details about the data types and scales, and the kind of you know understanding the kind of you know structuring restructuring. And that we have already discussed in the unit 2.

So now here is we directly start with you know some of the basic tools, that is under the descriptive analytics.

(Refer Slide Time: 03:49)



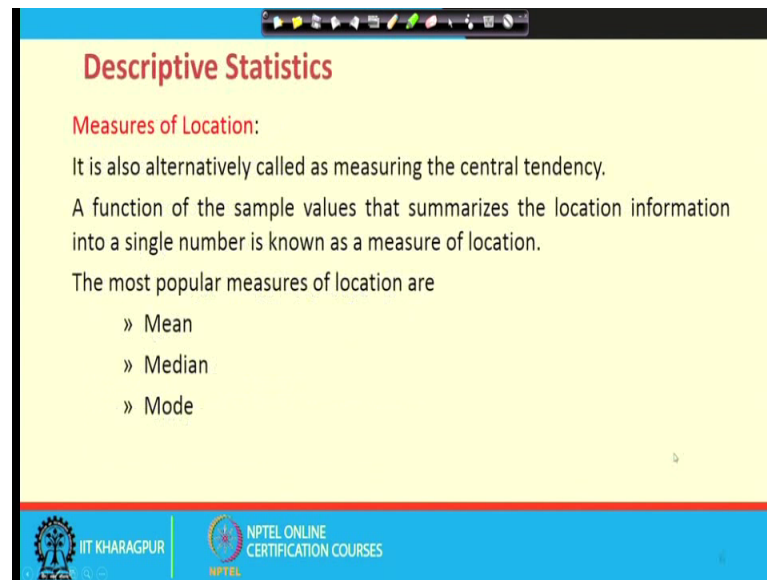
So, let me highlight what are the components we are going to discuss in this unit. So, that means, the idea is that you know what should be the unit coverage for this discussion. So, the coverage will be measures of location, measures of dispersion, measures of shape excel descriptive statistics tool, descriptive statistic for group data, descriptive statistics for categorical data, measures of associations and then something the basic requirements of probability and probability distributions. And then sampling with sampling distributions. So, that means, basically these are the coverage which will cover in this particular unit. And for that you must know the kind of you know problem. And to analyze the problem you must actually you know recognize the variables decision variables, and then you must have a data corresponding to each variables.

Depending upon the data or information corresponding to a particular variable, and your problems requirement or problem kind of you know structure. So, you usually or you like to use a particular you know tool, and then you can analyze the problem. In fact, descriptive analytics is not a kind of you know complex process, or you know kind of you know complex kind of you know environment. So, here you are supposed to know something so that you know that can be analyzed in a much better way through predictive analytics and prescriptive analytics. So, that is why first you know the descriptive structures, then we will go for predictive structure and prescriptive structure.



In some instances, the descriptive analytics you know can also help to analyze certain problems, and you may be in a position to get some kind of you know management decisions.

(Refer Slide Time: 05:39)



**Descriptive Statistics**


**Measures of Location:**


It is also alternatively called as measuring the central tendency.

A function of the sample values that summarizes the location information into a single number is known as a measure of location.

The most popular measures of location are

- » Mean
- » Median
- » Mode

 IIT KHARAGPUR

 NPTEL ONLINE CERTIFICATION COURSES

So, let us start with the kind of you know discussion. So, the first component under the descriptive statistic is a measure measures of location. And there are various ways of you know measuring this particular you know locations. So, standard 3 measures are you know mean median and mode, and this is or this otherwise called as a central tendency. And the entire particular structure is also called as you know summary of statistics.

So, here so, the idea is a you have a variables and corresponding to variable you have a plenty of you know information's; that is what we call as a sample observations. And you are supposed to know what is the central point of this particular you know spreadsheet or you know on data. So, for instance suppose you have actually 100 data points for a particular variables. And it is mandatory to know or it is the requirement of you know business analytics to know what is the average of this particular series, what is the maximum of that particular size, what is the minimum of that particular series of course, we have already discussed slightly in the use of you know excel sheet that is excel spreadsheet. But still you are supposed to know, whether you are using excel spreadsheet or any kind of you know statistical software's. So, the standard requirement is you should know the you know description about a particular variable, and the description of

a particular variable can be analyzed with respect to the descriptive statistic. That is what it is called as you know descriptive analytics.

So now these tools are frequently used to know the or to understand the nature of the data for a particular variables. So, here we are discussing about the particular you know measures. And the measures are nothing but called as you know a measures of you know central tendency. And the standard measures are called as you know mean median and mode.

(Refer Slide Time: 07:32)

**Measures of Location**

Arithmetic Mean

- ▶ For a population of size  $N$ :  
$$\mu = \frac{\sum_{i=1}^N x_i}{N}$$
- ▶ For a sample of  $n$  observations:  
$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$
- ▶ Excel function: `=AVERAGE(data range)`
- ▶ Property of the mean:  $\sum_i (x_i - \bar{x}) = 0$
- ▶ **Outliers** can affect the value of the mean.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, the these are all very simple, but you know still you have to know and you have to calculate and you have to report. So, that you know it will give you some kind of you know shape and some kind you know structure. So, since it is a kind of you know empirical kind of you know investigation process, or statistical kind of you know investigation process. So, always the game is with respect to population and sample. Population means it is say you know big set and sample is a part of it.

So, our entire structure or the problem discussion or problem analysis depends upon you know sample space particularly you know sample case only. So, it is sometimes you know very difficult to catch population the population, means the entire data for a particular variable which may not be feasible in a in a real-life scenario. So, that is why most of the instances we use you know samples and in the sample data, you know you have to analyze as per the kind of you know problem requirement. So, in the standard

measures of you know you know kind of you know mean is like this. So, it is something call as a sum of the total observations sum of the totals divided by you know sum of the total observations.

For instance,  $x$  is a variables. So, different items or different numbers are available with respect to  $x$ . And then we like to know what is the average of this item. So, just you know the simple structure is just add up all these variables, means values of the variables and then divide by number of observations. Then you will get actually simple statistic that is what it is called as actually mean and a whether it is a population whether it is sample. So, this the formula or the structure is almost all same. So, you have to just addons I mean say adding all these, but you know items. So, that is nothing but some of the values of the variables, and then divided by number of observations.

Then after that you will get a component called as you know average. So, here actually in the excel. So, this is the operation you have to apply. Just to go to the end of the data set and then put equal to signs and give the a command of you know average. And then specify the range particular range if you specify the entire range then it will give you the mean of the entire range if you specify mean of a particular range then it will also give you mean of a particular range. So, that means, it is the beauty of you know the particular statistics and the beauty of you know excel spreadsheet or any kind of you know statistical software. So, it will give you different kind of you know structures or the idea is it to find out the central point, and that can describes the entire things as per the a requirement.

But by the way so, the these are all typical futures when we will reporting the measures of you know locations. So, the mean should be actually in such a way that you know any day some of the deviation of the mean from the a you know observations should be equal to 0. And sometimes this may affect this particular you know feature, if there is a presence of outliers. In fact, we have discussed the concept of outlier in the unit 2. So, outlier means it is the data point which is a highly distance from other data point for instance. Suppose, there is a variable and the data entries are all are in single digit and double digit. But one data point may be in triple digit or you know or like that. Then in that case there is high chance that you know mean will not be stable. So, it will affect the particular process.

So, likewise so, there so, the likewise there are you know lots of other measures.

(Refer Slide Time: 11:11)

**Measures of Location**

Computing Mean Cost per Order

Using formula:  $\bar{x} = \frac{\sum_{j=1}^n x_j}{n}$

Mean =  $\$2,471,760/94$   
 $=$   
 $\$26,295.32$

Supplier	Order No	Item No.	Item Description	Item Cost	Quantity	Cost per order	AP Terms	Month/Order Date	Arrival Date
SpaceTime Technologies	A0111	6489	O-Ring	\$ 3.00	900	\$ 2,700.00	25	10/10/11	10/10/11
Steelgin Inc.	A0115	5319	Shielded Cable/ft.	\$ 1.10	17,500	\$ 19,250.00	30	08/20/11	08/31/11
Steelgin Inc.	A0123	4312	Bot-nut package	\$ 3.75	4,250	\$ 15,937.50	30	08/25/11	09/01/11
Steelgin Inc.	A0204	5319	Shielded Cable/ft.	\$ 1.10	18,500	\$ 20,350.00	30	09/15/11	10/05/11
Steelgin Inc.	A0205	5677	Side Panel	\$195.00	120	\$ 23,400.00	30	11/02/11	11/13/11
Steelgin Inc.	A0207	4312	Bot-nut package	\$ 3.75	4,200	\$ 15,750.00	30	09/01/11	09/10/11
Alum Sheeting	A0223	4224	Bot-nut package	\$ 3.95	4,500	\$ 17,775.00	30	10/15/11	10/20/11
Alum Sheeting	A0433	5417	Control Panel	\$255.00	500	\$ 127,500.00	30	10/20/11	10/27/11
Alum Sheeting	A0443	1243	Airframe fasteners	\$ 4.25	10,000	\$ 42,500.00	30	08/08/11	08/14/11
Alum Sheeting	A0446	5417	Control Panel	\$255.00	406	\$ 103,530.00	30	09/01/11	09/10/11
SpaceTime Technologies	A0533	9752	Gasket	\$ 4.05	1,500	\$ 6,075.00	25	09/20/11	09/25/11
SpaceTime Technologies	A0555	6489	O-Ring	\$ 3.00	1,100	\$ 3,300.00	25	10/05/11	10/10/11

IIT KHARAGPUR NPTEL ONLINE CERTIFICATION COURSES

So, this is actually standard examples this is a kind of you know purchase order data. And then. So, these are all entry. So, we have already discussed the excel kind of you know entry the kind of you know understanding all these things are there. So now, for instance with here the requirement is you mean cost per orders. So, that; so, here if we you know go through g columns, then this will give you in a cost of cost per order. So, then you are interested to know what is the mean? Cost per orders, just you go to the you know end of this particular series and put equal to sign. And then ask the excel to give the average, and then a by default you will get the value of this average is like this right.

