Statistics & Probablities

Sharique Nawaz

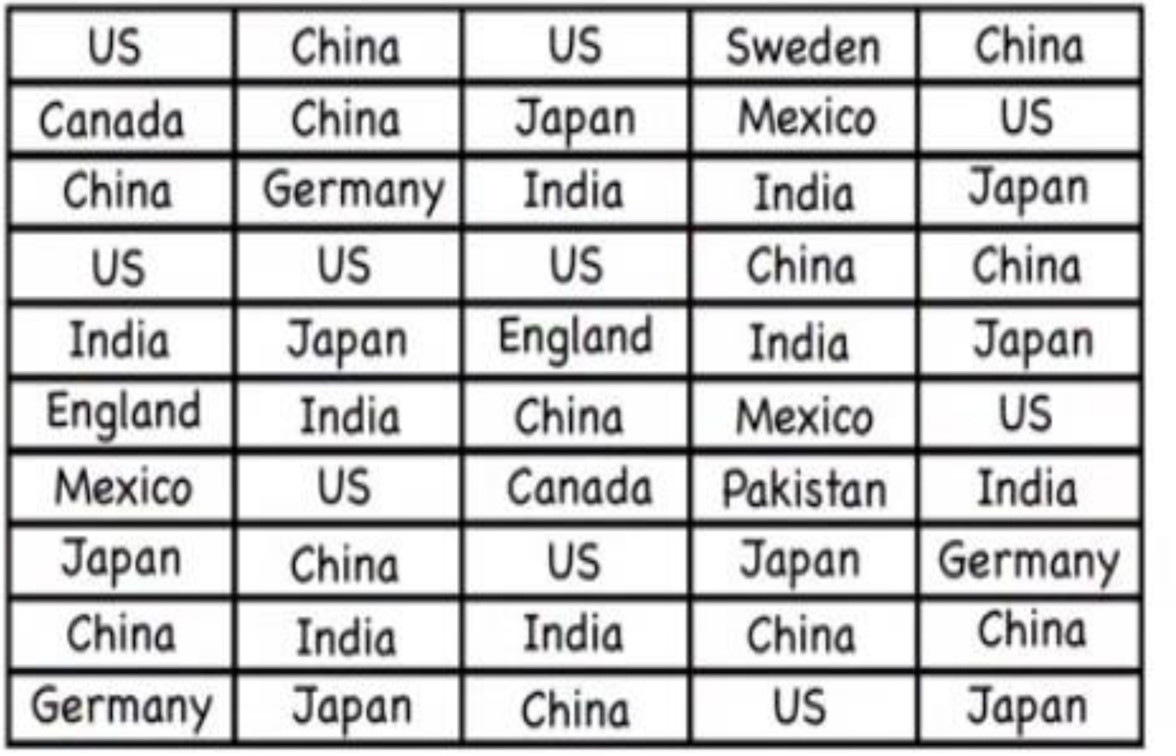
What is Statistics ?

The science of making decisions and knowing statistics can help you make better decisions through out life

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

Answer in 5 seconds

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



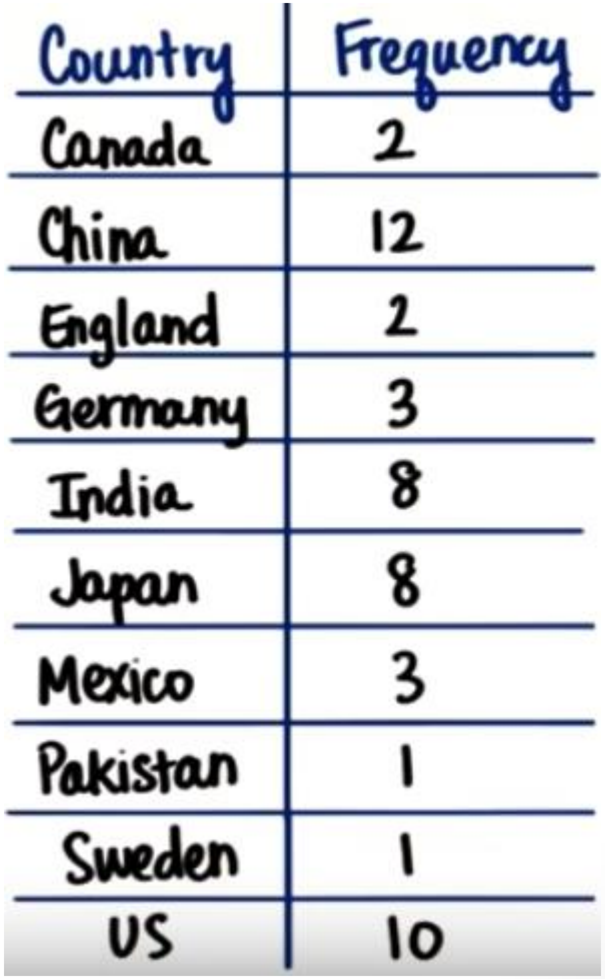
US	China	US	Sweden	China
Canada	China	Japan	Mexico	US
China	Germany	India	India	Japan
US	US	US	China	China
India	Japan	England	India	Japan
England	India	China	Mexico	US
Mexico	US	Canada	Pakistan	India
Japan	China	US	Japan	Germany
China	India	India	China	China
Germany	Japan	China	US	Japan

Frequency Table

Properties of RF

1. The range of proportions is between 0 and 1

2. Sum of relative frequencies =1



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

Case Study

Problem

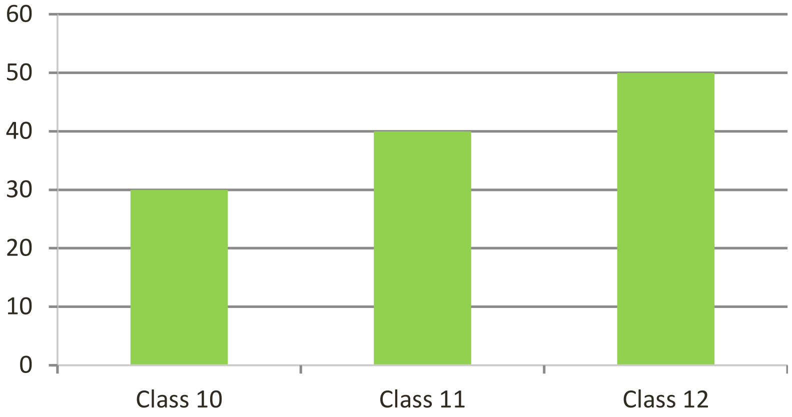
A parent changes school of their Son who is studying in 11th standard since his academic results are not good in the current School. They change Student A from **ABC school to XYZ school**

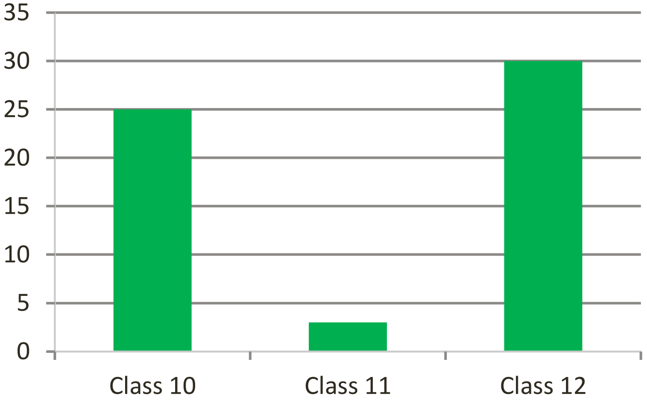
Results

1. Ranked 15th in ABC school
2. Ranked 2nd in XYZ school

What's the conclusion: Has the student improved ?

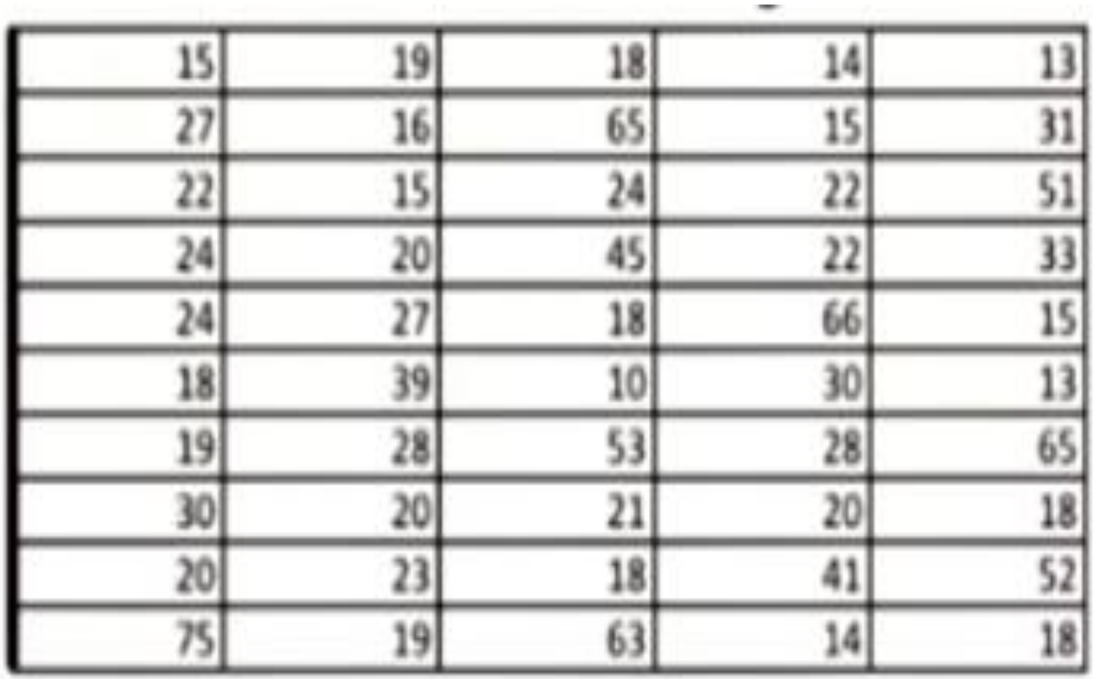
Number of Students


No of Students in ABC School


No of Students in XYZ School

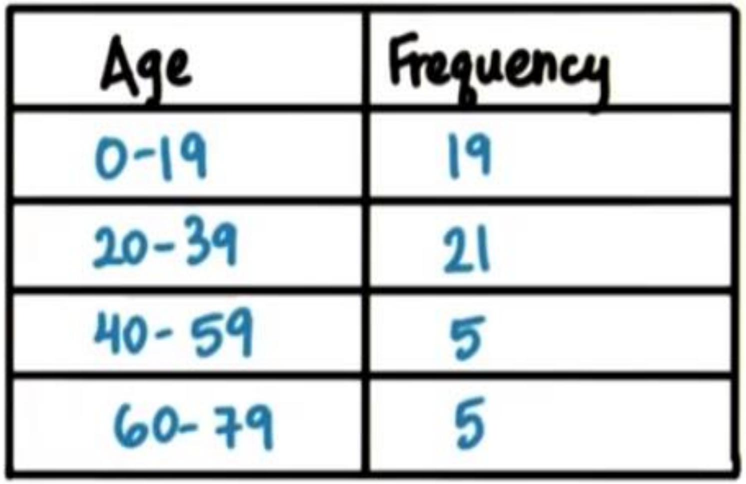
What's the most common Age ?

