Foundation of Data Science (IT305B)

Prepared by

Mr. Umesh B. Sangule
Assistant Professor

Department of Information Technology

Unit-III

LINEAR ALGEBRA

Course Objectives : *To apply the Linear Algebra for data science.*

Course Outcome(CO3) : *Apply the fundamentals of Linear Algebra on data***.**

Content

- Data measurements scale,
- Measures of central tendency,
- Measures of variation,
- Measures of shape

Data Types and Data Measurement Scale

“ Data have an important story to tell.

They rely on you to give them a voice ”

- ➢ Before you give them a voice, you have to understand the different data types.
- ➢ There are different ways to categorize data based on the way it has been collected or its structure.

Data Types and Data Measurement Scale

- There are different ways to categorize data based on the way it has been collected or its structure,

- Based on Structure. Another important way to classify data is based on their structure. It can be categorized into two types.

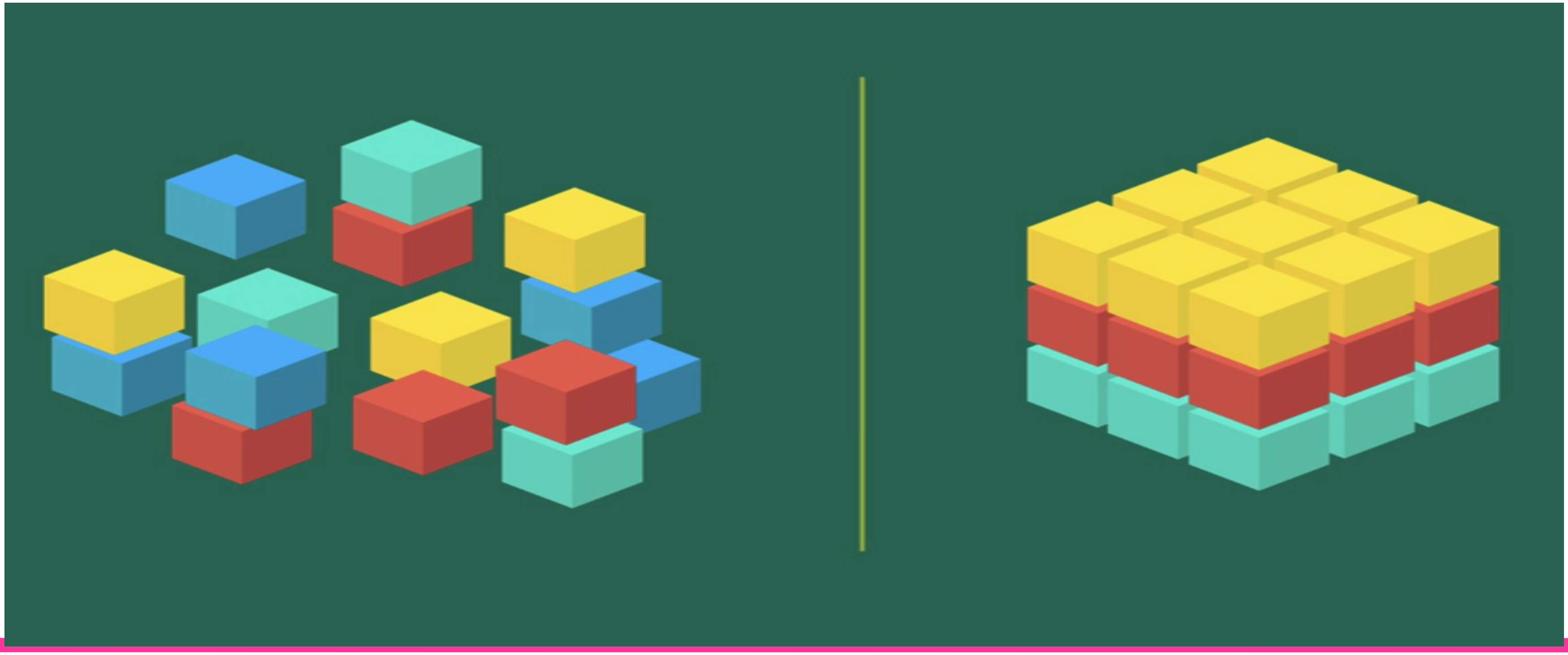
1) Structured Data: All the data points which have a specific structure and can be arranged in tabular form (also known as a matrix) with rows and columns are called structured data.

Ex: Salary of employees arranged with employee id.

Data Types and Data Measurement Scale

2) Unstructured Data: All the data points which are not arranged into any tabular format are unstructured data.

Ex: Emails, videos, clickstream data, etc.



Data Types and Data Measurement Scale

- ➢ Based on Data Collection: Data can be categorized into three types based on how data has been collected.

  1) ***Cross-Sectional Data***
  2) ***Time-Series Data***
  3) ***Panel Data***

- ➢ ***Cross-Sectional Data***: Any data points/values captured on multiple variables over one specific time period is termed as ***cross-sectional data.***

  Ex: attributes of the employee such as age, salary, level, team for the year 2019.

Data Types and Data Measurement Scale

- ***Time-Series Data***: Any data points/values captured on a single variable over multiple periods is called ***time-series data***.

  Ex: sales of smartphones on a monthly, quarterly, yearly basis.

- ***Panel Data:*** A combination of both the cross-sectional and time-series data is known as ***Panel data.***

  Ex: GDP of the various country over different periods

Data Measurement Scale:

- Scales of measurement in research and statistics are the different ways in which variables are defined and grouped into different categories,

- It describes the nature of the values assigned to the variables in a data set,

- ***Measurement*** is the process of recording observations collected as part of the research.

- Scaling is the assignment of objects to numbers or semantics.

Data Types and Data Measurement Scale

Data Measurement Scale:

- A measurement scale is used to qualify or quantify data variables,

- The properties evaluated are ***identity, magnitude, equal intervals*** and a ***minimum value of zero***

- ***Identity:*** Identity refers to each value having a unique meaning.

- ***Magnitude:*** Magnitude means that the values have an ordered relationship to one another, so there is a specific order to the variables.

Data Measurement Scale:

- ***Equal intervals***: Equal intervals mean that data points along the scale are equal, so the difference between data points one and two will be the same as the difference between data points five and six.

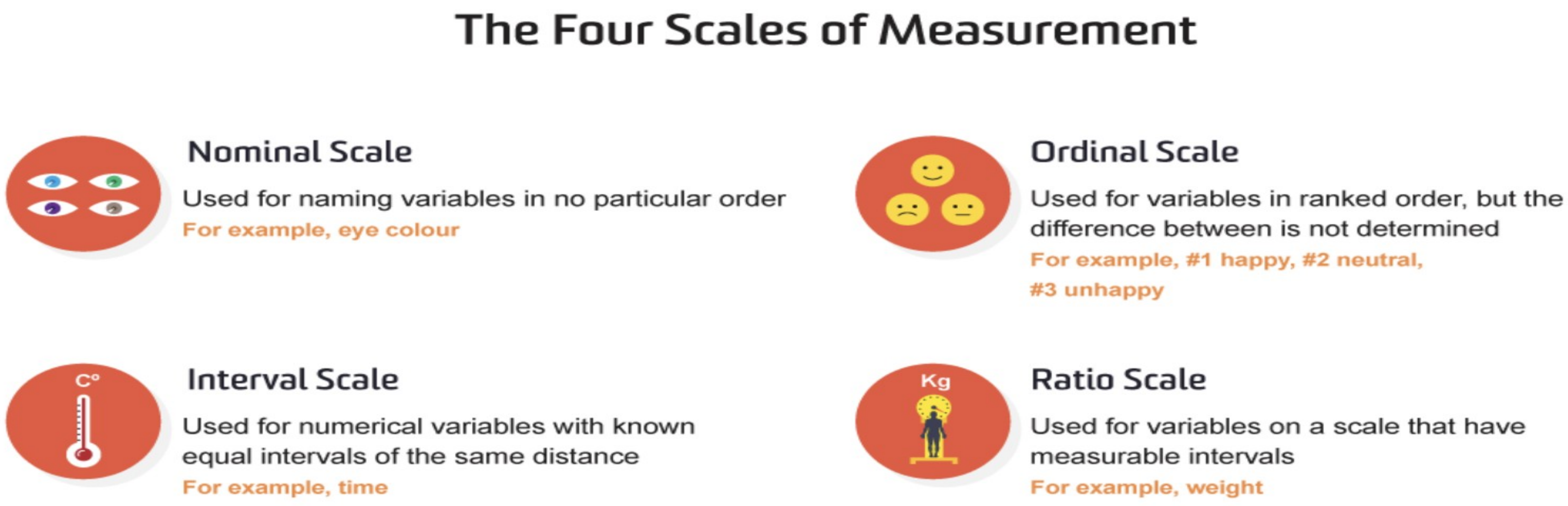
- ***A minimum value of zero***: A minimum value of zero means the scale has a true zero point.

  Degrees, for example, can fall below zero and still have meaning. But if you weigh nothing, you don't exist.

Data Types and Data Measurement Scale

Data Measurement Scale:

- ➢Data can be divided into four parts based on a measurement scale-



The Four Scales of Measurement

Nominal Scale

Used for naming variables in no particular order

For example, eye colour

Ordinal Scale

Used for variables in ranked order, but the difference between is not determined

For example, #1 happy, #2 neutral, #3 unhappy

Interval Scale

Used for numerical variables with known equal intervals of the same distance

For example, time

Ratio Scale

Used for variables on a scale that have measurable intervals

For example, weight

Data Measurement Scale:

1) Nominal Scale:

- The nominal scale is a scale of measurement that is used for identification purposes.

