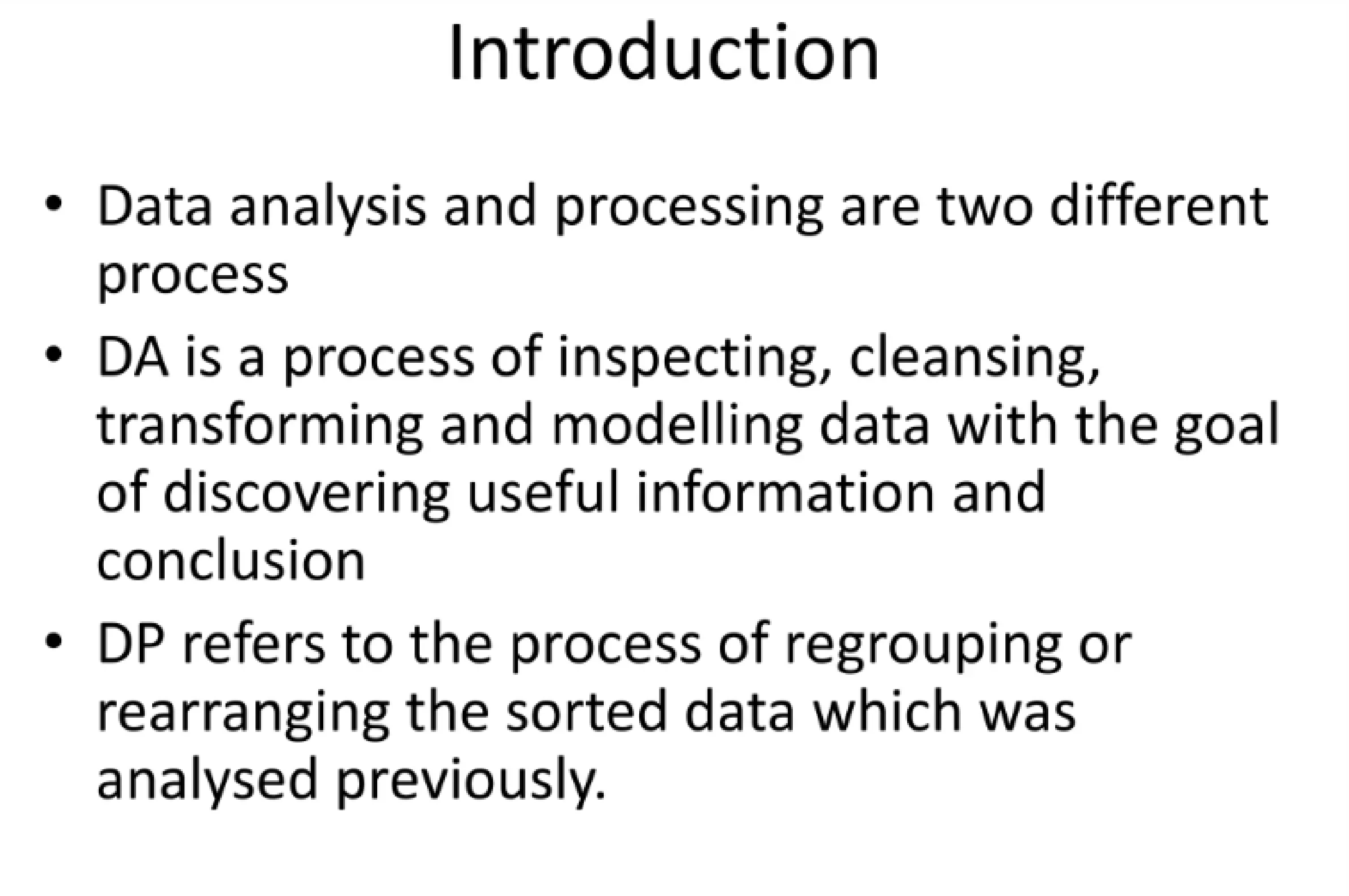
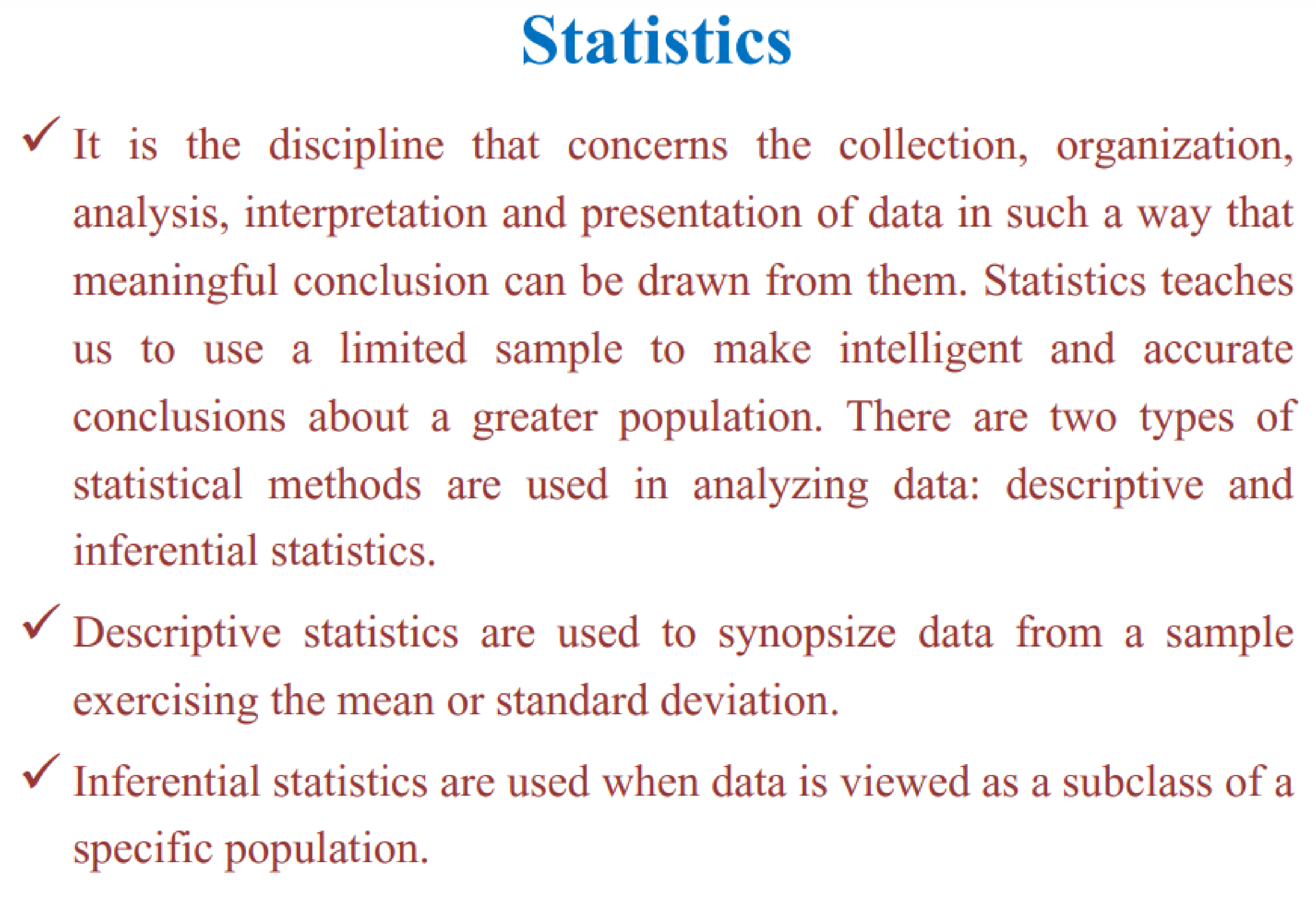


Introduction

- Data analysis and processing are two different process
- DA is a process of inspecting, cleansing, transforming and modelling data with the goal of discovering useful information and conclusion
- DP refers to the process of regrouping or rearranging the sorted data which was analysed previously.



Statistics

- ✓ It is the discipline that concerns the collection, organization, analysis, interpretation and presentation of data in such a way that meaningful conclusion can be drawn from them. Statistics teaches us to use a limited sample to make intelligent and accurate conclusions about a greater population. There are two types of statistical methods are used in analyzing data: descriptive and inferential statistics.
- ✓ Descriptive statistics are used to synopsize data from a sample exercising the mean or standard deviation.
- ✓ Inferential statistics are used when data is viewed as a subclass of a specific population.

Characteristics of statistics

✓Statistics are aggregates of facts.

✓Statistic are numerically expressed.

✓Statistic are affected to a marked extent by multiplicity of causes.

✓Statistic are enumerated or estimated according to a reasonable standard of accuracy.

Data

✓A piece of information which you collect through your study called data.

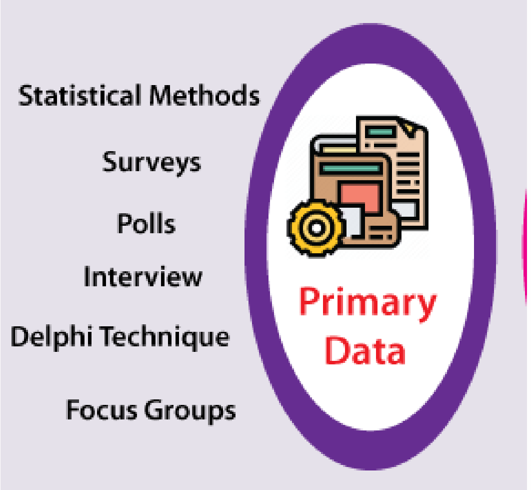
✓In other words it is a set of facts and figures which are useful in a particular purpose like a survey or an analysis. When one arrange data in an organized form, they refers to as information.

✓Data can be classified as qualitative and quantitative.

Unit –V :Data Analysis

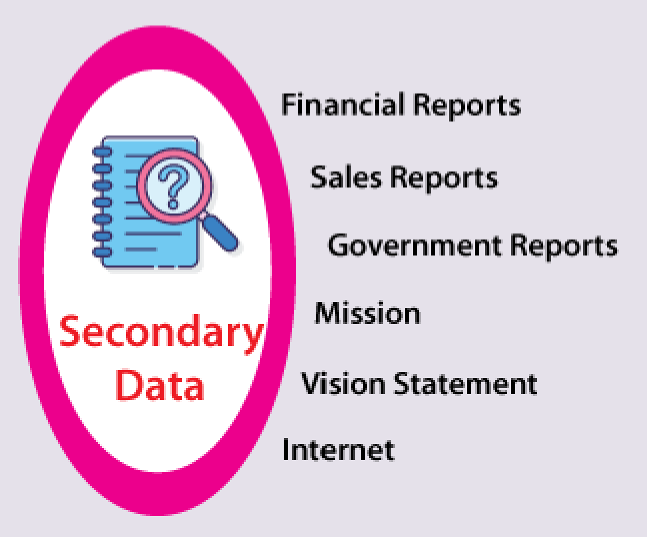
Statistical data can be classified into two categories, primary and secondary

- ***Primary data***:- primary data is one which is collected by the investigator himself for the purpose of a specific inquiry or study. Such data is original in character and is generated by surveys conducted by individuals or research institutions.



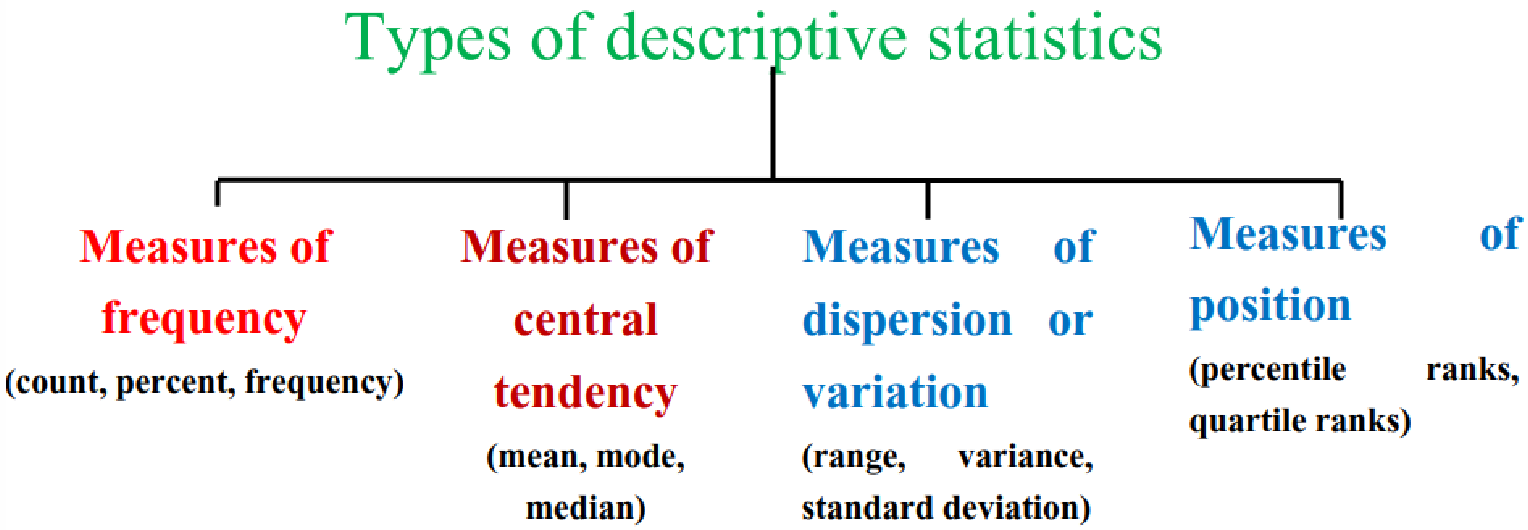
Secondary Data

- When an investigator used the data which has already been collected by others, such data is called secondary data. This data is primary data for the agency that collects it and becomes secondary data for someone else who uses this data for his own purposes.
- the secondary data can be obtained from journals, reports, government publications, publication of professional and research organizations and so on.



Descriptive statistics

- Descriptive statistics uses the data to provide descriptions of the population, entire through numerical calculations or graphs or tables.
- Descriptive statistics enables us to present the data in a more meaningful way, which allows simpler interpretation of the data.



Types of descriptive statistics

Measures of frequency
(count, percent, frequency)

Measures of central tendency
(mean, mode, median)

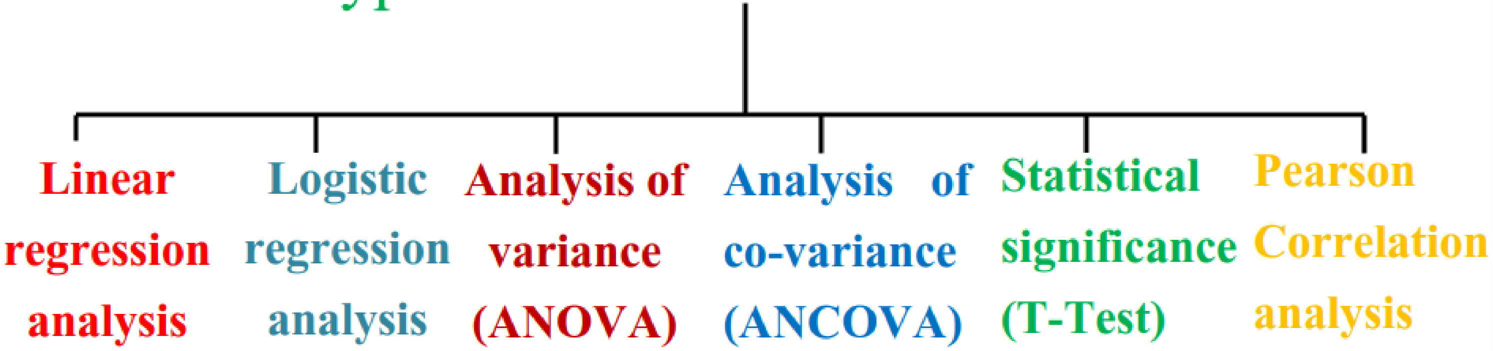
Measures of dispersion or variation
(range, variance, standard deviation)

Measures of position
(percentile ranks, quartile ranks)

Interferential statistics

- Interferential statistics makes interferences and predictions about a population based on a sample of data taken from the population in question.
- Through interferential statistics, one can take data from samples and make generalizations about a population.
- The most common methodologies in inferential statistics are hypothesis tests, confidence intervals, and regression analysis.