Students Age's

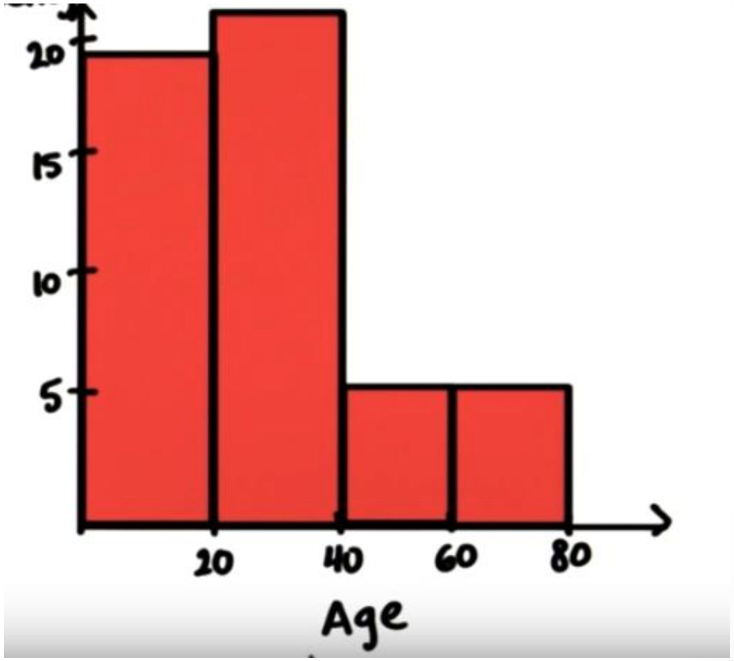


15	19	18	14	13
27	16	65	15	31
22	15	24	22	51
24	20	45	22	33
24	27	18	66	15
18	39	10	30	13
19	28	53	28	65
30	20	21	20	18
20	23	18	41	52
75	19	63	14	18

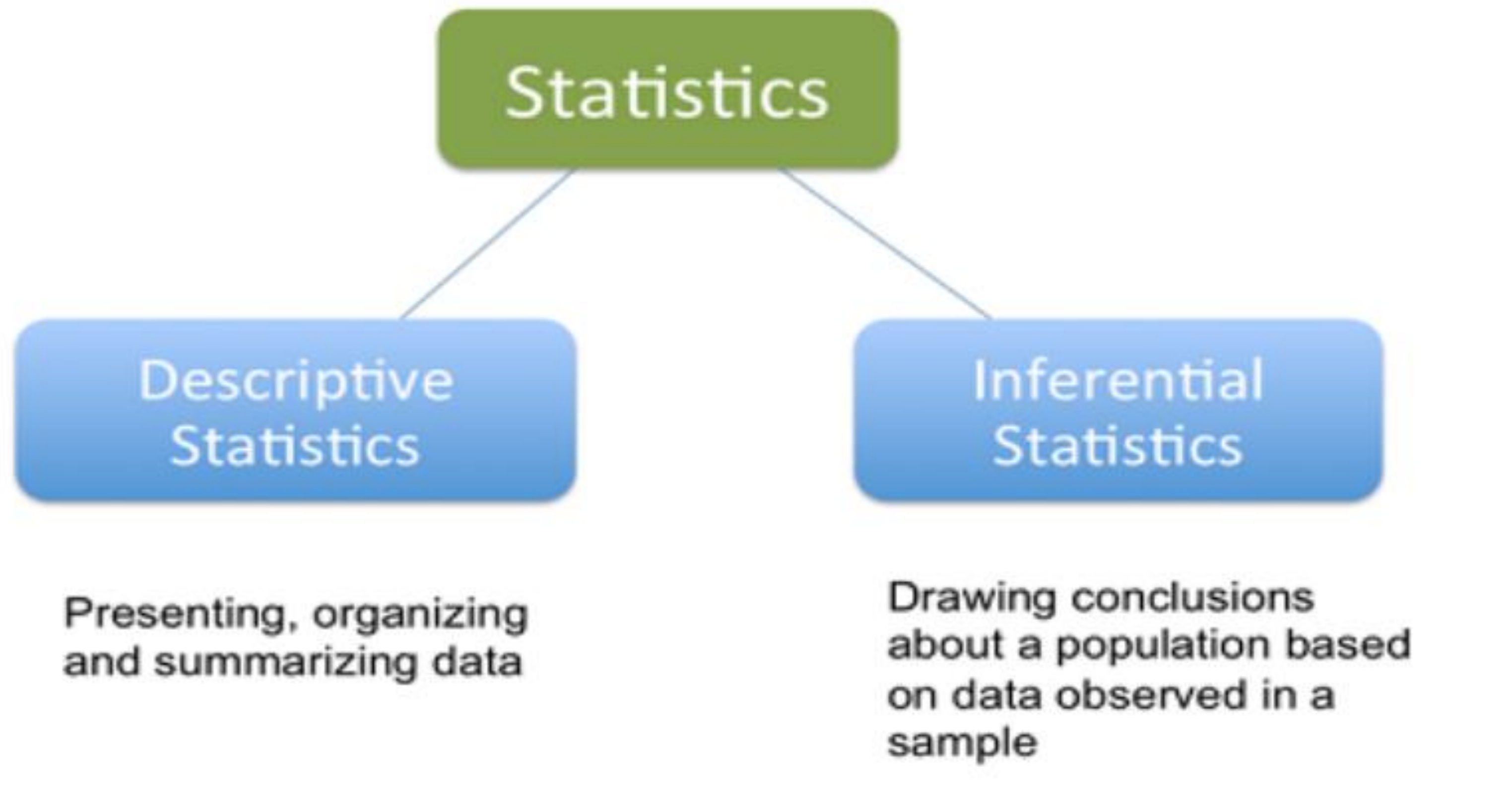
Converting Data to Range – Histogram plot