- It is also known as categorical scale, it assigns numbers to attributes for easy identity.

- These numbers are however not qualitative in nature and only act as labels.

Data Measurement Scale:

1) Nominal Scale:

- The only statistical analysis that can be performed on a nominal scale is the percentage or frequency count.
- It can be analyzed graphically using a bar chart and pie chart.
- Basic mathematical operations are meaningless on Nominal scale (e.g. subtraction: married -unmarried or ratio: married/unmarried)

Data Measurement Scale:

1) Nominal Scale:

- In the example below, the measurement of the popularity of a political party is measured on a nominal scale.

  Which political party are you affiliated with?

  Independent

  Republican

  Democrat

- Labeling Independent as “1”, Republican as “2” and Democrat as “3” does not in any way mean any of the attributes are better than the other. They are just used as an identity for easy data analysis.

Data Types and Data Measurement Scale

Data Measurement Scale:

2) Ordinal Scale:

- Ordinal Scale involves the ranking or ordering of the attributes depending on the variable being scaled.

- The items in this scale are classified according to the degree of occurrence of the variable in question.

- The attributes on an ordinal scale are usually arranged in ascending or descending order. It measures the degree of occurrence of the variable.

Data Measurement Scale:

2) Ordinal Scale:

- Ordinal scale can be used in market research, advertising, and customer satisfaction surveys.

- It uses qualifiers like very, highly, more, less, etc. to depict a degree.

- We can perform statistical analysis like median and mode using the ordinal scale, but not mean.

Data Measurement Scale:

2) Ordinal Scale:

- For example: A software company may need to ask its users:

- How would you rate our app?

  Excellent

  Very Good

  Good

  Bad

  Poor

- The attributes in this example are listed in descending order.

Data Measurement Scale:

3) Interval Scale:

- The interval scale of data measurement is a scale in which the levels are ordered and each numerically equal distances on the scale have equal interval difference.

- If it is an extension of the ordinal scale, with the main difference being the existence of equal intervals.

Data Measurement Scale:

3) Interval Scale:

- With an interval scale, you not only know that a given attribute A is bigger than another attribute B, but also the extent at which A is larger than B.

- Also, unlike ordinal and nominal scale, arithmetic operations can be performed on an interval scale.

- It is used in various sectors like in education, medicine, engineering, etc. Some of these uses include calculating a student's CGPA, measuring a patient's temperature, etc.

Data Measurement Scale:

3) Interval Scale:

- Example: A common example is measuring temperature on the Fahrenheit scale. It can be used in calculating mean, median, mode, range, and standard deviation.

- Example : Temperature (in centigrade), IQ level.

In such variables, addition or subtraction can be performed but division doesn’t make sense. As you can say Mumbai has 10 centigrade more than Bangalore, but you saying that Mumbai is twice hotter than Bangalore is not right, thus ratios don’t make sense here.

Data Measurement Scale:

4) Ratio Scale:

- Ratio Scale is the peak level of data measurement. It is an extension of the interval scale, therefore satisfying the four characteristics of the measurement scale; identity, magnitude, equal interval, and the absolute zero property.

- This level of data measurement allows the researcher to compare both the differences and the relative magnitude of numbers. Some examples of ratio scales include length, weight, time, etc.
- All the data points which are quantitative in nature falls in this category

Data Types and Data Measurement Scale

Data Measurement Scale:

4) Ratio Scale:

- With respect to market research, the common ratio scale examples are price, number of customers, competitors, etc. It is extensively used in marketing, advertising, and business sales.

- The ratio scale of data measurement is compatible with all statistical analysis methods like the measures of central tendency (mean, median, mode, etc.) and measures of dispersion (range, standard deviation, etc.).

Data Measurement Scale:

4) Ratio Scale:

- For example: A survey that collects the weights of the respondents.

Which of the following category do you fall in? Weight

more than 100 kgs

81 – 100 kgs

61 – 80 kgs

40 – 60 kgs

Less than 40 kgs

Measures of central tendency

Measures of central tendency:

- A measure of central tendency (also referred to as measures of centre or central location) is a summary measure that attempts to describe a whole set of data with a single value that represents the middle or centre of its distribution.

- Therefore, a measure of central tendency is a way to summarize a large set of numbers using one single score,

- There are three main measures of central tendency:

1. Mean 2.Median 3. Mode

Measures of central tendency:

- ***Mean:*** *The mean is the sum of the value of each observation in a dataset divided by the number of observations. This is also known as the arithmetic average.*

$$m = \frac{\text{sum of the terms}}{\text{number of terms}}$$

- **Advantage of the mean:**

The mean can be used for both continuous and discrete numeric data.

Measures of central tendency:

- **Limitations of the mean:**
- The mean cannot be calculated for categorical data, as the values cannot be summed,

- As the mean includes every value in the distribution the mean is influenced by outliers,

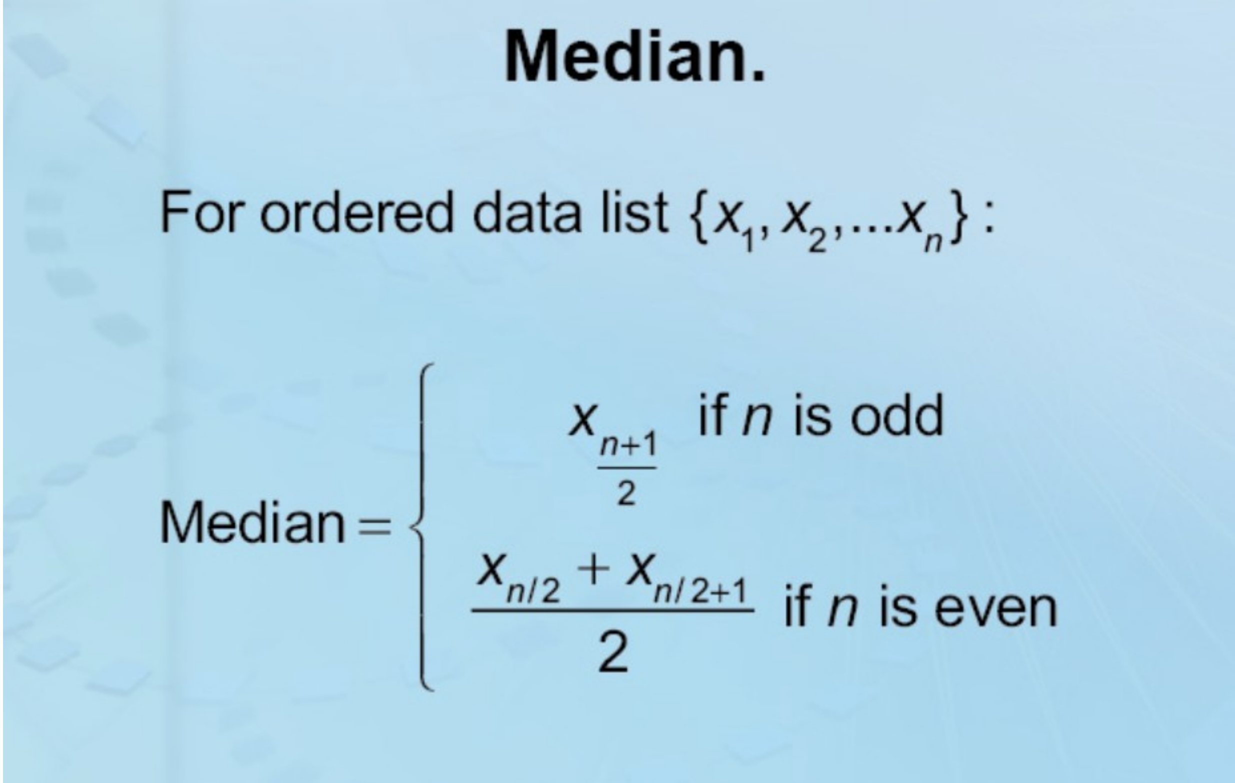
- Example: 54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60

  The mean is (623/11)= 56.6

Measures of central tendency

Measures of central tendency:

- ***Median:*** *The median is the middle value in distribution when the values are arranged in ascending or descending order.*



Median.

For ordered data list $\{x_1, x_2, \ldots x_n\}$:

$$\text{Median} = \begin{cases} x_{\frac{n+1}{2}} & \text{if } n \text{ is odd} \\ \dfrac{x_{n/2} + x_{n/2+1}}{2} & \text{if } n \text{ is even} \end{cases}$$

Measures of central tendency:

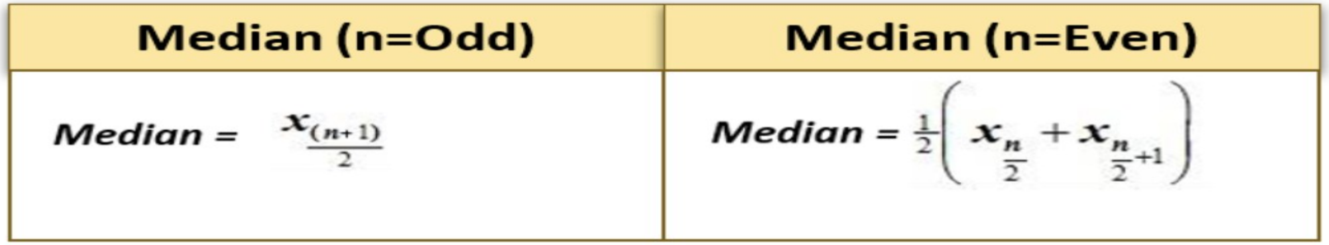
- ***Advantage of the median:***

The median is less affected by outliers and skewed data than the mean and is usually the preferred measure of central tendency when the distribution is not symmetrical.