Types of interferential statistics



Linear regression analysis

Logistic regression analysis

Analysis of variance (ANOVA)

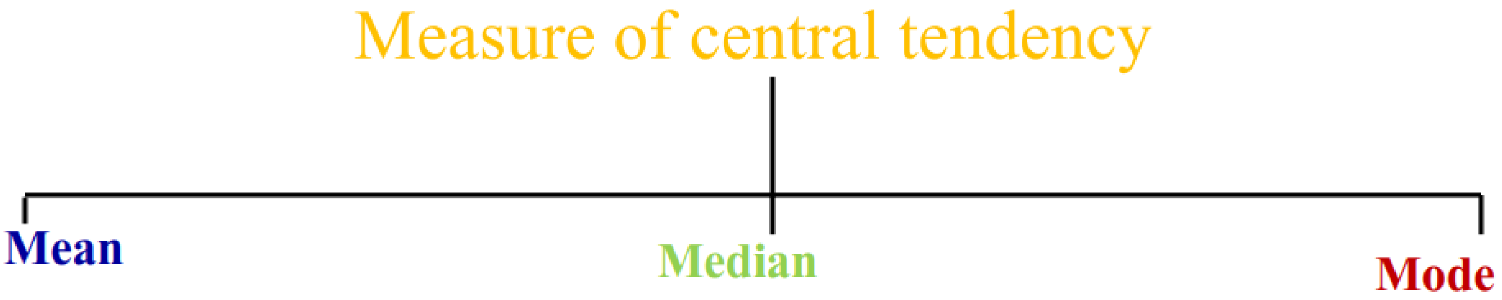
Analysis of co-variance (ANCOVA)

Statistical significance (T-Test)

Pearson Correlation analysis

Central tendency

A measure of central tendency in a single value that describes the way in which a group of data cluster around a central value. In other words, it is a way to describe the center of a data set. It may also be called a center or location of the distribution. There are three measures of central tendency: the mean, the median and the mode. Mean is the most frequently used measure of central tendency and generally considered the best measure of it. However , there are some situations where either median or mode are preferred. Median is the preferred measure of central tendency when: there are a few extreme scores in the distribution of the data.



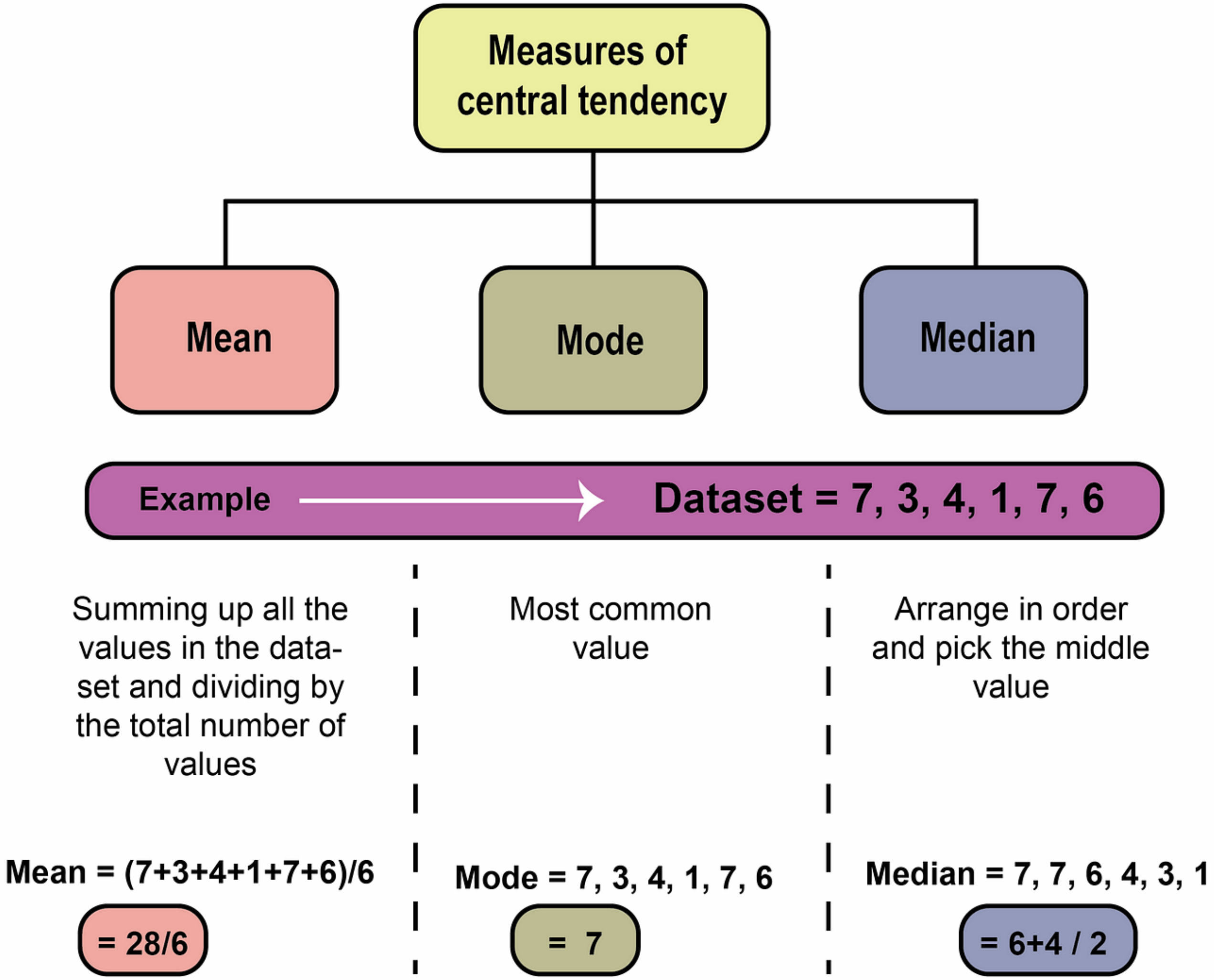
Measure of central tendency

Mean

Median

Mode

Unit –V :Data Analysis



Measures of central tendency

Mean

Mode

Median

Example → Dataset = 7, 3, 4, 1, 7, 6

Summing up all the values in the data-set and dividing by the total number of values

Mean = (7+3+4+1+7+6)/6

= 28/6

Most common value

Mode = 7, 3, 4, 1, 7, 6

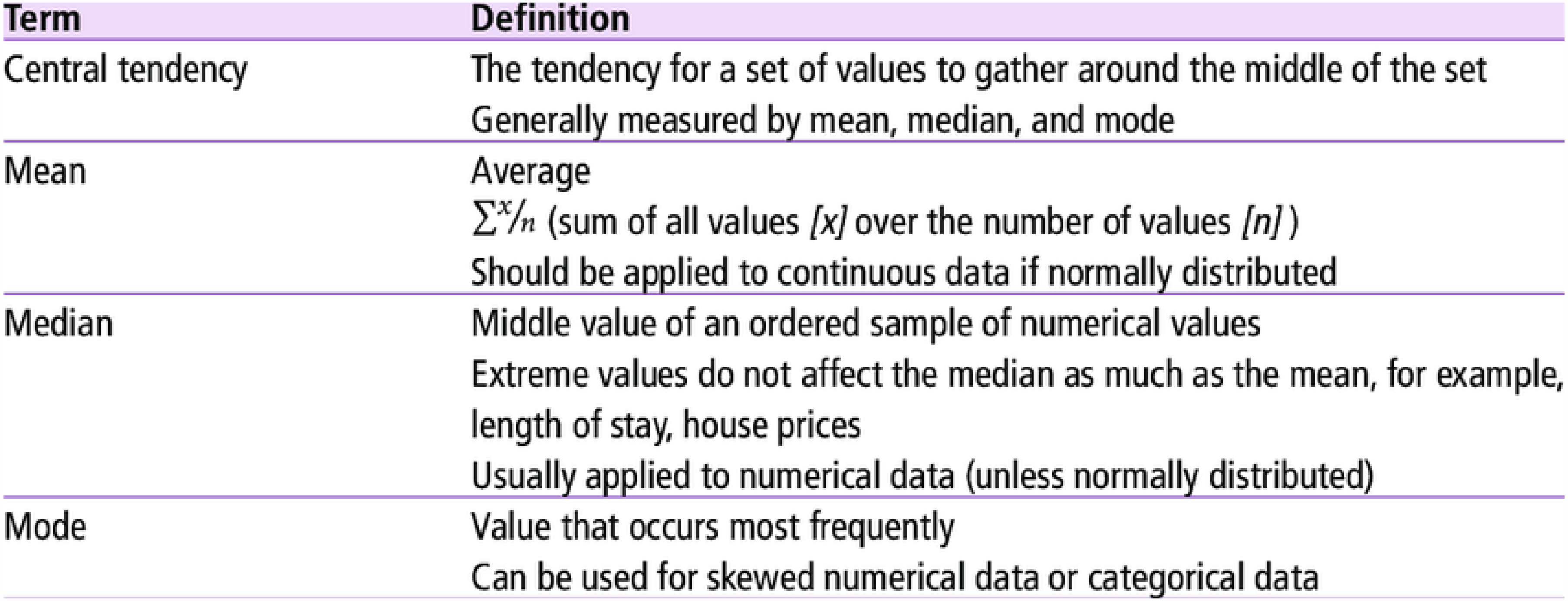
= 7

Arrange in order and pick the middle value

Median = 7, 7, 6, 4, 3, 1

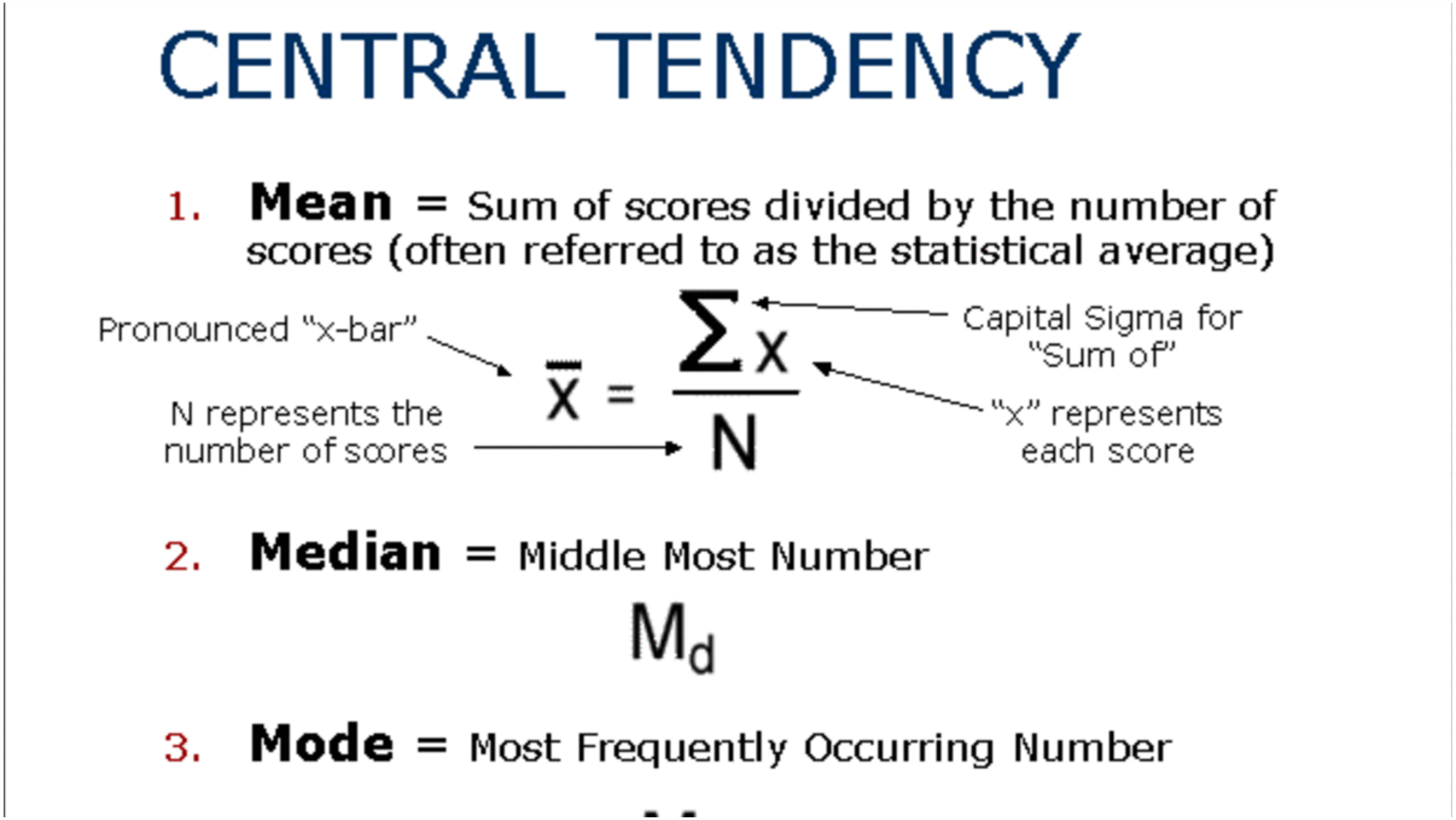
= 6+4 / 2

Unit –V :Data Analysis



Term	Definition
Central tendency	The tendency for a set of values to gather around the middle of the set Generally measured by mean, median, and mode
Mean	Average $\sum x/n$ (sum of all values *[x]* over the number of values *[n]*) Should be applied to continuous data if normally distributed
Median	Middle value of an ordered sample of numerical values Extreme values do not affect the median as much as the mean, for example, length of stay, house prices Usually applied to numerical data (unless normally distributed)
Mode	Value that occurs most frequently Can be used for skewed numerical data or categorical data

Unit –V :Data Analysis



CENTRAL TENDENCY

1. **Mean** = Sum of scores divided by the number of scores (often referred to as the statistical average)

Pronounced "x-bar"

$$\bar{X} = \frac{\sum X}{N}$$

Capital Sigma for "Sum of"

