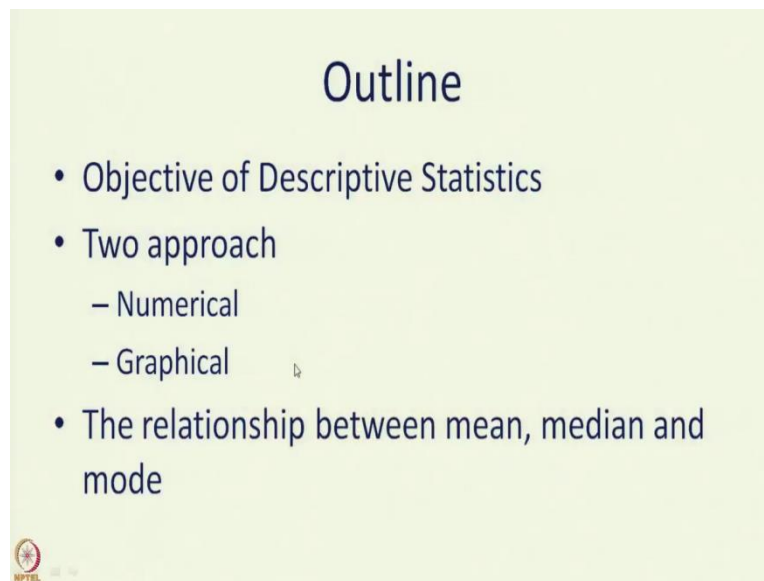


Dealing with materials of data: Collection, Analysis and Interpretation
Professor. Hina A. Gokhale
Department of Metallurgical and Material Science
Indian Institute of Technology, Bombay.
Lecture 01
Descriptive statistics-I

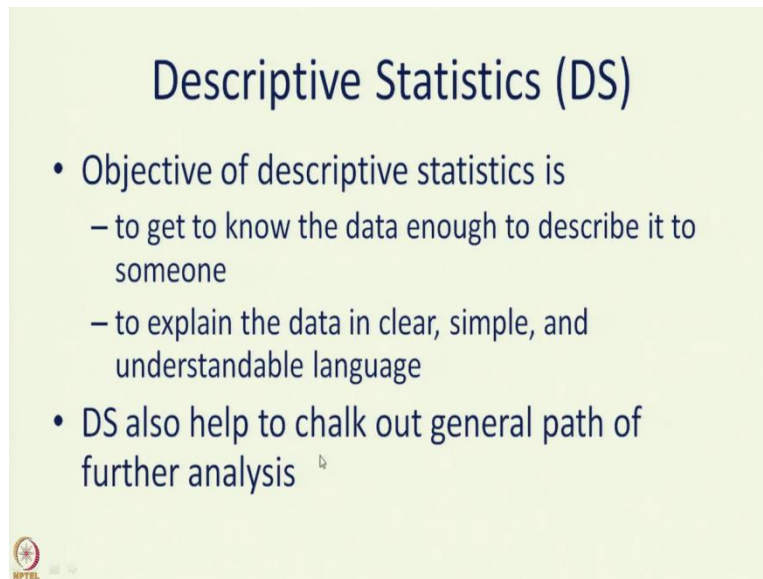
Hello and welcome to dealing with materials data course. Today we are going to cover the course or the activity called data description. This is a descriptive statistics. Let us see how we are going to cover it.

(Refer Slide Time: 0:42)



First we would like to know, the objective of finding descriptive statistics for the data. Second, the two approaches that we have, there are two approaches here, numerical and graphical and then we would like to know certain relationship between known descriptive statistics such as mean, median and mode.