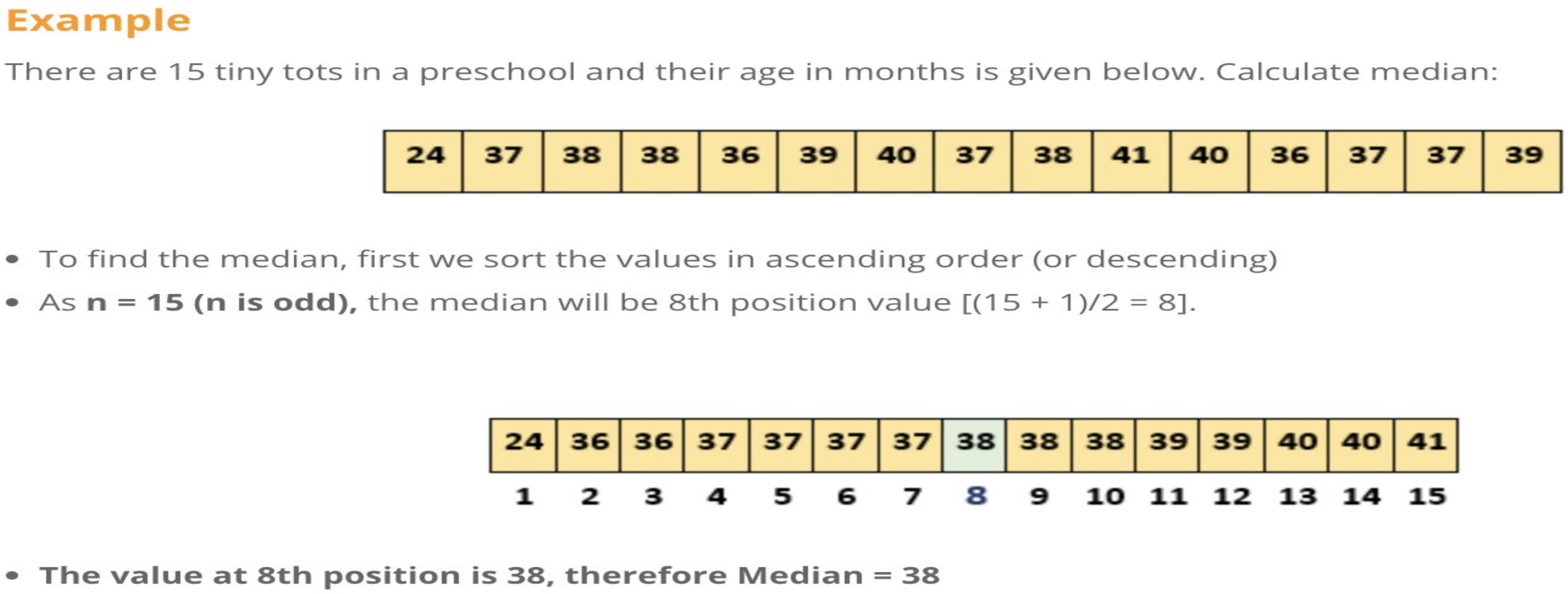
- ***Limitation of the median:***

The median cannot be identified for categorical nominal data, as it cannot be logically ordered.

Measures of central tendency:



Median (n=Odd)	Median (n=Even)
Median = $x_{\frac{(n+1)}{2}}$	*Median =* $\frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right)$



Example

There are 15 tiny tots in a preschool and their age in months is given below. Calculate median:

24	37	38	38	36	39	40	37	38	41	40	36	37	37	39

- To find the median, first we sort the values in ascending order (or descending)
- As **n = 15 (n is odd),** the median will be 8th position value [(15 + 1)/2 = 8].

24	36	36	37	37	37	37	38	38	38	39	39	40	40	41
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

- **The value at 8th position is 38, therefore Median = 38**

Measures of central tendency:

- ***Mode:*** *The mode is the most commonly occurring value in a distribution.*

- Consider this dataset showing the retirement age of 11 people, in whole years:

  54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60

- This list shows a simple frequency distribution of the retirement age data.

  54 – 3, 55 – 1, 56 – 1, 57 – 2, 58 – 2, 60 – 2

- The most commonly occurring value is 54, therefore the mode of this distribution is 54 years.

Measures of central tendency:

- ***Advantage of the mode:***

The mode has an advantage over the median and the mean as it can be found for both numerical and categorical (non-numerical) data.

- ***Limitations of Mode:***

It is also possible for there to be more than one mode for the same distribution of data, (bi-modal, or multi-modal). The presence of more than one mode can limit the ability of the mode in describing the centre or typical value of the distribution because a single value to describe the centre cannot be identified.

Measures of central tendency:

- ***Limitations of Mode:***

- In some cases, particularly where the data are continuous, the distribution may have no mode at all (i.e. if all values are different).

- In cases such as these, it may be better to consider using the median or mean or group the data into appropriate intervals and find the modal class.

Measures of central tendency

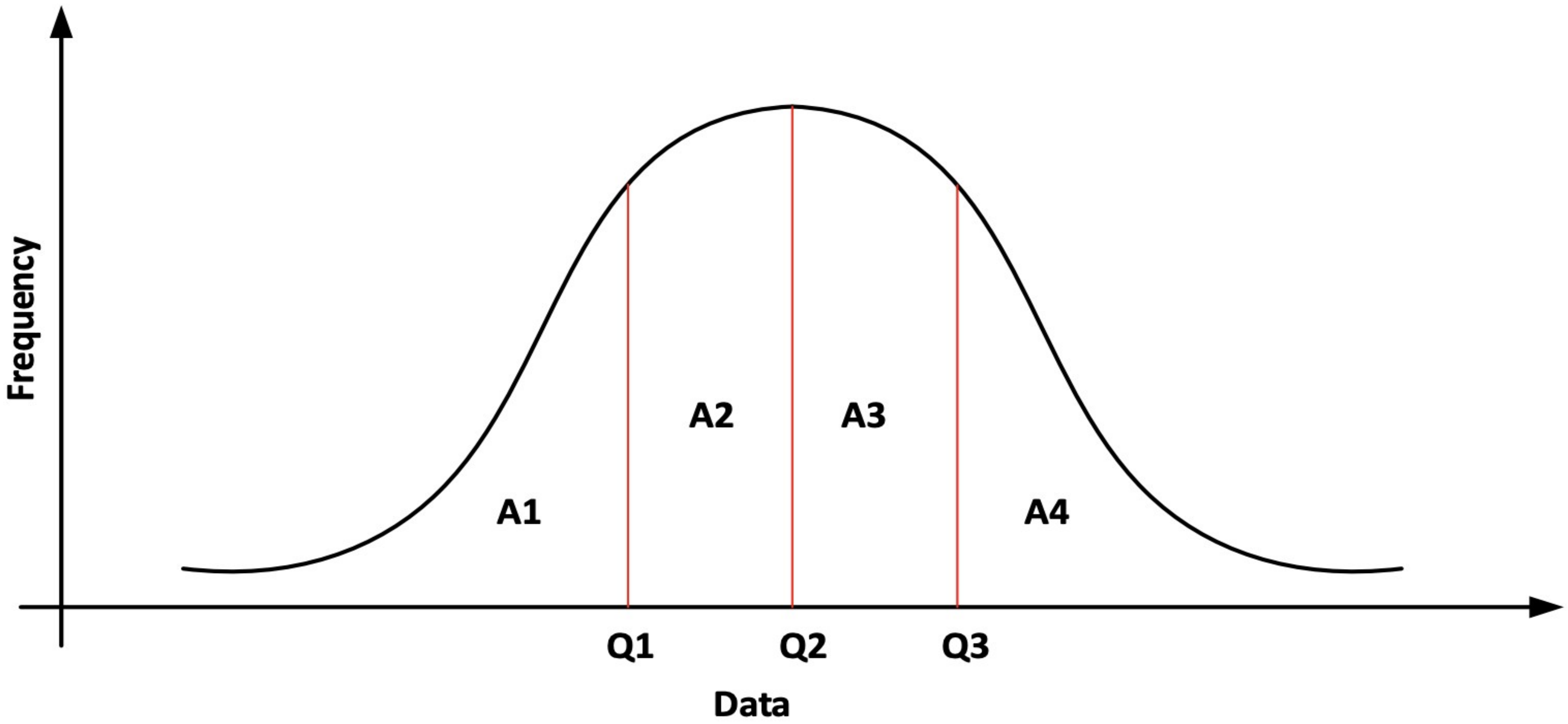
Percentile, decile, quartile:

- From the definition of median that it's the middle point in the axis frequency distribution curve, and it is divided the area under the curve for two areas have the same area in the left, and in the right.

