Dealing with Materials Data: Collection, Analysis and Interpretation

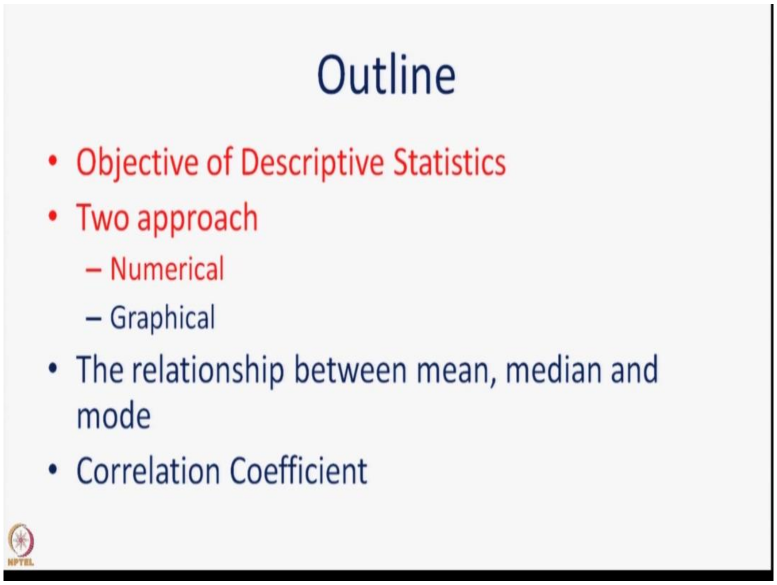
Professor. Hina. A Gokhale
Department of Metallurgical Engineering and Materials Science,
Indian Institute of Technology, Bombay.

Lecture 02

Descriptive Statistics – II

Hello and welcome to the course on Dealing with Materials Data. In the previous session, we were going through the subject of descriptive statistics. How to describe your data in terms of either numerical values or the graphical values.

(Refer Slide Time: 00:40)

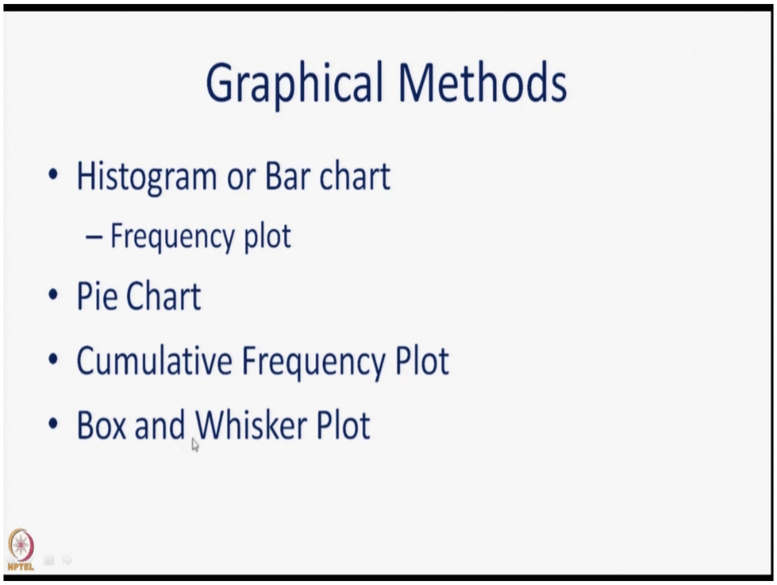


Outline

- Objective of Descriptive Statistics
- Two approach
  - Numerical
  - Graphical
- The relationship between mean, median and mode
- Correlation Coefficient

We covered the 3 numerical aspects of it, out of the two approaches. We first looked into what is the objective of descriptive statistics. Then we said there are 2 approaches, one is numerical and graphical and in the last session we covered up to numerical methods in which we covered the measure of central tendency and also the measure of dispersion. Today we are going to cover the graphical methods, the relationship between mean, median and mode and the correlation coefficient when there you have a data sets with 2 variables. Graphical methods, you are all very familiar with, you see them more often in the newspaper these days with the data journalism becoming very popular.

(Refer Slide Time: 01:43)

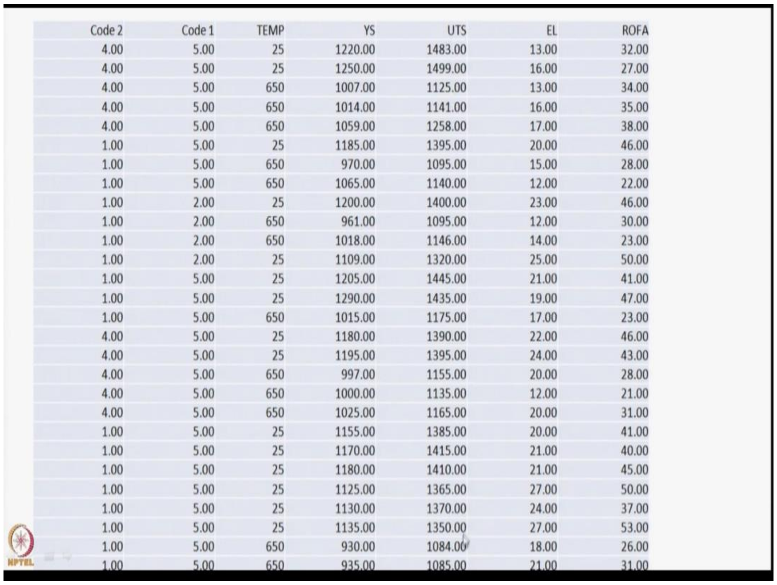


Graphical Methods

- Histogram or Bar chart
  - Frequency plot
- Pie Chart
- Cumulative Frequency Plot
- Box and Whisker Plot

