UNIT - II Data Types & Statistical Description :

Types of Data: Attributes and Measurement, What is an Attribute? The Type of an Attribute, The Different Types of Attributes, Describing Attributes by the Number of Values, Asymmetric Attributes, Binary Attribute, Nominal Attributes, Ordinal Attributes, Numeric Attributes, Discrete versus Continuous Attributes. Basic Statistical Descriptions of Data: Measuring the Central Tendency: Mean, Median, and Mode, Measuring the Dispersion of Data: Range, Quartiles, Variance, Standard Deviation, and Interquartile Range, Graphic Displays of Basic Statistical Descriptions of Data.

Data: It is how the data objects and their attributes are stored.

- A data object represents an entity—in a sales database, the objects may be customers, store items, and sales; in a medical database, the objects may be patients; in a university database, the objects may be students, professors, and courses. Data objects are typically described by attributes

- An **attribute** is an object's property or characteristics. For example. A person's hair colour, air humidity etc.
- An attribute set defines an **object**. The **object** is also referred to as a record of the instances or entity.
- For example, a sales data object may represent customers, sales, or purchases. When a data object is listed in a database they are called data tuples.

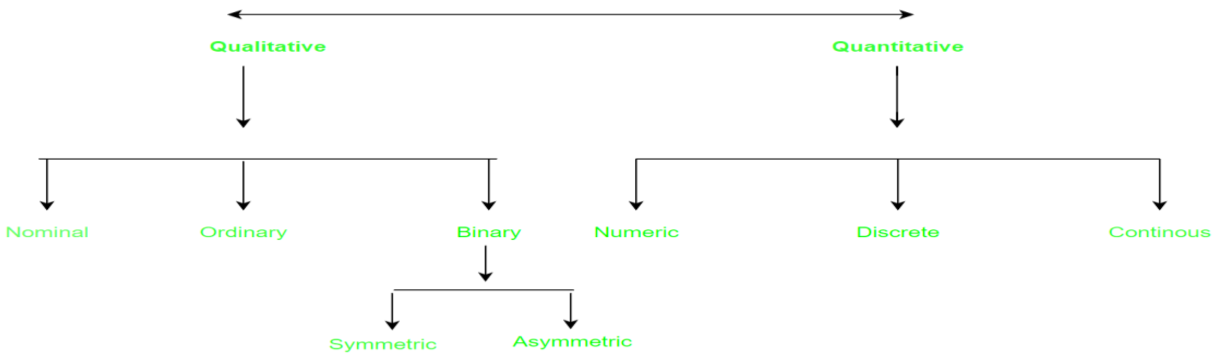
▪ **Attribute:**

It can be seen as a data field that represents the characteristics or features of a data object. For a customer, object attributes can be customer Id, address, etc. We can say that a **set of attributes used to describe a given object are known as attribute vector or feature vector.**

- An attribute is a data field, representing a characteristic or feature of a data object. The nouns attribute, dimension, feature, and variable are often used interchangeably.
- The term dimension is commonly used in data warehousing. Machine learning literature tends to use the term feature, while statisticians prefer the term variable. Data mining and database professionals commonly use the term attribute.
- Attributes describing a customer object can include, for example, customer ID, name, and address. Observed values for a given attribute are known as observations. A set of attributes used to describe a given object is called an attribute vector (or feature vector).

Type of attributes :

- This is the First step of Data-preprocessing. We differentiate between different types of attributes and then preprocess the data. So here is the description of attribute types.
- Qualitative (Nominal (N), Ordinal (O), Binary(B)).
- Quantitative (Numeric, Discrete, Continuous)

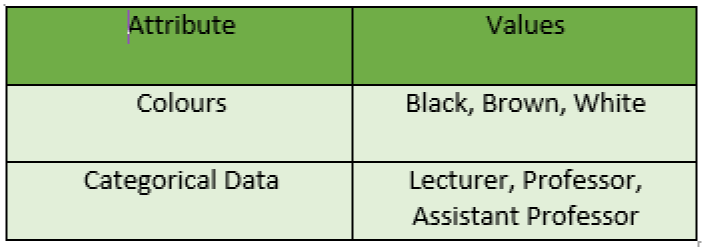


Qualitative Attributes:

1. Nominal Attributes – related to names: The values of a Nominal attribute are names of things, some kind of symbols. Values of Nominal attributes represents some category or state and that’s why nominal attribute also referred as categorical attributes and there is no order (rank, position) among values of the nominal attribute.
2. Nominal means “relating to names.” The values of a nominal attribute are symbols or names of things. Each value represents some kind of category, code, or state, and so nominal attributes are also referred to as categorical. The values do not have any meaningful order. In computer science, the values are also known as **enumerations**.

Example :

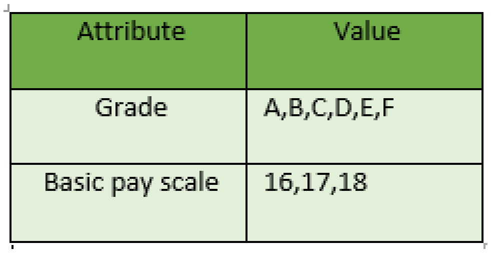
- **<u>Nominal attributes</u>**. Suppose that hair color and marital status are two attributes describing person objects. In our application, possible values for hair color are black, brown, blond, red, auburn, gray, and white. The attribute marital status can take on the values single, married, divorced, and widowed. Both hair color and marital status are nominal attributes. Another example of a nominal attribute is occupation, with the values teacher, dentist, programmer, farmer, and so on.



Attribute	Values
Colours	Black, Brown, White
Categorical Data	Lecturer, Professor, Assistant Professor

2. Ordinal Attributes : The Ordinal Attributes contains values that have a meaningful sequence or ranking(order) between them, but the magnitude between values is not actually known, the order of values that shows what is important but don’t indicate how important it is.

- Suppose that drink size corresponds to the size of drinks available at a fast-food restaurant. This nominal attribute has three possible values: small, medium, and large. The values have a meaningful sequence (which corresponds to increasing drink size).
- for example, assistant, associate, and full for professors Customer satisfaction had the following ordinal categories: 0: very dissatisfied, 1: somewhat dissatisfied, 2: neutral, 3: satisfied, and 4: very satisfied.



Attribute	Value
Grade	A,B,C,D,E,F
Basic pay scale	16,17,18

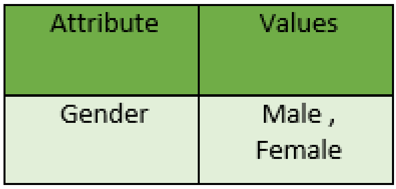
3. **Binary Attributes:** Binary data has only 2 values/states. For Example yes or no, affected or unaffected, true or false.

   A binary attribute is a nominal attribute with only two categories or states: 0 or 1, where 0 typically means that the attribute is absent, and 1 means that it is present. Binary attributes are referred to as Boolean if the two states correspond to true and false.

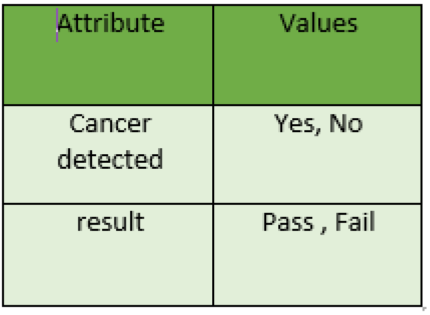
- **Symmetric:** Both values are equally important (Gender).
  - A binary attribute is symmetric if both of its states are equally

valuable and carry the same weight; that is, there is no preference on which outcome should be coded as 0 or 1. One such example could be the attribute gender having the states male and female.

- **Asymmetric:** Both values are not equally important (Result).
  - A binary attribute is asymmetric if the outcomes of the states are not equally impor- tant, such as the positive and negative outcomes of a medical test for HIV. By convention, we code the most important outcome, which is usually the rarest one, by 1 (e.g., HIV positive) and the other by 0 (e.g., HIV negative).