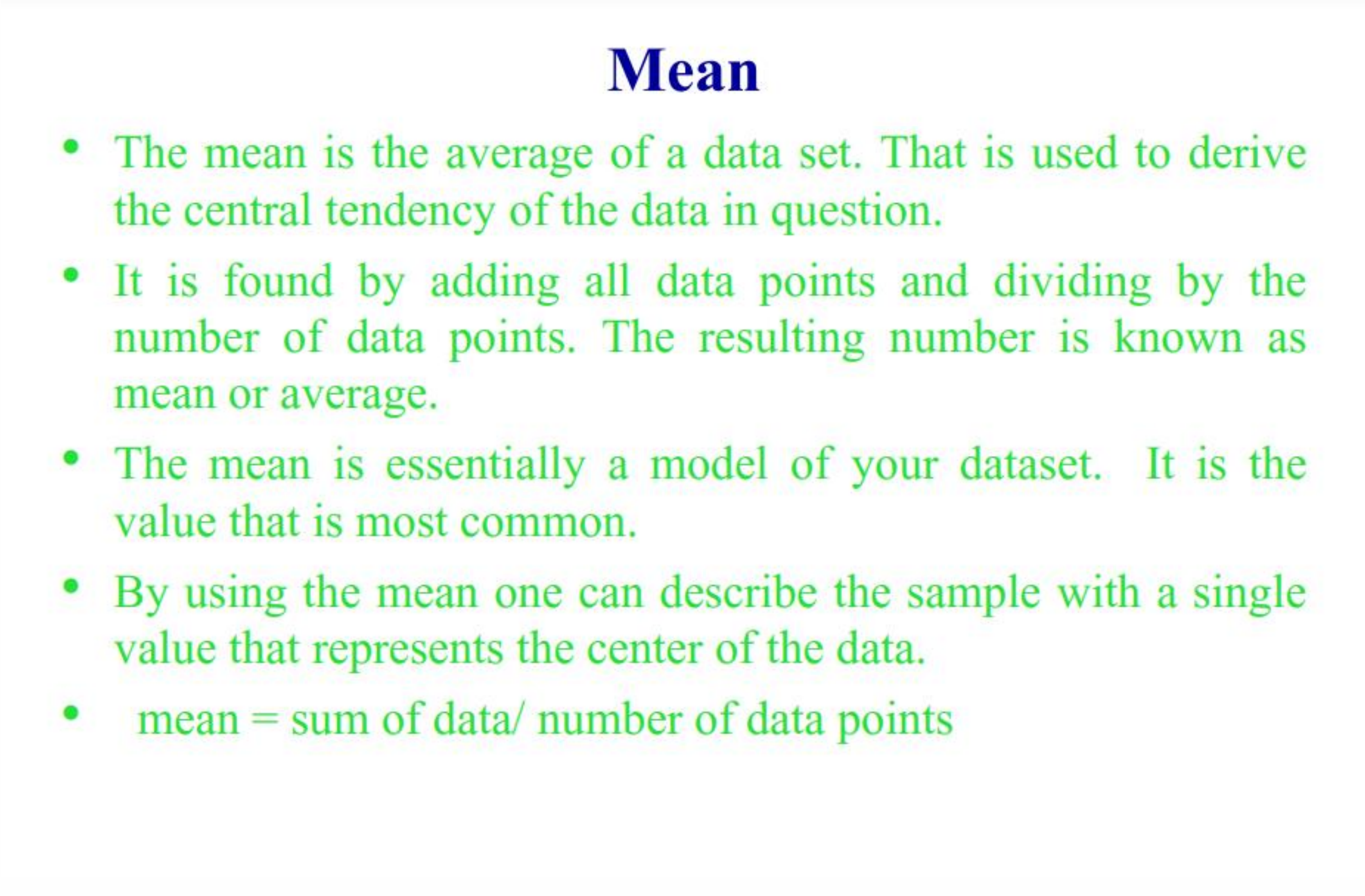
N represents the number of scores

"x" represents each score

2. **Median** = Middle Most Number

M_d

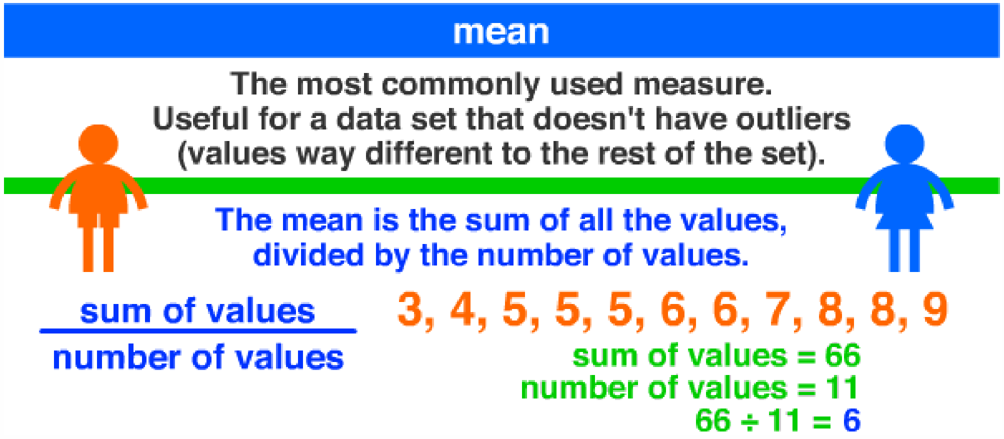
3. **Mode** = Most Frequently Occurring Number



Mean

- The mean is the average of a data set. That is used to derive the central tendency of the data in question.
- It is found by adding all data points and dividing by the number of data points. The resulting number is known as mean or average.
- The mean is essentially a model of your dataset. It is the value that is most common.
- By using the mean one can describe the sample with a single value that represents the center of the data.
- mean = sum of data/ number of data points

Unit –V :Data Analysis



mean

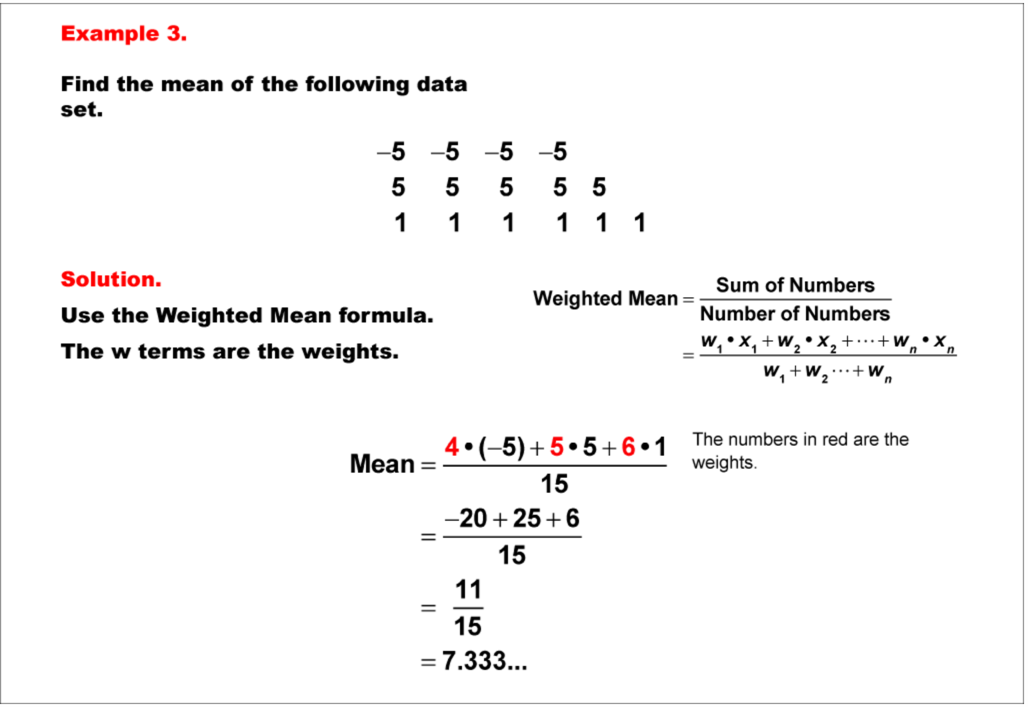
The most commonly used measure.
Useful for a data set that doesn't have outliers
(values way different to the rest of the set).

The mean is the sum of all the values,
divided by the number of values.

$$\frac{\text{sum of values}}{\text{number of values}}$$

3, 4, 5, 5, 5, 6, 6, 7, 8, 8, 9

sum of values = 66
number of values = 11
66 ÷ 11 = 6



Example 3.

Find the mean of the following data set.

–5	–5	–5	–5		
5	5	5	5	5	
1	1	1	1	1	1

Solution.

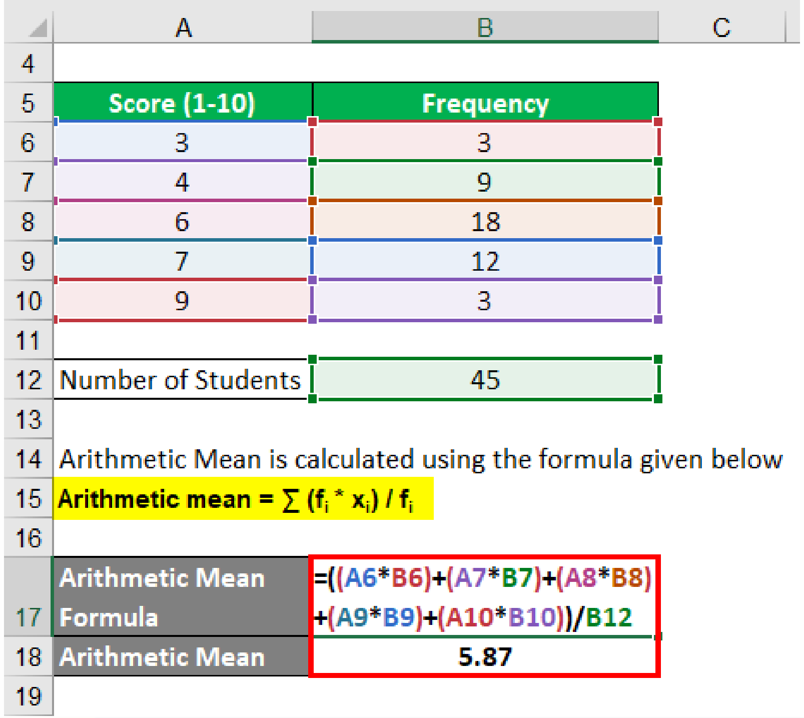
Use the Weighted Mean formula.

The w terms are the weights.

$$\text{Weighted Mean} = \frac{\text{Sum of Numbers}}{\text{Number of Numbers}} = \frac{w_1 \bullet x_1 + w_2 \bullet x_2 + \cdots + w_n \bullet x_n}{w_1 + w_2 \cdots + w_n}$$

$$\text{Mean} = \frac{4 \bullet (-5) + 5 \bullet 5 + 6 \bullet 1}{15} = \frac{-20 + 25 + 6}{15} = \frac{11}{15} = 7.333...$$

The numbers in red are the weights.



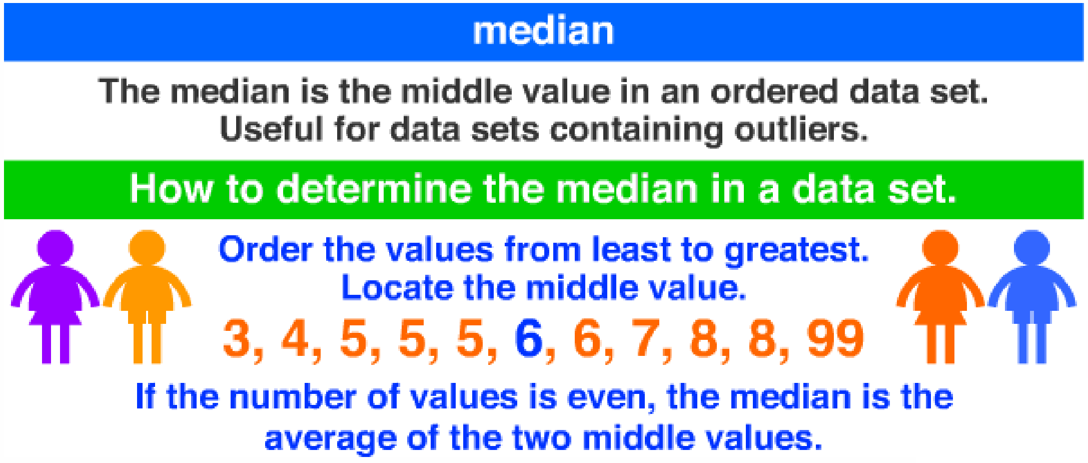
	A	B	C
4			
5	Score (1-10)	Frequency	
6	3	3	
7	4	9	
8	6	18	
9	7	12	
10	9	3	
11			
12	Number of Students	45	
13			
14	Arithmetic Mean is calculated using the formula given below		
15	Arithmetic mean = ∑ (f_i * x_i) / f_i		
16			
17	Arithmetic Mean Formula	=((A6*B6)+(A7*B7)+(A8*B8)+(A9*B9)+(A10*B10))/B12	
18	Arithmetic Mean	5.87	
19			

Median

- It is a simple measure of central tendency. It is the most suitable measure of average for data classified on an ordinal scale.
- The middle number , found by ordering all data points and picking out the one in the middle.
- It is the value which separating the lower half from higher half of a data sample.
- By comparing the median to the mean , one can get an idea of the distribution of a data set. When the mean and the median are same , the data set is more or less evenly distributed from lowest to highest values.

- If the number of observations is odd then the median is the observations that are ranked at position N+1/2(It is the middle value).
- If the number of observations is even then the median is the average value of the observations that are ranked at numbers N/2 and (N/2)+1(It is the average or mean of the two middle most values).
- Outliers and skewed data have a smaller effect on the median consequently , when some of the value s are more extreme , the effect on the median is smaller. When data distribution is skewed the median is better measure of the central tendency than the mean.

Unit –V :Data Analysis



median

The median is the middle value in an ordered data set.
Useful for data sets containing outliers.

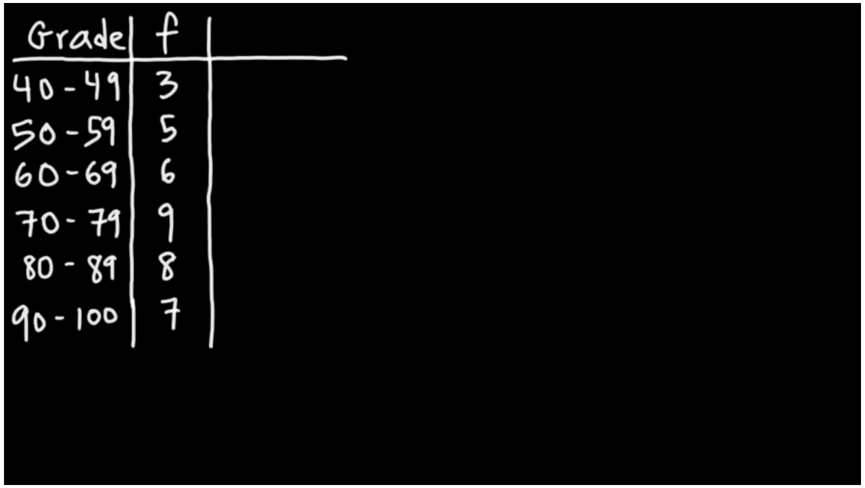
How to determine the median in a data set.

Order the values from least to greatest.
Locate the middle value.

3, 4, 5, 5, 5, **6**, 6, 7, 8, 8, 99

If the number of values is even, the median is the average of the two middle values.