So we have a histogram or a bar chart which is also called a frequency plot. Sometimes the variety of category of data that comes into it, it can also be shown as a pie chart in which you show different percentage of data which comes from different areas and then we have a cumulative frequency plot, which is very much close to the frequency plot that we are discussing here and then the last is box and whisker plot. It is actually a box and it has 2 whiskers and therefore it is called box and whisker plot. It is not by the name of any statistician or any scientist.

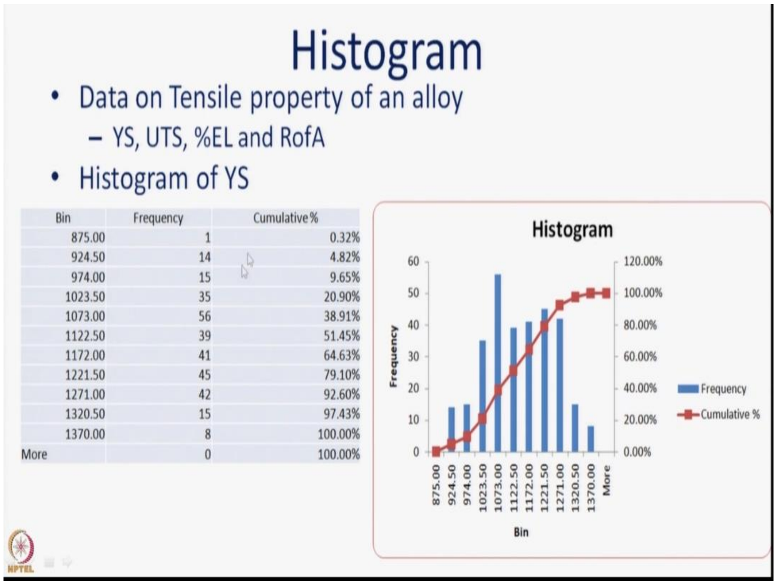
(Refer Slide Time: 02:28)



Code 2	Code 1	TEMP	YS	UTS	EL	ROFA
4.00	5.00	25	1220.00	1483.00	13.00	32.00
4.00	5.00	25	1250.00	1499.00	16.00	27.00
4.00	5.00	650	1007.00	1125.00	13.00	34.00
4.00	5.00	650	1014.00	1141.00	16.00	35.00
4.00	5.00	650	1059.00	1258.00	17.00	38.00
1.00	5.00	25	1185.00	1395.00	20.00	46.00
1.00	5.00	650	970.00	1095.00	15.00	28.00
1.00	5.00	650	1065.00	1140.00	12.00	22.00
1.00	2.00	25	1200.00	1400.00	23.00	46.00
1.00	2.00	650	961.00	1095.00	12.00	30.00
1.00	2.00	650	1018.00	1146.00	14.00	23.00
1.00	2.00	25	1109.00	1320.00	25.00	50.00
1.00	5.00	25	1205.00	1445.00	21.00	41.00
1.00	5.00	25	1290.00	1435.00	19.00	47.00
1.00	5.00	650	1015.00	1175.00	17.00	23.00
4.00	5.00	25	1180.00	1390.00	22.00	46.00
4.00	5.00	25	1195.00	1395.00	24.00	43.00
4.00	5.00	650	997.00	1155.00	20.00	28.00
4.00	5.00	650	1000.00	1135.00	12.00	21.00
4.00	5.00	650	1025.00	1165.00	20.00	31.00
1.00	5.00	25	1155.00	1385.00	20.00	41.00
1.00	5.00	25	1170.00	1415.00	21.00	40.00
1.00	5.00	25	1180.00	1410.00	21.00	45.00
1.00	5.00	25	1125.00	1365.00	27.00	50.00
1.00	5.00	25	1130.00	1370.00	24.00	37.00
1.00	5.00	25	1135.00	1350.00	27.00	53.00
1.00	5.00	650	930.00	1084.00	18.00	26.00
1.00	5.00	650	935.00	1085.00	21.00	31.00

So here for an example, I have taken a data. This data actually comes from the mechanical properties or the strength properties of super alloy. Here you can see that there are a couple of codes given to the data. There is a temperature attached to the data. This is the yield strength property of the data. This is ultimate tensile strength property of the data. This is elongation and this is reduction of area. Both of these are in percentage.