Attribute	Values
Gender	Male , Female



Attribute	Values
Cancer detected	Yes, No
result	Pass , Fail

<u>Quantitative Attributes</u>:

1. **Numeric:** A numeric attribute is quantitative because, it is a measurable quantity, represented in integer or real values. Numerical attributes are of 2 types, **interval**, and **ratio**.
2. **Ex:** e.g., **height, weight, temperature, blood glucose**, ...)

Numeric Attributes

- A numeric attribute is quantitative; that is, it is a measurable quantity, represented in integer or real values. Numeric attributes can be interval-scaled or ratio-scaled.

Interval-Scaled Attributes

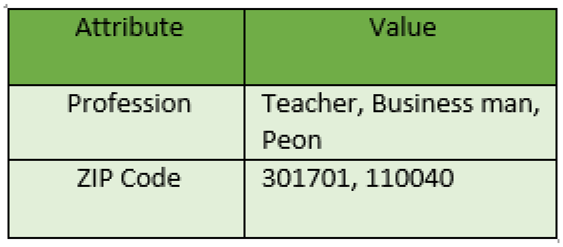
- Interval-scaled attributes are measured on a scale of equal-size units. The values of interval-scaled attributes have order and can be positive, 0, or negative. Thus, in addition to providing a ranking of values, such attributes allow us to compare and quantify the difference between values.

- An **interval-scaled** attribute has values, whose differences are interpretable, but the numerical attributes do not have the correct reference point, or we can call zero points. Data can be added and subtracted at an interval scale but can not be multiplied or divided. Consider an example of temperature in degrees Centigrade. If a day's temperature of one day is twice of the other day we cannot say that one day is twice as hot as another day.
- Eg: An interval scale is one where there is order and the difference between two values is meaningful. Examples of interval variables include: **temperature (Farenheit), temperature (Celcius),**
- A **ratio-scaled** attribute is a numeric attribute with a fix zero-point. If a measurement is ratio-scaled, we can say of a value as being a multiple (or ratio) of another value. The values are ordered, and we can also compute the difference between values, and the mean, median, mode, Quantile-range, and Five number summary can be given.

  E.g: **height, money, age, weight etc.**

3. **Discrete :** Discrete data have finite values it can be numerical and can also be in categorical form. These attributes has finite or countably infinite set of values.

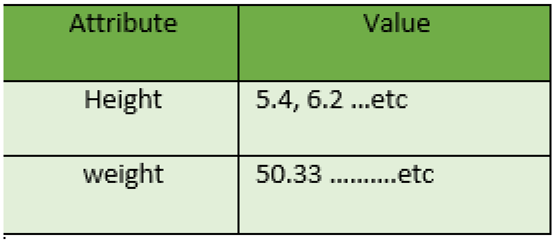
Example:



Attribute	Value
Profession	Teacher, Business man, Peon
ZIP Code	301701, 110040

3. Continuous: Continuous data have an infinite no of states. Continuous data is of float type. There can be many values between 2 and 3.

Example :



Attribute	Value
Height	5.4, 6.2 ...etc
weight	50.33 ....etc

Basic Statistical Descriptions of Data

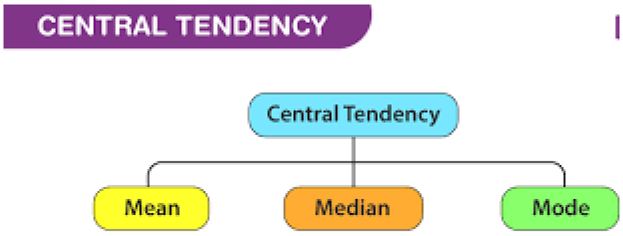
- **Statistical Analysis:** In statistics, data is collected, analyzed, explored, and presented to identify patterns and trends. Alternatively, it is referred to as quantitative analysis.
  - **Descriptive Statistics:** The purpose of descriptive statistics is to organize data and identify the main characteristics of that data. Graphs or numbers summarize the data. Average, Mode, SD(Standard Deviation), and Correlation are some of the commonly used descriptive statistical methods.

- **Inferential Statistics:** The process of drawing conclusions based on probability theory and generalizing the data. By analyzing sample statistics, you can infer parameters about populations and make models of relationships within data.

Measure of central tendency

- A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data.
- As such, measures of central tendency are sometimes called measures of central location.
- They are also classed as summary statistics. The mean (often called the average) is most likely the measure of central tendency that you are most familiar with, but there are others, such as the median and the mode.
- The central tendency is defined as the statistical measure that can be used to represent the entire distribution or a dataset using a single value called a measure of central tendency.

The mean, median and mode are all valid measures of central tendency,



- In statistics, the mean, median, and mode are the three most common measures of central tendency.
- Each one calculates the central point using a different method. Choosing the best measure of central tendency depends on the type of data you have.
- Measures of central tendency are summary statistics that represent the center point or typical value of a dataset.
- Examples of these measures include the mean, median, and mode. These statistics indicate where most values in a distribution fall and are also referred to as the central location of a distribution.

- Measures of central tendency are summary statistics that represent the center point or typical value of a dataset.
- Examples of these measures include the mean, median, and mode. These statistics indicate where most values in a distribution fall and are also referred to as the central location of a distribution.

Measures of Central Tendency Example

Example. The monthly salary of an employee for the 5 months is given in the table below,

Month	Salary
January	$105
February	$95
March	$105
April	$105
May	$100

- Suppose, we want to express the salary of the employee using a single value and not 5 different values for 5 months. This value that can be used to represent the data for salaries for 5 months here can be referred to as the measure of central tendency. The three possible ways to find the central measure of the tendency for the above data are,

- **Mean:** The mean salary of the given salary can be used as on of the measures of central tendency, i.e., $\bar{x}$ = (105 + 95 + 105 + 105 + 100)/5 = $102.
- **Mode:** If we use the most frequently occurring value to represent the above data, i.e., $105, the measure of central tendency would be mode.
- **Median:** If we use the central value, i.e., $105 for the ordered set of salaries, given as, $95, $100, $105, $015, $105, then the measure of central tendency here would be median.

Measuring the Central Tendency: Mean, Median, and Mode

✓ Mean is the average of the given numbers and is calculated by dividing the sum of given numbers by the total number of numbers.

Mean = (Sum of all the observations/Total number of observations)

Example:

What is the mean of 2, 4, 6, 8 and 10?

Solution:

First, add all the numbers.

2 + 4 + 6 + 8 + 10 = 30

Now divide by 5 (total number of observations).

Mean = 30/5 = 6

Mean Symbol (X Bar)

- The symbol of mean is usually given by the symbol ‘$\bar{x}$’. The bar above the letter x, represents the mean of x number of values.
- $\bar{X}$ = (Sum of values ÷ Number of values)
- $\bar{X} = (x_1 + x_2 + x_3 + \ldots + x_n)/n$

Mean = Sum of the Given Data/Total number of Data

- To calculate the arithmetic mean of a set of data we must first add up (sum) all of the data values (x) and then divide the result by the number of values (n). Since ∑ is the symbol used to indicate that values are to be summed (see Sigma Notation) we obtain the following formula for the mean ($\bar{x}$):

- **$\bar{x}=\sum x/n$**

Example:

In a class there are 20 students and they have secured a percentage of 88, 82, 88, 85, 84, 80, 81, 82, 83, 85, 84, 74, 75, 76, 89, 90, 89, 80, 82, and 83.

Find the mean percentage obtained by the class.

Solution:

Mean = Total of percentage obtained by 20 students in class/Total number of students

= [88 + 82 + 88 + 85 + 84 + 80 + 81 + 82 + 83 + 85 + 84 + 74 + 75 + 76 + 89 + 90 + 89 + 80 + 82 + 83]/20

= 1660/20

= 83

Hence, the **mean percentage of each student in the class is 83%.**

- **Median**
- A Median is a middle value for a sorted data. The sorting of the data can be done either in ascending order or in descending order. A median divides the data into two halves. The formula for median:
- If the number of values (n value) in the data set is odd then the formula to calculate median is:

$$\textbf{Median} = ((n + 1)/2)^{th} \textbf{ term}$$

If the number of values (n value) in the data set is even then the formula to calculate median is:

$$\textbf{Median} = [(n/2)^{th} \textbf{ term} + \{(n/2) + 1\}^{th} \textbf{ term}] / 2$$

Median

- The median is the mid-value or average of a data set. The data set must be sorted either in ascending or descending order.
- In other words, it is a middle value of a sorted data set. We find mean or average by using the median.