Weight in grams	Number of apples	Cumulative Frequency
410 – 420	14	14
420 – 430	20	34
430 – 440	42	76
440 – 450	54	130
450 – 460	45	175
460 – 470	18	193
470 – 480	7	200
Total	N = 200	



Grade	f	
40-49	3	
50-59	5	
60-69	6	
70-79	9	
80-89	8	
90-100	7	

$$\frac{N}{2} = \frac{200}{2} = 100.$$

Median class is 440 – 450

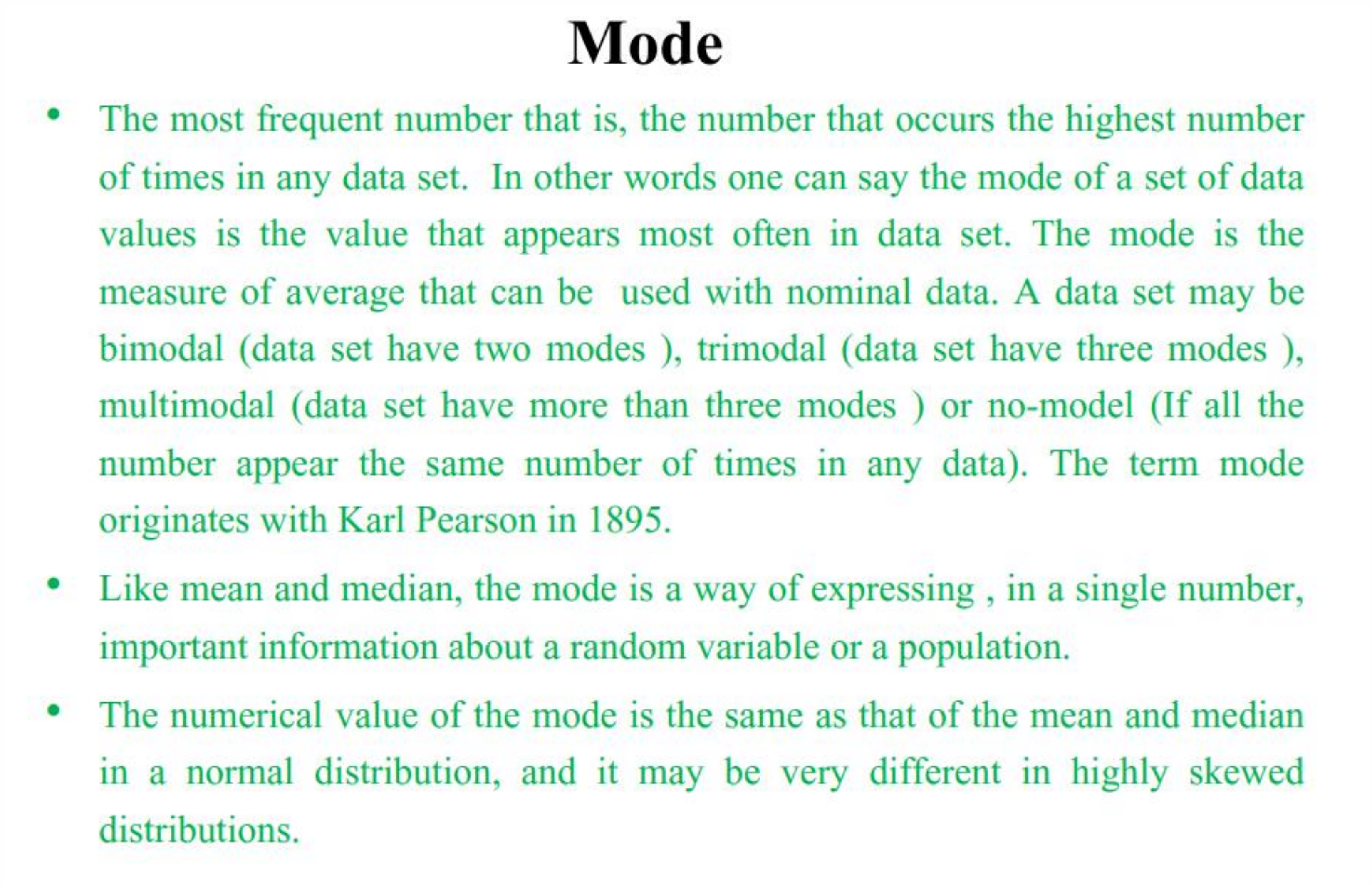
$$\text{Median } = l + \frac{\frac{N}{2} - m}{f} \times c$$

$l = 440, \quad \frac{N}{2} = 100, \quad m = 76, \quad f = 54, \quad c = 10$

$$\text{Median } = 440 + \frac{100-76}{54} \times 10$$

$$= 440 + \frac{24}{54} \times 10 = 440 + 4.44 = 444.44$$

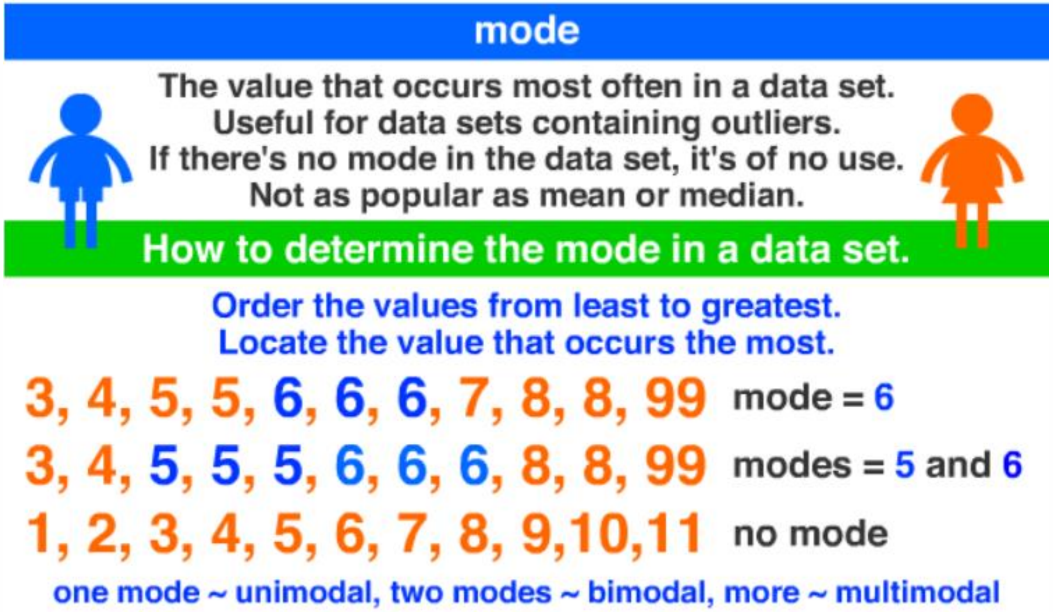
The median weight of the apple is 444.44 grams



Mode

- The most frequent number that is, the number that occurs the highest number of times in any data set. In other words one can say the mode of a set of data values is the value that appears most often in data set. The mode is the measure of average that can be used with nominal data. A data set may be bimodal (data set have two modes), trimodal (data set have three modes), multimodal (data set have more than three modes) or no-model (If all the number appear the same number of times in any data). The term mode originates with Karl Pearson in 1895.
- Like mean and median, the mode is a way of expressing , in a single number, important information about a random variable or a population.
- The numerical value of the mode is the same as that of the mean and median in a normal distribution, and it may be very different in highly skewed distributions.

Unit –V :Data Analysis



mode

The value that occurs most often in a data set.
Useful for data sets containing outliers.
If there's no mode in the data set, it's of no use.
Not as popular as mean or median.

How to determine the mode in a data set.

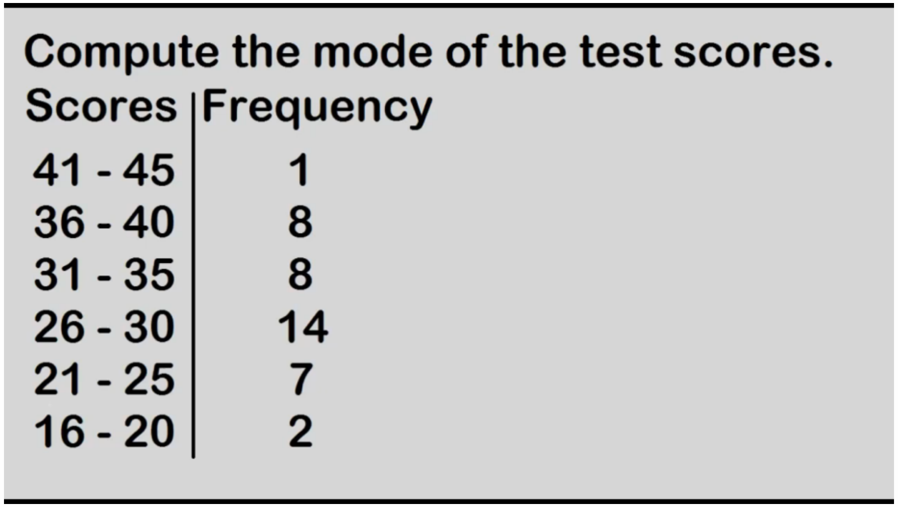
Order the values from least to greatest.
Locate the value that occurs the most.

3, 4, 5, 5, 6, 6, 6, 7, 8, 8, 99 mode = 6

3, 4, 5, 5, 5, 6, 6, 6, 8, 8, 99 modes = 5 and 6

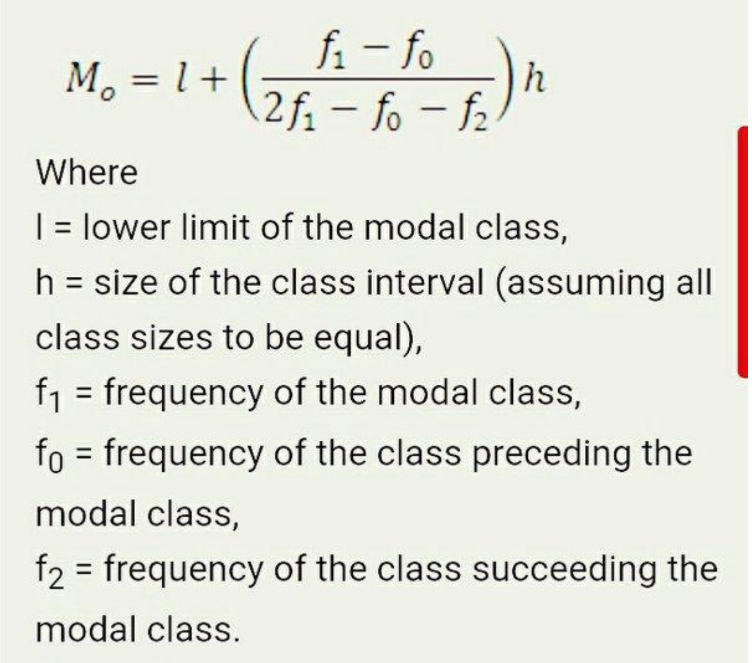
1, 2, 3, 4, 5, 6, 7, 8, 9,10,11 no mode

one mode ~ unimodal, two modes ~ bimodal, more ~ multimodal



Compute the mode of the test scores.

Scores	Frequency
41 - 45	1
36 - 40	8
31 - 35	8
26 - 30	14
21 - 25	7
16 - 20	2



$$M_o = l + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right)h$$

Where

l = lower limit of the modal class,

h = size of the class interval (assuming all class sizes to be equal),

f_1 = frequency of the modal class,

f_0 = frequency of the class preceding the modal class,

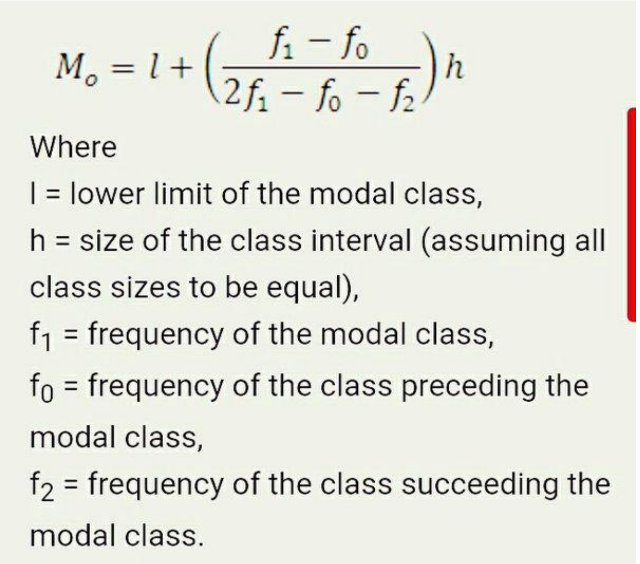
f_2 = frequency of the class succeeding the modal class.

Unit –V :Data Analysis

Example 5.

Find out the mode from the following data :

Class Interval	10-19	20-29	30-39	40-49	50-59	60-69	70-79
Frequency	10	12	18	30	16	6	8



$$M_o = l + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right)h$$

Where

l = lower limit of the modal class,

h = size of the class interval (assuming all class sizes to be equal),

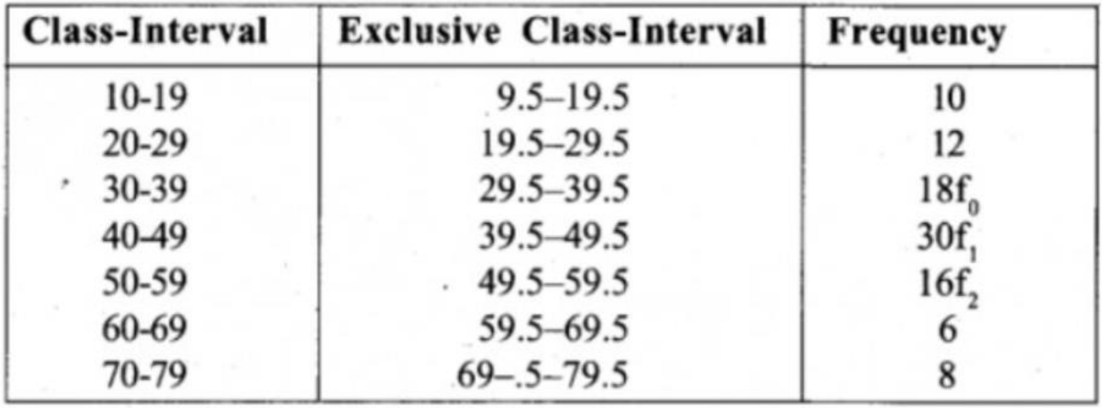
f_1 = frequency of the modal class,

f_0 = frequency of the class preceding the modal class,

f_2 = frequency of the class succeeding the modal class.

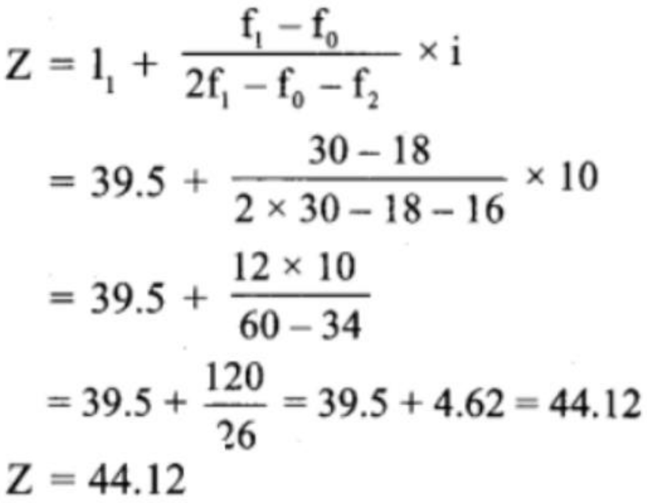
Solution:

Firstly we will convert inclusive series into exclusive series before solving the question. After that the mode will be calculated by the following way :



Class-Interval	Exclusive Class-Interval	Frequency
10-19	9.5–19.5	10
20-29	19.5–29.5	12
30-39	29.5–39.5	$18f_0$
40-49	39.5–49.5	$30f_1$
50-59	49.5–59.5	$16f_2$
60-69	59.5–69.5	6
70-79	69–.5–79.5	8

Class 39.5 – 49.5 has highest frequency and it is the mode-class.



$$Z = l_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

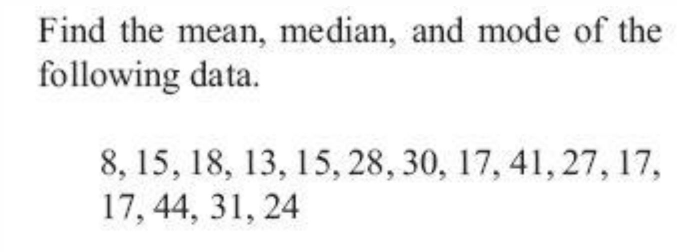
$$= 39.5 + \frac{30 - 18}{2 \times 30 - 18 - 16} \times 10$$

$$= 39.5 + \frac{12 \times 10}{60 - 34}$$

$$= 39.5 + \frac{120}{26} = 39.5 + 4.62 = 44.12$$

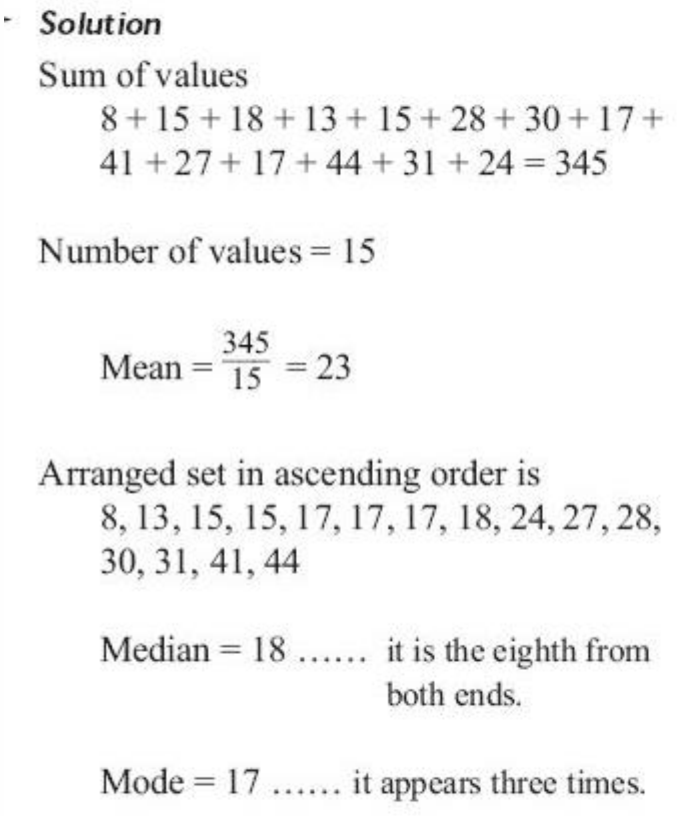
$$Z = 44.12$$

Unit –V :Data Analysis



Find the mean, median, and mode of the following data.

