

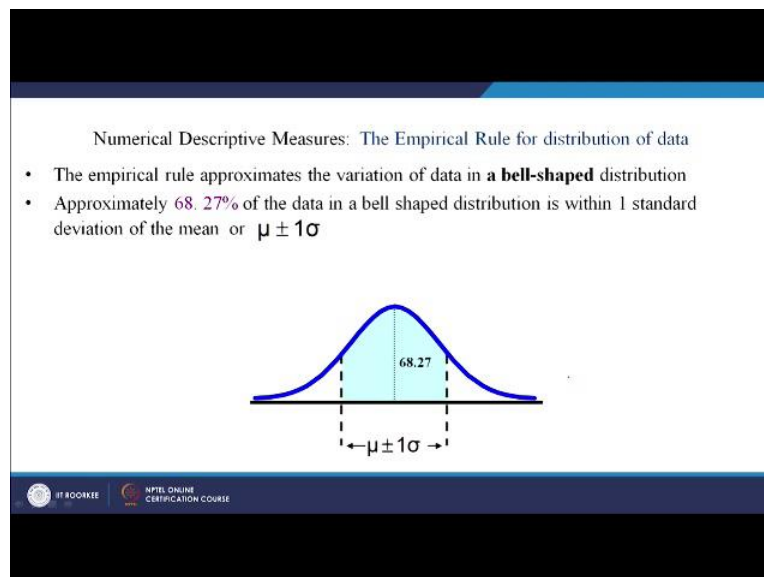
**Business Statistics**  
**Prof. M. K. Barua**  
**Department of Management Studies**  
**Indian Institute of Technology – Roorkee**

**Lecture – 11**  
**Numerical Descriptive Measures**

Good morning friends I welcome you all in this session as you are aware in previous session we discussed couple of exercises related to measures of Central tendency and measures of dispersion we also seen what is Z score and apart from Z score we have seen what are different population parameters and what are the population parameters. Population parameters are mean, population variance and population standard deviation.

Let us move on to today's session it is on numerical descriptive measures and we will see couple of insights about normal distribution curve.

**(Refer Slide Time: 01:18)**

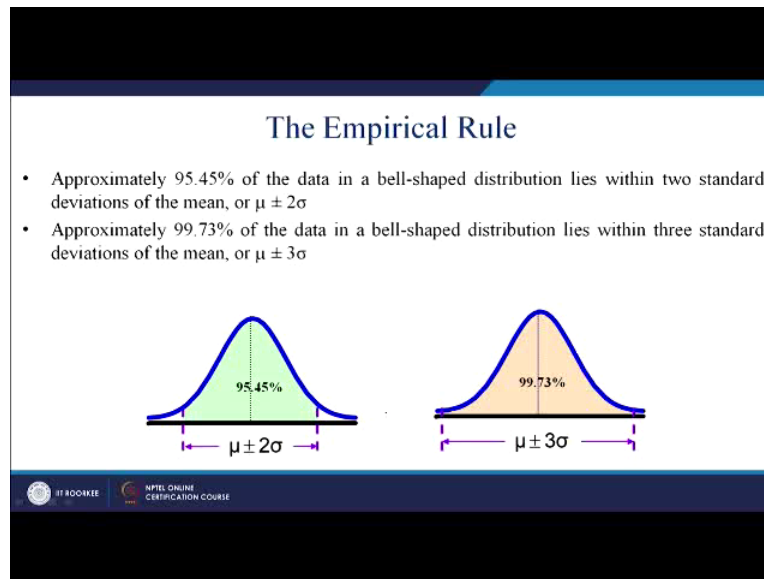


So, the empirical rule approximately variation of data in a bell shaped distribution. So, this is your bell-shaped distribution now it has got certain property the first thing is this is middle point of this and this side is 50% and this side is 50%. So, area under curve is 100% or we call it unity just one. Now this is your population means you are denoting it by  $\mu$  and this is your standard deviation Sigma. So, will tell that approximately 68.27% of the data in bell shaped distribution

fall within one standard deviation  $\pm 1$  standard deviation, so this is  $\pm 1$  standard deviation now if I ask you a question what is this area?