- From this may be divided the area under the curve for four equally area and this called ***quartiles***,
- In the same procedure divided the area for ten equally pieces of area is called ***deciles***,
- Finally where divided the area for hundred equally pieces of area is called ***percentiles,***

Percentile, decile, quartile:

- ***Quartile Example:***

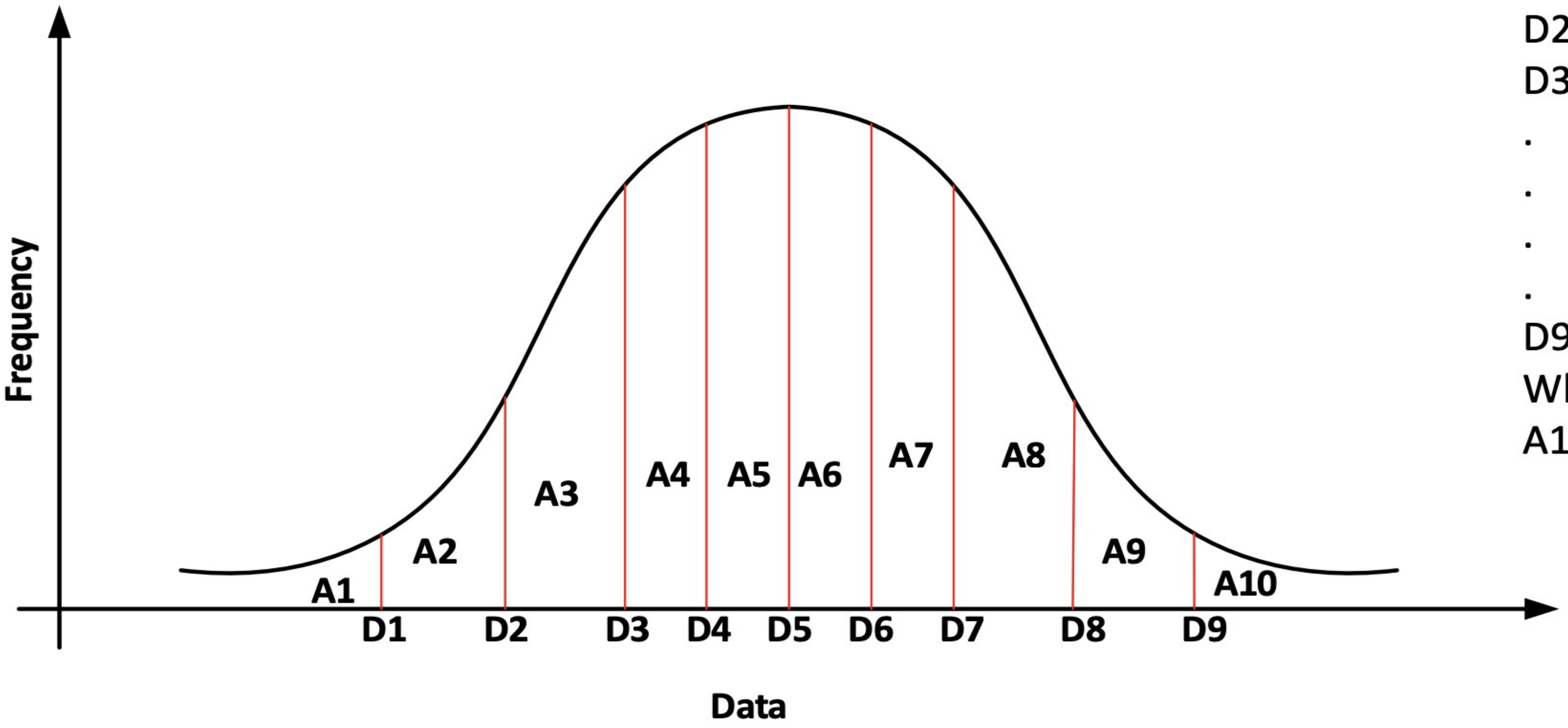


Q1 = first quartile
Q2 = second quartile
Q3 = third quartile
Where:
A1 = A2 = A3 = A4

Measures of central tendency

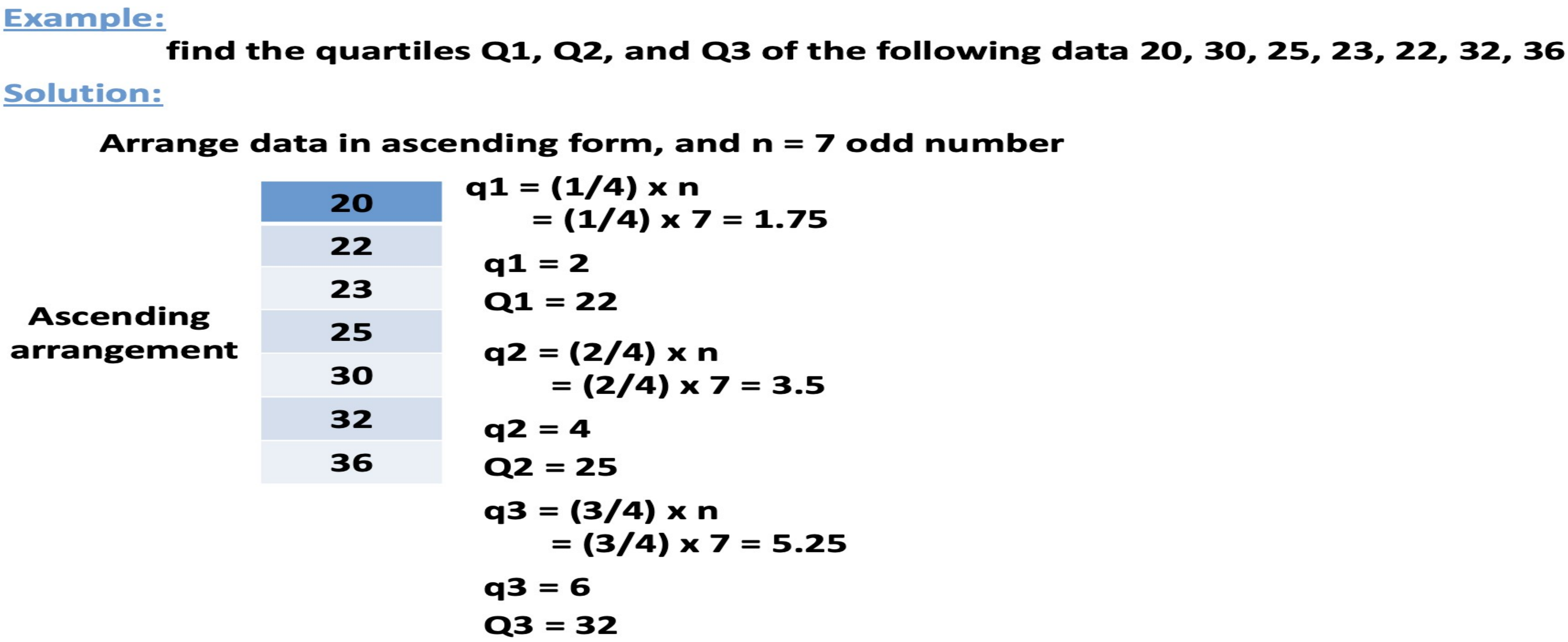
Percentile, decile, quartile:

- ***Decile Example:***



D1 = first decile
D2 = second decile
D3 = third decile
.
.
.
.
D9 = ninthe decile
Where:
A1 = A2 = A3 = A4= . . . .=A10

Percentile, decile, quartile:



Example:

find the quartiles Q1, Q2, and Q3 of the following data 20, 30, 25, 23, 22, 32, 36

Solution:

Arrange data in ascending form, and n = 7 odd number

Ascending arrangement

20
22
23
25
30
32
36

q1 = (1/4) x n
= (1/4) x 7 = 1.75

q1 = 2

Q1 = 22

q2 = (2/4) x n
= (2/4) x 7 = 3.5

q2 = 4

Q2 = 25

q3 = (3/4) x n
= (3/4) x 7 = 5.25

q3 = 6

Q3 = 32

Measures of central tendency

Percentile, decile, quartile:

Example:

find the quartiles Q1, Q2, and Q3 of the following data 20, 30, 25, 23, 22, 32, 36, 18

Solution:

Arrange data in ascending form, and n = 8 even number

Ascending arrangement

18
20
22
23
25
30
32
36

q1 = (1/4) x n
= (1/4) x 8 = 2

q1 = mean of (2), and (3)

Q1 = (20+22)x(1/2) = 21

q2 = (2/4) x n
= (2/4) x 8 = 4

q2 = mean of (4), and (5)

Q2 = (23+25)x(1/2) = 24

q3 = (3/4) x n
= (3/4) x 8 = 6

q3 = mean of (6), and (7)

Q3 = (30+32)x(1/2) = 31

Measures of variation

Variation or Dispersion:

- The degree to which numerical data tend to spread about an average value is called the ***dispersion, or variation***, of the data.

- Various measures of this dispersion (or variation) are available, the most common-
  1. Range,
  2. Interquartile Distance(IQD),
  3. Variance
  4. Standard deviation.

Measures of variation

➢ ***Range:***

- The range of a set of numbers is the difference between the largest and smallest numbers in the set.

- Example:

  The range of the set 2, 3, 3, 5, 5, 5, 8, 10, 12 is 12 - 2 = 10.

  Sometimes the range is given by simply quoting the smallest and largest numbers;

  In the above set, for instance, the range could be indicated as

  2 to 12, or 2–12.

Measures of variation

➢ ***Interquartile Distance (IQD):***

- The midpoint of data distribution, or the middle of your four quartiles, is referred to as the interquartile range (IQR), which is in the middle of the lower and upper quartiles.