8, 15, 18, 13, 15, 28, 30, 17, 41, 27, 17, 17, 44, 31, 24



- ***Solution***

Sum of values

8 + 15 + 18 + 13 + 15 + 28 + 30 + 17 + 41 + 27 + 17 + 44 + 31 + 24 = 345

Number of values = 15

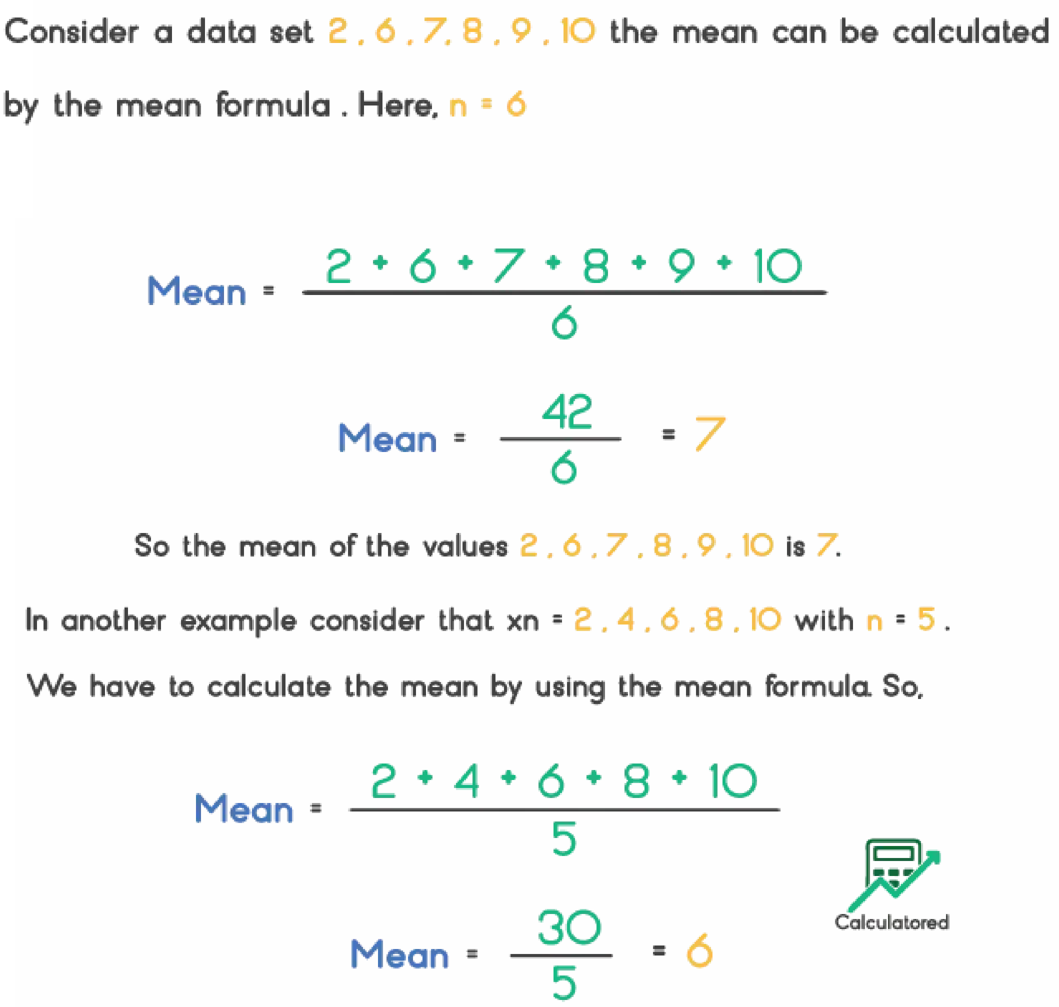
$$\text{Mean} = \frac{345}{15} = 23$$

Arranged set in ascending order is

8, 13, 15, 15, 17, 17, 17, 18, 24, 27, 28, 30, 31, 41, 44

Median = 18 ...... it is the eighth from both ends.

Mode = 17 ...... it appears three times.



Consider a data set 2 , 6 , 7, 8 , 9 , 10 the mean can be calculated by the mean formula . Here, n = 6

$$\text{Mean} = \frac{2 + 6 + 7 + 8 + 9 + 10}{6}$$

$$\text{Mean} = \frac{42}{6} = 7$$

So the mean of the values 2 , 6 , 7 , 8 , 9 , 10 is 7.

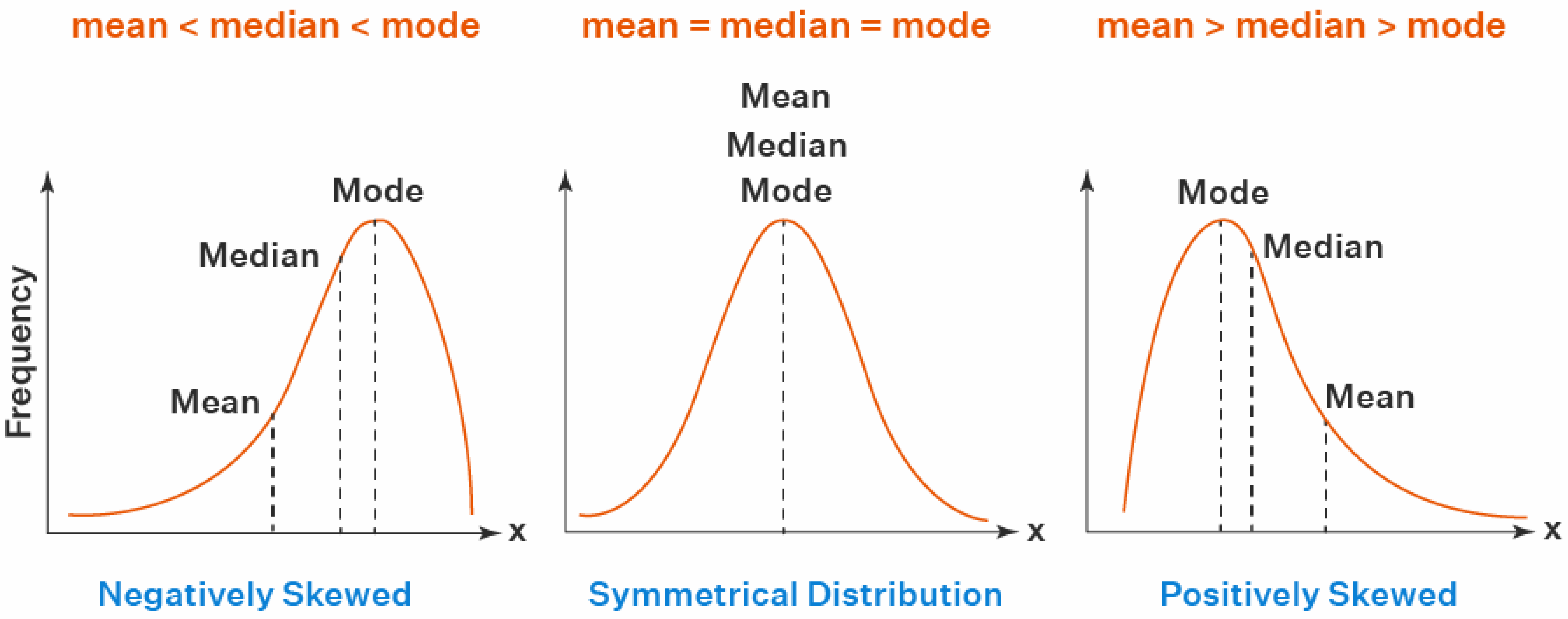
In another example consider that xn = 2 , 4 , 6 , 8 , 10 with n = 5 .

We have to calculate the mean by using the mean formula. So,

$$\text{Mean} = \frac{2 + 4 + 6 + 8 + 10}{5}$$

$$\text{Mean} = \frac{30}{5} = 6$$

Unit –V :Data Analysis



mean < median < mode

Negatively Skewed

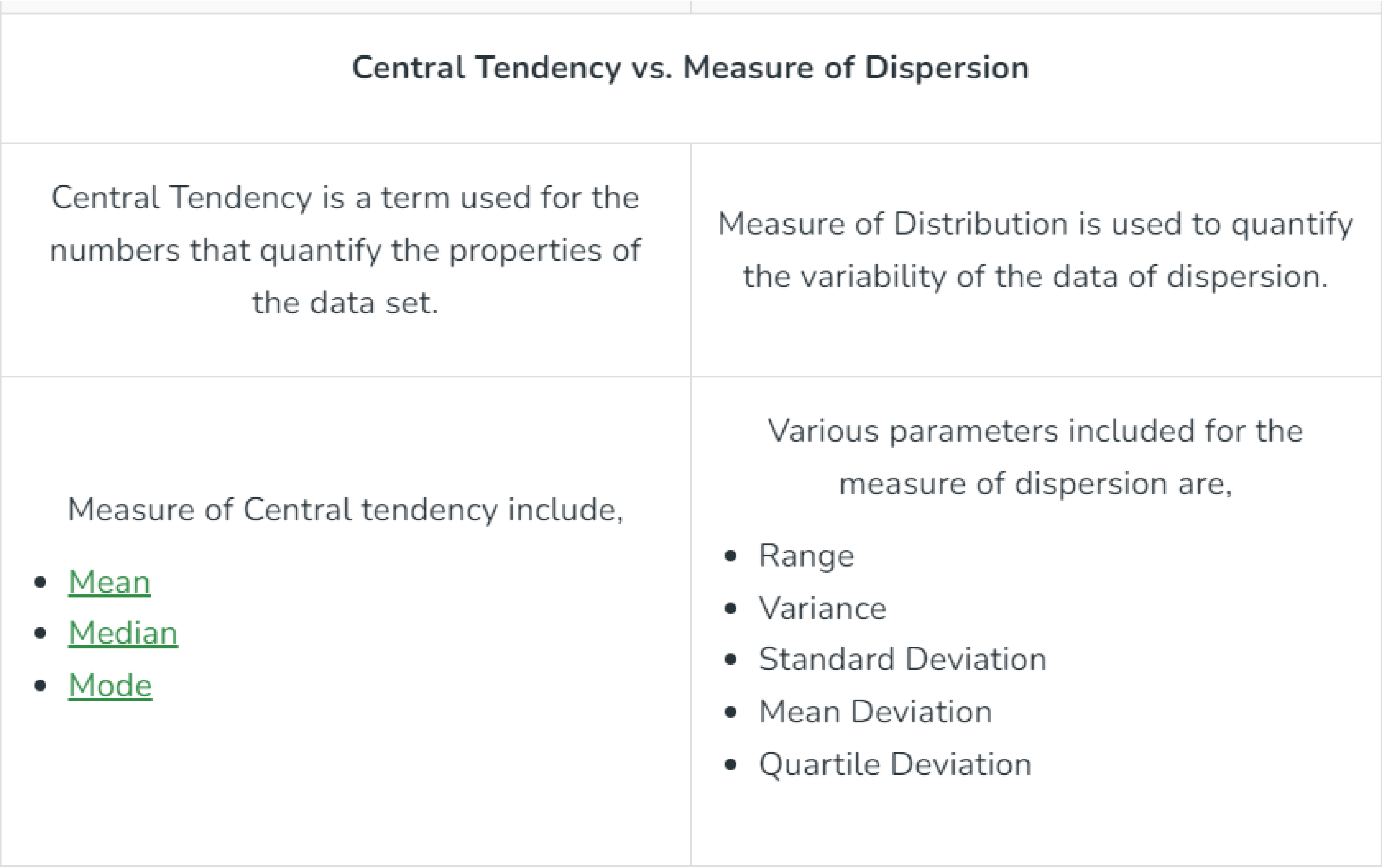
mean = median = mode

Symmetrical Distribution

mean > median > mode

Positively Skewed

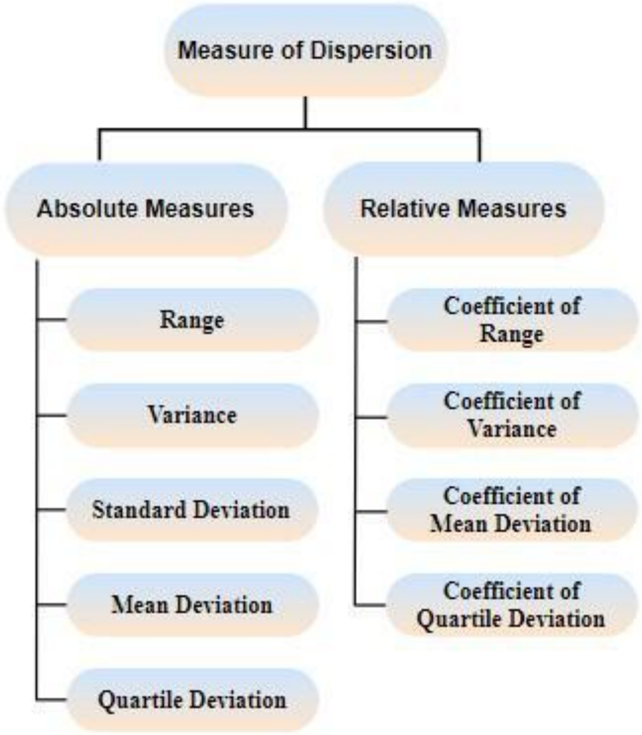
Unit –V :Data Analysis

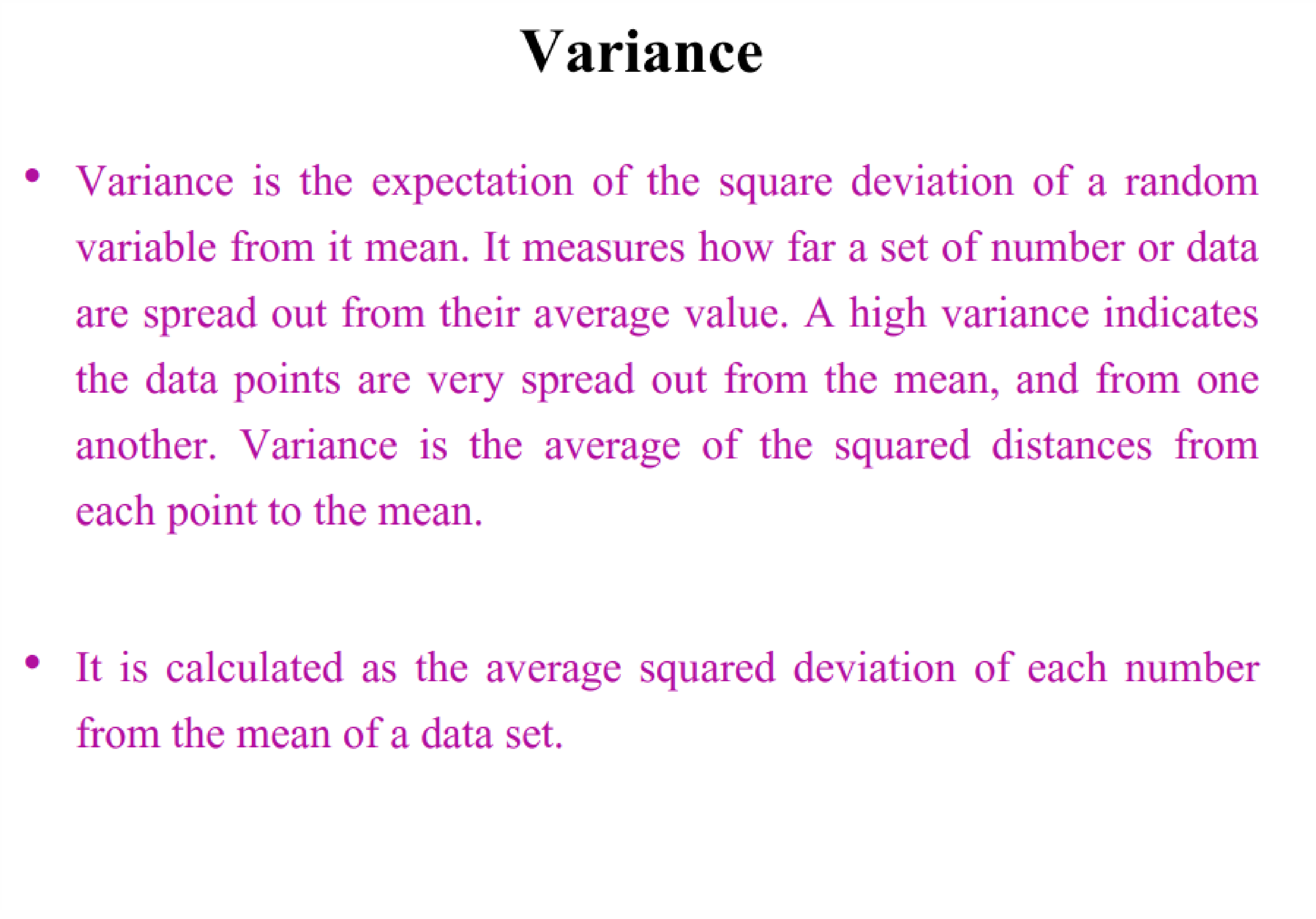


Central Tendency vs. Measure of Dispersion	
Central Tendency is a term used for the numbers that quantify the properties of the data set.	Measure of Distribution is used to quantify the variability of the data of dispersion.
Measure of Central tendency include, • Mean • Median • Mode	Various parameters included for the measure of dispersion are, • Range • Variance • Standard Deviation • Mean Deviation • Quartile Deviation

