Statistics
&
Probablities

By

Laxminarayen

Descriptive Statistics

Agenda-

In this session you will learn about

- Basics of Statistics
- Types of Variables
- Measure of Central Tendancy
- Measure of Dispersion
- Case studies of Central tendencies and Dispersion
- Percentile/Quartile & Correlation and Covariance
- Central Limit Theorem
- Data Visualization and distribution

What is Statistics ?

WHAT IS STATISTICS?

A branch of mathematics taking and transforming numbers into useful information for decision makers.

What is Statistics

Statistics is a way to get information from data.

Why Learn Statistics?

Case 1 - Answer in 5 seconds !

Case 1 - Answer in 5 seconds !

A college in US has students from the following countries for a Masters degree. Which country is in majority ?

Case 1 - Answer in 5 seconds !

A college in US has students from the following countries. Which country is in majority ?

US	China	US	Sweden	China
Canada	China	Japan	Mexico	US
China	Germany	India	India	Japan
US	US	US	China	China
India	Japan	England	India	Japan
England	India	China	Mexico	US
Mexico	US	Canada	Pakistan	India
Japan	China	US	Japan	Germany
China	India	India	China	China
Germany	Japan	China	US	Japan

Frequency Table

Country	Frequency
Canada	2
China	12
England	2
Germany	3
India	8
Japan	8
Mexico	3
Pakistan	1
Sweden	1
US	10

Case 2

Problem

A parent changes school of their Son who is studying in 11^{th} standard since his academic results are not good in 10^{th} Standard in his current School.

They change Student A from ABC school to XYZ school

Case 2

Problem

A parent changes school of their Son who is studying in 11^{th} standard since his academic results are not good in 10^{th} Standard in his current School.

They change Student A from ABC school to XYZ school

Results

1. Ranked 15^{th} in ABC school
2. Ranked 2^{nd} in XYZ school

What's the conclusion ?

Case 2

Problem

A parent changes school of their Son who is studying in 11^{th} standard since his academic results are not good in 10^{th} Standard in his current School.

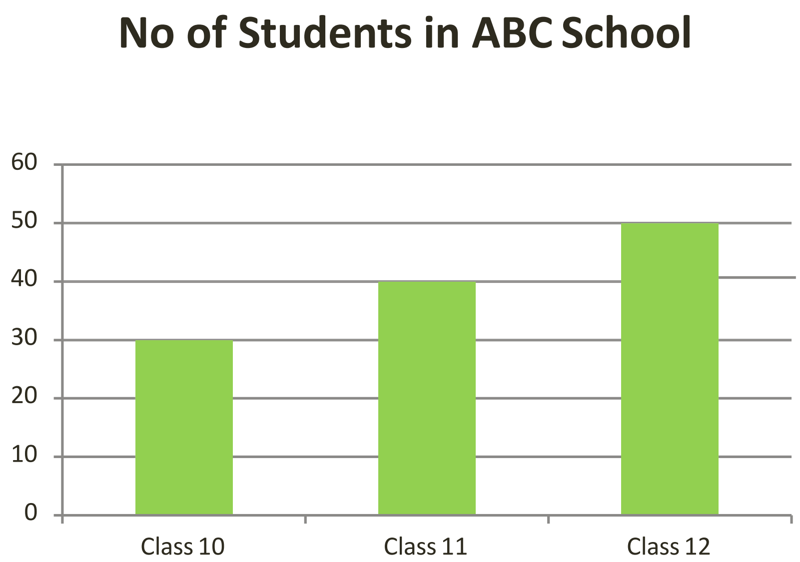
They change Student A from ABC school to XYZ school

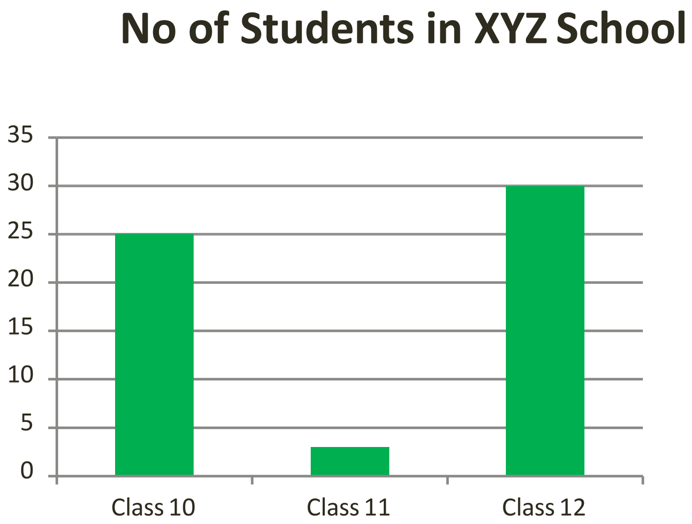
Results

1. Ranked 15^{th} in ABC school
2. Ranked 2^{nd} in XYZ school

What's the conclusion: Has the student improved ?

Number of Students





Why Learn Statistics?

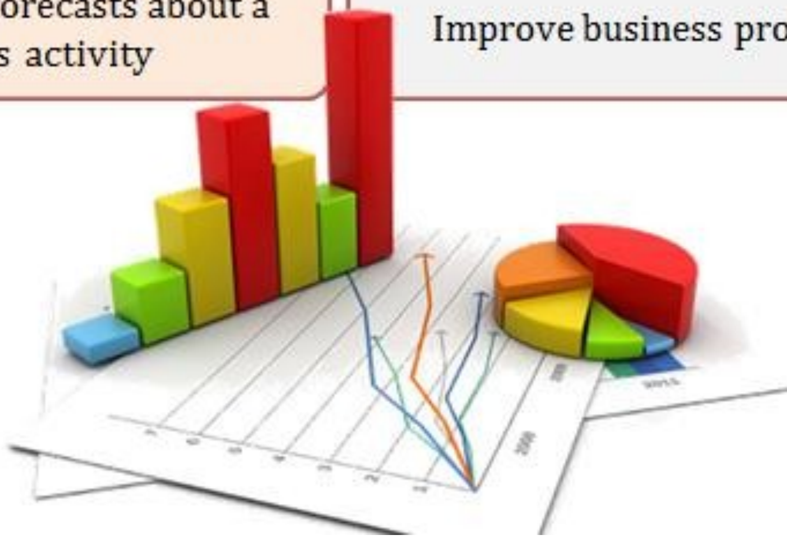
Why Learn Statistics?

Knowledge of Statistics allows you to make better sense of the ubiquitous use of numbers.

Why Learn Statistics ?

Decision Makers Use Statistics for Various Purposes:

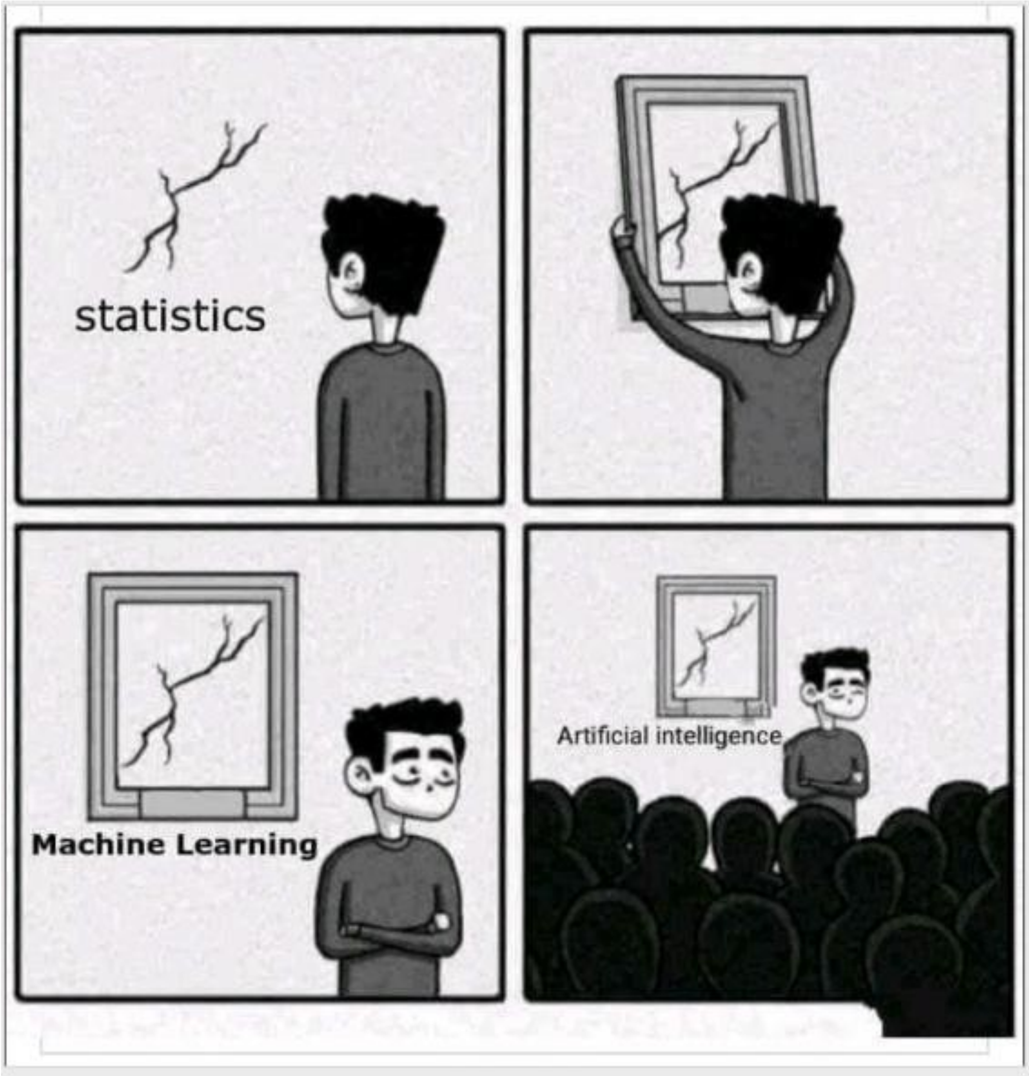
- Present and describe business data and information properly
- Draw conclusions about large sets using information collected from subsets
- Make reliable forecasts about a business activity
- Improve business processes



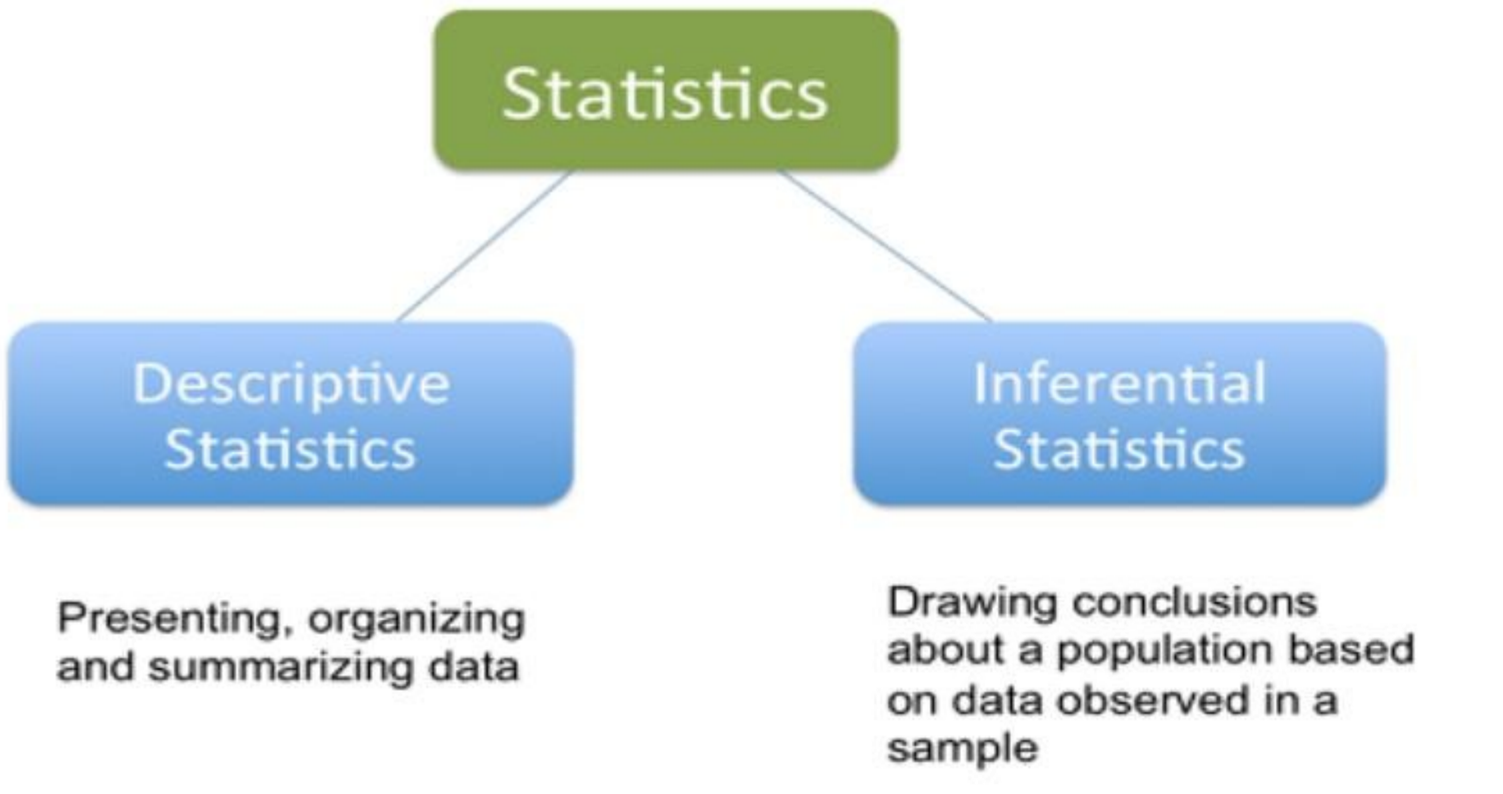
Statistics is …

1. *Collecting Data*
2. *Analyzing Data*
3. *Interpreting Data*
4. *Presenting Data*

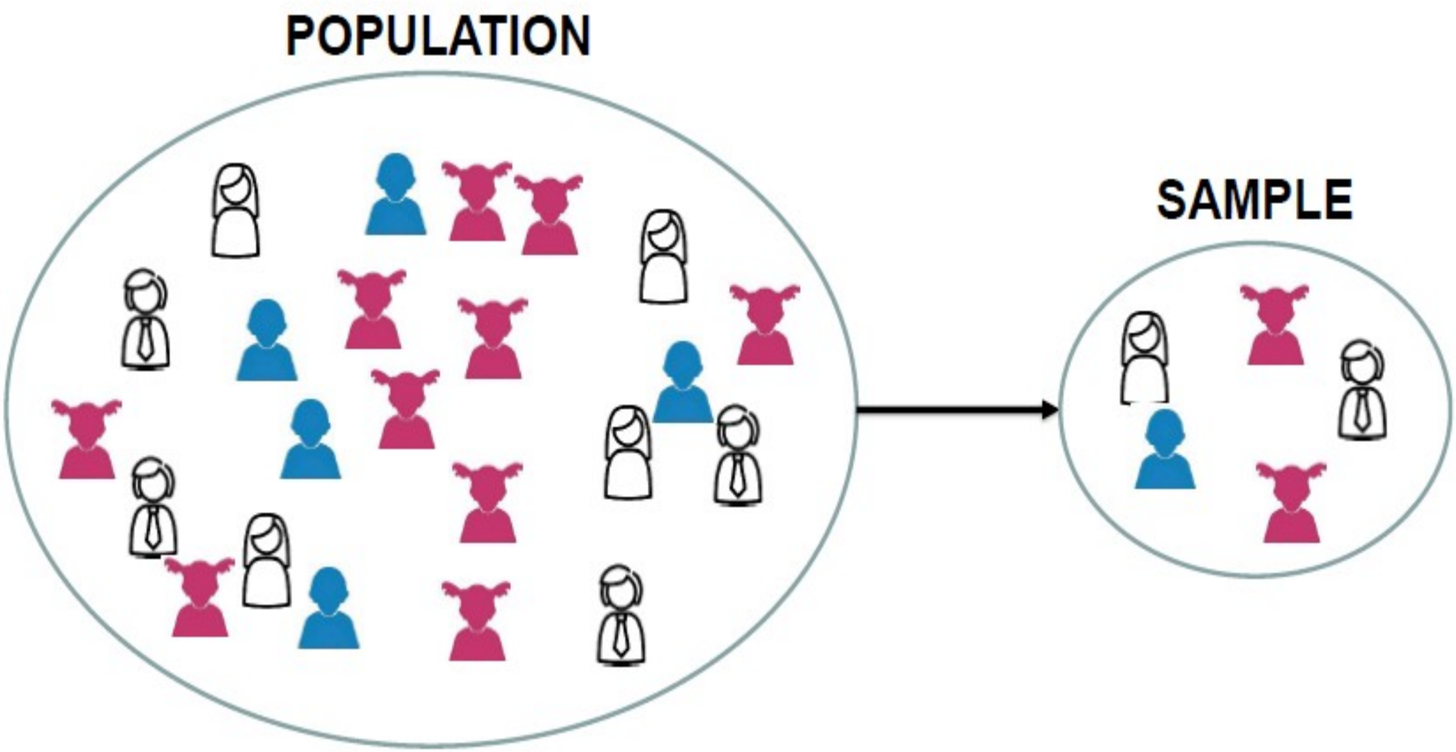
What does it Tell?