How to Find Median

To determine the median of a data set, the values of the data set must be sorted or arranged in either ascending or descending order. The data may be in two formats:

- Ungrouped Frequency Distribution
- Grouped Frequency Distribution

Ungrouped Frequency Distribution

In an ungrouped frequency distribution, the data may of two types:

- When an odd number of the frequency distribution is given
- When even number of the frequency distribution is given

When an odd number of the frequency distribution is given

- To find the median of odd frequency distribution follow the steps given below. But remember that data must be sorted. After sorting the data, use the following formula:

$$\text{Median} = \left(\frac{n+1}{2}\right)\text{th item}$$

Where **n** is the total number of items in the data set.

Another quick method to find the median is:

- First, sort the values or items.
- Pick the middle value as the median.

Example 1: Find the median of 23, 2, 12, 33, 65, 45, and 9.

Solution:

First, we sort the given data set.

2, 9, 12, **23**, 33, 45, 65

There is a total of 7 values, so the mid-value (4th) will be median, i.e. 23.

Similarly, we can find the median using the formula:

$$\text{Median} = \left(\frac{n+1}{2}\right)\text{th item}$$

Putting the value of n in the formula, we get:

$$\text{Median} = \left(\frac{7+1}{2}\right)\text{th item}$$

$$\text{Median} = 4\text{th item}$$

The 4th item or value will be median, i.e. 23.

Hence, the median of the given data set is 23.

When even number of the frequency distribution is given

- To find the median of the data set that contains even number of frequency distribution, we must follow the steps given below:
- Sort the values of the data set.
- Find the middle pair and its values.
- Sum up the values and divide it by 2.
- The value that we get on dividing is the median of the given data set.
- We can also write the above steps in terms of the formula:

$$\text{Median} = \frac{\left(\frac{\text{N}}{2}\right)\text{th item} + \left(\frac{\text{N}}{2}\right) + 1\text{ th item}}{2}$$

Where **N** is the total number of items in the data set.

Let's understand the example of even frequency distribution.

Example 2: Find the median of the following list:

1, 5, 77, 32, 65, 12, 44, 21, 90, 34, 8, 56, 4, 99

Solution:

Step 1: Sort the given list.

1, 4, 5, 8, 12, 21, **32, 34,** 44, 56, 65, 77, 90, 99

There are total 14 values in the list.

Step 2: **Find the middle pair and its values.**

The middle pair terms of the list are **7th** and **8th** and its values are **32** and **34,** respectively.

Step 3: Sum up the values and divide it by 2.

$$\frac{32+34}{2} = \frac{66}{2} = 33$$

- ❖ A point to be noticed here is that **33** is not in the list. But it indicates that half values in the list are less than 33, and half values are greater than 33.
- ❖ Let's find the median through the formula which we have learned above.

$$\text{Median} = \frac{\left(\frac{\text{N}}{2}\right)\text{th item} + \left(\frac{\text{N}}{2}\right) + 1\text{ th item}}{2}$$

$$\text{Median} = \frac{\left(\frac{14}{2}\right)\text{th item} + \left(\frac{14}{2}\right) + 1\text{ th item}}{2}$$

$$\text{Median} = \frac{7\text{th item} + (7) + 1\text{ th item}}{2}$$

$$\text{Median} = \frac{7\text{th item} + 8\text{th item}}{2}$$

Count the **7th** and **8th** item in the list and put the values.

$$\text{Median} = \frac{32+34}{2} = \frac{66}{2} = 33$$

Hence, the median of the given list is 33.

Example 1:

The median of the data 30, 40, 10, 20, 50 is:

Step 1: *Order the given data in ascending order as:*

10, 20, 30, 40, 50

Step 2: *Check n (number of terms of data set) is even or odd and find the median of the data with respective 'n' value.*

Step 3: *Here, n = 5 (odd) then Median = [(n + 1)/2]*th *term 10, 20, 30, 40, 50*

The median of the data is [(5 + 1)/2]th *term is 30.*

Example 2:

25, 12, 5, 24, 15, 22, 23, 25

Step 1: *Order the given data in ascending order as:*

5, 12, 15, 22, 23, 24, 25, 25

Step 2: *Check n (number of terms of data set) is even or odd and find the median of the data with respective 'n' value.*

Step 3: *Here, n = 8 (even) then,*

Median = [(n/2)th *term + {(n/2) + 1)*th *term] / 2*

Median = [(8/2)th *term + {(8/2) + 1}*th *term] / 2 = (22+23) / 2 = 22.5*

Mode

A mode is the most frequent value or item of the data set. A data set can generally have one or more than one mode value. If the data set has one mode then it is called "Uni-modal". Similarly, If the data set contains 2 modes then it is called "Bimodal" and if the data set contains 3 modes then it is known as "Trimodal". If the data set consists of more than one mode then it is known as "multi-modal"(can be bimodal or trimodal). There is no mode for a data set if every number appears only once.

Example 1:

If the data set is {1, 2, 2, 3, 3, 4, 5} then it has 2 modes i.e, 2 and 3 (bi-modal). Since, both the values 2 and 3 are repeating twice in the data set.

Example 2:

If the data set is {15, 42, 65, 65, 95} then the mode is 65 (uni-modal). Since 65 is the only repeating value in the data set.

Range

It is the difference between the highest value and the lowest value. It is a way to understand how the numbers are spread in a data set. Formula to find Range is:

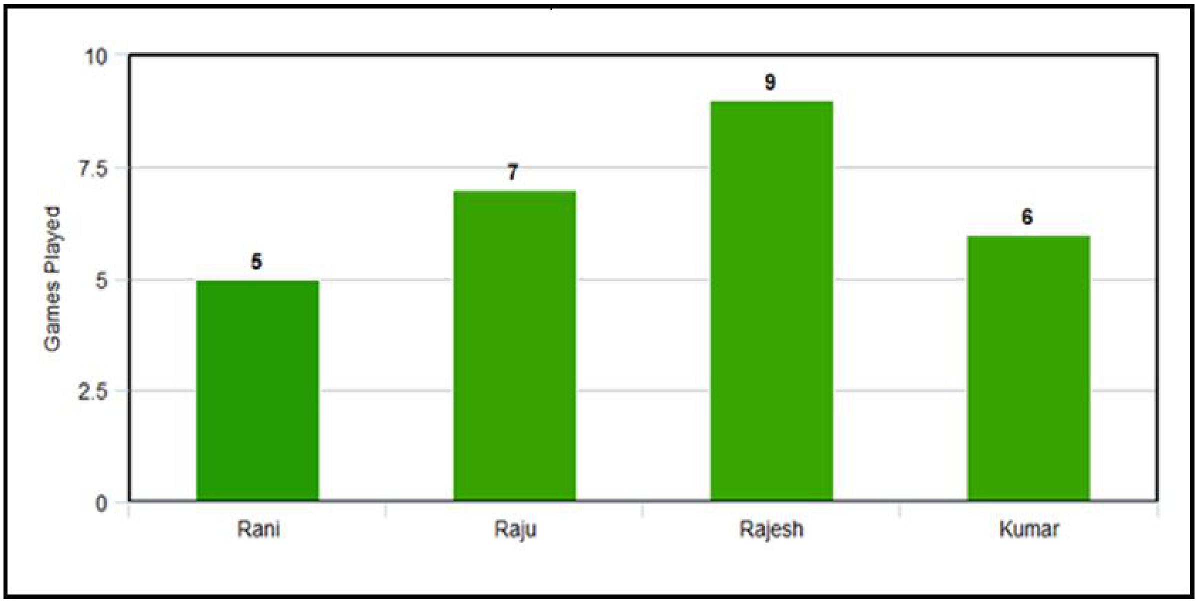
Range = Highest value – Lowest Value

Example:

If the data set is {12, 19, 6, 2, 15, 4} then the lowest value is 2 and the highest value is 19.