Age	Frequency
0-19	19
20-39	21
40-59	5
60-79	5



Classification



Statistics

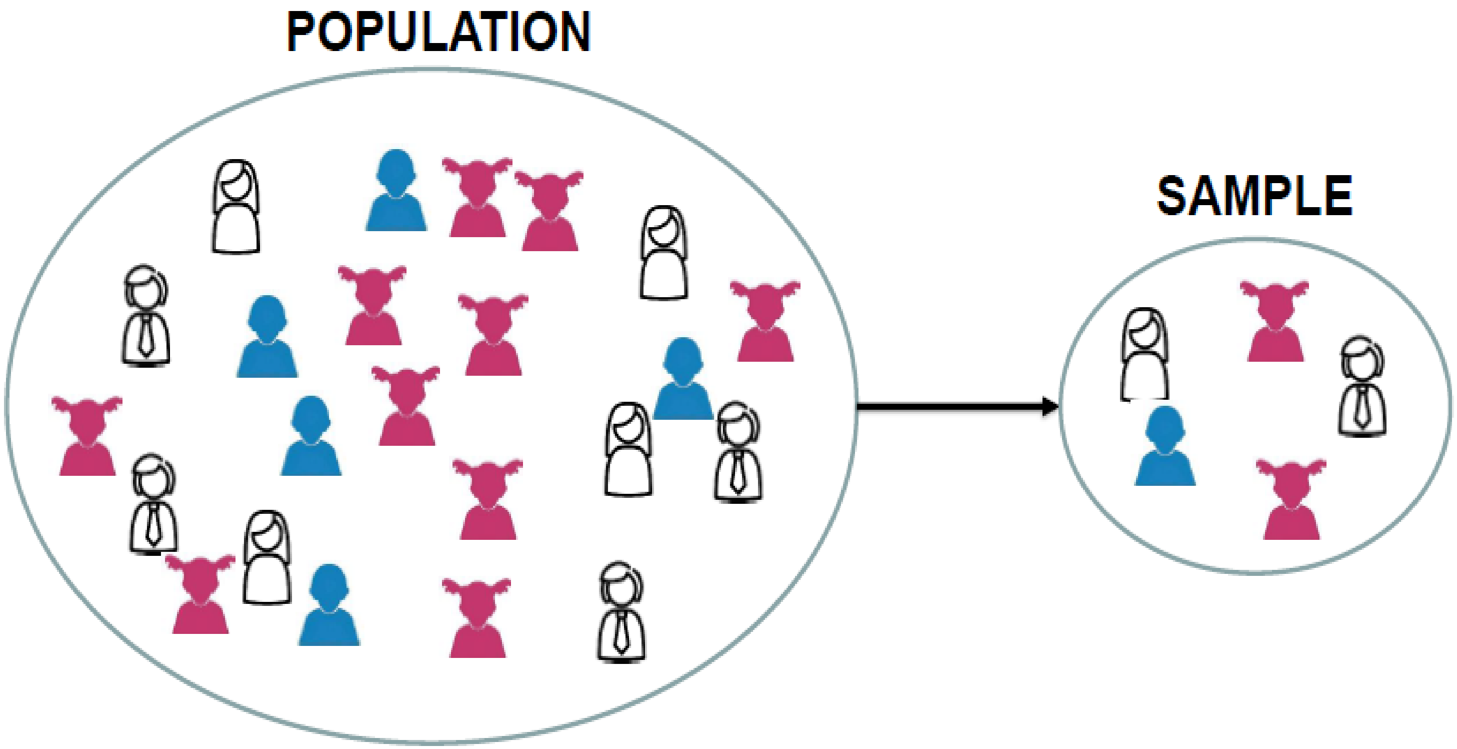
Descriptive Statistics

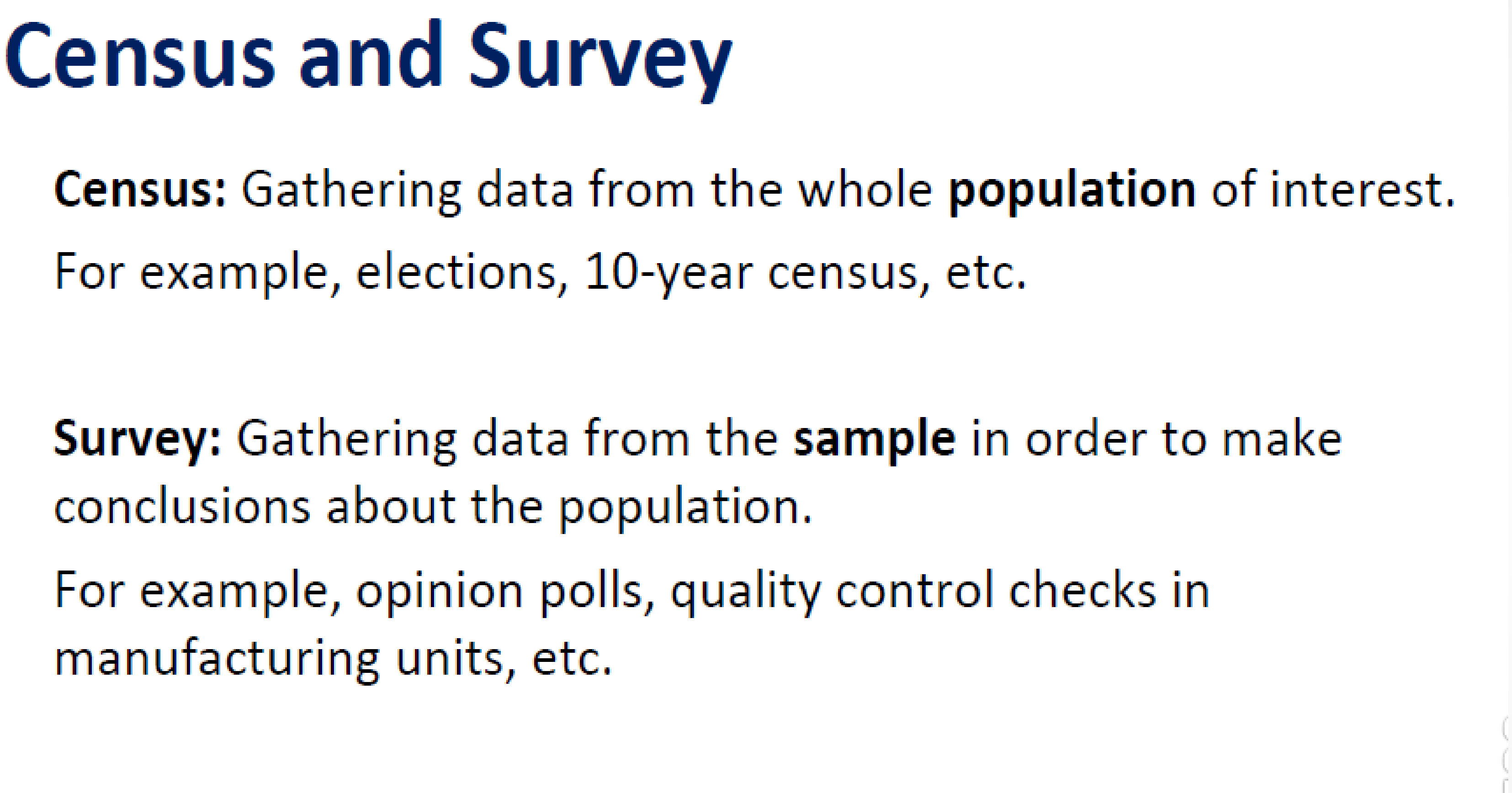
Inferential Statistics

Presenting, organizing and summarizing data

Drawing conclusions about a population based on data observed in a sample

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

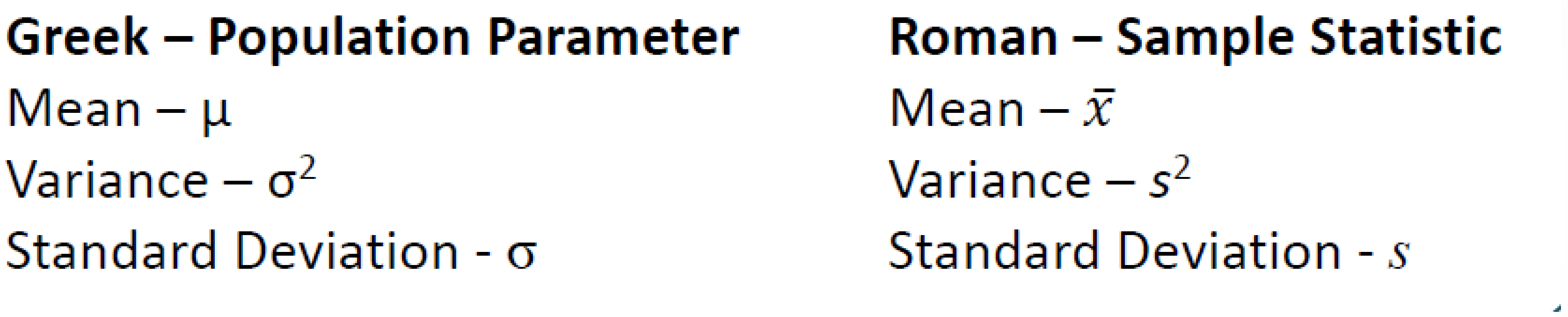
Parameter: A descriptive measure of the **population**.

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.

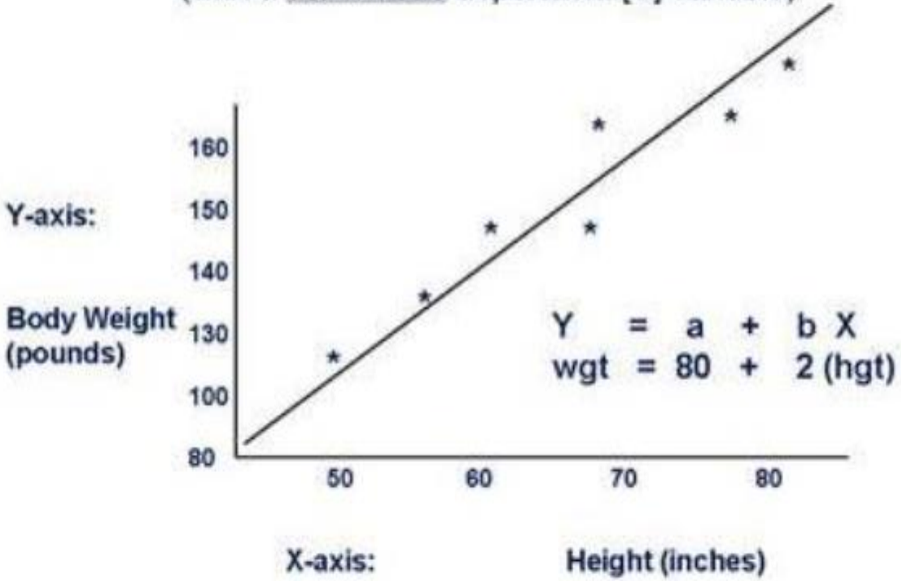
Statistical Notations



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

Variables - Dependent and Independent

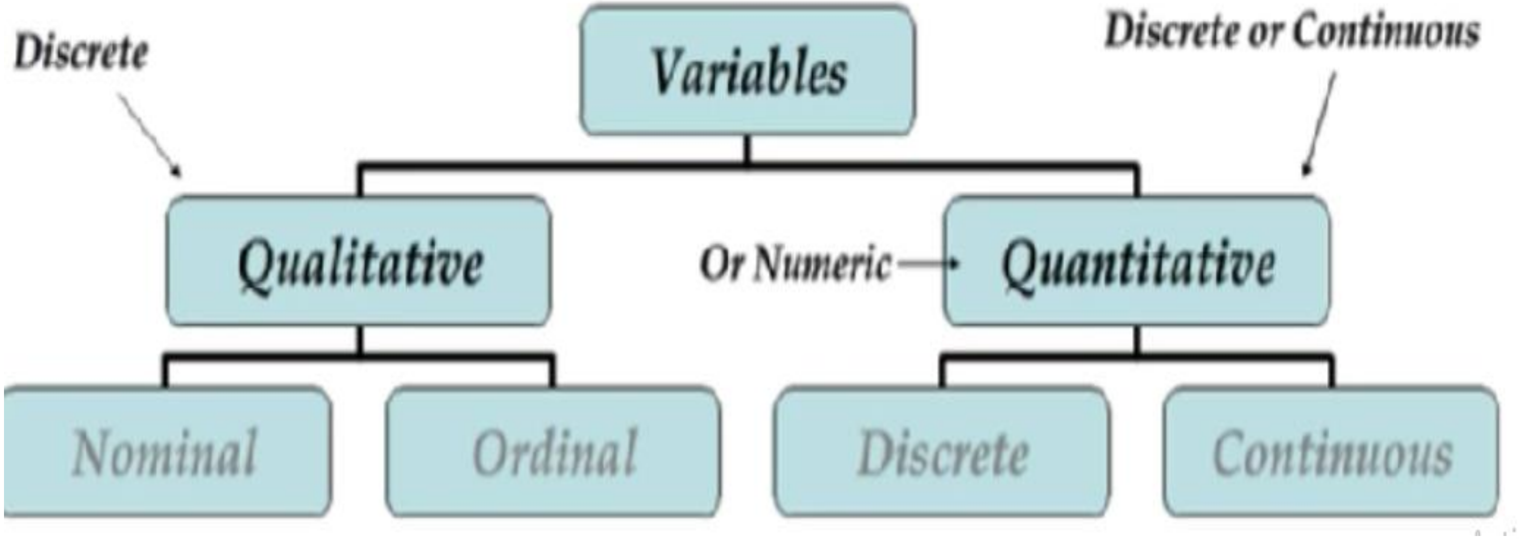
- Dependent variables on y-axis and Independent on x-axis.
- Dependent variable also called Target variable or Class variable.



Simple Linear Regression

(with a *continuous* dependent [Y] variable)

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

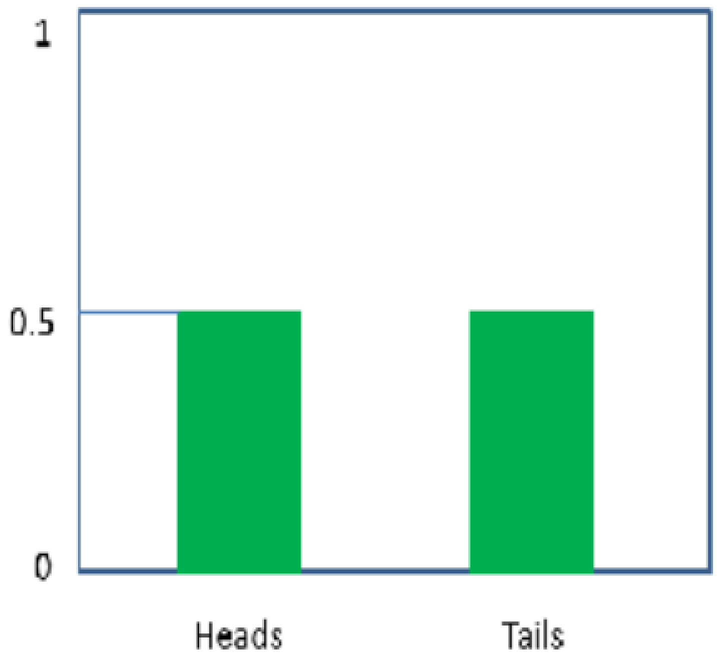
Ordinal

Examples

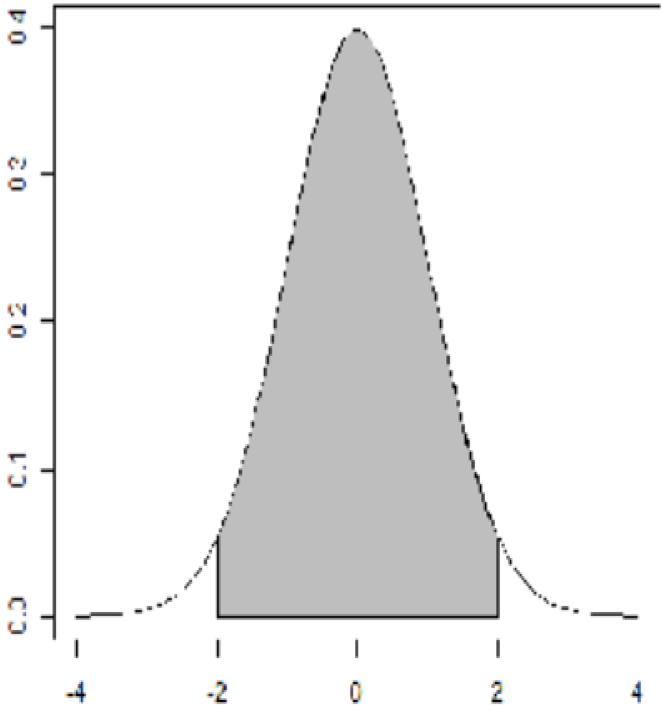
- Mutual fund risk ratings • Fortune 50 rankings
- Movie ratings

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

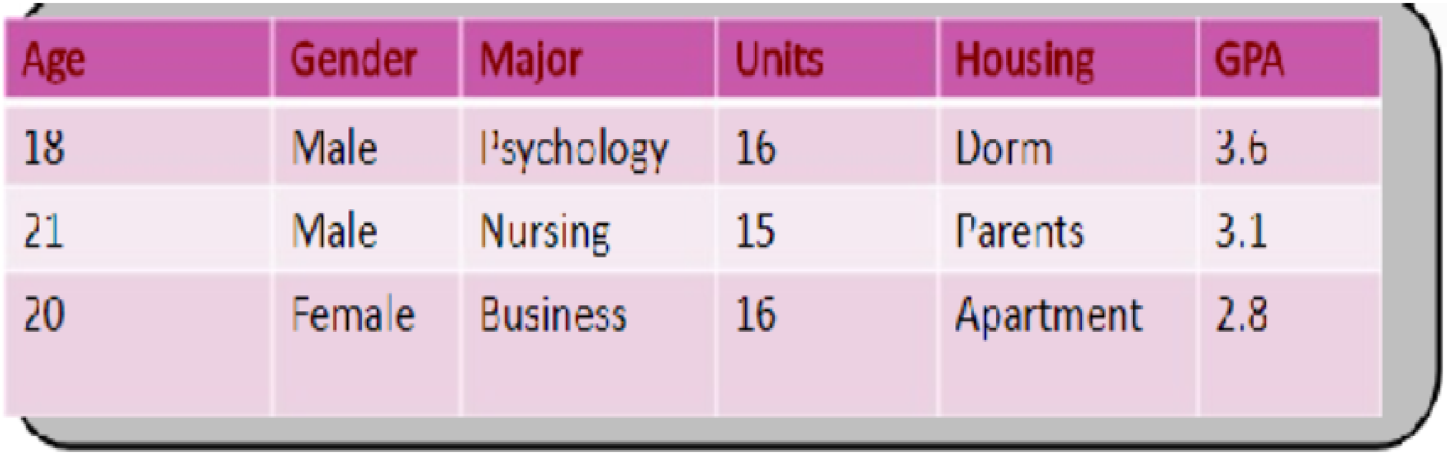
  Discrete
- Lengths of newly designed automobiles -

  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?