Classification



Population and Sample



Census and Survey

Census: Gathering data from the whole **population** of interest.

For example, elections, 10-year census, etc.

Survey: Gathering data from the **sample** in order to make conclusions about the population.

For example, opinion polls, quality control checks in manufacturing units, etc.

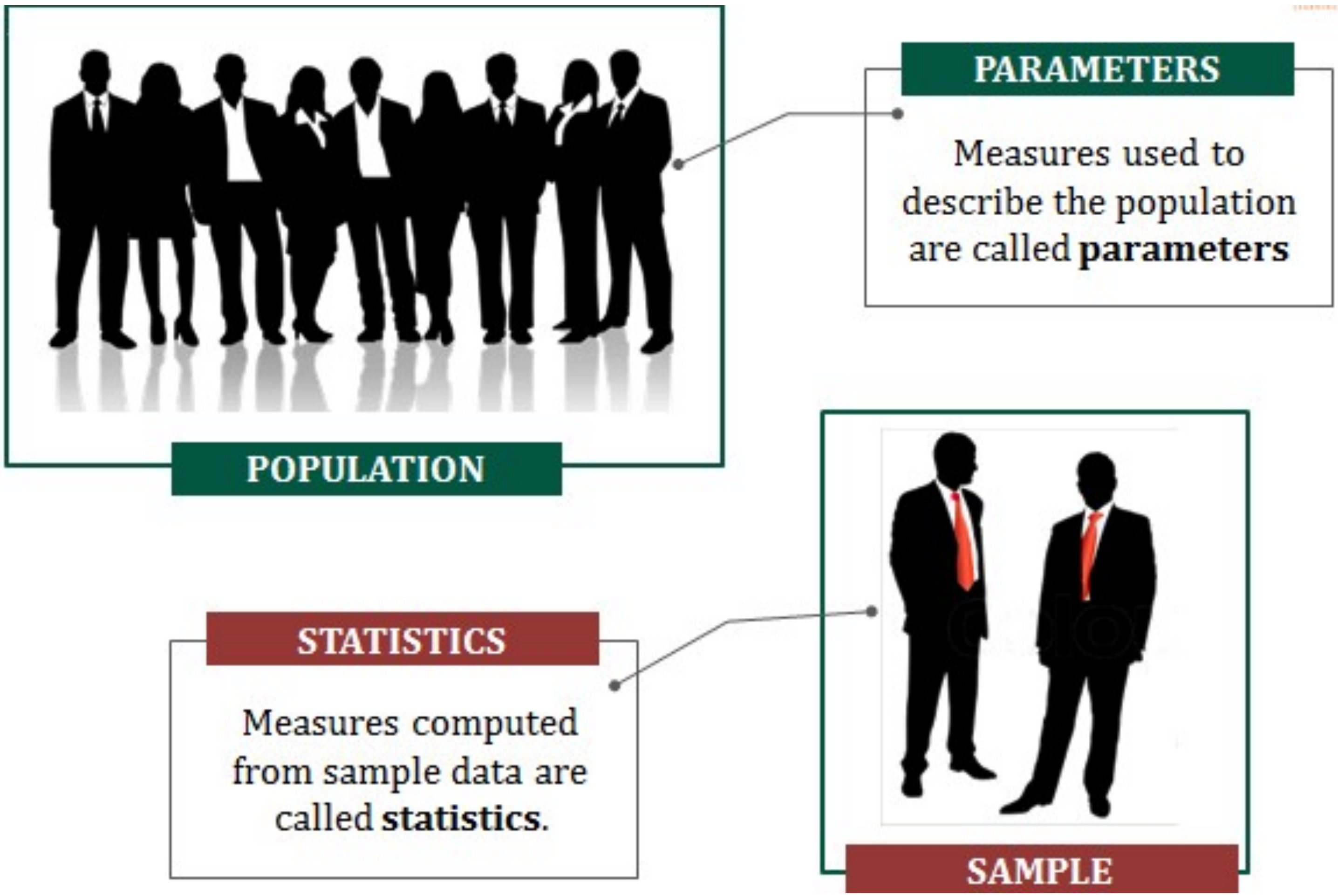
Parameter and Statistic

Parameter: A descriptive measure of the **population**.

For example, population mean, population variance, population standard deviation, etc.

Statistic: A descriptive measure of the **sample**.

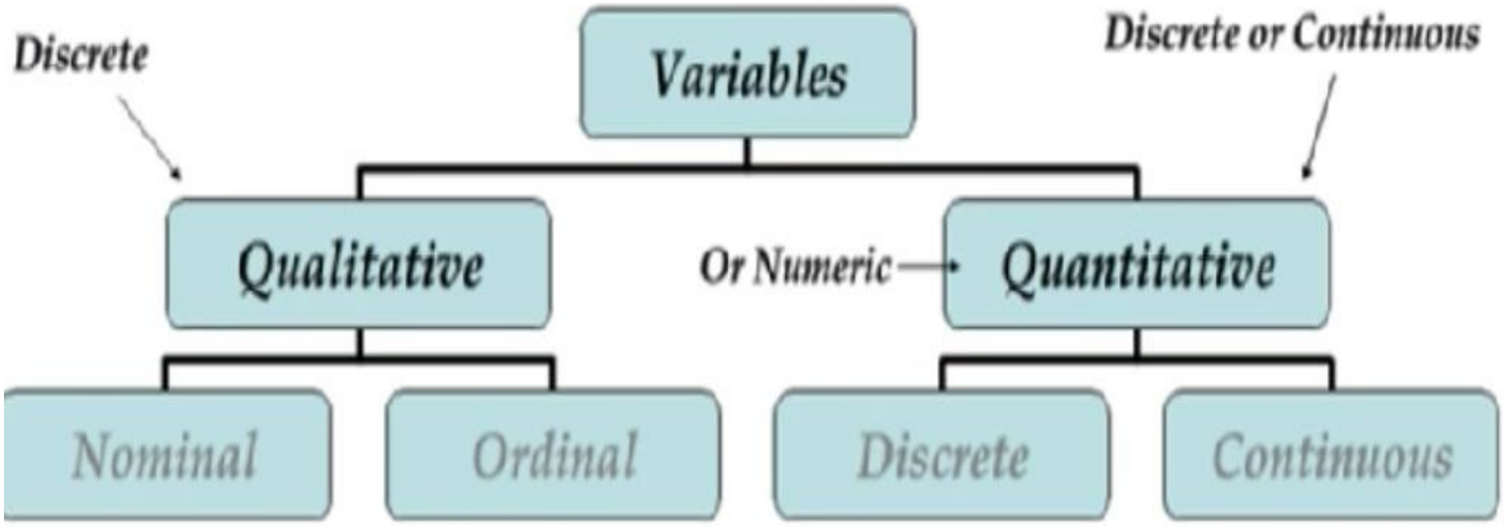
For example, sample mean, sample variance, sample standard deviation, etc.