(Refer Slide Time: 1:03)



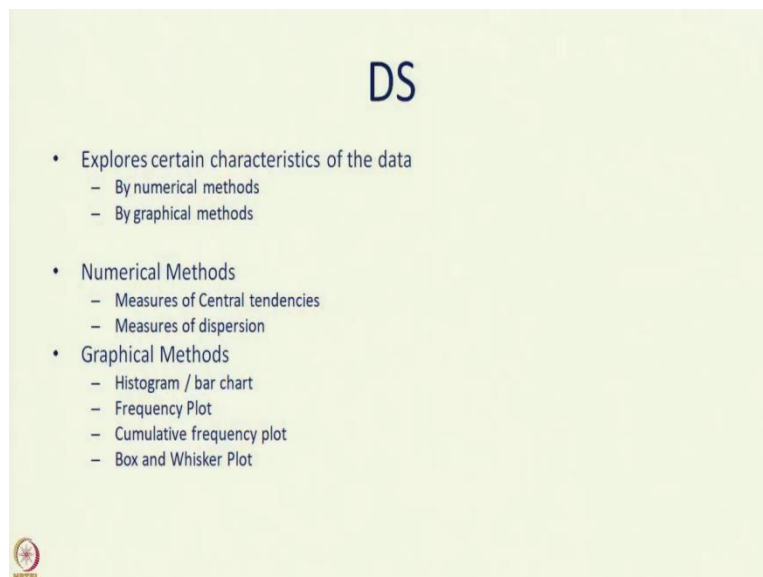
Descriptive Statistics (DS)

- Objective of descriptive statistics is
 - to get to know the data enough to describe it to someone
 - to explain the data in clear, simple, and understandable language
- DS also help to chalk out general path of further analysis

NPTEL

The descriptive statistics, primary objective is to describe your data to a totally unknown person in a clear, simple and understandable language. It also helps when you go through this descriptive statistics. It gives you an idea how you are going to move in future for further analysis in statistics.

(Refer Slide Time: 1:42)



DS

- Explores certain characteristics of the data
 - By numerical methods
 - By graphical methods
- Numerical Methods
 - Measures of Central tendencies
 - Measures of dispersion
- Graphical Methods
 - Histogram / bar chart
 - Frequency Plot
 - Cumulative frequency plot
 - Box and Whisker Plot

NPTEL


What does descriptive analysis do? It explores a certain characteristics of the data. It explores this one by numerical method or it also explores it by graphical methods. Among the numerical methods, the second methods play a very important role. They are called measures of central tendencies and the measure of dispersion. While graphical methods you must have seen in

various newspapers in which they give a plot. They give a histogram or a bar chart, they give a frequency plot, they give a cumulative frequency plot and we do not get to see this very often in the general literature, but box and whisker plot is very common among the scientifically literature to represent the data.


(Refer Slide Time: 2:52)

Numerical Methods

- Mean / Average
 - Let X_1, X_2, \dots, X_n be n data points, then mean of the data is defined as

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$


- Mean provides the central value about which the data is spread out.
- Drawing an equivalence from Physics Mean value of data is like centre of gravity of the matter



So let us move on. First we would like to discuss different numerical methods to describe the data or to explore the data. Let us assume now that $X_1, X_2, X_3, \dots, X_n$ be n data points and then we all know that the mean value or the average value of the data is defined as

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

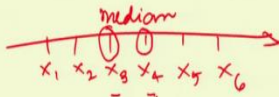
This is called a mean or arithmetic average or simply average of the data points. It gives a kind of a central value if you look at it carefully. What it really does is that if you imagine that you have data values on this straight line distributed all over the place, then generally the mean value \bar{X} is fixed right in the center of the data. Your data is spreaded all over.

It sits in the center of the data and thus it is called a measure of center tendency. If you look at it carefully, it actually draws an equivalence from the physics of what they call? Physics we just call the gravity of matter, center of gravity of the matter. Just as in any matter there is a center of gravity. Similarly for any data, it is centered around its mean value and thus it has a certain similarity between the physics where we say center of gravity and mean value of the data in statistics.