From Mean 1 Sigma towards left so how would you find out so just think for a while? We are aware that this area this total area is 68.27 to the 50% of this would be 34. Let say 135 is it not. This is area within 1 standard deviation.

**(Refer Slide Time: 03:01)**




Approximately 94.45 of the data in the bell shaped distribution will lie within 2 standard deviation so  $\mu \pm 2$  Sigma. Again if I ask you to calculate this area then what it would be it would be  $95.45/ 2$ . Similarly approximately 99.73 data points would lie within 3 Sigma limits.


**(Refer Slide Time: 03:41)**

Using the Empirical Rule

- Suppose that the variable Math SAT scores is bell-shaped with a **mean of 500 and a standard deviation of 90**. Then,
  - 68.27% of all test takers scored between 410 and 590     ???
  - 95.45% of all test takers scored between 320 and 680     ???
  - 99.73% of all test takers scored between 230 and 770     ???

2

 NPTEL  
IT #001K02

 NPTEL ONLINE  
CERTIFICATION COURSE

This is empirical rule and it has got several advantages. Let us take one example wherein we have applied one empirical rule. Suppose that the variable maths Z score is bell shaped with mean in this and standard deviation, how would you draw this? This is your Distribution this is mean 500 and this is standard deviation 90. So will say that 68.27% of the test rankers was appeared in test would score between 410 to 590 how did you get this very simple  $500 \pm 1 * 90$ ; 90 is your standard deviation.

Similarly within two Sigma limits are when the area is this you will have  $500 \pm 2 * 90$ , so this is 320 lower limit and upper limit is 680. Similarly  $500 \pm 3$  Sigma, sigma is 90. So, you can find out how much person how many data points lie within 1 Sigma, 2 Sigma and 3 Sigma limits.

**(Refer Slide Time: 05:06)**

### Numerical Descriptive Chebyshev Rule

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

Now apart from empirical rule there is one more theorem is called chebyshev rule or chebyshev theorem. Now one of the assumptions are one of the properties of Normal Distribution are what we have seen here a bell shaped distribution is that the distribution is normal then only this work right. 68.27, 95.45, 99.73 all these are true when your distribution is normal but in real life situation most of the time you will not have normal distribution.

So, if distribution is not normal then according to Chebyshev rule at least this much percentage of values will fall within test standard deviation of the mean here  $k$  is ranging from 1 to any other number right 1, 2, 3, 4 whatever it is. So, let us look at how many data points lie within 2 standard deviations this  $(1 - 1/2^2) k = 2$ . So, will say that within two Sigma limits 75% of data points lie, what was the case in earlier example where we had normal distribution it was 95. Some 4 points approximately this much.

But here it is only 75% when  $k = 3$  put  $k = 3$  in this equation you will get 89%. So, 89% of the data points will lie within 3 Sigma limits according to Chebyshev rule. And according to our empirical rule where are the distribution was normal it was approximately 99.73 right but here it is 89%.

**(Refer Slide Time: 07:23)**

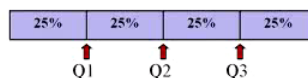
Another way of describing numerical data is through an exploratory data analysis that includes:  
**Quartile,  
Five number summary,  
and the Box plot.**

So, there are some other ways in which we can describe our numerical data it is called exploratory data analysis is called EDA. It includes quartiles, five number summary and box plot and we will see these exploratory data analysis tools.

(Refer Slide Time: 07:51)

### Quartiles

Quartiles split the ranked **data into 4 segments with** an equal number of values per segment