(Refer Slide Time: 03:06)



Histogram

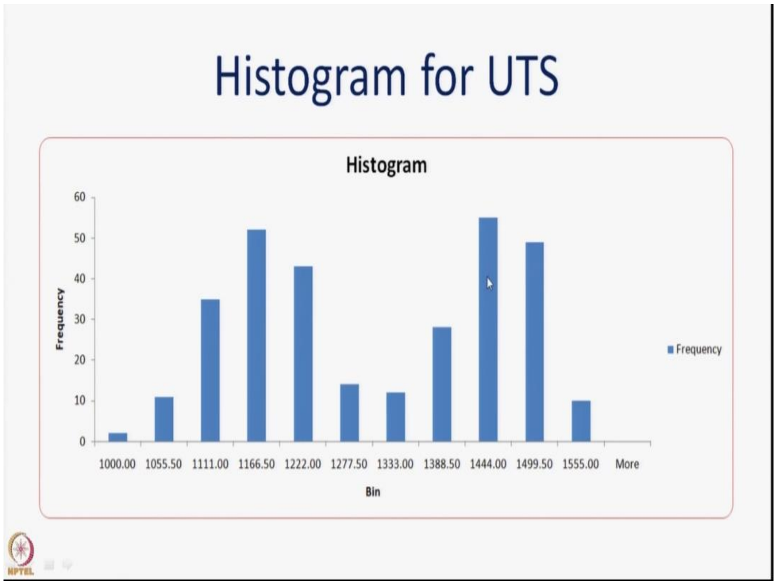
- Data on Tensile property of an alloy
  - YS, UTS, %EL and RofA
- Histogram of YS

Bin	Frequency	Cumulative %
875.00	1	0.32%
924.50	14	4.82%
974.00	15	9.65%
1023.50	35	20.90%
1073.00	56	38.91%
1122.50	39	51.45%
1172.00	41	64.63%
1221.50	45	79.10%
1271.00	42	92.60%
1320.50	15	97.43%
1370.00	8	100.00%
More	0	100.00%

Now from this data if I want to plot a histogram, what I do is I generally find the lowest value and the highest value of the data and I divide it into a number of bins. So taking a number of bins, there is no thumb rule. There are no written rules, but there are some thumb rules. Here, generally, I like when the data is very large as it is the case because in this case we have 311 data points and therefore we have divided it into 10 bins and these values you see its 875 MPa or 924.5 MPa, etc. These data is a yield strength midpoint of the bin value and here is the count. The count number of data points that falls in it and here is a cumulative frequency that is, this is the 1, then this is 15 out of 311. Then this is 15 plus 1530 out of 311.

Like that, this is the cumulative percentage of data that falls below this particular bin value and here is a plot. This plot is called a histogram in which at every bin value it has shown the length of a bar which corresponds to the frequency of the data and this red plot is actually a cumulative frequency curve. The y axis on the right side shows the percentage as in terms of the cumulative percentage and the left side y axis actually gives the frequency. This is called a histogram.

(Refer Slide Time: 04:58)

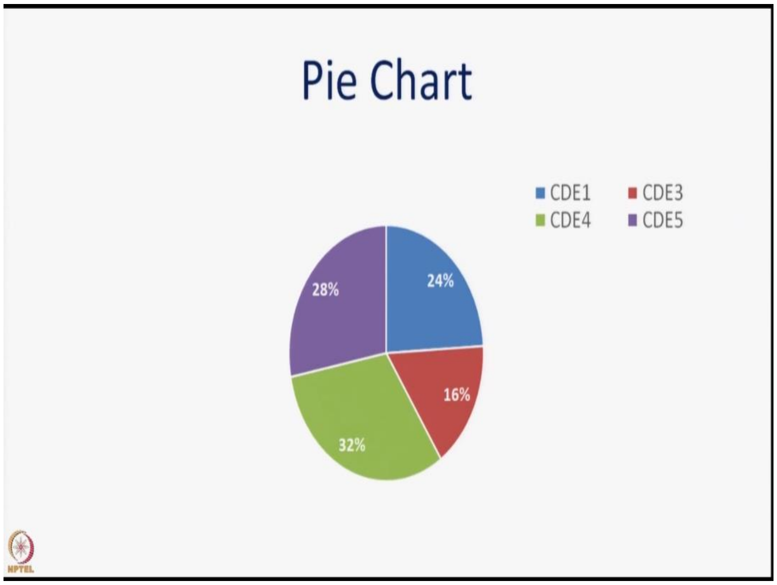


Histogram for UTS

Histogram gives us a lot of information. For example, when I make the similar plot for the ultimate tensile strength, I find that these are bimodal. You remember the different definition of mode. Mode of a distribution is where the data point occurs with the highest frequency and you can see that here 1000 or rather 1166.5 MPa is the highest frequency here is while here it is 1444 MPa is the highest frequency. So you find that this is a bimodal data and it can be shown that this happens because there are two temperatures.

If you look back the data has temperature values of 25 degrees C which is room temperature and high temperature is taken as 650 degrees C and therefore you see the two peaks. One belongs to the room temperature the other belongs to the high temperature. So what I am trying to say here is that histogram actually gives out a lot of information in the beginning about the data so that you know, when you do the statistical analysis of the data how to deal with the data.