(Refer Slide Time: 4:47)

Numerical Methods

- Median is the value which divides the data in two halves
 - Let X_1, X_2, \dots, X_n be n data points,
 - Order the data values $Y_1 \leq Y_2 \leq \dots \leq Y_n$ 50% ← → 50%
 - If the number of data points is odd, sample median is the value in the position $(n+1)/2$
 - If the number of data points is even, sample median is the average of values in positions $n/2$ and $(n+1)/2$



There is another such value and that is called the median. Now median is a, another center tendency value in which it actually divides the data into two half. So if you have $X_1, X_2, X_3, \dots, X_n$ to be your n data points.

We ordered this value. Remember when you get the X_1, X_2, \dots, X_n value, they need not to be ascending or descending order. So suppose you ordered them down and call their Y_1 less than or equal to Y_2 , less than or equal to n . Basically X_1, X_2, X_3 and Y_1, Y_2, Y_n are same but Y_1, Y_2, Y_n are ordered. Once you ordered them, If you have odd number of data points, that is if your N is odd, you pick exactly the middle value.

You pick, $(n+1)/2$ value, that value is called the median of the data. In other words, it will divide the data. This side, 50 percent of the data will be there and this side also 50 percent of the data will be there. So this is called the median value.

Now suppose if we have the point are even in number, then you have to take an average value because it does not have any middle value. There are two middle values. So when there are two middle values, you have to take an average of the value which is sitting at $n/2$ point and the value that sits with $(n+1)/2$. So for example, if we have X_1, X_2, X_3, X_4 and X_5 . These are all ordered because this is a real line and therefore these are all ordered, then this is your median value. But if you have one another point X_6 then the median value is average of X_3 and X_4 this is what it shows.

(Refer Slide Time: 7:23)

Mean or Median?

- Both the measures provide “middle” value of the data, so how do they compare?
 - Median is robust against extreme values in the data,
 - While Mean is affected by extreme value
- Example: Let 8.0, 9.0, 10.0, 11.0, 12.0 be five data points.
 - Mean = 10.0 and Median = 10.0
 - Replace 12.0 by 18.0
 - Mean = 11.2, but the median = 10.0



Now, the question is that mean or median? What to choose? A very common question because both of them in a way provide a middle value of the data. One divides the data in 50-50 while the other divides the data in such a way that it becomes a center of the data. We find that when you compare from their value, it makes a difference. That median is called a robust against extreme values in the data. While mean is affected by the extreme values of the data.

Here we have shown it through a very simple example. Let us take five data points in our data which are 8, 9, 10, 11, and 12. What is the mean value? Average is going to be 10 and the median value, because there are five data points, the middle value we have to pick up, so the median is also 10. Now just replace this 12 by 18.0 extreme value, it says.

So extreme value it means that either it is very large or it is very small. In this example, we are changing the larger side value to an even further large value. Now you see the mean value will become 11.2 while the median is always a middle value and therefore it always remains 10. So this is what it says, that median is robust, against the extreme value while the mean value tends to get affected by the extreme value.

So this is the difference. These kinds of differences one needs to use very selectively or very thoughtfully, when you apply the centre tendency value. For example, if you are considering a case in which you are grading a person, among 10 people or among 12 people and you would like to know that 12 people have graded a single individual and have given different grades.

Now if someone wants to favour the individual may suddenly give a very high value or if he does not want to favour the person, she may give him a very small value. In that case, if you

take average as your central tendency, the mean value, then that person's perception, which is a sort of a biased perception will effect. But if you take in such cases a median value, it will not affect. So such in such situation where you know board is examining a person or student or a candidate for a post or a candidate for a promotion. It is wordy, very common to use median as a measure against mean.

(Refer Slide Time: 10:57)

Numerical Methods

- Mode:
 - Mode is a value in data that occurs with highest frequency
 - It's the most probable value of the data
 - It is possible to have data that has more than one Mode value. Such a data is called multimodal.
- Other statistics
 - Percentiles
 - Order the data set in ascending order
 - Then, P_k is called 1st percentile if 1% of points lie below this value
 - Similarly, P_k is called kth percentile if k% of data points lie below this value, where $0 \leq k \leq 100$
 - Quartiles
 - P_{25} is also called 1st quartile Q_1
 - P_{75} is also called 3rd quartile Q_3
 - P_{50} is median

Handwritten notes: mean, median, mode (underlined in red).
 Diagrams: $Y_1 \leq Y_2 \leq Y_3 \dots \leq Y_n$ with P_k and Q_1, Q_2, Q_3 marked. A bar chart with a peak at the median.

