Statistical Inference

Unit-2

Outline

- **Need of statistics in Data Science & Big Data Analytics**

- **Measures of Central Tendency**

Mean, Median, Mode, Mid-range

- **Measures of Dispersion**

Range, Variance, Mean Deviation, Standard Deviation

- **Bayes theorem, Basics and need of hypothesis and hypothesis testing, Pearson Correlation, Sample Hypothesis testing, Chi-Square Tests, t-test.**

contents

Unit-II	Statistical Inference	07 hrs
Contents	Need of statistics in Data Science, Measures of Central Tendency: Mean, Median, Mode, Mid-range. Measures of Dispersion: Range, Variance, Mean Deviation, Standard Deviation. Bayes theorem, Basics and need of hypothesis and hypothesis testing, Pearson Correlation, Sample Hypothesis testing, Chi-Square Tests, t-test.	
Case Study	For an employee dataset, create a measure of central tendency and its measure of dispersion for statistical analysis of given data	
CO	Apply statistics for Data Analytics	

Components of DATA SCIENCE

01 Data

Data is a collection of factual information.
Types: Structured Data and Unstructured Data

02 Big Data

Big Data is enormously big data sets, various V's such as, volume, variety, velocity, vision, value etc.

03 Machine Learning

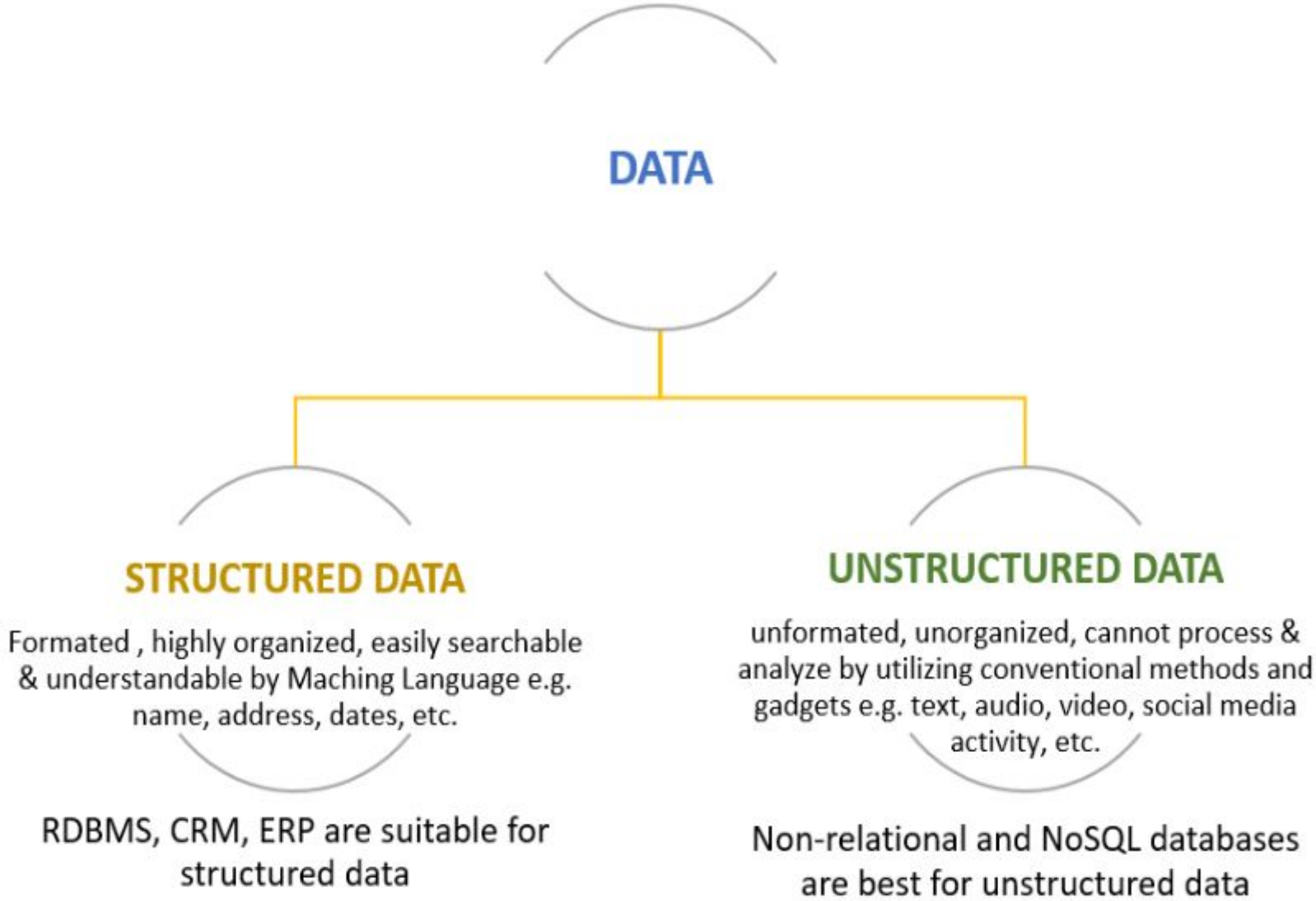
Further three types, supervised learning, unsupervised learning and reinforcement learning.

04 Statistics and Probability

The numerical foundation of data science is insights and likelihood.

05 Programming languages

Generally, data organization and investigation is finished by computer programming i.e. Python, R,


DATA
STRUCTURED DATA
Formated , highly organized, easily searchable & understandable by Maching Language e.g. name, address, dates, etc.
RDBMS, CRM, ERP are suitable for structured data
UNSTRUCTURED DATA
unformated, unorganized, cannot process & analyze by utilizing conventional methods and gadgets e.g. text, audio, video, social media activity, etc.
Non-relational and NoSQL databases are best for unstructured data

Big Data

big data image

Machine Learning

- **Machine Learning is the part of Data Science** which enables the system to process datasets autonomously without any human interference by utilizing various algorithms to work on massive volume of data generated and extracted from numerous sources.
- It makes prediction, analysis patterns and gives recommendations.
- Machine learning is frequently being used in fraud detection and client retention.

Example

- A social media platform i.e. Facebook is a decent example of machine learning implementation where fast and furious algorithms are used to gather the behavioral information of every user on social media and recommend them appropriate articles, multimedia files and much more according to their choice.
- **Machine learning is also the part of Artificial Intelligence** where the requisite information is achieved after utilizing various algorithms and techniques, such as Supervised and Un-supervised Machine Learning Algorithm

Need of statistics in Data Science

- The heart of data science lies the discipline of statistics, which plays a crucial role in extracting meaningful information from vast datasets.

Why Data science?

- It encompasses various techniques such as data mining, machine learning, and statistical analysis to uncover patterns, trends, and correlations within data.
- Data scientists utilize their expertise to generate actionable insights and drive evidence-based decision-making across industries.

What are Statistics?

- It is a branch of mathematics that deals with the collection, organization, analysis, interpretation, and presentation of data.
- It provides techniques and tools to summarize and make sense of data, enabling researchers and analysts to draw conclusions and make predictions.

- Statistics encompass both descriptive and inferential methods

The Role of Statistics in Data Science

1. Data Exploration and Preprocessing:

- Statistics helps in understanding
  - the distribution,
  - central tendency,
  - and variability of data.
- Exploratory data analysis techniques like mean, median, standard deviation, and correlation coefficients allow data scientists to gain initial insights and identify data quality issues.

2. Hypothesis Testing:

- Statistical hypothesis testing enables data scientists to validate assumptions, make inferences, and determine the significance of relationships between variables.
- It provides a framework to assess the validity of claims and draw conclusions based on evidence from sample data.

3. **Regression Analysis:**

- Regression analysis is a statistical technique that helps establish relationships between dependent and independent variables.
- Data scientists use regression models to make predictions, understand the impact of different factors, and uncover patterns within the data.

4. **Experimental Design**

- Statistics guides the design and implementation of experiments in data science.
- It enables researchers to select appropriate sample sizes, control for confounding variables, and analyze experimental results to draw accurate conclusions.

5. **Sampling Techniques**

- Statistics provides methods for selecting representative samples from large datasets.
- Sampling techniques such as random sampling, stratified sampling, and cluster sampling help data scientists obtain reliable insights without analyzing the entire population.

6. Data Visualization:

- Statistics enhances data visualization by providing graphical techniques to represent data effectively.
- Visualization tools like histograms, scatter plots, and box plots enable data scientists to communicate insights visually, making complex information more accessible to stakeholders.

7.Machine Learning and Statistics:

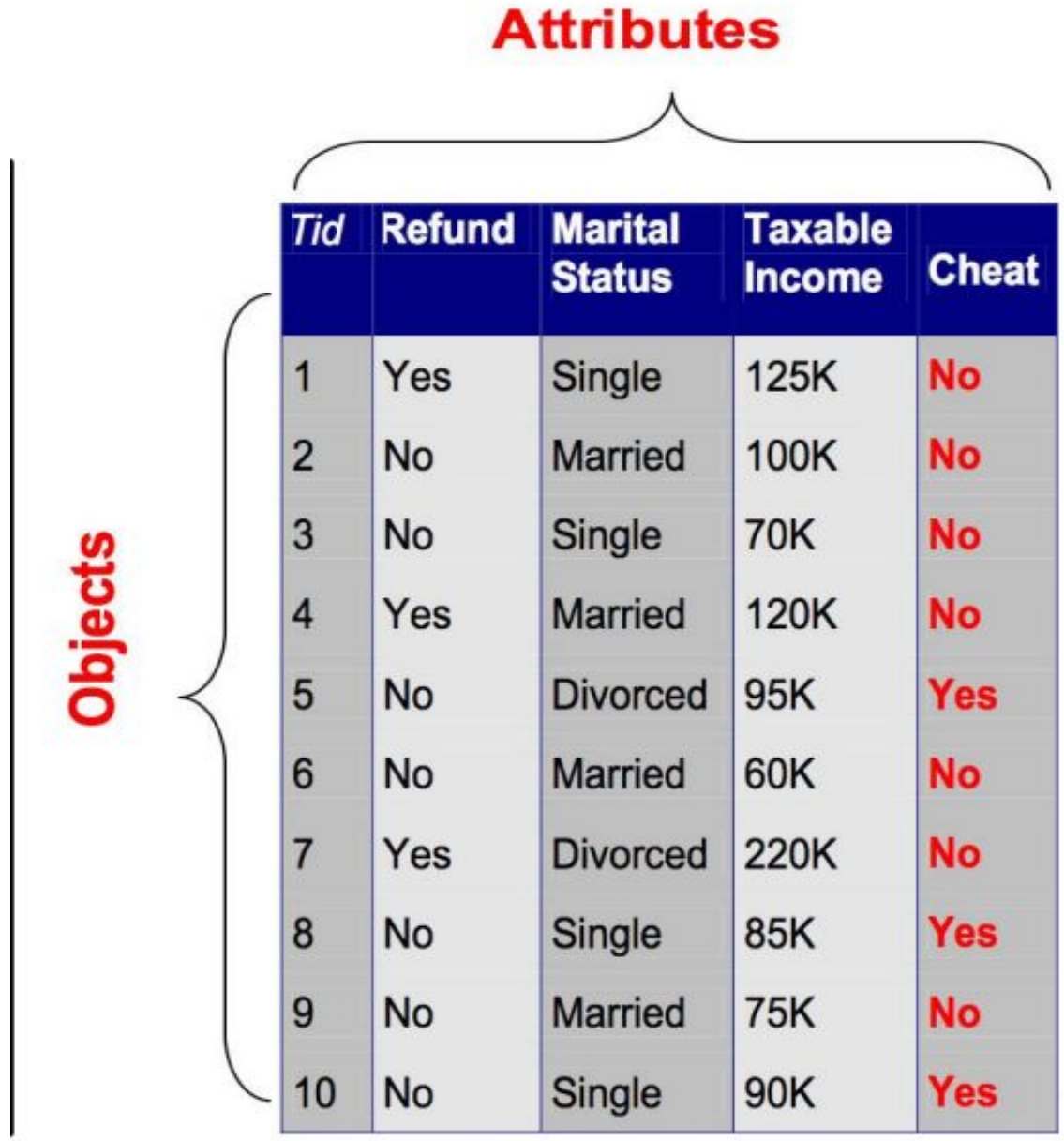
- Machine learning algorithms often leverage statistical concepts to make predictions and classify data.
- Techniques like decision trees, support vector machines, and neural networks incorporate statistical principles to analyze patterns and make accurate predictions based on training data.
- Statistics plays a fundamental role in model evaluation, validation, and the interpretation of machine learning results.

8. Ethical Considerations in Data Science:

- In the realm of data science, ethical considerations are of paramount importance. Statistics plays a critical role in ensuring ethical practices throughout the data science lifecycle.
- It helps address issues such as bias, fairness, and privacy when collecting, analyzing, and interpreting data.
- Statistical techniques like anonymization, de-identification, and statistical disclosure control help protect sensitive information and ensure that data-driven decisions are fair and unbiased.

Know your data

Statistics serve as the backbone of data science, providing the necessary tools and techniques to extract meaningful insights from data


Attributes
Objects
Tid
Refund
Marital Status
Taxable Income
Cheat
1 Yes Single 125K No
2 No Married 100K No
3 No Single 70K No
4 Yes Married 120K No
5 No Divorced 95K Yes
6 No Married 60K No
7 Yes Divorced 220K No
8 No Single 85K Yes
9 No Married 75K No
10 No Single 90K Yes

Attribute and object

- An attribute is a property or characteristic of an object

- A collection of attributes describe an object
- Objects are also known as records, points, cases, samples, entities, or instance

1. Univariate data –

- This type of data consists of **only one variable**.

Heights (in cm)	**164**	**167.3**	**170**	**174.2**	**178**	**180**	**186**

Bivariate data

- This type of data involves **two different variables**.
- The analysis of this type of data deals with causes and relationships and the analysis is done to find out the relationship among the two variables.
- Example of bivariate data can be temperature and ice cream sales in summer season.

TEMPERATURE(IN CELSIUS)	ICE CREAM SALES
20	2000
25	2500
35	5000
43	7800

Multivariate data

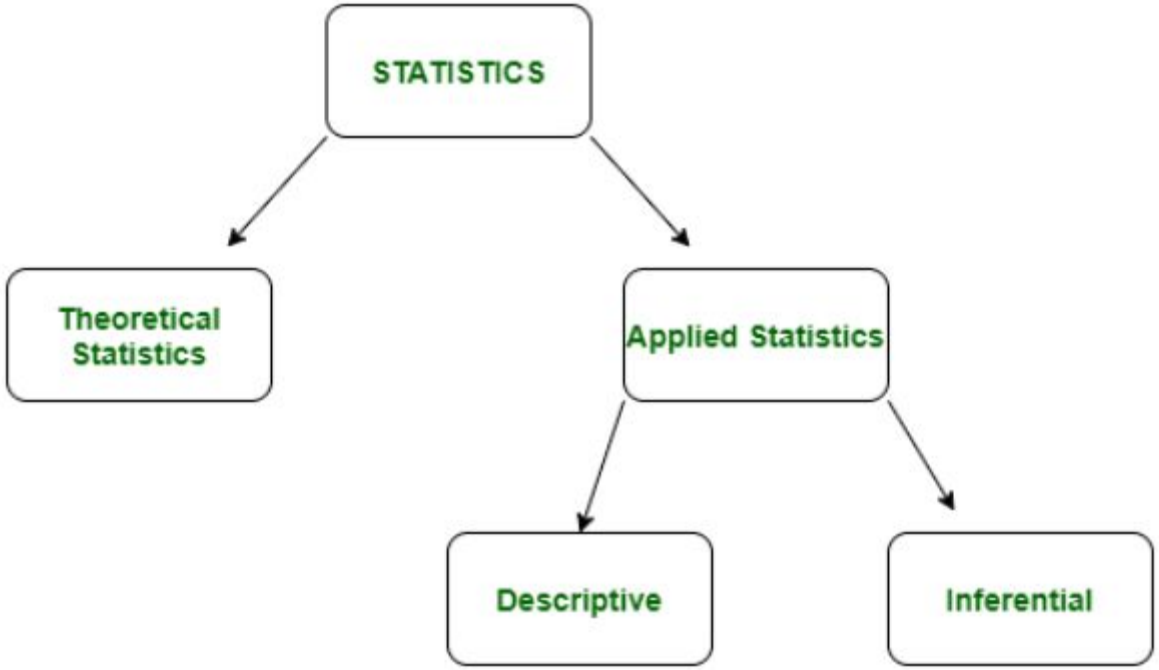
- When the data involves **three or more variables**, it is categorized under multivariate.
- Example of this type of data is suppose an advertiser wants to compare the popularity of four advertisements on a website, then their click rates could be measured for both men and women and relationships between variables can then be examined

Website_address, rating, gender

Introduction of Statistics

- Statistics is a branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data. It is basically a collection of quantitative data.
- https://www.biologyforlife.com/skew.html