In fact, mathematical you can calculate, but spreadsheet by default will give you such value. And by this particular process you can able to know what is the structure of the data. And then you can also verify the data from the center you know you know actual point to a mean point. So, if the sum of this particular deviation is coming 0. So, that means, there is a high accuracy in this data. And that data can be you know analyze in a nice way to represent something.

(Refer Slide Time: 12:34)

**Weighted mean**

It is also called weighted arithmetic mean or weighted average.

**Definition: Weighted mean**

When each sample value  $x_i$  is associated with a weight  $w_i$ , for  $i = 1, 2, \dots, n$ , then it is defined as

$$\bar{x} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$$

**Note :** When all weights are equal, the weighted mean reduces to simple mean.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

And then so likewise so, you have actually another kind of you know structure here a called as you know weighted mean. So, like it is a it is a part of you know mean structure. And here the idea is summation of  $x$  divided by  $n$  instead of doing that. So, here it is the summation of you know  $w \times x$  divided by summation  $w$ . So, that basically it is nothing but actually a like you know frequency distributions.

So, that means, earlier we are you know adding all the values of the variable and dividing by some of the other observations and to know, what is the center of this particular you know spreadsheets. So, this will give you the you know you know idea, what is the shape of this particular you know data or the position of the data. Your problem may be different, but you know let us say the example is you know it is a kind of you know cells data. In various locations for a particular point of time. So, you like to know what is the average cells in these locations. So, then accordingly you can actually get to know which particular location has a high cells which can which location as a low cells, and then after reporting the mean you can find out the particular you know difference.

If the difference is a high or you know low, then the then the kind of you know problem you know requirement is also analyzed accordingly. So now, this is another kind of you know structure where you can also calculate the mean and this is nothing but called as a weighted average. And here is agonist each variable you can actually give some kind of

you know weight factors. So, as a result after you know adjust with you weight, then you can calculate the a means. So, this is this and this is sometimes you know you know this is one kind of you know requirement for some kind of you know investigation process.

(Refer Slide Time: 14:19)

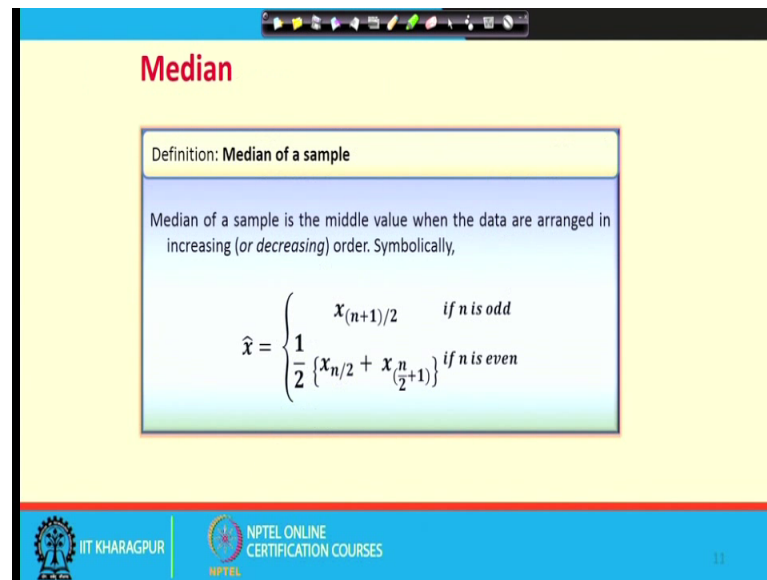
**Other measures of mean**

<p>» Arithmetic Mean (AM)</p> <ul style="list-style-type: none"> <li>- <math>S: \{x_1, x_2\}</math></li> <li>- <math>\bar{x} = \frac{x_1 + x_2}{2}</math></li> <li>- <math>\bar{x} - x_1 = x_2 - \bar{x}</math></li> </ul>	<p>Harmonic Mean (HM)</p> <ul style="list-style-type: none"> <li>- <math>S: \{x_1, x_2\}</math></li> <li>- <math>\hat{x} = \frac{2}{\frac{1}{x_1} + \frac{1}{x_2}}</math></li> <li>- <math>\frac{2}{\hat{x}} = \frac{1}{x_1} + \frac{1}{x_2}</math></li> </ul>
<p>» Geometric mean (GM)</p> <ul style="list-style-type: none"> <li>- <math>S: \{x_1, x_2\}</math></li> <li>- <math>\tilde{x} = \sqrt{x_1 \cdot x_2}</math></li> <li>- <math>\frac{x_1}{\tilde{x}} = \frac{\tilde{x}}{x_2}</math></li> </ul>	

NPTEL ONLINE CERTIFICATION COURSES

So, then and there is other measures of you know means, like you know arithmetic mean, harmonic means, then geometric means. So, these are all you know different ways of you know reporting the mean, but ultimately the idea behind this particular structure is a you have to find out the center points. And then you would like to check how did you know, you know values of the variables are you know, means what is the exact difference values of the values of the each variable or you know each data point from the mean data points.

(Refer Slide Time: 14:51)



**Median**

Definition: Median of a sample

Median of a sample is the middle value when the data are arranged in increasing (or decreasing) order. Symbolically,

$$\hat{x} = \begin{cases} x_{(n+1)/2} & \text{if } n \text{ is odd} \\ \frac{1}{2} \{x_{n/2} + x_{(n/2+1)}\} & \text{if } n \text{ is even} \end{cases}$$

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, this is another kind of you know measures, another kind of you know measures. So, where you know we like to know what is the middle of the particular you know series.

So, median is the kind of you know figure which can divide the series into 2 equal parts. What the mean may not divide it into equal parts, but it will give you a center points you know through which the entire you know variables. Can you know lying some may be high some may be low, but it will give you a specification where you know 50 percent of data will be left side, and 50 percent will be our right side or 50 percent will be up and 50 percent will be down. So, this is how the kind of you know structure you have to follow. And this is the formula which you can you know use to calculate the median. So, if it is the series is a can say, or then simply divide by 2. So, then the you can get the value you know you know value of the medians. Even if it is even number, then you can actually divide you know the nearest 2 values. And you know means adding nearest 2 value then divided by a 2.

So, then you can get the a value of the median. So, this is another measure of you know kind of you know location and for group data.

(Refer Slide Time: 15:59)

**Median of a sample**

Definition : Median of a grouped data

Median of a grouped data is given by

$$\hat{x} = l + \left( \frac{\frac{N}{2} - cf}{f} \right) h$$

where  $h$  = width of the median class  
 $N = \sum_{i=1}^n f_i$   
 $f_i$  is the frequency of the  $i^{th}$  class, and  $n$  is the total number of groups  
 $cf$  = the cumulative frequency  
 $N$  = the total number of samples  
 $l$  = lower limit of the median class

**Note**  
A class is called **median class** if its cumulative frequency is just greater than  $N/2$

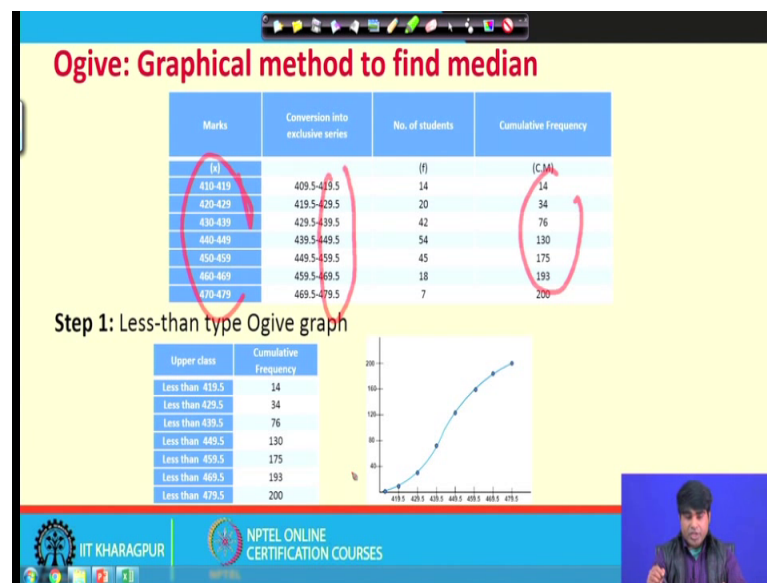
IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, you know for instance something kind of you know interval kind of you know situation. So, then you know you can use this particular formula. But ultimately, we are not here to mathematically calculate all these value, but this is a I am just showing that you know for every kind of you know tools. You know, analytics tools descriptive analytic tools, you have actually mathematical formula and software by default will give you all these values. But directly, and then you can get this values of the variable or values of the statistics. And then with the help of these you know statistics you can analyze the problem.