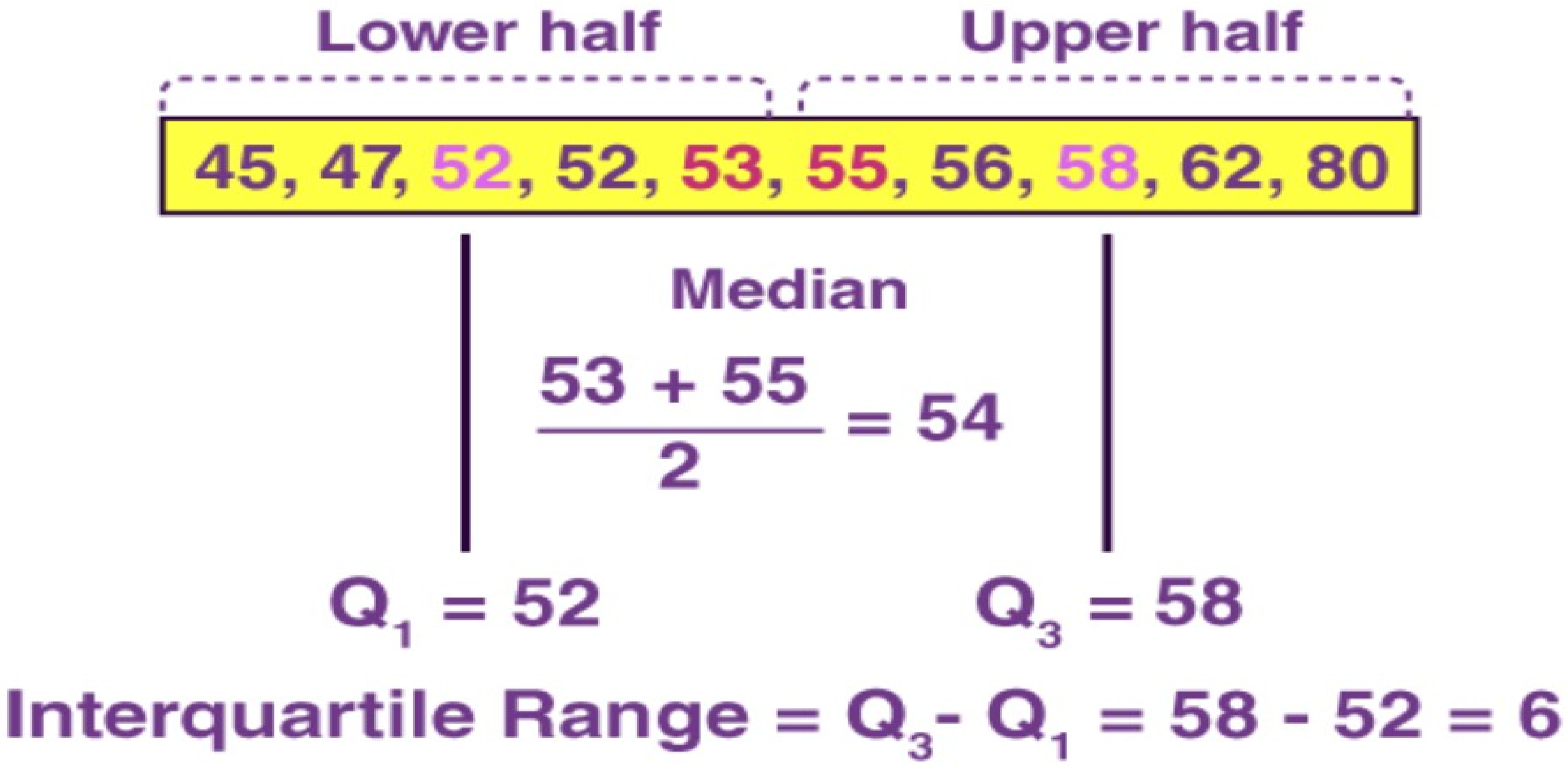
- The IQD is a measurement of how evenly the data is distributed around the average.
- The formula for Interquartile Range is given below:

  Interquartile Distance(IQD)= $Q3 - Q1$

- The IQD is a useful measure for identifying outliers in data,

Measures of variation

➢ ***<u>Interquartile Distance (IQD):</u>***



45, 47, 52, 52, 53, 55, 56, 58, 62, 80

Median

$$\frac{53 + 55}{2} = 54$$

$Q_1 = 52$ $Q_3 = 58$

Interquartile Range = $Q_3 - Q_1 = 58 - 52 = 6$

Measures of variation

- ***Variance:***

- Variance is a measure of variability in the data from mean value,

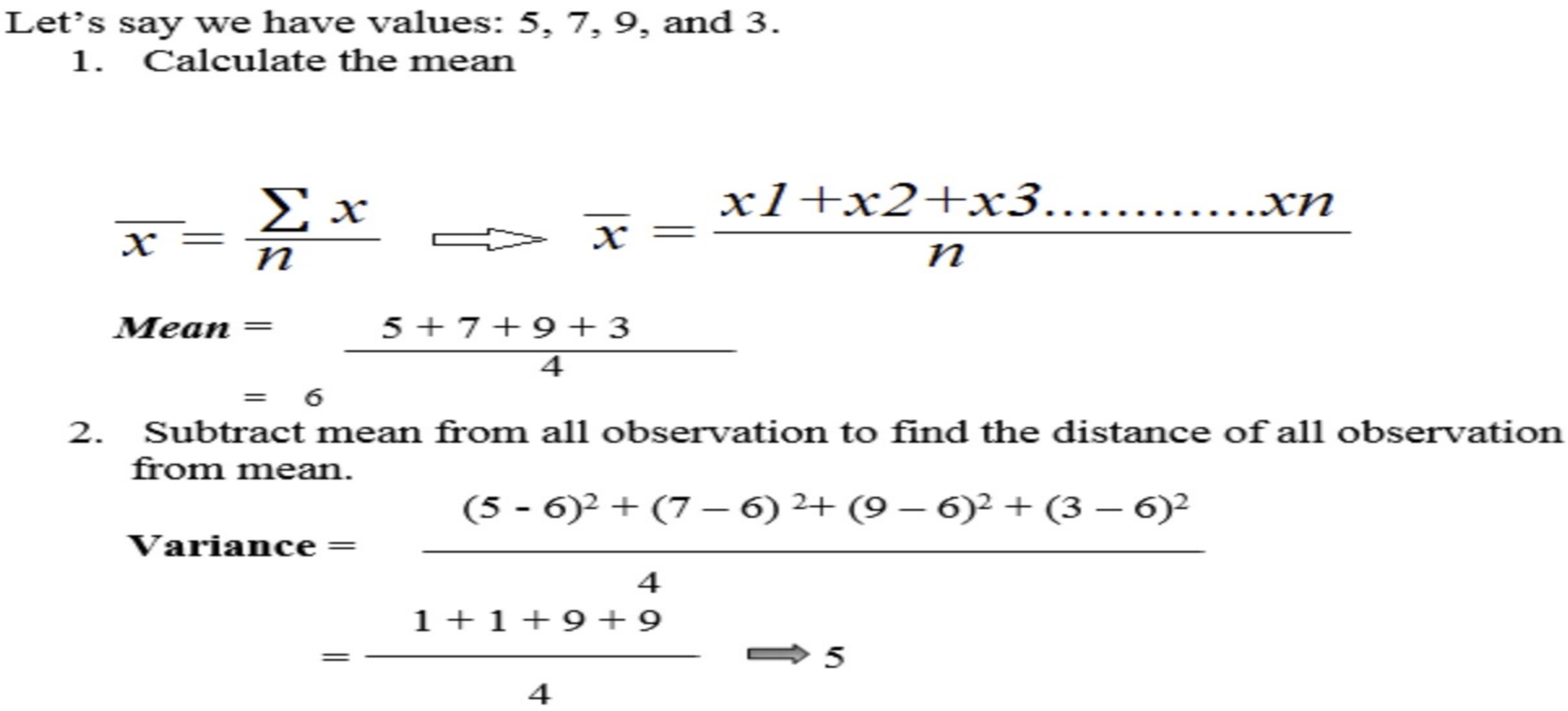
- It compares every piece of value to the mean, which is why variance differs from the other measures of variation.

- Variance also displays the spread of the data set,

- variance to compare pieces of data to one another to see how they relate,

Measures of variation

➢ Variance:



Let’s say we have values: 5, 7, 9, and 3.

1. Calculate the mean

$$\overline{x} = \frac{\sum x}{n} \quad \Rightarrow \quad \overline{x} = \frac{x1 + x2 + x3 \ldots\ldots\ldots\ldots xn}{n}$$