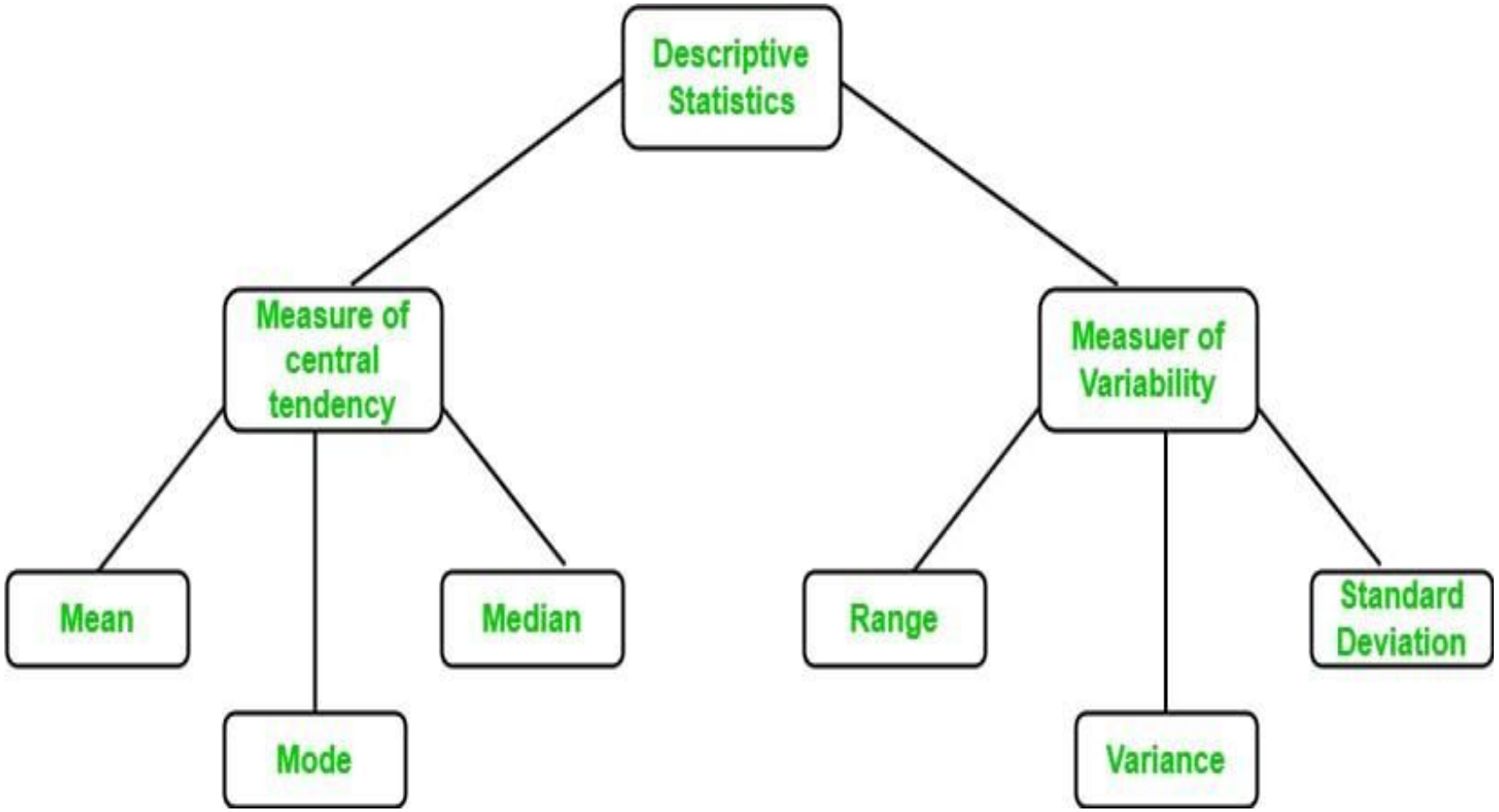
Types of Statistics



Descriptive statistics

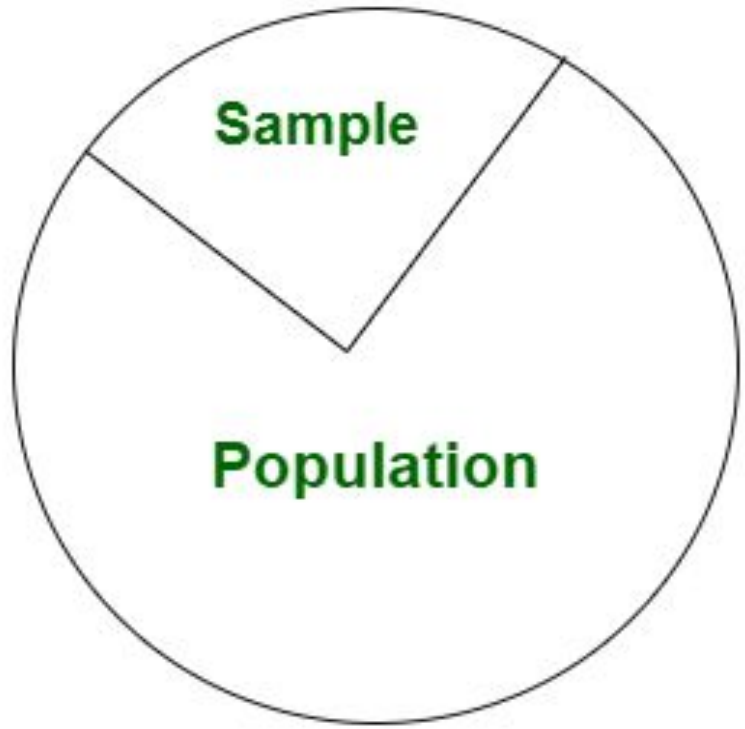
- It is a term given to the analysis of data that helps to describe, show and summarize data in a meaningful way.
- It is a simple way to describe our data
- Descriptive statistics is very important to present our raw data in effective/meaningful way using numerical calculations or graphs or tables
- This type of statistics is applied to already known data.

Types of Descriptive Statistics



inferential statistics

- Predictions are made by taking any group of data in which you are interested.
- It can be defined as a random sample of data taken from a population to describe and make inferences about the population.
- Any group of data that includes all the data you are interested in is known as population.
- It basically allows you to make predictions by taking a small sample instead of working on the whole population.

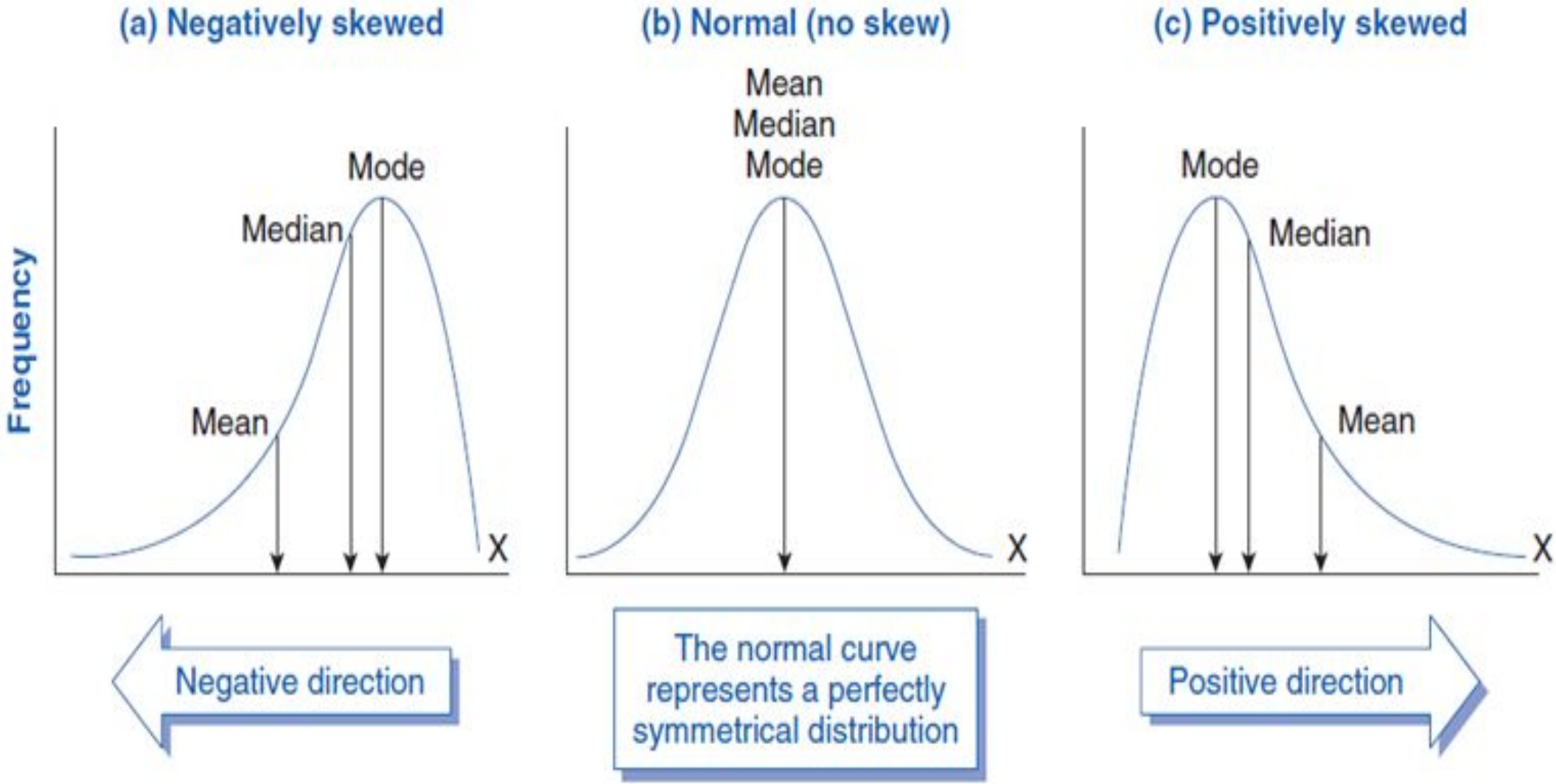

Sample
Population

S.No.	Descriptive Statistics	Inferential Statistics
1.	It gives information about raw data which describes the data in some manner.	It makes inferences about the population using data drawn from the population.
2.	It helps in organizing, analyzing, and to present data in a meaningful manner.	It allows us to compare data, and make hypotheses and predictions.
3.	It is used to describe a situation.	It is used to explain the chance of occurrence of an event.
4.	It explains already known data and is limited to a sample or population having a small size.	It attempts to reach the conclusion about the population.
5.	It can be achieved with the help of charts, graphs, tables, etc.	It can be achieved by probability.

Measures of Central Tendency in Statistics

- **Central Tendencies in Statistics** are the numerical values that are used to represent mid-value or central value a large collection of numerical data.
- These obtained numerical values are called **central** or **average values** in Statistics.

- The representative value of a data set, generally the central value or the most occurring value that gives a general idea of the whole data set is called Measure of Central Tendency.


(a) Negatively skewed
(b) Normal (no skew)
(c) Positively skewed
Frequency
Mode
Median
Mean
Mean
Median
Mode
Mode
Median
Mean
X
X
X
Negative direction
The normal curve represents a perfectly symmetrical distribution
Positive direction

Mean

- Mean in general terms is used for the arithmetic mean of the data, but other than the arithmetic mean there are geometric mean and harmonic mean as well that are calculated using different formulas.

Mean for Ungrouped Data

- Arithmetic mean ($\bar{x}$) is defined as the sum of the individual observations (x_i) divided by the total number of observations N.
- In other words, the mean is given by the sum of all observations divided by the total number of observations.
- $\bar{x} = \frac{\sum x_i}{N}$
- ***Mean = Sum of all Observations ÷ Total number of Observations***

Example

- **If there are 5 observations,** we have the AQI values for five cities: **60, 75, 80, 90, 100**. Mean is-----

Mean for Grouped Data

- Mean ($\bar{x}$) is defined for the grouped data as the sum of the product of observations (x_i) and their corresponding frequencies (f_i) divided by the sum of all the frequencies (f_i).
- $\bar{x} = \frac{\sum f_i x_i}{\Sigma f_i}$

Example

x_i	4	6	15	10	9
f_i	5	10	8	7	10

Solution

$\bar{x} = (4\times5 + 6\times10 + 15\times8 + 10\times7 + 9\times10) \div (5 + 10 + 8 + 7 + 10)$

$\Rightarrow \bar{x} = (20 + 60 + 120 + 70 + 90) \div 40$

$\Rightarrow \bar{x} = 360 \div 40$

$\Rightarrow \bar{x} = 9$

Advantage and Limitations of Mean

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

Limitations:

- 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 and skewed distributions.

Median

- The median is the middle value in distribution when the values are arranged in ascending or descending order.
- The median divides the distribution in half (there are 50% of observations on either side of the median value).
- In a distribution with an odd number of observations, the median value is the middle value.

- Looking at the retirement age distribution (which has 11 observations), the median is the middle value, which is 57 years:
- 54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60

- When the distribution has an even number of observations, the median value is the mean of the two middle values.
- In the following distribution, the two middle values are 56 and 57, therefore the median equals 56.5 years:
- 52, 54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60

Advantage and Limitations of 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

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

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

- Frequency Table:

Age	Frequency
54	3
55	1
56	1
57	2
58	2
60	2

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.

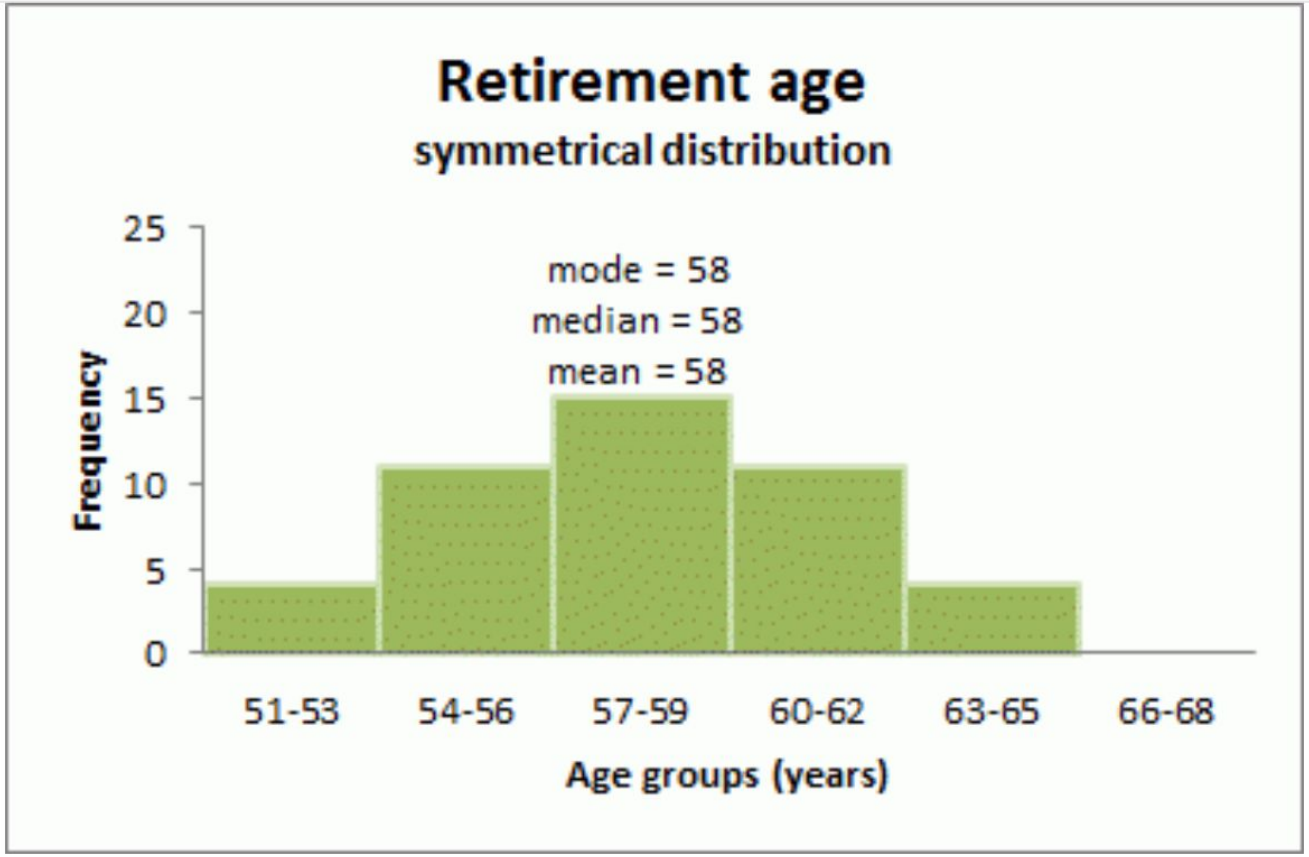
Limitations

- In some distributions, the mode may not reflect the center of the distribution very well.
- When the distribution of retirement age is ordered from lowest to highest value, it is easy to see that the center of the distribution is 57 years, but the mode is lower, at 54 years.
- 54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60