So, this this is a kind of you know formula, and here this formula is simply used for you know calculating medians. So,  $l$  stands for you know lower limit of this series, then  $n$  stands for sample observations, then  $f$  is the frequency a frequency, and a  $c f$  is the cumulative frequency and  $h$  is nothing but called as you know width of the median class. So, then accordingly you can calculate. I will take a example then I will show you how to calculate in the excel sheet, right.



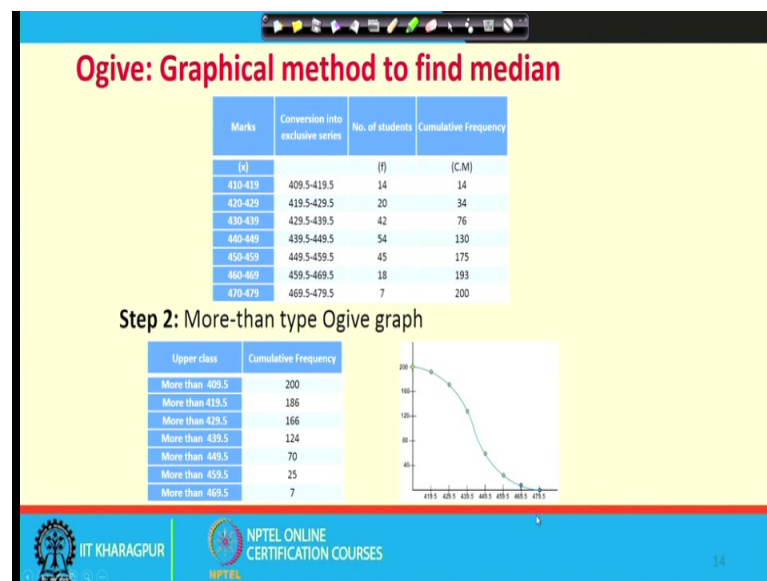
(Refer Slide Time: 17:05)



So, this is how this speak space specific examples, what I have already mentioned. So, this is the kind of you know data structure. So now, here this is the range of the data. And you know you can actually put lower range and upper range. Then you can streamline the particular process, and this is what the frequency, and this is cumulative frequency. And the particular structure can be represent graphically also graphically to find out the medians that is nothing but called as you know ogive. And then we can do throw less than methods, and then by you know kind of you know more than methods.

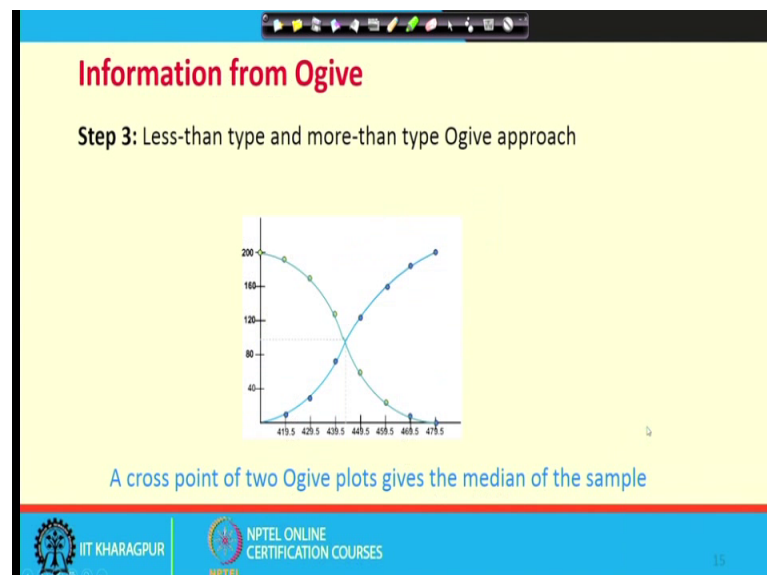
In the less than methods you are supposed to check you know this part of the series. And then the cumulative frequency. So, then you if you plot then the figure will be coming like this, then the other side if you go if you go to the other part, that is the more than type methods.

(Refer Slide Time: 18:02)



So, you have to take the lower limit this this part this lower part. And then and then you have to use the cumulative frequency. Again, you can you know with respect to this range you can you know plot this particular you know items. So, then you will get the figure like this, then if you joined the or you know the first one and the second one.

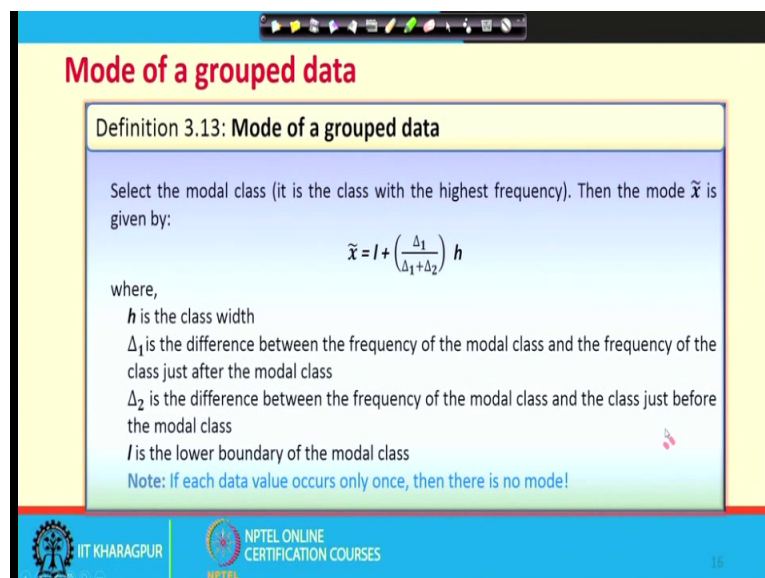
(Refer Slide Time: 18:24)



Then the intersection point will give you something called as you know or median. So, the intersection point will give you something called as you know medians.

So now so, these are the things you can actually, you can report from this you know descriptive statistics.

(Refer Slide Time: 18:45)



**Mode of a grouped data**

**Definition 3.13: Mode of a grouped data**

Select the modal class (it is the class with the highest frequency). Then the mode  $\tilde{x}$  is given by:

$$\tilde{x} = l + \left( \frac{\Delta_1}{\Delta_1 + \Delta_2} \right) h$$

where,

- $h$  is the class width
- $\Delta_1$  is the difference between the frequency of the modal class and the frequency of the class just after the modal class
- $\Delta_2$  is the difference between the frequency of the modal class and the class just before the modal class
- $l$  is the lower boundary of the modal class

Note: If each data value occurs only once, then there is no mode!

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

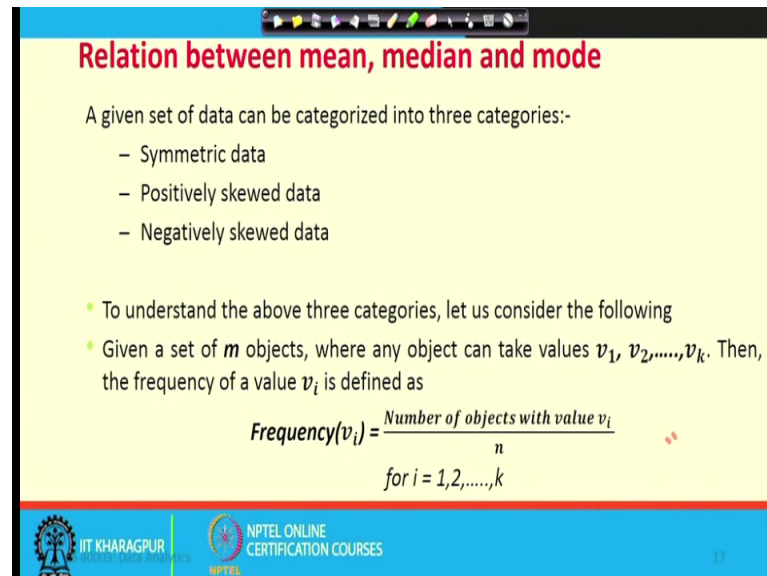
And then accordingly can actually analyze the particular problem. This is another kind of you know measures called as you know, mode and highest frequency distribution or the highest frequency in a particular you know obser you know particular you know set, data set can give you the kind of you know mode representations. So, like you know mean and medians mode is also kind of you know measures which can describes you know the values of the variables, you know that that is specifically again you know a you know indicate to you that you know what is the distance of you know individual points from the central point.