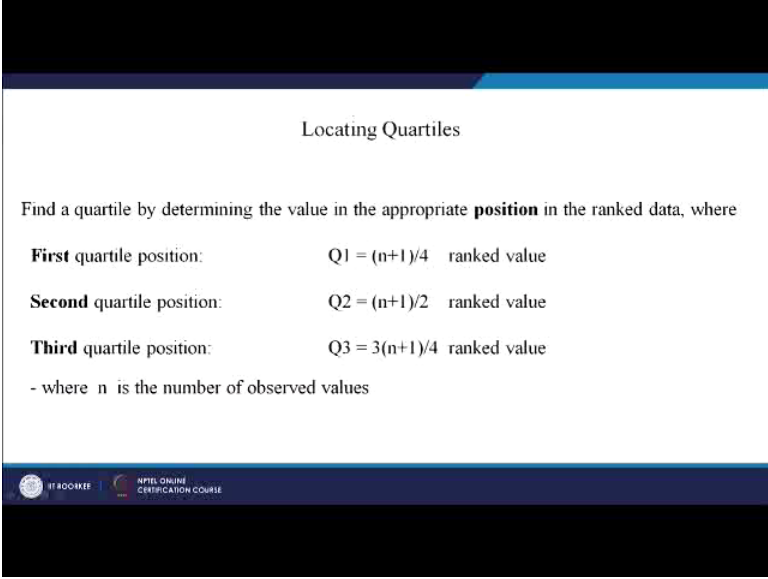


- The first quartile,  $Q_1$ , is the value for which **25% of the observations are smaller** and 75% are larger
- $Q_2$  is the same as the median (50% of the observations are smaller and 50% are larger)
- Only 25% of the observations are **greater than the third quartile**

Let us look at quartile, what is quartile and effect in one of the classes I have already defined. So, if you divide your data into let say 10 equal parts then each part would be a deciles. If you divide data into 4 equal parts then each part would be the quartile. If you divide data into 100 equal parts then each part would be a percentile. So, quartile split rank data into four segments with equal number of values per segment. So, let us look at this so the first quartile is  $Q_1$  with which means 25% of the data points one side and remaining 75% of the data points other side.

Q2 is just middle point right it is median so 50% of the observation one side and 50% of the observation other side ok. Now only 25% of the observations are greater than third quartile. So, this is your first quartile 25% of data, second quartile this side 50% and this side 50% third quartile this side 75% and this side 25% is it not.

**(Refer Slide Time: 09:19)**



Locating Quartiles

Find a quartile by determining the value in the appropriate **position** in the ranked data, where

<b>First</b> quartile position:	$Q1 = (n+1)/4$ ranked value
<b>Second</b> quartile position:	$Q2 = (n+1)/2$ ranked value
<b>Third</b> quartile position:	$Q3 = 3(n+1)/4$ ranked value

- where  $n$  is the number of observed values

IT ROOMEE NPTEL ONLINE CERTIFICATION COURSE

So, we need to find out the location of these quartiles. The first quartile is  $(n + 1)/ 4$  so  $n$  is how many;  $n$  is number of data points in your dataset. Let us look at second quartile it just 50% so it is just  $(n + 1)/ 2$  right just when you are divided by 4 means 25% and divided by 2, 50% when it is  $3/4$  it means 75% isn't it? So, this is how we calculate positions.

**(Refer Slide Time: 09:59)**

### Calculation Rules

- When calculating the ranked position use the following rules
  - If the result is a **whole** number then it is the **ranked** position to use
  - If the result is a **fractional half** (e.g. 2.5, 7.5, 8.5, etc.) then **average** the two corresponding data values.
  - If the result is **not a whole number or a fractional half** then **round the result to the nearest integer** to find the ranked position.

So, now we calculate rank positions by using the following rules right. There are these 3 rules so if the result is a whole number then is called Ranked position to use. But if it is fractional and fractional is like this 2.5, 3.5, 17.5, 95.5 and so on. If the result is fractional half then average is to be taken of the two corresponding data values. If it is not fractional half any other number then round it to the integer value towards higher side towards positive side so, this how you can calculate these quartile positions.

(Refer Slide Time: 11:04)

### Locating Quartiles

Sample Data in Ordered Array: 11 12 13 16 16 17 18 21 22

(n = 9)

$Q_1$  is in the  $(9+1)/4 = 2.5$  position of the ranked data

so use the value half way between the 2<sup>nd</sup> and 3<sup>rd</sup> values.

so  $Q_1 = 12.5$

$Q_1$  and  $Q_3$  are measures of non-central location  
 $Q_2$  = median, is a measure of central tendency

So, let us solve this example so this is a case where in there are 9 data points 11 12 13 16 16 17 18 21 22. So the first quartile is what is  $n + 1$  by 4, so  $n$  is 9 this is  $10/4$  is 2.5 so what we have said it is 2.5 a fraction is one half right. So, we have to take the average of these 2 right 12 and

13. So,  $Q_1$  is 12.5. Why did we take average of this because the result is the fractional half? Let us look at how to find out  $Q_2$ , so before this let me tell you that  $Q_1$  and  $Q_3$  are measures of noncentral location of course they are not Central location in fact  $Q_2$  is a Central location.