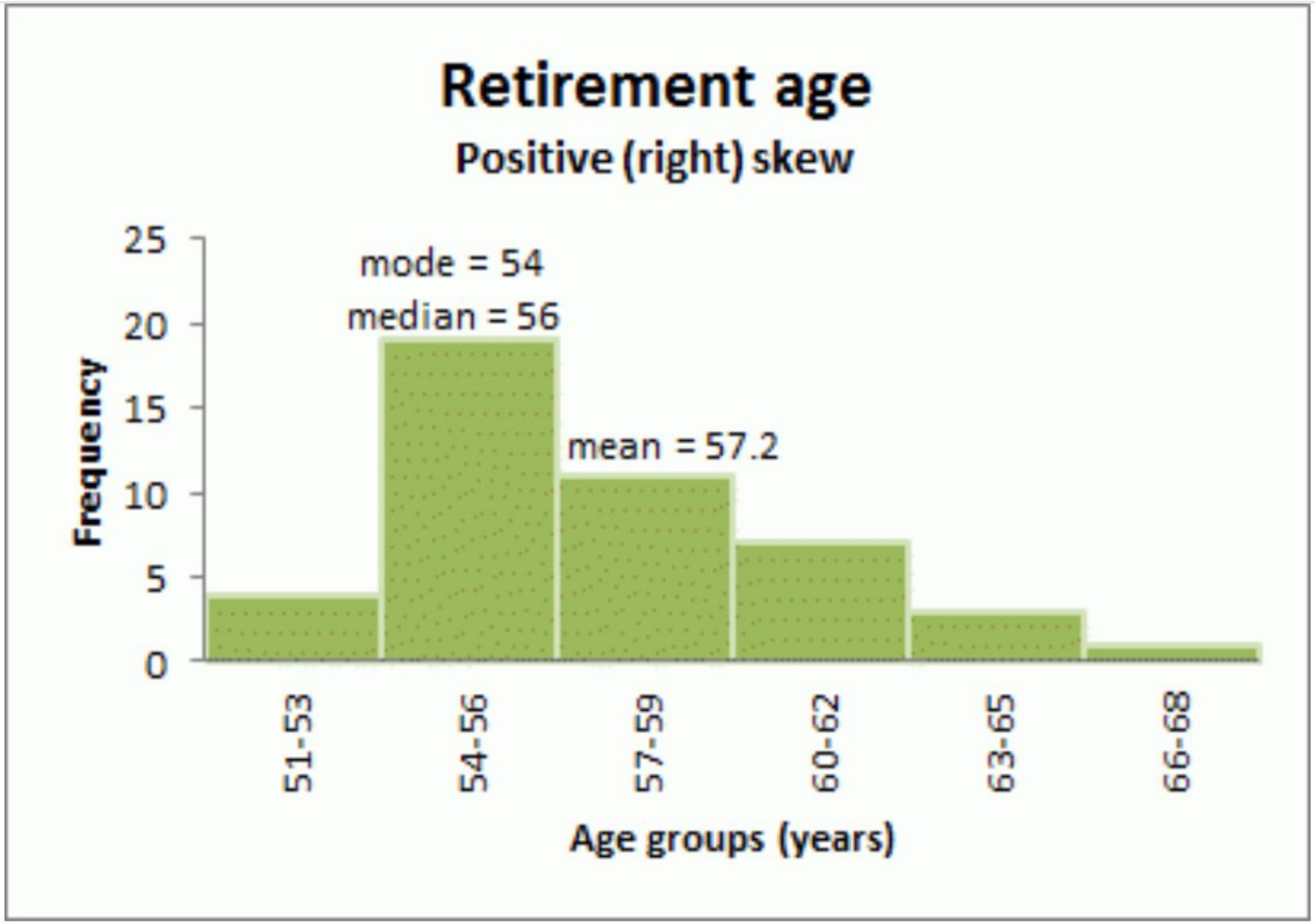
- It is also possible for there to be more than one mode for the same distribution of data, (bi-modal, or multi-modal).
- The presence of more than one mode can limit the ability of the mode in describing the center or typical value of the distribution because a single value to describe the center cannot be identified
- In some cases, particularly where the data are continuous, the distribution may have no mode at all (i.e. if all values are different).

Impact of shape of distribution on measures of central tendency

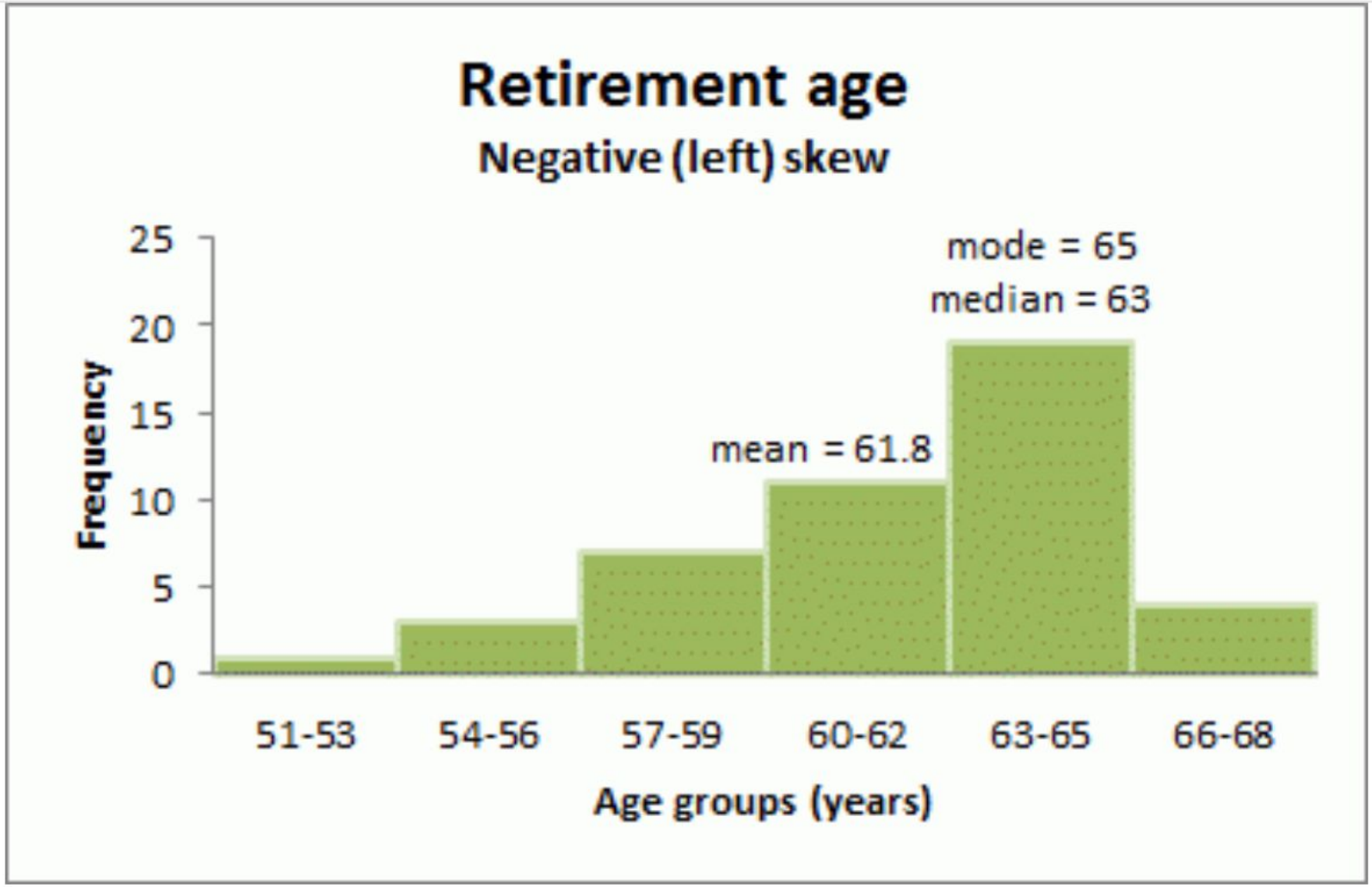
Retirement age: Symmetrical distribution



Retirement age: Positive (right) skew



Retirement age: Negative (left) skew



Outliers influence on measures of central tendency

- Outliers are extreme, or atypical data value(s) that are notably different from the rest of the data.
- It is important to detect outliers within a distribution, because they can alter the results of the data analysis.
- The mean is more sensitive to the existence of outliers than the median or mode.

Example

- Consider the initial retirement age dataset again, with one difference; the last observation of 60 years has been replaced with a retirement age of 81 years.
- This value is much higher than the other values, and could be considered an outlier.
- However, it has not changed the middle of the distribution, and therefore the median value is still 57 years.
- 54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 81

MEASURES OF DISPERSION

- Different measures of central tendency, discussed earlier, but t it gives no idea about the nature of scatter or spread.
- For example, the observations 10, 30 and 50 have mean 30 while the observations 28, 30, 32 also have mean 30.
- Both the distributions are spread around 30. But it is observed that the variability among units is more in the first than in the second.
- In other words, there is greater variability or dispersion in the first set of observations in comparison to other.
- Measure of dispersion is calculated **to get an idea about the variability in the data.**

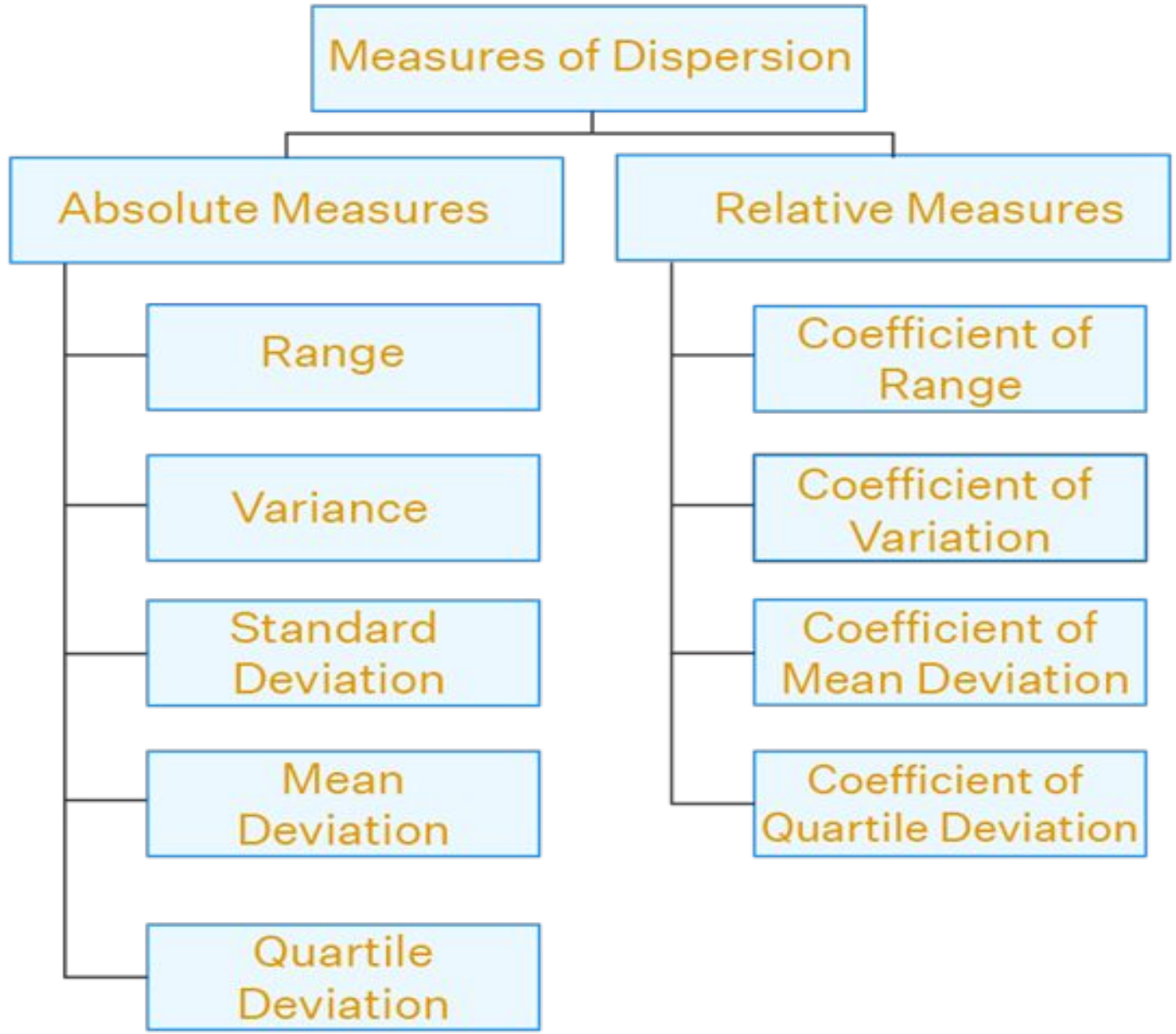
Measure of Dispersion

Spread

- The degree to which numerical data tend to spread about an average value is called the variation or dispersion of data.
- Actually, there are two basic kinds of a measure of dispersion (i) Absolute measures and (ii) Relative measures.
- The absolute measures of dispersion are used to measure the variability of a given data expressed in the same unit,
- while the relative measures are used to compare the variability of two or more sets of observations

Types of Measures of Dispersion

cuemath
THE MATH EXPERT



Objectives

- conceptualize dispersion
- explain the utility of a measure of dispersion
- explain the properties of a good measure of dispersion
- explain methods of calculation of different types of measures of dispersion along with their merits and demerits
- solve the numerical problems related to the measures of dispersion.

Following are the different measures of dispersion:

- Range
- Quartile Deviation
- Mean Deviation
- Standard Deviation
- Variance

Significance of Measures of Dispersion

- One of real time purpose of measuring variation, the variation in the quality of product in the process form of production can be checked by quality control department by identifying the reason for the variations in the quality of product.
- Thus, measurements of dispersion are helpful to control the causes of variation.

RANGE

- Range is the simplest measure of dispersion.
- It is defined as the difference between the maximum value of the variable and the minimum value of the variable in the distribution
- $R = X_{max} - X_{min}$
- Example, Find the range of the distribution 6, 8, 2, 10, 15, 5, 1, 13

Range for Grouped Data

- **Example: 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

- *For Largest Value: Taking the higher limit of Highest Class = 40*
- *For Smallest Value: Taking the lower limit of Lowest Class = 0*

  Range = 40 – 0
- *Thus, the range of the given data set is, Range = 40*

Practice

Class	0-10	10-20	20-30	30-40	40-50
Frequency	2	6	12	7	3

Merits and Demerits of Range

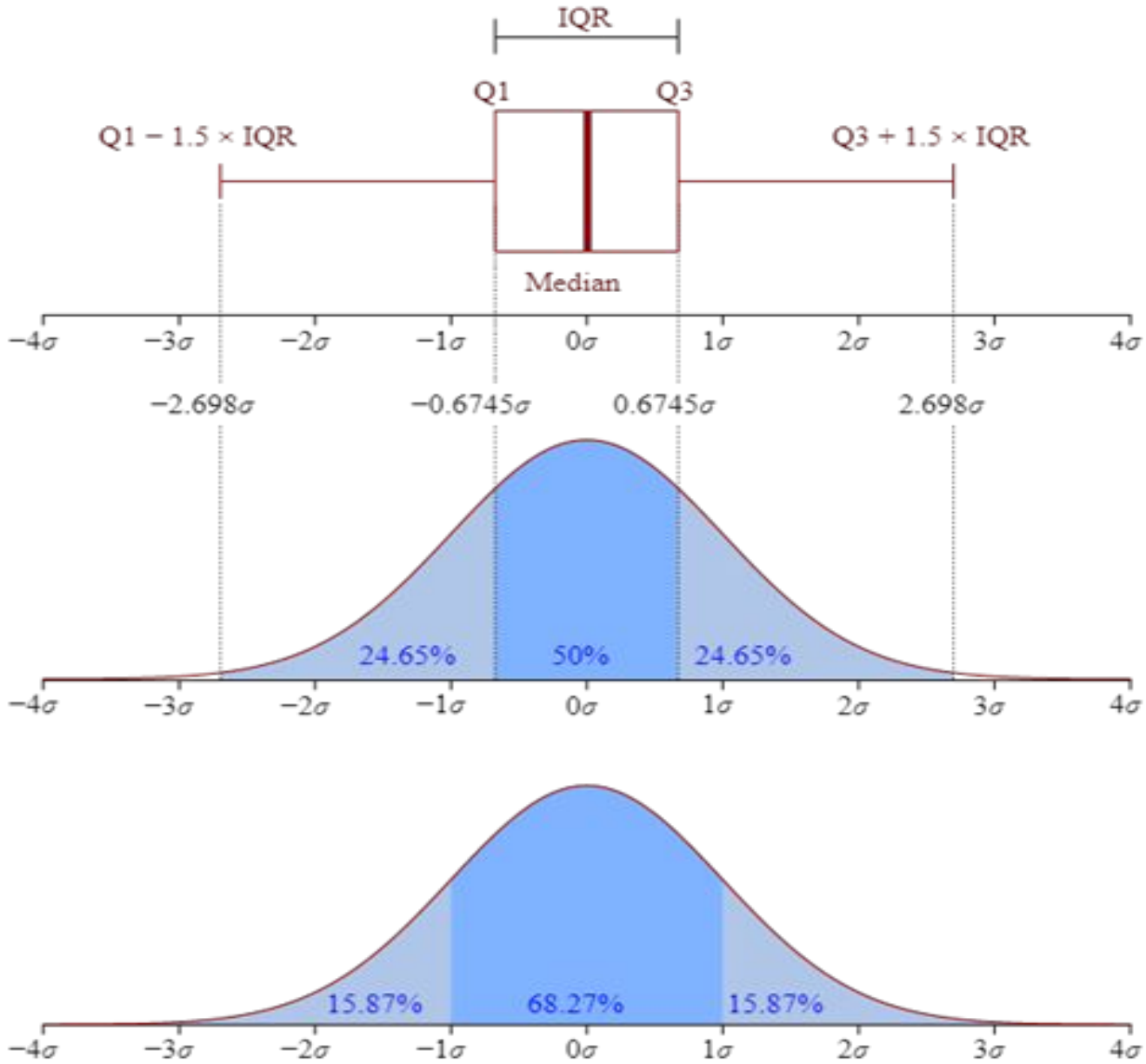
- It can be visually obtained since one can detect the largest and the smallest observations easily and can take the difference without involving much calculations

Demerit:

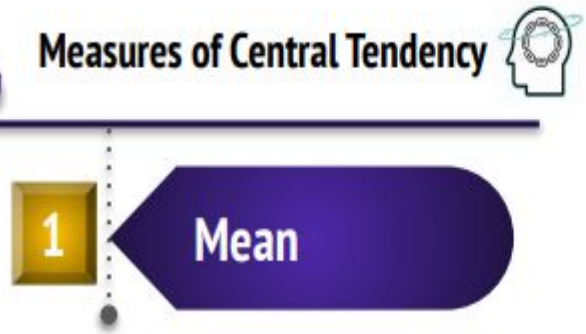
It utilizes only the maximum and the minimum values of variable in the series and gives no importance to other observations

Quartile Deviation (Semi-Interquartile Range)

- **Quartile Deviation** is a measure of statistical dispersion, indicating the spread of the middle 50% of a dataset. It is calculated as
- Quartile Deviation=$\frac{Q3-Q1}{2}$


IQR
Q1
Q3
Q1 − 1.5 × IQR
Q3 + 1.5 × IQR
Median
−4σ
−3σ
−2σ
−1σ
0σ
1σ
2σ
3σ
4σ
−2.698σ
−0.6745σ
0.6745σ
2.698σ
24.65%
50%
24.65%
15.87%
68.27%
15.87%

Measures of Central Tendency



1 Mean

$2 + 2 + 5 + 6 + 7 + 8 = 30$

$30 \div 6 = 5$

The mean number is

5

Measures of Central Tendency

1 Mean

- The mean represents the average value of the dataset.
- n is sample size and N is population size.
- $$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

$$\mu = \frac{\sum x}{n}$$

Measures of Central Tendency

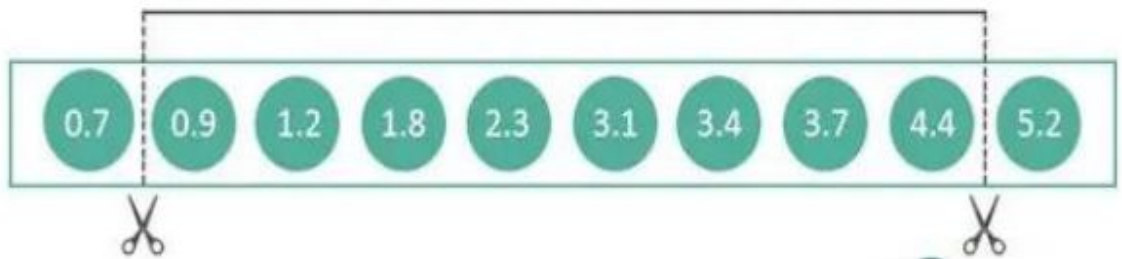
1 Mean

Trimmed Mean

"Trimmed mean is a central tendency measure that cuts down the smallest and highest values before applying the standard averaging formula for greater accuracy."