Dispersion

Dispersion is the extent to which a distribution is stretched or squeezed. It is also called variability, scatter or spread. Common example of measures of dispersion in statistic are the variance, standard deviation and interquartile range. Measure of statistical dispersion is a non negative that is zero if all the data are the same and increases as the data become more diverse.

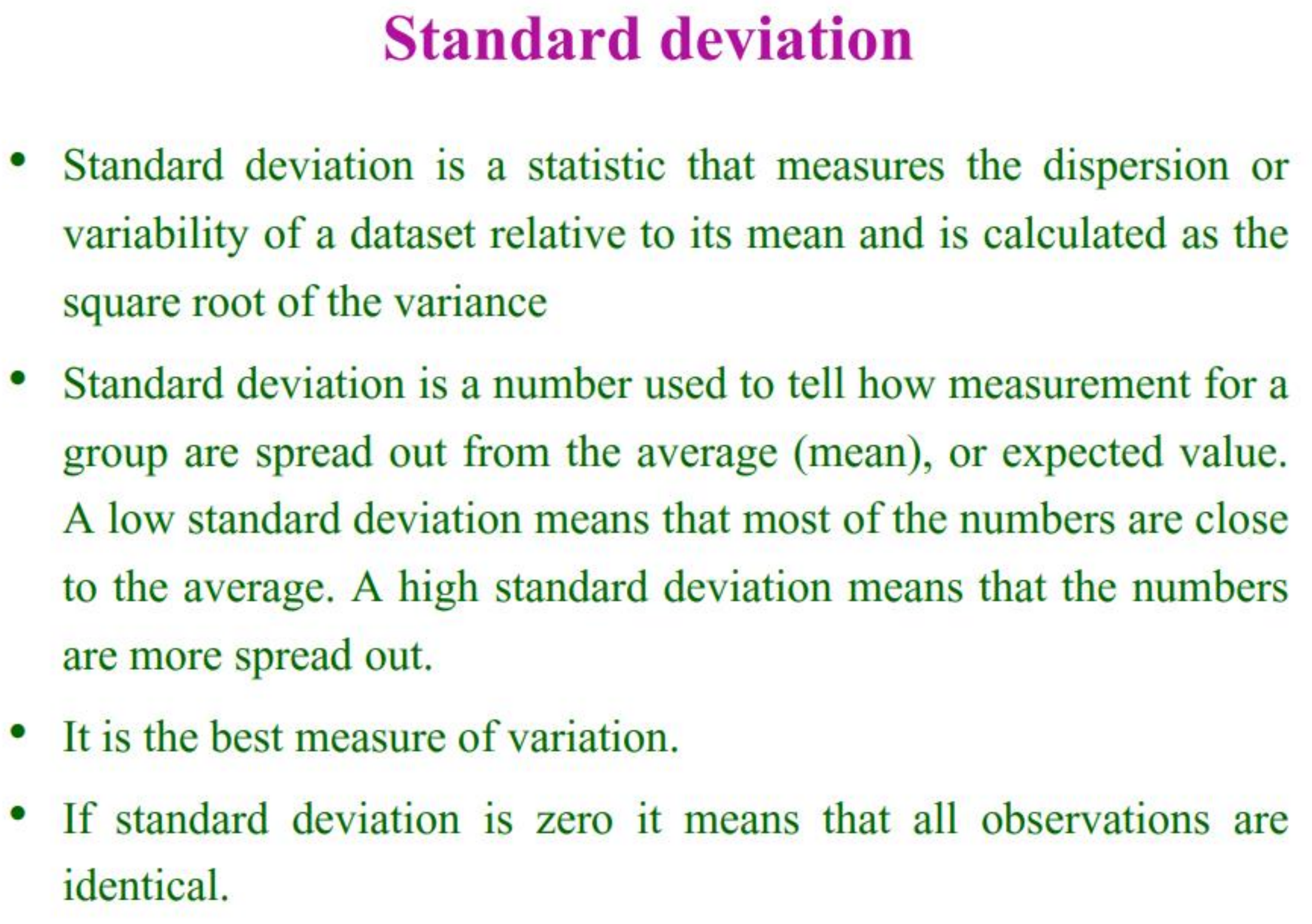




Variance

- Variance is the expectation of the square deviation of a random variable from it mean. It measures how far a set of number or data are spread out from their average value. A high variance indicates the data points are very spread out from the mean, and from one another. Variance is the average of the squared distances from each point to the mean.

- It is calculated as the average squared deviation of each number from the mean of a data set.



Standard deviation

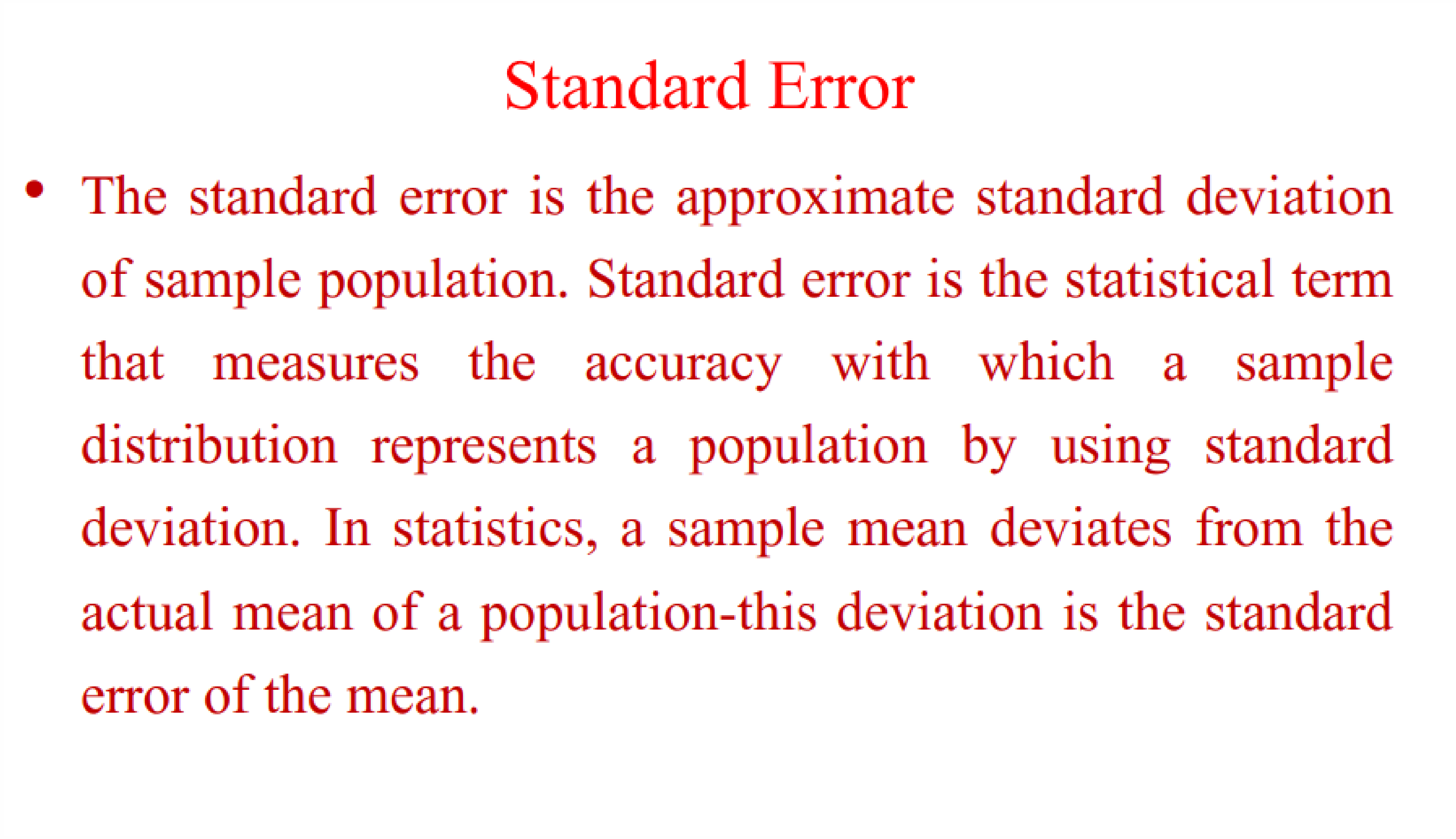
- Standard deviation is a statistic that measures the dispersion or variability of a dataset relative to its mean and is calculated as the square root of the variance
- Standard deviation is a number used to tell how measurement for a group are spread out from the average (mean), or expected value. A low standard deviation means that most of the numbers are close to the average. A high standard deviation means that the numbers are more spread out.
- It is the best measure of variation.
- If standard deviation is zero it means that all observations are identical.

Standard deviation vs. variance

- Standard deviation and variance are the most commonly used measures of spread.

- Standard deviation looks at how spread out a group of numbers is from the means, by looking at the square root of the variance . The variance measures the average degree to which each point differs from the mean(the average of all data point).

- Standard deviation symbolized by ‘S^{2}’ and variance symbolized by ‘S’

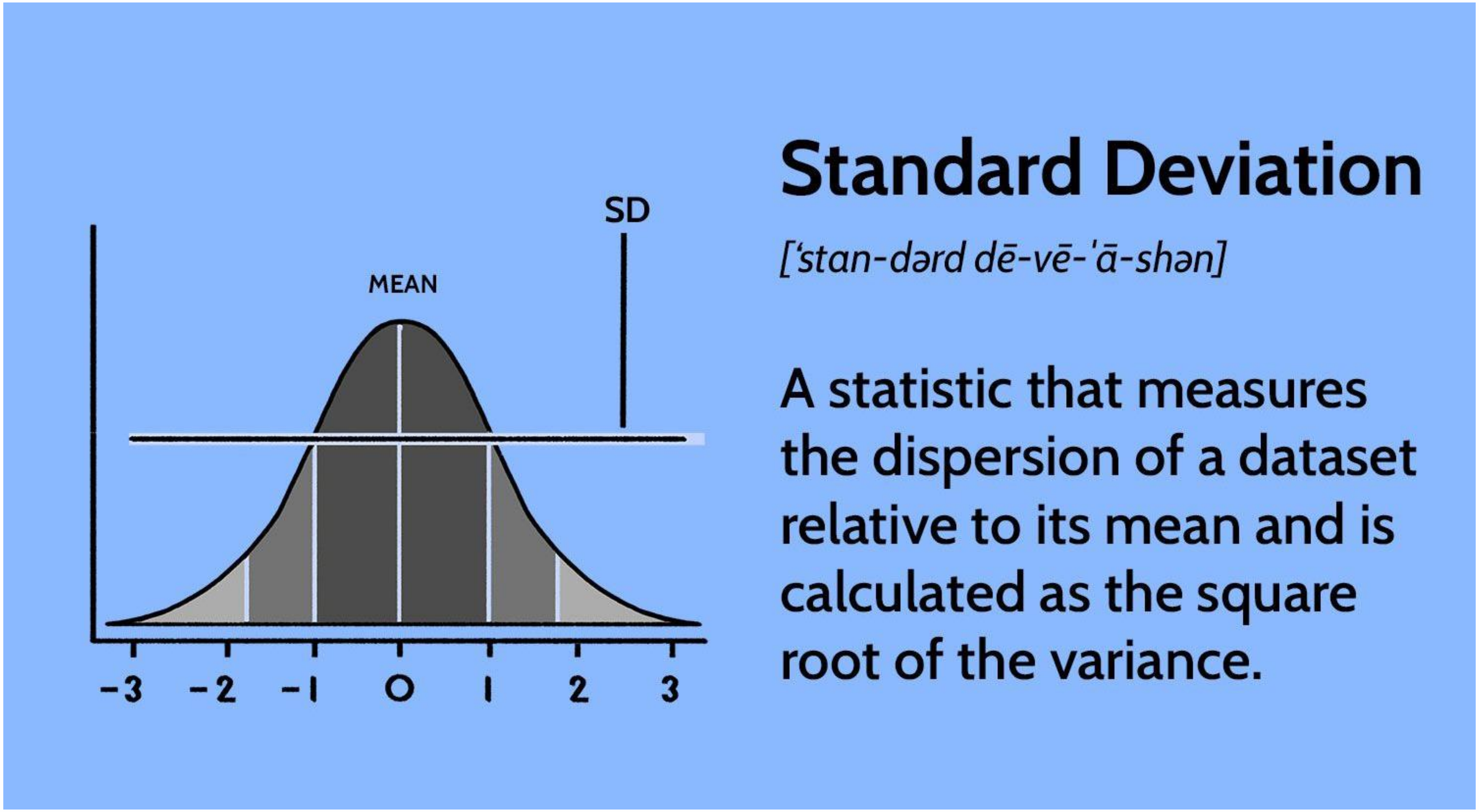


Standard Error

- The standard error is the approximate standard deviation of sample population. Standard error is the statistical term that measures the accuracy with which a sample distribution represents a population by using standard deviation. In statistics, a sample mean deviates from the actual mean of a population-this deviation is the standard error of the mean.

Coefficient of variation

- The coefficient of variation also known as relative standard deviation, is a standardized measure of dispersion of a probability distribution or frequency distribution.
- It is useful static for comparing the degree of variation from one data series to another. The greater the SD value the less precise the data because it increases the acceptable range within the deviation. Value of CV between 2%-3% is good and acceptable.
- It is often expressed as a percentage, and is defined as the ratio of the standard deviation to the mean.



Standard Deviation

['stan-dərd dē-vē-'ā-shən]

A statistic that measures the dispersion of a dataset relative to its mean and is calculated as the square root of the variance.