So, this is the formula which you can you know use for calculating mode. And in fact, you know when we have a data and with respect to time or any kind of you know cross sectional unit. So, then you know the idea is you just you have to visualize the data by different plotting. And then here the idea is you have to check the kind of you know distributions. Usually we follow a kind of you know normal distributions, whether the data is follows a normal distributions or not normal distribution. So, we have a lots of you know distribution which you discuss you know after some time, but you know usually and if the data is you know in the kind of you know normally distributed, and then it will give you some kind of you know consistency results. And most of the things

can be in a in a right path. So, that is how we are always try to do some kind of you know adjustment. So, that you know the data should be normally distributed.

If not actually normally distributed, then it will be called as you know skewed data.

(Refer Slide Time: 20:19)



**Relation between mean, median and mode**

A given set of data can be categorized into three categories:-

- Symmetric data
- Positively skewed data
- Negatively skewed data

• To understand the above three categories, let us consider the following

• Given a set of  $m$  objects, where any object can take values  $v_1, v_2, \dots, v_k$ . Then, the frequency of a value  $v_i$  is defined as

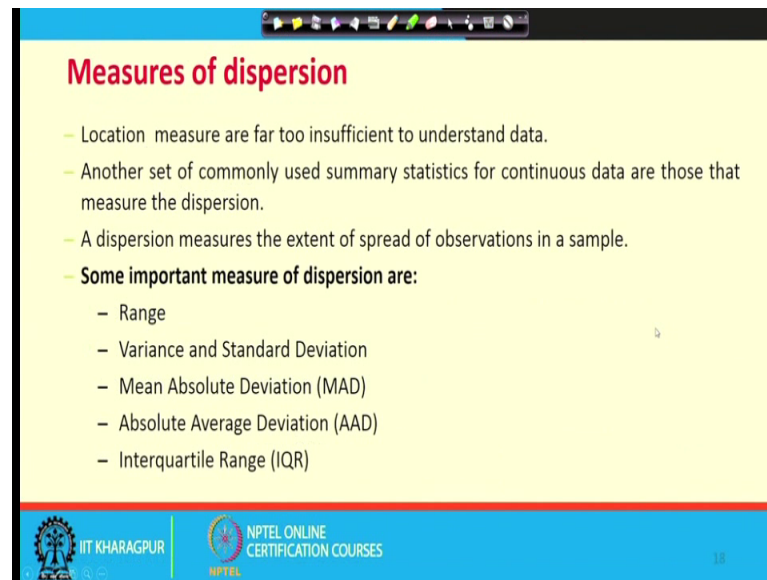
$$\text{Frequency}(v_i) = \frac{\text{Number of objects with value } v_i}{n}$$

for  $i = 1, 2, \dots, k$

NPTEL ONLINE CERTIFICATION COURSES

It may be right skewed it may be left skewed. So, for instance here the idea is a symmetrical data. So, this is what actually when the data is asymmetrical means; so, it is actually equally kind of you know distribution. If you put you know mean median modes that will be coincide, then you know the data will be equally spread. So, but if not then you know few data will be more in the right side or few data will be more in the left side. If it is more in the right side, then it will call as a positive skewed or it if not then it will be called as you know left skewed or you know negatively skewed right. So, this is how the particular structure through which you can you know calculate all these mean median and mode.

(Refer Slide Time: 21:04)



**Measures of dispersion**

- Location measure are far too insufficient to understand data.
- Another set of commonly used summary statistics for continuous data are those that measure the dispersion.
- A dispersion measures the extent of spread of observations in a sample.
- **Some important measure of dispersion are:**
  - Range
  - Variance and Standard Deviation
  - Mean Absolute Deviation (MAD)
  - Absolute Average Deviation (AAD)
  - Interquartile Range (IQR)

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, the other parts of you know this particular you know measures of location easy measures of dispersion. You know so, that means, in the descriptive analytics. So, one part is a measures of you know location, this is a measure of dispersion; that means, you like to know what is the kind of you know structure of data once against to analyze this problem. So, there are various you know tools are there is like earlier mean median mode we have here range variant standard deviations, mean absolute deviations absolute average deviation interquartile range so many things are there. And here the idea is again to non-the spread of this particular you know series the spread of that particular series will give you enough exposure to understand the problems, and understand the variable and then you can predict the particular you know environment right.

(Refer Slide Time: 21:55)

### Measures of dispersion

**Example :** Suppose, two samples of fruit juice bottles from two companies **A** and **B**. The unit in each bottle is measured in litre.

Sample A	0.97	1.00	0.94	1.03	1.06
Sample B	1.06	1.01	0.88	0.91	1.14

- Both samples have same mean. However, the bottles from company A with more uniform content than company B.
- We say that the dispersion (or variability) of the observation from the average is less for A than sample B.
  - The variability in a sample should display how the observation spread out from the average
  - In buying juice, customer should feel more confident to buy it from A than B

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, this is a kind of you know slight examples. So now, particularly when there is a kind of you know comparison. So, because we are we are dealing with you know business related problems. And then you have to always have some kind of you know comparisons. Because we are living in the competitive you know environment. So, let us say there are 2 series here sample A and sample B A sample A and sample B. And in this case, we like to know which particular series is more consistent. Then you know this is the you know tools which you know this is the tool which you can use to solve this you know you know problems. Or here to justify which one is more consistent. So, just to find out the variations or variability. So, either through standard deviation or variance, then you can report actually what the variance is high and that is you know less consistent, and where the standard deviation is low so, it is called as a more consistent.

So, likewise you can you know solve some of the business problems depending upon the particular you know requirement.

(Refer Slide Time: 22:56)

**Range of a sample**

Definition : **Range of a sample**

Let  $X = x_1, x_2, x_1, \dots, x_n$  be  $n$  sample values that are arranged in increasing order.

The range  $R$  of these samples are then defined as:

$$R = \max(X) - \min(X) = x_n - x_1$$

Range identifies the maximum spread, it can be misleading if most of the values are concentrated in a narrow band of values, but there are also a relatively small number of more extreme values.

The variance is another measure of dispersion to deal with such a situation.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, this is another kind of you know, tool tools descriptive tools that is called as a range; which can give you the difference between maximum and minimum. If the difference is actually 0 so, that means, this particular you know data set will not be useful for any kind of you know investigation process. So, your range should be always actually some positive value some positive value. And then you know if it is equal to 0. So, then a analytic tools may not be a use for further kind of this is actually a mean component and this is the individual observations.

(Refer Slide Time: 23:29)

**Variance and Standard Deviation**

Definition: **Variance and Standard Deviation**

Let  $X = \{ x_1, x_2, x_1, \dots, x_n \}$  are sample values of  $n$  samples. Then, variance denoted as  $\sigma^2$  is defined as :-

$$\sigma^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

where,  $\bar{x}$  denotes the mean of the sample

The standard deviation,  $\sigma$ , of the samples is the square root of the variance  $\sigma^2$

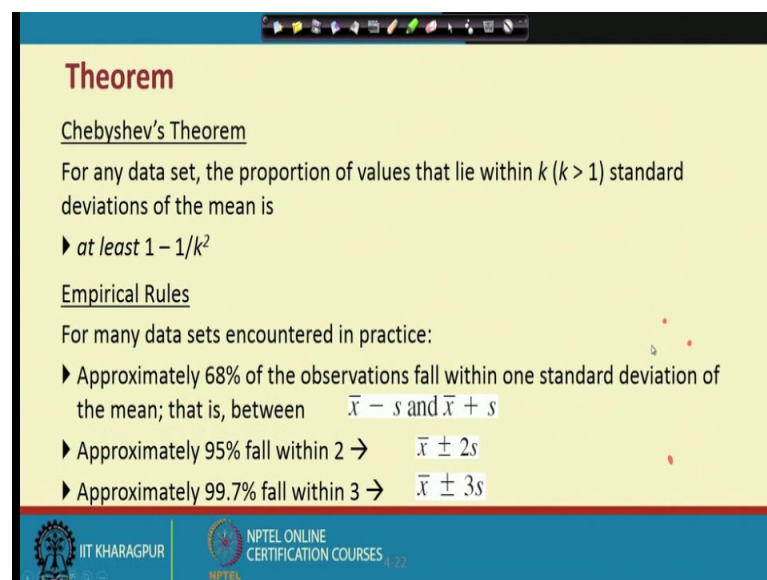
IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES



And squaring these you know distance from the actual point to mean point, and then adjusted with you know in a degree of freedom, that is the sample observations. And we like to know what is the kind of you know (Refer Time: 23:46).