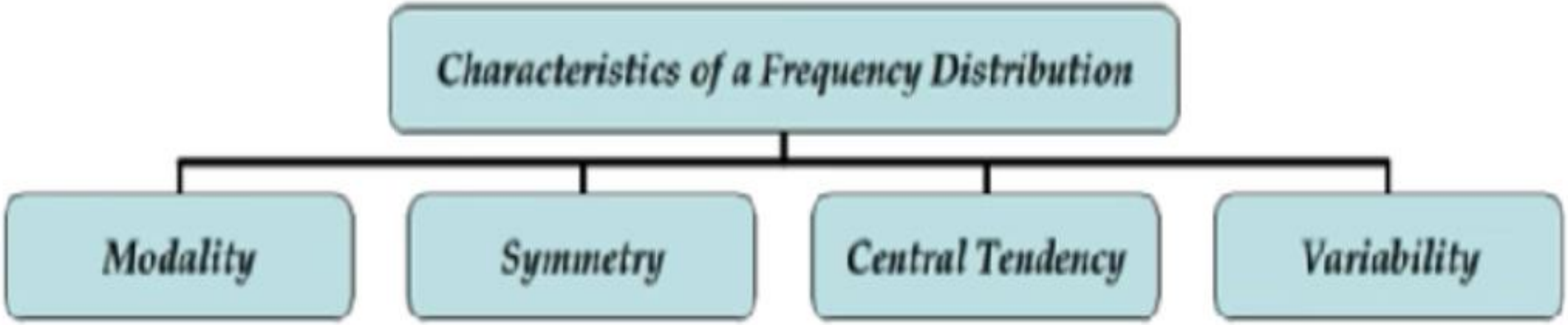


Age	Gender	Major	Units	Housing	GPA
18	Male	Psychology	16	Dorm	3.6
21	Male	Nursing	15	Parents	3.1
20	Female	Business	16	Apartment	2.8

