Basics of Mechanical Engineering-3

Prof. J. Ramkumar

Prof. Amandeep Singh Oberoi

Department of Mechanical Engineering

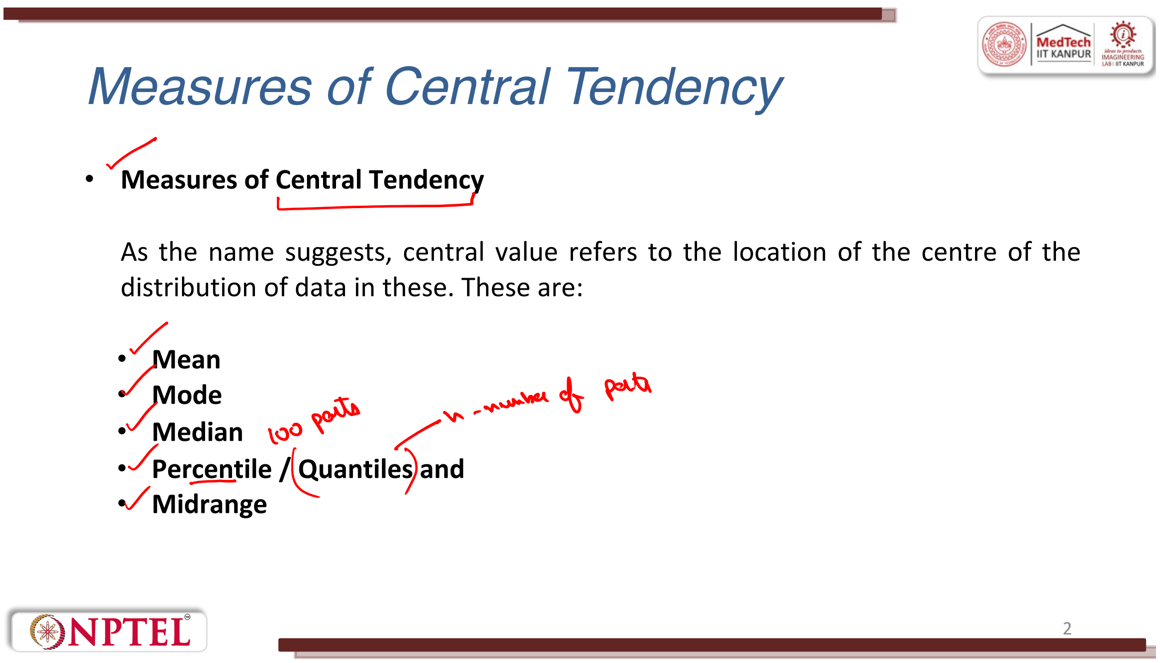
Indian Institute of Technology, Kanpur

Week 12

Lecture 51: Basics of Descriptive Statistics

Welcome to the second part of the lecture series on Engineering Statistics. We talked about the basics of engineering statistics in part one. We will talk about the kinds of statistics, the steps in conducting a statistical analysis, and introduce you to statistical terms as well. In this part, I will try to talk about the basics of descriptive statistics. As we discussed in the previous part, what is descriptive statistics?

It only describes the data. It helps in organizing and analyzing the data in its present form. But in a meaningful manner, we explain the already known data, limited to a sample or population, whatever we are trying to present, and we try to display the data in the form of graphs or tables. Now, how do we present that in terms of graphs and tables? And what are the general calculations that we do in descriptive statistics? I will try to go through them in this lecture.



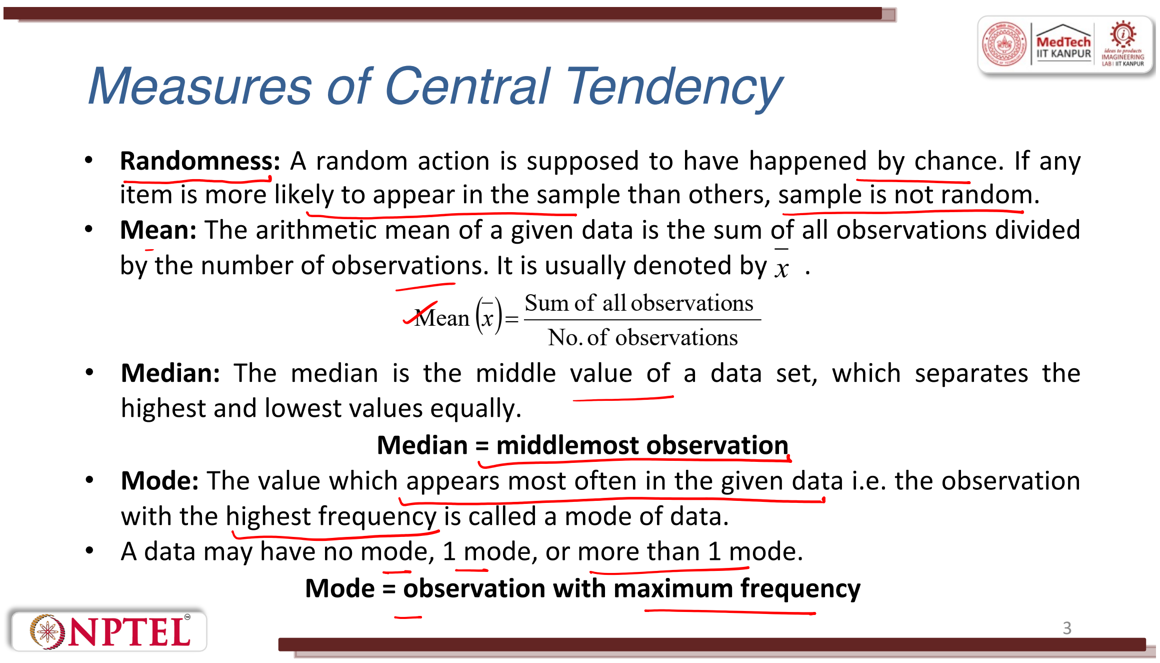
Measures of Central Tendency

- **Measures of Central Tendency**

  As the name suggests, central value refers to the location of the centre of the distribution of data in these. These are:

  - **Mean**
  - **Mode**
  - **Median** 100 parts
  - **Percentile / Quantiles and** n - number of parts
  - **Midrange**

First is measures of central tendency. I talked about measures of central tendency in the previous lecture as well. When we talk about the centering of the data, to what point is the data centered? It could be based upon mean, mode, median, percentiles or quantiles, and mid-range. Percentile is about cent, percent. It is about 100 parts. Quantile is a number of parts. If it is a quartile, it is four parts. This we will discuss.



Measures of Central Tendency

- **Randomness:** A random action is supposed to have happened by chance. If any item is more likely to appear in the sample than others, sample is not random.
- **Mean:** The arithmetic mean of a given data is the sum of all observations divided by the number of observations. It is usually denoted by $\bar{x}$.

$$\text{Mean}\left(\bar{x}\right) = \frac{\text{Sum of all observations}}{\text{No. of observations}}$$

- **Median:** The median is the middle value of a data set, which separates the highest and lowest values equally.

Median = middlemost observation

- **Mode:** The value which appears most often in the given data i.e. the observation with the highest frequency is called a mode of data.
- A data may have no mode, 1 mode, or more than 1 mode.

Mode = observation with maximum frequency

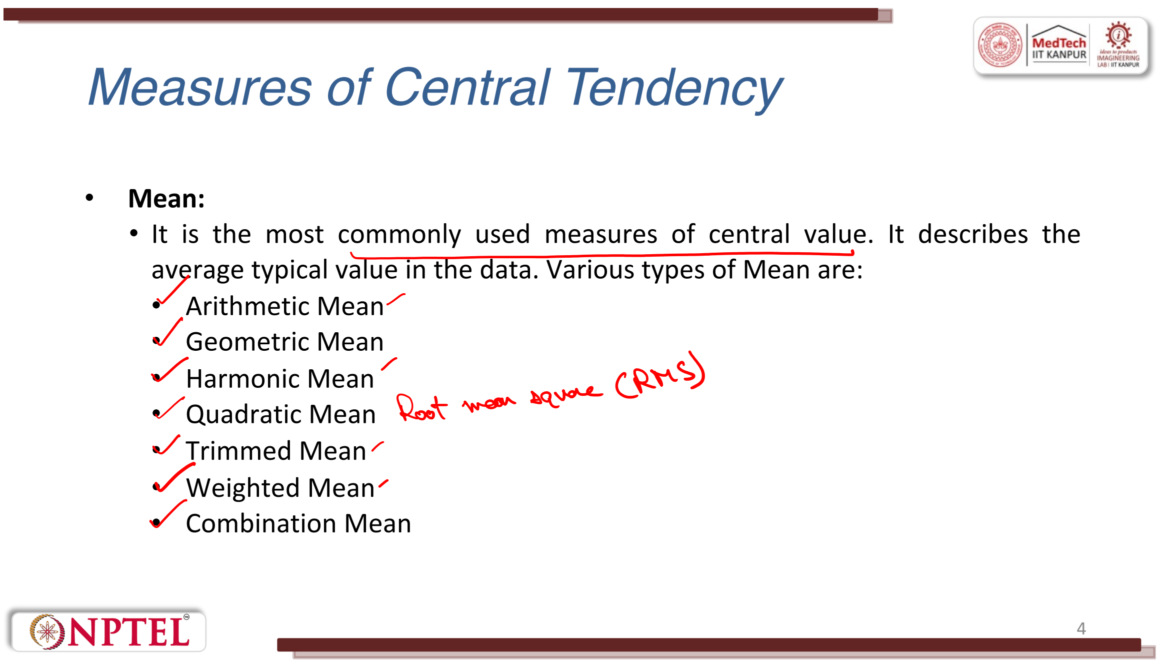
Let me first go through some terms in the measures of central tendency. Randomness is the most important term here. A random action is supposed to have happened by chance. If any item is more likely to appear in the sample than others, then the sample is not random. So, whenever we take any data set, whenever we are trying to take the mean of any set of observations that we are taking,

We always consider that observations are taken at random. That is, one observation is not dependent upon another. For instance, if I say the roughness value is 2.5 micrometers, I have been taking this as an average of the 10 samples which have been taken; all the samples are now considered to be completely random. If I take, suppose, the sample manufactured first in the morning and the sample manufactured last in the evening, There could be a difference depending upon the machine performance.

So then it doesn't stay random. We have to give more weight to the sample in the evening than to the one in the morning. Then the kind of mean we take there is not simply a mean; it could be a weighted mean. Then, the mean is the arithmetic mean given by the data:

$$\text{Mean}\left(\bar{x}\right)=\frac{\text{Sum of all observations}}{\text{No. of observations}}$$

I will give you an example. The median is the middle term, the middle value; that is, the median is the middlemost observation. The mode is the value that appears most frequently in the given data. The observation with the highest frequency is called the mode. The data may have no mode, one mode, or more than one mode. The mode = observation with the maximum frequency.



Measures of Central Tendency

- **Mean:**
  - It is the most commonly used measures of central value. It describes the average typical value in the data. Various types of Mean are:
    - ✓ Arithmetic Mean
    - ✓ Geometric Mean
    - ✓ Harmonic Mean
    - ✓ Quadratic Mean Root mean square (RMS)
    - ✓ Trimmed Mean
    - ✓ Weighted Mean
    - ✓ Combination Mean

What is the mean? It is the most commonly used measure of central value. It describes the average or typical value in the data. Various types of means are the arithmetic mean, or whatever mean you generally talk about, like the arithmetic mean. You simply take the sum of the observations and divide it by the number of observations. That is the arithmetic mean. Geometric mean. Geometric mean, in place of the sum, we take the product of the values. So, the geometric mean is the nth root of the product of n numbers. For example, if there are 10 numbers,

we multiply them. Then, we take the 10th root of that. That is the geometric mean. Then comes the harmonic mean. The harmonic mean is generally the mean of the ratios or the reciprocal of the arithmetic mean.

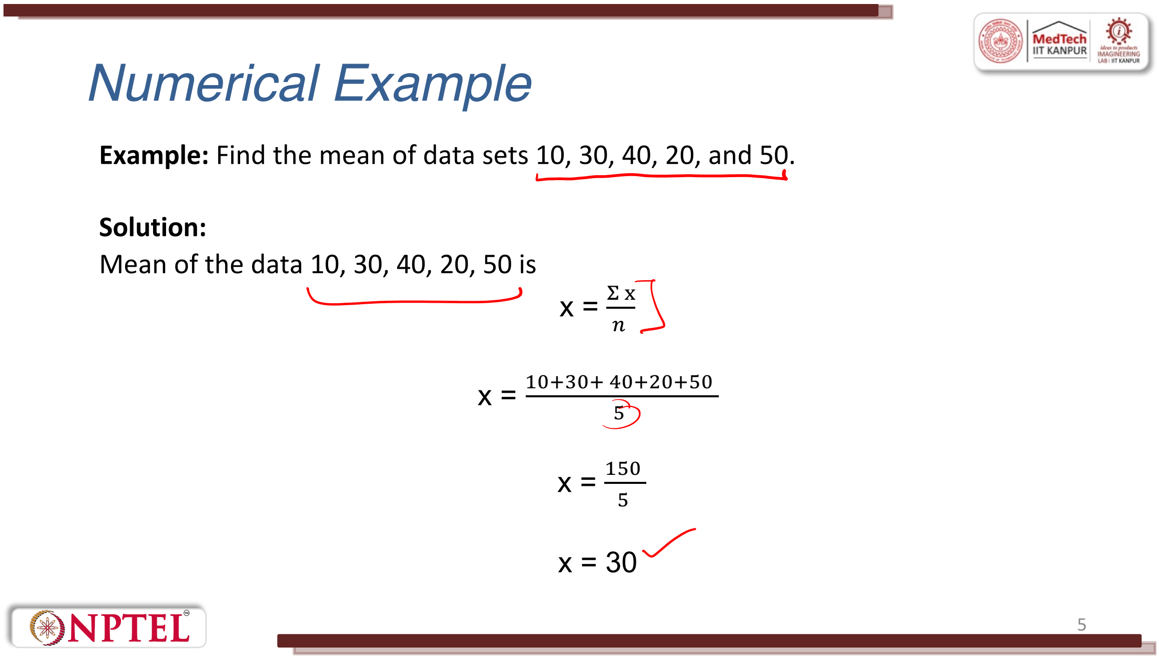
When I say mean of the ratios, it refers to terms that have a ratio between two numbers. For example, speed is equal to distance divided by time. For example, density is equal to mass per unit volume. There, the harmonic mean would come into play. Quadratic mean.