So the range is 19 − 2 = 17.

Reading Bar Charts: Putting it Together with Central Tendency



Question 1. Finding Mean for the above bar chart.

Mean = (sum of all data values) / (number of values)

Mean = (5 + 7 + 9 + 6) / 4 = 27 / 2 = 6.75

Question 2. Finding the Median for the above bar chart:

Order the given data in ascending order as: 5, 6, 7, 9

Here, n = 4 (number of students which is even)

Median = [(n/2)th term + {(n/2) + 1}th term] / 2

Median = (6 + 7) / 2 = 6.5

Question 3. Finding Mode for the above bar chart:

Mode = most frequent value = 9 (highest value)

Question 4. Finding the range for the above bar chart:

Range = highest value – lowest value

Range = 9 – 5 = 4

Measuring the Dispersion of Data: Range, Quartiles, Variance, Standard Deviation, and Interquartile Range,

- ✓ Dispersion or variability describes how items are distributed (scattered) from each other and the center of a distribution.
- ✓ Example: Height of students
- ✓ In statistics, dispersion helps to understand the distribution of the data.

- ✓ Dispersion in statistics is a way of describing how to spread out a set of data is. Dispersion is the state of data getting dispersed, stretched, or spread out in different categories. It involves finding the size of distribution values that are expected from the set of data for the specific variable. The meaning of dispersion in statistics is “numeric data that is likely to vary at any instance of average value assumption”.

Measures of Dispersion

- ✓ Statistical methods that help to know about the distribution or the spread of the data points in the datasets are known as Measures of Dispersion.

- ✓ There are 4 methods to measures the dispersion of the data:
- ✓ Range
- ✓ Interquartile Range
- ✓ Variance
- ✓ Standard Deviation

❖ Range

- ✓ The range is the easiest measure of dispersion. It is simply calculated by subtracting the highest value from the lowest value.

✓ **Range = Highest Value – Lowest Value**

- ✓ EX:

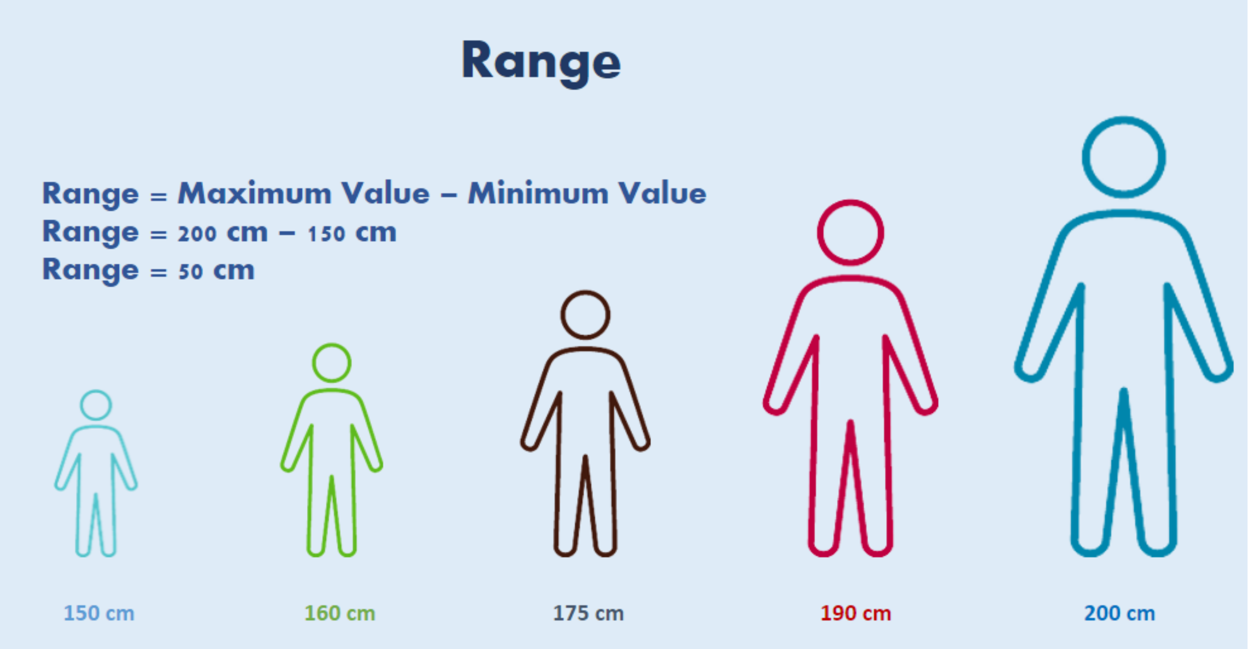
- ✓ Problem Statement:

- ✓ Let there be 5 students in the class having heights of 150cm, 160cm, 175cm, 190cm and 200 cm.

- ✓ Calculate the range of heights?

Range = 200cm – 150cm

Hence, Range = 50cm



Range for ungrouped data:
Question 1: Find out the range for the following observations.

20, 24, 31, 17, 45, 39, 51, 61

Solution:

The largest value in the given observations is 61 and the smallest value is 17.

The Range is 61 – 17 = 44

Range for grouped data:

Question 2: Find out the range for the following frequency distribution table for the marks scored by class 10 students.

Marks Intervals	**Number of Students**
0-10	5
10-20	8

20-30	15
30-40	9

Solution:

For the largest value – Take higher limit of the highest class = 40

For the smallest value – Take lower limit of the lowest class = 0

Range = 40 – 0

Range = 40

❖ Interquartile Range

✓ Before defining the interquartile range, let's discuss the quartiles and five-number summary

Quartiles: Quartiles divide the set into 4 equal parts.

There are three quartiles Q1, Q2 and Q3, where Q2 is the median of the distribution.

Five number summary:

Every dataset can be described using these 5 numbers

- Lowest value
- Q1: 25 percentile
- Q2: Median
- Q3: 75 Percentile
- Highest Value

Interquartile Range

Interquartile Range: Interquartile range is defined as the range between 75 percentile (Q3) and 25 percentile (Q1).

IQR = Q3 – Q1

Let’s understand Q1, Q2, Q3 and the Interquartile range by an example.

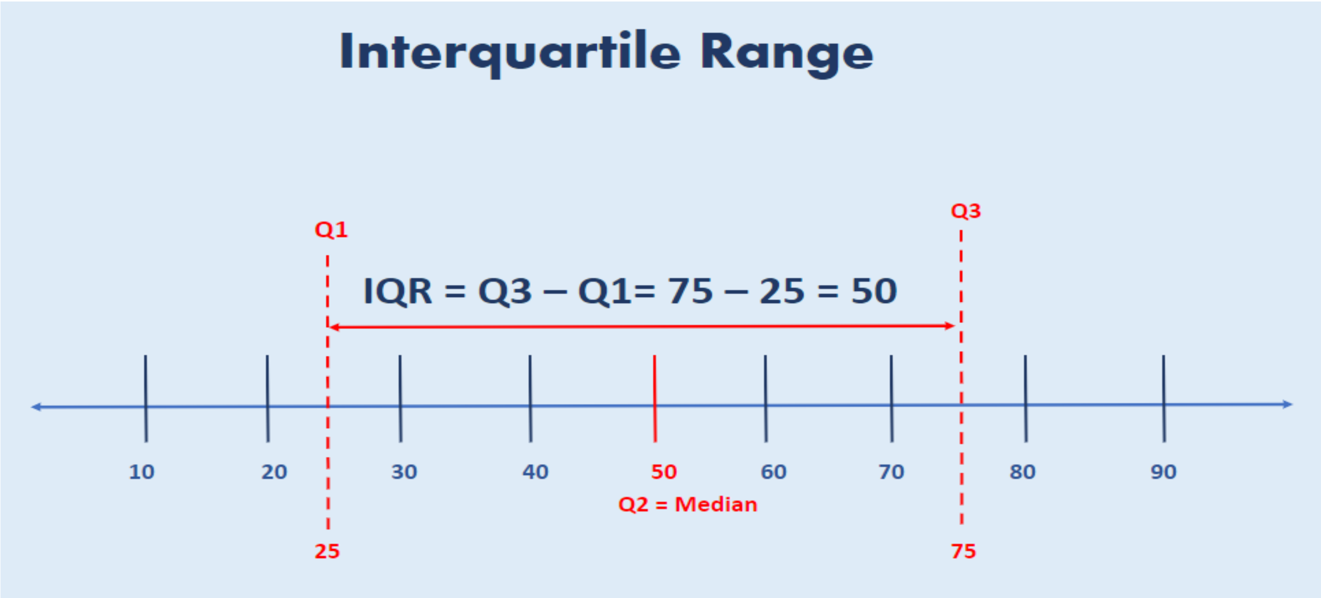
Problem Statement:

Let there are 8 numbers between 10 and 90 which are equally distributed.

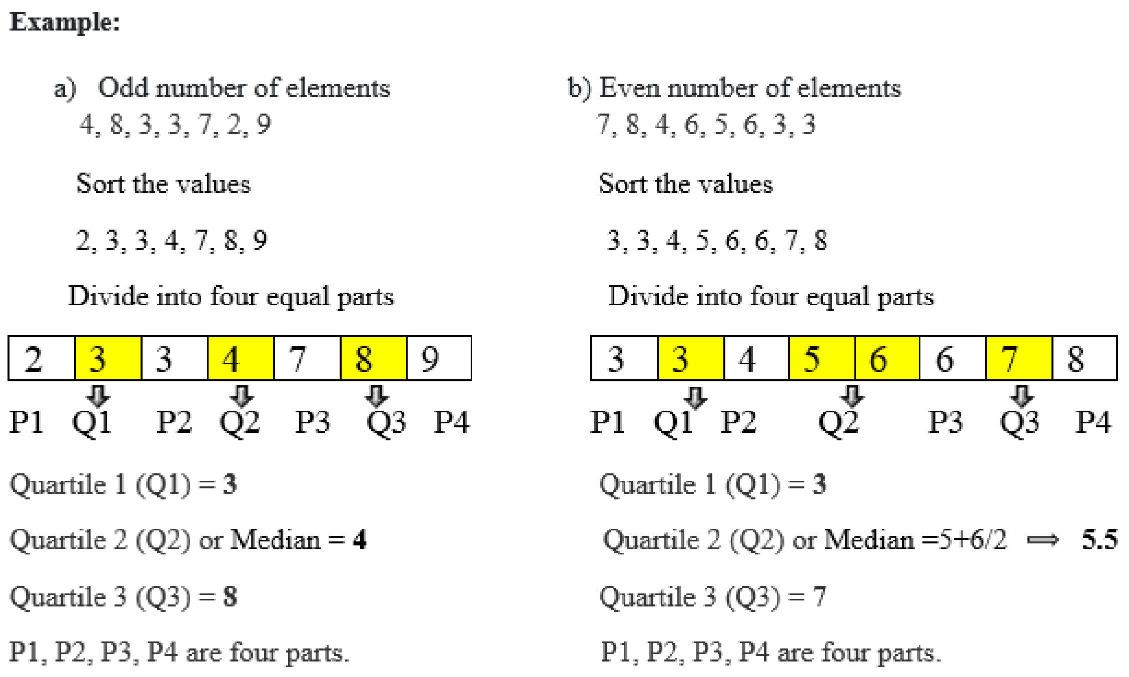
Define the five-number summary and find the Interquartile Range?

- Lowest value : 10
- Q1 (25 percentile) : 25
- Q2 (50 percentile) : 50
- Q3 (75 percentile) : 75
- Highest value : 90
- Interquartile Range(IQR) = Q3 – Q1 = 75 – 25 = 50

Interquartile Range = 50



Ex: 2



Example:

a) Odd number of elements

4, 8, 3, 3, 7, 2, 9

Sort the values

2, 3, 3, 4, 7, 8, 9

Divide into four equal parts

2	3	3	4	7	8	9

P1 Q1 P2 Q2 P3 Q3 P4

Quartile 1 (Q1) = **3**

Quartile 2 (Q2) or Median = **4**

Quartile 3 (Q3) = **8**

P1, P2, P3, P4 are four parts.

b) Even number of elements

7, 8, 4, 6, 5, 6, 3, 3

Sort the values

3, 3, 4, 5, 6, 6, 7, 8

Divide into four equal parts

3	3	4	5	6	6	7	8

P1 Q1 P2 Q2 P3 Q3 P4

Quartile 1 (Q1) = **3**

Quartile 2 (Q2) or Median =5+6/2 ⟹ **5.5**

Quartile 3 (Q3) = 7

P1, P2, P3, P4 are four parts.

Variance

Definition

- ✓ Variance is a measure of how data points differ from the mean. According to Layman, a variance is a measure of how far a set of data (numbers) are spread out from their mean (average) value.

- ✓ Variance means to find the expected difference of deviation from actual value. Therefore, variance depends on the standard deviation of the given data set

Ex:

How to Calculate Variance

Variance can be calculated easily by following the steps given below:

- Find the mean of the given data set. Calculate the average of a given set of values
- Now subtract the mean from each value and square them
- Find the average of these squared values, that will result in variance

Say if x_1, x_2, x_3, x_4, …,x_n are the given values.

Therefore, the mean of all these values is:

$\bar{x} = (x_1+x_2+x_3+\ldots+x_n)/n$

Now subtract the mean value from each value of the given data set and square them.

$(x_1-\bar{x})^2$, $(x_2-\bar{x})^2$, $(x_3-\bar{x})^2$,…….,$(x_n-\bar{x})^2$

Find the average of the above values to get the variance.

$\text{Var}(X) = [(x_1-\bar{x})^2+ (x_2-\bar{x})^2+ (x_3-\bar{x})^2+\ldots\ldots.+(x_n-\bar{x})^2]/n$

Hence, the variance is calculated.

Example of Variance

Let's say the heights (in mm) are 610, 450, 160, 420, 310.

Mean and Variance is interrelated. The first step is finding the mean which is done as follows,

Mean = (610+450+160+420+310)/ 5 = 390

So the mean average is 390 mm.

To calculate the Variance, compute the difference of each from the mean, square it and find then find the average once again.

So for this particular case the variance is :

$= (220^2 + 60^2 + (-230)^2 +30^2 + (-80)^2)/5$

= (48400 + 3600 + 52900 + 900 + 6400)/5

Final answer : Variance = 22440

Example: Find the variance of the numbers 3, 8, 6, 10, 12, 9, 11, 10, 12, 7.

Solution:

Given,

3, 8, 6, 10, 12, 9, 11, 10, 12, 7

Step 1: Compute the mean of the 10 values given.

Mean = (3+8+6+10+12+9+11+10+12+7) / 10 = 88 / 10 = 8.8

Step 2: Make a table with three columns, one for the X values, the second for the deviations and the third for squared deviations. As the data is not given as sample data so we use the formula for population variance. Thus, the mean is denoted by μ.

Value X	X – μ	$(X – μ)^2$
3	-5.8	33.64
8	-0.8	0.64
6	-2.8	7.84
10	1.2	1.44
12	3.2	10.24
9	0.2	0.04
11	2.2	4.84
10	1.2	1.44
12	3.2	10.24
7	-1.8	3.24
Total	0	73.6

Step 3:

σ2=∑(X−μ)2N
= 73.6 / 10

= 7.36

Standard Deviation

- ✓ Standard deviation is a metric that represents the amount to which various values of a statistical series tend to fluctuate or disperse from its mean or median. It describes how the values are distributed over the data sample and is a measure of the data points' deviation from the mean.
- ✓ The square root of the variance of a sample, statistical population, random variable, data collection, or probability distribution is its standard deviation.