If larger the deviations larger is the means lower is the consistency. And lower is the you know standard deviation higher is the consistency. But it should not be equal to 0 against. Like you know range should not be equal to 0, standard deviance deviations should not be also equal to 0 if standard deviation equal to 0. Then again, this this particular data set may not be used for any kind of you know investigation process. So, you must be very careful how you have to deal with these problems, right.

(Refer Slide Time: 24:19)



**Theorem**

Chebyshev's Theorem

For any data set, the proportion of values that lie within  $k$  ( $k > 1$ ) standard deviations of the mean is

- ▶ at least  $1 - 1/k^2$

Empirical Rules

For many data sets encountered in practice:

- ▶ Approximately 68% of the observations fall within one standard deviation of the mean; that is, between  $\bar{x} - s$  and  $\bar{x} + s$
- ▶ Approximately 95% fall within 2 →  $\bar{x} \pm 2s$
- ▶ Approximately 99.7% fall within 3 →  $\bar{x} \pm 3s$

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES 4.22

So, this is a kind of you know you know theorems when we go for some kind of you know investigation process by using data. So, then you know you are interested to know the particular you know distributions, then sometimes you know. So, the data can be actually normally distributed. And the theorem says that you know the proportion of values that lie within you can say  $k$ ; where  $k$  greater than 1, and that to standard deviation of the mean. That is nothing but actually 1 minus 1 by  $k$  squares. So, that means, actually most of the observations you know follows like this. So, if approximately 60 percent, then it will be a kind of you know interval. Either otherwise it is a 95 percent interval, or it is a 99 percent kind of you know.



So, it depends upon you know mean with you know, standard deviation only mean with one standard deviation, mean with the 2-standard deviation, and mean with the 3-standard deviation. Mean with the plus minus 1 standard deviation means, 68 percent data will be in the particular diligence. Otherwise if it is mean plus minus 2 standard deviation means, 95 percent of the data will be lying. Then mean plus minus 3 standard deviation means 99 percent data will be lying in these particular structures.

(Refer Slide Time: 25:33)

**Coefficient variation**

**Basic properties**

- $\sigma$  measures spread about mean and should be chosen only when the mean is chosen as the measure of central tendency.
- $\sigma = 0$  only when there is no spread, that is, when all observations have the same value, otherwise  $\sigma > 0$

**Coefficient variation**

A related measure is the coefficient of variation **CV**, which is defined as follows

$$CV = \frac{\sigma}{\bar{x}} \times 100$$

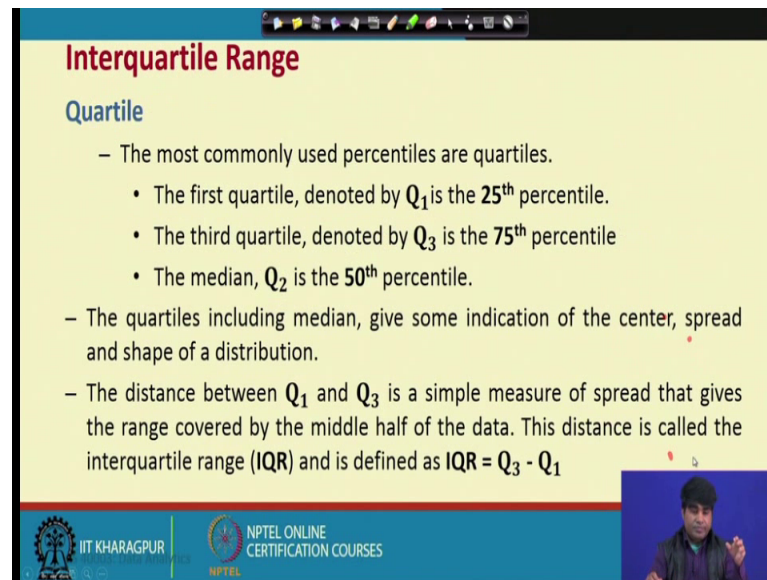
This gives a ratio measure to spread.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, this is another measure of you know dispersions coefficient variations, this is actually relative measure, and the entire structure can be actually you know you know transfer into a percentage format. And then you can make a some kind of you know comparative analysis. It is just you know the ratio between standard deviation to mean followed by you know percentage, right.

So, this will be very easy to have some kind of you know comparative analysis.

(Refer Slide Time: 26:00)



**Interquartile Range**

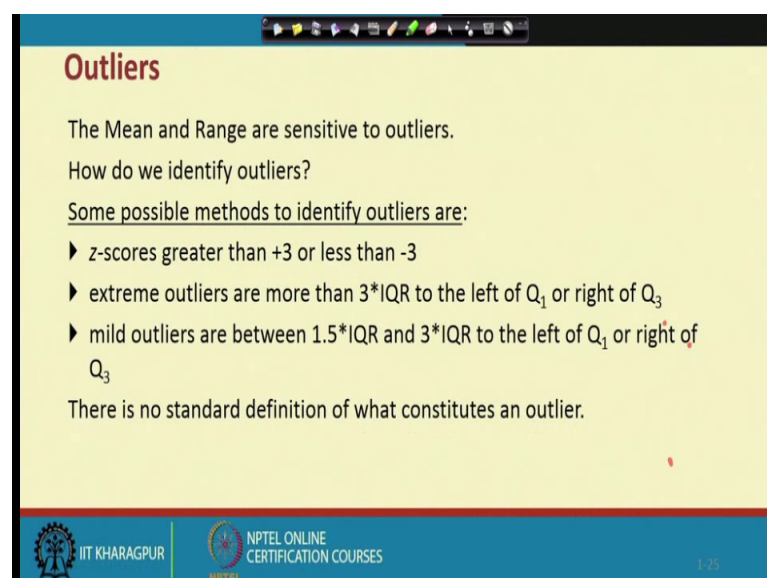
**Quartile**

- The most commonly used percentiles are quartiles.
  - The first quartile, denoted by  $Q_1$  is the 25<sup>th</sup> percentile.
  - The third quartile, denoted by  $Q_3$  is the 75<sup>th</sup> percentile
  - The median,  $Q_2$  is the 50<sup>th</sup> percentile.
- The quartiles including median, give some indication of the center, spread and shape of a distribution.
- The distance between  $Q_1$  and  $Q_3$  is a simple measure of spread that gives the range covered by the middle half of the data. This distance is called the interquartile range (IQR) and is defined as  $IQR = Q_3 - Q_1$

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Then the next measure is the interquartile may interquartile range, and that is the difference between third quartile and the first quartile. So, when you are you know plotting the data. So, you have a 25 percent data, that is the first quartile then 50 percent of data that is the second quartile, then third quartile is the 75 percent of data. So, that means, say so if we actually data is equally spread, then you know this will be actually give you some kind of you know better structures. So, interquartile range that is nothing but called as IRQ. So, we will give you some kind of you know snap shot whether the data is a well spread or not well spread.

(Refer Slide Time: 26:46).



**Outliers**

The Mean and Range are sensitive to outliers.

How do we identify outliers?

Some possible methods to identify outliers are:

- ▶ z-scores greater than +3 or less than -3
- ▶ extreme outliers are more than  $3 \times IQR$  to the left of  $Q_1$  or right of  $Q_3$
- ▶ mild outliers are between  $1.5 \times IQR$  and  $3 \times IQR$  to the left of  $Q_1$  or right of  $Q_3$

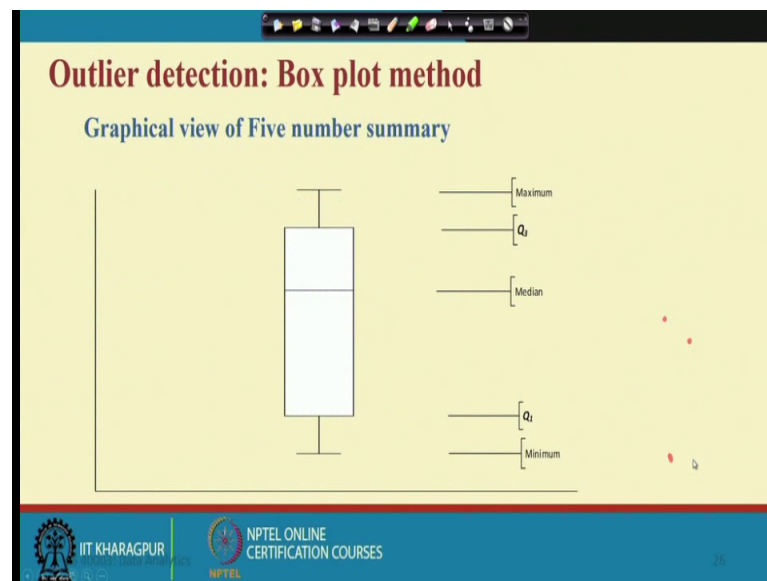
There is no standard definition of what constitutes an outlier.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