POPULATION
PARAMETERS
Measures used to describe the population are called **parameters**
SAMPLE
STATISTICS
Measures computed from sample data are called **statistics**.

Statistical Notations

Greek – Population Parameter	**Roman – Sample Statistic**
Mean – μ	Mean – $\bar{x}$
Variance – σ^2	Variance – s^2
Standard Deviation - σ	Standard Deviation - s

Variables


Discrete
Variables
Discrete or Continuous
Qualitative
Or Numeric
Quantitative
Nominal
Ordinal
Discrete
Continuous

Categorical Data (Qualitative)

Nominal Examples

- Employee ID
- Gender
- Religion
- Ethnicity
- Pin codes
- Place of birth
- Aadhaar numbers

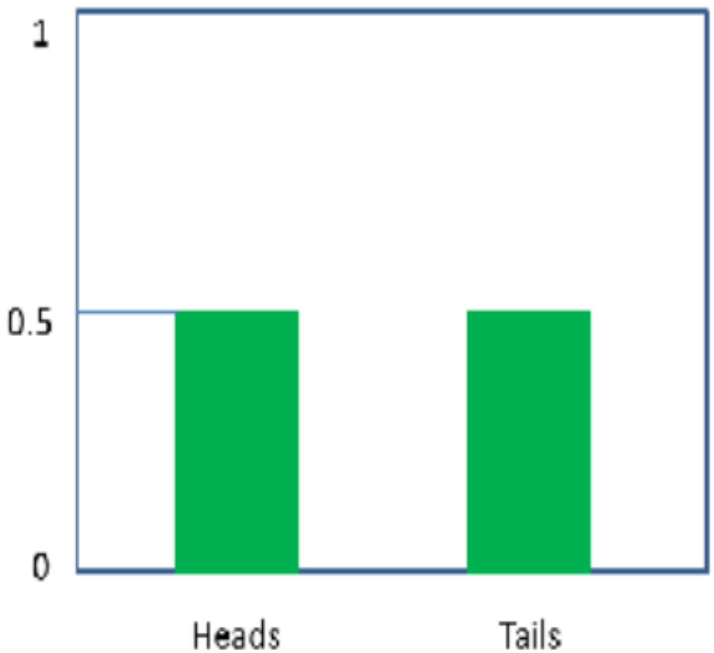
Ordinal

Examples

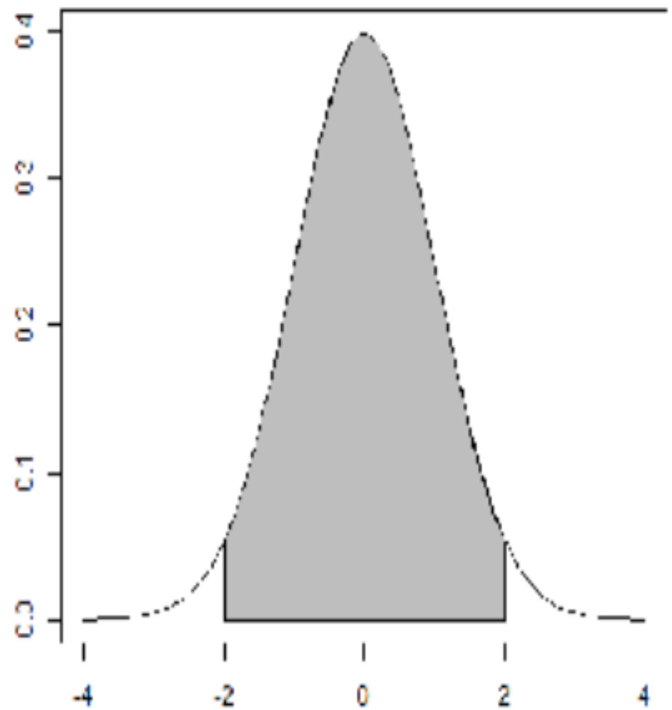
- Mutual fund risk ratings • Fortune 50 rankings
- Movie ratings

While there is an order, difference between consecutive levels are not always equal.

Discrete and Continuous



Countable



Measurable

Discrete or Continuous?

- Time between customer arrivals at a retail outlet
  Continuous
- Sampling 100 voters in an exit poll and determining how many voted for the winning candidate
  Discrete
- Lengths of newly designed automobiles -
  Continuous
- No. of customers arriving at a retail outlet during a five- minute period
  Discrete
- No. of defects in a batch of 50 items
  Discrete

Numerical or Categorical?

Age	Gender	Major	Units	Housing	GPA
18	Male	Psychology	16	Dorm	3.6
21	Male	Nursing	15	Parents	3.1
20	Female	Business	16	Apartment	2.8

- Numerical
- Categorical

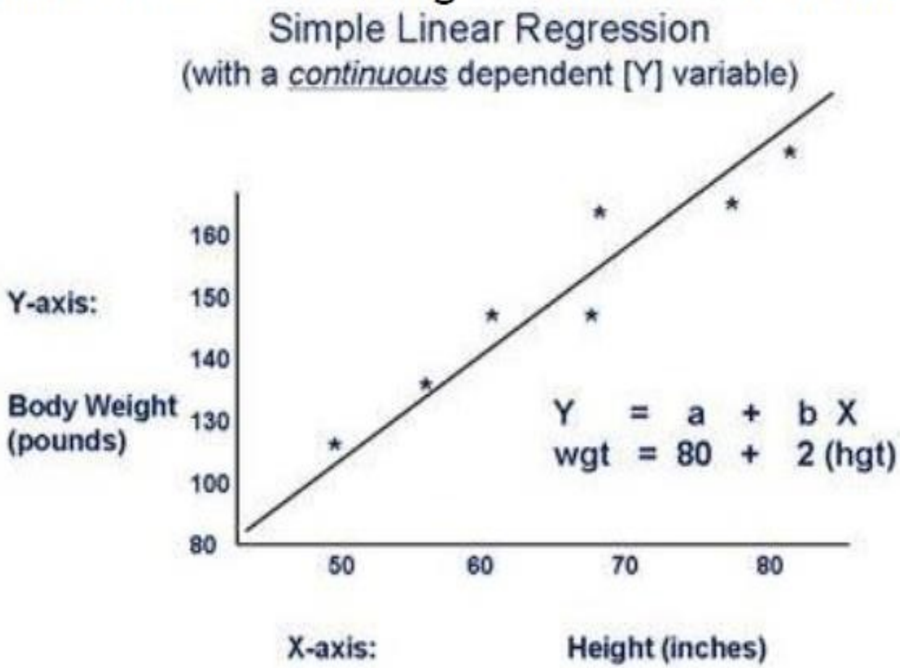
Numerical or Categorical?

Age	Gender	Major	Units	Housing	GPA
18	Male	Psychology	16	Dorm	3.6
21	Male	Nursing	15	Parents	3.1
20	Female	Business	16	Apartment	2.8

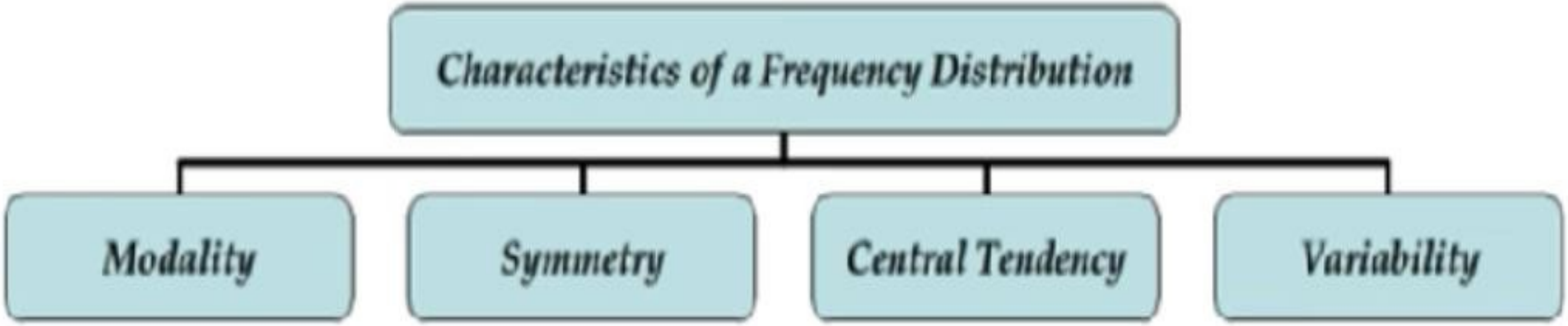
- Numerical
  - Age
  - Units
  - GPA
- Categorical
  - Gender
  - Major
  - Housing