Quadratic mean—the word itself says it is root mean square. That is, we call it the RMS value. Root mean square value—when we talk about standard deviation, root mean square is also used there sometimes. So, it is calculated by squaring each value, averaging these values, then taking the square root of them. Then comes trimmed mean.

When I say trimmed, trimmed means we remove a predefined percentage of the lowest and highest values from the dataset. For example, if I say in the morning, the machine is completely fresh, it is performing really well. Units are being produced, and so the nozzles are being produced in one day. In the evening, its performance—let me consider—is going lower. Then whatever happens between 4 to 5 p.m. in the evening, that I'm trimming off.

That is, I clean the data that is trimmed. Then comes weighted mean. Another way to deal with a dataset that is not completely random is weighted mean. I give more weight to the nozzles manufactured in the morning and less weight to those manufactured in the evening if the machine performance declines from morning to evening. That is weighted mean.

Then comes combination mean, a combination between any two types of the means. It could be between arithmetic and harmonic mean, or between trimmed and weighted mean, so that we can describe the data more appropriately for our requirements. These are the kinds of means. You might have gone through arithmetic, geometric, harmonic, trimmed, and weighted means, but combination mean might be new for you.



Numerical Example

Example: Find the mean of data sets 10, 30, 40, 20, and 50.

Solution:

Mean of the data 10, 30, 40, 20, 50 is

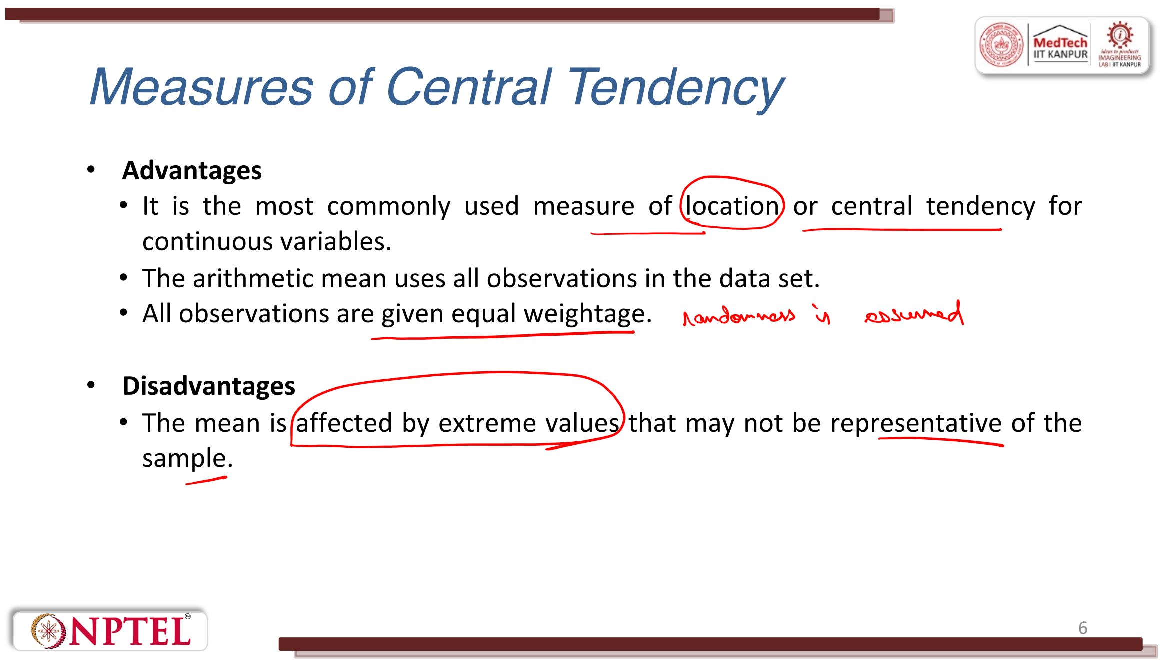
$$x = \frac{\Sigma x}{n}$$

$$x = \frac{10+30+ 40+20+50}{5}$$

$$x = \frac{150}{5}$$

$$x = 30$$

Here is a simple dataset: 10, 30, 40, 20, 50. To find the mean, we take the sum and divide it by the number of observations. The mean is 30. It is a simple calculation.



Measures of Central Tendency

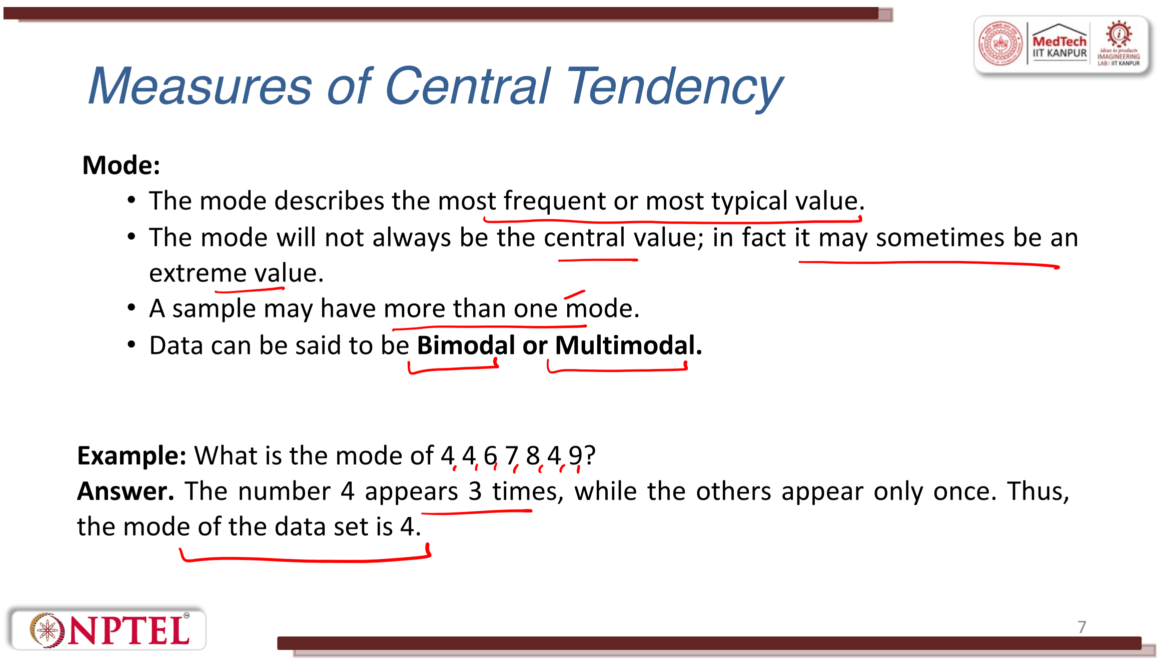
- **Advantages**
  - It is the most commonly used measure of location or central tendency for continuous variables.
  - The arithmetic mean uses all observations in the data set.
  - All observations are given equal weightage. randomness is assumed

- **Disadvantages**
  - The mean is affected by extreme values that may not be representative of the sample.

The advantages of the mean are: it is the most commonly used measure of location or central tendency. Measure of location is also one of the terms used for measure of central tendency for continuous variables. Arithmetic mean uses all observations in the dataset. All observations are given equal weightage; that is, randomness is assumed.

When we take the mean, when we talk about the mean, we generally or always refer to the arithmetic mean throughout this week. A disadvantage is that the mean is affected by extreme values that may not be representative of the sample. For example, there are students in the class. If I take the heights of the students, the average height in India is around 5 feet 8 inches. If there is a student who is 7 feet tall, including that student in a class average of maybe 20 students, the overall average will go very high.

Or it could be the other way around. There could be one student who is of the height of maybe 4 feet 5 inches—a dwarf, right? Now, here, the overall average height of the students in the class is about 5 feet 8 inches. This one student out of the 20 students, if included, would reduce the overall average. So, extreme values affect the sample; this is a disadvantage of the mean.



Measures of Central Tendency

Mode:

- The mode describes the most frequent or most typical value.
- The mode will not always be the central value; in fact it may sometimes be an extreme value.
- A sample may have more than one mode.
- Data can be said to be **Bimodal or Multimodal.**

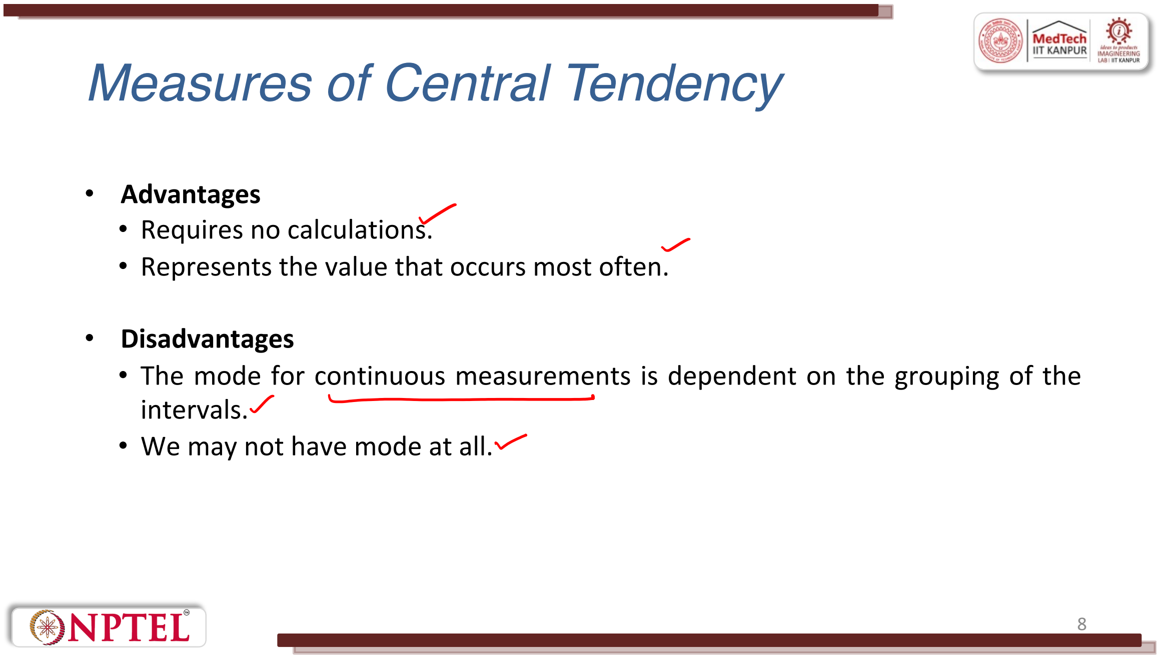
Example: What is the mode of 4 4 6 7 8 4 9?

Answer. The number 4 appears 3 times, while the others appear only once. Thus, the mode of the data set is 4.

Now comes mode. Mode describes the most frequent or most typical value. The mode will not always be the central value. In fact, it may sometimes be an extreme value. For example, 5 feet 8 inches is the average.

There could be 27 students who have the exact height of maybe 5 feet 6 inches. So that is a mode, the maximum number of repetitions in a set of samples. A sample may have more than one mode. For example, there are four students of 5 feet 6 inches. There are four students of 5 feet 7 inches.

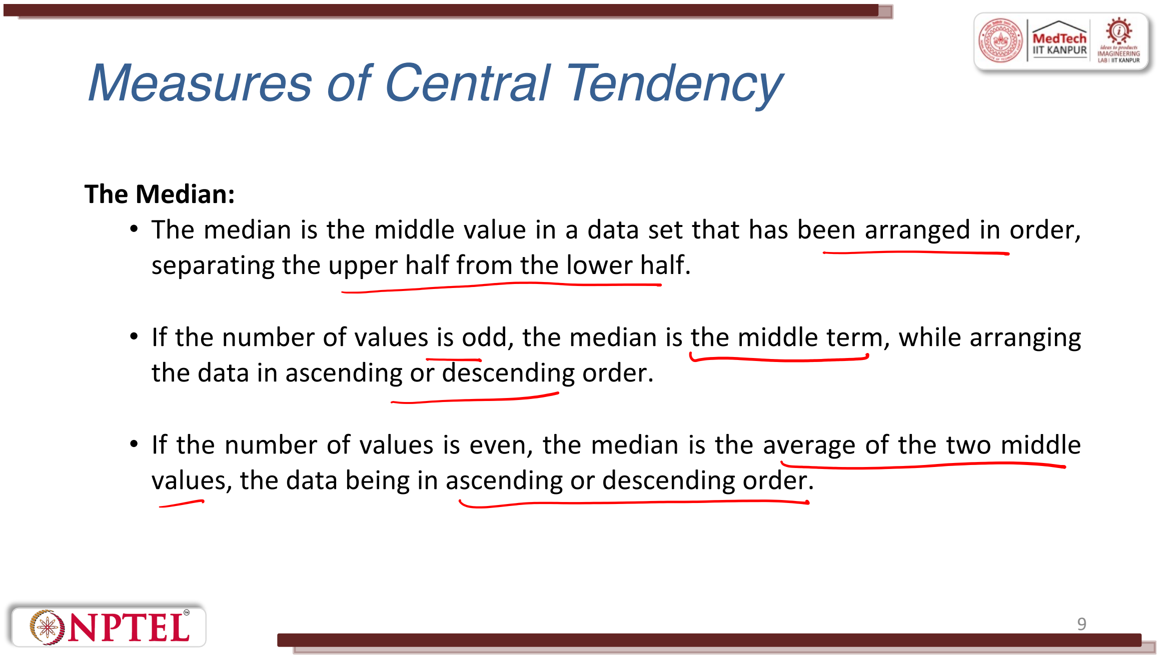
Both of them are the largest values of the repetition of the number of heights. So, that becomes more than one mode. So, the data is said to be bimodal. In this case, it could be multimodal. For example, here, in this set: 4, 4, 6, 7, 8, 4, 9. 4 is repeated three times, while others appear only once. Thus, the mode is 4.



Measures of Central Tendency

- **Advantages**
  - Requires no calculations.
  - Represents the value that occurs most often.
- **Disadvantages**
  - The mode for continuous measurements is dependent on the grouping of the intervals.
  - We may not have mode at all.

Advantages: it requires no calculations. It represents the values that occur most often. Disadvantages: the mode for continuous measurement is dependent upon the grouping of the intervals. It may not have a mode at all.