Unit –V :Data Analysis

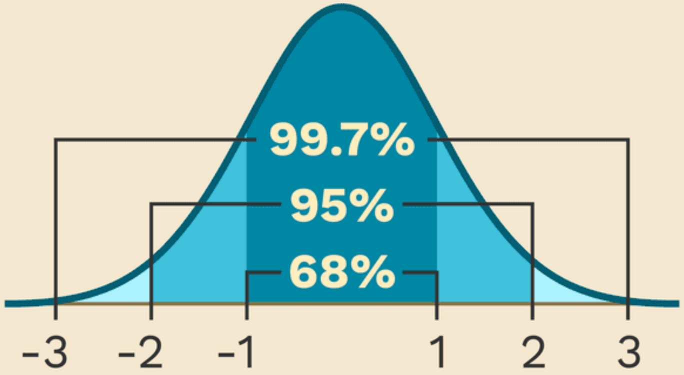
Calculating Standard Deviation

$$S_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

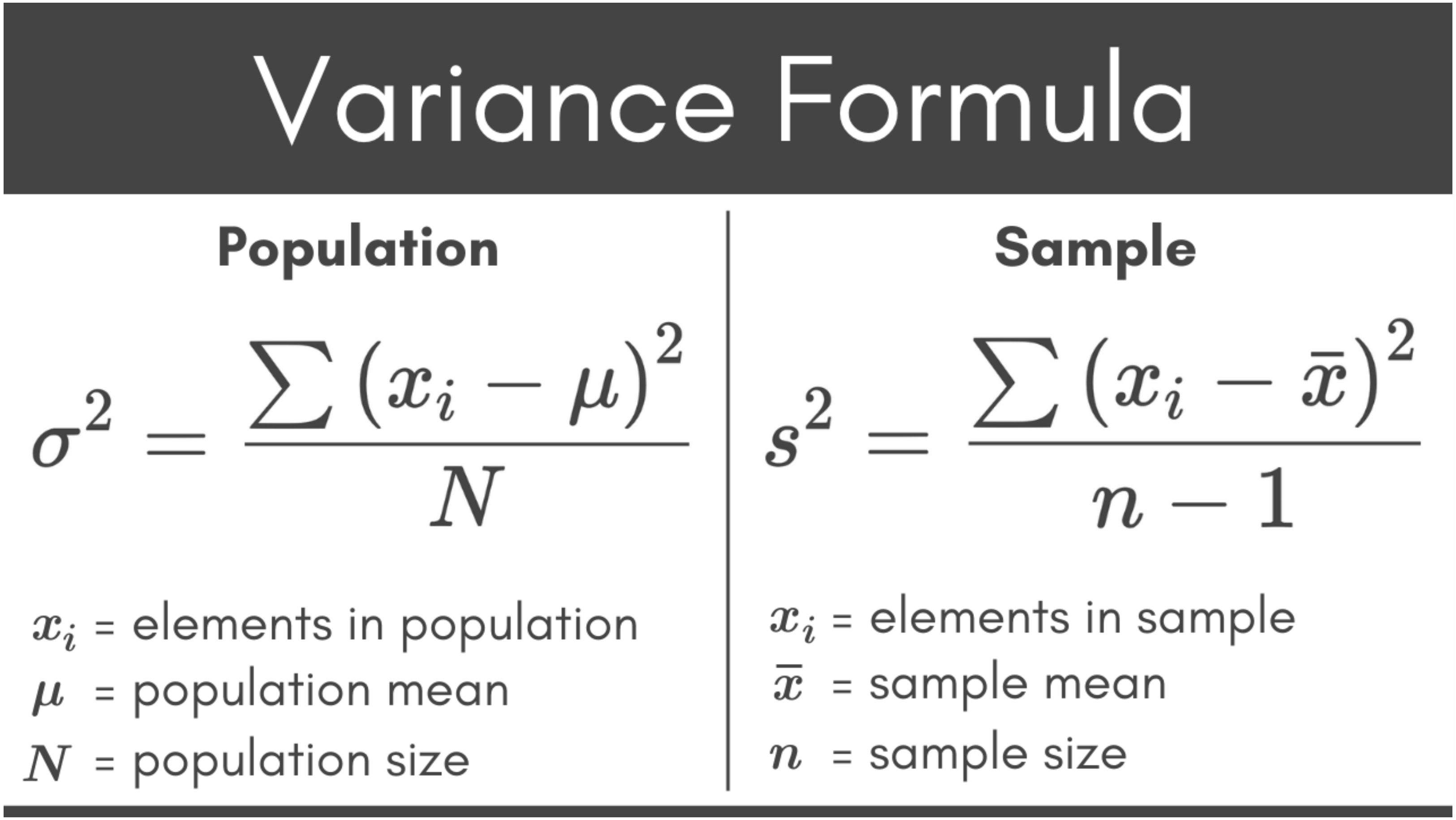
n = The number of data points

x_i = Each of the values of the data

$\bar{x}$ = The mean of x_i



Normal Distribution Curve



Variance Formula

Population

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

x_i = elements in population

μ = population mean

N = population size

Sample

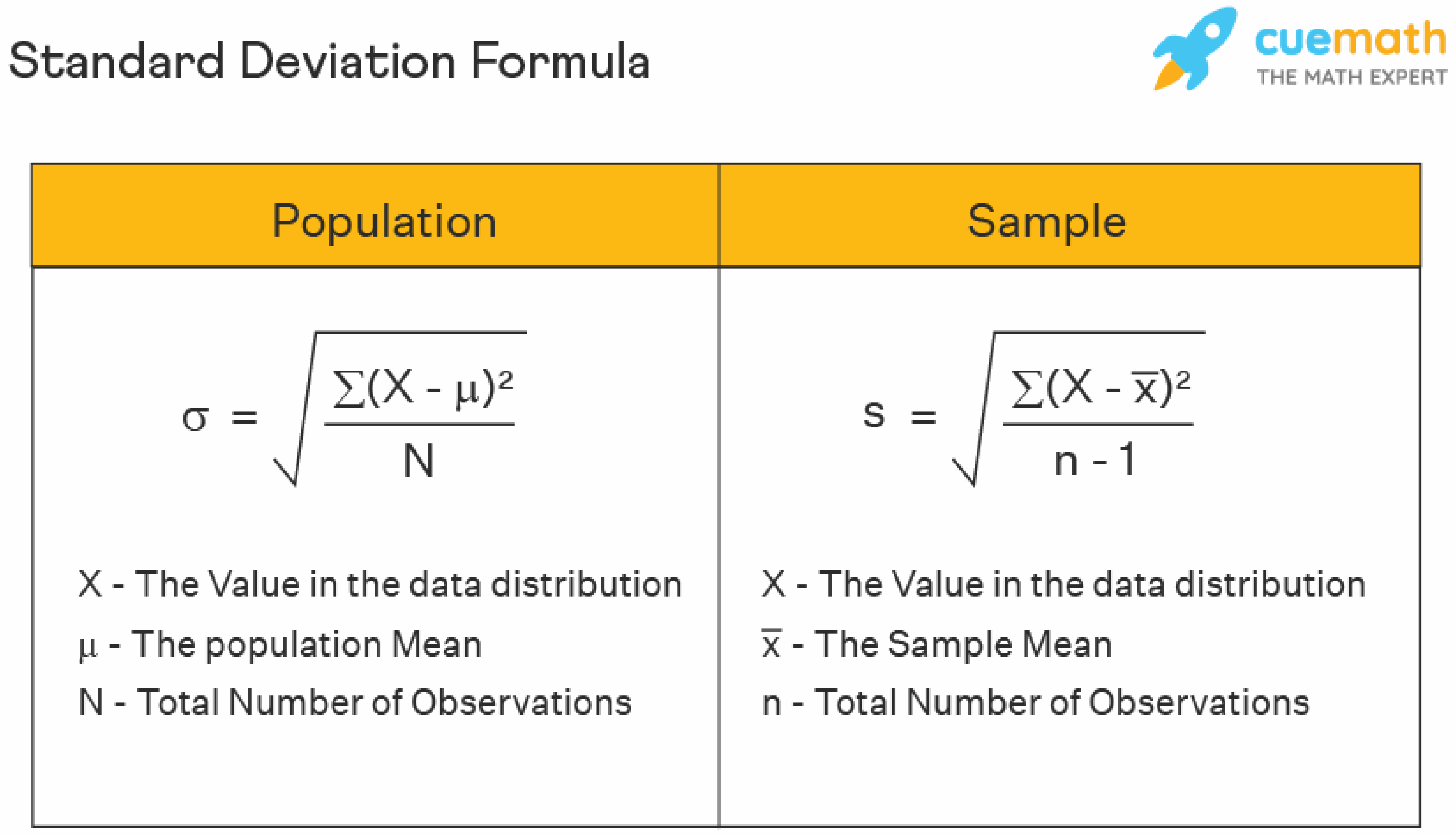
$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

x_i = elements in sample

$\bar{x}$ = sample mean

n = sample size

Unit –V :Data Analysis

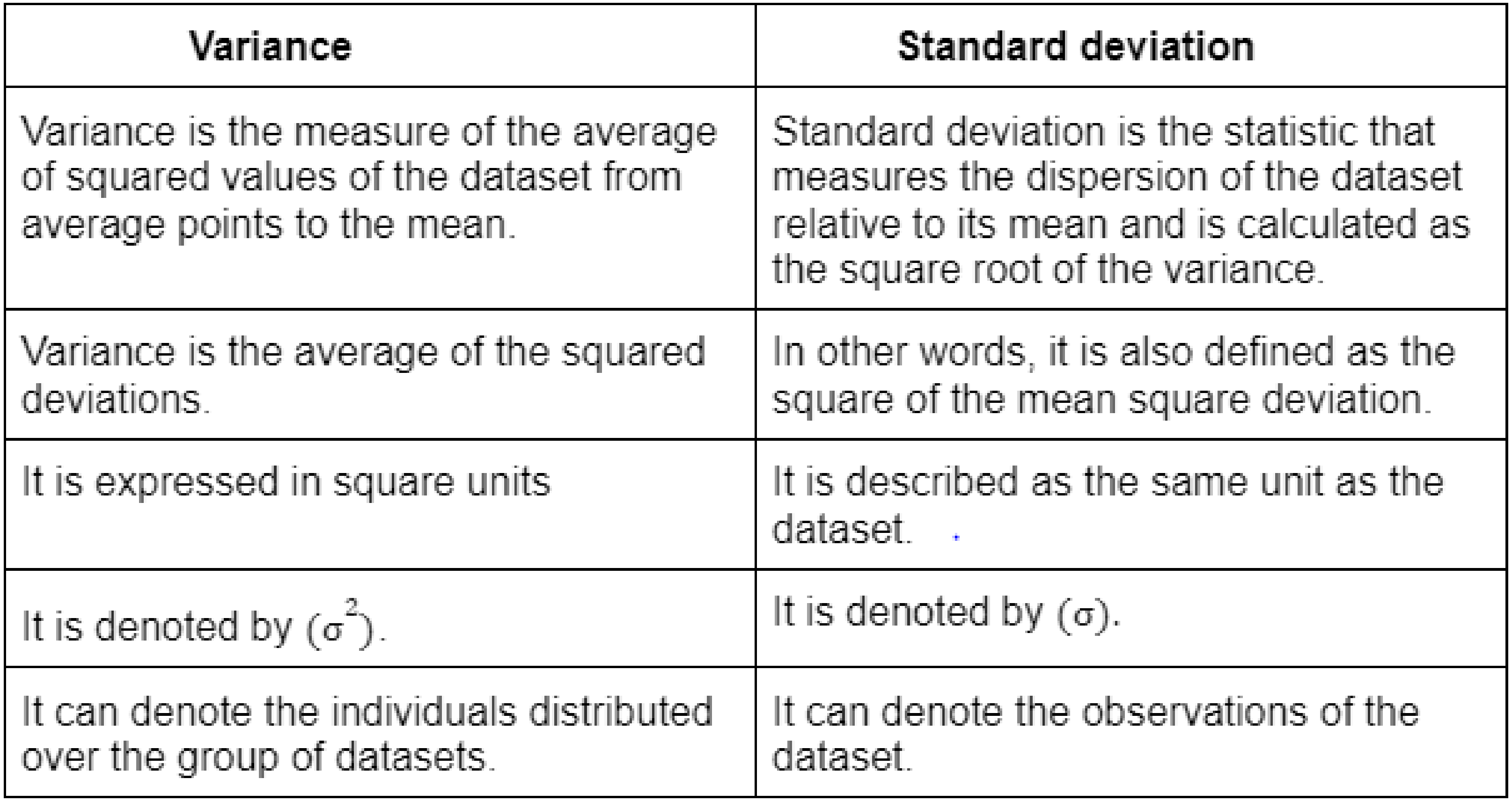


Standard Deviation Formula

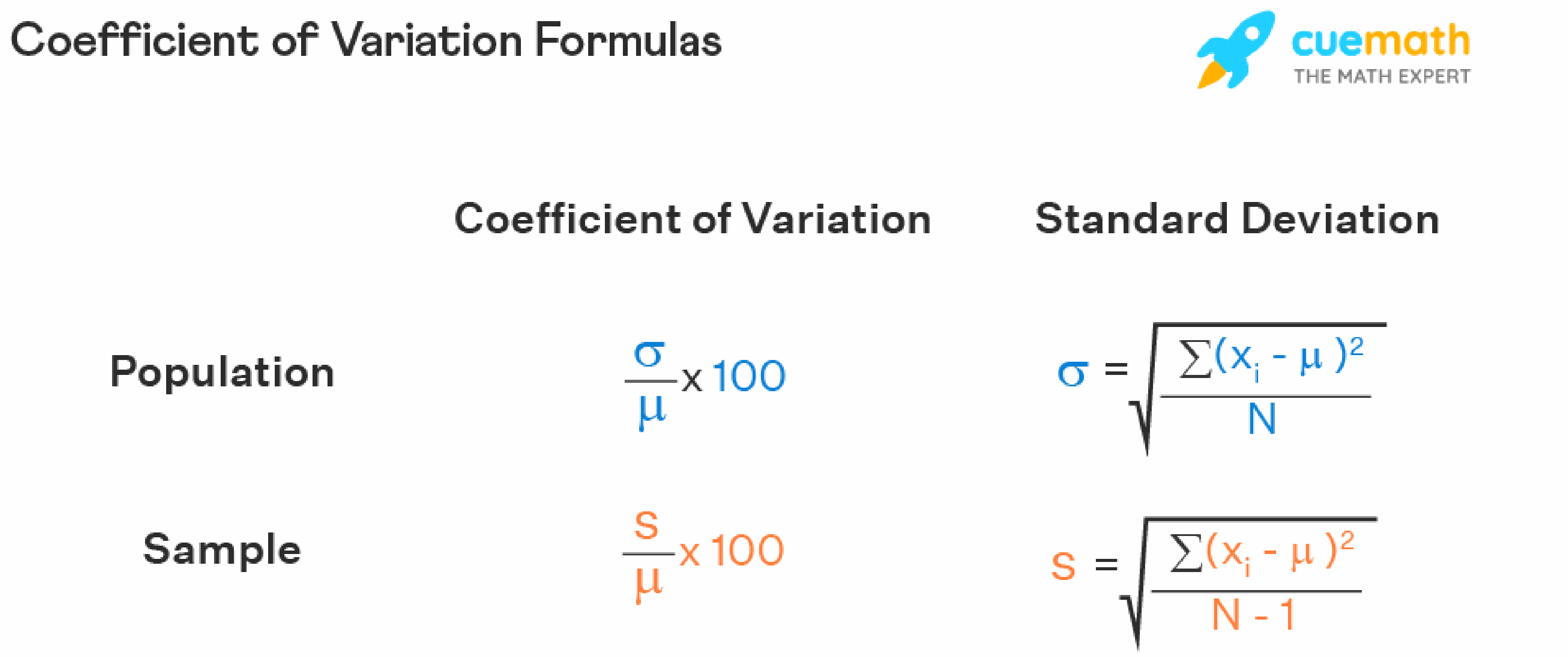
cuemath
THE MATH EXPERT

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

Unit –V :Data Analysis



Variance	Standard deviation
Variance is the measure of the average of squared values of the dataset from average points to the mean.	Standard deviation is the statistic that measures the dispersion of the dataset relative to its mean and is calculated as the square root of the variance.
Variance is the average of the squared deviations.	In other words, it is also defined as the square of the mean square deviation.
It is expressed in square units	It is described as the same unit as the dataset.
It is denoted by (σ^2).	It is denoted by (σ).
It can denote the individuals distributed over the group of datasets.	It can denote the observations of the dataset.