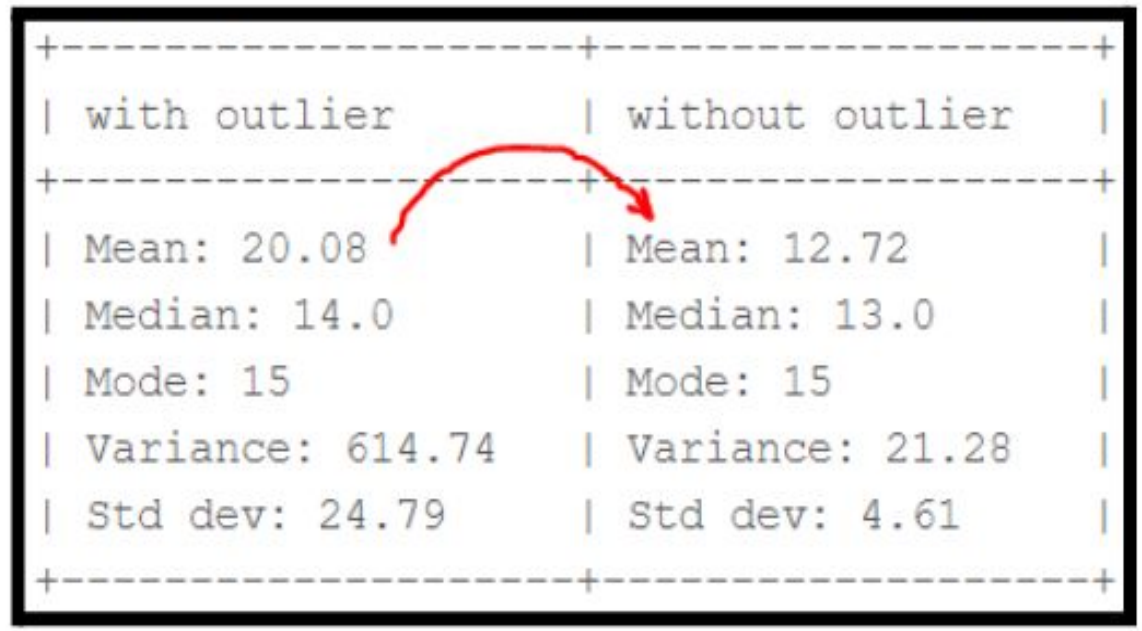
Loss of Information

10% Trimmed Mean = 2.6



- 10% samples will be cut down from each side
- Three commonly applied trim percentages, i.e., 5%, 10%, and 20%

- Mean’ is the only measure of central tendency that is affected by the outliers which in turn impacts Standard deviation.
- Example: Consider a small dataset, sample= [15, 101, 18, 7, 13, 16, 11, 21, 5, 15, 10, 9]. By looking at it, one can quickly say ‘101’ is an outlier that is much larger than the other values.


with outlier
without outlier
Mean: 20.08
Median: 14.0
Mode: 15
Variance: 614.74
Std dev: 24.79
Mean: 12.72
Median: 13.0
Mode: 15
Variance: 21.28
Std dev: 4.61

Measures of Central Tendency

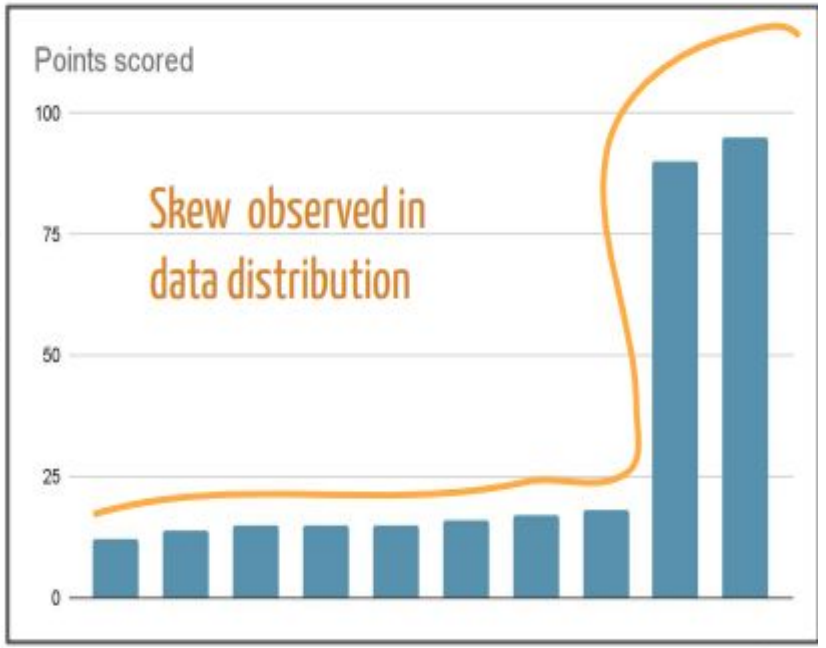
1 **Mean**

- Sort in descending order

12	14	15	15	15	16	17	18	90	95

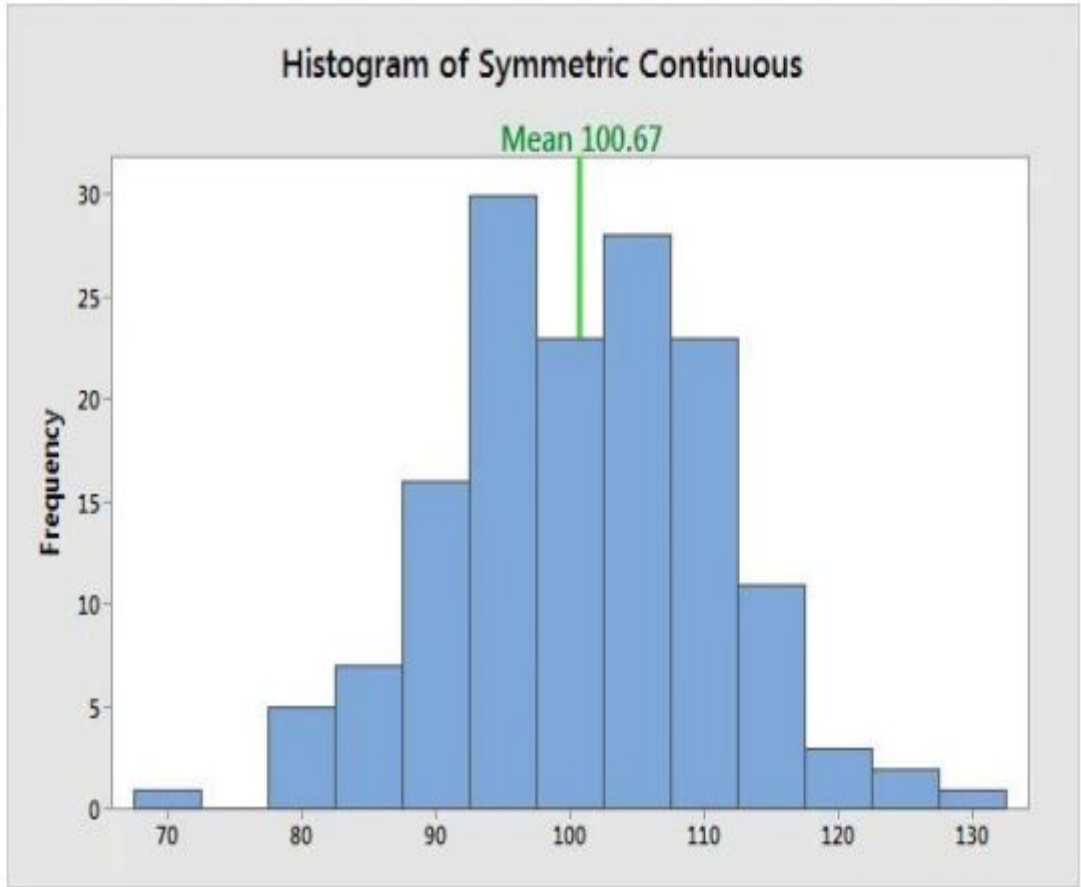
When mean fails...

Staff	1	2	3	4	5	6	7	8	9	10
Salary	15k	18k	16k	14k	15k	15k	12k	17k	90k	95k



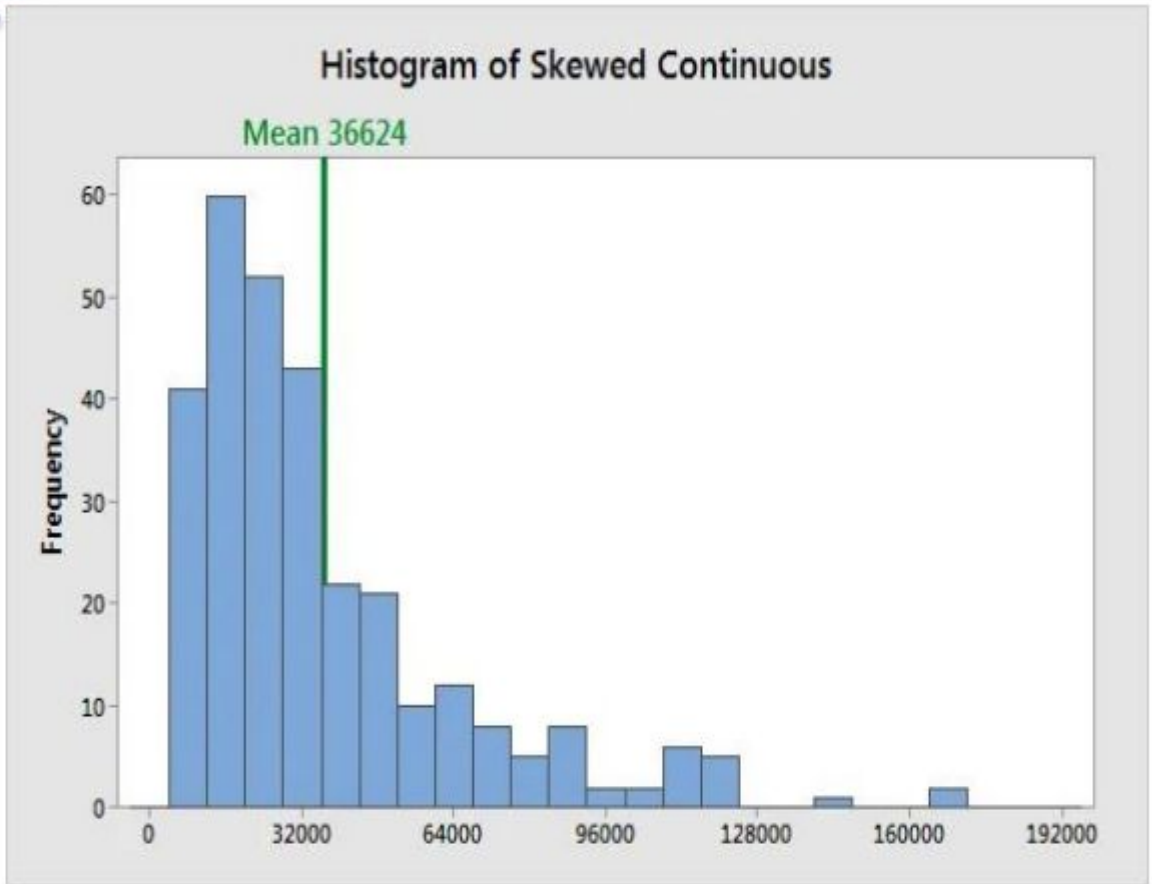
Measures of Central Tendency

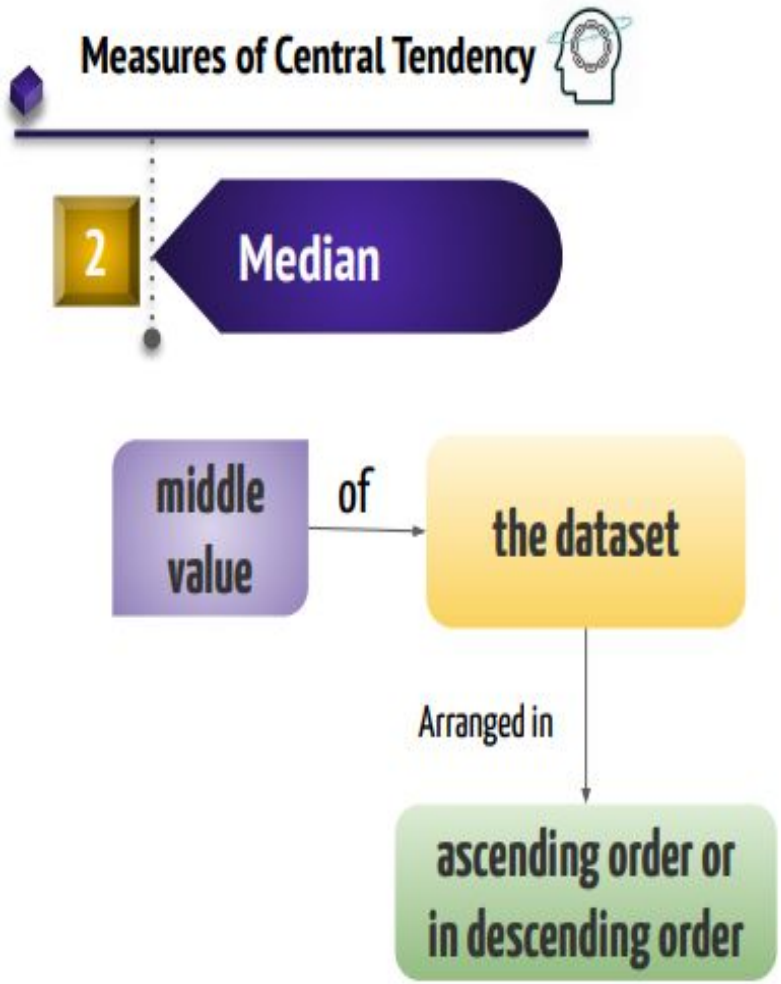
1 **Mean**

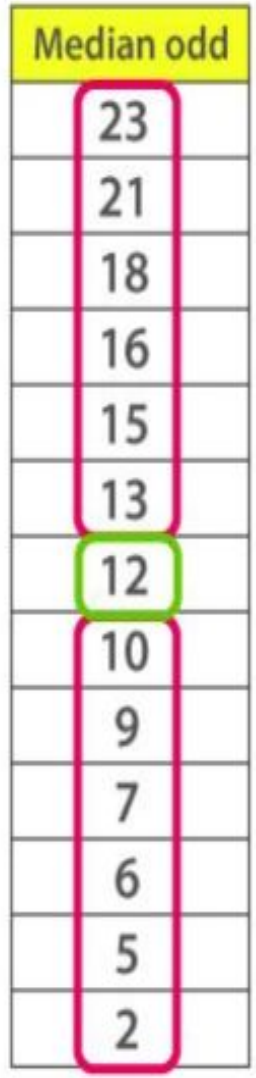


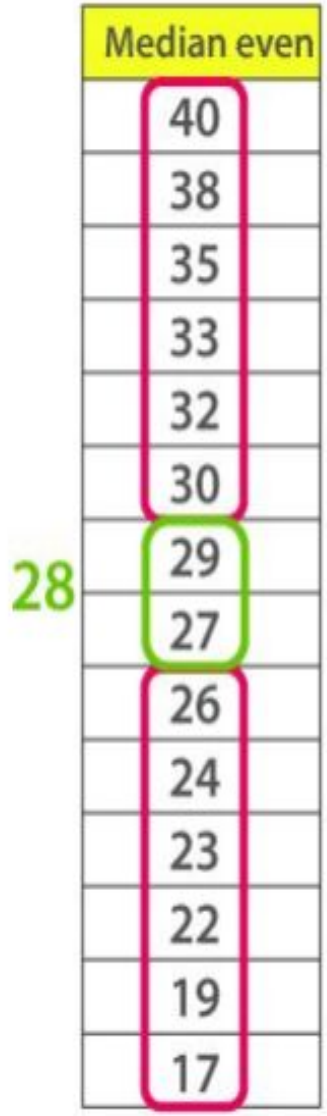
Measures of Central Tendency

1. Mean




Measures of Central Tendency
2
Median
middle value
of
the dataset
Arranged in
ascending order or in descending order