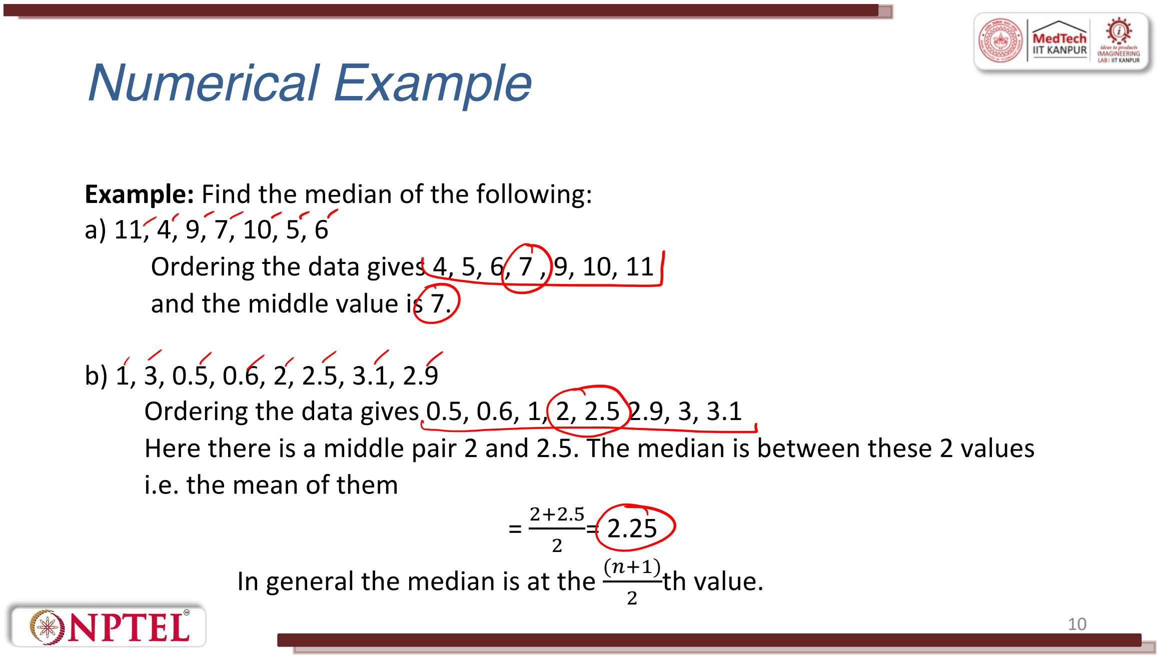


Measures of Central Tendency

The Median:

- The median is the middle value in a data set that has been arranged in order, separating the upper half from the lower half.
- If the number of values is odd, the median is the middle term, while arranging the data in ascending or descending order.
- If the number of values is even, the median is the average of the two middle values, the data being in ascending or descending order.

Next comes the median. The median is the middle value in a data set that has been arranged in order, separating the upper half from the lower half. If the number of values is odd, the median is the middle term while arranging the data in ascending or descending order. If the number of values is even, the median is the average of the two middle values, the data being in ascending or descending order.



Numerical Example

Example: Find the median of the following:

a) 11, 4, 9, 7, 10, 5, 6

Ordering the data gives 4, 5, 6, 7 , 9, 10, 11

and the middle value is 7.

b) 1, 3, 0.5, 0.6, 2, 2.5, 3.1, 2.9

Ordering the data gives 0.5, 0.6, 1, 2, 2.5 2.9, 3, 3.1

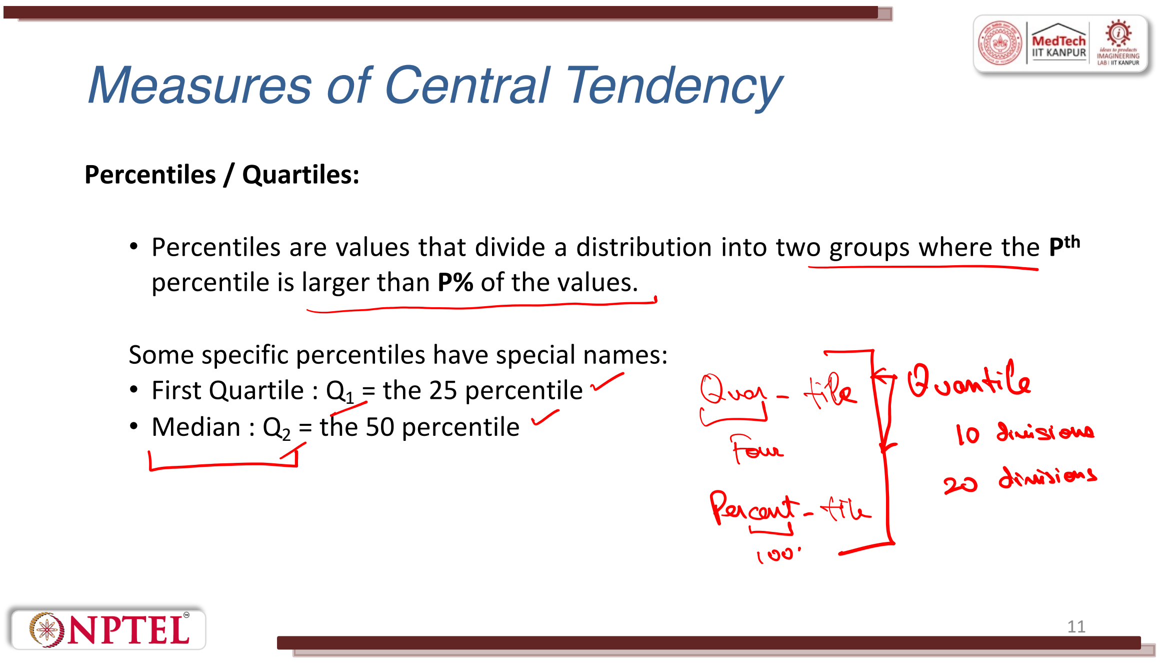
Here there is a middle pair 2 and 2.5. The median is between these 2 values

i.e. the mean of them

$$= \frac{2+2.5}{2} = 2.25$$

In general the median is at the $\frac{(n+1)}{2}$th value.

To see an example quickly, these are the terms: 11, 4, 9, 7, 10, 5, 6. So, here we see a total of 7 values. So, the middle value here, while putting them in ascending order, is 7; this is the median. So, here: 1, 3, 0.5, 0.6, 2, 2.5, 3.1, 2.9; the number of values here is 8. So, the median would be, when you put them in ascending order, the two middle terms, which are 2 and 2.5; the average of them is the median, that is 2.25.



Measures of Central Tendency

Percentiles / Quartiles:

- Percentiles are values that divide a distribution into two groups where the **Pth** percentile is larger than **P%** of the values.

Some specific percentiles have special names:

- First Quartile : Q_1 = the 25 percentile
- Median : Q_2 = the 50 percentile

Quar - tile

Four

Percent - tile

100

Quantile

10 divisions

20 divisions

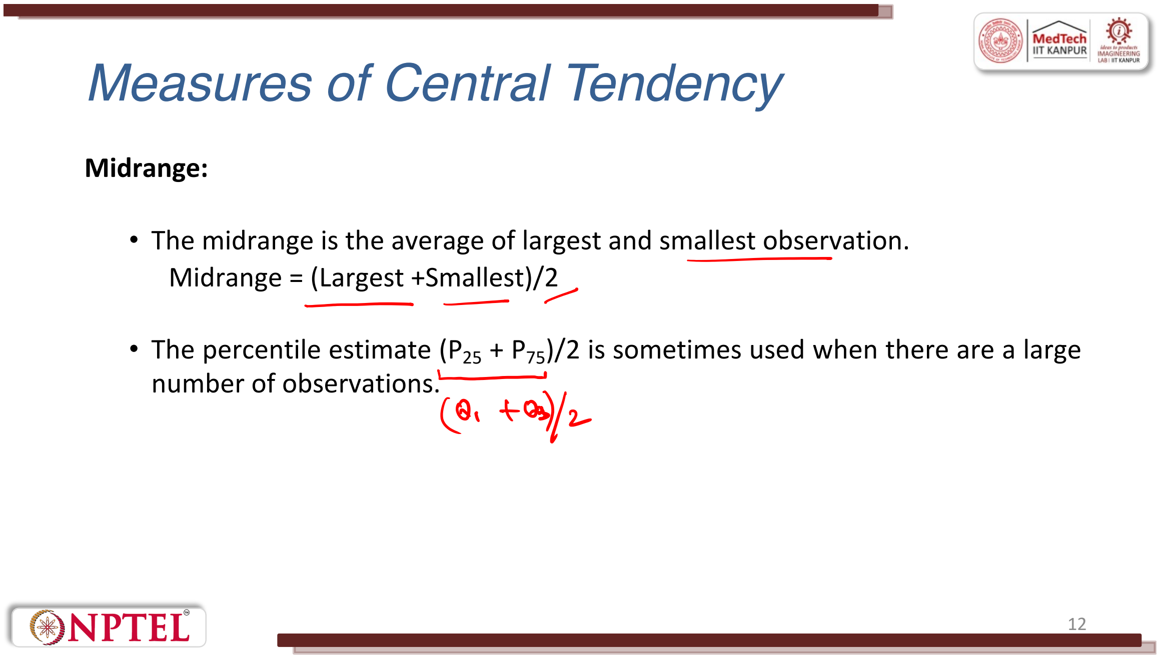
Percentiles or quartiles—talking about the median; we have to talk about quartiles as well. Percentiles are values that divide a distribution into two groups where the pth percentile is larger than p% of the values. Some specific percentiles have specific names. For example, the first quartile is Q1, and the second quartile is Q2. The first quartile is the 25th percentile, and the second quartile is the 50th percentile.

Out of the 20 values, when we talked about the number of students in the class—20 students—arranged in ascending order of height, or you say the nozzles; measurements that you have taken of the diameter, there will be some change in the diameter, some small difference in the diameter.

You arrange them in ascending order out of the 100 values; the first 25 numbers are Q1, quartile 1. Number 50 is Q2, which is the median itself. It has been given; the median is Q2. Then the 75th percentile—that is, before that, 75% of the data lies—is Q3, which is the quartile. It is Q-U-A-R; 'quar' means 4, and 'quartile' means the 4th position. Then we have percent.

'Cent,' I said, is 100. Percentile—we say it becomes 'percentile,' which means 100 divisions. Similarly, all of this is derived from a common term called 'quantile.' This quantile could be a quartile, it could be a percentile, or it could be any. We will say 10 divisions, or we may say 20 divisions.

So, that could also be taken. So, this also gives me the central location of the data and what we are trying to talk about.



Measures of Central Tendency

Midrange:

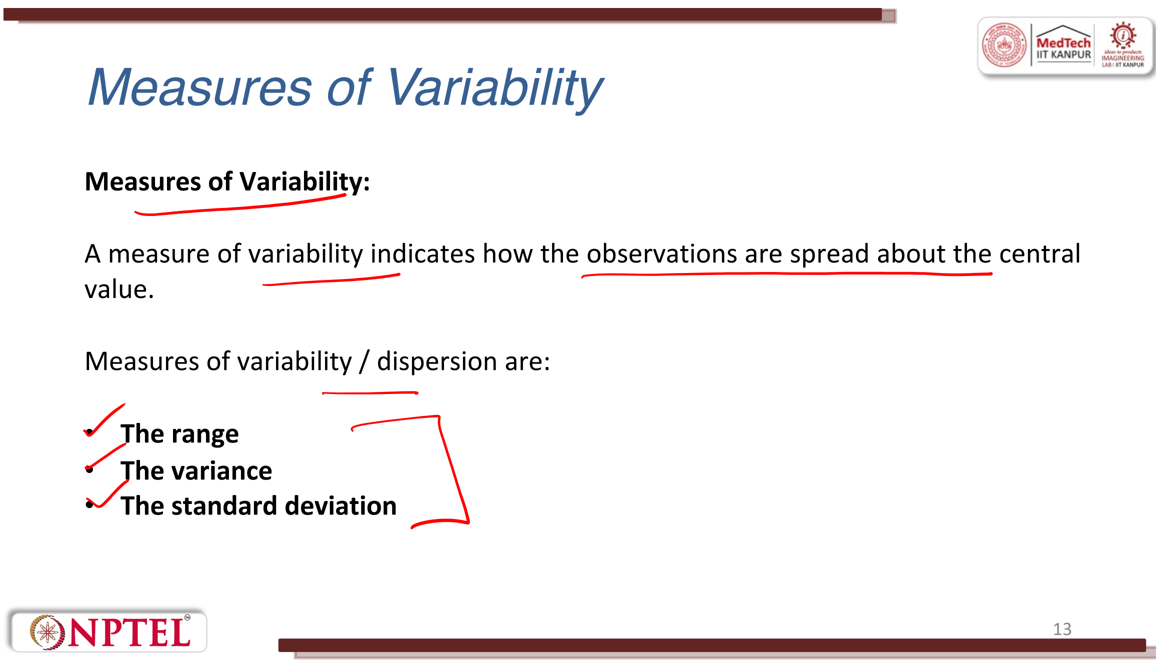
- The midrange is the average of largest and smallest observation.
  Midrange = (Largest +Smallest)/2

- The percentile estimate $(P_{25} + P_{75})/2$ is sometimes used when there are a large number of observations.

$(Q_1 + Q_3)/2$

Mid-range. Mid-range is the average of the largest and the smallest observation. It is the largest value plus the smallest value divided by 2. From the percentile estimate, mid-range is sometimes taken as the $(P_{25} + P_{75})/2$, that is (Q1 + Q3)/2.

This is sometimes used when there are a large number of observations. So, this range is also taken to have a better estimate of the mid-range. When I talk about mid-range, I am still referring to the central tendency only. It is not about range. Range would be the difference between the largest value and the smallest value. We will discuss that when we talk about the measures of variability.