$$\textit{Mean} = \frac{5 + 7 + 9 + 3}{4}$$

$$= \quad 6$$

2. Subtract mean from all observation to find the distance of all observation from mean.

$$\text{Variance} = \frac{(5 - 6)^2 + (7 - 6)^2 + (9 - 6)^2 + (3 - 6)^2}{4}$$

$$= \frac{1 + 1 + 9 + 9}{4} \Rightarrow 5$$

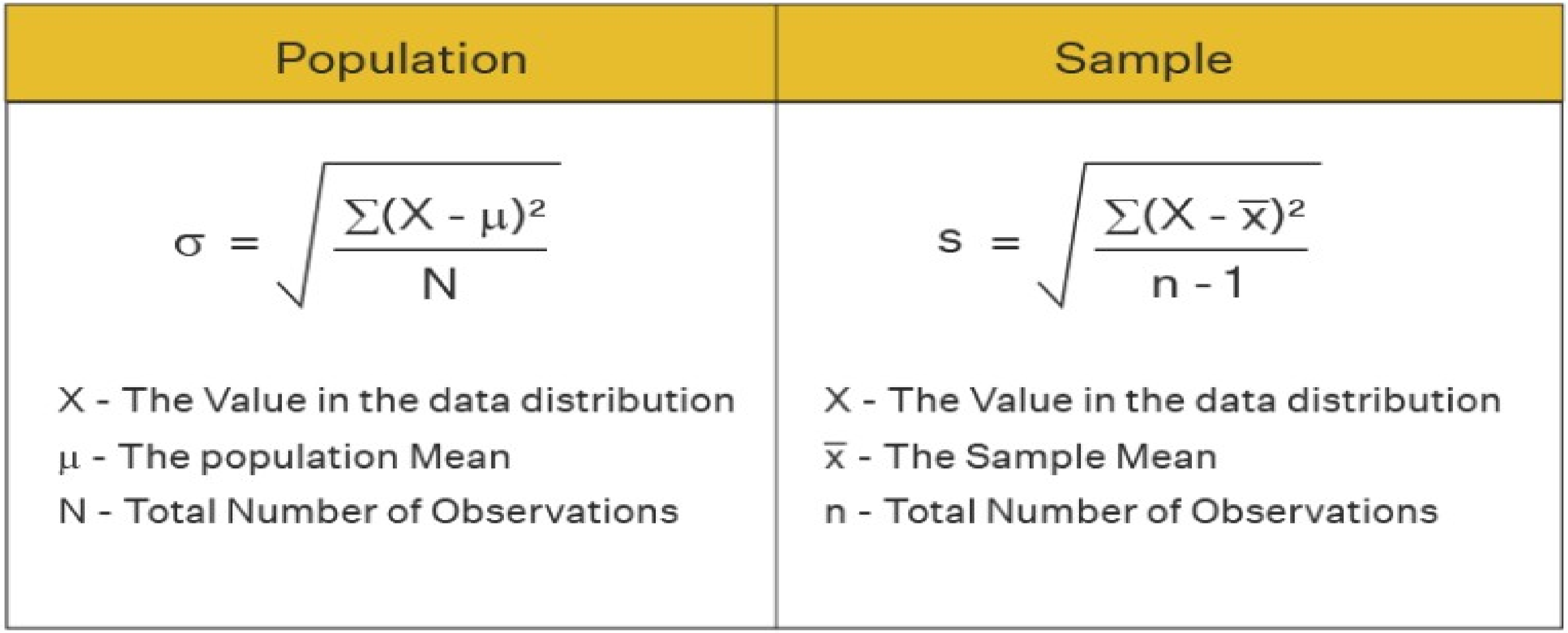
Measures of variation

➢ ***Standard Deviation:***

- ➢ Standard deviation is a squared root of the variance to get original values.
- ➢ Low standard deviation indicates data points close to mean.
- ➢ Standard deviation uses the square root of the variance to get original values.
- ➢ Standard deviation calculates the extent to which the values differ from the average.
- ➢ Standard Deviation, the most widely used measure of dispersion, is based on all values

Measures of variation

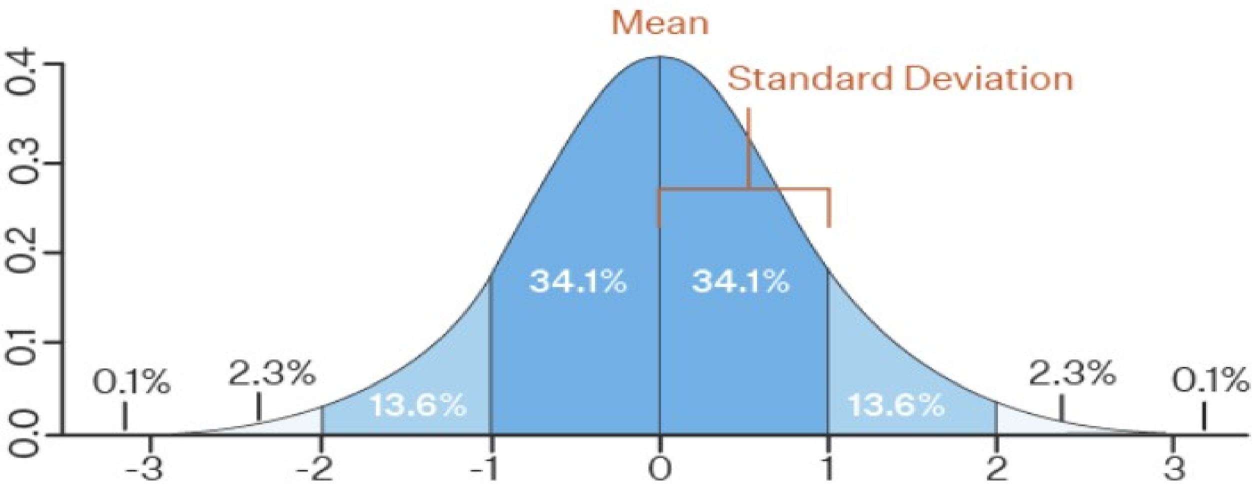
➢ ***Standard Deviation:***

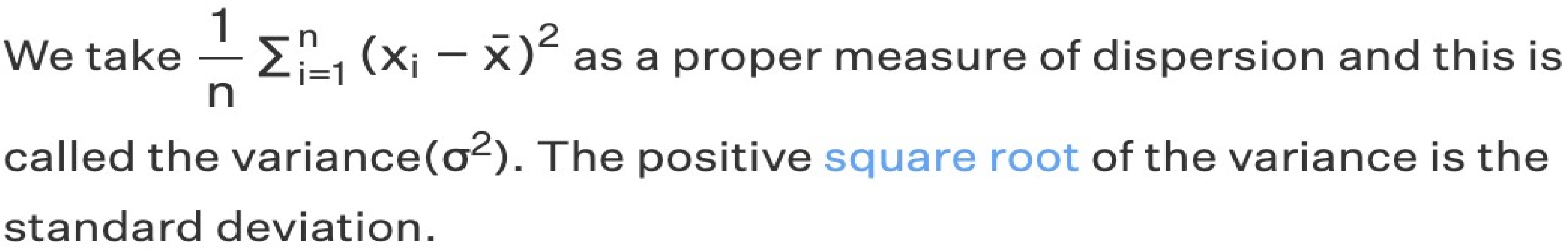


Population	Sample
$\sigma = \sqrt{\dfrac{\sum(X - \mu)^2}{N}}$	$s = \sqrt{\dfrac{\sum(X - \overline{x})^2}{n - 1}}$
X - The Value in the data distribution μ - The population Mean N - Total Number of Observations	X - The Value in the data distribution $\overline{x}$ - The Sample Mean n - Total Number of Observations

Measures of variation

➢ ***Standard Deviation:***





We take $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ as a proper measure of dispersion and this is called the variance(σ^2). The positive square root of the variance is the standard deviation.

Measures of variation

- ***Standard Deviation:***
- The procedure to calculate the standard deviation is given below:

Step 1: Compute the mean for the given data set.

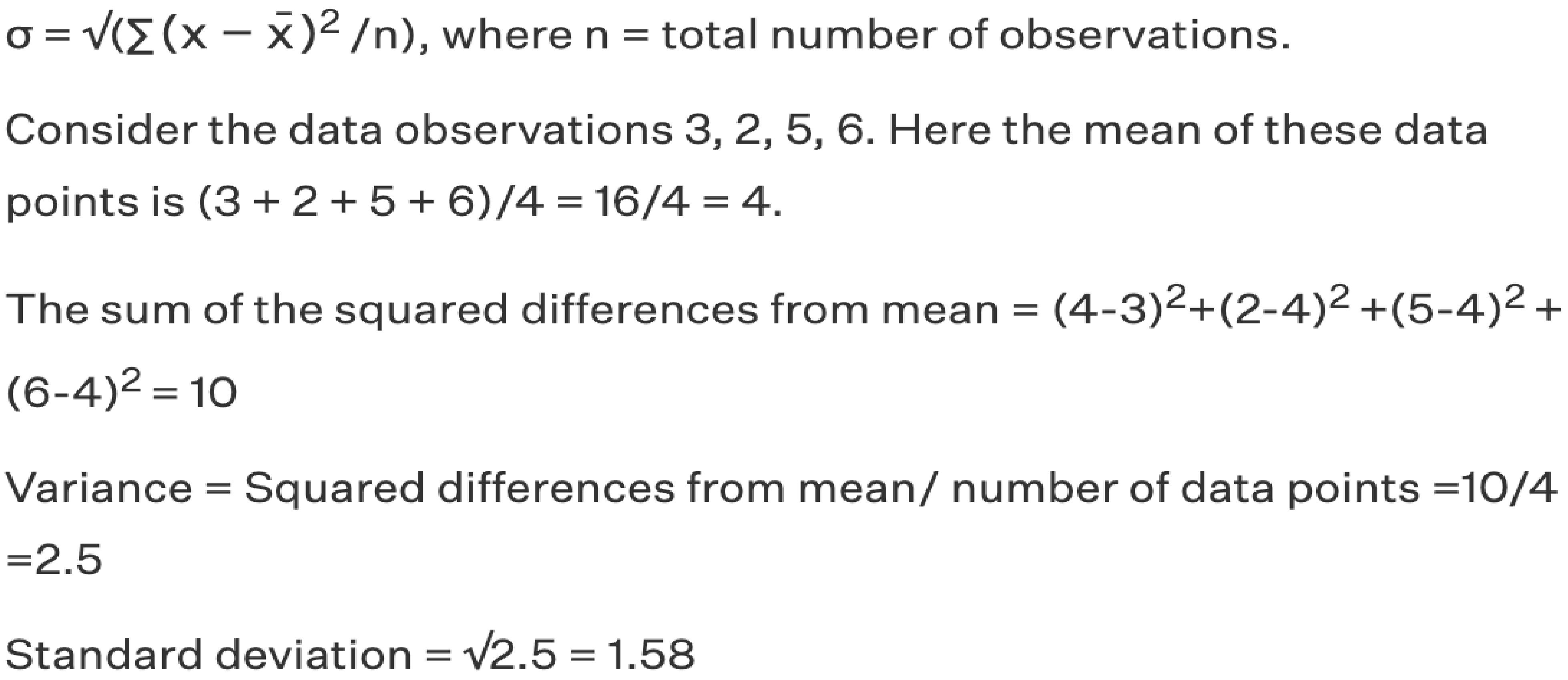
Step 2: Subtract the mean from each observation and calculate the square in each instance.