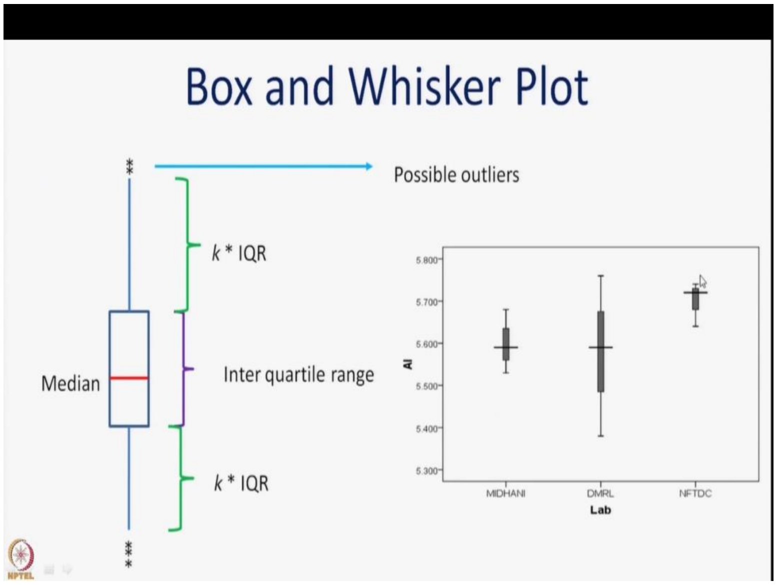
(Refer Slide Time: 06:24)



Pie Chart

Here is a pie chart. As I said, there were some codes defined. So code 1,4,3,5 and here it shows that how much data belongs to which code value. One code value is defined as per the ASTM number for the grain size so therefore it is divided into this manner. So this is a pie chart.

(Refer Slide Time: 06:54)



Box and Whisker Plot

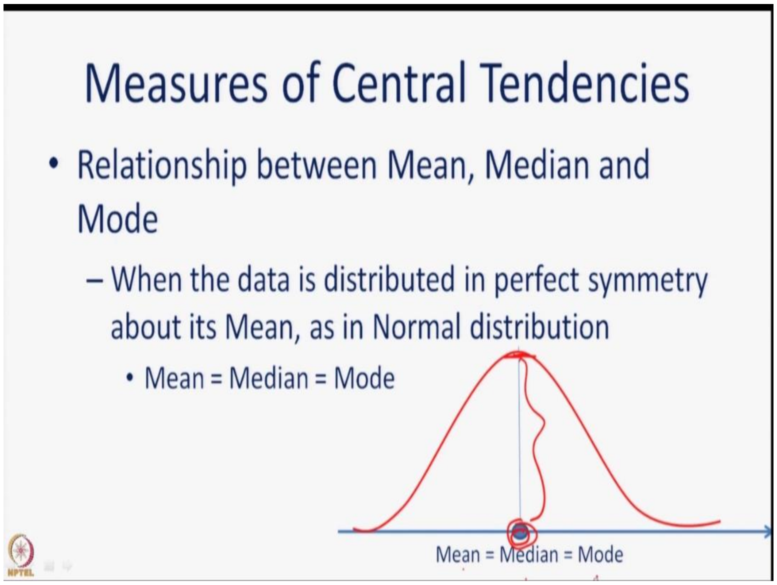
Box and whisker plot. This is a one plot which gives away a lot of information. As the box and whisker name says there is a box to it and there are whiskers attached to it. You see there are the whiskers and this is the box. The box is made in this way. You have the inter quartile range given here. So if we understand that this is the minimum and this is the maximum value then this is q1. Remember what is inter quartile range? It is (q3 - q1). Q1 says that 25 percent of the data is below q1 and q3 says that 75 percent of the data is below the q3 or 25 percent of data is above q3.

Please remember this is what I say smallest value to largest value and therefore the below and above look the opposite in this particular figure. This plot also has 1 median line shown which is actually the median value of the product of the data. Here is the inter quartile range. Then the whiskers lengths vary. Number of times whiskers length is given in terms of the maximum value and the maximum value here and the minimum value here but it all depends on different softwares and different approaches. It can be said that it is generally taken as a k times the inter quartile range on both the sides.

Typically here I am showing you a box and whisker plot which I have obtained for a chemical analysis of aluminum in certain alloy and I have 3 laboratories at where this alloy has been chemical analysis has been tested or analysis has been done. One is Mishra Dhatu Nigam, other is DMRL and the third is NFTDC and this shows the results obtained in either case as you can see here. Here when they show it like this, it means that it has taken maximum and minimum value. When the whisker has a short horizontal lines on top and bottom it is maximum and minimum values. Here, of course, it goes from minimum to maximum. So what I described here it is upside down here, please remember.

So here it shows that the maximum value is somewhere in between 5.5 ppa to somewhere around 5.7 ppa while in the case of DMRL, it is a very wide range data. This shows the median and this can also give you why have I shown this, this also gives you a way to compare the data from 3 different sources. The same data has been collected from 3 different sources and this is the plot which shows you the 3 different ways of expressing the data.