Measures of Variability

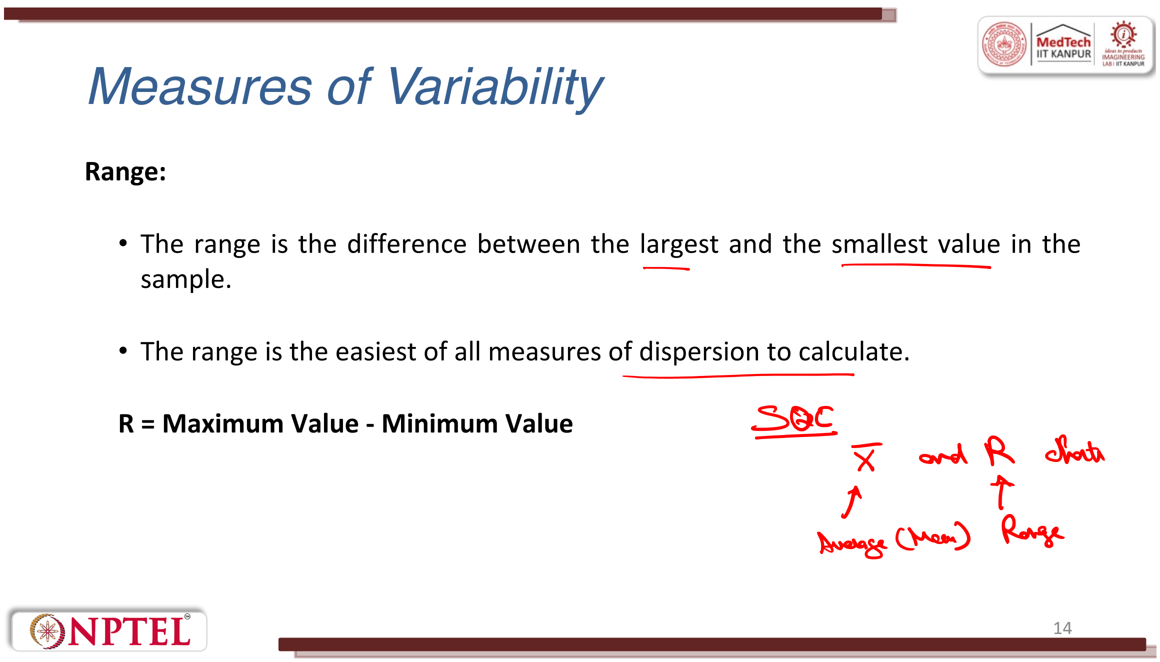
Measures of Variability:

A measure of variability indicates how the observations are spread about the central value.

Measures of variability / dispersion are:

- ✓ **The range**
- ✓ **The variance**
- ✓ **The standard deviation**

Measures of variability, as I mentioned, the central tendency is the location of the data. Then comes the spread of the data. That talks about variability. In variability, we talk about the range, the variance, and the standard deviation. Measures of variability indicate how the observations are spread about the central value. There is dispersion, there is variability, and these are the measures.



Measures of Variability

Range:

- The range is the difference between the largest and the smallest value in the sample.
- The range is the easiest of all measures of dispersion to calculate.

R = Maximum Value - Minimum Value

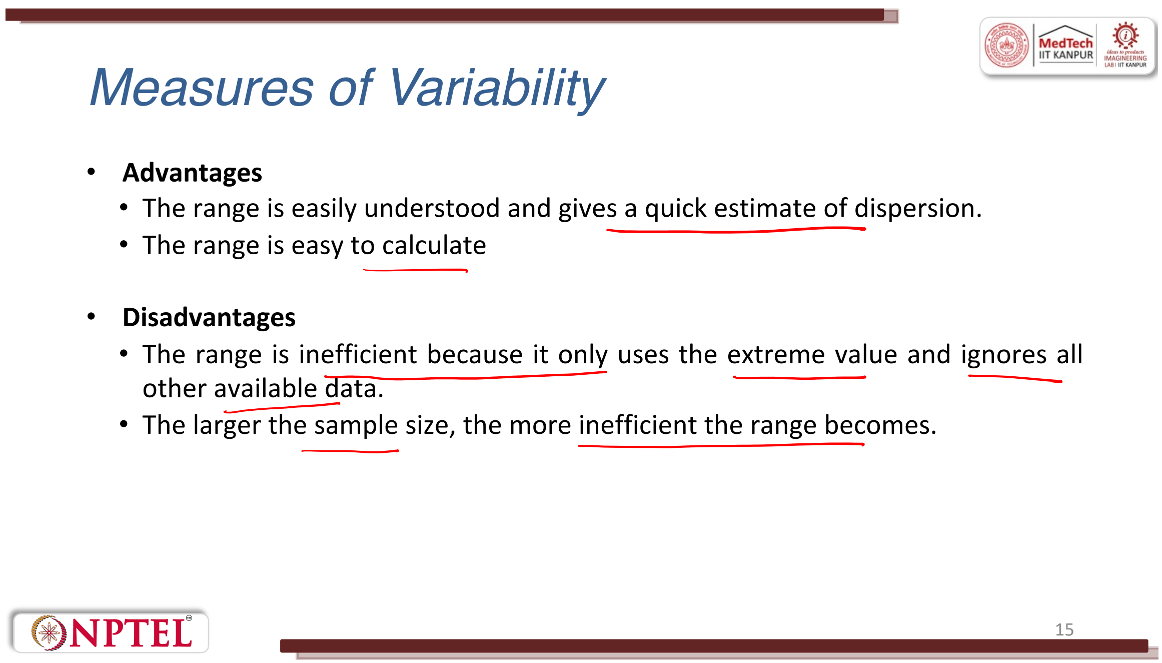
SQC

$\bar{X}$ and R charts

Average (Mean) Range

Range: The range is the difference between the largest and the smallest value in the sample. The range is the easiest way to measure dispersion. To calculate the range, we just need to subtract the minimum value from the maximum value. In SQC, which stands for Statistical Quality Control in manufacturing, when we try to determine whether our output, the manufacturing process, the performance of the boiler, or the performance of the running turbine is within control or not, we plot a chart. The chart is sometimes known, for variables, as the X-bar and R chart. Here, X-bar represents the average or mean, and R represents the range.

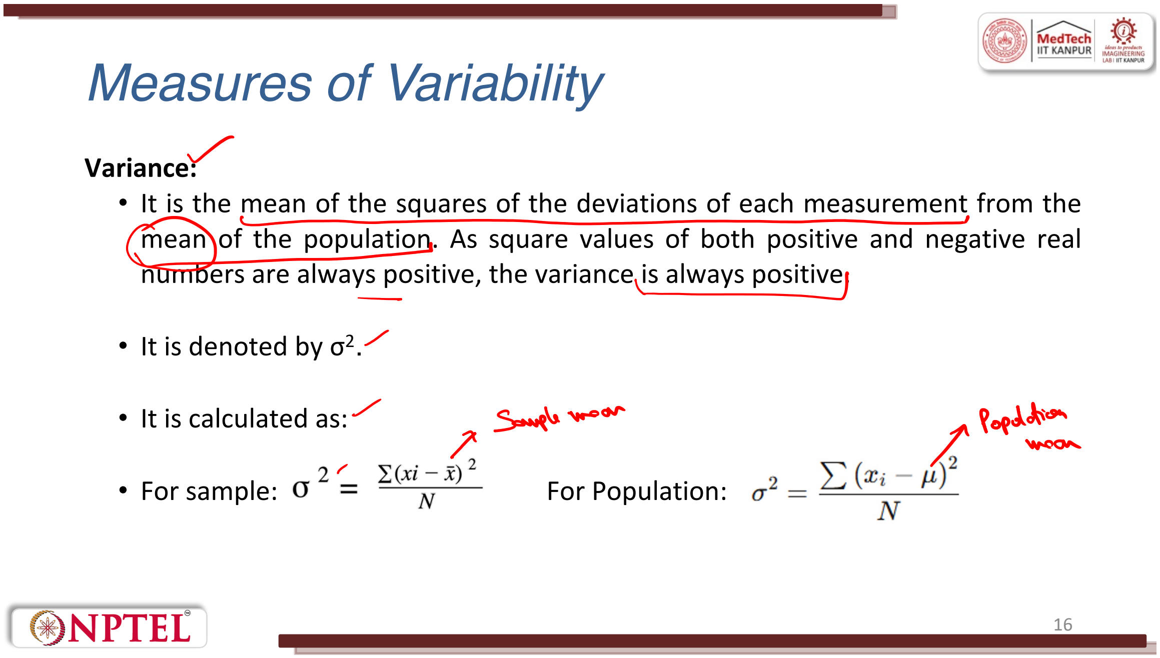
These are very important charts. What is the minimum value? What is the largest value? What is the total difference? What is the overall spread of the data? So, this is one of the measures used for variability.



Measures of Variability

- **Advantages**
  - The range is easily understood and gives a quick estimate of dispersion.
  - The range is easy to calculate
- **Disadvantages**
  - The range is inefficient because it only uses the extreme value and ignores all other available data.
  - The larger the sample size, the more inefficient the range becomes.

The advantages of the range are: it is easily understood and provides a quick estimate of dispersion. The range is easy to calculate, as it is simply the difference between the largest and smallest values. The disadvantages are that the range is inefficient because it only uses extreme values and ignores all other available data. The larger the sample size, the more inefficient the range becomes.



Measures of Variability

Variance:

- It is the mean of the squares of the deviations of each measurement from the mean of the population. As square values of both positive and negative real numbers are always positive, the variance is always positive
- It is denoted by σ^2.
- It is calculated as:

Sample mean

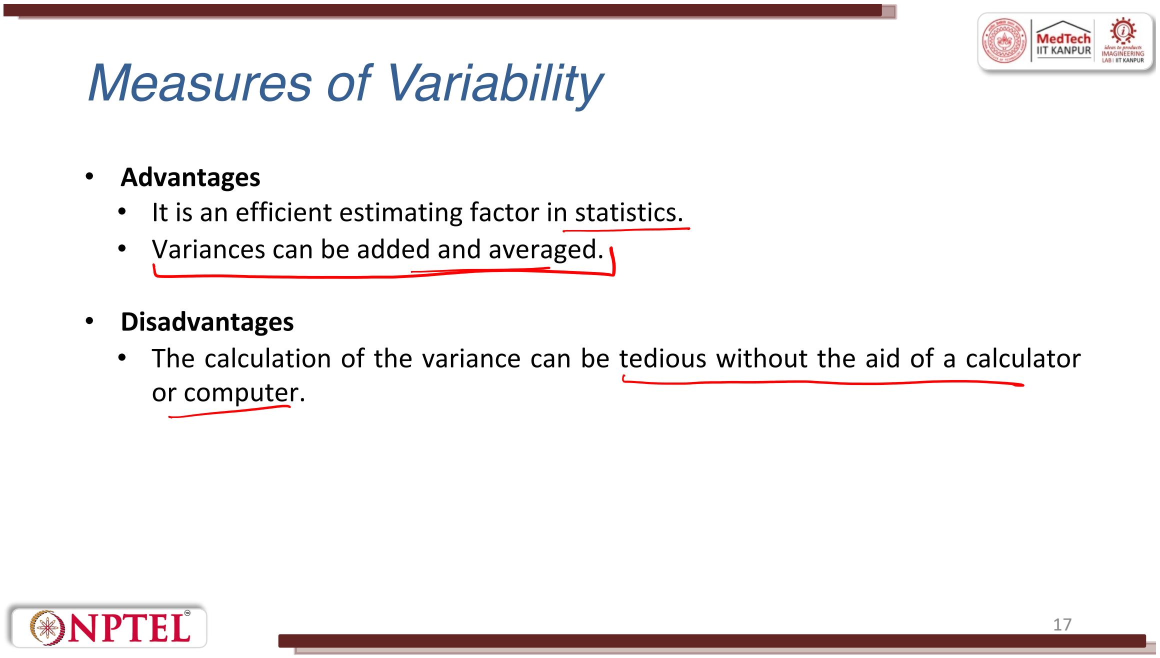
Population mean

- For sample: $\sigma^2 = \frac{\Sigma(xi - \bar{x})^2}{N}$ For Population: $\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$

Then there is a close variability-determining parameter that is calculated, which is variance. It is the mean of the squares of the deviations of each measurement from the mean of the population. So there could be variance with respect to mean, variance with respect to median. But generally, when we talk about variance, it is always about the variance with respect to the mean. It is given here.

It is from the mean of the population; the mean of the squares of the deviation. The square values of both positive and negative real numbers are always positive. The variance is always positive because it is squared.

For sample: $\sigma^2 = \frac{\Sigma(xi - \bar{x})^2}{N}$ For Population: $\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$

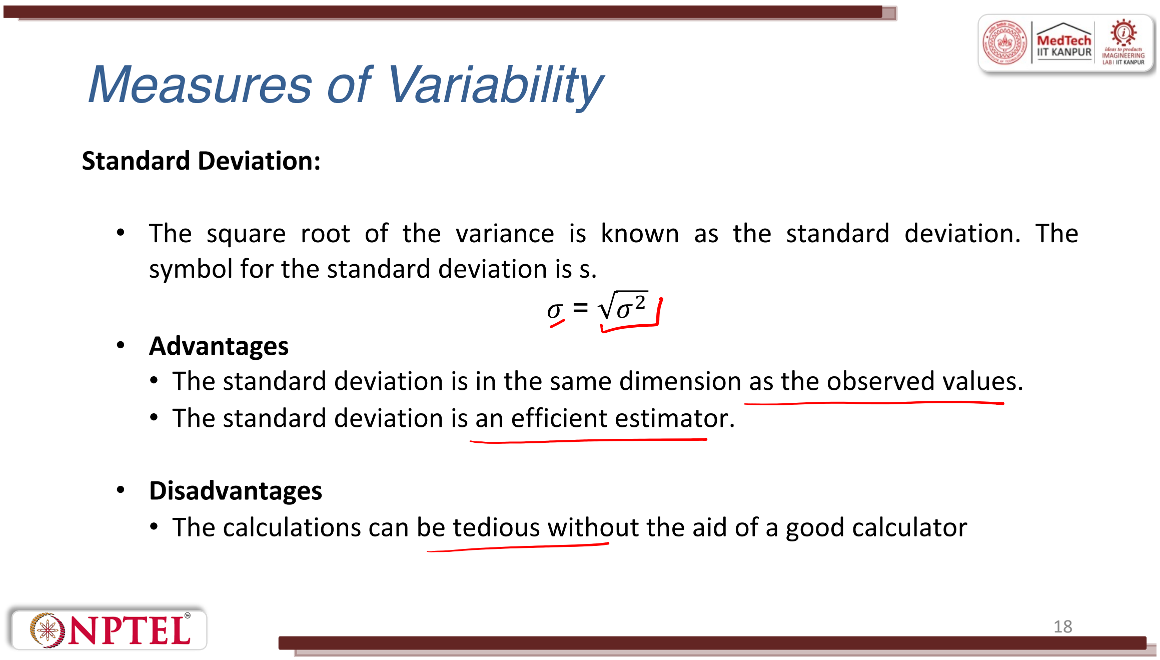


Measures of Variability

- **Advantages**
  - It is an efficient estimating factor in statistics.
  - Variances can be added and averaged.
- **Disadvantages**
  - The calculation of the variance can be tedious without the aid of a calculator or computer.

Advantages: it is an efficient estimating factor in statistics. Variances can be added and averaged. That is why variance is always the most used factor for determining the spread, because variances have additive properties. Standard deviation is the square root of the variance.