Median odd
23
21
18
16
15
13
12
10
9
7
6
5
2


Median even
40
38
35
33
32
30
29
27
26
24
23
22
19
17

28

Measures of Central Tendency

2 Median

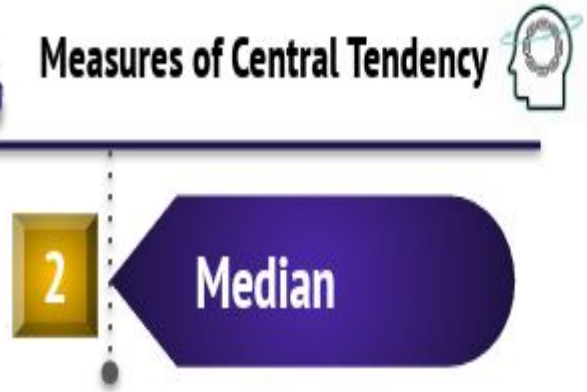
Odd number of Samples:

$$\text{Median} = \textit{value of } (n+1/2)\textit{th observation}$$

Even number of Samples:

$$\text{Median} = \frac{\textit{value of } \left(\frac{n}{2}\right)^{th} \textit{observation} + \textit{value of } \left(\frac{n}{2}+1\right)^{th} \textit{observation}}{2}$$

Measures of Central Tendency



2 Median

The number of rooms in the seven five stars hotel in Chennai city is 71, 30, 61, 59, 31, 40 and 29. Find the median number of rooms:

Step 1	Arrange the data in ascending order : 29, 30, 31, 40, 59, 61,71
Step 2	n = 7 (odd)
Step 3	Median = 7+1 / 2 = 4th positional value
Step 4	Median = 40 rooms

Measures of Central Tendency

2 Median

Median for Discrete grouped data:

i. Calculate the cumulative frequencies

ii. Find $(N+1)/2$, where $N=\Sigma f$=total frequencies

iii. Identify the cumulative frequency just greater than $(N+1)/2$

iv. The value of x corresponding to that cumulative frequency is the $(N+1)/2$ median

Measures of Central Tendency

2 Median

The following data are the weights of students in a class. Find the median weights of the students

Weight(kg)	10	20	30	40	50	60	70
Number of Students	4	7	12	15	13	5	4

Weight (kg) x	Frequency f	Cumulative Frequency $c.f$
10	4	4
20	7	11
30	12	23
40	15	38
50	13	51
60	5	56
70	4	60
Total	N = 60	

Weight (kg) x	Frequency f	Cumulative Frequency $c.f$
10	4	4
20	7	11
30	12	23
40	15	38
50	13	51
60	5	56
70	4	60
Total	N = 60	

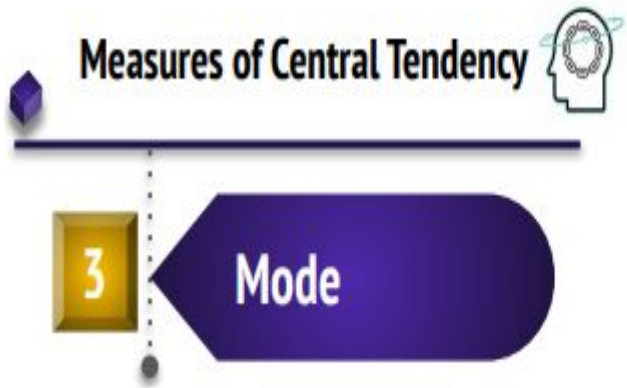
Step 1 N= 60

Step 2 (N+1)/2= (60+1)/2= 30.5

Step 3 Cumulative frequency >30.5 is 38

Step 4 Value of x corresponding to 38 is 40

Step 5 The median weight of students is 40



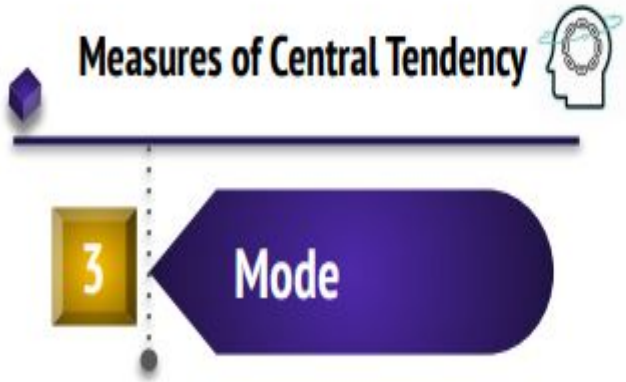
The following are the marks scored by 20 students in the class. Find the mode

90, 70, 50, 30, 40, 86, 65, 73, 68, 90, 90, 10, 73, 25, 35, 88, 67, 80, 74, 46

The marks 90 occurs the maximum number of times

Mode=90

Measures of Central Tendency

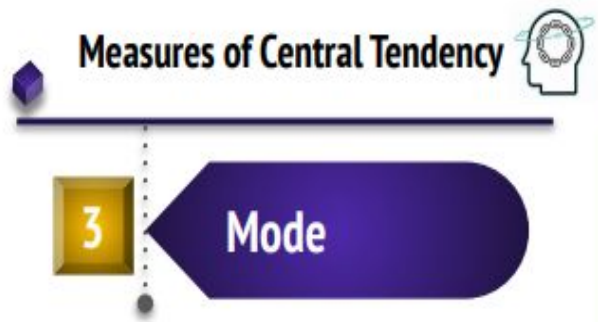


A doctor who checked 9 patients' sugar level is given below. Find the mode value of the sugar levels

80, 112, 110, 115, 124, 130, 100, 90, 150, 180

Each values occurs only once

there is no mode



Compute mode value for the following observations.

7, 10, 12, 10, 19, 2, 11, 3, 12

the observations 10 and 12 occurs twice in the data set

the modes are 10 and 12

Measures of Central Tendency

3 Mode

Mode for Continuous data

$$\text{Mode} = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times c$$

Modal class is the class which has maximum frequency.

f1 = frequency of the modal class

f0 = frequency of the class preceding the modal class

f2 = frequency of the class succeeding the modal class

c = width of the class limits

Measures of Central Tendency

3 Mode

Income (`)	0-100	100-200	200-300	300-400	400-500	500-600	600-700
No.of persons	5	7	12	18	16	10	5

The given data relates to the daily income of families in an urban area. Find the modal income of the families.

Income (`)	No.of persons (f)	
0-100	5	
100-200	7	
200-300	12	f_0
300-400	18	f_1
400-500	16	f_2
500-600	10	
600-700	5	

Income (`)	No.of persons (f)	
0-100	5	
100-200	7	
200-300	12	f_0
300-400	18	f_1
400-500	16	f_2
500-600	10	
600-700	5	

Step 1 Highest Frequency is 18. Modal class is 300-400

Step 2 l= lower boundary of 300-400 = 300

Step 3 f1= frequency of 300-400 =18

Step 4 f0= frequency of 200-300 =12

Step 5 f2= frequency of 440-500 =16

Step 6

Mode= 300+ (18-12)/(2*18-12-16)*100

=300+6/(36-28) *100

=300+ 600/8 = 300+75= **375**

Data Attributes and Measure of Central Tendency

Levels of measurement

Scale	Mode	Median	Mean
Nominal	√		
Ordinal	√	√	
Interval	√	√	√
Ratio	√	√	√

Example		
Problem	**Find the range and midrange for the following set of numbers: 2, 4, 7, 10, 14, 35.**	
	range: 35 – 2 = 33	Subtract the least value from the greatest value to find the range.
	midrange: $\frac{35+2}{2} = \frac{37}{2} = 18.5$	Add together the greatest value and the least value and divide by 2.
Answer	The range is 33. The midrange is 18.5.	

Relationship among mean, median and Mode

No of days spend in training	
Team 1	4
Team 2	5
Team 3	6
Team 4	6
Team 5	6
Team 6	7
Team 7	7
Team 8	7
Team 9	7
Team 10	7
Team 11	7
Team 12	8
Team 13	8
Team 14	8
Team 15	9
Team 16	10

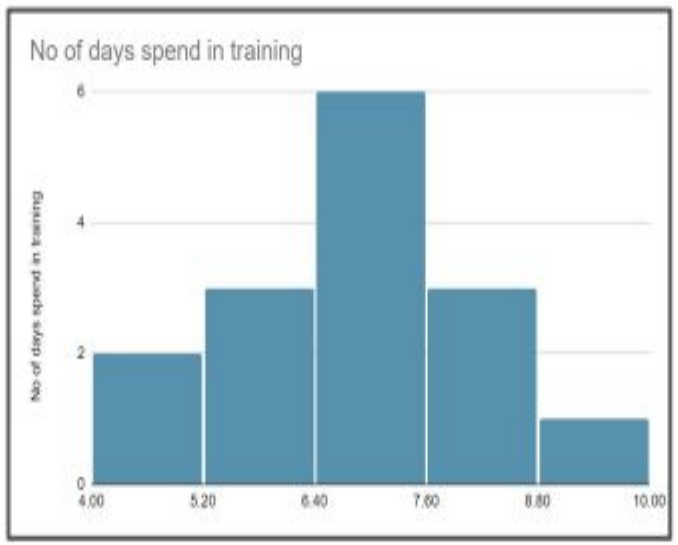
No of days spend in training	
Team 1	4
Team 2	5
Team 3	6
Team 4	6
Team 5	6
Team 6	7
Team 7	7
Team 8	7
Team 9	7
Team 10	8

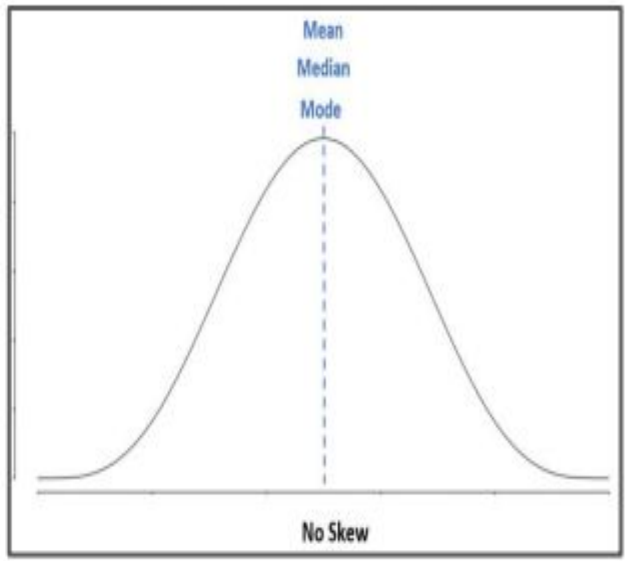
No of days spend in training	
Team 1	6
Team 2	7
Team 3	7
Team 4	7
Team 5	7
Team 6	8
Team 7	8
Team 8	8
Team 9	9
Team 10	10

No of days spend in training	
Team 1	4
Team 2	5
Team 3	6
Team 4	6
Team 5	6
Team 6	7
Team 7	7
Team 8	7
Team 9	7
Team 10	7
Team 11	7
Team 12	8
Team 13	8
Team 14	8
Team 15	9
Team 16	10

Mean	7
Median	7
Mode	7

Mean= Median=Mode



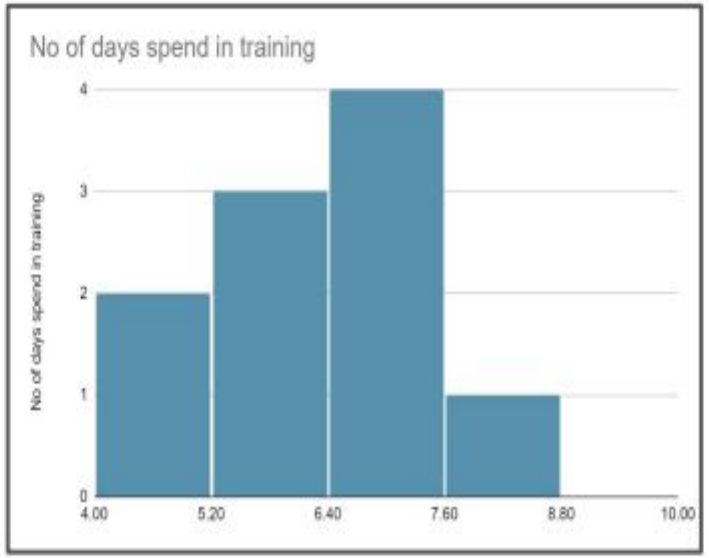


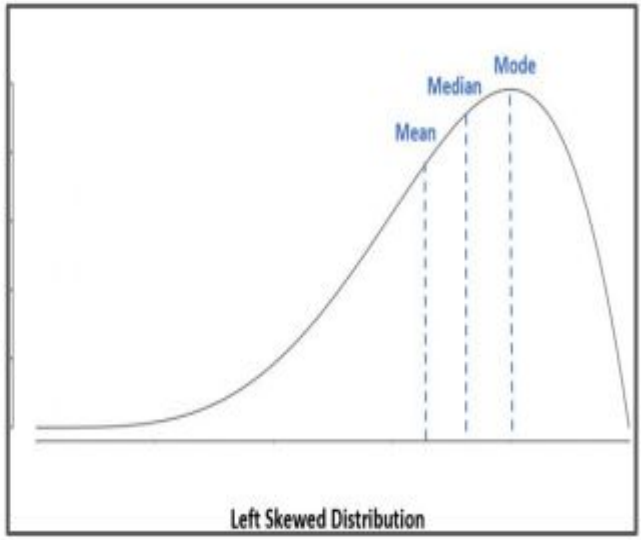
Symmetric Data Distribution

No of days spend in training	
Team 1	4
Team 2	5
Team 3	6
Team 4	6
Team 5	6
Team 6	7
Team 7	7
Team 8	7
Team 9	7
Team 10	8

Mean	6.3
Median	6.5
Mode	7

Mean < Median < Mode



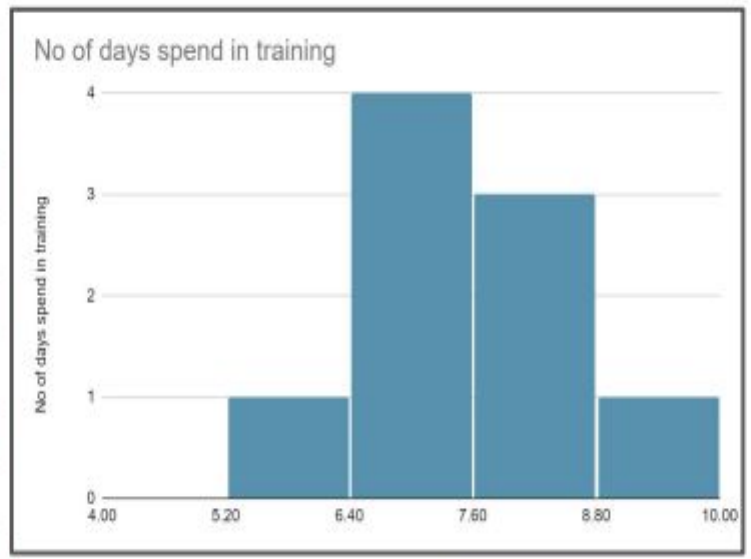


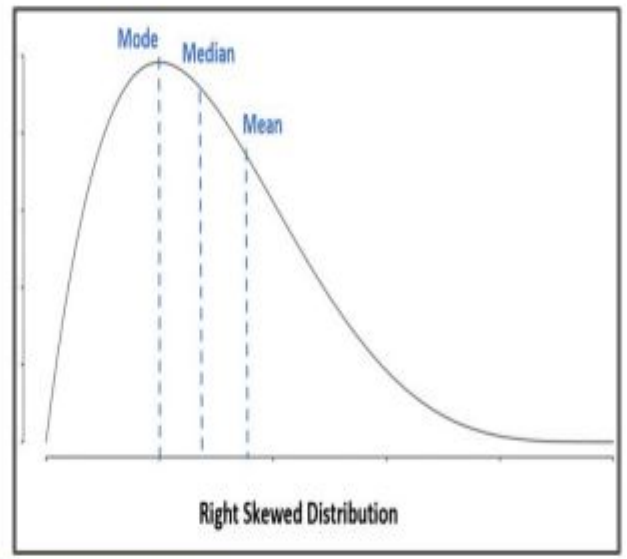
Left Skew data distribution

No of days spend in training	
Team 1	6
Team 2	7
Team 3	7
Team 4	7
Team 5	7
Team 6	8
Team 7	8
Team 8	8
Team 9	9
Team 10	10