(Refer Slide Time: 10:34)



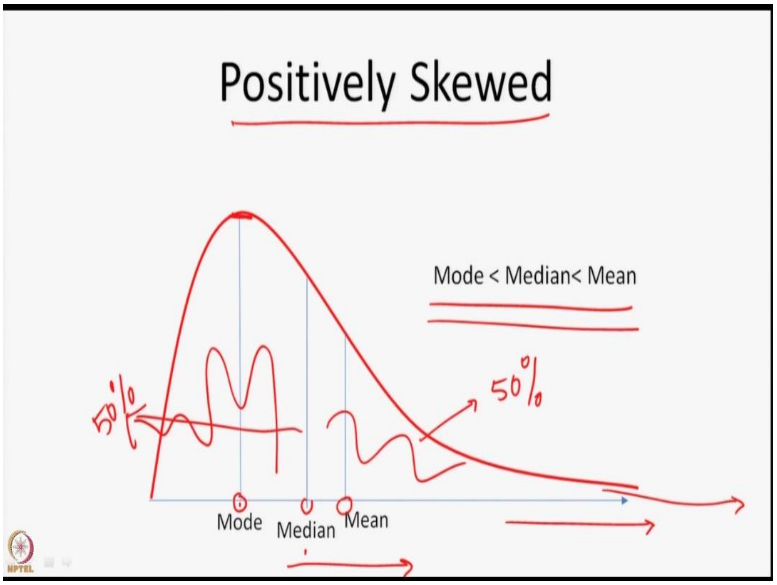
Measures of Central Tendencies

- Relationship between Mean, Median and Mode
  - When the data is distributed in perfect symmetry about its Mean, as in Normal distribution
    - Mean = Median = Mode

Now we come back to measures of central tendencies something that we had left out. Why because this relationship is best shown in terms of frequency plot. Now you recall, for example, if you look at this, if you join these points together, you get a plot. This is called a frequency plot and here I have shown a nice bell-shaped frequency plot. This is the data which has a distributed in perfect symmetry such as it can happen in normal distribution. So when a data distributed in a perfect symmetry the mean is equal to median is equal to mode. It means that all these 3 value reside in the same point.

You can see that it resides in the same point and where the mode is because mode is very easy to find out from a frequency plot. It is the highest frequency or the highest frequency the value occurs and therefore mode is at the highest frequency level and here you can see that mean, median and mode are, say, if the data is in perfect symmetry.

(Refer Slide Time: 12:10)



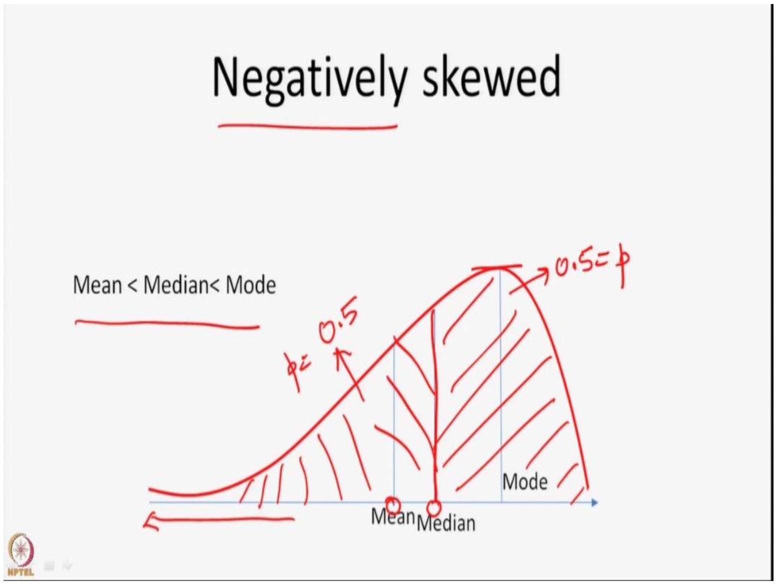
Positively Skewed

Mode < Median< Mean

What happens if the data is skewed, now you see this long tail, this long tail on the right side is identified as positively skewed data. When a frequency plot has a long tail on the right side, it is called positively skewed. Now when you have a long tail instead of symmetric, you have a long tail on right side. It means that the frequency of data occurring on the right side is higher than what it would have been in a perfectly symmetric curve and therefore the mean value gets shifted towards the right. The median remains in the center because it divides the data into two halves.

Please remember median divides the data into two halves. So 50 percent of data is on this side while 50 percent of the data is on this side. Mode always remains at the highest frequency point. So mode is very easy to find out and the mode is here, median is here and the mean is here. Therefore the relationship becomes in the positively skewed distribution mode < median < mean. This relationship helps when you actually find mode, median and mean in descriptive statistics and you find that they are strikingly different and having this relationship. You already have an idea that your data is positively skewed.