Though the overall degree comes equivalent to the mean, we squared the values, we took the square root, the degree becomes equivalent, but that cannot be added or subtracted. So, these variances can be added and averaged; that is, they have an additive property, which is true, and that is why these are used the most. Disadvantages: the calculation of the variances can be tedious without the aid of a calculator or computer.



Measures of Variability

Standard Deviation:

- The square root of the variance is known as the standard deviation. The symbol for the standard deviation is s.

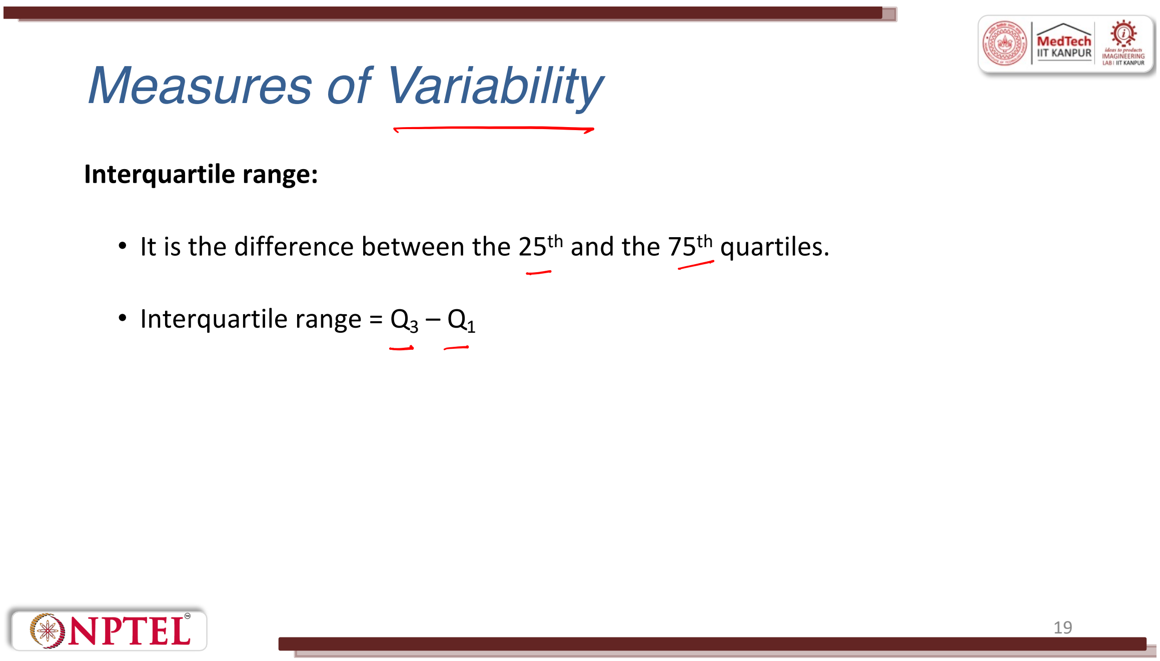
$$\sigma = \sqrt{\sigma^2}$$

- **Advantages**
  - The standard deviation is in the same dimension as the observed values.
  - The standard deviation is an efficient estimator.
- **Disadvantages**
  - The calculations can be tedious without the aid of a good calculator

18

But nowadays, you can use an Excel sheet where you simply have the data relation about the variance. You can try to calculate it very directly. Standard deviation is the square root of the variance.

It has advantages, such as it is $\sigma = \sqrt{\sigma^2}$. Advantages are that it is in the same dimension as the observed value. As I said, the dimension stays the same. We took the square. We took the square root. Standard deviation is an efficient estimator. Disadvantages: the calculation can be tedious without the aid of a good calculator.



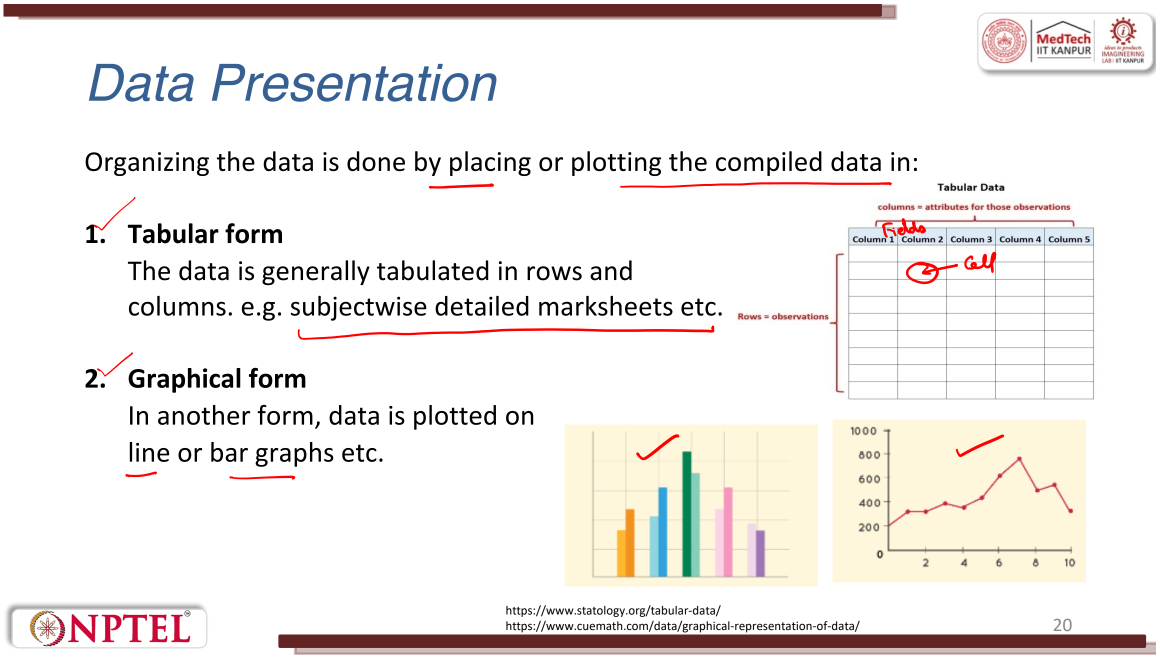
Measures of Variability

Interquartile range:

- It is the difference between the 25th and the 75th quartiles.
- Interquartile range = $Q_3 – Q_1$

Interquartile range, like I talked about the range, is the largest and the smallest value. Interquartile range is the first quartile, the 25th value, and the 75th value when we talk about percentages. So, the 25th percentile and 75th percentile, when we talk about quartiles, Q3 - Q1 is the interquartile range. This is also one of the measures of variability. Now, this is about how we describe the data. These measures are not only used in descriptive statistics. Whatever is used in descriptive statistics is also used in inferential statistics as well.

But in inferential statistics, we add more things. We have hypothesis testing. We have analysis of variance. All these variables and parameters can be used to describe the data. These could also be used to infer something. Now, descriptive statistics and data calculations we talked about. Now, let me talk about how you present the data.



Data Presentation

Organizing the data is done by placing or plotting the compiled data in:

1. **Tabular form**
   The data is generally tabulated in rows and columns. e.g. subjectwise detailed marksheets etc.

Tabular Data

columns = attributes for those observations

Fields

Rows = observations

Column 1	Column 2	Column 3	Column 4	Column 5

Cell

2. **Graphical form**
   In another form, data is plotted on line or bar graphs etc.

https://www.statology.org/tabular-data/
https://www.cuemath.com/data/graphical-representation-of-data/

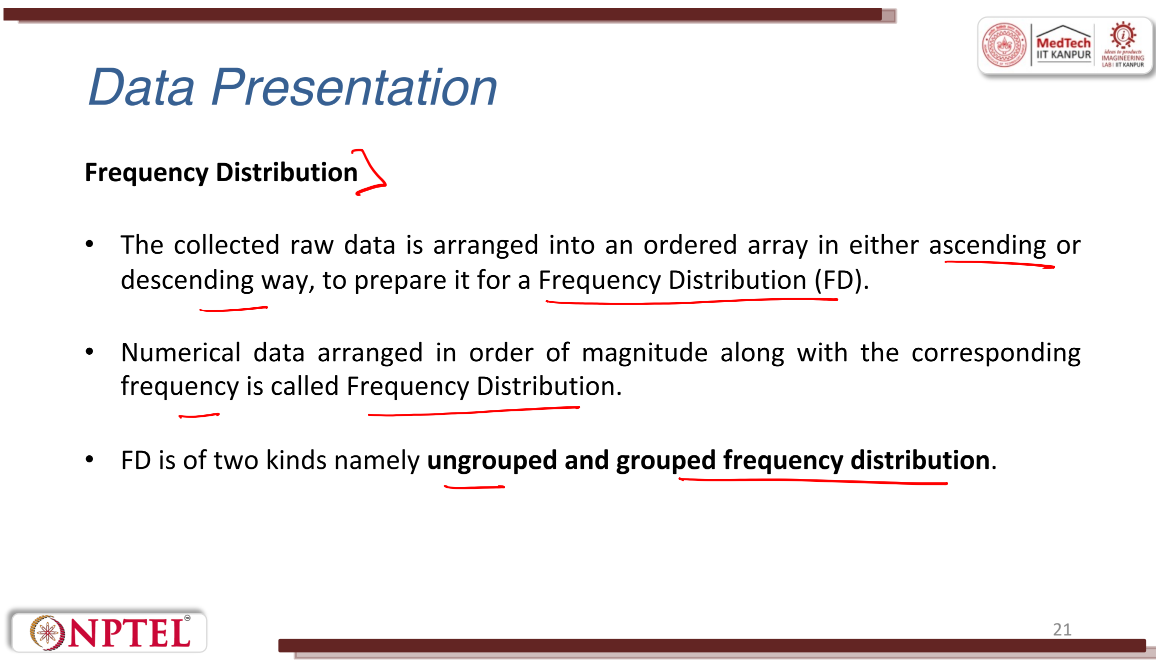
To present the data, there could be two forms. That is, organizing the data is done by placing or plotting the compiled data in a tabular form or in a graphical form. The data is generally tabulated in rows and columns.

That is a subject-wise detailed mark sheet. You see, there are columns. We also call them fields. And we have rows where we have observations, and each of them is known as a cell. Graphical form.

In other forms, data is plotted in the form of line or bar graphs. You see, these are the bar graphs. These are line diagrams. So, this way, data is also presented. It is said an illustration or a figure is worth a thousand words.

You tell a complete story in a thousand words. For example, you need to see the weather forecast. You need to see what the scenario of the present manufacturing system is. You need to see what the present behavior or health of the running turbine is. You can either write a full paragraph on that or present it in an illustration—a figure showing this is the growth that is happening.

This was the last year's thing. And this is a bar graph that is displayed. This is a line diagram showing the trends. So that is a graphical form, which is the most important thing. How to present the data?

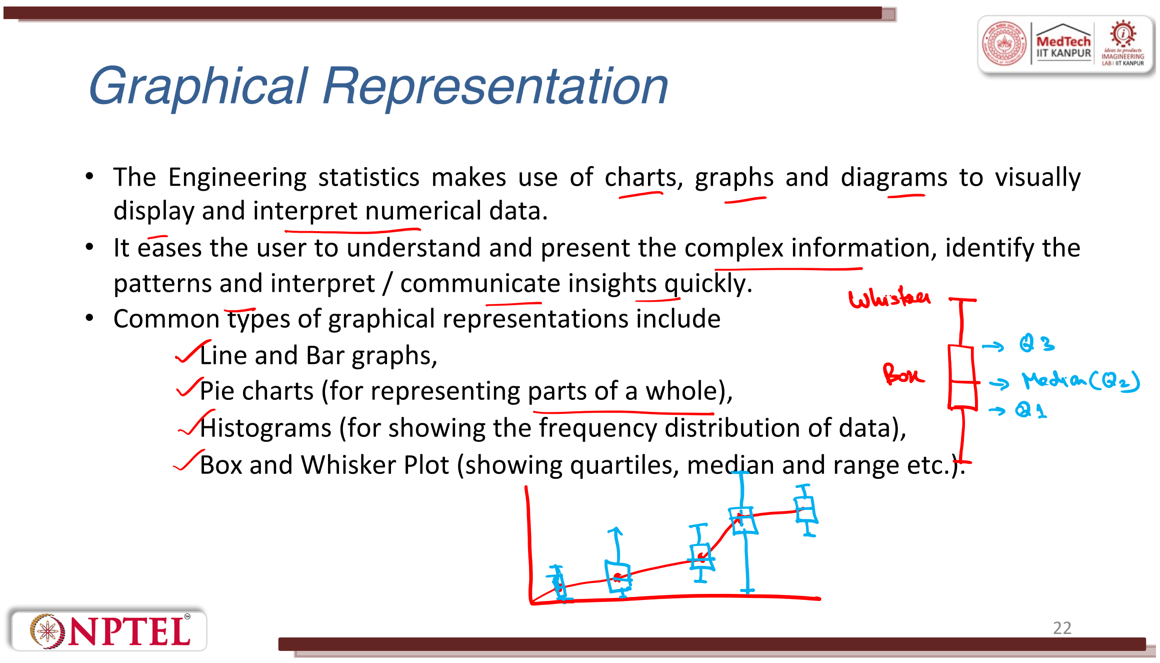


Data Presentation

Frequency Distribution

- The collected raw data is arranged into an ordered array in either ascending or descending way, to prepare it for a Frequency Distribution (FD).
- Numerical data arranged in order of magnitude along with the corresponding frequency is called Frequency Distribution.
- FD is of two kinds namely **ungrouped and grouped frequency distribution**.

Let me first talk about the tabular form. In the tabular form, we have one term known as frequency distribution. The collected raw data is arranged into an ordered array in either ascending or descending order to prepare it for frequency distribution. Numerical data arranged in order of magnitude along with the corresponding frequency is called frequency distribution. Frequency distribution could be ungrouped or grouped. I will talk about this more when I discuss normal probability distribution in the next lecture.



Graphical Representation

- The Engineering statistics makes use of charts, graphs and diagrams to visually display and interpret numerical data.
- It eases the user to understand and present the complex information, identify the patterns and interpret / communicate insights quickly.
- Common types of graphical representations include
  - ✓Line and Bar graphs,
  - ✓Pie charts (for representing parts of a whole),
  - ✓Histograms (for showing the frequency distribution of data),
  - ✓Box and Whisker Plot (showing quartiles, median and range etc.).