Mean	7.7
Median	7.5
Mode	7

Mode < Median < Mean





Right Skew data distribution

Measures of Central Tendency

Empirical Relationship among mean, median and mode

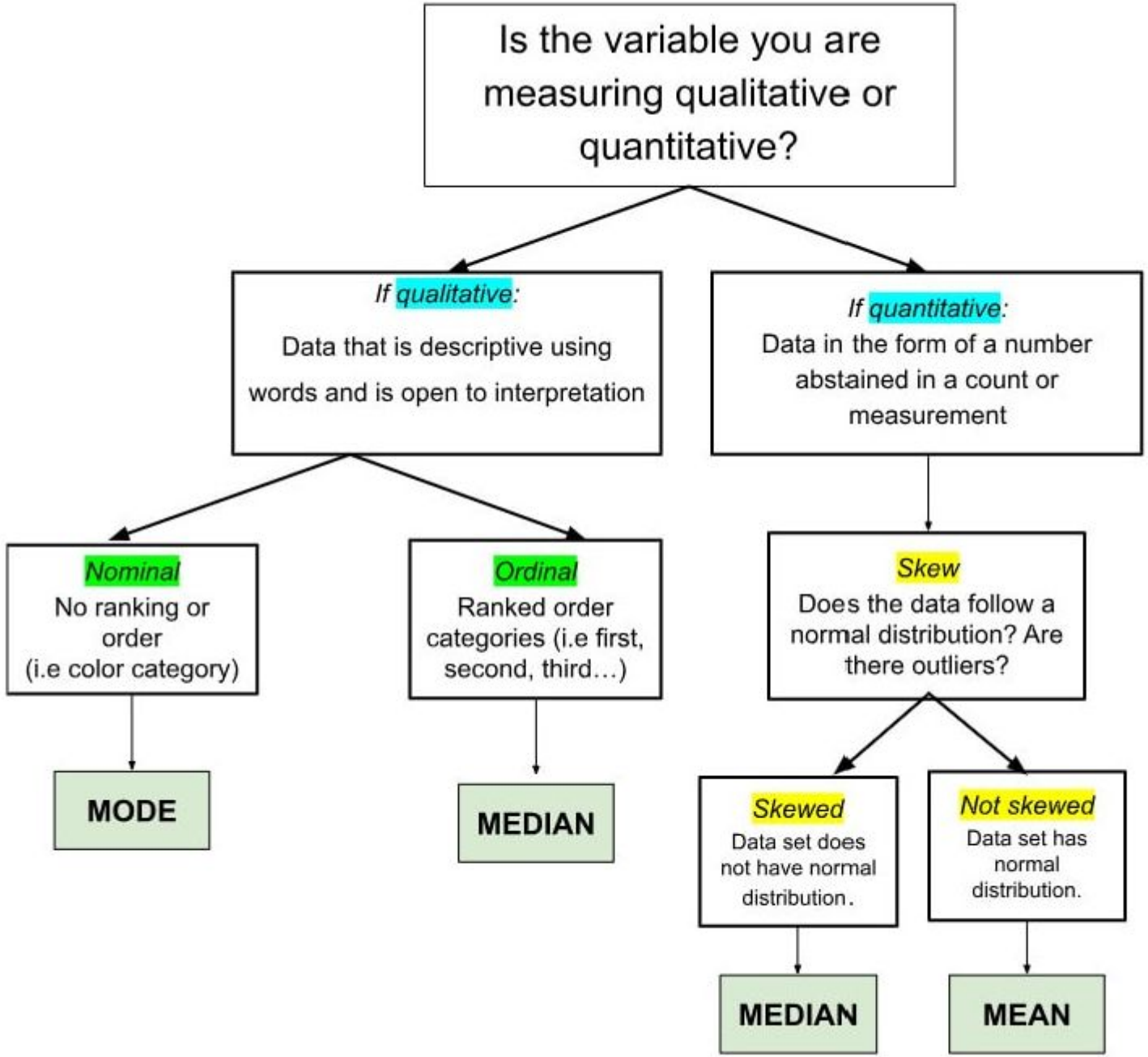
Mean- Median= ⅓(Mean-Mode)

7.7 -7.5= ⅓ (7.7-7)

0.2 ≈ 0.26

For Right Skew Data

Mean	7.7
Median	7.5
Mode	7


Is the variable you are measuring qualitative or quantitative?
If qualitative:
Data that is descriptive using words and is open to interpretation
If quantitative:
Data in the form of a number abstained in a count or measurement
Nominal
No ranking or order (i.e color category)
Ordinal
Ranked order categories (i.e first, second, third…)
Skew
Does the data follow a normal distribution? Are there outliers?
MODE
MEDIAN
Skewed
Data set does not have normal distribution.
Not skewed
Data set has normal distribution.
MEDIAN
MEAN

Type of Data for Responding Variable	Best Measure of Central Tendency
Qualitative Nominal (no ranking or order)	Mode
Qualitative Ordinal (has ranking or order)	Median
Quantitative With majority of skew values less than |1.0|	Mean
Quantitative With majority of skew values more than |1.0|	Median

Measures of Dispersion

4. Check the assumptions for the test

χ^2 Table

Right-tail area	df = 1	df = 2	df = 3	df = 4	df = 5
>0.100	< 2.70	< 4.60	< 6.25	< 7.77	< 9.23
0.100	2.70	4.60	6.25	7.77	9.23
0.095	2.78	4.70	6.36	7.90	9.37
0.090	2.87	4.81	6.49	8.04	9.52
0.085	2.96	4.93	6.62	8.18	9.67
0.080	3.06	5.05	6.75	8.33	9.83
0.075	3.17	5.18	6.90	8.49	10.00
0.070	3.28	5.31	7.06	8.66	10.19
0.065	3.40	5.46	7.22	8.84	10.38
0.060	3.53	5.62	7.40	9.04	10.59
0.055	3.68	5.80	7.60	9.25	10.82
0.050	3.84	5.99	7.81	9.48	11.07
0.045	4.01	6.20	8.04	9.74	11.34
0.040	4.21	6.43	8.31	10.02	11.64
0.035	4.44	6.70	8.60	10.34	11.98
0.030	4.70	7.01	8.94	10.71	12.37
0.025	5.02	7.37	9.34	11.14	12.83
0.020	5.41	7.82	9.83	11.66	13.38
0.015	5.91	8.39	10.46	12.33	14.09
0.010	6.63	9.21	11.34	13.27	15.08
0.005	7.87	10.59	12.83	14.86	16.74
0.001	10.82	13.81	16.26	18.46	20.51
<0.001	>10.82	>13.81	>16.26	>18.46	>20.51

Right-tail area	df = 6	df = 7	df = 8	df = 9	df = 10
>0.100	<10.64	<12.01	<13.36	<14.68	<15.98
0.100	10.64	12.01	13.36	14.68	15.98
0.095	10.79	12.17	13.52	14.85	16.16
0.090	10.94	12.33	13.69	15.03	16.35
0.085	11.11	12.50	13.87	15.22	16.54
0.080	11.28	12.69	14.06	15.42	16.75
0.075	11.46	12.88	14.26	15.63	16.97
0.070	11.65	13.08	14.48	15.85	17.20
0.065	11.86	13.30	14.71	16.09	17.44
0.060	12.08	13.53	14.95	16.34	17.71
0.055	12.33	13.79	15.22	16.62	17.99
0.050	12.59	14.06	15.50	16.91	18.30
0.045	12.87	14.36	15.82	17.24	18.64
0.040	13.19	14.70	16.17	17.60	19.02
0.035	13.55	15.07	16.56	18.01	19.44
0.030	13.96	15.50	17.01	18.47	19.92
0.025	14.44	16.01	17.53	19.02	20.48
0.020	15.03	16.62	18.16	19.67	21.16
0.015	15.77	17.39	18.97	20.51	22.02
0.010	16.81	18.47	20.09	21.66	23.20
0.005	18.54	20.27	21.95	23.58	25.18
0.001	22.45	24.32	26.12	27.87	29.58
<0.001	>22.45	>24.32	>26.12	>27.87	>29.58

Steps for Chi-Square Test with an example:

- Consider a data-set where we have to determine why customers are leaving the bank, let's perform a Chi-Square test for two variables.
- ***Gender*** of a customer with values as Male/Female as the predictor and ***Exited*** describes whether a customer is leaving the bank with values Yes/No as the response. In this test we will check *is there any relationship between Gender and Exited*.

- Steps to perform the Chi-Square Test:

1. Define Hypothesis.
2. Build a Contingency table.
3. Find the expected values.
4. Calculate the Chi-Square statistic.
5. Accept or Reject the Null Hypothesis.

- **1.Define Hypothesis**
- Null Hypothesis (H0): Two variables are independent.
- Alternate Hypothesis (H1): Two variables are not independent.

- **2. Contingency table**
- A table showing the distribution of one variable in rows and another in columns. It is used to study the relation between two variables.

Exited\ Gender	Yes	No	Total
Male	38	178	216
Female	44	140	184
Total	82	318	400

Contingency table for observed values

- Degrees of freedom for contingency table is given as (r-1) * (c-1) where r, c are rows and columns. Here df = (2–1) * (2–1) = 1.

Exited\Gender	Yes	No
Male	44	172
Female	38	146

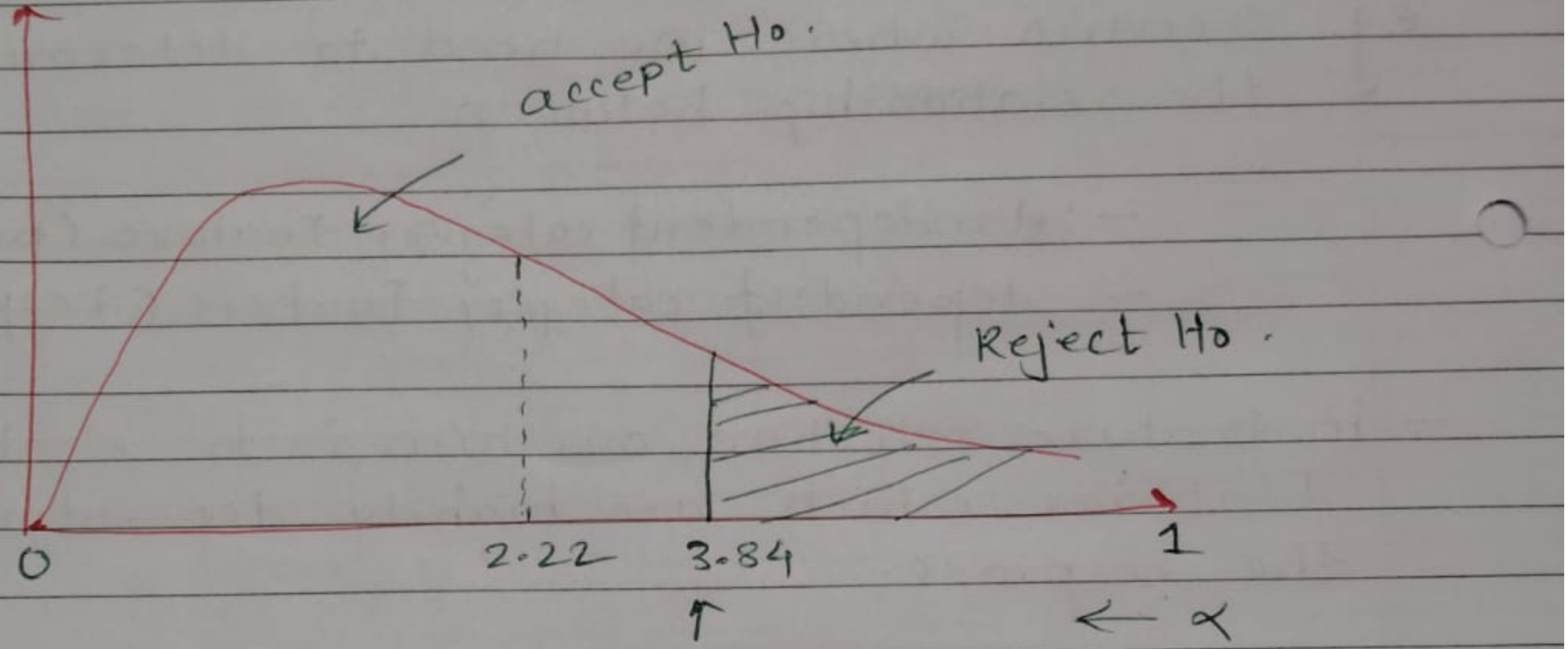
Expected values

Gender,Exited	O	E	O-E	Square of O-E	(Square of O-E) / E
Male,Yes	38	44	-6	36	0.818181818
Male,No	178	172	6	36	0.209302326
Female,Yes	44	38	6	36	0.947368421
Femal,No	140	146	-6	36	0.246575342
Chi Square Value					2.221427907

- Having degrees of freedom =1(calculated with contingency table) and alpha =0.05 the Chi-Square value is 3.84.
- *The chi-square distribution is the right side since the difference in Observed and Expected is large.*

$\alpha = 0.05$, $df = 1$, $x^2 = 3.84$

calculated $x^2 = 2.22$



critical Chi-square value.

α - Rages from

- **5. Accept or Reject the Null Hypothesis**

With 95% confidence that is alpha = 0.05, We will check the calculated Chi-Square value falls in the acceptance or rejection region.

- There are three main types of chi-square tests commonly used in statistics:
- **Pearson's Chi-Square Test**: This test is used to determine if there is a significant association between two categorical variables in a single population. It compares the observed frequencies in a contingency table with the expected frequencies assuming independence between the variables.
- **Chi-Square Goodness of Fit Test:** This test is used to assess whether observed categorical data follows an expected distribution. It compares the observed frequencies with the expected frequencies specified by a hypothesized distribution.
- **Chi-Square Test of Independence**: This test is used to examine if there is a significant association between two categorical variables in a sample from a population. It compares the observed frequencies in a contingency table with the expected frequencies assuming independence between the variables.

T-test example

- Consider the following example. The weights of 25 obese people were taken before enrolling them into the nutrition camp. The population mean weight is found to be 45 kg before starting the camp. After finishing the camp, for the same 25 people, the sample mean was found to be 75 with a standard deviation of 25. Did the fitness camp work?

T-test

- A t-test is a type of inferential statistic test used to determine if there is a significant difference between the means of two groups. It is often used when data is normally distributed and population variance is unknown.
- The t-test is used in hypothesis testing to assess whether the observed difference between the means of the two groups is statistically significant or just due to random variation.

P-value:

- The p-value is the probability of observing a test statistic (or something more extreme) given that the null hypothesis is true.
- A small p-value (typically less than the chosen significance level) suggests that the observed data is unlikely to have occurred by random chance alone, leading to the rejection of the null hypothesis.
- A large p-value suggests that the observed data is likely to have occurred by random chance, and there is not enough evidence to reject the null hypothesis.

Degree of freedom (df):

- The degree of freedom represents the number of values in a calculation that is free to vary. The degree of freedom (df) tells us the number of independent variables used for calculating the estimate between 2 sample groups.

Significance Level:

- The significance level is the predetermined threshold that is used to decide whether to reject the null hypothesis. Commonly used significance levels are 0.05, 0.01, or 0.10. A significance level of 0.05 indicates that the researcher is willing to accept a 5% chance of making a Type I error (incorrectly rejecting a true null hypothesis).

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291

Bayes Theorem

- Bayes Theorem is the extension of Conditional probability. Conditional probability helps us to determine the probability of A given B, denoted by P(A|B).
- So Bayes' theorem says if we know P(A|B) then we can determine P(B|A), given that P(A) and P(B) are known to us.