(Refer Slide Time: 14:16)



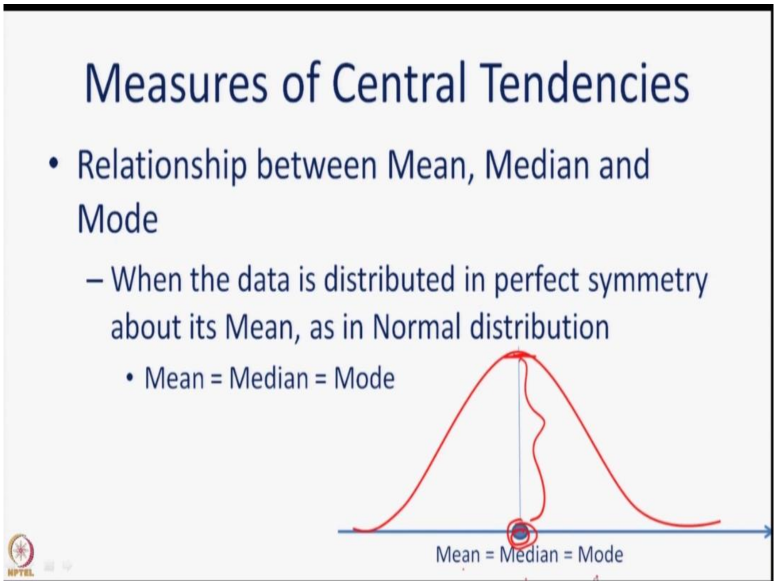
Negatively skewed

Mean < Median< Mode

On the other hand, if the data is negatively skewed you see that a long tail is on the left side and therefore such a data is called negatively skewed. Again because on the left side, the frequency of data has increased, the mean has moved to the left of the center, median remains at the center. So again this side of the data is 0.5 or 50 percent and this side of data is also 0.5 or 50 percent and this side of data is also 0.5 probability. This is p is equal to 0.5 probability because that is how the median is at the center.

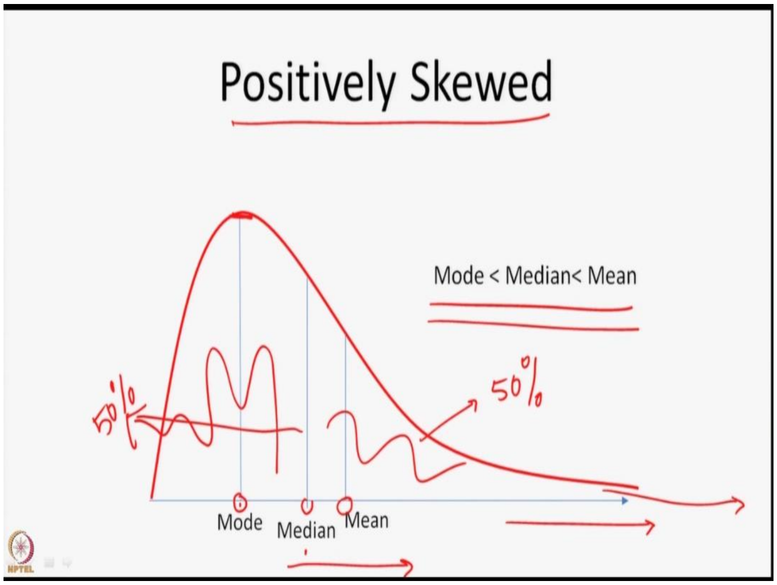
Mode is always the highest frequency and therefore when the data is negatively skewed, you will find that mean < median < mode. So once again from the descriptive statistics if we come to know that there is a large difference and the relation is mean < median < mode then you have an idea that you are going to deal with a negatively skewed distribution. This you can further confirm by plotting a histogram and have a look at it how much skewed it is but this is how the relationship between mean, median and mode.

(Refer Slide Time: 16:20)



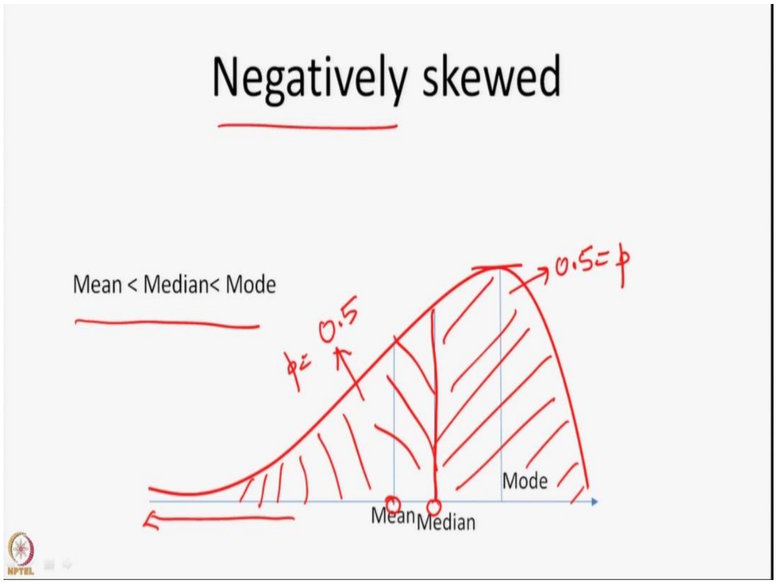
Measures of Central Tendencies

- Relationship between Mean, Median and Mode
  - When the data is distributed in perfect symmetry about its Mean, as in Normal distribution
    - Mean = Median = Mode