- Numerical
  - Age
  - Units
  - GPA

- ❑ Categorical
  - ❑ Gender
  - ❑ Major
  - ❑ Housing

Summarizing Data



Characteristics of a Frequency Distribution

Modality

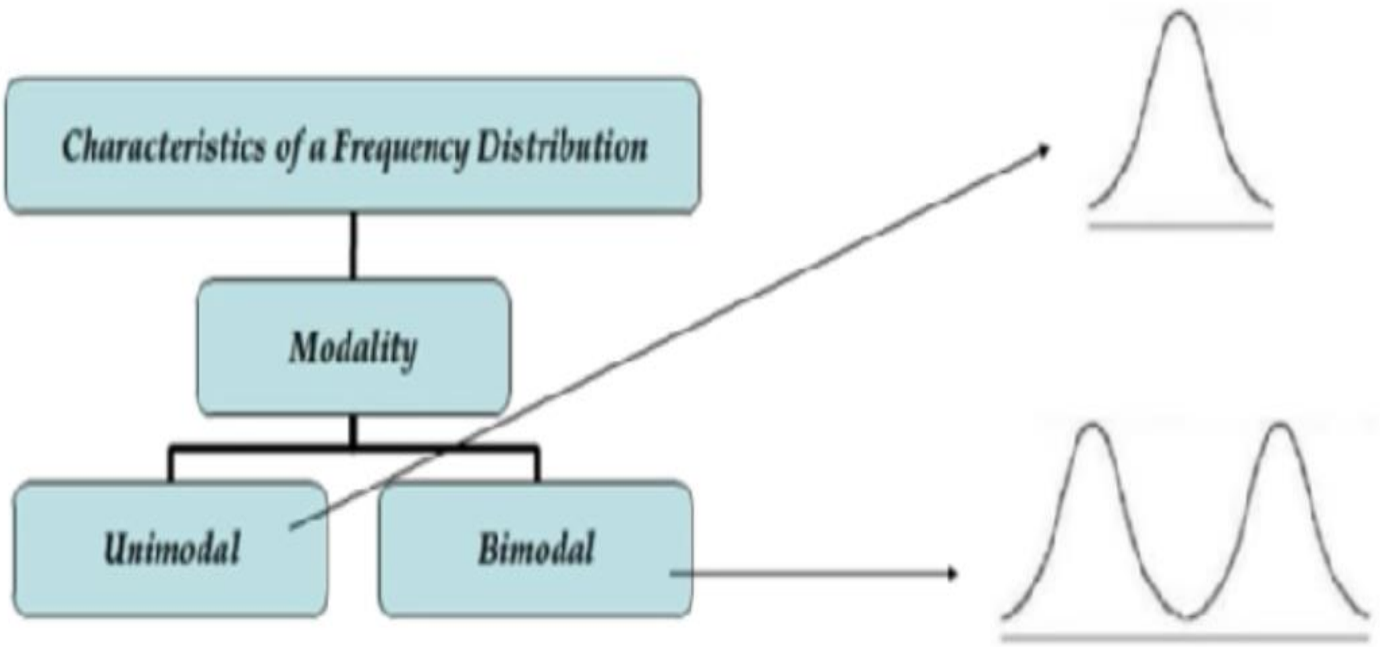
Symmetry

Central Tendency

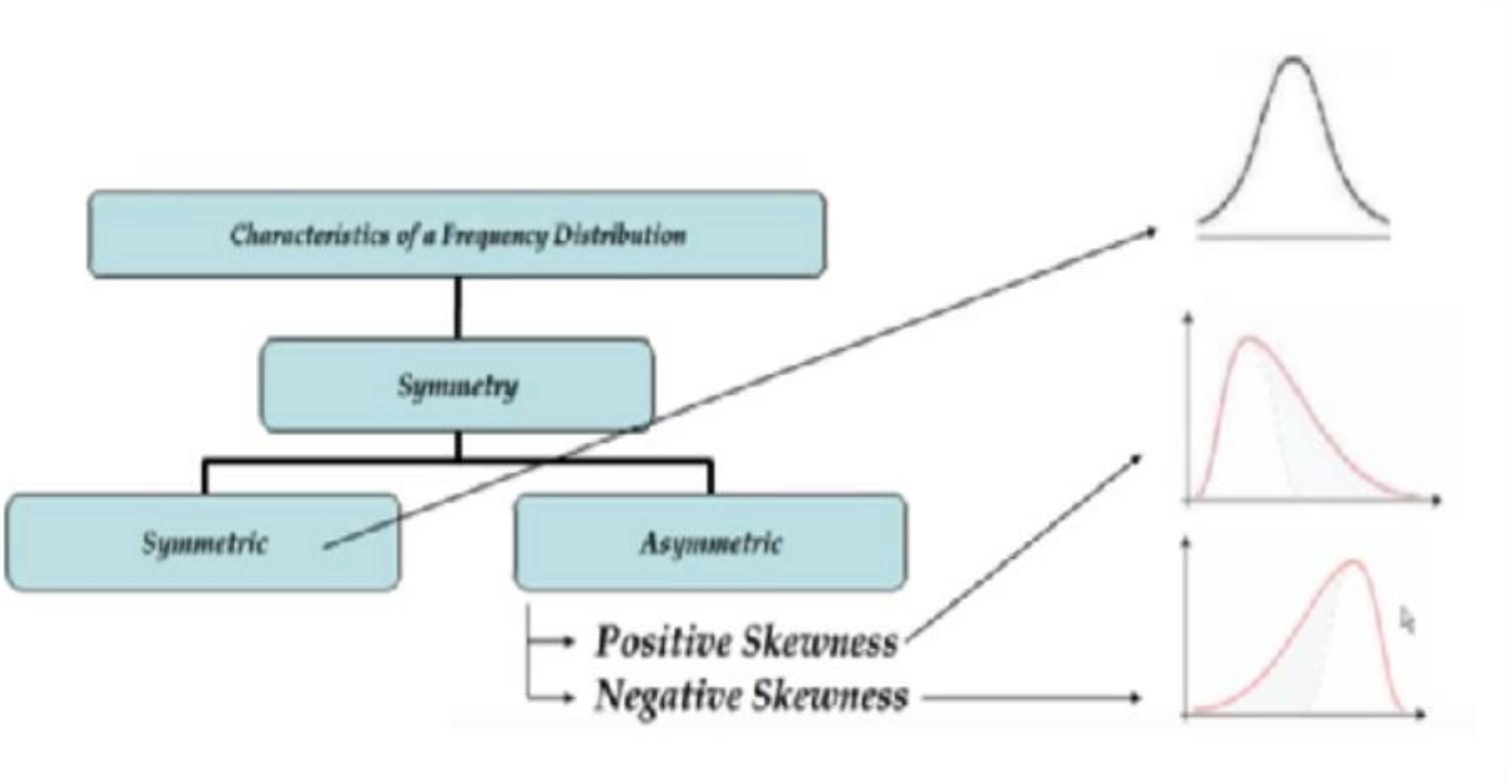
Variability

1. Frequency Distribution
2. Bar Chart
3. Histogram

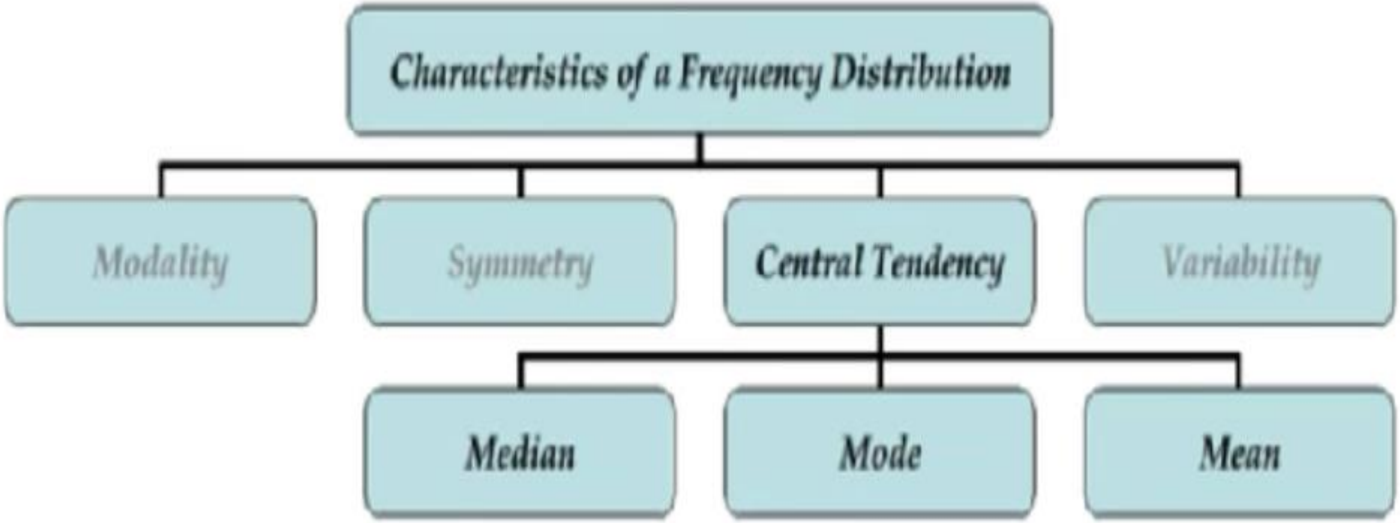
Modality


Characteristics of a Frequency Distribution
Modality
Unimodal
Bimodal

Symmetry


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

Central Tendency



Characteristics of a Frequency Distribution

Modality

Symmetry

Central Tendency

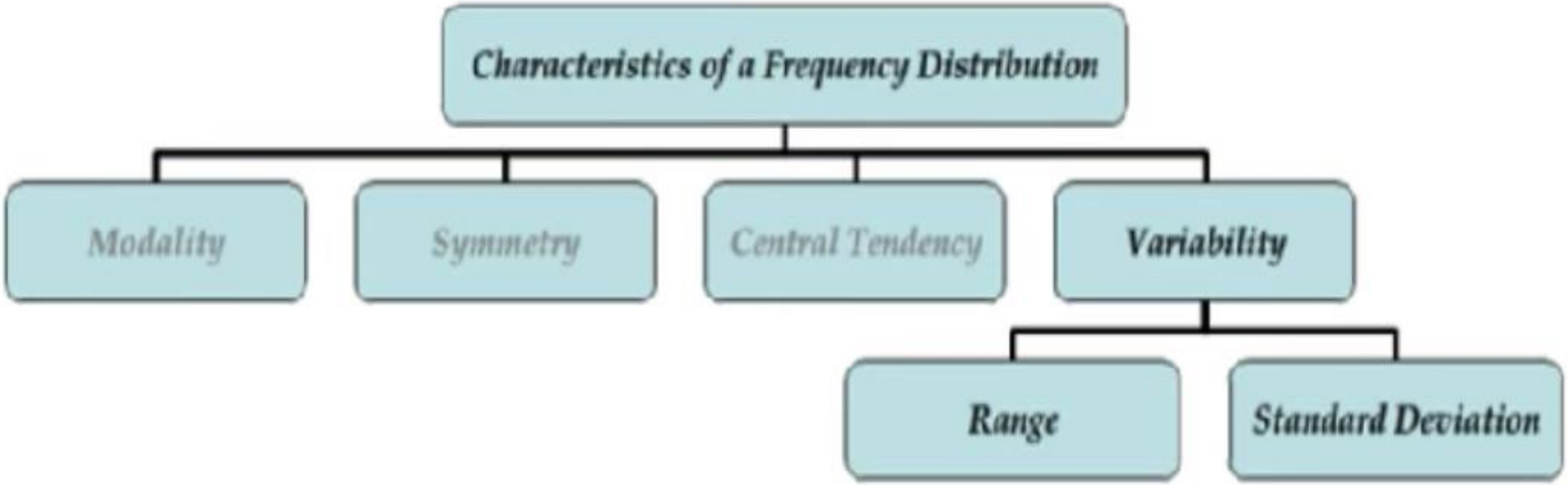
Variability

Median

Mode

Mean

Variability



Central Tendencies

- The reliable quantity