1-25

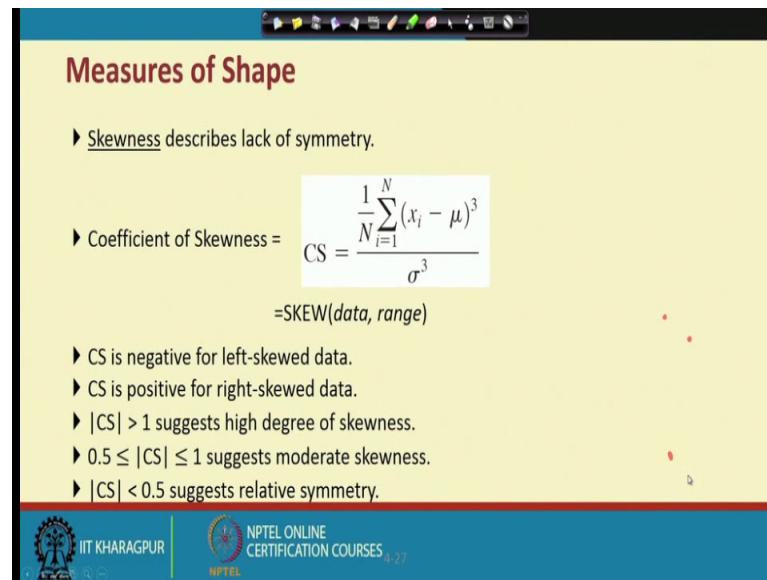
So, this this will be this will be helping lot again for further kind of you know investigations. Then if there is a outliers the outliers will be will be problematic to you know make the kind of you know statistics more consistent. So, when will be when we like to use all these you know statistic to report something. So, be careful that you know outliers should not be in the process. If there is outlier you try to remove the outlier, and then analyze the problem. Otherwise the entire result will be inconsistent altogether.

(Refer Slide Time: 27:14)



So, this is the standard examples of outliers by box diagram you can find here the median is the location is here, but you know a high point height you know structure this side and lower structure this side. So, that means, outlier will affect the particular you know process. So, what is the best requirement that you know remove the outlier, then against you know restructure the process, right.

(Refer Slide Time: 27:35)



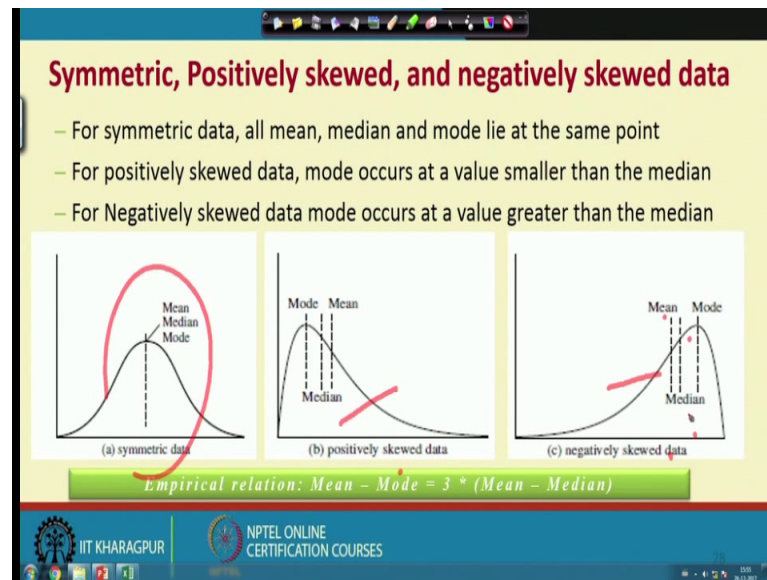
**Measures of Shape**

- ▶ Skewness describes lack of symmetry.
- ▶ Coefficient of Skewness = 
$$CS = \frac{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^3}{\sigma^3}$$
  
=SKEW(data, range)
- ▶ CS is negative for left-skewed data.
- ▶ CS is positive for right-skewed data.
- ▶  $|CS| > 1$  suggests high degree of skewness.
- ▶  $0.5 \leq |CS| \leq 1$  suggests moderate skewness.
- ▶  $|CS| < 0.5$  suggests relative symmetry.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES 4.27

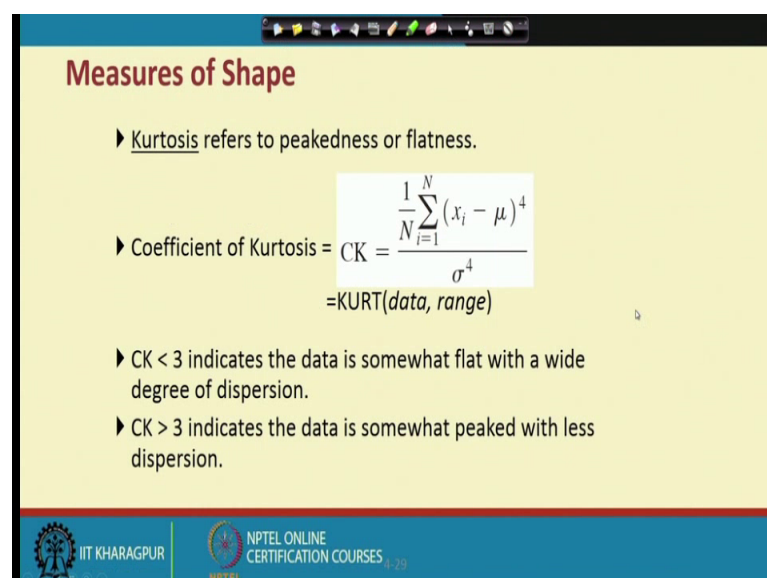
So, this is the third measures of you know descriptive analytics. So, that is called as a measures of you know says shape. And here there are 2 type of you know requirement that is called as a skewness and kurtosis. So, skewness you know it will give you actually spreadness of a particular you know distribution, which I have already discussed about the concept called as a normal distribution. And when this you know distribution is a symmetrical then, and then you know mean median mode will be coincide. If the distribution is not actually symmetrical, then it will be either right skewed or you know left skewed. And that that is what actually the structure called as you know shape of the distributions. So now, skewness and kurtosis will give you some kind of you know better exposure to know these shape of this distribution.

(Refer Slide Time: 28:22)



So, this is the standard examples. And here is this one is actually called as a equal spread. And this is actually a not skewed. And this is in actually not skewed. This is called as a positively skewed. And this is called as you know negatively skewed. So, that means, this is not well distribution, but this is actually perfect distribution. And when your data will be like this, then your analysis will be very perfect, and your findings will be very perfect, and you can have a perfect decision. So, if your data is not actually very accurate, then it will give you some kind of you know wrong decision. So, this is a another measure of you know measures of you know shape.

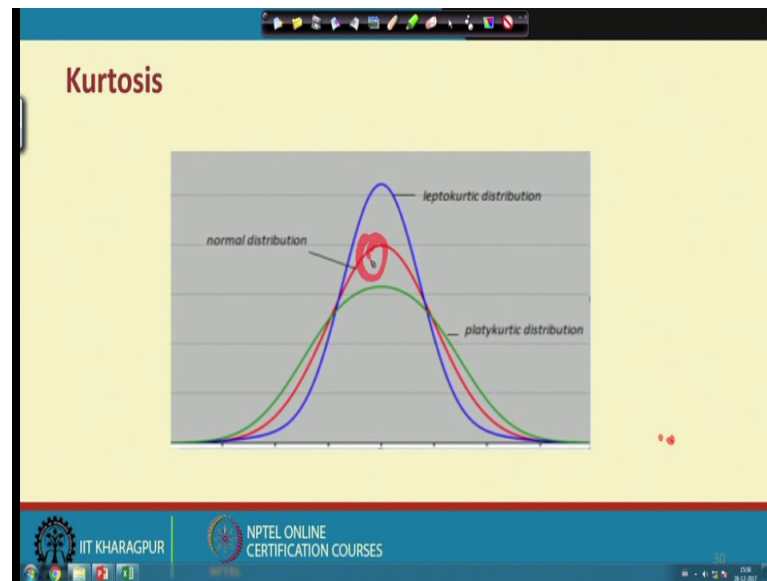
(Refer Slide Time: 28:59)



And same as you know excel will help you lot to get these to get these values of the you know variables.