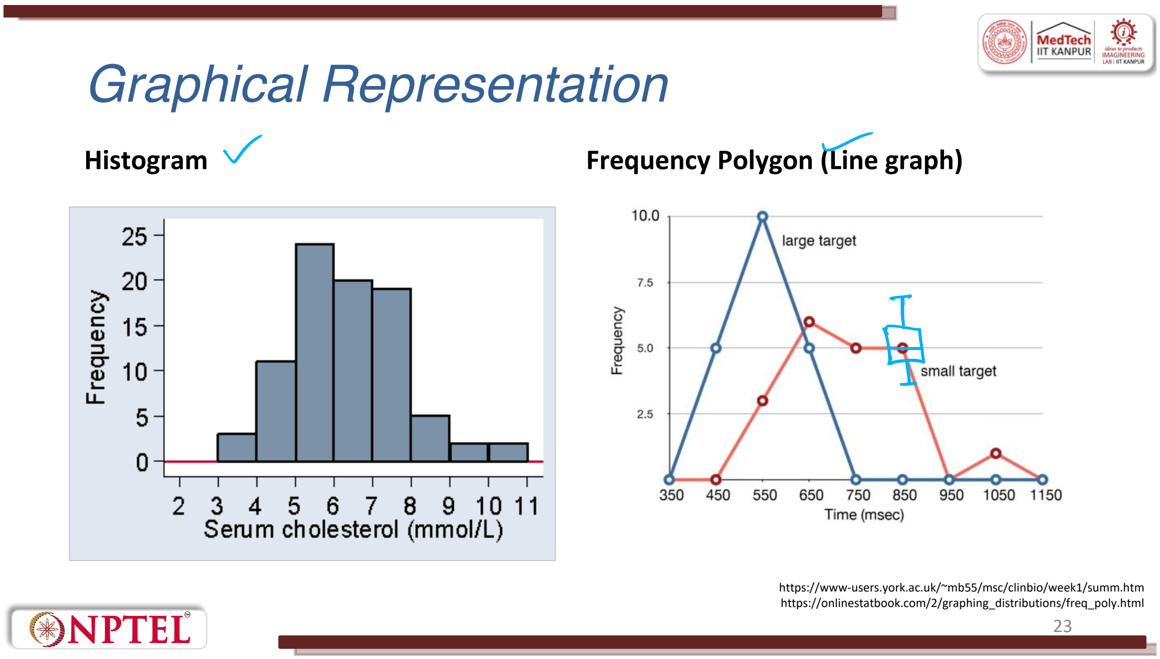
Let me first talk about the graphical representation of the data. In engineering statistics, we create charts, graphs, and diagrams to visually display and interpret numerical data. It helps the user understand and present complex information, identify patterns, and interpret or communicate insights quickly. The common types of graphical representations include line or bar graphs and pie charts, which represent parts of a whole.

If the whole 100% of the population is to be represented, we say this percentage is correct, this percentage is the other way, then we use pie charts. Histograms are used for showing the frequency distribution of the data. Box and whisker plots are used for showing quartiles, median, and range; a box and whisker plot looks something like this. This is known as the box. These are the whiskers.

Technically, this is the median or quartile 2. This is quartile 1, and this is quartile 3. This is a box and whisker plot. It is generally used when we have a line graph. For example, there is a ζ plotted here.

We have a line graph here, and this is plotted like this. Each point is an average of some of the observations. Then we can have a box and whisker plot everywhere. For example, here we can have this as the mid value. This is how the whisker is there.

This is the mid value. This is how the whisker is positioned. It depends on the kind of thing. The whisker could be very large at some points. At some points, it could be very small. This is a box-and-whisker plot to show the range or the error in each observation.

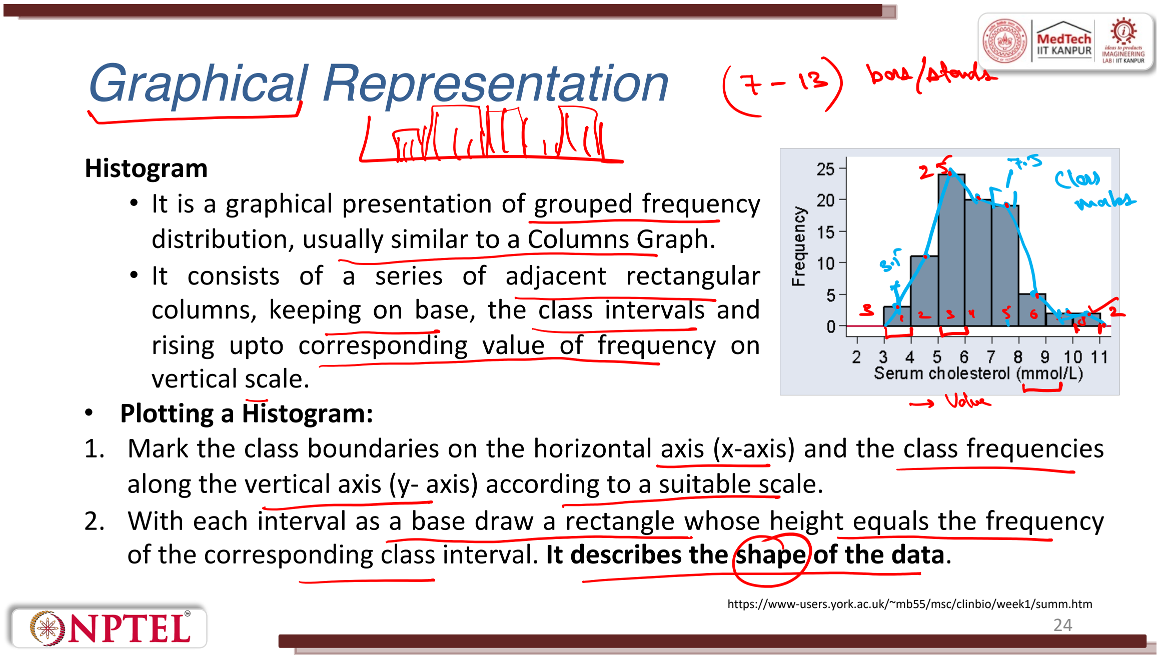


Graphical Representation

Histogram ✓ **Frequency Polygon (Line graph)**

https://www-users.york.ac.uk/~mb55/msc/clinbio/week1/summ.htm
https://onlinestatbook.com/2/graphing_distributions/freq_poly.html

Let me talk about histograms or frequency polygons and their uses. I drew the box-and-whisker plot here in the frequency polygon itself.



Graphical Representation

(7 - 13) bars/stands

Histogram

- It is a graphical presentation of grouped frequency distribution, usually similar to a Columns Graph.
- It consists of a series of adjacent rectangular columns, keeping on base, the class intervals and rising upto corresponding value of frequency on vertical scale.

- **Plotting a Histogram:**

1. Mark the class boundaries on the horizontal axis (x-axis) and the class frequencies along the vertical axis (y- axis) according to a suitable scale.
2. With each interval as a base draw a rectangle whose height equals the frequency of the corresponding class interval. **It describes the shape of the data**.

https://www-users.york.ac.uk/~mb55/msc/clinbio/week1/summ.htm

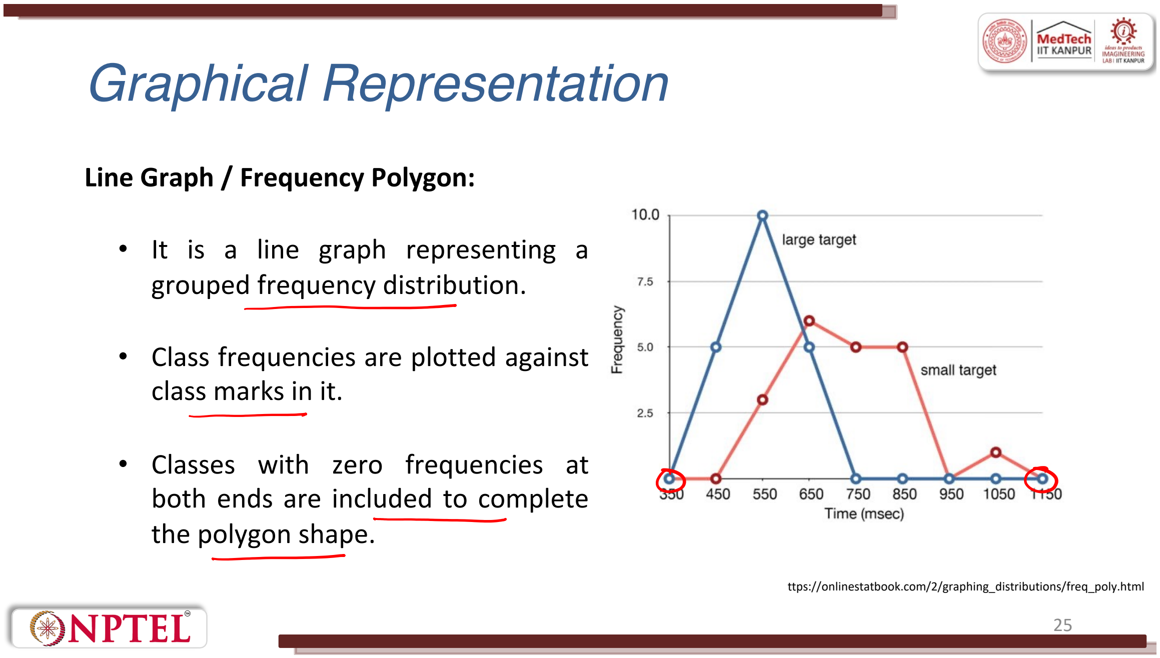
Histogram. It is a graphical presentation of grouped frequency distribution, usually similar to a column graph. It consists of a series of adjacent rectangular columns based on class intervals, rising to the corresponding frequency value on the vertical scale. For example, this is serum cholesterol in millimoles per liter, and this is the frequency of occurrence. This is the largest value, between 10 to 11 millimoles per liter. This is the smallest class interval, between 2 to 4 millimoles per liter. The maximum number of them are between 5 to 6 millimoles per liter.

That is around 25 observations here. And, for instance, only 3 observations are here. Only 2 observations are here. This is frequency. This is the value.

Plotting a histogram: mark the class boundaries on the horizontal axis and the class frequencies on the vertical axis according to a suitable scale. With each interval as a base, draw a rectangle whose height equals the frequency of the corresponding class interval. It describes the shape of the data for a good histogram that should be visible. What do you think? Should the class intervals be selected so that we get 100 lines? Should it be selected so that we get 500 lines?

Or should we select so that we have only 4 lines? There is a range or number of stands which should be there in a histogram—vertical stands—that is, in general, between 7 to 13 bars, or I call them stands, so that it is visually appealing. Here we have 1, 2, 3, 4, 5, 6, 7, 8. It is visible. If I suppose we get 100 bars like this.

This doesn't make much sense. It is better to combine them and get the visual data, with a maximum of up to 13 bars, so that it is visually appealing. The purpose is graphical representation, which should visually convey the shape or behavior of the data.



Graphical Representation

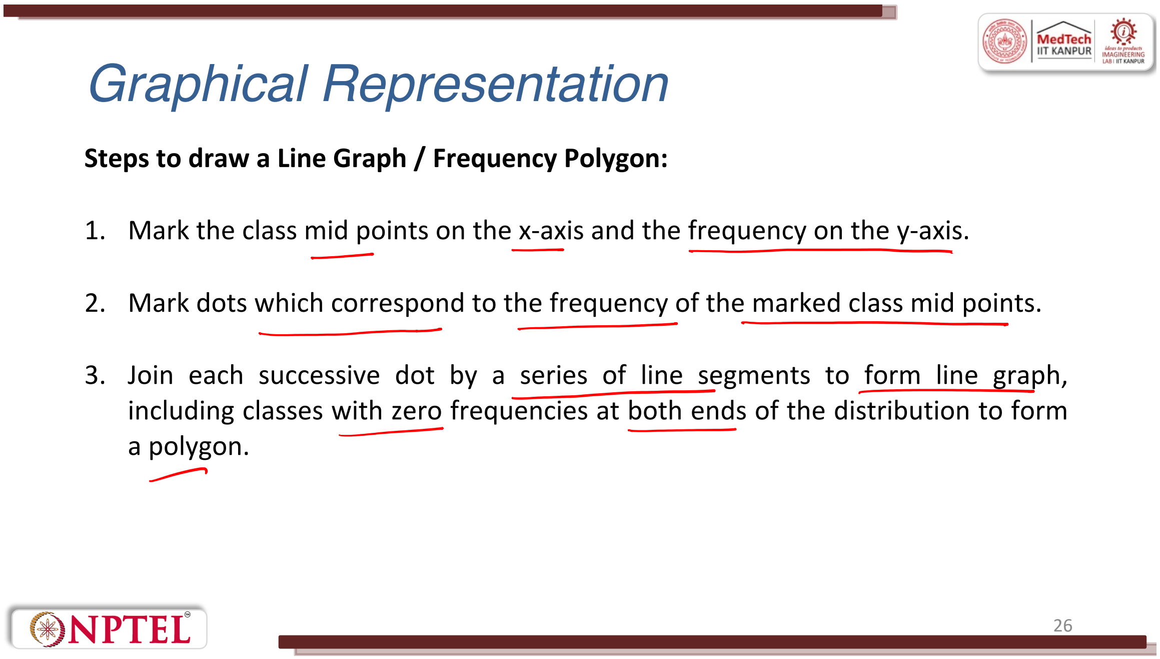
Line Graph / Frequency Polygon:

- It is a line graph representing a grouped frequency distribution.
- Class frequencies are plotted against class marks in it.
- Classes with zero frequencies at both ends are included to complete the polygon shape.

ttps://onlinestatbook.com/2/graphing_distributions/freq_poly.html

Line graph or frequency polygon. It is a line graph representing a grouped frequency distribution here. Class frequencies are plotted against class marks in it. Class marks, if you say, here if the range or interval is between 3 to 4, the class mark would be 3.5. So, here the class mark between 7 and 8 would be 7.5. These are class marks. So, these class marks should give you one point now. One point here, one point here.

At each position, you get one point. So, if you join these points—these class marks—this converts my histogram into my frequency polygon. Classes with zero frequencies at both ends are included to complete the polygon shape. That is, it touches here at the bottom.