Variables - Dependent and Independent

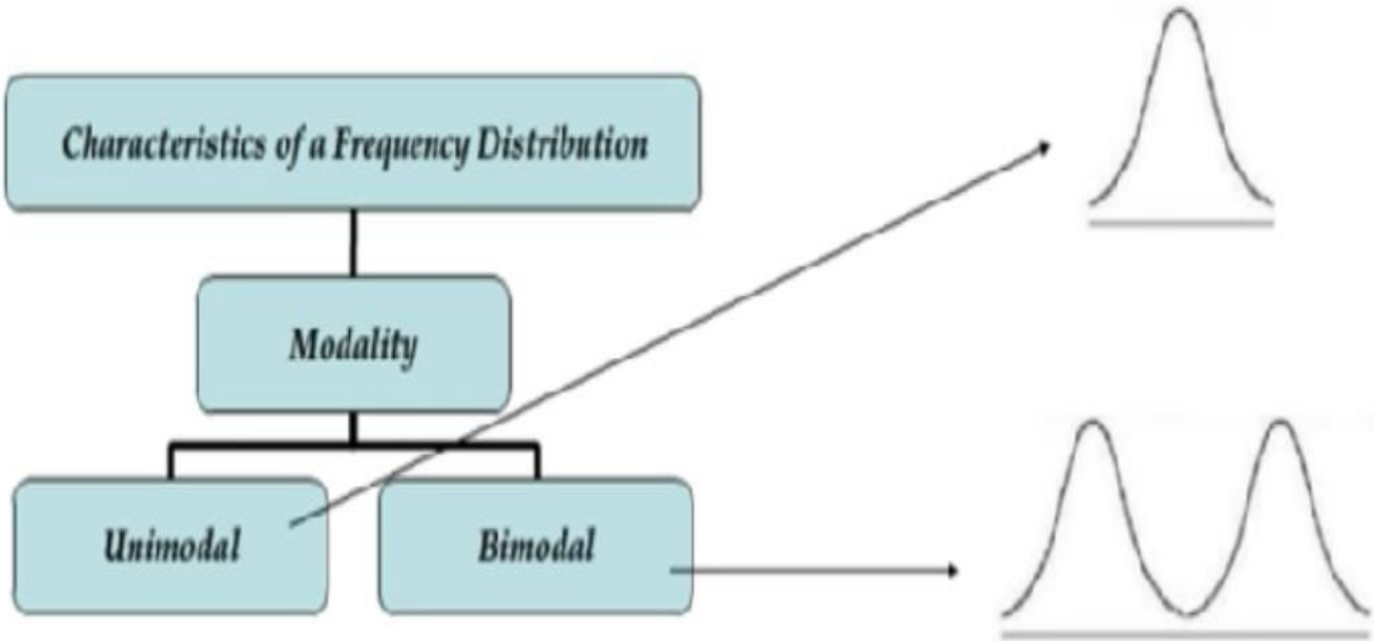
- Dependent variables on y-axis and Independent on x-axis.
- Dependent variable also called Target variable or Class variable.