Positively Skewed

Mode < Median< Mean

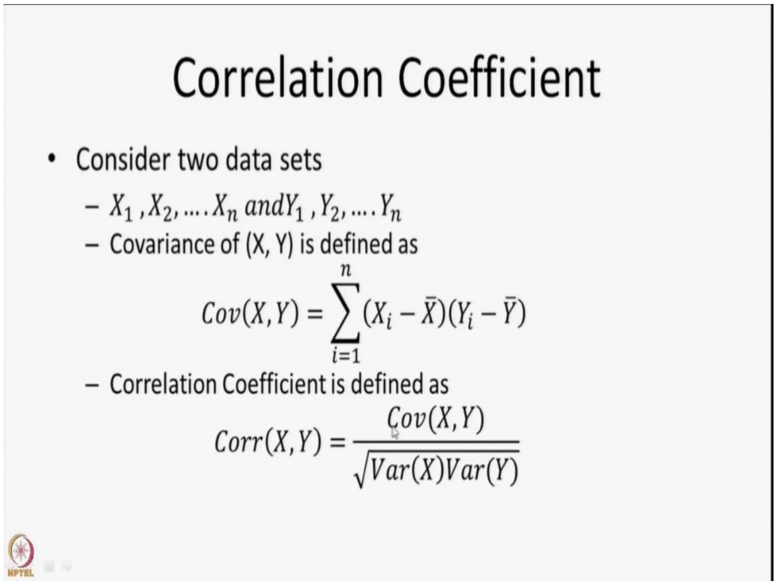


Negatively skewed

Mean < Median< Mode

So if we quickly go through it, if it is a symmetric curve, perfect symmetry occur distributed in a perfect symmetry mean = median = mode. If it is positively skewed then mode is smaller than median then is smaller than mean and if it is negatively skewed then mean is smaller than median is smaller than mode.

(Refer Slide Time: 16:47)



Correlation Coefficient

- Consider two data sets
  - $X_1, X_2, \ldots. X_n \; and Y_1, Y_2, \ldots. Y_n$
  - Covariance of (X, Y) is defined as

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

  - Correlation Coefficient is defined as

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

Next, we would like to define something called correlation coefficient. So far we were dealing with one data set. Now suppose you have two data sets. We have two data sets: X_1, X_2, X_3, X_n and Y_1, Y_2, Y_3, Y_n. The covariance of X and Y. You know that what is the variance of X which is a variation within X. Variance of y shows the variation within Y with respect to its mean value $\bar{Y}$ and here it is with respect to its mean value $\bar{X}$. Covariance means that how do the two together vary and therefore the covariance is defined

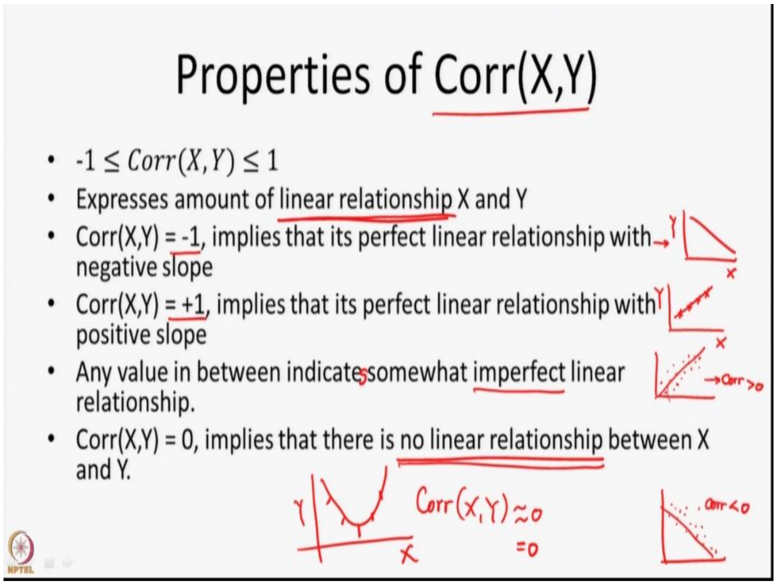
$$Cov(X,Y) = \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$$

As it can be seen that this value is not necessarily positive. Actually, we find that covariance can be negative and it can be positive. It can be anything. Why? The correlation coefficient is then defined as

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

Remember variance is a square of unit of X. Variance of Y unit of variance of Y is a square of unit of Y and therefore when you take a square root it becomes a multiplication of unit of X with unit of Y. Same unit with covariance and therefore correlation is a unitless quantity.