Formula Derivation

From conditional probability, we know that

- P(A|B) = P(A and B)/P(B)
- P(A and B) = P(B) * P(A|B) — — — — –[1]

Similarly

- P(B|A) = P(B and A)/P(A) = P(A and B)/P(A) [In Joint Probability order does not matter]
- P(A and B) = P(A) * P(B|A) — — ---------[2]

- From equation [1] and [2],
- P(B) * P(A|B) = P(A) * P(B|A)
- **P(A|B) = P(A) * P(B|A) / P(B)**
- **Which mean if we know P(A|B) then we can easily determine P(B|A) and vice versa. Assuming we know the total probabilities P(A) and P(B).**

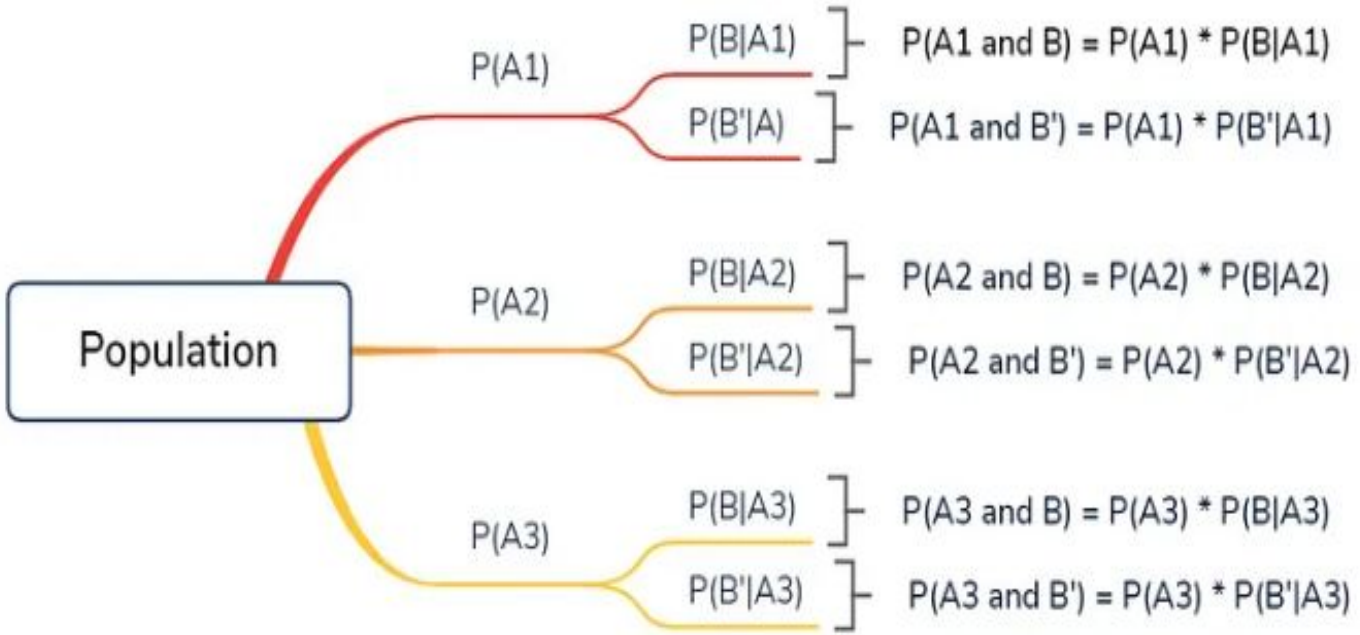
1. Prior Probability

- Prior Probability is the probability of occurring an event before the collection of new data.
- It is the best logical evaluation of the probability of an outcome which is based on the present knowledge of the event before the inspection is performed

2. Posterior Probability

- When new data or information is collected then the Prior Probability of an event will be revised to produce a more accurate measure of a possible outcome.
- This revised probability becomes the Posterior Probability and is calculated using Bayes' theorem

Tree representation of the Bayes' Theorem



Bayes Theorem: To Find Reverse Probabilities

$P(A1|B) = P(A1) * P(B|A1) / P(B)$

P(A1) and P(B) are known as marginal probabilities.
P(B|A1) and P(A1) is given to us.
P(B) can be calculated as
$P(B) = P(A1)*P(B|A1) + P(A2)*P(B|A2) + P(A3)*P(B|A3)$ and also known as Total Probability

Bayes Theorem: Find Reverse Probabilities

$P(A1|B') = P(A1)*P(B'|A1 / P(B')$

P(B') can be calculated as

$P(B') = P(A1)*P(B'|A1) + P(A2)*P(B'|A2) + P(A3)*P(B'|A3)$

Example

- Technicians regularly make repairs when breakdowns occur on an automated production line.
- Janak, who services 20% of the breakdowns, makes an incomplete repair 1 time in 20.
- Tarun, who services 60% of the breakdowns, makes an incomplete repair 1 time in 10.
- Gautham, who services 15% of the breakdowns, makes an incomplete repair 1 time in 10
- and Prasad, who services 5% of the breakdowns, makes an incomplete repair 1 time in 20.
- For the next problem with the production line diagnosed as being due to an initial repair that was incomplete, what is the probability that this initial repair was made by Janak?

Visualize Bayes Tree

Reverse Probability

P(Janak | Incomplete) = ?

Solution:

P(Janak | Incomplete) = P(Repaired by Janak)*P(Incomplete | Repaired by Janak)/P(Incomplete)

P(Janak | Incomplete) = 0.20*0.05/[0.20*0.05 + 0.60*0.1 + 0.15*0.1 + 0.05*0.05]

Example

- SpamAssassin works as a mail filter to identify the spam in which users train the system. In emails, it considers patterns in the words which are marked as spam by the users.

For Example, it may have learned that the word “release” is marked as spam in 30% of the emails. Concluding 0.8% of non-spam mails which includes the word “release” and 40% of all emails which are received by the user is spam. Find the probability that a mail is a spam if the word “release” seems in it.

- SpamAssassin works by having users train the system. It looks for patterns in the words in emails marked as spam by the user. For example, it may have learned that the word “release” appears in 30% of the emails marked as spam. Assuming 0.8% of non-spam mail includes the word “release” and 40% of all emails received by the user is spam, find the probability that a mail is a spam if the word “release” appears in it.

- SpamAssassin works by having users train the system. It looks for patterns in the words in emails marked as spam by the user. For example, it may have learned that the word “free” appears in 20% of the emails marked as spam. Assuming 0.1% of non-spam mail includes the word “free” and 50% of all emails received by the user is spam, find the probability that a mail is a spam if the word “free” appears in it.

- **Data Given:**
- P(Free | Spam) = 0.20
- P(Free | Non Spam) = 0.001
- P(Spam) = 0.50 => P(Non Spam) = 0.50
- P(Spam | Free) = ?

- **Using Bayes' Theorem:**
- P(Spam | Free) = P(Spam) * P(Free | Spam) / P(Free)
- P(Spam | Free) = 0.50 * 0.20 / (0.50 * 0.20 + 0.50 * 0.001)
- P(Spam | Free) = 0.995

Pearson Correlation

- Pearson correlation coefficient is a measure of the strength of a linear association between two variables
- **denoted by r**

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Where,

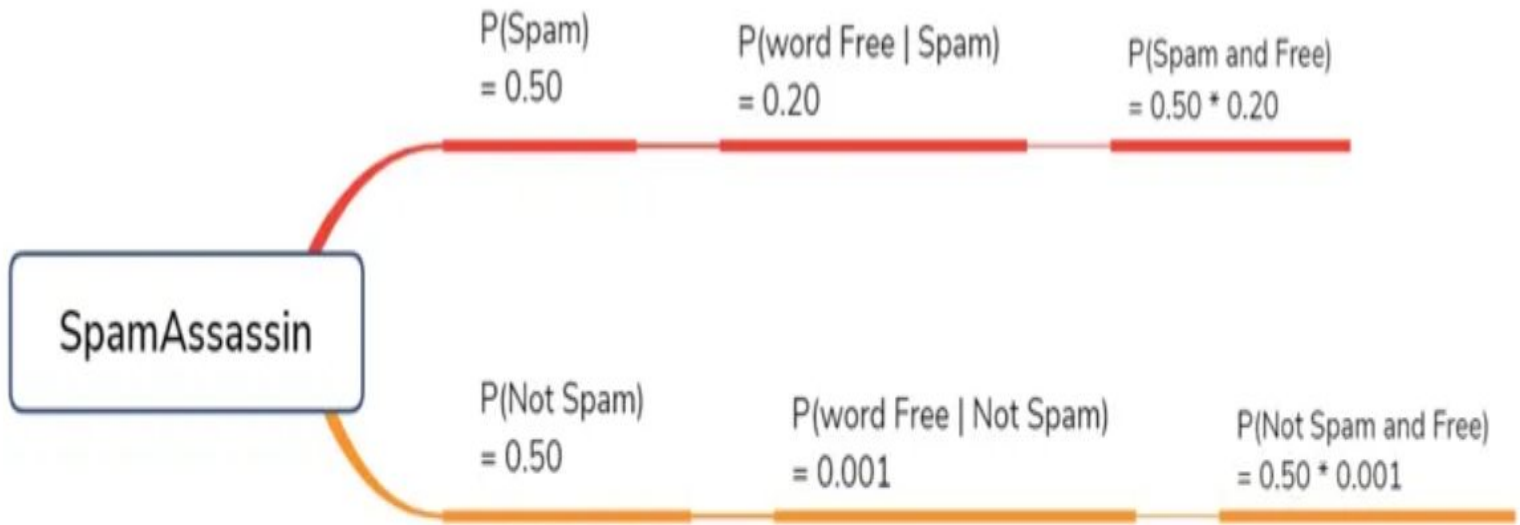
r = Pearson Correlation Coefficient

x_i = x variable samples

y_i = y variable sample

$\bar{x}$ = mean of values in x variable

$\bar{y}$ =mean of values in y variable



Using Bayes' Theorem:

- P(Spam | Free) = P(Spam) * P(Free | Spam) / P(Free)
- P(Spam | Free) = 0.50 * 0.20 / (0.50 * 0.20 + 0.50 * 0.001)
- P(Spam | Free) = 0.995

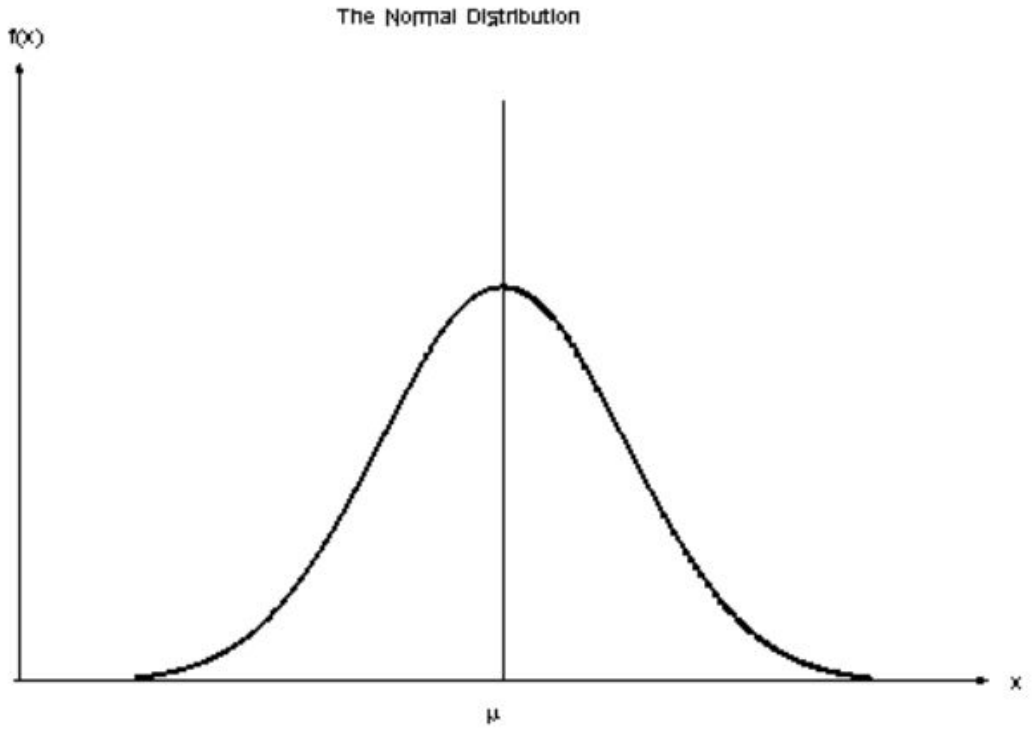
Why Correlation?

- Is there a statistically significant relationship between age and height?
- Is there a relationship between temperature and ice cream sales?
- Is there a relationship among job satisfaction, productivity, and income?
- Which two variable have the strongest correlation between age, height, weight, size of family and family income

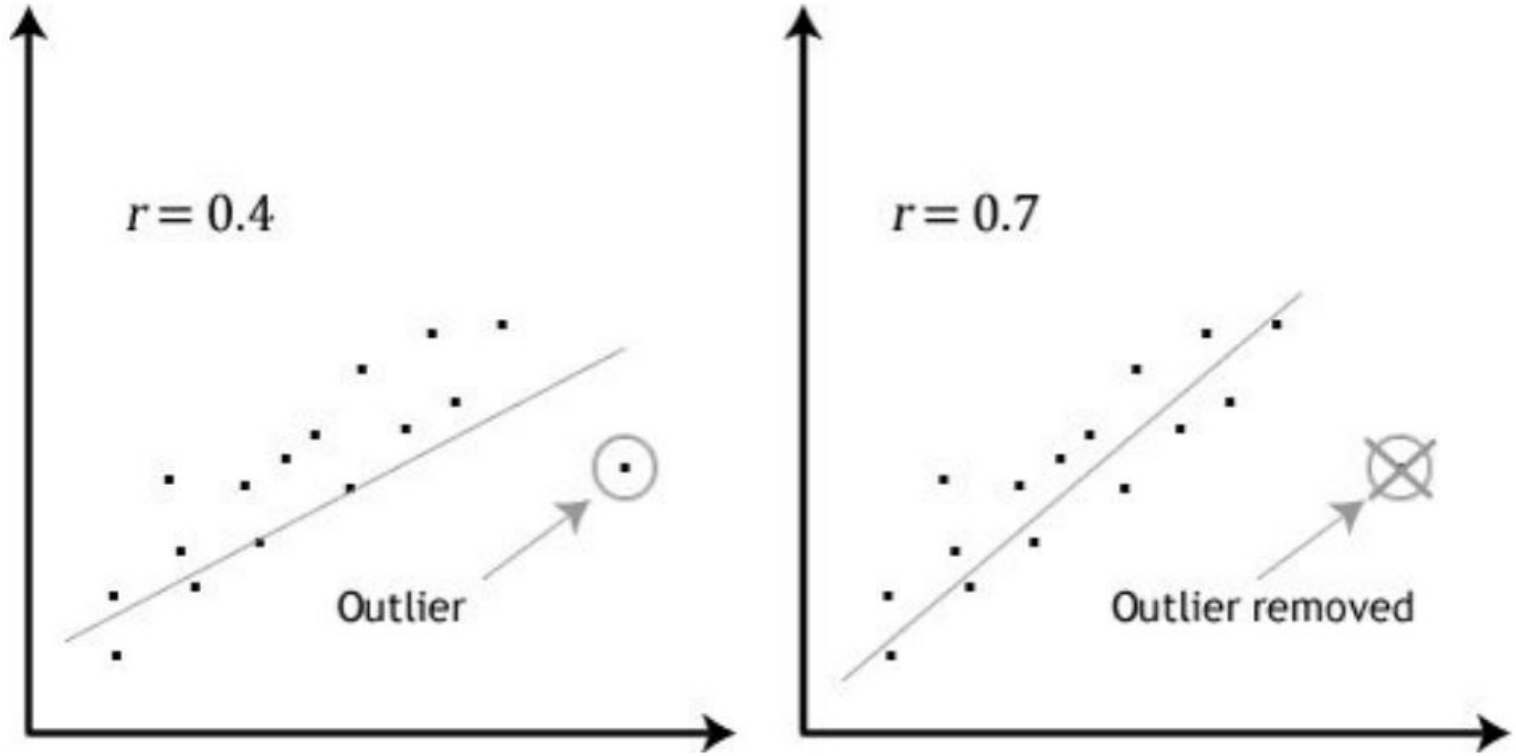
Assumptions for a Pearson Correlation

1. Both variables should be normally distributed.

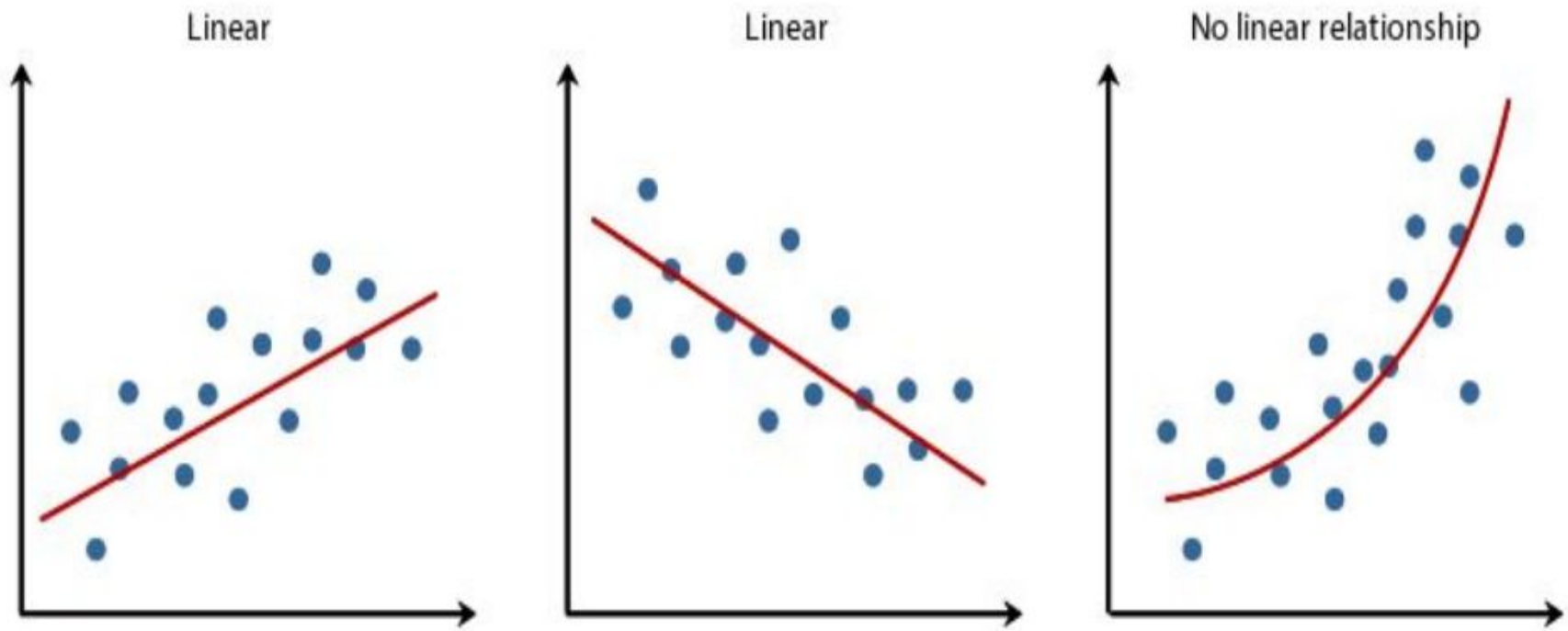
   This is sometimes called the 'Bell Curve' or the 'Gaussian Curve'