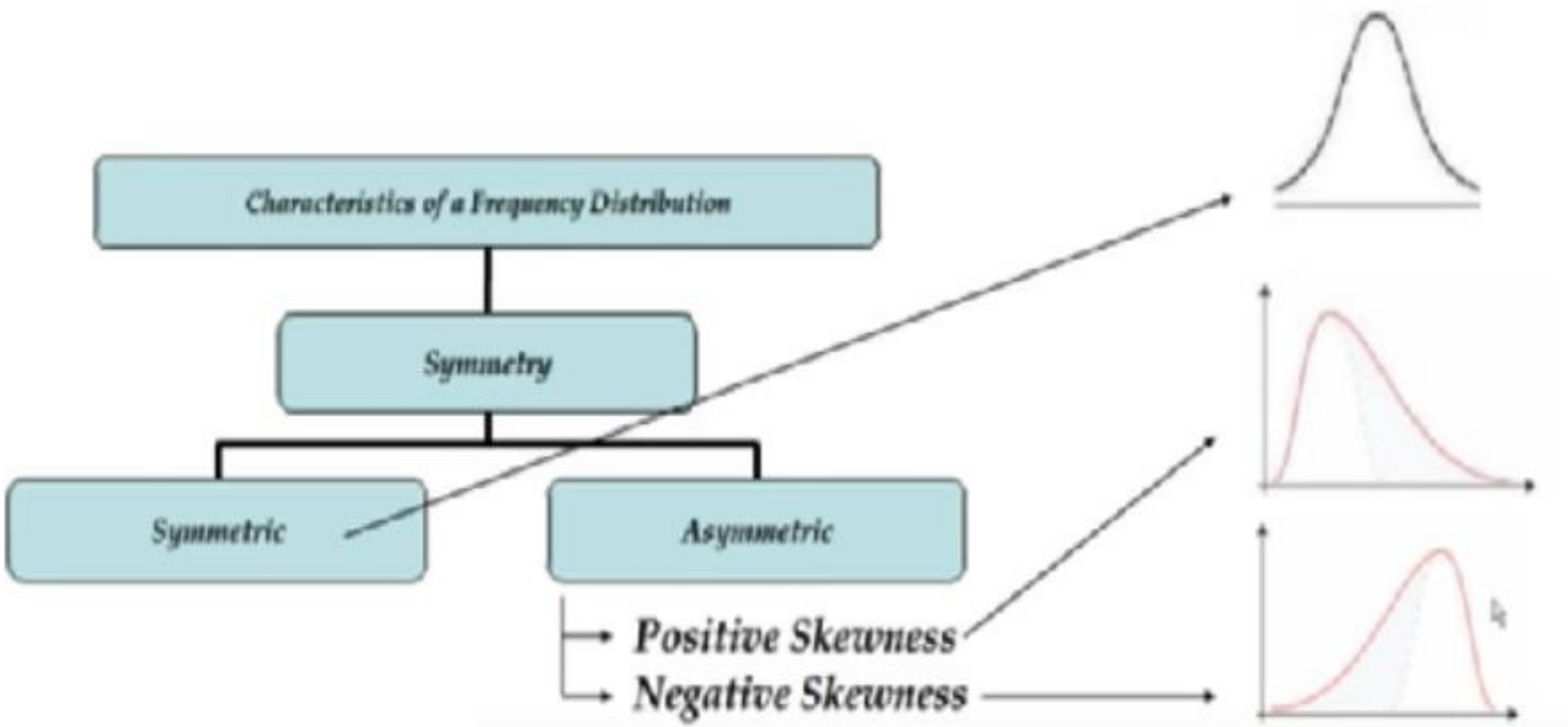
Summarizing Data



Modality

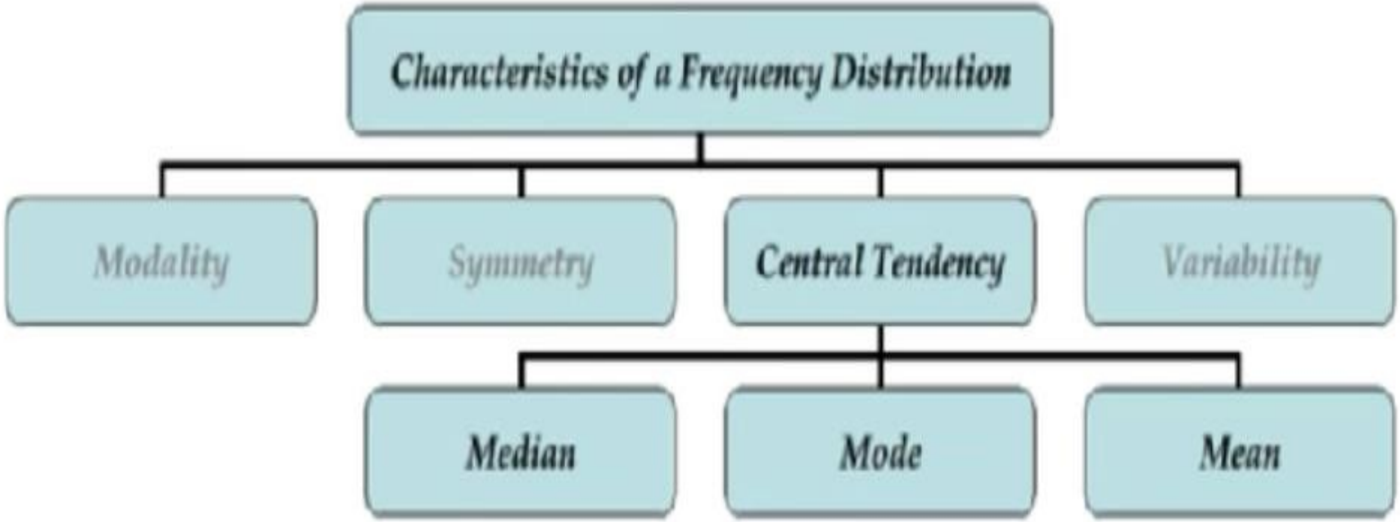

Characteristics of a Frequency Distribution
Modality
Unimodal
Bimodal

Symmetry

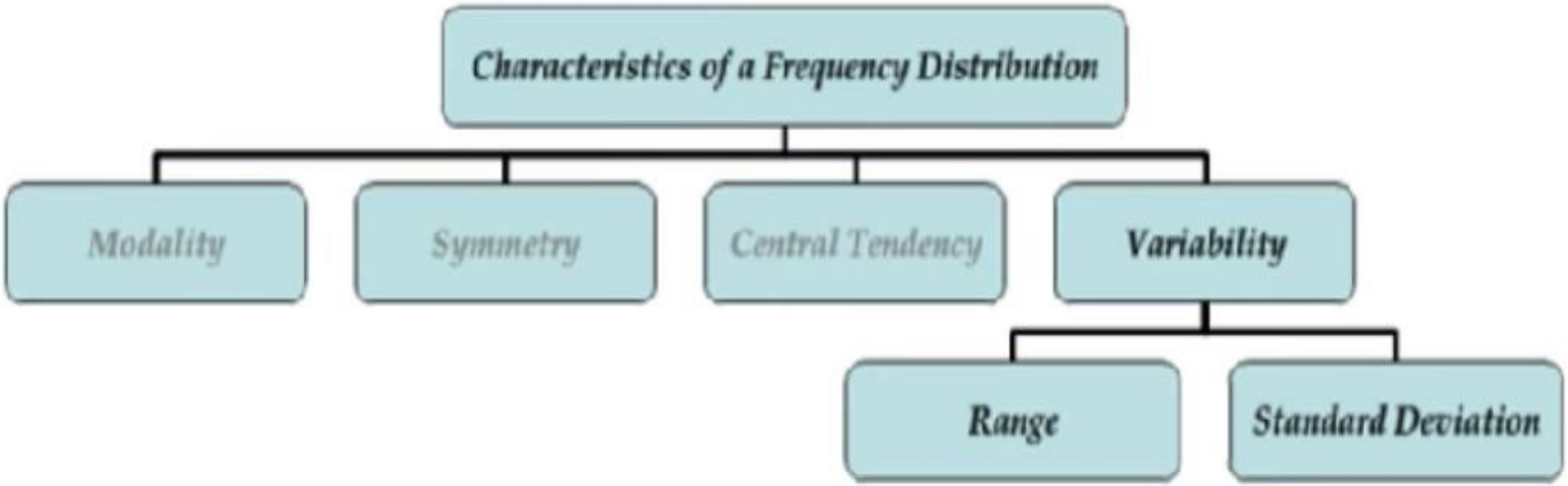


Characteristics of a Frequency Distribution
Symmetry
Symmetric
Asymmetric
Positive Skewness
Negative Skewness

Central Tendency



Variability



Central Tendency

A measure of **Central Tendency** is a single value that attempts to describe a set of data **by identifying the central position** within that set of data. In other words, the Central Tendency computes the “center” around which the data is distributed.

- **The reliable quantity**

Mean

$$\text{Mean, } \mu = \frac{\Sigma x}{n}$$

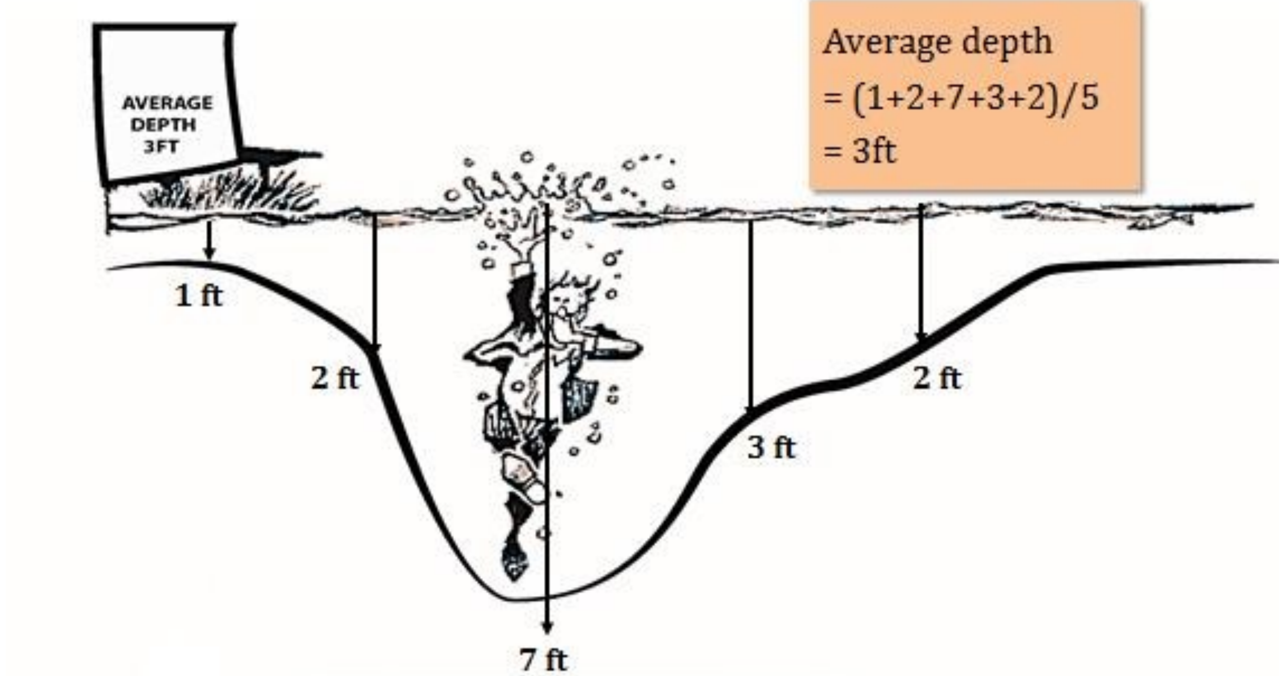
Alan went for a trek. On the way, he had to cross a stream. As Alan did not know swimming, he started exploring alternate routes to cross over.

Suddenly he saw a sign-post, which said "Average depth 3 feet". Alan was 5'7" tall and thought he could safely cross the stream.

Alan never reached the other end and drowned in the stream.

Why did Alan Drown?

Why did Alan Drown?



Average depth
= (1+2+7+3+2)/5
= 3ft

Beware of Averages!!!

The "Hotshot" Sales Executive

Kurt works as a sales manager at vsellhomes.com. In the monthly sales review, Kurt reports that he will achieve his quarterly target of $1M.

Kurt claims his average deal size is $100,000 and he has 10 deals in his pipeline. Kurt's boss Ross is very delighted with his numbers.

At the end of quarter, even after closing 8 deals Kurt fails to meet his target number and falls short by more than $500,000.

Discussion

Why did Kurt fail to achieve his quarterly target?