(Refer Slide Time: 12:11)

**Quartile Example**

**Sample Data in Ordered Array:** 11 12 13 16 16 17 18 21 22

$(n = 9)$

$Q_1$  is in the  $(9+1)/4 = 2.5$  position of the ranked data.  
so  $Q_1 = (12+13)/2 = 12.5$

$Q_2$  is in the  $(9+1)/2 = 5^{\text{th}}$  position of the ranked data.  
so  $Q_2 = \text{median} = 16$

$Q_3$  is in the  $3(9+1)/4 = 7.5$  position of the ranked data.  
so  $Q_3 = (18+21)/2 = 19.5$

$Q_1$  and  $Q_3$  are measures of **non-central location**  
 $Q_2$  = median, is a measure of **central tendency**

NPTEL ONLINE CERTIFICATION COURSE

Let us find out what is  $Q_2$  right,  $Q_2$  is what  $n+1$  by 2 isn't it? This is 5th position, 5th position is what? First, second, third, fourth and fifth so this is  $Q_2$  which is also median right.  $Q_3$   $(n + 1) * 3$  by 4 so this is 7.5 again fractional half so we have to take the average of 7th and 8th data points. So, 1st 2nd 3rd 4th 5th 6th 7th and 8th 18 + 21, so it is 19.5, so, we will say that  $Q_1$  is this and  $Q_2$  is this and  $Q_3$  is this. This is how you can find out quartiles.

Now one of the important measures of dispersion is range. So, you can; though range is not a good measure because it takes into account maximum and minimum value but there is something called interquartile range. Now interquartile range can be calculated like this.

(Refer Slide Time: 13:38)



### Quartile Measures: The Interquartile Range (IQR)

- The IQR is  $Q_3 - Q_1$  and measures the spread in the **middle 50%** of the data
- The IQR is also called the **midspread** because it covers the middle 50% of the data
- The IQR is *a measure of variability that is not influenced by outliers* or extreme values
- Measures like  $Q_1$ ,  $Q_3$ , and IQR that are *not influenced by outliers are called resistant measures*

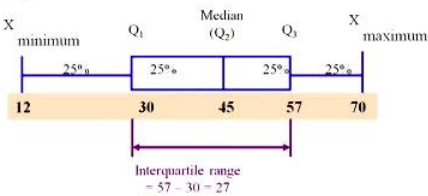
This is the difference between  $Q_3 - Q_1$ , in other words the middle 50% of the data so this is your distribution this is your  $Q_1$  this is  $Q_3$  and this is your  $Q_2$  so middle 50% of the data is interquartile range, it is called IQR. IQR is also called as mid spread because it covers middle 50% of the data. So, you can call IQR and mid spread data. IQR is a measure of variability that is noninfluenced by outlier most important characteristics of IQR why? Because it is not affected by outliers or extreme values why, why not affected by extreme values because the extreme values would be somewhere either in this  $Q_1$  part in less than  $Q_1$  part or more than  $Q_3$  part right.

So, that is why it is not like you are not influenced by outliers. So, measures like  $Q_1$ ,  $Q_3$  and IQR that are not influenced by outliers are also called resistant measure. So, you call them resistant measures as well. So, let us look at this if we want to find out IQR as I said it is middle 50% of the data.

**(Refer Slide Time: 15:14)**

## Calculating The Interquartile Range

Example:



So, let say you have got a data set and in which the minimum value is 12 maximum 70 and you have found that the Q1 is 30, right Q2 45, Q3 47 how did we find Q1 it just  $\frac{1}{4}$ ,  $n + (1 / 4)$ . So, after putting n here in this formula we calculated Q1 and Q1 was 30, Q3 57 so IQR would be 27 different between Q3 – Q1 is IQR. So, this was about quartile. Let us move on to five number summary right so what we are discussing their discussing exploratory data analysis tools so we have seen quartile the second one is five number summery is very simple one.