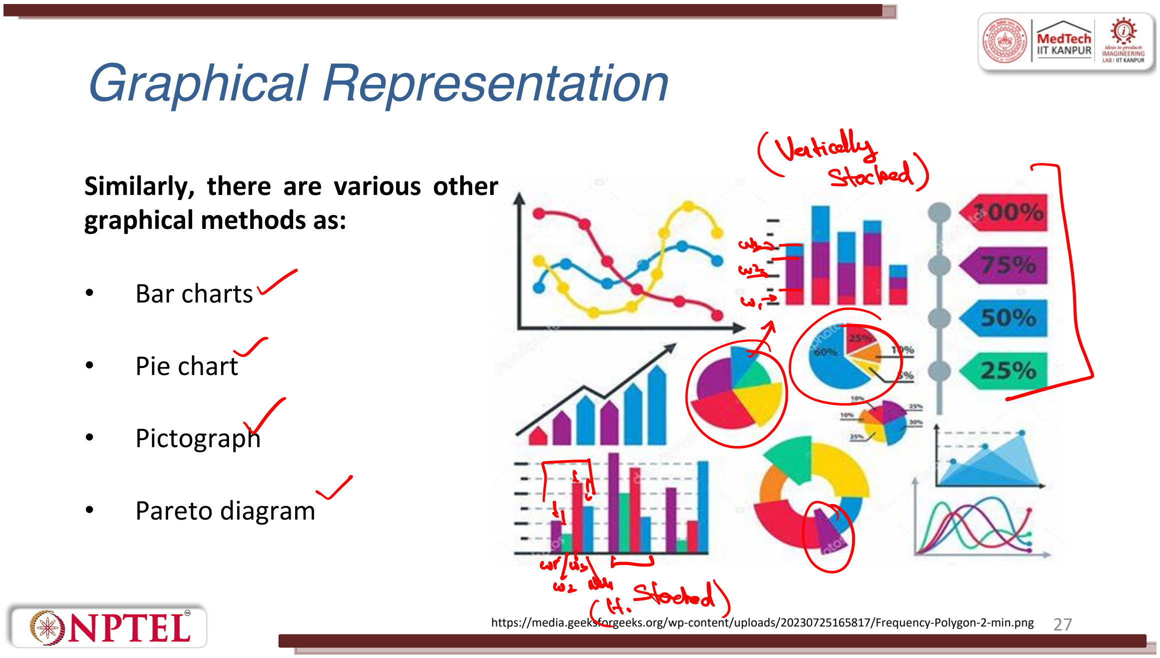
Graphical Representation

Steps to draw a Line Graph / Frequency Polygon:

1. Mark the class mid points on the x-axis and the frequency on the y-axis.
2. Mark dots which correspond to the frequency of the marked class mid points.
3. Join each successive dot by a series of line segments to form line graph, including classes with zero frequencies at both ends of the distribution to form a polygon.

Steps to draw a line graph for a frequency polygon. Mark the class midpoints on the axis and the frequency on the y-axis, as I just did for the histogram plot. Mark the dots that correspond to the frequency and the marked class midpoints. Join each successive dot by a series of line segments to form a line graph, including classes with zero frequencies at both ends of the distribution, to form a polygon. This is what I did here in blue color. I marked the average values of the class interval for these points.

Then, I joined these points. This is how the frequency polygon is constructed from the histogram.



Graphical Representation

Similarly, there are various other graphical methods as:

- Bar charts ✓
- Pie chart ✓
- Pictograph ✓
- Pareto diagram ✓

https://media.geeksforgeeks.org/wp-content/uploads/20230725165817/Frequency-Polygon-2-min.png

Other than the frequency polygon and histogram, there are various graphical methods. For example, these are the bar charts. These are the four things. It could be, for example, in a month: week 1, this green is week 2, then red is week 3, then blue is week 4. The total production in these weeks. These are the bars. This could also be a stacked bar. For example, week 1 is here, week 2 is here, week 3 is here.

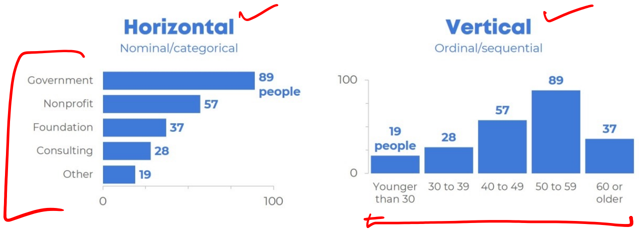
Total production of the month: week 1, week 2, week 3, total added over each other. It is vertically stacked. These are bar charts here.

Then comes the pie chart. I talked about pie charts when we have 100% of the data. Out of the 100% of the data, if we wish to plot it, pie charts could take many forms. You can see a three-dimensional pie chart. You can see a pie chart with some pie protruding out to highlight a specific output or point of interest. Then we have a pictograph. In pictorial form, we can show the 25th percentile, 50th percentile, 75th percentile, or 100th percentile. Where does the data lie? Then we have a Pareto diagram, which I will discuss in the coming slides.

Graphical Representation

Bar chart

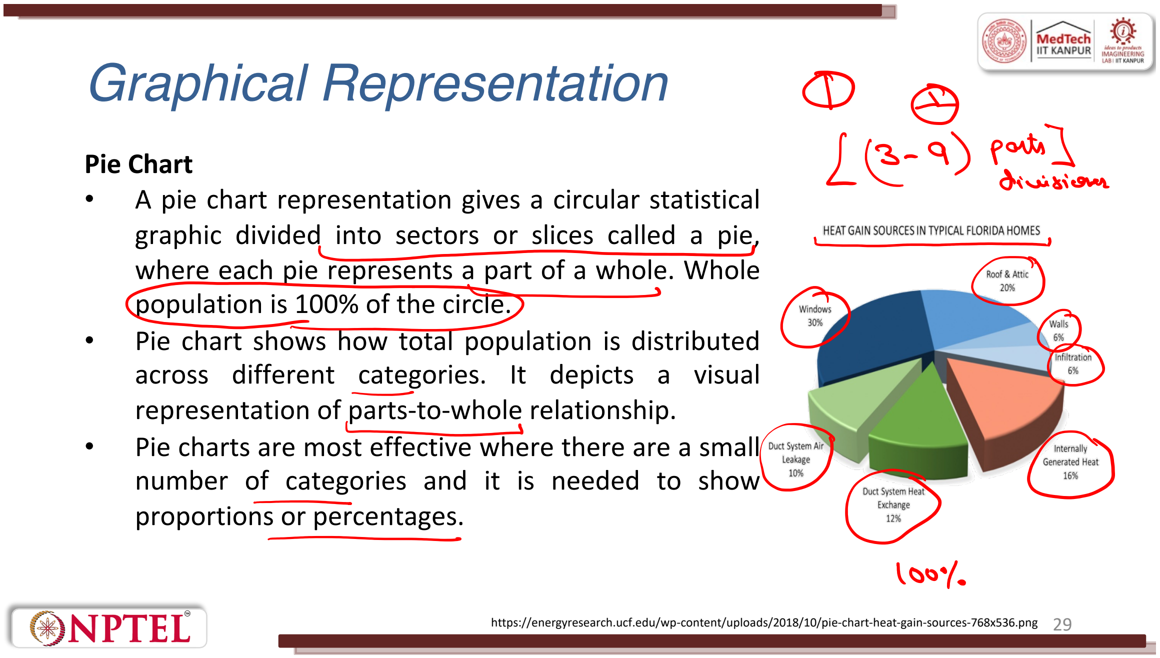
- Similar to a Histogram, a bar chart is a graphical tool that uses rectangular bars to represent and compare data, height of each bar being proportional to the value it displays.
- Bar charts can be oriented vertically or horizontally and are ideal for visualizing trends, comparisons, and distributions across different groups or over time.



https://depictdatastudio.com/when-to-use-horizontal-bar-charts-vs-vertical-column-charts/

Bar charts. Similar to a histogram, a bar chart is a graphical tool that uses rectangular bars to represent and compare data, with the height of each bar proportional to the value it displays. Bar charts can be oriented vertically or horizontally and are ideal for visualizing trends, comparisons, and distributions across different groups over time.

You see, this is a horizontal bar chart. This is a vertical bar chart, which I showed you in the previous slide. These were horizontally stacked. And these were vertically stacked bars. So you see government, non-profit, foundation, consulting, other. The kind of people being talked about, or we talked about younger between different ages, this is the way a vertical stacked bar is there. Between ages, less than 30 to more than 60, this many number of people are there. Then comes the pie chart, pie chart.



Graphical Representation

Pie Chart

- A pie chart representation gives a circular statistical graphic divided into sectors or slices called a pie, where each pie represents a part of a whole. Whole population is 100% of the circle.
- Pie chart shows how total population is distributed across different categories. It depicts a visual representation of parts-to-whole relationship.
- Pie charts are most effective where there are a small number of categories and it is needed to show proportions or percentages.

[(3-9) parts] divisions

HEAT GAIN SOURCES IN TYPICAL FLORIDA HOMES
Windows 30%
Roof & Attic 20%
Walls 6%
Infiltration 6%
Internally Generated Heat 16%
Duct System Heat Exchange 12%
Duct System Air Leakage 10%
100%

https://energyresearch.ucf.edu/wp-content/uploads/2018/10/pie-chart-heat-gain-sources-768x536.png

Representation gives a circular statistical graphic divided into sectors or slices called a pie. Where each pie represents a part of a whole, the whole population is 100% of the circle. A pie chart shows how the total population is distributed across different categories. It depicts a visual representation of parts-to-whole relationships. You can see heat gain sources in typical Florida homes.

Where is the heat gain happening? 30% through windows, 20% through roof and attic homes. Through walls, it is only 6%. Infiltration is 6%. Internally generated heat is 16%.

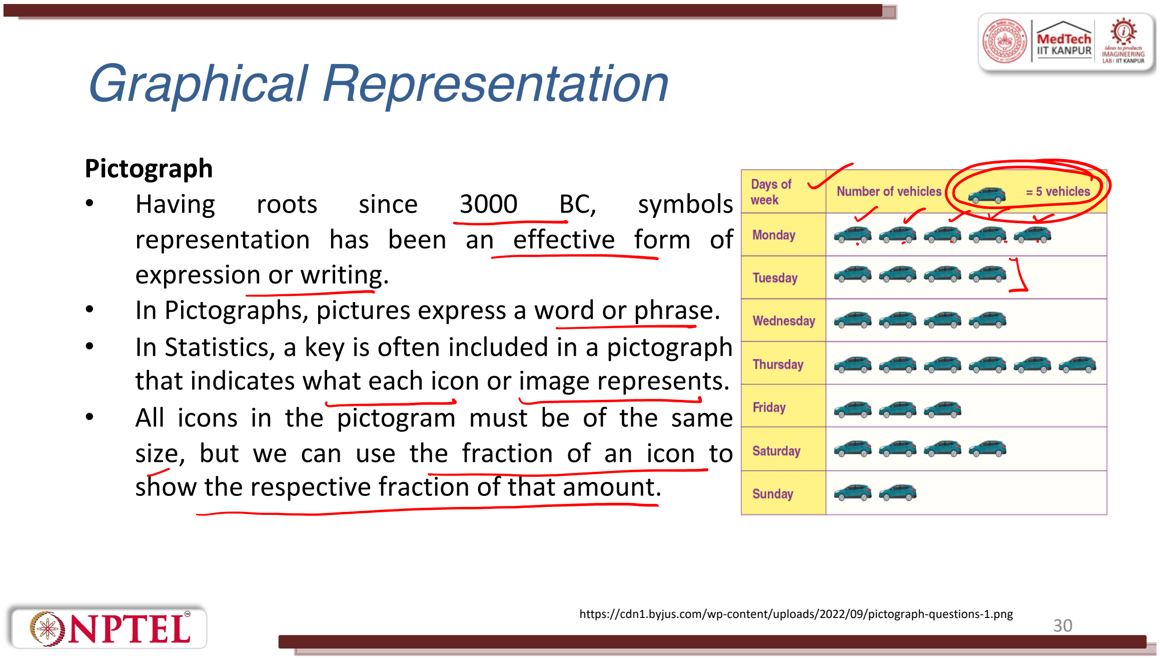
Duct system heat exchange is 12%. Duct system leakage, that is, air leakage is 10%. In total, 100% of the heat gained in homes is divided into these different sources. Pie charts are most effective when there are a small number of categories and it is necessary to show proportions or percentages. When I say a small number of categories, it is not always advisable to use a pie chart when we have 100 distributions.

For example, having 100 distributions for each percentile is not ideal. Generally, a pie chart is used when we have between 3 to a maximum of 9 divisions, if I may say. For example, 2 divisions would only form a semicircle. If there are 3, there could be differences between the angles. So, that is between 3 to 9 parts or divisions.

Then a pie chart is used. Pictograph. A pictograph is not something plotted through statistics. It is just a display or a board. For example, in warehouses.

In warehouses, whenever one specific carton is completely filled, they just say they put the key there. There is a key number for that specific carton. Key 1, 2, 3, 4. That is hung up there. This one is complete.

And when the supervisor comes, they count the number of keys. It could be anything, like a number of tokens. That is a kind of pictorial representation of the overall status. A total of five cartons of this size, maybe very big room-sized cartons, are fixed to be shifted or transported out of the factory. This is a pictograph.



Graphical Representation

Pictograph

- Having roots since 3000 BC, symbols representation has been an effective form of expression or writing.
- In Pictographs, pictures express a word or phrase.
- In Statistics, a key is often included in a pictograph that indicates what each icon or image represents.
- All icons in the pictogram must be of the same size, but we can use the fraction of an icon to show the respective fraction of that amount.