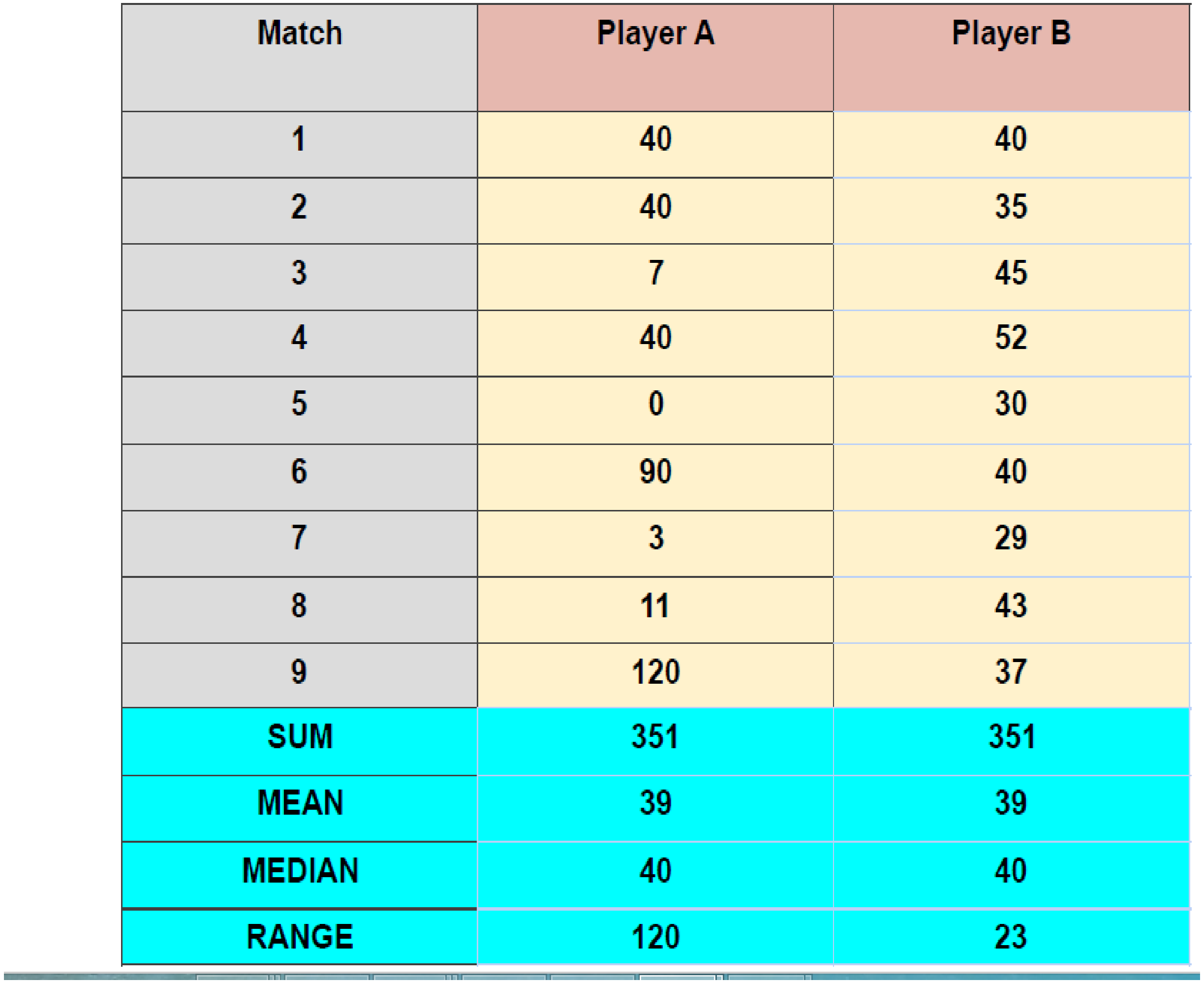
Mean – Median - Mode

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

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

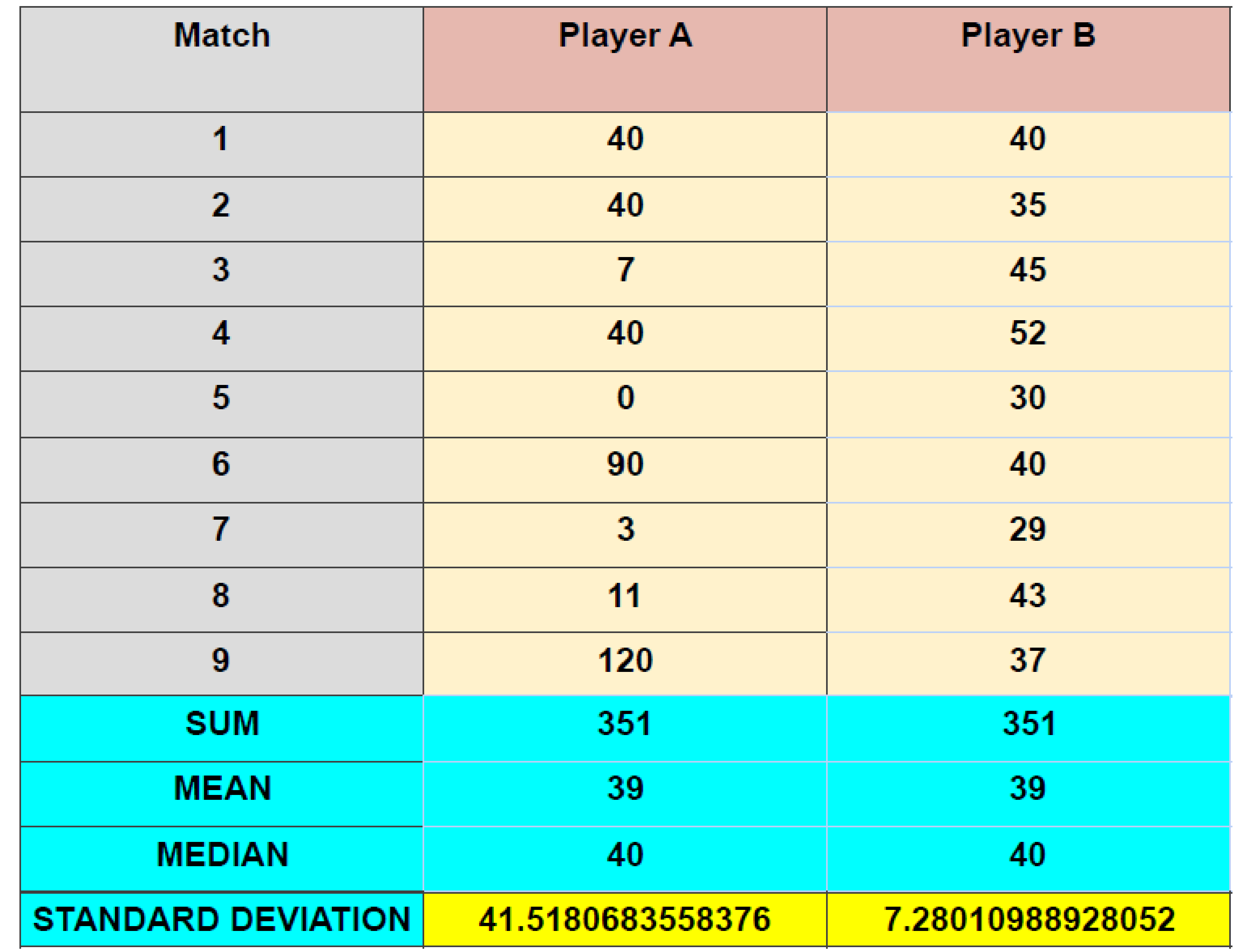
Mode – the most frequently occurring

Player A vs Player B



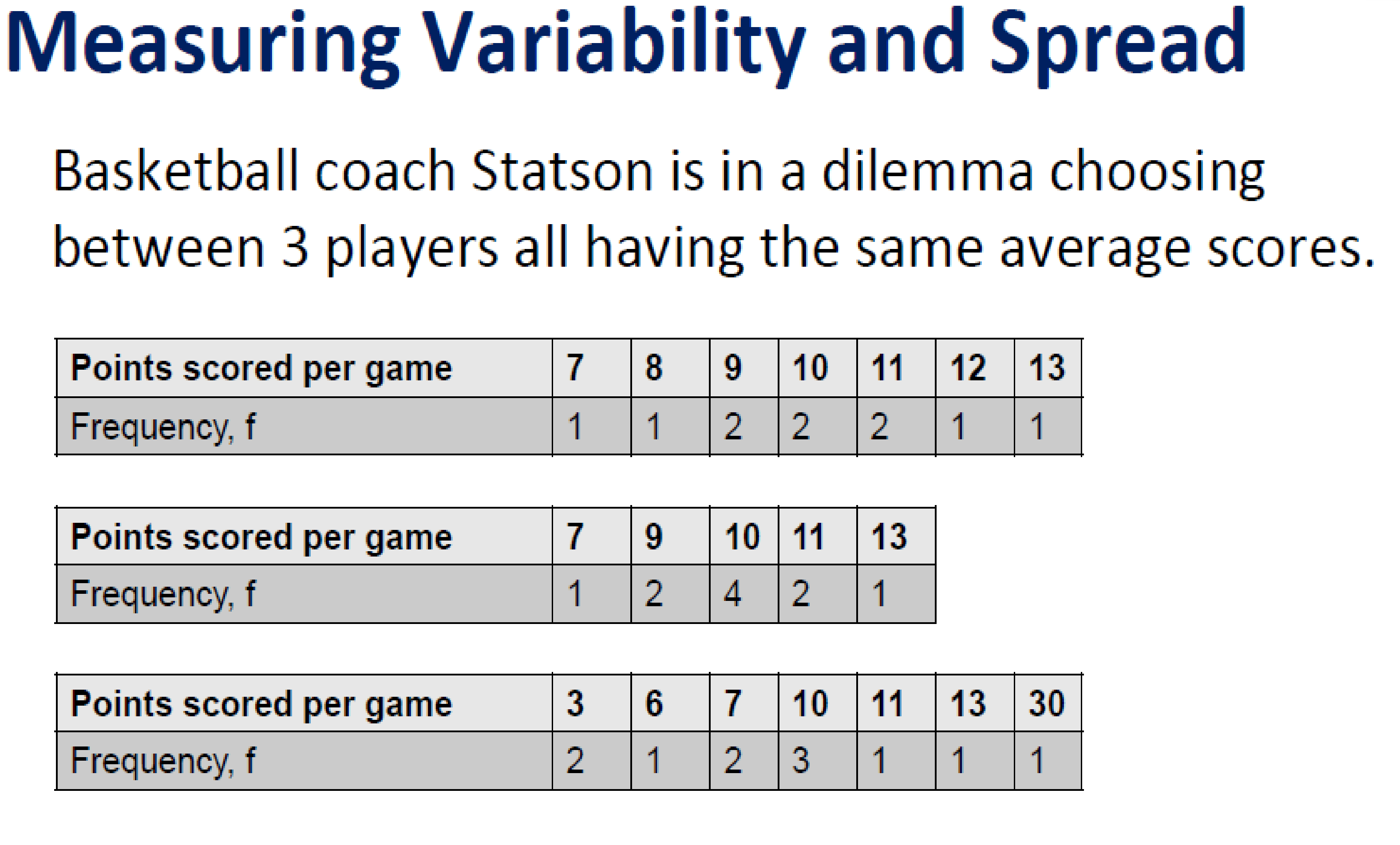
Match	Player A	Player B
1	40	40
2	40	35
3	7	45
4	40	52
5	0	30
6	90	40
7	3	29
8	11	43
9	120	37
SUM	351	351
MEAN	39	39
MEDIAN	40	40
RANGE	120	23

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

Spread of Data



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.

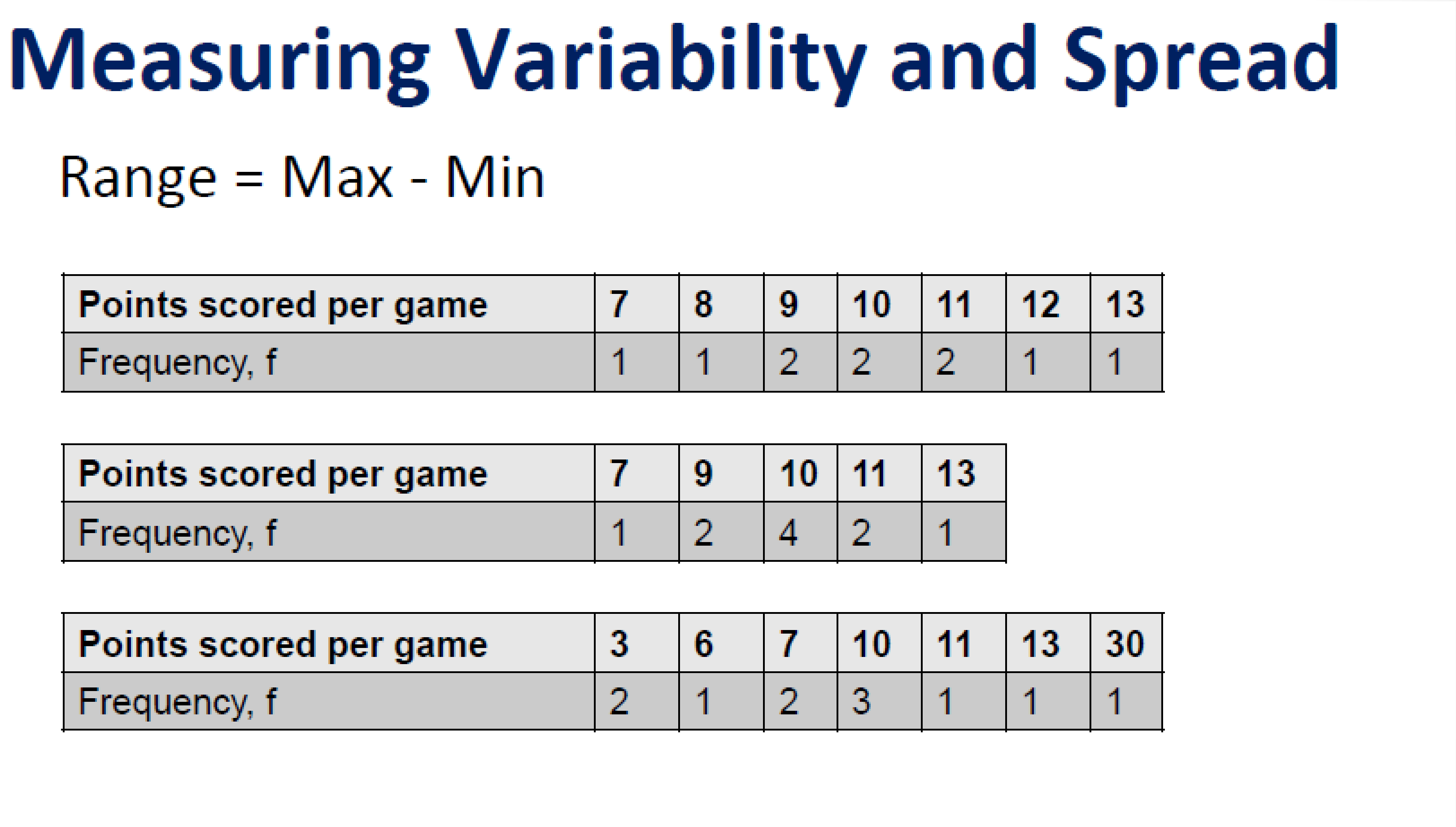
Points scored per game	**7**	**8**	**9**	**10**	**11**	**12**	**13**
Frequency, f	1	1	2	2	2	1	1

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.