**(Refer Slide Time: 16:16)**

## The Five Number Summary

The five numbers that help describe the **center, spread and shape of data** are:

- $X_{\text{smallest}}$
- First Quartile ( $Q_1$ )
- Median ( $Q_2$ )
- Third Quartile ( $Q_3$ )
- $X_{\text{largest}}$

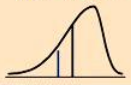
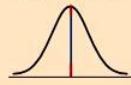
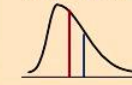
And there are 5 number summary means there are 5 numbers which would be telling us about shape of the distribution. So, this first number smallest and last one is largest number in a dataset Q1 Q2 and Q3 so these are five number that is why it is called five member summary. Now if

you look at these five number then you will be able to know where data are concentrated in a distribution and there is a relationship between five number summary and distribution shape.

(Refer Slide Time: 17:01)

Relationships among the five-number summary and distribution shape		
Left-Skewed	Symmetric	Right-Skewed
Median - $X_{\text{smallest}}$ >	Median - $X_{\text{smallest}}$ =	Median - $X_{\text{smallest}}$ <
$X_{\text{largest}}$ - Median >	$X_{\text{largest}}$ - Median =	$X_{\text{largest}}$ - Median <
$Q_1$ - $X_{\text{smallest}}$ >	$Q_1$ - $X_{\text{smallest}}$ =	$Q_1$ - $X_{\text{smallest}}$ <
$X_{\text{largest}}$ - $Q_3$ >	$X_{\text{largest}}$ - $Q_3$ =	$X_{\text{largest}}$ - $Q_3$ <
Median - $Q_1$ >	Median - $Q_1$ =	Median - $Q_1$ <
$Q_3$ - Median >	$Q_3$ - Median =	$Q_3$ - Median <

<b>Left-Skewed (Negative)</b> <b>Mean &lt; Median</b> 	<b>Symmetric</b> <b>Mean = Median</b> 	<b>Right-Skewed (Positive)</b> <b>Median &lt; Mean</b> 
---	---	---

Five number summary and distribution shape. If you look at this distribution shape for example this one is right skewed distribution right. This one is left skewed or negatively skewed or this is also known as positively skewed this symmetric data perfectly bell shaped curve right. So let us find out relationship will look at only for right skewed distribution. So what is the relationship, whatever is median if you subtract smallest value then that would be always less than largest value minus median, so this the relationship.

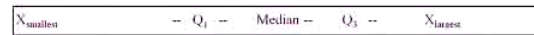
The second relationship is whatever is your  $Q_1$  value subtract smallest value from it that would always be less than or equal to largest value minus  $Q_3$ , this is another relationship. Third one is median minus  $Q_1$  will always be less than  $Q_3$  minus median. So, these are 3 relationships for right skewed distribution. Similarly you can have 3 for left skewed and 3 for symmetric. But you need to know what is right skewed distribution or positive skewed distribution keep in mind that when you say positive skewed so this is your distribution let us say.

This is your 0 point, this is +1, +2 like this right, this is -1 and -2 so since this tail is skewed towards positive side that is why we are calling it positive skewed.

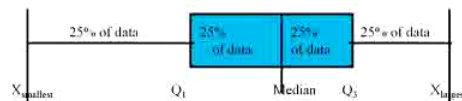
(Refer Slide Time: 19:11)

## Five Number Summary and The Boxplot

- **The Boxplot:** A Graphical display of the data based on the five-number summary:



Example:

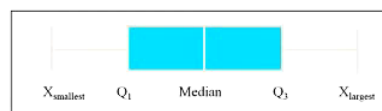


Five number summary and the box plot so these are; so this is your five number summary smallest value  $Q_1$   $Q_3$  median and largest and this is the box plot. In box plot what you will have all these 5 points smallest value largest value  $Q_1$  and  $Q_3$  and median right we have always said that this  $Q_1$  is first 25% of data median first 50% right this is the middle point in  $Q_3$  first 75% box plot. So, this is box plot.