(Refer Slide Time: 19:20)



Properties of Corr(X,Y)

- $-1 \le Corr(X,Y) \le 1$
- Expresses amount of linear relationship X and Y
- Corr(X,Y) = -1, implies that its perfect linear relationship with negative slope
- Corr(X,Y) = +1, implies that its perfect linear relationship with positive slope
- Any value in between indicates somewhat imperfect linear relationship.
- Corr(X,Y) = 0, implies that there is no linear relationship between X and Y.

It can be very easily shown that correlation lies between - 1 and + 1. It actually also expresses the amount of linear relationship, very important, it actually expresses the linear relationship between X and Y. I will talk about it a little later. The correlation of X and Y is minus 1. It implies that it is a perfect linear relationship with a negative slope. So if I want to draw a small picture here. This case means that if you draw the perfect relationship, this is X and this is Y, then the relationship will be like this when the correlation is perfectly - 1. When the correlation is perfectly 1 between X and Y the relationship will be a straight line with a positive slope.

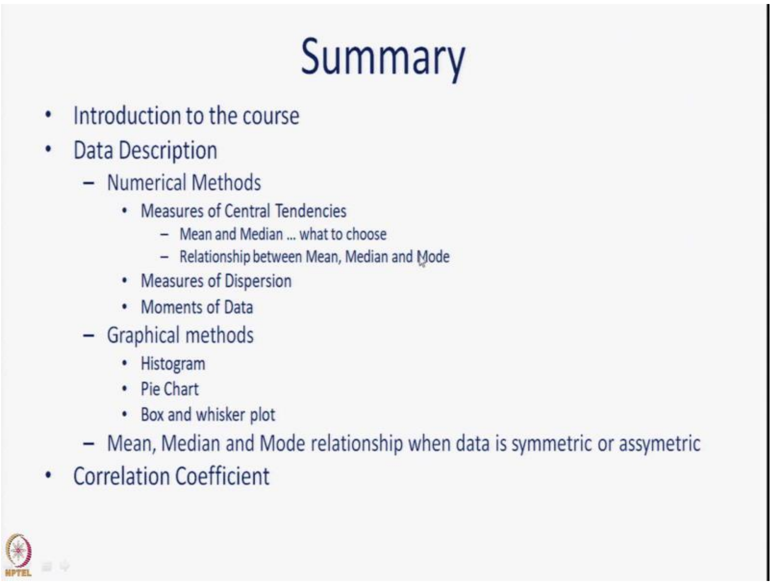
Any value in between indicates, any here there is a spelling mistake it should be indicates. Somewhat imperfect linear relationship. So for example, if we have a set of data which goes in this manner you can see that there is a linearity in its trend but it is not perfect linear. If it is a perfect linear, you will find that here the data would fall perfectly on the line but it is not perfectly on the line. So there will be some approximate line going on and that line may have some relationship and that is it is called imperfect relationship. It is an imperfect relationship.

When correlation is 0 it implies that there is no linear relationship between X and Y. So this is very important. There is no linear relationship. That is why I have earlier also noted that it is a linear relationship and this says that there is no linear relationship. For example if I take a perfect parabola, you will find that correlation between X and Y would be approximately 0. This is X and this is Y. If you take exact symmetric data points on this parabola, you will find it will become exactly 0. So it is very important to realize that correlation, the correlation coefficient explains only linear relationship between X and Y.

If it is - 1, it is perfectly negative slope. If it is + 1 it is perfectly positive slope. Anything in between is somewhat imperfect relationship. Again, if the imperfect relationship is like this, your correlation coefficient will be greater than 0 but less than 1 and if you have a case in which your data is distributed somewhat like this, then it is going to be correlation is going to be less than 0 but of course greater than - 1.

So it gives you an idea whether relationship is with a positive side or negative side, but it will not give you perfect 1 or perfect - 1 and when it gives you perfect 0, please understand that it only says that there is no linear relationship.

(Refer Slide Time: 24: 04)



Summary

- Introduction to the course
- Data Description
  - Numerical Methods
    - Measures of Central Tendencies
      - Mean and Median ... what to choose
      - Relationship between Mean, Median and Mode
    - Measures of Dispersion
    - Moments of Data
  - Graphical methods
    - Histogram
    - Pie Chart
    - Box and whisker plot
  - Mean, Median and Mode relationship when data is symmetric or assymetric
- Correlation Coefficient

So with this, we complete the sections on descriptive statistics. Let us quickly go through it. First, we had an introduction to the course and then we had a data description. We studied the numerical methods in which we looked after measures of central tendencies mean median. We also talked what to choose when. We also studied the relationship between mean, median and mode in which you can decide whether it is a positively skewed data or a negatively skewed data. If it is positively skewed data mode is smaller than mean is smaller than median is smaller than mean.

If it is negatively skewed data then mean is smaller than median is smaller than mode. Then we studied the measure of dispersions such as range standard deviation or variance and we also learned about the inter quartile range. We talked something about the moments of the data. Then we studied the histogram the pie chart. These are the graphical methods box and whisker plot, how they are useful in explaining the data before for the others to know what is the data looks like. We again it is a repetition mean median relationship of symmetric and estimated distributions. We used by doing the histogram and then finally, we introduced the correlation coefficient. With this, we conclude the chapter on descriptive statistics.