Range



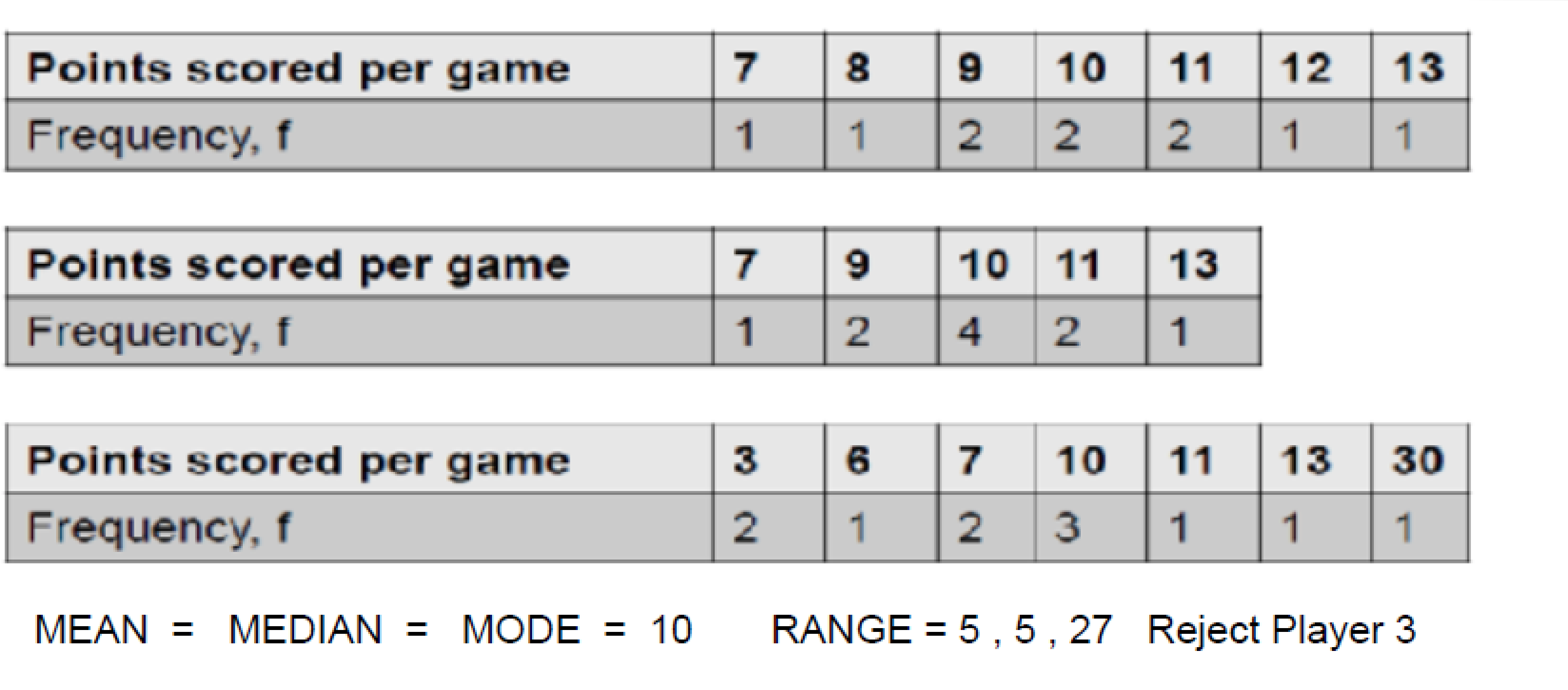
Measuring Variability and Spread

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



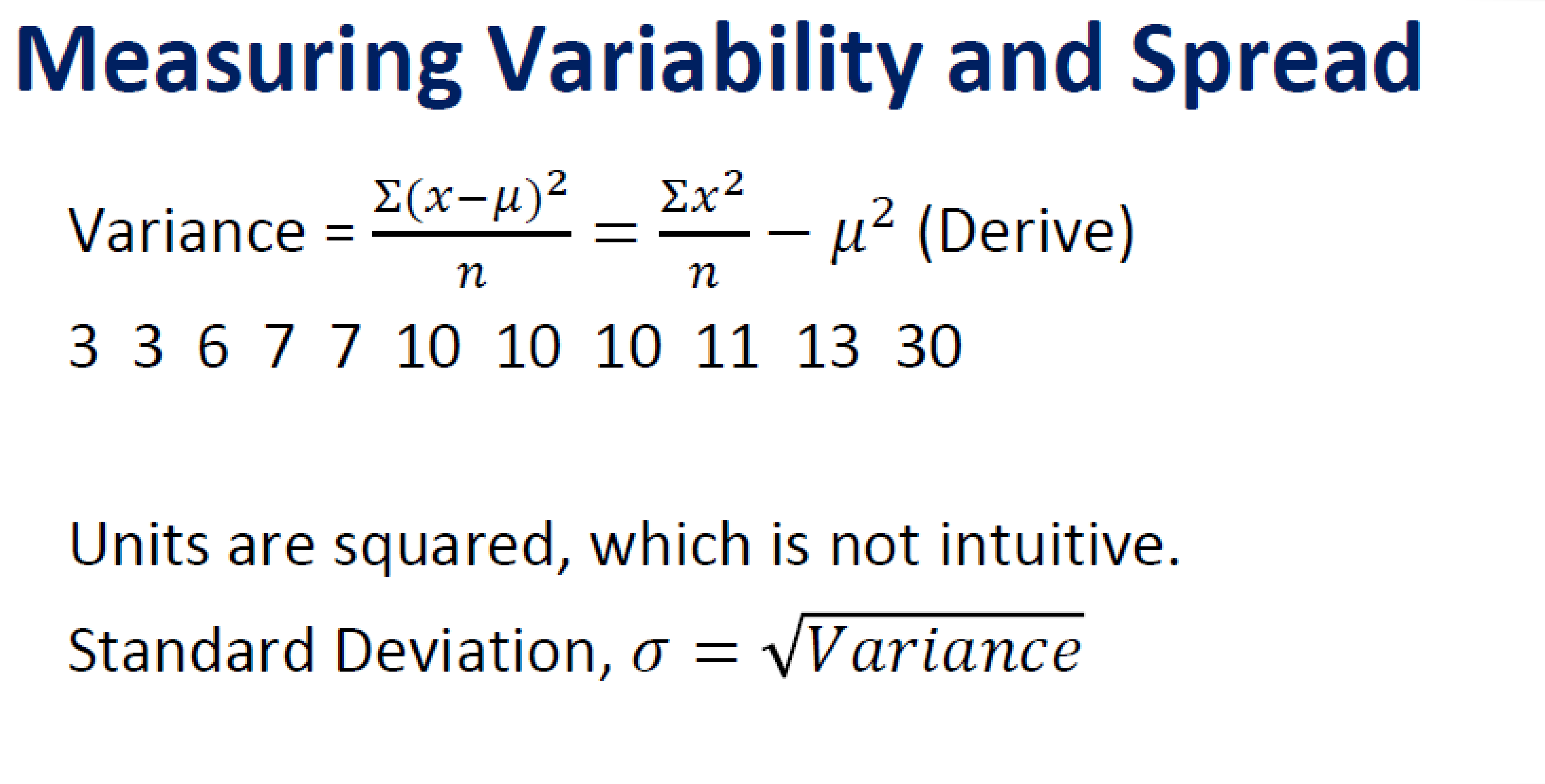
Points scored per game	7	8	9	10	11	12	13
Frequency, f	1	1	2	2	2	1	1

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 Reject Player 3

SD and Variance



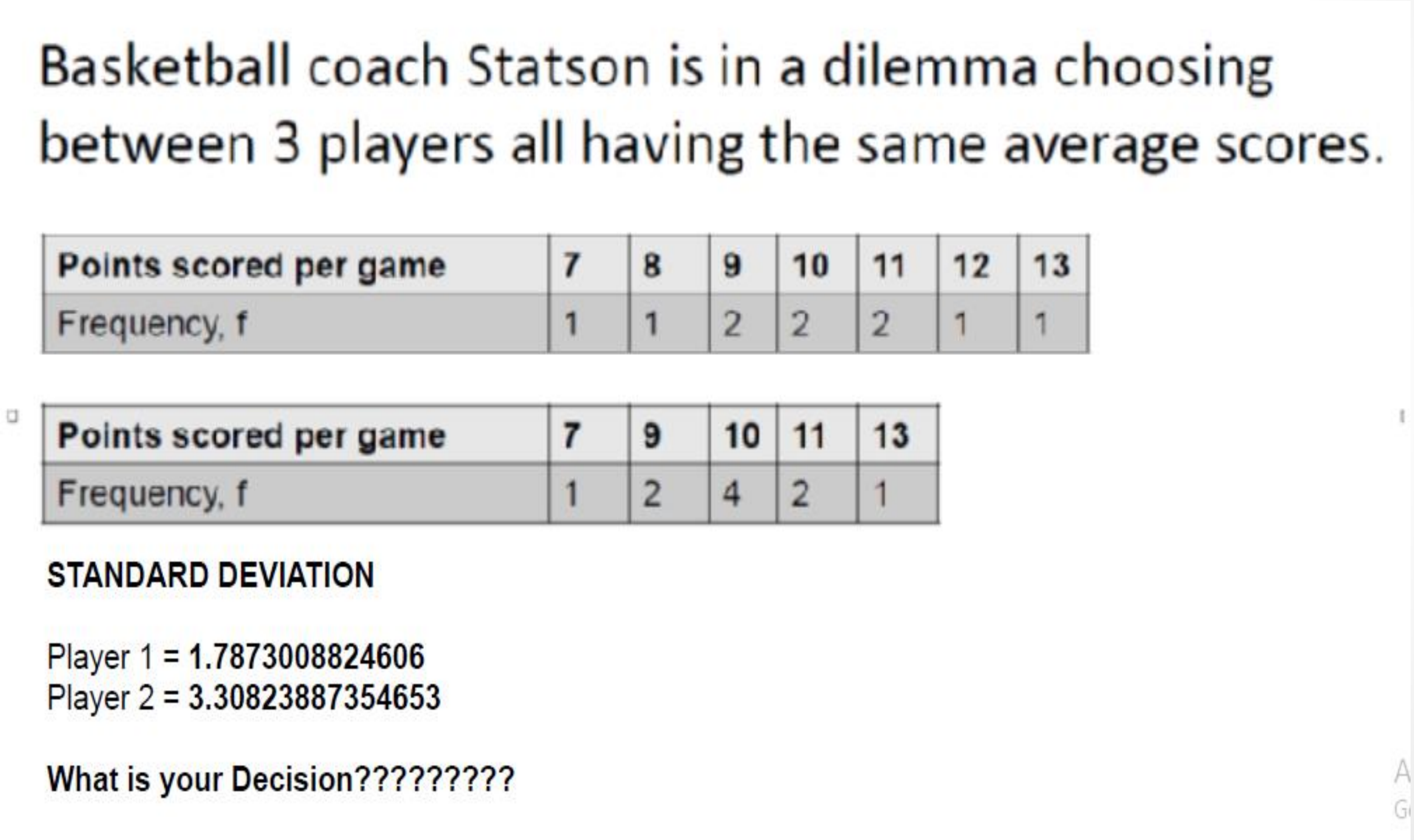
Measuring Variability and Spread

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

3 3 6 7 7 10 10 10 11 13 30

Units are squared, which is not intuitive.

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



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

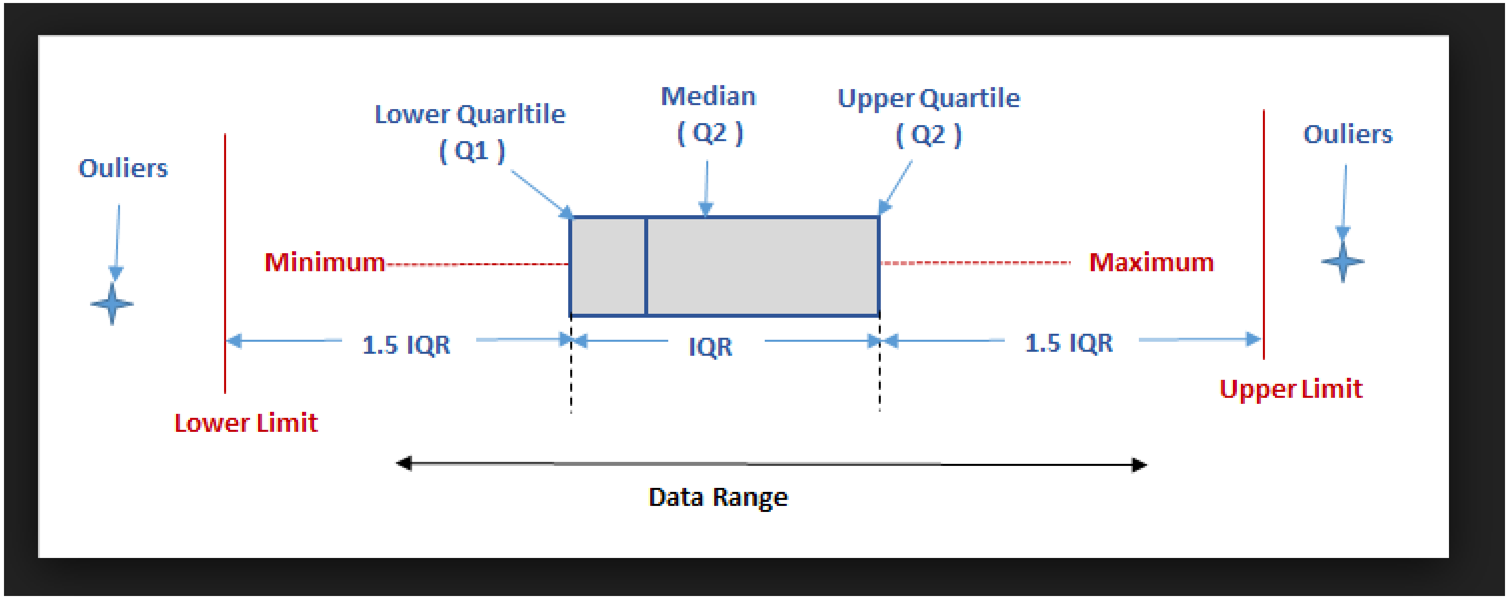
Player 1 = **1.7873008824606**
Player 2 = **3.30823887354653**