(Refer Slide Time: 19:56)

## Five Number Summary: Shape of Boxplots

- If data are **symmetric around** the median then the box and central line are **centered between the endpoints**



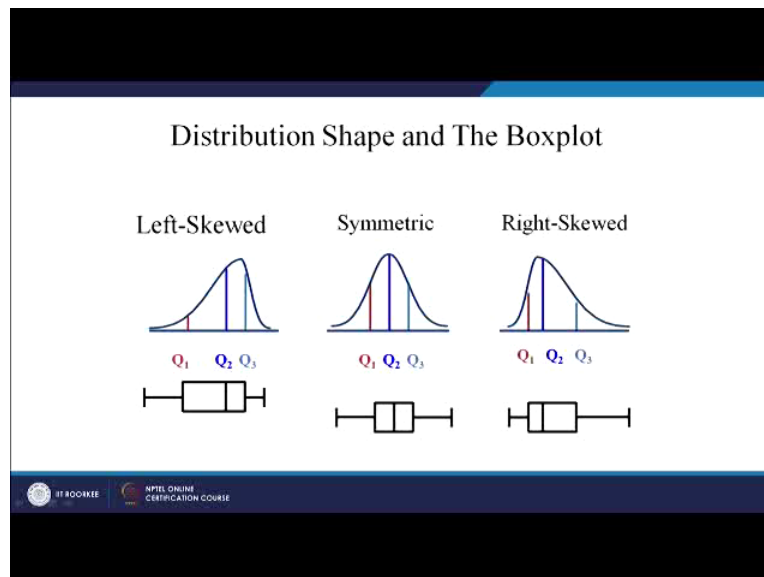
- A Boxplot can be shown in either a **vertical or horizontal** orientation

Now shape of the box plot will tell you how data distributed. Let us look at this box plot so that data are symmetric then the box plot and centre line centered between the endpoints right. So, this is your box plot and this is your median. So, this line would be in centre of  $Q_1$  and  $Q_3$ , it means it is a symmetric distribution. A box plot can be shown in either vertical or horizontal

position. So, you can have a box plot like this also isn't it? So, this is you are; let us say largest value smallest value this is your median.

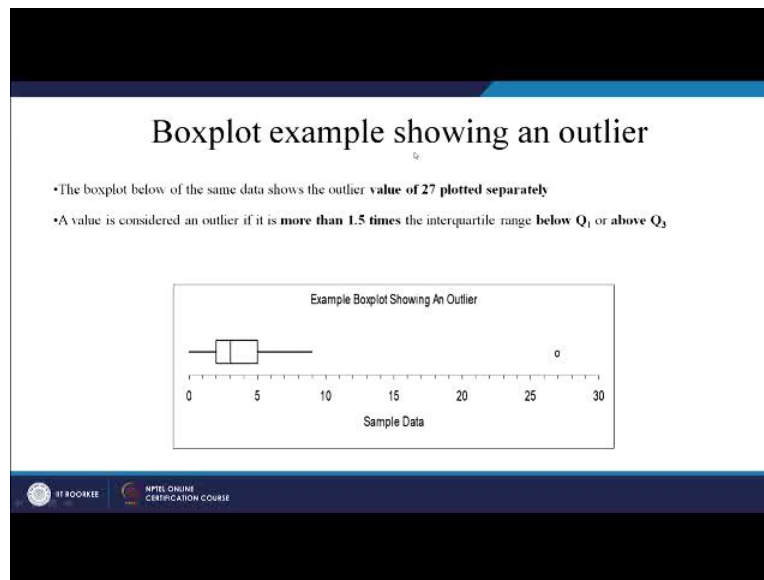
So, you can have this is your  $Q_1$  and  $Q_3$ . In fact  $Q_3$  should be towards the largest value so this is  $Q_1$  is median and this is  $Q_3$  right. So, you can have a vertical box plot as well but you have to see this point very carefully. If this is towards  $Q_1$  then it would be a positively skewed distribution if it is towards this negatively skewed distribution.

**(Refer Slide Time: 21:33)**



Just see this is your box plot and this is a right skewed curve right. So, this is your box plot this is your  $Q_1$  this is  $Q_3$  so and this is median. Since this is towards  $Q_1$  it is right skewed. Now here if you look at this is towards  $Q_3$ , so it is left skewed and of course in case of symmetric it would be in the middle of the box plot right. So, this is how there is a relationship between box plot and distribution shape.