Step 3: Find the mean of those squared deviations.

Step 4: Finally, take the square root obtained mean to get the standard deviation.

Measures of variation

➢ ***Standard Deviation:***



$\sigma = \sqrt{(\sum(x - \bar{x})^2 / n)}$, where n = total number of observations.

Consider the data observations 3, 2, 5, 6. Here the mean of these data points is (3 + 2 + 5 + 6)/4 = 16/4 = 4.

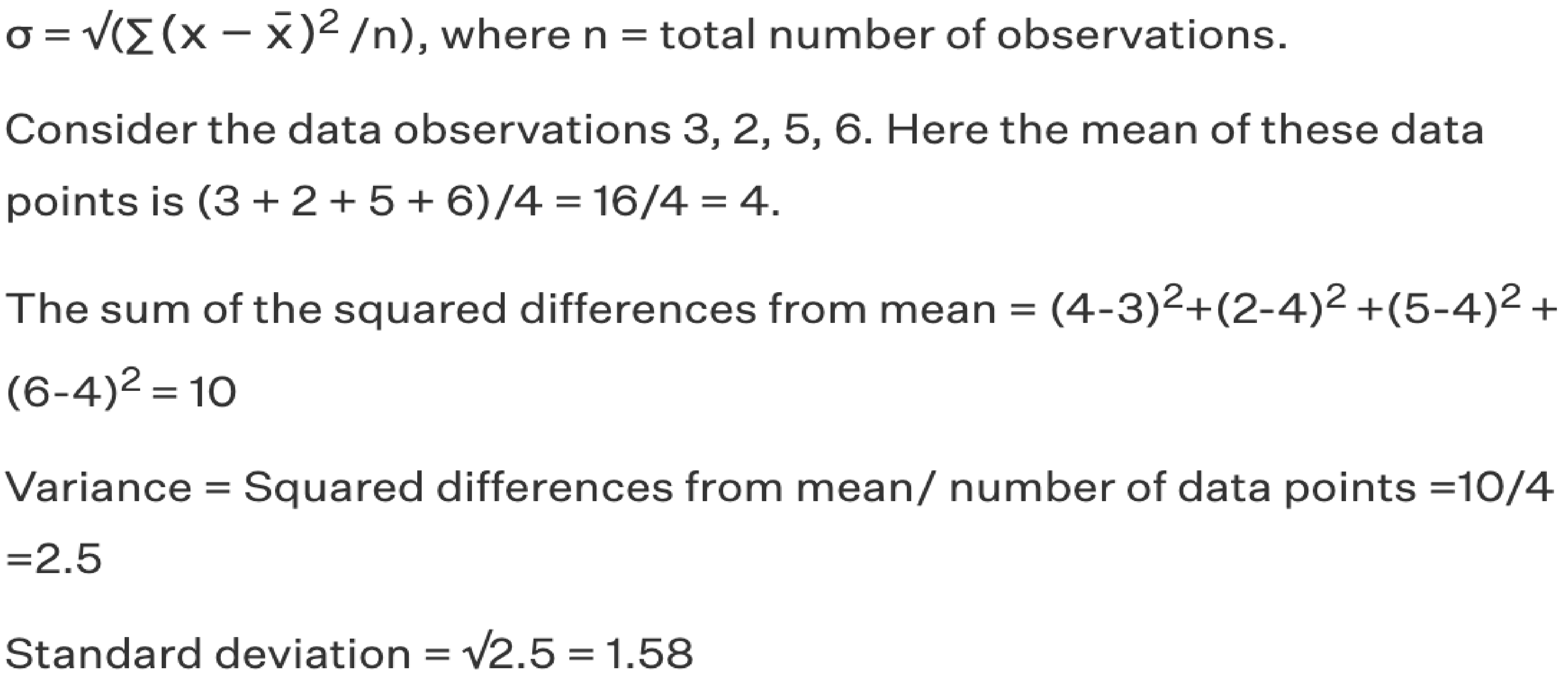
The sum of the squared differences from mean = $(4\text{-}3)^2+(2\text{-}4)^2+(5\text{-}4)^2+(6\text{-}4)^2 = 10$

Variance = Squared differences from mean/ number of data points =10/4 =2.5

Standard deviation = √2.5 = 1.58

Measures of variation

➢ ***Standard Deviation:***



$\sigma = \sqrt{(\sum (x - \bar{x})^2 / n)}$, where n = total number of observations.

Consider the data observations 3, 2, 5, 6. Here the mean of these data points is $(3 + 2 + 5 + 6)/4 = 16/4 = 4$.

The sum of the squared differences from mean = $(4-3)^2+(2-4)^2+(5-4)^2+(6-4)^2 = 10$

Variance = Squared differences from mean/ number of data points =10/4 =2.5

Standard deviation = $\sqrt{2.5} = 1.58$

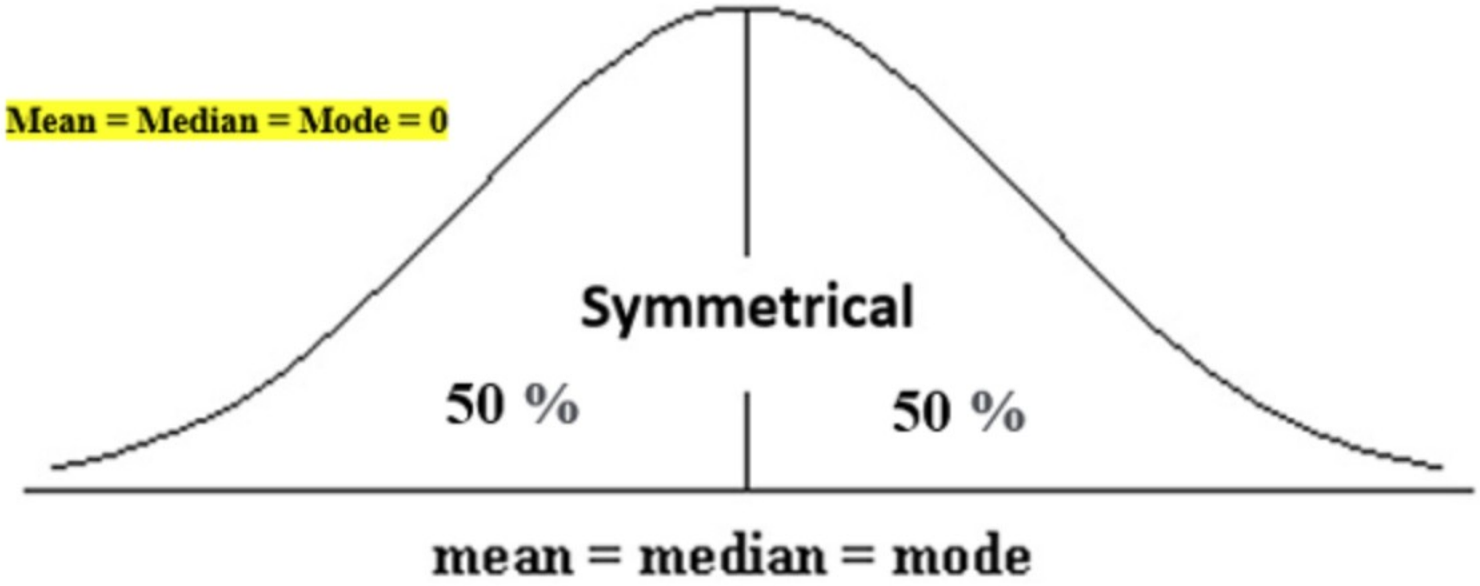
Measures of Shape

➢ ***Skewness:***

- ➢ Skewness is a statistical measure that assesses the asymmetry of a probability distribution. It quantifies the extent to which the data is skewed or shifted to one side.
- ➢ Positive skewness indicates a longer tail on the right side of the distribution, while negative skewness indicates a longer tail on the left side. Skewness helps in understanding the shape and outliers in a dataset
- ➢ If the values of a specific independent variable (feature) are skewed, depending on the model, skewness may violate model assumptions or may reduce the interpretation of feature importance.

Measures of Shape

- ***Skewness:***
- The symmetrical distribution has zero skewness as all measures of a central tendency lies in the middle.