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

Data Visualization - Plots

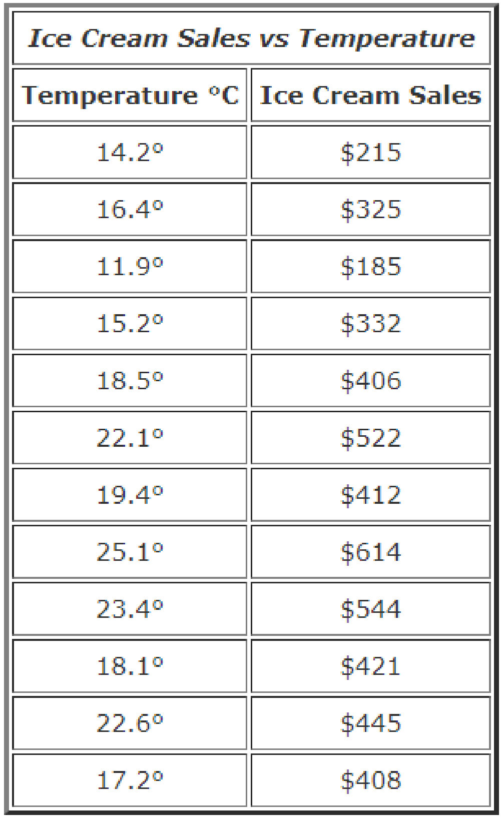
1. *Box Plot*
2. *Scatter plot*
3. *Density Plot*

Box Plot

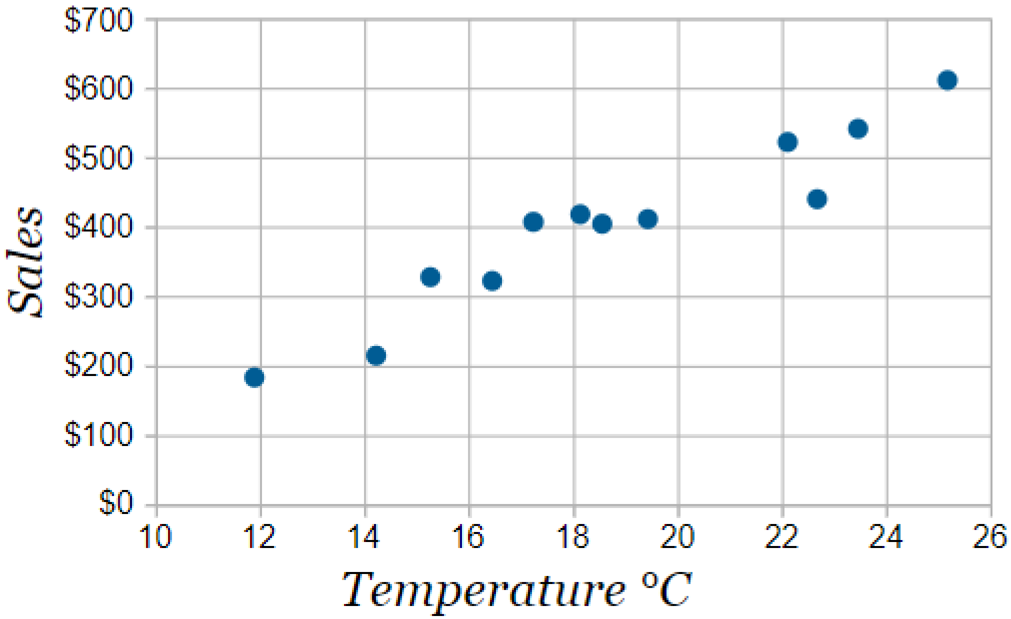


- Shows the data spread for individual columns

Scatter Plot

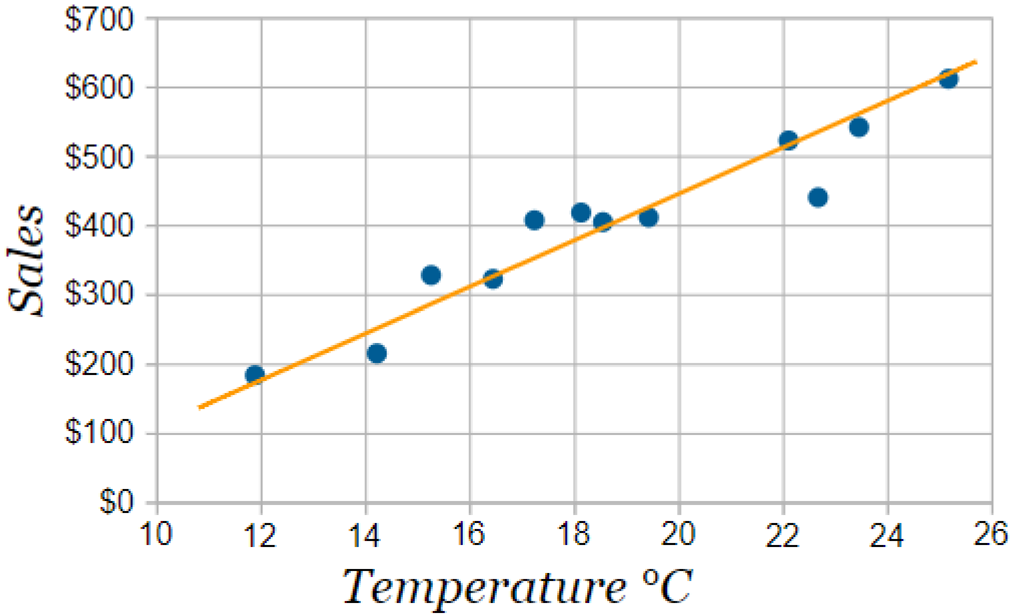


Ice Cream Sales vs Temperature	
Temperature °C	Ice Cream Sales
14.2°	$215
16.4°	$325
11.9°	$185
15.2°	$332
18.5°	$406
22.1°	$522
19.4°	$412
25.1°	$614
23.4°	$544
18.1°	$421
22.6°	$445
17.2°	$408

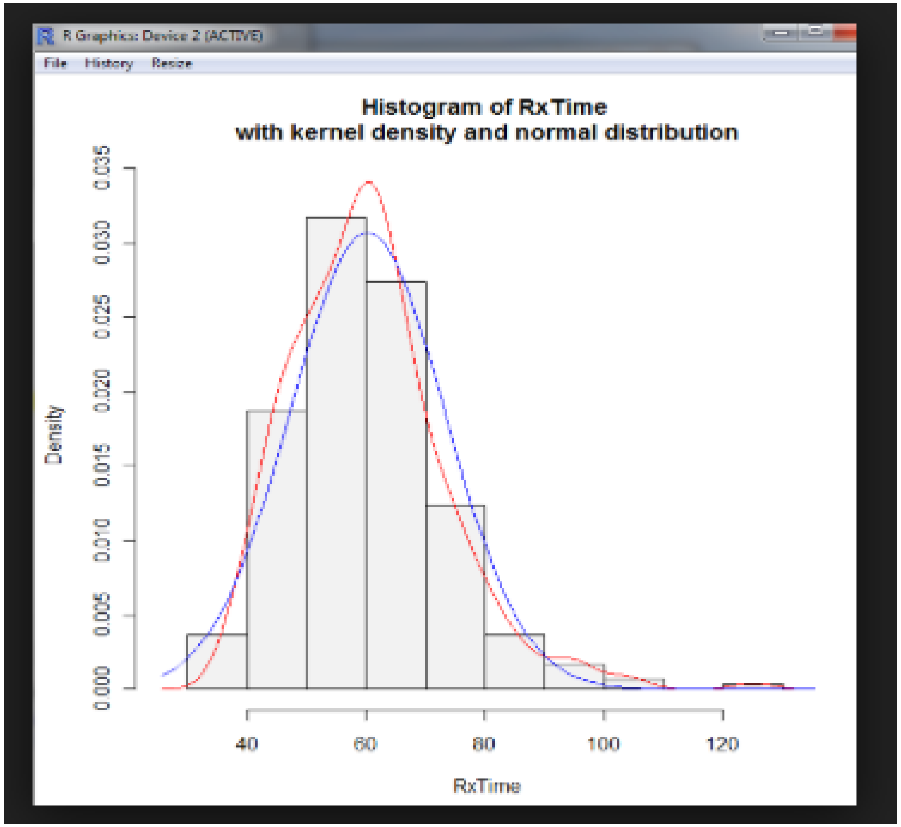


- **Shows relationship between 2 columns**

Line of Best Fit



Density Plot



- Shows the distribution of data

Statistical simulation link

http://www.shodor.org/interactivate/activities/

Percentile & Quartile

Nth percentile – States that there are atleast N% of values less than or equal to this value and (100-N) 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

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

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 %

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

(**Also called as Standard Error - SE**) Standard deviation of sample mean = **(population standard deviation/square root(n))**

Mean of mean sample distribution = Population mean

As n increases SE decreases – SE is inversely proportional to n

Measure of association 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

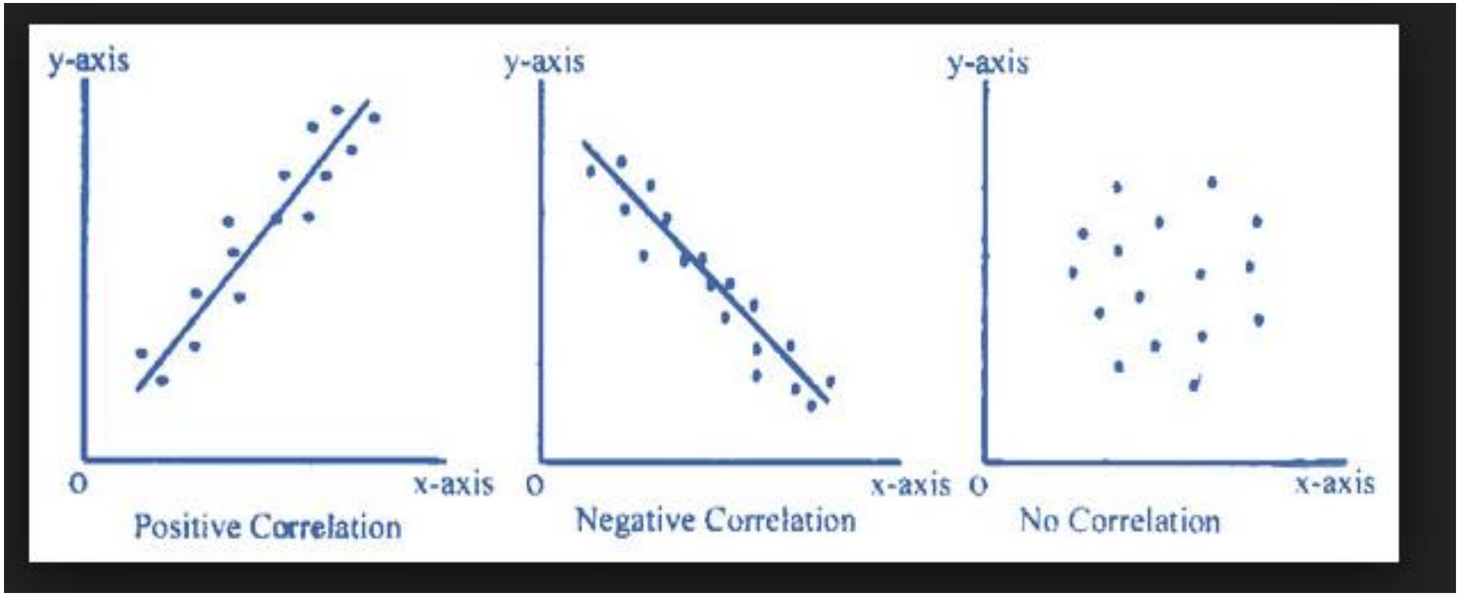
$$r_{xy} = \frac{\mathrm{Cov}(x, y)}{S_x \times S_y}$$

Key Points

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

2. Ranges from -1 to +1

Types of Correlation


y-axis
O
x-axis
Positive Correlation
y-axis
O
x-axis
Negative Correlation
y-axis
O
x-axis
No Correlation