Steps to Calculate Standard Deviation

- Find the mean, which is the arithmetic mean of the observations.
- Find the squared differences from the mean. (The data value - mean)2
- Find the average of the squared differences. (Variance = The sum of squared differences ÷ the number of observations)

- Find the square root of variance. (Standard deviation = √Variance)

Ex: 1

Consider the data set: 2, 1, 3, 2, 4. The mean and the sum of squares of deviations of the observations from the mean will be 2.4 and 5.2, respectively. Thus, the standard deviation will be √(5.2/5) = 1.01.

Ex: 2

For example: Take the values 2, 1, 3, 2 and 4.

1. Determine the mean (average):

2 + 1 +3 + 2 + 4 = 12
12 ÷ 5 = **2.4 (mean)**

2. Subtract the mean from each value:

2 - 2.4 = **-0.4**

1 - 2.4 = **-1.4**

3 - 2.4 = **0.6**

2 - 2.4 = **-0.4**

4 - 2.4 = **1.6**

3. Square each of those differences:

-0.4 x -0.4 = **0.16**

-1.4 x -1.4 = **1.96**

0.6 x 0.6 = **0.36**

-0.4 x -0.4 = **0.16**

1.6 x 1.6 = **2.56**

4. Determine the average of those squared numbers to get the variance.

0.16 + 1.96 + 0.36 + 0.16 + 2.56 = **5.2**

5.2 ÷ 5 = **1.04 (variance)**

5. Find the square root of the variance.

Square root of 1.04 = **1.01**

The standard deviation of the values 2, 1, 3, 2 and 4 is **1.01.**

EX: 3

A class of students took a math test. Their teacher wants to know whether most students are performing at the same level, or if there is a high standard deviation.

1. The scores for the test were 85, 86, 100, 76, 81, 93, 84, 99, 71, 69, 93, 85, 81, 87, and 89. When the teacher adds them together, she gets 1279. She divides by the number of scores (15) to get the mean score.

1279 ÷ 15 =**85.2 (mean)**

2. 85.2 is a high score, but is everyone performing at that level? To find out, the teacher subtracts the mean from every test score.

85 - 85.2 = **-0.2**
86 - 85.2 = **0.8**
100 - 85.2 = **14.8**
76 - 85.2 = **-9.2**
81 - 85.2 = **-4.2**
93 - 85.2 = **7.8**
84 - 85.2 = **-1.2**
99 - 85.2 = **13.8**
71 - 85.2 = **-14.2**
69 - 85.2 = **-16.2**
93 - 85.2 = **7.8**
85 - 85.2 = -**0.2**
81 - 85.2 = **-4.2**
87 - 85.2 = **1.8**
89 - 85.2 = **3.8**

3. She squares each difference:

-0.2 x -0.2 = **0.04**
0.8 x 0.8 = **0.64**
14.8 14.8 = **219.04**
-9.2 x -9.2 = **84.64**
-4.2 x -4.2 = **17.64**
7.8 x 7.8 = **60.84**
-1.2 x -1.2 = **1.44**
13.8 x 13.8 = **190.44**
-14.2 x -14.2 = **201.64**
-16.2 x -16.2 = **262.44**
7.8 x 7.8 = **60.84**
-0.2 x -0.2 = **0.04**
-4.2 x -4.2 = **17.64**

1.8 x 1.8 = **3.24**
3.8 x 3.8 = **14.44**

4. The teacher finds the variance, which is the average of the squares:

0.04 + 0.64 + 219.04 + 84.64 + 17.64 + 60.84 +1.44 +190.44 +201.64 +262.44 + 60.84 + 0.04 + 17.64 + 3.24 + 14.44 = 1135

830.64 ÷ 15 = **75.6 (variance)**

5. Last, the teacher finds the square root of the variance:

Square root of 75.6 = **8.7 (standard deviation)**

The standard deviation of these tests is **8.7** points out of 100. Since the variance is somewhat low, the teacher knows that most students are performing around the same level.

EX:4

A market researcher is analyzing the results of a recent customer survey that ranks a product from 1 to 10. He wants to have some measure of the reliability of the answers received in the survey in order to predict how a larger group of people might answer the same questions.

Because this is a sample size, the researcher needs to subtract 1 from the total number of values in step 4.

1. The scores for the survey are 9, 7, 10, 8, 9, 7, 8, and 9. The mean is **8.4.**
2. The researcher subtracts the mean from every score (differences: 0.6, -1.4, 1.6, -0.4, 0.6, -1.4, -0.4, 0.6).
3. He squares each number (0.36, 1.96, 2.56, 0.16, 0.36, 1.96, 0.16, 0.36).
4. Because this is a sample of responses, the researcher subtracts one from the number of values (8 values -1 = 7) to average squares and find the variance: **1.12 (variance)**

5. Last, the researcher finds the square root of the variance: **1.06 (standard deviation)**

The standard deviation is 1.06, which is somewhat low. The researcher now knows that the results of the sample size are probably reliable.

Graphic Displays of Basic Statistical Descriptions of Data.

.Graphical Representation of Data

- ✓ In today's world of the internet and connectivity, there is a lot of data available and some or the other method is needed for looking at large data, the patterns, and trends in it.
- ✓ There is an entire branch in mathematics dedicated to dealing with collecting, analyzing, interpreting, and presenting the numerical data in visual form in such a way that it becomes easy to understand and the data becomes easy to compare as well, the branch is known as **Statistics**.
- ✓ There are two ways of representing data,
- ✓ **Tables**
- ✓ **Pictorial Representation through graphs.**

- ✓ They say, "A picture is worth the thousand words". It's always better to represent data in graphical format.
- ✓ Study the graphic displays of basic statistical descriptions.
- ✓ These include

  - ❖ **quantile plots,**
  - ❖ **quantile–quantile plots, (Q-Q)**

- **histograms,**
- **and scatter plots.**

Such graphs are helpful for the visual inspection of data, which is useful for data preprocessing

Scatter Plot Chart

- ✓ Scatter Plot refers to a two-dimensional chart that visually represents supplied data in real-time.
- ✓ Generally, the scatter plot visualizes two sets of data on the X and Y axis that are co-related
- ✓ This type of chart displays many points on vertical and horizontal axes as per the supplied data sets, and it is mainly used to show relationships between two variables.
- ✓ A scatter plot works by placing one variable on the vertical axis and a different variable on the horizontal axis.
- ✓ Each piece of data is then plotted as a discrete point on the chart. Both the X and Y axis display values in a scatter plot, which means that the scatter chart has no category axis.
- ✓ By convention, the X-axis represents arbitrary values that do not depend on another variable, called the independent variable. Besides, Y values are placed on the vertical axis and represent the dependent variable.