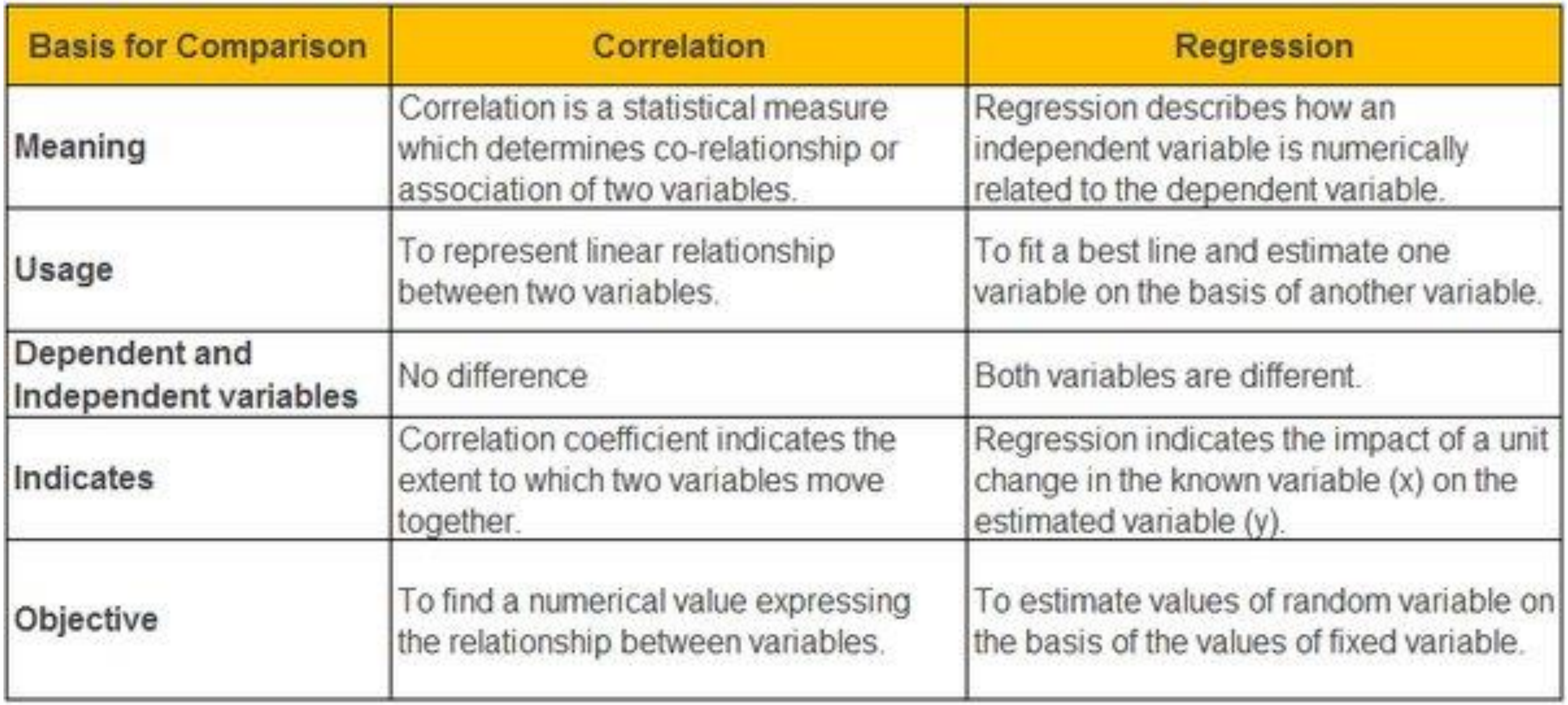


Coefficient of Variation Formulas

cuemath
THE MATH EXPERT

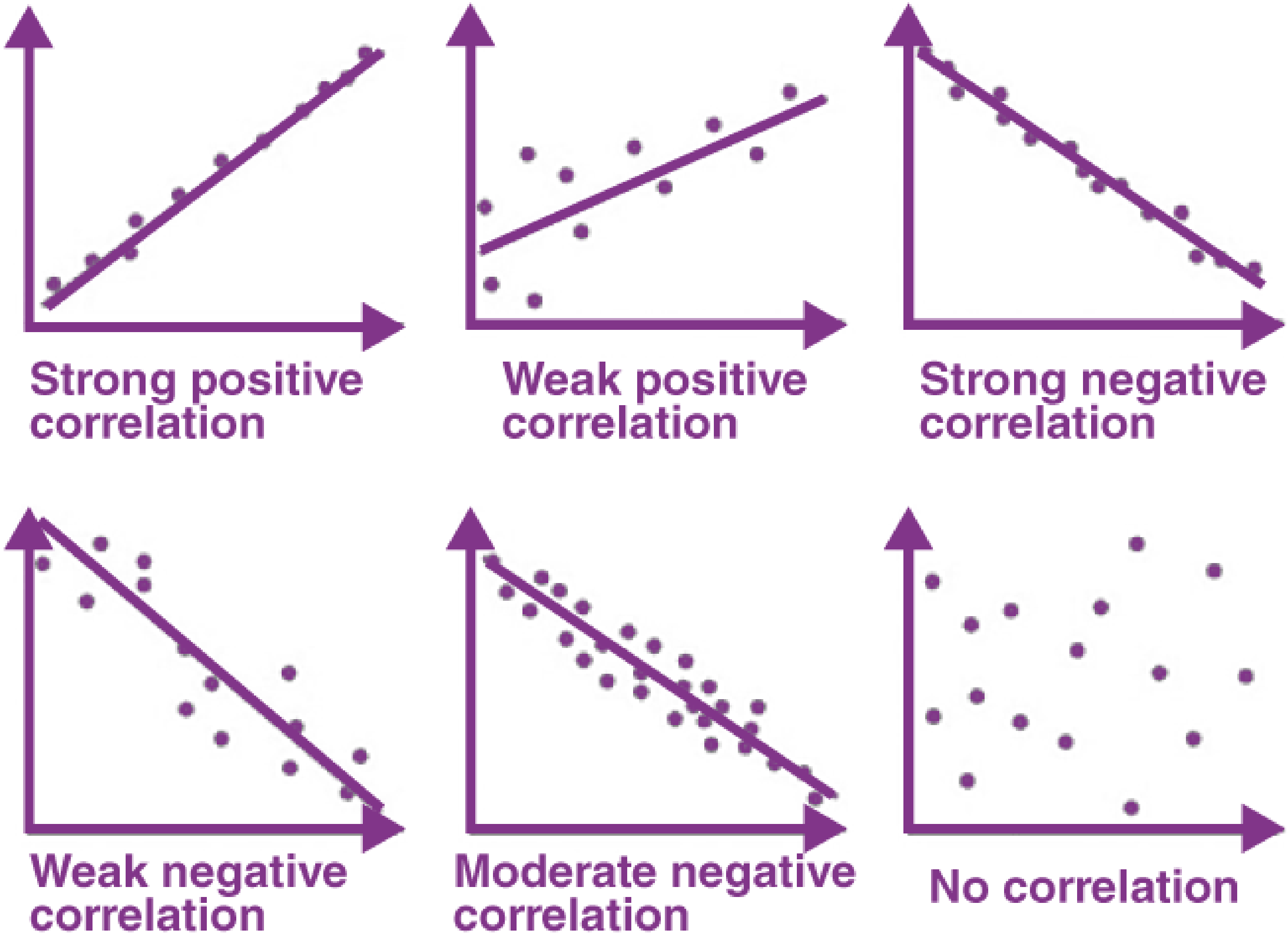
	Coefficient of Variation	Standard Deviation
Population	$\frac{\sigma}{\mu} \times 100$	$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$
Sample	$\frac{s}{\mu} \times 100$	$s = \sqrt{\frac{\sum(x_i - \mu)^2}{N - 1}}$

Unit –V :Data Analysis



Basis for Comparison	Correlation	Regression
Meaning	Correlation is a statistical measure which determines co-relationship or association of two variables.	Regression describes how an independent variable is numerically related to the dependent variable.
Usage	To represent linear relationship between two variables.	To fit a best line and estimate one variable on the basis of another variable.
Dependent and Independent variables	No difference	Both variables are different.
Indicates	Correlation coefficient indicates the extent to which two variables move together.	Regression indicates the impact of a unit change in the known variable (x) on the estimated variable (y).
Objective	To find a numerical value expressing the relationship between variables.	To estimate values of random variable on the basis of the values of fixed variable.

Unit –V :Data Analysis



Strong positive correlation

Weak positive correlation

Strong negative correlation

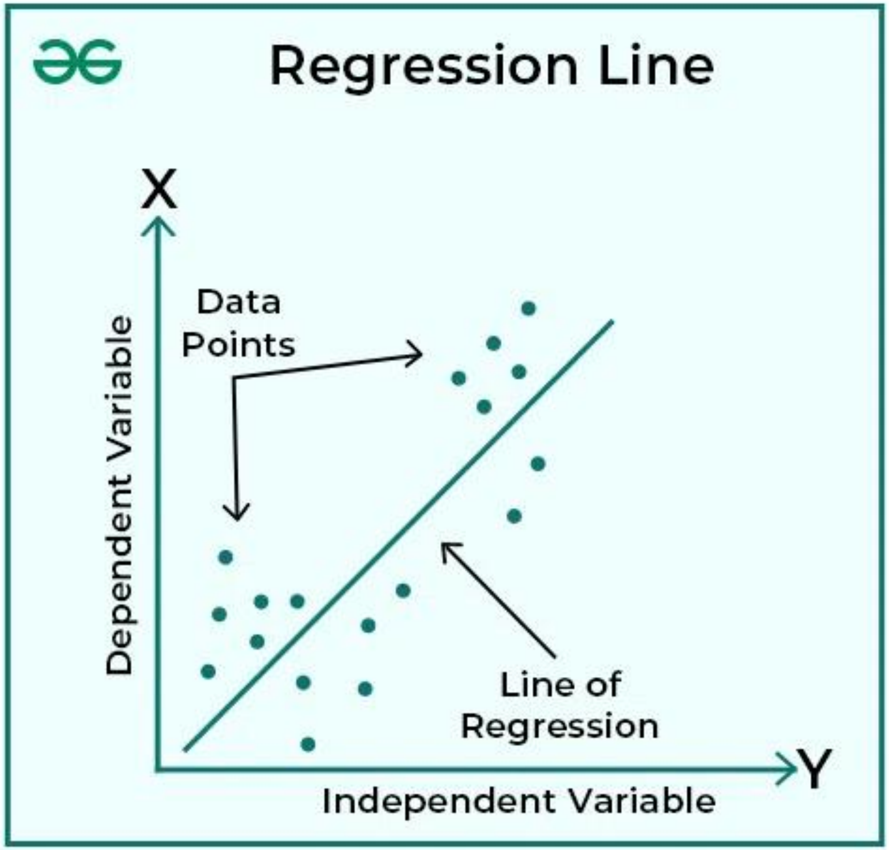
Weak negative correlation

Moderate negative correlation

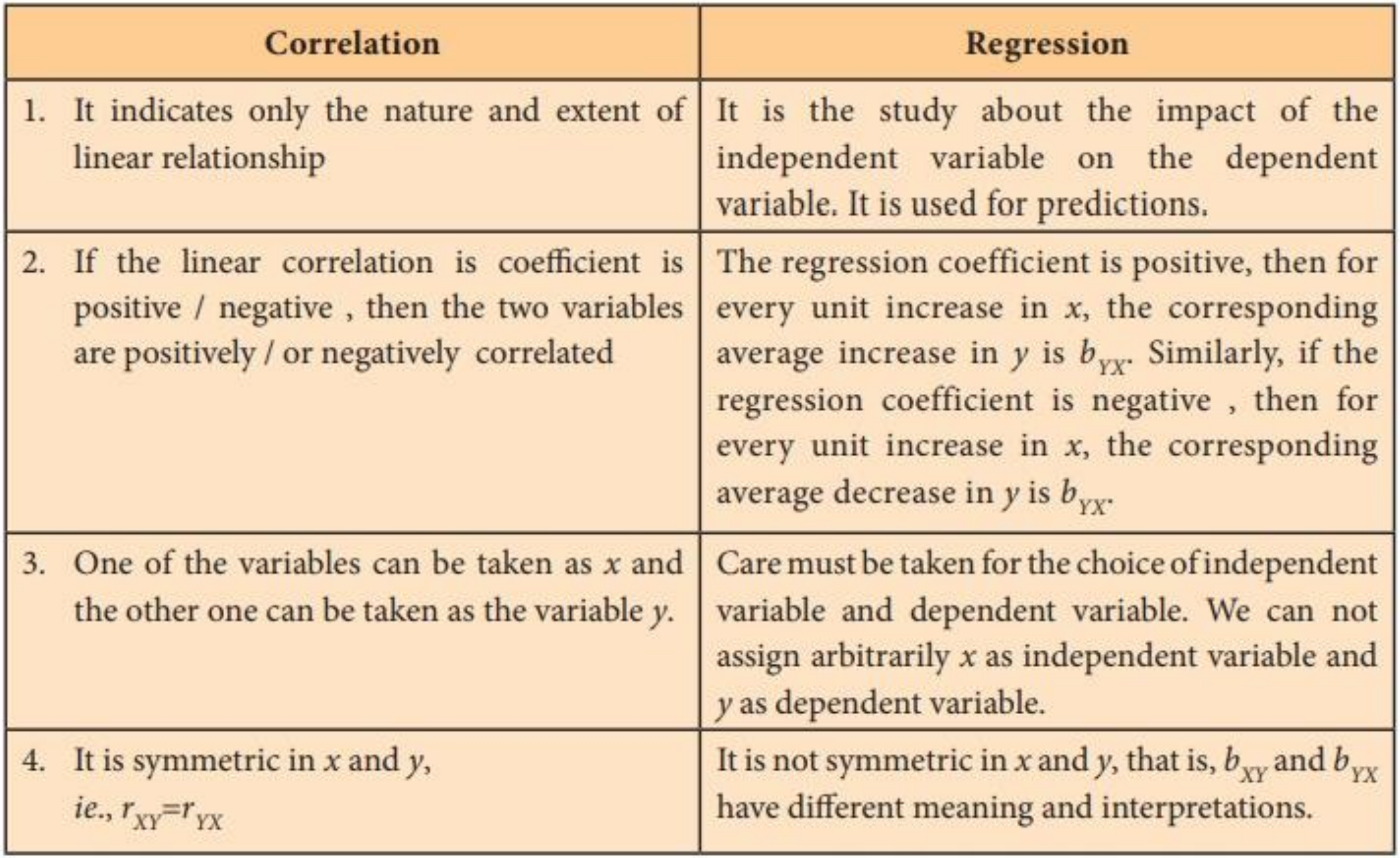
No correlation

What is Regression Analysis?

Regression analysis is a form of predictive modeling technique which investigates the relationship between a dependent (target) and independent variable (s) (predictor). This technique is used for forecasting, time series modeling and finding the causal effect relationship between the variables. For example, relationship between rash driving and number of road accidents by a driver is best studied through regression.



Unit –V :Data Analysis



Correlation	Regression
1. It indicates only the nature and extent of linear relationship	It is the study about the impact of the independent variable on the dependent variable. It is used for predictions.
2. If the linear correlation is coefficient is positive / negative , then the two variables are positively / or negatively correlated	The regression coefficient is positive, then for every unit increase in x, the corresponding average increase in y is b_{YX}. Similarly, if the regression coefficient is negative , then for every unit increase in x, the corresponding average decrease in y is b_{YX}.
3. One of the variables can be taken as x and the other one can be taken as the variable y.	Care must be taken for the choice of independent variable and dependent variable. We can not assign arbitrarily x as independent variable and y as dependent variable.
4. It is symmetric in x and y, *ie.*, $r_{XY}=r_{YX}$	It is not symmetric in x and y, that is, b_{XY} and b_{YX} have different meaning and interpretations.