With 10 deals in pipeline and with average deal size of $100,000 and converting 7 of those deals, how did he fail?

The Reality of the “Hotshot” Salesman

- Average deal size in pipeline

  = $100,000

Deal #	Deal Value	Deal Status
1	70,000	Open
2	50,000	Closed
3	55,000	Closed
4	60,000	Closed
5	55,000	Closed
6	50,000	Closed
7	50,000	Closed
8	60,000	Closed
9	50,000	Closed
10	5,00,000	Open

The Reality of the “Hotshot” Salesman

- Average deal size in pipeline

  = $100,000

- Deal #10 is of significantly higher value than all the other deals and impacts the average calculation

Deal #	Deal Value	Deal Status
1	70,000	Open
2	50,000	Closed
3	55,000	Closed
4	60,000	Closed
5	55,000	Closed
6	50,000	Closed
7	50,000	Closed
8	60,000	Closed
9	50,000	Closed
10	5,00,000	Open

Median

Median

Median: Arrange data in increasing order and find the mid-point $\frac{(n+1)}{2}$.

The Reality of the "Hotshot" Salesman

- Average deal size in pipeline

  = $100,000

- Deal #10 is of significantly higher value than all the other deals and impacts the average calculation

- Median = $55,000 more realistic measure

Deal #	Deal Value	Deal Status
1	70,000	Open
2	50,000	Closed
3	55,000	Closed
4	60,000	Closed
5	55,000	Closed
6	50,000	Closed
7	50,000	Closed
8	60,000	Closed
9	50,000	Closed
10	5,00,000	Open

The Reality of the “Hotshot” Salesman

- Average deal size in pipeline = $100,000
- Deal #10 is of significantly higher value than all the other deals and impacts the average calculation
- Median = $55,000 more realistic measure

Deal #	Deal Value	Deal Status
1	70,000	Open
2	50,000	Closed
3	55,000	Closed
4	60,000	Closed
5	55,000	Closed
6	50,000	Closed
7	50,000	Closed
8	60,000	Closed
9	50,000	Closed
10	5,00,000	Open

Median is less susceptible to the influence of Outliers.

Mode

Mode

Mode – the most frequently occurring

Central Tendency: Example

- Timing for the Men's 500-meter Speed Skating event in Winter Olympics is tabulated.
- The Central Tendency measures are computed below:

Year	Time
1928	43.4
1932	43.4
1936	43.4
1948	43.1
1952	43.2
1956	40.2
1960	40.2
1964	40.1
1968	40.3
1972	39.44
1976	39.17
1980	38.03
1984	38.19
1988	36.4

Mean
= (43.4+...+36.4)/14
= 568.53/14
= 40.61

Year	Time
1988	36.4
1980	38.03
1984	38.19
1976	39.17
1972	39.44
1964	40.1
1956	40.2
1960	40.2
1968	40.3
1948	43.1
1952	43.2
1928	43.4
1932	43.4
1936	43.4

Median
= (7^{th} + 8^{th} Value)/2
= (40.2+40.2)/2
= 40.2

Year	Time
36.4	1
38.03	1
38.19	1
39.17	1
39.44	1
40.1	1
40.2	2
40.3	1
43.1	1
43.2	1
43.4	3

Mode
= Value with highest frequency
= 43.4

Player_A Vs Player_B – Who is Better ?

Match	Player A	Player B
1	40	40
2	40	35
3	7	45
4	40	52
5	0	30
6	90	40
7	3	29
8	11	43
9	120	37

Player_A Vs Player_B – Who is Better ?

Match	Player A	Player B
1	40	40
2	40	35
3	7	45
4	40	52
5	0	30
6	90	40
7	3	29
8	11	43
9	120	37
SUM	351	351

Player_A Vs Player_B – Who is Better ?

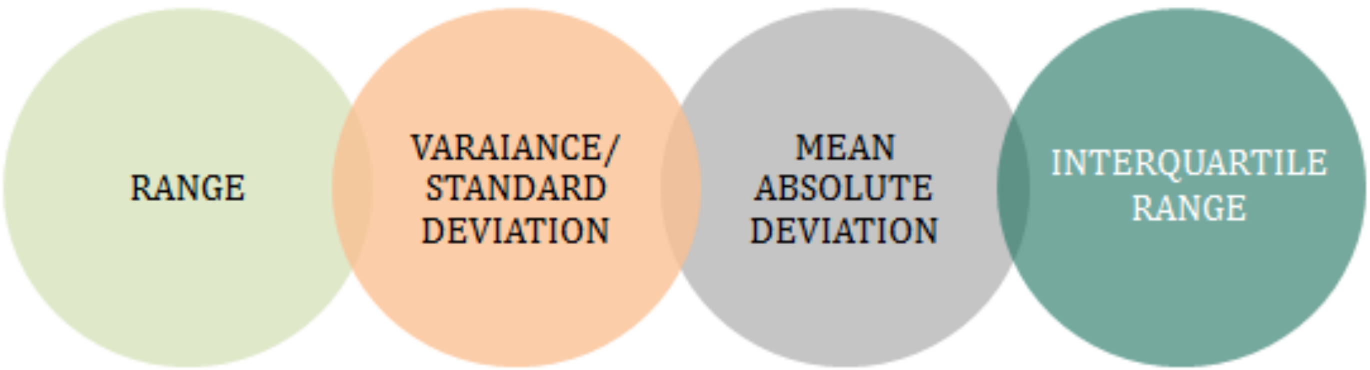
Match	Player A	Player B
1	40	40
2	40	35
3	7	45
4	40	52
5	0	30
6	90	40
7	3	29
8	11	43
9	120	37
SUM	351	351
MEAN	39	39

Player_A Vs Player_B – Who is Better ?

Match	Player A	Player B
1	40	40
2	40	35
3	7	45
4	40	52
5	0	30
6	90	40
7	3	29
8	11	43
9	120	37
SUM	351	351
MEAN	39	39
MEDIAN	40	40

Dispersion Measures

Measures of Dispersion describe the data spread or how far the measurements are from the center.



Spread of Data - Range

Range = Max - Min

Spread of Data - SD and Variance

$$\text{Variance} = \frac{\Sigma(x-\mu)^2}{n}$$

$$\text{Standard Deviation, } \sigma = \sqrt{Variance}$$

Who's Best?

Match	Player A	Player B
1	40	40
2	40	35
3	7	45
4	40	52
5	0	30
6	90	40
7	3	29
8	11	43
9	120	37
SUM	351	351
MEAN	39	39
MEDIAN	40	40
STANDARD DEVIATION	41.5180683558376	7.28010988928052

Measuring Variability and Spread

Basketball coach Statson is in a dilemma choosing between 3 players all having the same average scores.

Points scored per game	**7**	**8**	**9**	**10**	**11**	**12**	**13**
Frequency, f	1	1	2	2	2	1	1

Points scored per game	**7**	**9**	**10**	**11**	**13**
Frequency, f	1	2	4	2	1

Points scored per game	**3**	**6**	**7**	**10**	**11**	**13**	**30**
Frequency, f	2	1	2	3	1	1	1

Measuring Variability and Spread

Basketball coach Statson is in a dilemma choosing between 3 players all having the same average scores.

Points scored per game	**7**	**8**	**9**	**10**	**11**	**12**	**13**
Frequency, f	1	1	2	2	2	1	1

Points scored per game	**7**	**9**	**10**	**11**	**13**
Frequency, f	1	2	4	2	1

Points scored per game	**3**	**6**	**7**	**10**	**11**	**13**	**30**
Frequency, f	2	1	2	3	1	1	1

Mean = Median = Mode = 10 for all 3.