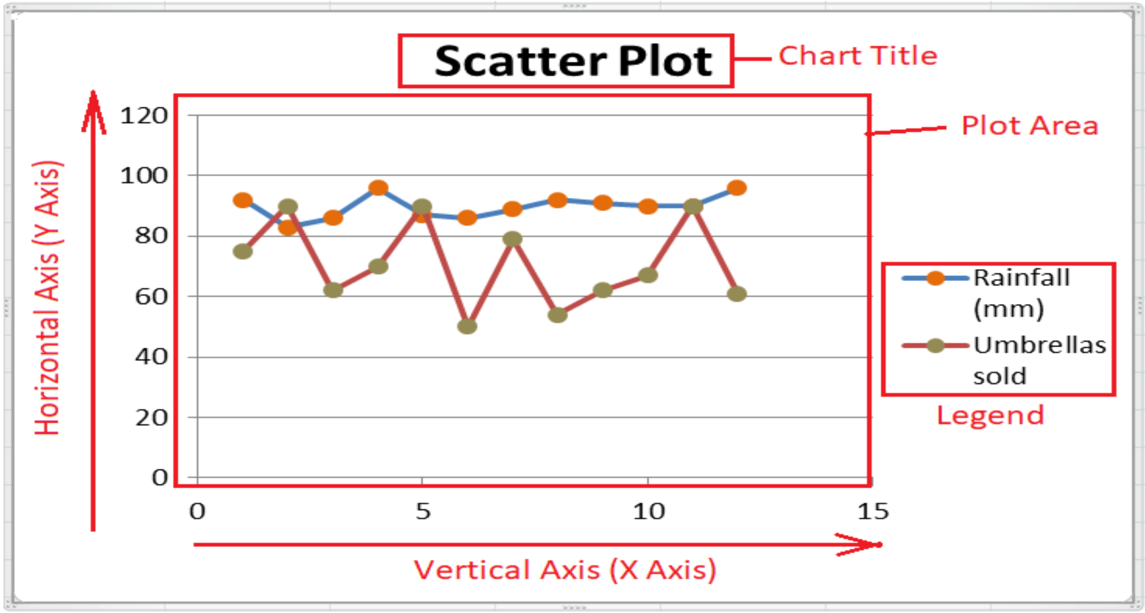
These charts are also known by many other names, such as '**Scatter Graphs, Scatter Charts, Scattergrams, Scatter Diagrams, XY Graph,** etc.'

Components of Scatter Plot Chart

There are mainly five components in a Scatter Plot Chart, as listed below:

- **Plot Area:** A graphical form/area within the sheet where the data is drawn is called the Plot Area.

- **Chart Title:** A chart title represents the subject of the plotted chart that primarily helps determine the chart's topic or motive. The text in the chart title can be edited, and the position can be arranged accordingly.
- **Vertical Axis:** An axis that lies vertically in the chart window is called the vertical axis, and it is located on the bottom area of the plot area. Since the vertical axis typically represents the measurement values across X-axis, it is known as the X-axis.
- **Horizontal Axis:** An axis that lies horizontally in the chart window is called the horizontal axis, and it is located on the left side of the plot area. Since the horizontal axis represents the different data categories across Y-axis, it is also known as the Y-axis. We can group series data on the horizontal axis.
- **Legend:** The legend is another useful component of the chart that helps list and distinguish various data groups. We can move the legend or change the legend's position accordingly, and it can be placed on any side in the chart window.



Advantages of using Scatter Plots

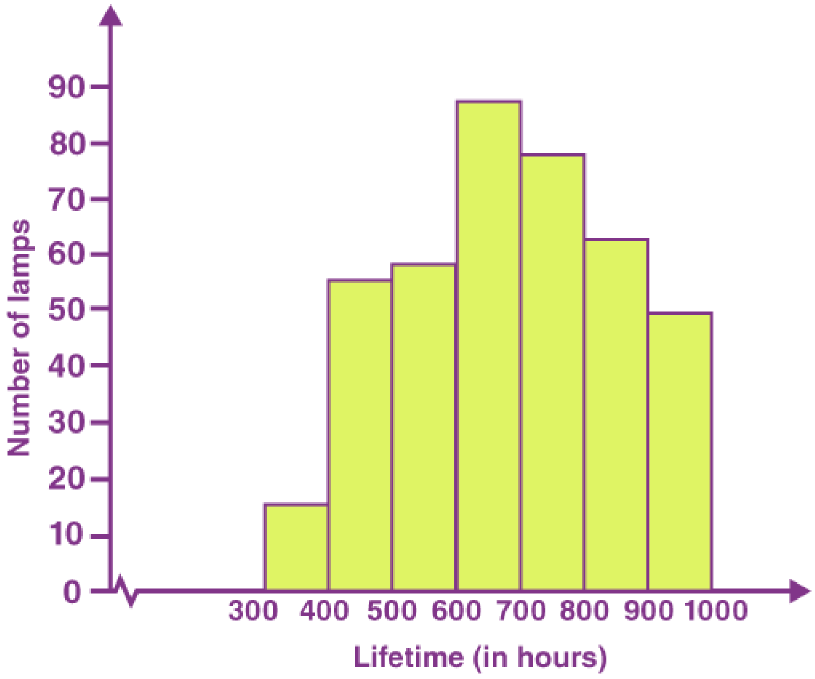
- The scatter charts help determine the relationship between two or more variables, and they mainly showcase the relationship of one variable concerning another.
- The scatter plots can show correlations visually.
- It is easy to analyze the maximum and minimum values (high and low) in scatter charts on the data flow range.
- The scatter charts are used for various scientific analyses because plotting these charts is moderately easy, and perception and readings are accurate.

Histogram

- ✓ A **histogram** is a graphical representation of the frequency distribution of **continuous** series using rectangles.
- ✓ The **x-axis of the graph represents the class interval**, and the **y-axis shows the various frequencies corresponding to different class intervals**.
- ✓ A histogram is a **two-dimensional diagram** in which the width of the rectangles shows the **width** of the class intervals, and the length of the rectangles depicts the corresponding frequency.
- ✓ There are no gaps between two consecutive rectangles based on the fact that histograms can be drawn when data are in the form of the frequency distribution of **continuous series.**

- **Question:** The following table gives the lifetime of 400 neon lamps. Draw the histogram for the below data.

Lifetime (in hours)	**Number of lamps**
300 – 400	14
400 – 500	56
500 – 600	60
600 – 700	86
700 – 800	74
800 – 900	62
900 – 1000	48



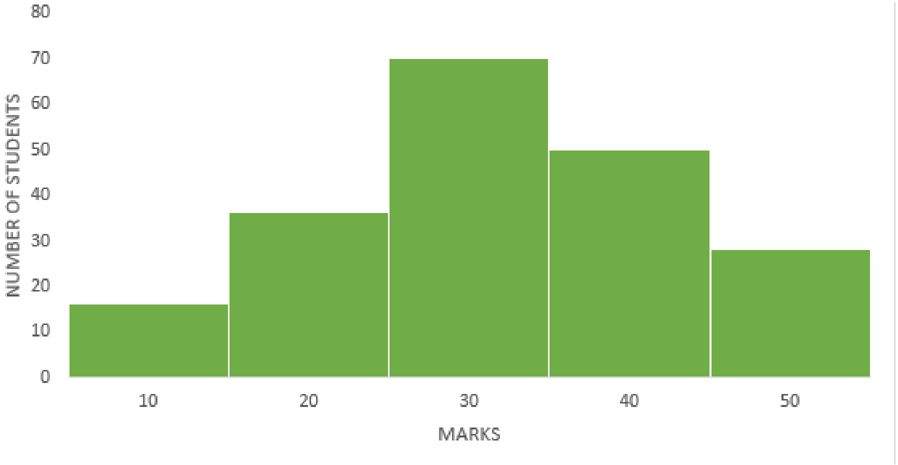
- Example:

Present the following information in the form of a Histogram:

Marks	0-10	10-20	20-30	30-40	40-50
Number of students	16	36	70	50	28

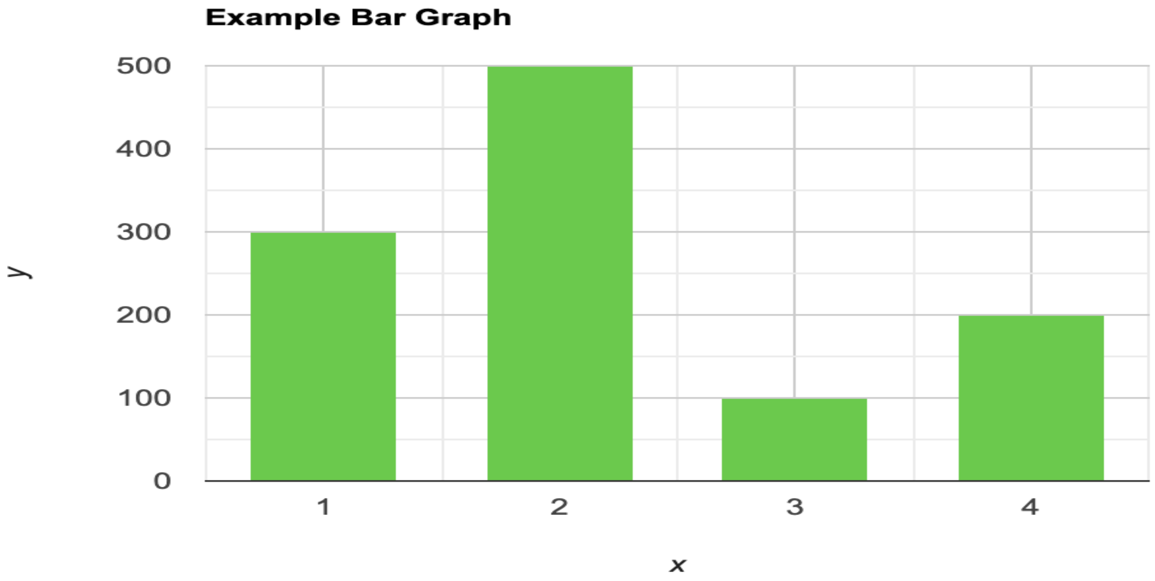
Solution

- It is visible that the set of data given is of the equal class interval; i.e., the difference between the upper limit and the lower limit of each class interval is 10. So, drawing a Histogram is feasible.
- The X-axis represents the marks (class intervals), and Y-axis represents the number of students (frequency distribution).



bar graph

- ✓ A bar graph is a type of graphical representation of the data in which bars of uniform width are drawn with equal spacing between them on one axis (x-axis usually), depicting the variable. The values of the variables are represented by the height of the bars.

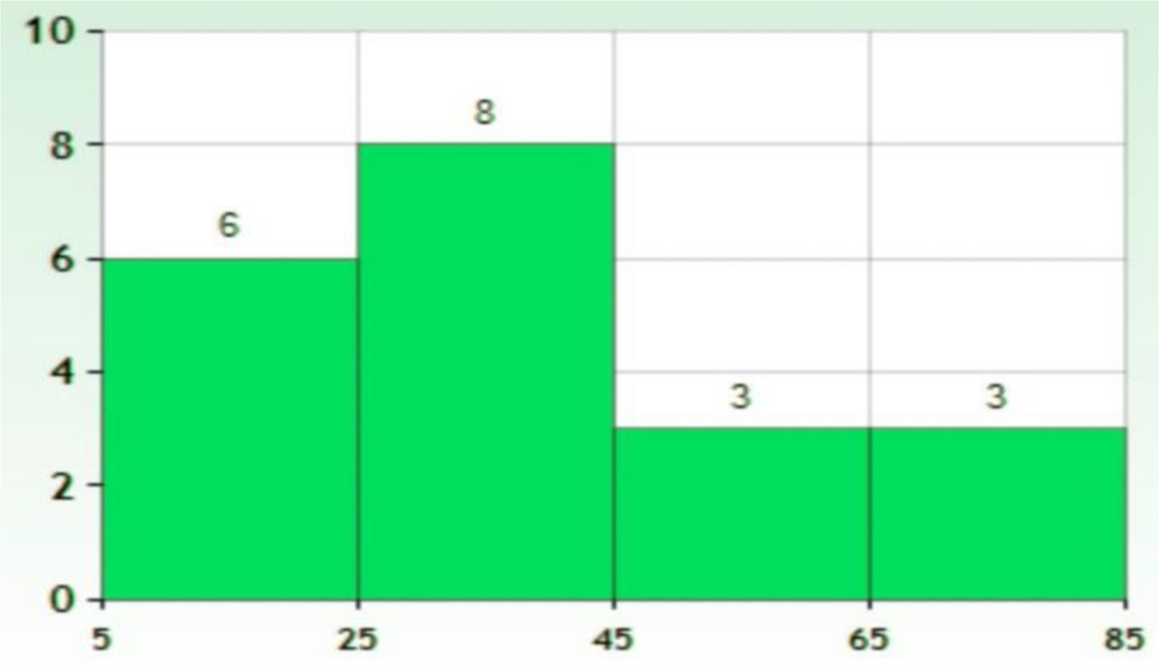


Note :

Histograms

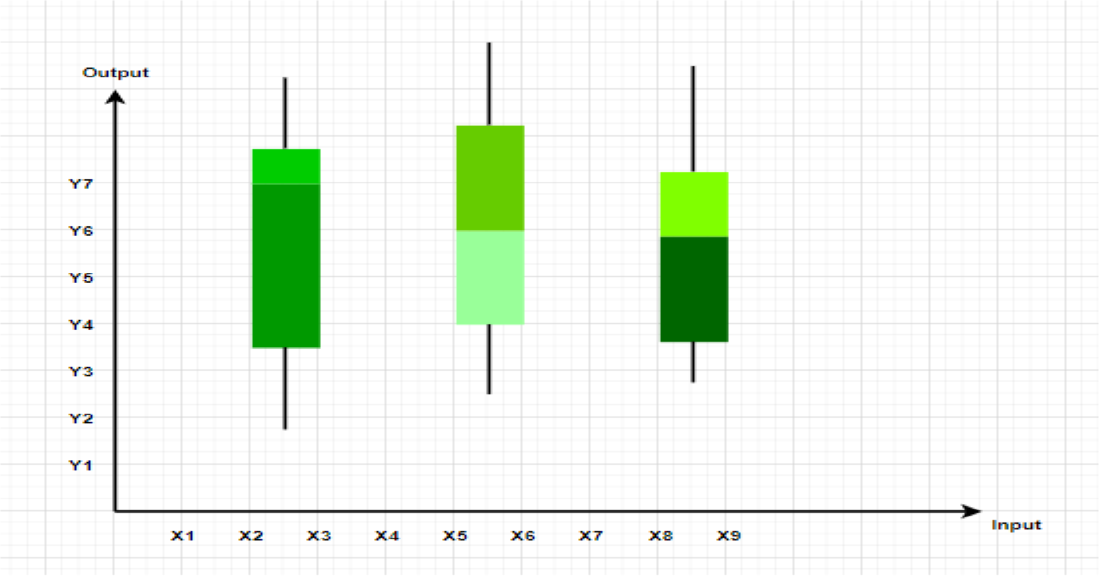
“Histos” means pole or mast, and “gram” means chart, so a histogram is a chart of poles. Plotting histograms is a graphical method for summarizing the distribution of a given attribute,

This is similar to bar graphs, but it is based frequency of numerical values rather than their actual values. The data is organized into intervals and the bars represent the frequency of the values in that range. That is, it counts how many values of the data lie in a particular range.



Box and Whisker Plot

- ✓ These plots divide the data into four parts to show their summary. They are more concerned about the spread, average, and median of the data.



quantile-quantile (q-q)plot

- ✓ The quantile-quantile plot is a graphical method for determining whether two samples of data came from the same population or not. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value.
- ✓ Q Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it.
- ✓ The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

A normal distribution, sometimes called the bell curve, is a distribution that occurs naturally in many situations. For example, the bell curve is seen in tests like the SAT and GRE. The bulk of students will score the average (C), while smaller numbers of students will score a B or D. An even smaller percentage of students score an F or an A. This creates a distribution that resembles a bell (hence the nickname). The bell curve is symmetrical. Half of the data will fall to the left of the mean; half will fall to the right.

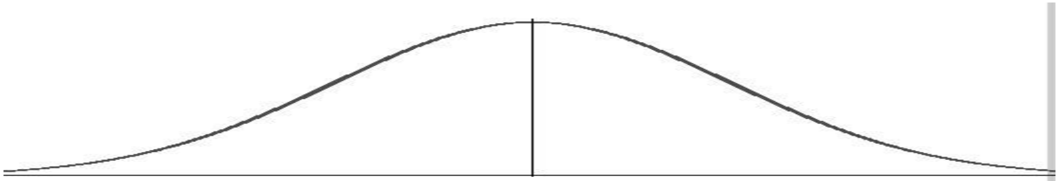


Fig. The bell curve is symmetrical