**(Refer Slide Time: 22:14)**



Look at one more example in fact apart from knowing what is the shape of distribution from box plot you can get one more information and check whether how many points are out outliers or which points are outliers in a dataset so that can also be known from IQR from interquartile range. So here you have got some data points so you can easily draw box plot.



So, let us see how to calculate outliers in a dataset? So, there are multiple methods but the very first method is using IQR interquartile range. So, a value is considered an outlier if it is more than 1.5 times the interquartile range below  $Q_1$  and above  $Q_3$ . So  $Q_1 - 1.5*(IQR)$  is nothing but those data points would be outliers. Similarly  $Q_3 + 1.5*(IQR)$  all those data points would be outliers.

**(Refer Slide Time: 23:46)**

Find outliers?

850	875	4700	4900	5300	5700	6700	7300	7700	8100
8300	8400	8700	8700	8900	9300	9500	9500	9700	10000
10300	10500	10700	10800	11000	11300	11300	11800	12700	12900
13100	13500	13800	14900	16300	17200	18500	20300	21310	21315

$Q1 = 8100$   
 $Q3 = 12900$   
 $IQR = 12900 - 8100 = 4800$   
 $1.5 * IQR = 7200$   
 Outliers-  $Q1 - 7200 = 8100 - 7200 = 900$   
 Outliers-  $Q3 - 7200 = 12900 - 7200 = 20100$   
 Any data point below 900 and above 20100 are outliers

 NPTEL  
 NPTEL ONLINE  
 CERTIFICATION COURSE

Let us look at this example so we have to find out outliers. So, here you got how many data points 1 2 3 4 5 6 7 8 9 10 so there are 40 data points right and you to find out outliers so first of all you need first of all you need Q1 and Q3 right. So, Q1 is this 8100 just see Q3 is 12900 what is median 10,000 right. So, IQR is what?  $Q3 - Q1$  so this is 4800 is IQR. So, first of all you to take 1.5 IQR is this, 7200 is 1.5 times of IQR.

So, what is outlier I have set  $Q1 - 1.5 * IQR$  so all those points which are below this would be outliers you just see this is  $Q1 - 1.5 * IQR$  so Q1 is this 7200 so all those points which are less than or equal to 900 they are outliers. So, is there any outliers over here, is there any data point which is less than or equal to 900 yes these two are outliers. Now in fact you just keep in mind that for finding out Q1 and Q3 first of all you need to arrange data in ascending order right, I think you know this point.

Q3 let us look at the outliers towards higher side,  $Q3 - 1.5 * IQR$  so this is 20100 so all those points beyond this or more than this will be outliers. Is there any data points. Do we have any data point? Which is more than this, yes it is there these three are outliers. This is 20300, 21310, 21315 so these three. So, in this data set how many outliers total 2 + 3 right 5 outliers 2 + 3 ok. Before moving on to next slide let me summarize what we did today.

In today's session we view empirical rule for a normally distributed data and in which we said 1 Sigma limit, 2 Sigma limit and 3 Sigma limit you have got approximately 68.45 data 95.45 data and 99.73 data. We have also seen Chebyshev rule and the advantage this Chebyshev rule is that it can be applied to nonnormal distribution as well. So, if you look at let us say  $\mu \pm 2\sigma$  then approximately 90% of the approximately 89% of the dataset would lie in 2 sigma limit. While in case of normal distribution that value was 95 point something right.

Then we have also seen exploratory data analysis tools. We have seen quartile using five number summary and we have also seen IQR right and box plot so what is quartile it is basically you are dividing dataset into 4 equal parts so for our IQR and interquartile range you need to subtract  $Q3 - Q1$  which is also known as the middle 50% of the dataset which is not influenced by outliers. And we have seen box plot and five number summary. In five number summary we seen largest value smallest value  $Q1$  median and  $Q3$ .

And what is the use of box plot what does it give one information does box plot give? It tells us whether data are normally distributed are positively distributed or negatively distributed just by looking at the centre value of the box plot right. So, let us say this is  $Q1$  this is  $Q3$  this would be if it is exactly in centre then it would be normally distributed data and we have seen how to find out outliers.

So,  $Q1 - 1.5 * IQR$  those values which are less than this would be outliers in  $Q3 + 1.5$  times of IQR all those values more than this would be again outliers. So, with this let me finish today's session thank you very much.