There is one more method which is called mode. Mode is very simple. You must have seen the frequency plots. We are going to talk about it at some later time. There is, you can see that there are, you come with a bar chart in which the frequency plots are given like this. In that case it says that mode is the value where the highest frequency is measured. It's the most probable value of their data because it has the highest probability of occurrence and therefore it has the most probable value. It is called the most probable value of the data.

It is possible to have more than one modal value and such in a data and such data is called multimodal. In future, in this course, now itself when we come to frequency plots, we will show you one plot from marterials data where we do see the bi-modal and this two modal distribution actually tells us that how to look into the analysis further how to, you know, segregate the data in future. There are other statistics involved in it. Such as there is this percentiles. It means that once again, you ordered the statistics in an ascending order. So as I said, suppose you ordered them as, $Y_1 \leq Y_2 \leq Y_3 \dots Y_n$

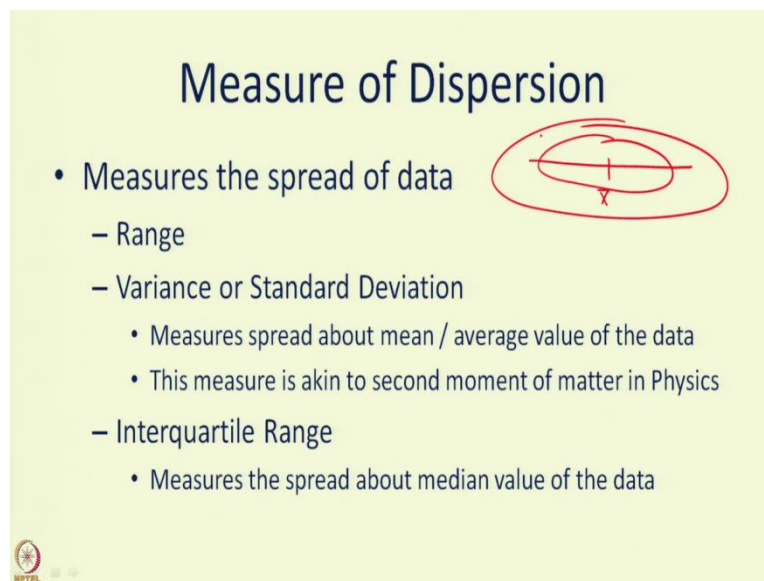
Then the first percentile, P_1 , P_1 is first percentile where 1 percent values lie below P_1 . So for example, if n had been ten, then first percentile would have been probably only one data point

should lie below that and therefore Y_2 would have been the first percentile. So likewise, you can have a K_{th} percentile where you pick up some value here. I call it P_K , where the number of data here are K percent of n .

The values below that value is about K percent of N . There are special names to these percentiles. P_{25} is called the first quartile Q_1 , because it divides the 25th percent data. P_{75} is third quartile and of course P_{50} is what we call median. So if you look at the data, if data is on this straight line, 50 percent of the data is divided by this data point, then it is median or we call it P_{50} or it is called Q_2 .

Here when you say this is Q_1 , it means that on this side there are 25 percent of the data and here if you call it Q_3 it means that this side it is 25 percent, this side is 75 percent of the data. So these are some of them methods measures of central tendency. There is a relationship between the three measures that we have learned. That is mean, median and mode. Before we go into it, let us talk about some graphical methods.