- There should be no significant outliers



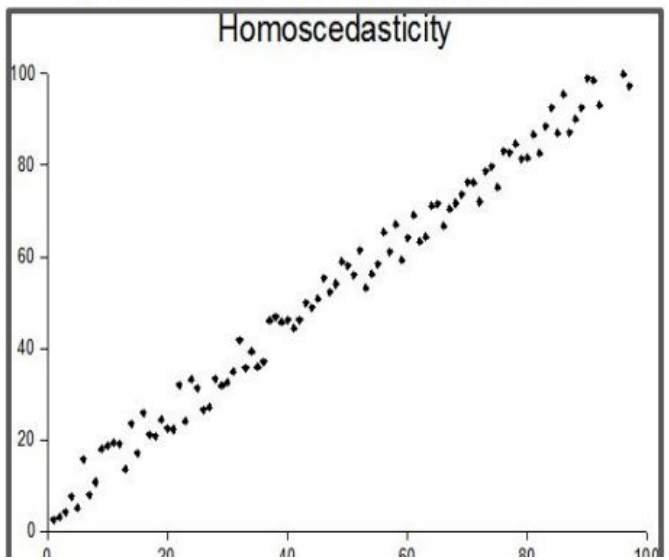
3. Each variable should be continuous i.e. interval or ratios for example weight, time, height, age etc
4. The two variables have a linear relationship

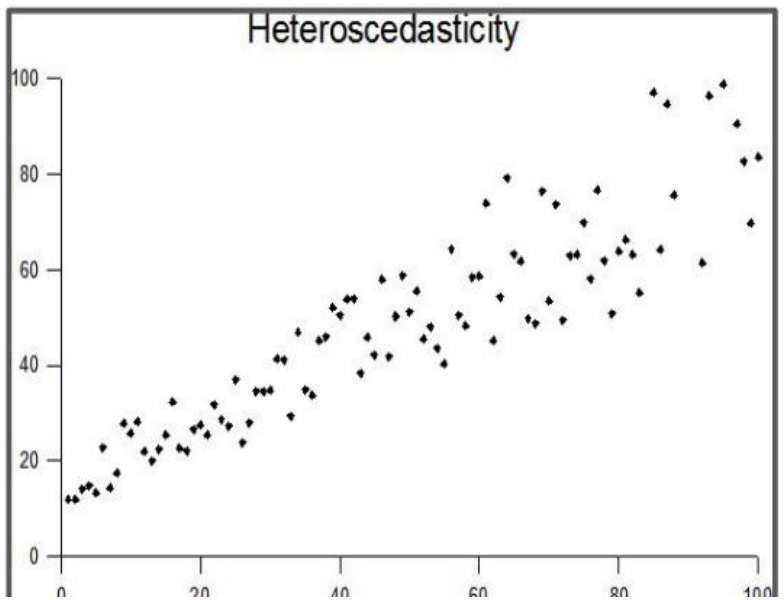

Linear
Linear
No linear relationship

5. The observations are **paired observations**. That is, for every observation of the independent variable, there must be a corresponding observation of the dependent variable.

For example if you're calculating the correlation between age and weight. If there are 12 observations of weight, you should have 12 observations of age. i.e. no blanks.

6. There should be **Homoscedasticity,** which means **the variance around the line of best fit should be similar.** Homoscedasticity describes a situation in which **the error term is the same across all values of the independent variables. A scatter-plot makes i**t easy to check for this. If the points lie equally on both sides of the line of best fit, then the data is homoscedastic.





Pearson correlation example

	A	B	C	D	E	F	G	H
5								
6	Hours Played Sport	Test Score	$x - \bar{x}$	$y - \bar{y}$	$(xi - \bar{x})^2$	$(yi - \bar{y})^2$	$(xi - \bar{x})*(yi - \bar{y})$	
7	x	y						
8	3	74	0.43	1.71	0.18	2.94	0.73	
9	1	68	-1.57	-4.29	2.47	18.37	6.73	
10	1	66	-1.57	-6.29	2.47	39.51	9.88	
11	3	72	0.43	-0.29	0.18	0.08	-0.12	
12	4	80	1.43	7.71	2.04	59.51	11.02	
13	2	68	-0.57	-4.29	0.33	18.37	2.45	
14	4	78	1.43	5.71	2.04	32.65	8.16	
15								
16		$\bar{x}$ (Mean of x)	$\bar{y}$ (Mean of y)					
17	Mean	**2.57**	**72.29**					
18								

	A	B	C	D	E	F	G	H
5								
6	Hours Played Sport	Test Score	$x - \bar{x}$	$y - \bar{y}$	$(xi - \bar{x})^2$	$(yi - \bar{y})^2$	$(xi - \bar{x})*(yi - \bar{y})$	
7	x	y						
8	3	74	0.43	1.71	0.18	2.94	0.73	
9	1	68	-1.57	-4.29	2.47	18.37	6.73	
10	1	66	-1.57	-6.29	2.47	39.51	9.88	
11	3	72	0.43	-0.29	0.18	0.08	-0.12	
12	4	80	1.43	7.71	2.04	59.51	11.02	
13	2	68	-0.57	-4.29	0.33	18.37	2.45	
14	4	78	1.43	5.71	2.04	32.65	8.16	
15								
19	Sum is calculated as							
20								
21		$(xi - \bar{x})^2$	$(yi - \bar{y})^2$	$(xi - \bar{x})*(yi - \bar{y})$				
22	Formula	=SUM(E8:E14)	=SUM(F8:F14)	=SUM(G8:G14)				
23	Sum	9.71	171.43	38.86				
24								

	A	B	C	D	E
20					
21		$(xi - \bar{x})^2$	$(yi - \bar{y})^2$	$(xi - \bar{x})*(yi - \bar{y})$	
22	Sum	9.71	171.43	38.86	
23					
24	Standard Deviation is calculated as				
25					
26		σx	σy		
27	Formula	=SQRT(B22)	=SQRT(C22)		
28	Standard Deviation	3.12	13.09		
29					

	A	B	C	D	E
20					
21		$(xi - \bar{x})^2$	$(yi - \bar{y})^2$	$(xi - \bar{x})*(yi - \bar{y})$	
22	Sum	9.71	171.43	38.86	
25					
26		σx	σy		
27	Standard Deviation	3.12	13.09		
28					
29	Pearson Correlation Coefficient is calculated using the formula given below				
30	**Pearson Correlation Coefficient = $\rho(x,y) = \Sigma[(xi - \bar{x}) * (yi - \bar{y})] / (\sigma x * \sigma y)$**				
31					
32	Pearson Correlation Coefficient Formula	=D22/(B27*C27)			
33	Pearson Correlation Coefficient	0.95			
34					

- Pearson Correlation Coefficient = 38.86/(3.12*13.09)
- Pearson Correlation Coefficient = 0.95

We have an output of 0.95; this indicates that when **the number of hours played to increase, the test scores also increase. These two variables are positively correlated.**

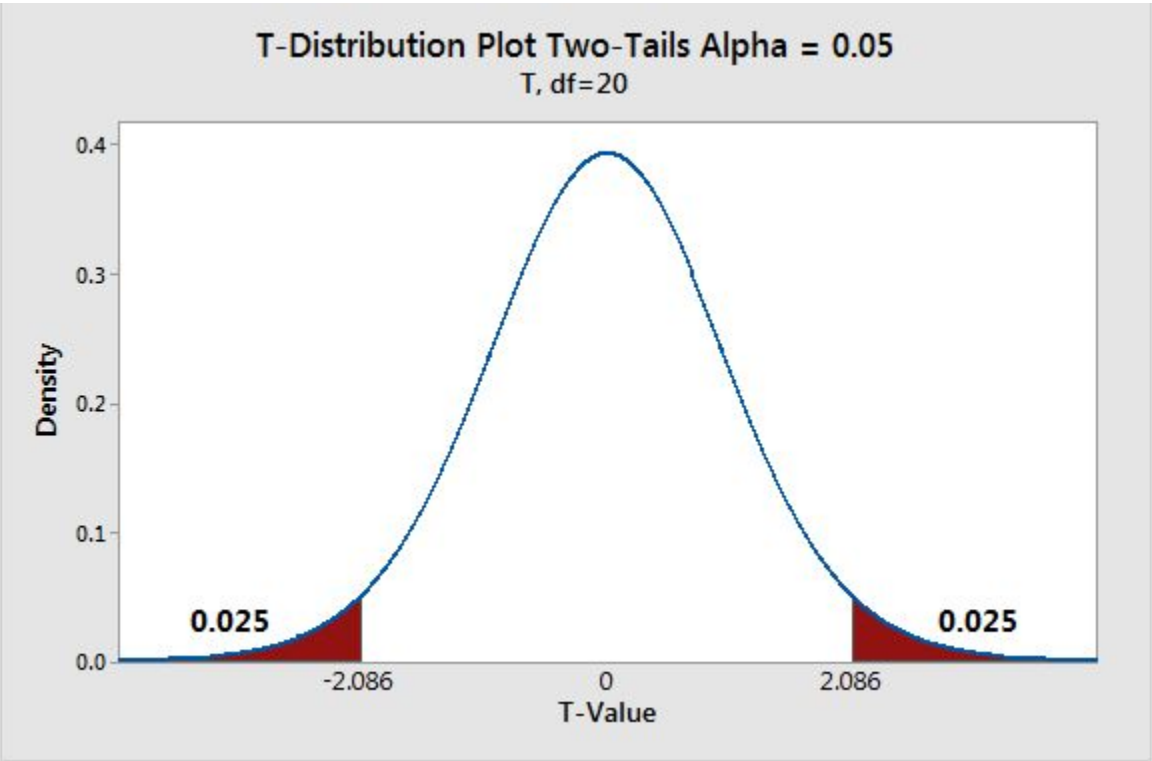
Type-I and Type-II Errors

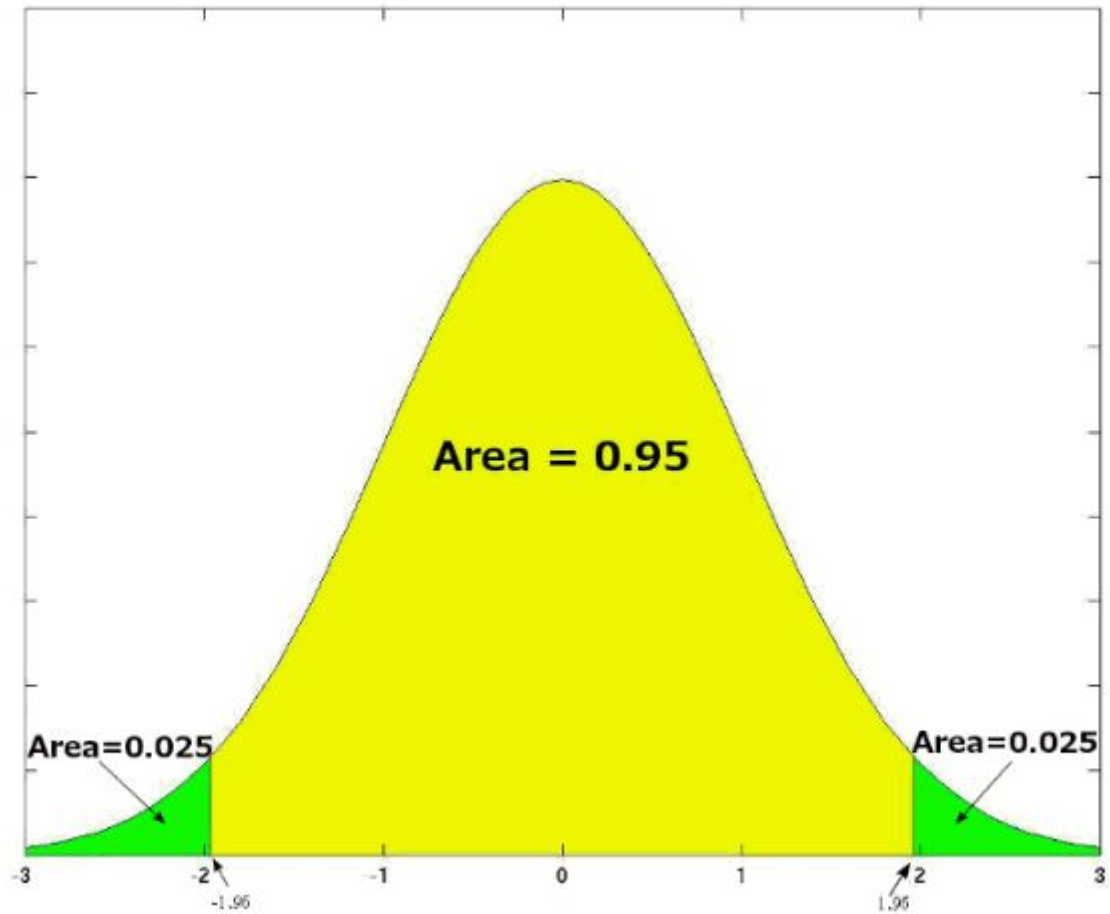
- **Type I and Type II errors** are subjected to the result of the null hypothesis
- In case of Type-I error, the null hypothesis is rejected though it is true
- Type II error, the null hypothesis is not rejected even when the alternative hypothesis is true

- Both the error type-i and type-ii are also known as “**false negative**”

Error Types	When H_0 is True	When H_0 is False
Don't Reject	Correct Decision (True negative) Probability = $1 - \alpha$	Type II Error (False negative) Probability = β
Reject	Type II Error (False Positive) Probability = α	Correct Decision (True Positive) Probability = $1 - \beta$

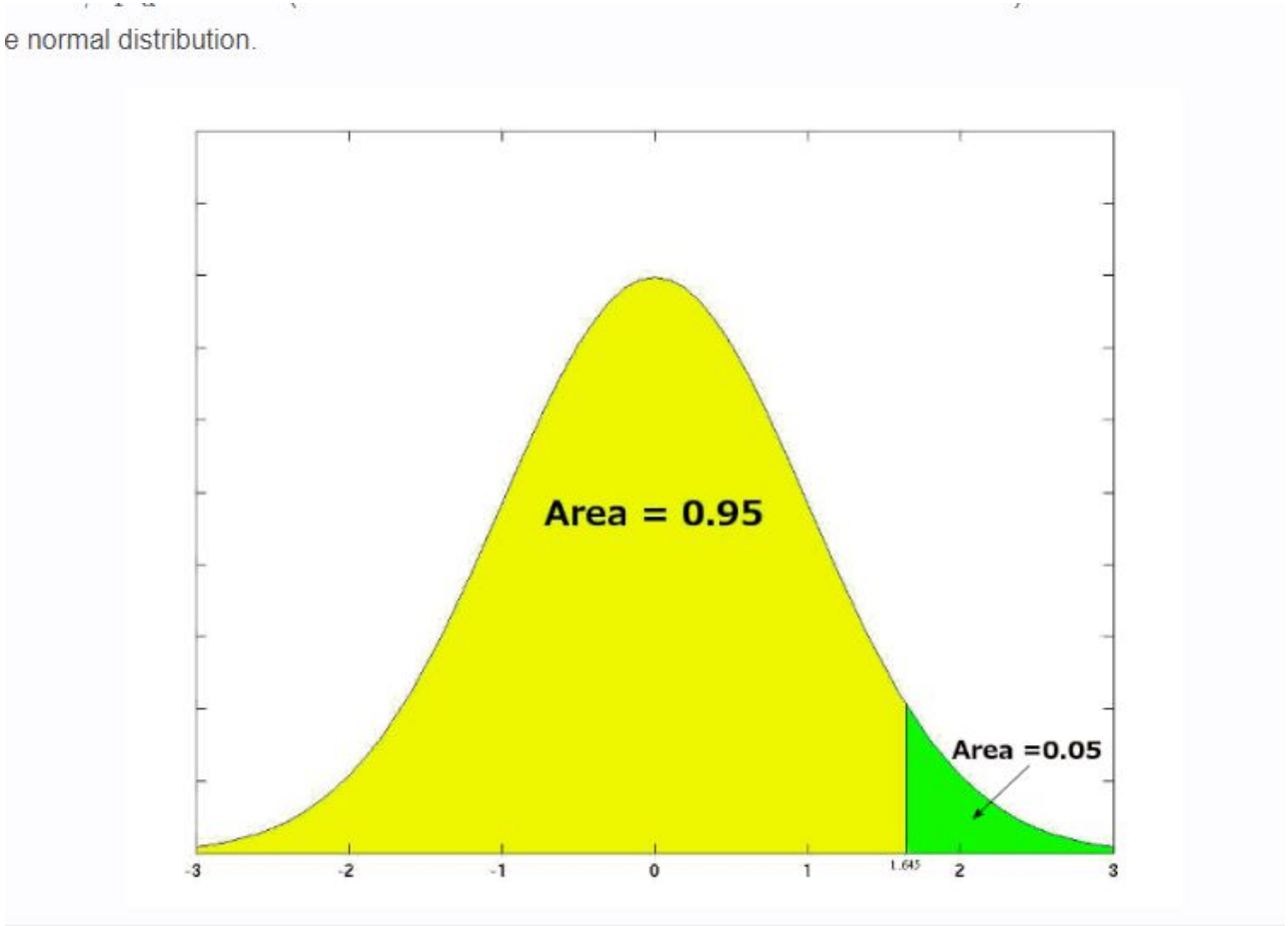
Two tail Test




Area = 0.95
Area=0.025
Area=0.025
-3
-2
-1
0
1
2
3
-1.96
1.96

One Tail test

e normal distribution.



one-tailed test

- A one-tailed test may be either left-tailed or right-tailed.
- A *left-tailed* test is used when the alternative hypothesis states that the true value of the parameter specified in the null hypothesis is less than the null hypothesis claims.
- A *right-tailed* test is used when the alternative hypothesis states that the true value of the parameter specified in the null hypothesis is greater than the null hypothesis claims