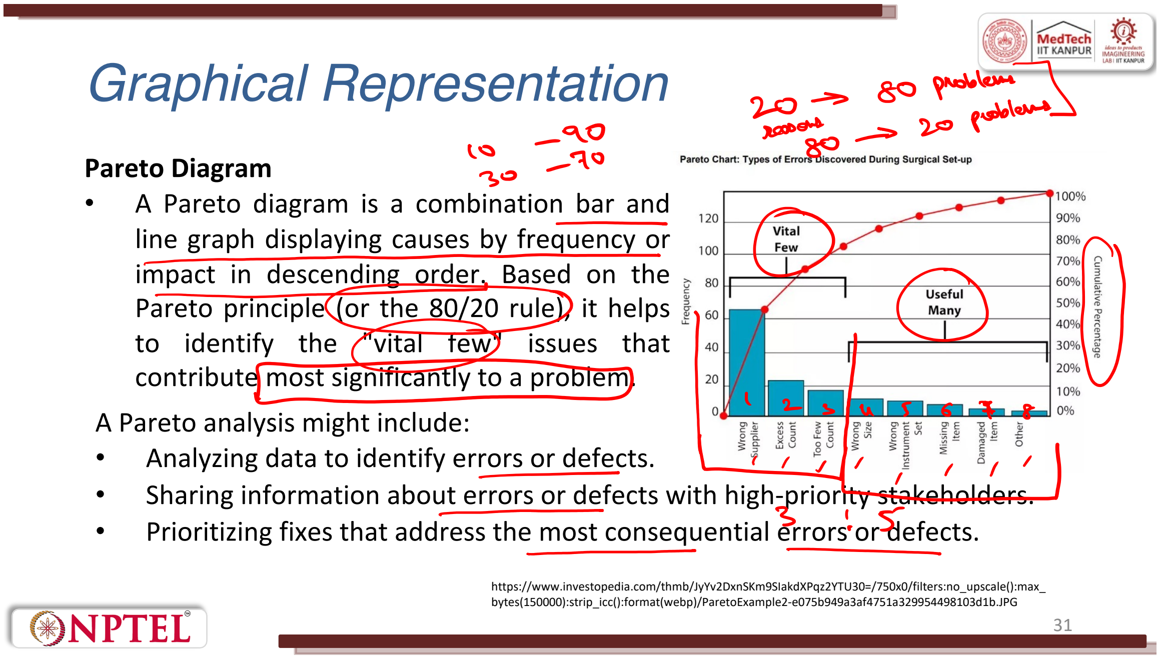
Days of week	Number of vehicles 🚗 = 5 vehicles
Monday	🚗 🚗 🚗 🚗 🚗
Tuesday	🚗 🚗 🚗 🚗
Wednesday	🚗 🚗 🚗 🚗
Thursday	🚗 🚗 🚗 🚗 🚗 🚗
Friday	🚗 🚗 🚗
Saturday	🚗 🚗 🚗 🚗
Sunday	🚗 🚗

https://cdn1.byjus.com/wp-content/uploads/2022/09/pictograph-questions-1.png

So, having roots since 3000 BC, symbolic representation has been an effective form of expression in writing. For example, the number of vehicles or days of the week. This is, let me say, a service station for Hyundai cars in the city of Kanpur. In that city, each day when cars come, the number of cars being serviced is marked with a sticker. One car is serviced, then a second, third, fourth, and fifth.

Here they say, one car equals five vehicles. When they put 1, that means 5 cars are serviced, it means. That means 1, 2, 3, 4, 5. 5 into 5, 25 cars are serviced on Monday, 20 cars are serviced on Tuesday. This is a pictorial representation so that the supervisor or the top management could see how and what the progress of the work is. In pictographs, pictures express a word or phrase.

In statistics, a key is often included in a pictograph that indicates what each icon or image represents. This is known as a key. One picture is equal to five vehicles. All icons in the pictogram must be of the same size, but we can use a fraction of an icon to show the respective fraction of that amount. That could also be possible. Another graphical representation that I would like to talk about is the Pareto diagram.



Graphical Representation

Pareto Diagram

- A Pareto diagram is a combination bar and line graph displaying causes by frequency or impact in descending order. Based on the Pareto principle (or the 80/20 rule), it helps to identify the "vital few" issues that contribute most significantly to a problem.

A Pareto analysis might include:

- Analyzing data to identify errors or defects.
- Sharing information about errors or defects with high-priority stakeholders.
- Prioritizing fixes that address the most consequential errors or defects.

Pareto Chart: Types of Errors Discovered During Surgical Set-up

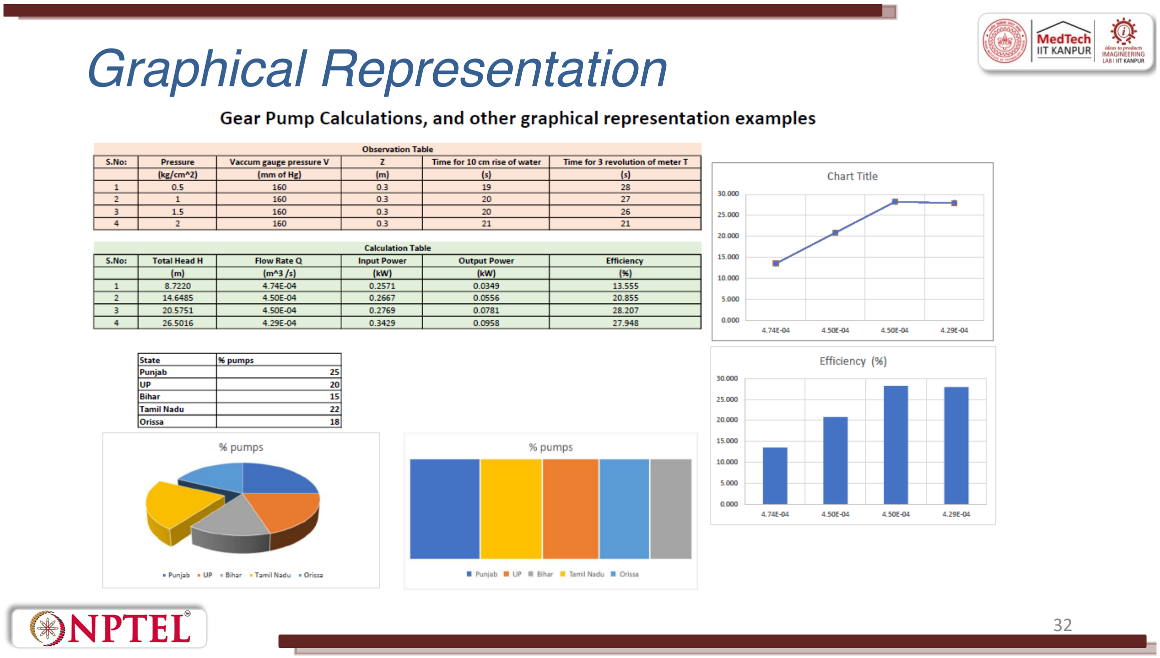
https://www.investopedia.com/thmb/JyYv2DxnSKm9SIakdXPqz2YTU30=/750x0/filters:no_upscale():max_bytes(150000):strip_icc():format(webp)/ParetoExample2-e075b949a3af4751a329954498103d1b.JPG

A Pareto diagram is something where we take or put all the observations in ascending order and try to see the cumulative frequencies. You see, a Pareto diagram is a combination of a bar and line graph displaying causes by frequency or impact in descending order, based upon the Pareto principle, that is, the 80-20 rule. It helps to identify the vital few issues that contribute more significantly to a problem. For instance, this is a plot that shows the number of errors which are there during a surgical setup.

Wrong supplier, excess count, too few counts, wrong size, wrong instrument set, missing item, damaged item, and others. These are plotted in a cumulative percentage way. Now you see there are a few issues out of 1, 2, 3, 4, 5, 6, 7, 8. Out of 8 issues, 3 issues are contributing very highly to the wrong surgical setup. And there are 5 issues which are contributing very little.

So, these are the vital few which are identified and the useful many which are also identified. That is, 20 percent of the items contribute to 80 percent of the problems—20 percent of, I would say, reasons. On the other hand, 80 percent of the reasons contribute to 20 percent of the problems. This is the Pareto rule. It is not exactly 80-20.

Here in this case, it is 3 and 5. It could be 10-90. It could be 30-70. But the Pareto rule says this is the way we try to identify what the vital few issues are that contribute most significantly to a problem. The parallel analysis might include analyzing data to identify errors or defects, sharing information about errors or defects with high-priority stakeholders, and prioritizing fixes that address the most consequential errors or defects. We had talked about so many data presentation methods and graphs.



Graphical Representation

Gear Pump Calculations, and other graphical representation examples

Observation Table

S.No:	Pressure	Vaccum gauge pressure V	Z	Time for 10 cm rise of water	Time for 3 revolution of meter T
	(kg/cm^2)	(mm of Hg)	(m)	(s)	(s)
1	0.5	160	0.3	19	28
2	1	160	0.3	20	27
3	1.5	160	0.3	20	26
4	2	160	0.3	21	21

Calculation Table

S.No:	Total Head H	Flow Rate Q	Input Power	Output Power	Efficiency
	(m)	(m^3 /s)	(kW)	(kW)	(%)
1	8.7220	4.74E-04	0.2571	0.0349	13.555
2	14.6485	4.50E-04	0.2667	0.0556	20.855
3	20.5751	4.50E-04	0.2769	0.0781	28.207
4	26.5016	4.29E-04	0.3429	0.0958	27.948

State	% pumps
Punjab	25
UP	20
Bihar	15
Tamil Nadu	22
Orissa	18

32

Let me move to an Excel sheet itself. I have taken here the data that we plotted or that we saw using the virtual demonstration of the GAR pump. So here, efficiency is there, and I told you to plot the efficiency against the flow rate. Let me try to use the kinds of plots which are there.

Let me see. For example, I have the Insert tab here. Here, we can plot using the recommended charts. Or if I need to plot between two columns, I can just select the columns. I'm selecting the columns of flow rate and efficiency, and I will pick a scatter plot here.

You see, this is a scatter plot that you can see plotted here. On this scatter plot, we have efficiency on the y-axis and flow rate on the x-axis. This is plotted here. I can also convert this into a frequency polygon. A frequency polygon, just like this line, or if you wish to have dots between the specific readings, I can use another kind of set.

This is a frequency polygon. Let me put it here on the side. So this is a frequency polygon. I could have also drawn a histogram. Let me once again select flow rate and efficiency and try to first have a scatter plot.

This scatter plot could be converted to a histogram here as well. You can see here the efficiency histogram with respect to the flow rate, showing how efficiency is varying. Also, just to show you how to draw a pie chart—a very quick introduction—let me say, for example, because 100% of data is to be used in a pie chart. Let me make a small table here. Let me say in this state, the percentage of pumps.

I'm talking about the percentage of pumps because we're discussing gear pumps in a specific area in a state. Let me try to just put some values here. Let me say in Punjab, I'm putting the percentage as 25%. In UP, let me say 20%. Bihar, let me say 15%. 60 is done. I'm going to make it total 100. Then let me say Tamil Nadu, 22%.

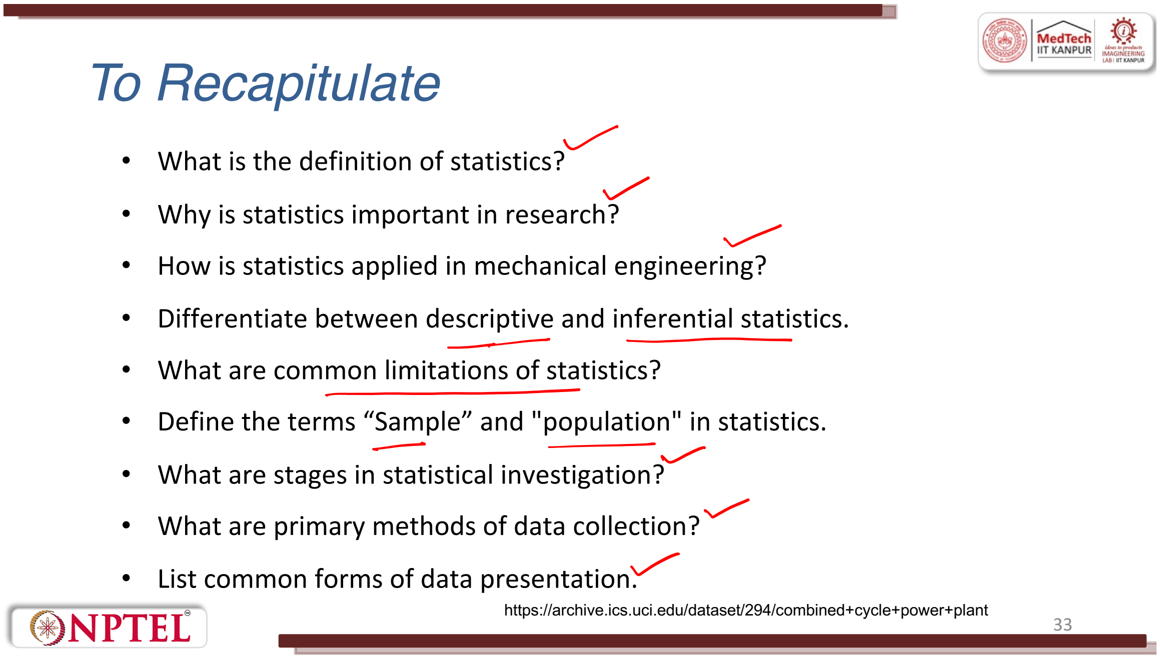
And let me select another state, Odisha, 18%. The total sum should be 100. Yes, the sum is 100. So, let's put this in a table. Now, this data is already selected.

I will insert a pie chart here. The kind of pie chart you would like to see, for example, is shown here. This is a simple two-dimensional percentage of pumps in Punjab, UP, Bihar, Tamil Nadu, and Russia. We can also display this in a slightly three-dimensional way. I have separated them.

I can bring them closer. I can separate them. Any way you like, I can separate one of the pie slices. For example, I have separated Tamil Nadu from the others. Tamil Nadu is shown in yellow color.

Here is the legend that is given. So this is simply using Excel. The data presentation you can do. A lot of plots are there. If you see the insert chart, if I come to the recommended charts, amongst all charts there are recent templates, column, line, pie, bar.

These are vertical and horizontal bars, the area XY scatter map; you can also have stock, surface, radar, or a kind of web-like chart that is there. A tree map is also similar to a pie chart. For example, if I try to draw a tree map amongst this, I will just select this, okay. A tree map is also very similar to what a pie chart is. The tree map is also showing wherever the maximum value is there: 25% is Punjab, it is showing Punjab in blue, then Tamil Nadu is 22%, Tamil Nadu in yellow. This is a tree map using the same data. I will also plot the same pie chart for you, and this sheet will be provided in the lecture notes.



To Recapitulate

- What is the definition of statistics?
- Why is statistics important in research?
- How is statistics applied in mechanical engineering?
- Differentiate between descriptive and inferential statistics.
- What are common limitations of statistics?
- Define the terms “Sample” and "population" in statistics.
- What are stages in statistical investigation?
- What are primary methods of data collection?
- List common forms of data presentation.

https://archive.ics.uci.edu/dataset/294/combined+cycle+power+plant

33

With this, I'm recapitulating the first two parts. We talked about the statistics, definitions of statistics, and the role of statistics in research, the role of statistics in mechanical engineering, then the differentiation between descriptive and inferential statistics. We tried to have some examples on certain applications that we saw throughout this course or throughout the series of courses in the previous sessions as well.

Then common limitations of statistics, the difference between sample and population in statistics, what are the stages in statistical investigation, primary and secondary methods of data collection, and a list of the common forms of data presentation. With this, I am closing this lecture. I will meet you in the next lecture, where I will talk about the basics of inferential statistics.

Thank you.