So now like you know skewness kurtosis will be also give you some kind of you know exposure to know the shape of the distributions.

(Refer Slide Time: 29:22)



So, the structure will be this structure will be like this it is kind of you know flatness of the curve, right. So, this is actually a these are all various shapes of the distribution. And kurtosis will give you some kind of you know shape like this, but; that means, this is actually the correct kind of you know specification. For instance; so, the red one, red one is actually the correct shape of the distribution. If your data will be in this particular structure then it will be actually having high kind of you know impact and the decision will be very perfect. Otherwise it will be little bit you know inconsistent kind of you know and decision-making process.

(Refer Slide Time: 30:03)



### Descriptive Statistics for Grouped Data

Example: Computing Statistical Measures from Frequency Distributions  
(Computer Repair Times)

	A	B	C	D	E	F
1	Computer Repair Times					
2						
3	Days (x)	Frequency (f)	Frequency*Days	Days - Mean	(Days - mean)*2	Frequency*(Days - Mean)*2
4	0	0	0	-14.912	222.368	0.000
5	1	0	0	-13.912	193.544	0.000
6	2	0	0	-12.912	166.720	0.000
7	3	0	0	-11.912	141.896	0.000
43	39	1	39	24.088	580.232	580.232
44	40	1	40	25.088	629.408	629.408
45	41	0	0	26.088	680.584	0.000
46	42	0	0	27.088	733.760	0.000
47	Sum	250	3728			8840.064
48	Mean		14.912	Variance		35.50226506

$$\bar{x} = \frac{\sum f x_i}{n}$$

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

NPTEL ONLINE  
CERTIFICATION COURSES

4-31

So, this is the standard examples.



(Refer Slide Time: 30:07)

### Descriptive Statistics for Grouped Data

Example: Computing Descriptive Statistics for a Grouped Frequency  
Distribution

We can use group midpoints as approximate percentages of household income spent on rent (except in rows 13, 14).

	A	B	C
1	Gross Rent as a Percentage of Household Income in 1999		
2	Source: US Census Bureau		
3			
4	Group	Number of Households	
5	Less than 10 percent	2,239,346	
6	10 to 14 percent	4,130,917	
7	15 to 19 percent	5,037,981	
8	20 to 24 percent	4,498,604	
9	25 to 29 percent	3,666,233	
10	30 to 34 percent	2,585,327	
11	35 to 39 percent	1,809,948	
12	40 to 49 percent	2,364,443	
13	50 percent or more	6,209,568	
14	Not computed	2,657,135	

NPTEL ONLINE  
CERTIFICATION COURSES

4-32

And then, and these are all you know various ways you have to report the descriptive statistics.



(Refer Slide Time: 30:11)

## Descriptive Statistics for Grouped Data

Example (continued)

Our calculations indicate that the typical renter spends about 30% of household income on rent.

	A	B	C	D	E	F	G
2	Group	Percent (n)	Number (n)	P * x	x - mean	(x - mean)^2	P(x - mean)^2
3	Less than 10 percent	5%	2,239,345	111967.30	-24.8645%	0.0618	138446.0126
4	10 to 14 percent	12%	4,130,917	495710.04	-17.8645%	0.0319	131834.1452
5	15 to 19 percent	17%	5,037,981	856456.77	-12.8645%	0.0165	83376.1701
6	20 to 24 percent	22%	4,498,604	989692.88	-7.8645%	0.0062	27823.9852
7	25 to 29 percent	27%	3,666,233	989882.91	-2.8645%	0.0008	3008.2636
8	30 to 34 percent	32%	2,585,327	827304.64	2.1355%	0.0005	1179.0089
9	35 to 39 percent	37%	1,809,948	669680.76	7.1355%	0.0051	9215.4310
10	40 to 49 percent	44.50%	2,364,443	1052177.14	14.6355%	0.0214	50645.9048
11	50 percent or more	60%	6,209,568	3726740.80	30.1355%	0.0908	563921.1249
12	Sum		32,542,367	9718613.24			1009450.0462
13							
14			Mean	29.86%		Variance	0.031019565
15						Standard Dev.	17.61%

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES 4:33

(Refer Slide Time: 30:12)

## Descriptive Statistics for Categorical Data: The Proportion

Example Computing a Proportion

► Proportion of orders placed by Spacetime Technologies

=COUNTIF(A4:A97, "Spacetime Technologies")/94

= 12/94 = 0.128

	A	B	C	D	E	F	G	H	I	J
1	Purchase Orders									
2	Supplier	Order No	Item No.	Item Description	Item Cost	Quantity	Cost per order	A/P Terms	(Month)	Order Date
4	Spacetime Technologies	A0111	6489	O-Ring	\$ 3.00	900	\$ 2,700.00	25		10/10/11
5	Sheelap Inc.	A0115	5319	Shielded Cable/Rt.	\$ 1.10	17,500	\$ 19,250.00	30		08/20/11
6	Sheelap Inc.	A0123	4312	Bolt-nut package	\$ 3.75	4,250	\$ 15,937.50	30		08/25/11
7	Sheelap Inc.	A0204	5319	Shielded Cable/Rt.	\$ 1.10	16,500	\$ 18,150.00	30		09/15/11
8	Sheelap Inc.	A0205	5677	Side Panel	\$195.00	120	\$ 23,400.00	30		11/02/11
9	Sheelap Inc.	A0207	4312	Bolt-nut package	\$ 3.75	4,200	\$ 15,750.00	30		09/01/11
10	Alum Sheeling	A0223	4224	Bolt-nut package	\$ 3.95	4,500	\$ 17,775.00	30		10/15/11
11	Alum Sheeling	A0433	5417	Control Panel	\$255.00	500	\$ 127,500.00	30		10/20/11
12	Alum Sheeling	A0443	1243	Airframe Fasteners	\$ 4.25	10,000	\$ 42,500.00	30		08/08/11
13	Alum Sheeling	A0446	5417	Control Panel	\$255.00	406	\$ 103,530.00	30		09/01/11
14	Spacetime Technologies	A0533	9752	Gasket	\$ 4.05	1,500	\$ 6,075.00	25		09/20/11
15	Spacetime Technologies	A0555	6489	O-Ring	\$ 3.00	1,100	\$ 3,300.00	25		10/05/11

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES 4:34



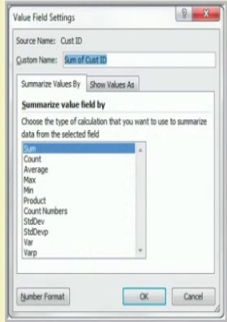
(Refer Slide Time: 30:13)

## Statistics in Pivot Tables

### Statistical Measure Choices in PivotTables

Under *Value Field Settings*:

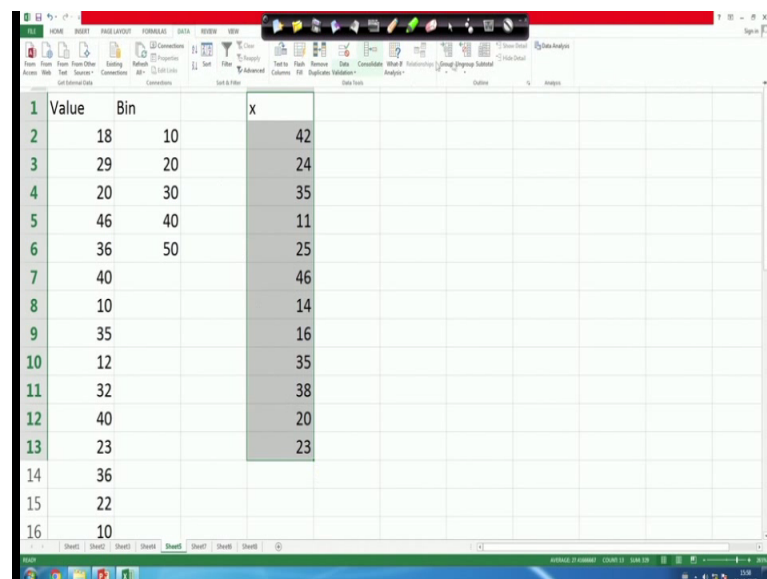
- ▶ Average
- ▶ Max and Min
- ▶ Product
- ▶ Standard deviation
- ▶ Variance



IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES 4:35

And so, let me let me take you to a examples, and show you how these actually the shape of the you know all these descriptive statistic can be reported. For instance let us say so, this is actually series here. And what will be do here?