Measuring Variability and Spread

Range = Max - Min

Points scored per game	**7**	**8**	**9**	**10**	**11**	**12**	**13**
Frequency, f	1	1	2	2	2	1	1

Points scored per game	**7**	**9**	**10**	**11**	**13**
Frequency, f	1	2	4	2	1

Points scored per game	**3**	**6**	**7**	**10**	**11**	**13**	**30**
Frequency, f	2	1	2	3	1	1	1

Points scored per game	7	8	9	10	11	12	13
Frequency, f	1	1	2	2	2	1	1

Points scored per game	7	9	10	11	13
Frequency, f	1	2	4	2	1

Points scored per game	3	6	7	10	11	13	30
Frequency, f	2	1	2	3	1	1	1

MEAN = MEDIAN = MODE = 10   RANGE = 5 , 5 , 27

Points scored per game	7	8	9	10	11	12	13
Frequency, f	1	1	2	2	2	1	1

Points scored per game	7	9	10	11	13
Frequency, f	1	2	4	2	1

Points scored per game	3	6	7	10	11	13	30
Frequency, f	2	1	2	3	1	1	1

MEAN = MEDIAN = MODE = 10 RANGE = 5 , 5 , 27 Reject Player 3

Basketball coach Statson is in a dilemma choosing between 3 players all having the same average scores.

Points scored per game	7	8	9	10	11	12	13
Frequency, f	1	1	2	2	2	1	1

Points scored per game	7	9	10	11	13
Frequency, f	1	2	4	2	1

STANDARD DEVIATION

Player 1 = **1.7873008824606**
Player 2 = **3.30823887354653**

What is your Decision?????????

Percentile & Quartile

<u>N**th** percentile</u> states that there are atleast N% of values less than or equal to this value **and** <u>(100-N) </u>values are greater or equal to this value

i = (N/100)*n

N – The percentile you are interested

n – Number of values

Key points

1. If i is decimal then round off to next value
2. If i is integer then take average of **i and i+1** value

Let's calculate 85th percentile

Data:

3310 3355 3450 3480 3480 3490 3520 3540 3550 3650 3730 3925

Calculate 85th percentile ?

Quartile

Data:

3310 3355 3450 3480 3480 3490 3520 3540 3550 3650 3730 3925

Quartile

Dividing data into ¼ – 4 parts

Q1 – First Quartile – 25^{th} percentile

Q2 – Second Quartile – 50^{th} percentile (Median)

Q3 – Third Quartile – 75^{th} percentile

IQR (Inter Quartile Range) = Q3 – Q1

Inter Quartile Range

Quartile

Dividing data into ¼ – 4 parts

Q1 – First Quartile – 25th percentile

Q2 – Second Quartile – 50th percentile (Median)

Q3 – Third Quartile – 75th percentile

IQR (Inter Quartile Range) = Q3 – Q1

Case Study

In an Under 19 World Cup selection squad for 2018 the BCCI needs to select 1 player based on the current performance in 2017 – 2018 Ranji Trophy. There are 2 players with similar stats and the board is not sure whom to select.

- Can you help the board members with your analysis ?

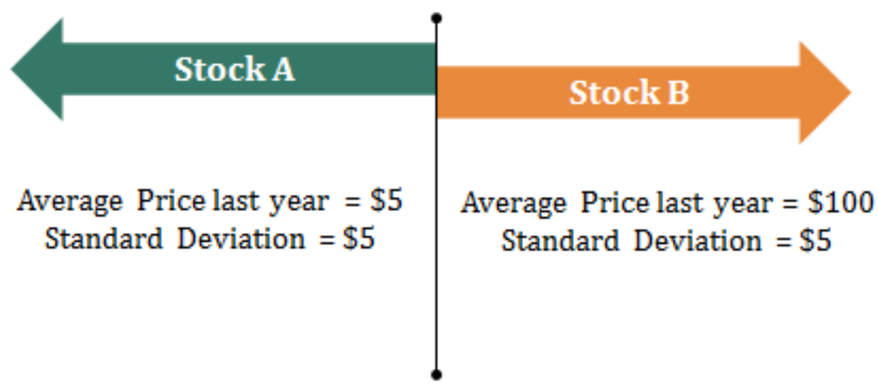
Stats - Player X & Y

Runs scored by both players in last 14 matches

Player X	Player Y
40	35
20	40
5	7
20	23
10	20
75	26
100	12
25	30
15	27
15	102
20	18
17	17
11	14
5	7

Coefficient of Variation

Coeff of Variation = (Standard deviation/ Mean) * 100 %



Coefficient of Variation:

Stock A: CV = 100%

(5/5*100=100%)

Stock B: CV = 5%

(5/100*100=5%)

$$CV = \left(\frac{S}{\overline{X}} \right) \cdot 100\%$$

Coefficient of Variation

Calculate the descriptive statistics of both players and if the coefficient of variation is greater than 85% then drop that player

Coeff of Variation = (Standard deviation/ Mean) * 100 %

Measures ofassociation between 2 variables

1. Covariance

2. Correlation coefficient

Covariance

$$Cov(X,Y) = \frac{\sum (X_i - \overline{X}) * (Y_i - \overline{Y})}{n}$$

Higher the value stronger the relation between them

Correlation coefficient

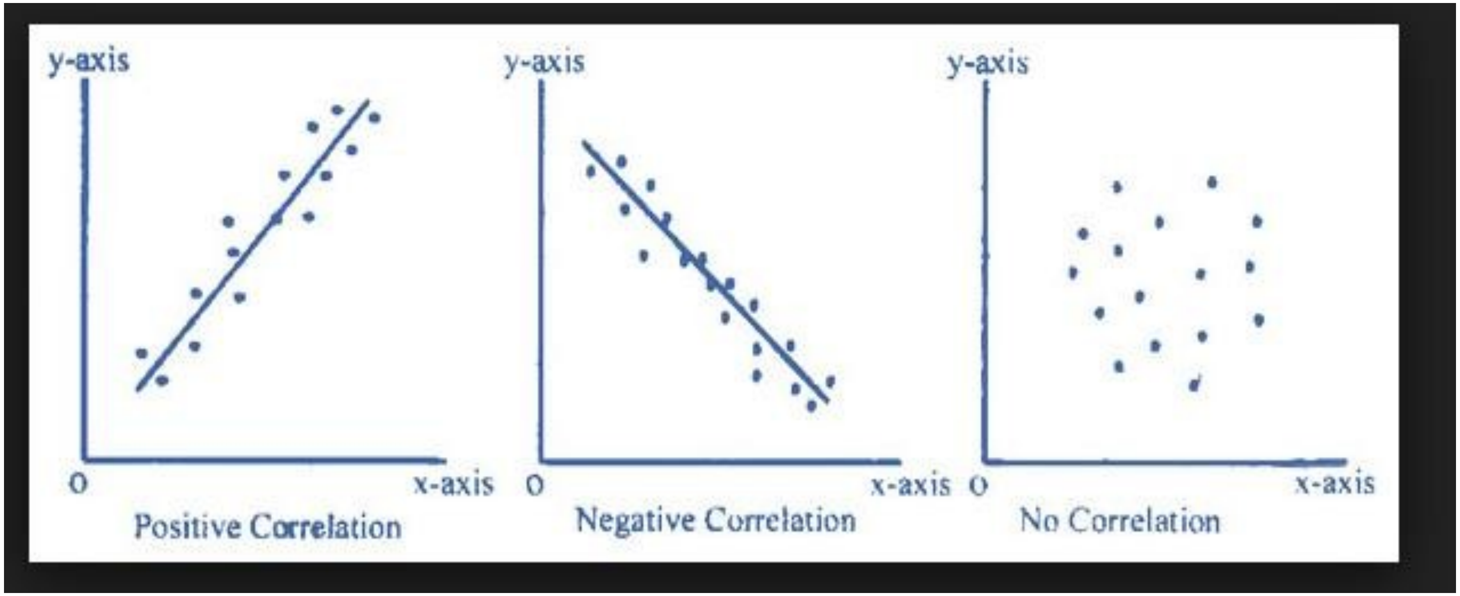
$$r_{xy} = \frac{\mathrm{Cov}(x, y)}{S_x \times S_y}$$

Key Points

1.A measure of relationship not affected by the units of measurements

2. Ranges from -1 to +1

Types of Correlation



Central Limit Theorem

When samples of size **n>=30** are drawn from a population and distributed with individual samples mean then any distribution changes to normal distribution

$$\frac{\sigma}{\sqrt{n}}$$

Key Points

1. Also called as Standard Error (SE)

Standard deviation of sample mean = **(population standard deviation/square root(n))**

2. Mean of sample means distribution = **Population mean**

NOTE: As n increases SE decreases - SE is inversely proportional to n