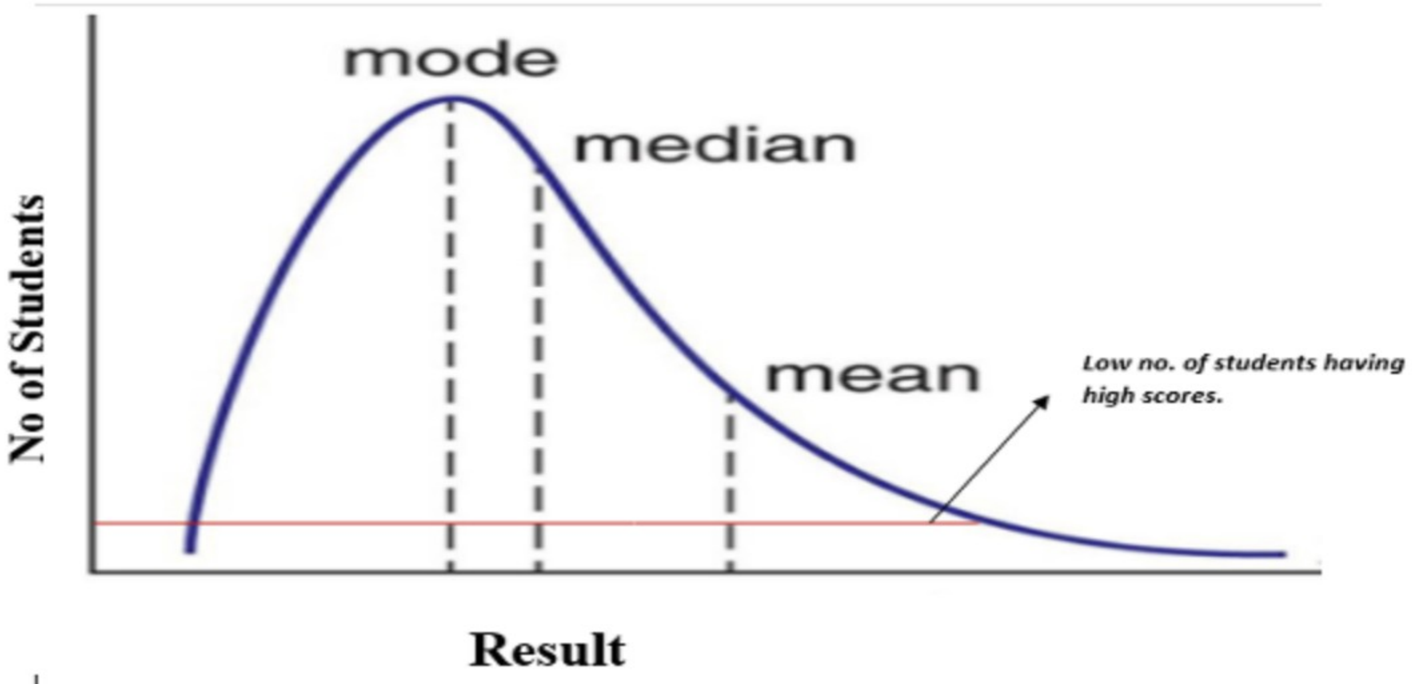


Measures of Shape

➢ ***Types of Skewness***

Positive Skewed or Right-Skewed (Positive Skewness)

In statistics, a positively skewed or right-skewed distribution has a long right tail



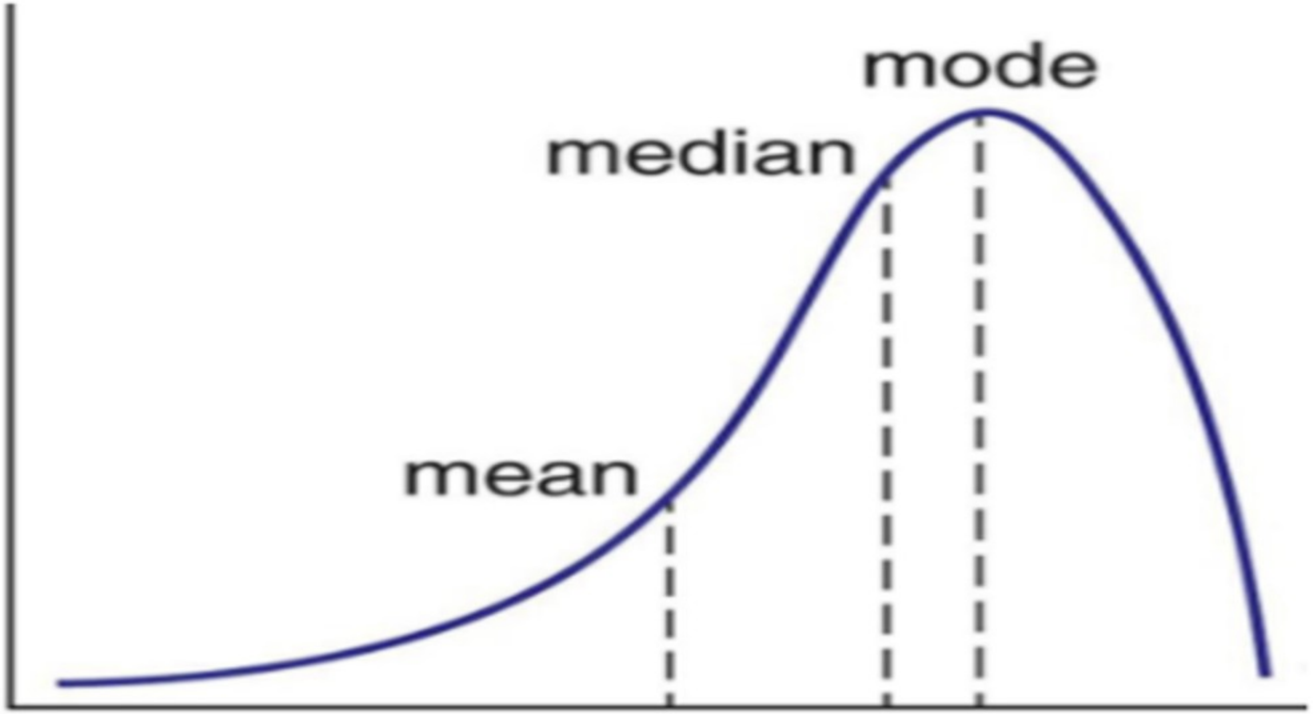
Mean > Median > Mode

Measures of Shape

➢ <u>Types of Skewness</u>

Negative Skewed or Left-Skewed (Negative Skewness)

A negatively skewed or left-skewed distribution has a long left tail;



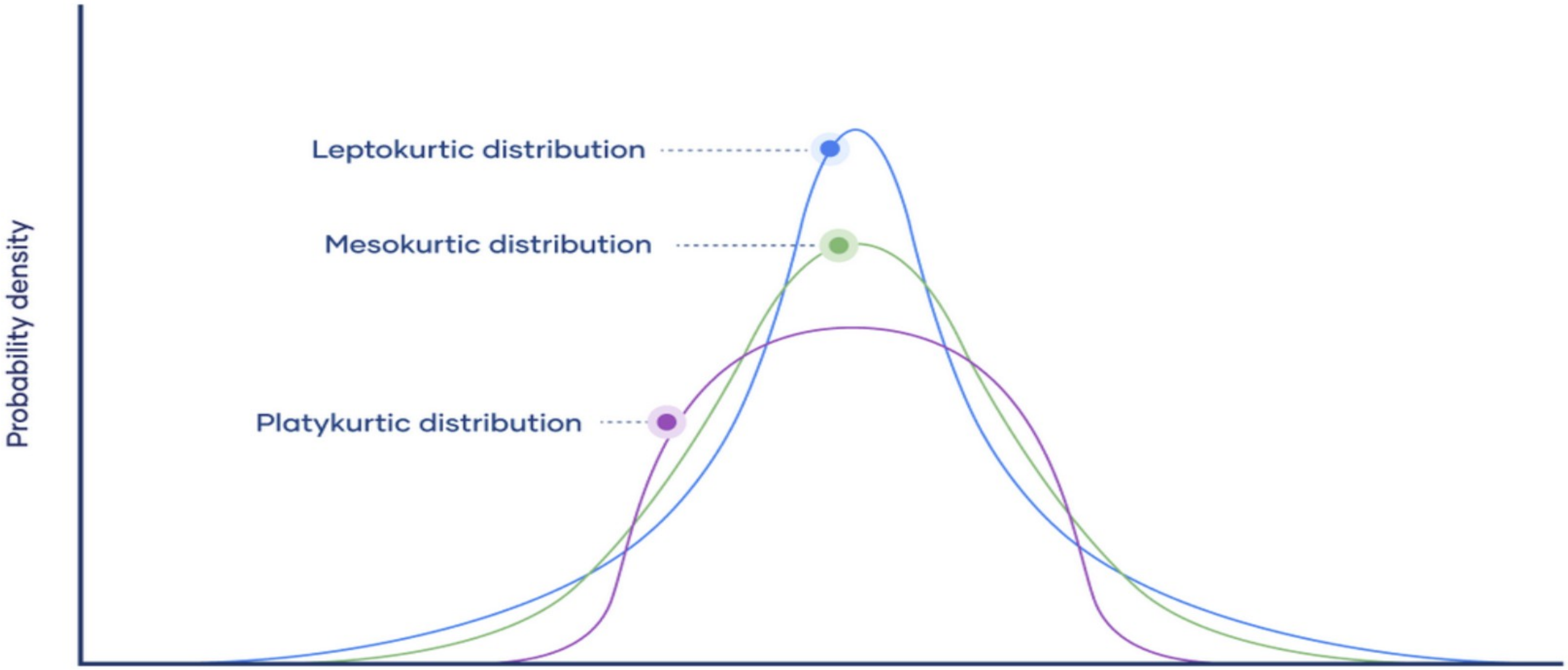
Mode > Median > mean

Measures of Shape

- ***Kurtosis:***
- The excess kurtosis is used in statistics and probability theory to compare the kurtosis coefficient with that normal distribution.
- Excess kurtosis can be positive (Leptokurtic distribution), negative (Platykurtic distribution), or near zero (Mesokurtic distribution),
- Leptokurtic or heavy-tailed distribution (kurtosis more than normal distribution).
- Mesokurtic (kurtosis same as the normal distribution).
- Platykurtic or short-tailed distribution (kurtosis less than normal distribution)

Measures of Shape

➢ <u>Kurtosis:</u>



THANK YOU