(Refer Slide Time: 30:36)



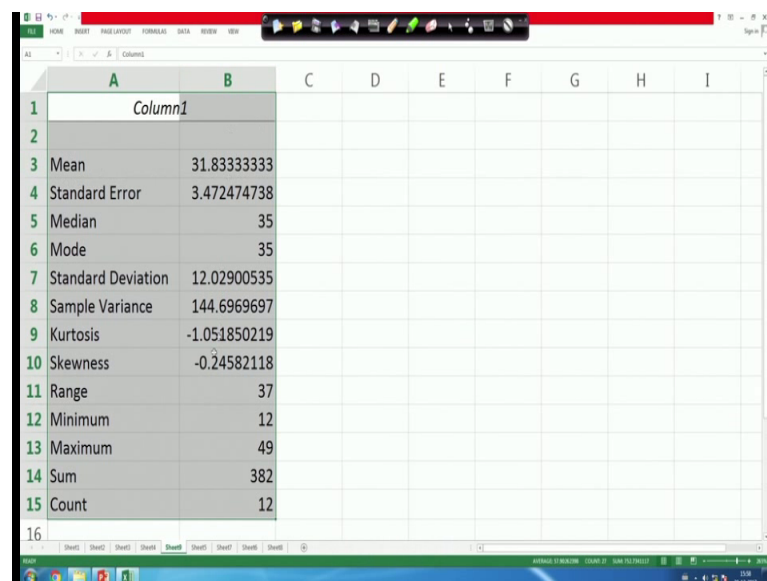
	Value	Bin	x
1			
2	18	10	42
3	29	20	24
4	20	30	35
5	46	40	11
6	36	50	25
7	40		46
8	10		14
9	35		16
10	12		35
11	32		38
12	40		20
13	23		23
14	36		
15	22		
16	10		

So, let it be big one. So, let us say this is a series here, I have actually randomly generated here. 10 to 10 to sorry, 15 to yes, 10 to 50 here. So, that means, it is just you create a let us say x is a variable here. And then you just randomly generated data something else. You know, a random data between you know say 10 to 50. And then

close the loop and make enters. So, this will be then you take some series of the data like you know let us say this much of series you have generated. And the range is between actually 10 to 50. Because it is a you know you know random data. And then our idea is just to know how the descriptive statistic is all about in this particular you know series.

So, what did it do actually? So, just you know highlights this data. So, highlights this data, and go to the you know data here. And then here there is actually data analysis package. And you just go to the data analysis package. And then here there is a component called as you know descriptive statistics. Just you put, and then by default it will ask you to put specify the range. For instance, our range is here, you know how much d 2 to say you know d like you know 13, the entire range. Then you just you know see here so many items are there. And then you just click summary statistics and then put.

(Refer Slide Time: 32:07)

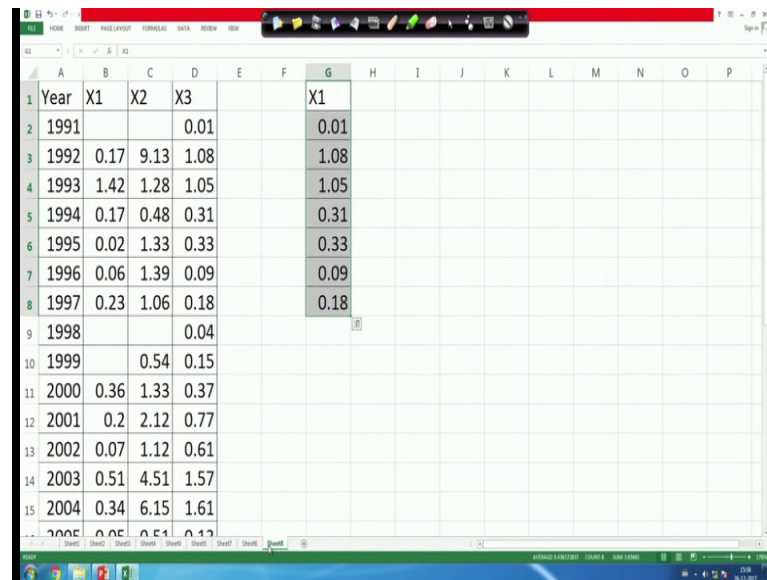


The screenshot shows an Excel spreadsheet with a summary statistics table. The table is located in the range A3:B15. The first column (A) lists the statistical measures, and the second column (B) shows the corresponding values. The data series is named 'Column1'.

	Column1
Mean	31.83333333
Standard Error	3.472474738
Median	35
Mode	35
Standard Deviation	12.02900535
Sample Variance	144.6969697
Kurtosis	-1.051850219
Skewness	-0.24582118
Range	37
Minimum	12
Maximum	49
Sum	382
Count	12

And by default, you will get you know standard results like this. So, just you know I will I will be enhance say. So, you can get to know actually. So, these are the actually descriptive statistic which you have already discussed. So, you know from the data set, this is this is actually variable x.

(Refer Slide Time: 32:29)



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
	Year	X1	X2	X3			X1									
1	1991			0.01			0.01									
2	1992	0.17	9.13	1.08			1.08									
3	1993	1.42	1.28	1.05			1.05									
4	1994	0.17	0.48	0.31			0.31									
5	1995	0.02	1.33	0.33			0.33									
6	1996	0.06	1.39	0.09			0.09									
7	1997	0.23	1.06	0.18			0.18									
8	1998			0.04												
9	1999		0.54	0.15												
10	2000	0.36	1.33	0.37												
11	2001	0.2	2.12	0.77												
12	2002	0.07	1.12	0.61												
13	2003	0.51	4.51	1.57												
14	2004	0.34	6.15	1.61												
15	2005	0.05	0.51	0.13												

And the original data set is a, original data set sorry, so this is the kind of you know statistics. And so, the star the values of the variable you know differing from 10 to 50, but you know the mean value is here 31. And standard error is this much median is it 35; that means, we have a we have calculated for you know some of the observations at 12 observations. So, median is showing 35, mode is 35 and mean is 31. And then standard deviation is at 12, then sample variance is 144.7 and kurtosis is minus 1.05 skewness is minus 0.24. And the range is 37 minimum is a 12 maximum is 49. And the total sum is it 382.

So, that means, these are you know basic statistic which we have reported through you know you know descriptive analytics. So, that means, the variable variables you know information's are readily available from you know particular point to particular point. So, you want to know what is the nature of this particular you know data corresponding to this particular variable. So, so these 2 this descriptive statistic will give you some kind of you know exposure to know about to know more about this particular you know variables. If the data structure is not perfectly, then it will not help you lot to take some kind of you know management decision. And a as a result descriptive statistics you know through descriptive statistic, you can get to know whether you know the data which is available for this particular variables or you know a few variables are consistent or not.

So, this is the reporting for a particular variable, if you have a series of variable, then every variable can be analyzed in you know with these you know test are distinct then descriptive analytics tools. So, in any kind of you know be you know problem investigation, when you have actually variables and the data. So, it is the standard requirement is that you know you have to report the descriptive statistic. And get to know details about these particular variables of origin of past data is a constant. We have discussed the concept called as you know and descriptive analytics, where you know to know what happened in the past.

So, that means so, this is the kind of you know summary statistics. And this summary statistic will give you some kind of you know snapshot what was actually happened in the past. So, that means, what is the average kind of you know structure, what is the maximum kind of you know structure, a minimum kind of you know structure, what is the variations or how is the kind of you know range, all these things are you know you know you can explores with the available you know historical data, that is what we called as you know past data.

Once you know all details, then you will be in a position to go for you know further kind of you know investigation through predictive analytics and prescriptive analytics. But it is the mandatory requirement, that you know you are supposed to know details about this particular variable so far as you know, data availability is concerns. So, whatever shape of the data does not matters, whether it is a 10 observation, or 100 observation, or 1,000 observation just you know just you know allow the kind of you know software to report the descriptive statistics. Then software will it by default will give you values of the descriptive statistics. And with the help of you know descriptive statistics you can analyze measure of location measures of dispersions and measures of shape.

So, that means, these are the 3 major requirements we are supposed to report in any kind of you know problem investigations. If you if you are you know aware about the measure of location measures of dispersion a measures of shape, and by default your 50 percent of the things are you know reported there. So, you are in your in a absolutely you are in a right position to describe the particular structure in a much better way. Until unless you know the kind of you know descriptive statistics. So, you are not in a position to pick up a you know complex kind of you know tools to investigate to investigate the problems in depth means more in depth, and to get you know better inference, and then

that may that may help you lot you know means a if you are you know exploring or you know all details by this kind of you know structure.

So, the further kind of you know investigations by using any complex analytical tools will be give you some kind of you know better inference, and better judgment, and then you know the decision will be actually more or less perfect and consistent. So, that is how the standard requirement is a for any kind of you know investigation probe you know a process, you are supposed to identify the variables, have the data, after having the data first you report the descriptive statistic, check the descriptive statistic, understand the situations, and after understanding the situation, then you look for you know further in in depth kind of you know investigation in a further kind of you know testing or something like that then you will be get some kind of you know better inference. And that will help you lot to you know to solve some of the problems. And that to take some kind of you know good management decisions. With this we will be stop here.

Thank you very much. Have a nice time.