(Refer Slide Time: 15:33)



Measure of Dispersion

- Measures the spread of data
 - Range
 - Variance or Standard Deviation
 - Measures spread about mean / average value of the data
 - This measure is akin to second moment of matter in Physics
 - Interquartile Range
 - Measures the spread about median value of the data

The slide features a hand-drawn red diagram of a bell curve with a vertical line through its center and a horizontal line below it, with an 'x' marking the center. A small logo is visible in the bottom left corner.

So here we are and before going to graphical methods, we have to cover measure of dispersion. See, in the central tendency we have covered the measures which actually decides the center point where the data is centered. So we have mean which is the average value. We have a median which divides the data in 2 parts 50-50. And we have a mode which gives you the maximum data, maximum frequency of a data.

Now we would also like to know how the data is spreaded. For example, why is it important? Because if you know what is the mean value, it is insufficient. As we saw that, you know what is the maximum and what is the minimum within which area it is spreaded. It is like this or is it like this? You will see that sometimes you have the same mean value but the data could be spreaded on a larger scale so we must know what is called a measure of dispersion. There are three major dispersions we are going to discuss here. They are called range, variance or standard deviation and the interquartile range.

(Refer Slide Time: 17:04)

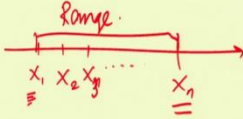
Measure of Dispersion

- Range = $M - m$, where,
 - $M = \max\{X_1, X_2, \dots, X_n\}$
 - $m = \min\{X_1, X_2, \dots, X_n\}$
- Variance

$$- S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{(n-1)}$$

mean squared distance of data from mean

 - Standard Deviation = $S = \sqrt{S^2}$
- Interquartile Range : $Q_3 - Q_1$



Let us go to each one of them one after the other. Range as it says, if M is the maximum of your data value and small m is the minimum of the data value, then maximum minus minimum gives you the total range of data this is one measure of dispersion. So if you have a data on this straight line which is X_1, X_2 . Now please note that this is, I am writing it as an ordered data and this is X_n then this X_n minus X_1 is the range of the data.

Please note that I have assumed that maximum is X_n and minimum is X_1 which need not be true all the time but this is for the simplicity I have shown it in this way. Another measure which is most commonly used measure is Variance. Variance is defined as the mean squared distance of data from mean.

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{(n - 1)}$$

Now this (n-1), why are we taking (n - 1)? You can roughly understand that out of n data points freedom that we had, 1 parameter of \bar{X} is already been calculated and therefore we are

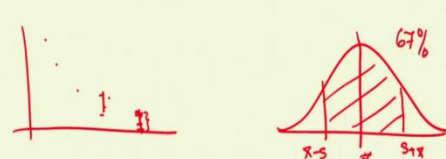
dividing it by $(n - 1)$ but for time being we take it as a formula and this is the formula for S^2 square and the standard deviation is the square root of a S^2 .

The next measure is interquartile range. In the previous slide we defined what is the third and the first quartile. The distance between two is called an interquartile range. Roughly speaking variance gives you dispersion around the mean value and interquartile range gives your dispersion around median value.

(Refer Slide Time: 19:57)

Standard Deviation

- Standard Deviation (sd) is most commonly used measure of dispersion.
 - Under the assumption of Normality the range of $\bar{X} \pm S$ covers 67% of the data.
 - Hence this is commonly used to show possible error in the observed value of data \bar{X}



The slide contains two hand-drawn diagrams in red. The left diagram shows a scatter plot with a vertical error bar extending from a central point. The right diagram shows a normal distribution curve with a shaded area between $\bar{x} - S$ and $\bar{x} + S$, labeled 67%.

This standard deviation, as I said is most commonly used measure of dispersion and in this course we would like to emphasize it because the data representation with an error bar actually represents the standard deviation that you found in the data. So under the assumption of normality that is or distribution is, the data is distributed very nicely as a bell shape curve where the mean value lies right in the center and in that case it says that if you take $\bar{X} + S$ value and you take $\bar{X} - S$ value in this area about 67 percent of your data will lie.

So when you measure any value of your experimental result, then that value if you give, if you find it standard deviation and in plotting on a graph paper, suppose you want to show your data points like this, you can show it with an error bar in this manner, which actually shows that your data lies right in the center and at the most it can make an error of plus or minus 1 Sigma. And that way 67 percent of your data is covered in this. Therefore, it is very common to show an error bar against every data point and this error bar is generally 1 standard deviation of away.

(Refer Slide Time: 22:16)


Graphical Methods

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



Now we move on to graphical methods. We will cover mainly four kinds of graphical methods, which are most commonly used. Histogram or bar chart, which is also known as frequency plot, Pie chart, You must have seen it number of times and cumulative frequency plot and finally box and whisker plot. As I said, the histogram, Pie chart these are very commonly seen in any general literature such as newspaper or any general scientific journal or any business journal. These are very commonly used graphical methods. Cumulative frequency plot you get to see more frequently in any many of the financial matters. Box and whisker is a plot which is very commonly used among scientific data representation.

(Refer Slide Time: 23:29)

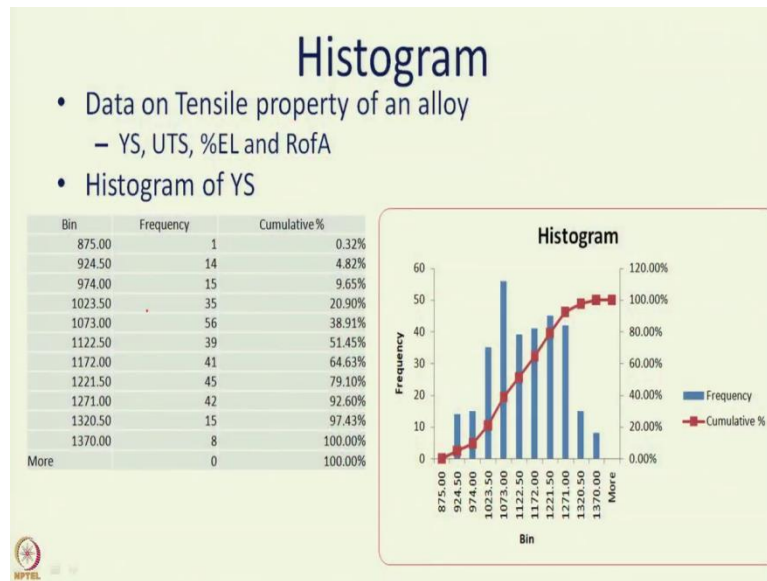


Code 2	Code 1	TEMP	YS	UTS	EL	ROFA
4.00	5.00	25	1220.00	1483.00	13.00	32.00
4.00	5.00	25	1250.00	1499.00	16.00	27.00
4.00	5.00	650	1007.00	1125.00	13.00	34.00
4.00	5.00	650	1014.00	1141.00	16.00	35.00
4.00	5.00	650	1059.00	1258.00	17.00	38.00
1.00	5.00	25	1185.00	1395.00	20.00	46.00
1.00	5.00	650	970.00	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 let us move throughout this graphical representation we are going to use this data. This shows a very partial data which covers certain aspects of our data. There is a code 1, code 2, some codes are given. There is a temperature given. These are the materials, yield strength, ultimate tensile strength, percentage elongation and percentage reduction of area. So this is a typical materials data and we will cover the graphical methods using this particular set of data.

Before going further, I would like to summarize. What we have done is we are discussing to use descriptive statistics to describe the data to a common person. There are two methods of doing it. So far we have covered the method, which is the method of numerical. In other words, numerically you can calculate certain values to describe the data. The first was central tendency in which we covered mean or arithmetic average, median, and more. Then we also covered how the data is spreaded and that we covered through measure of dispersion by covering the range, the standard deviation and the interquartile range.

(Refer Slide Time: 25:22)



In the next session, we will cover the graphical methods, which will be histogram, frequency plots, Pie chart and Box and whisker plot. Thank you.