MOHAMED IMRAN
DATA SCIENTIST
SHELL

Descriptive Statistics

What is Statistics?

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

Why do we learn statistics?

Scenario 1:

A college in US has students from the following countries for a Masters degree. 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

Scenario 2:

A parent decided to switch school of their son who is going for 11th standard since his academic results are not good in 10th standard in his current School.

And, they changed him from ABC school to XYZ school.

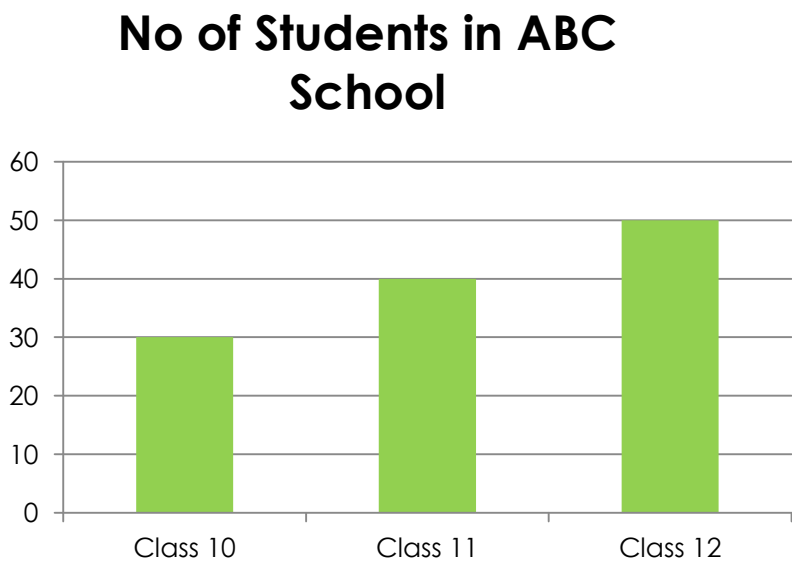
- Below is the rank as a result of school change.

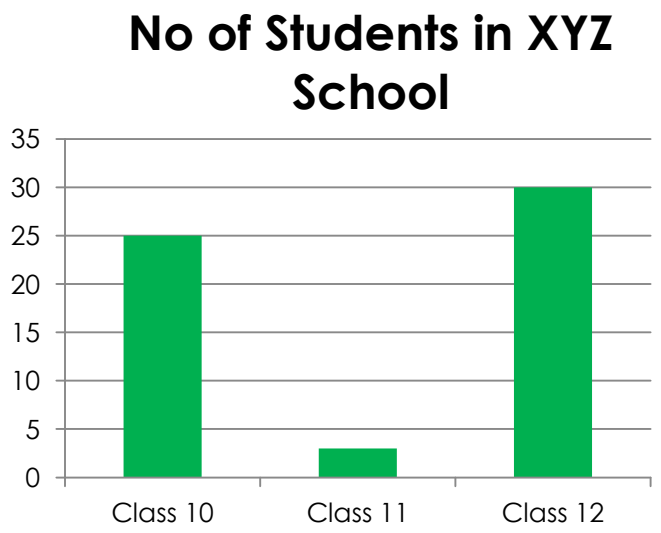
Results

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

What's the conclusion: Has the student improved ?

Alright, now look at this.





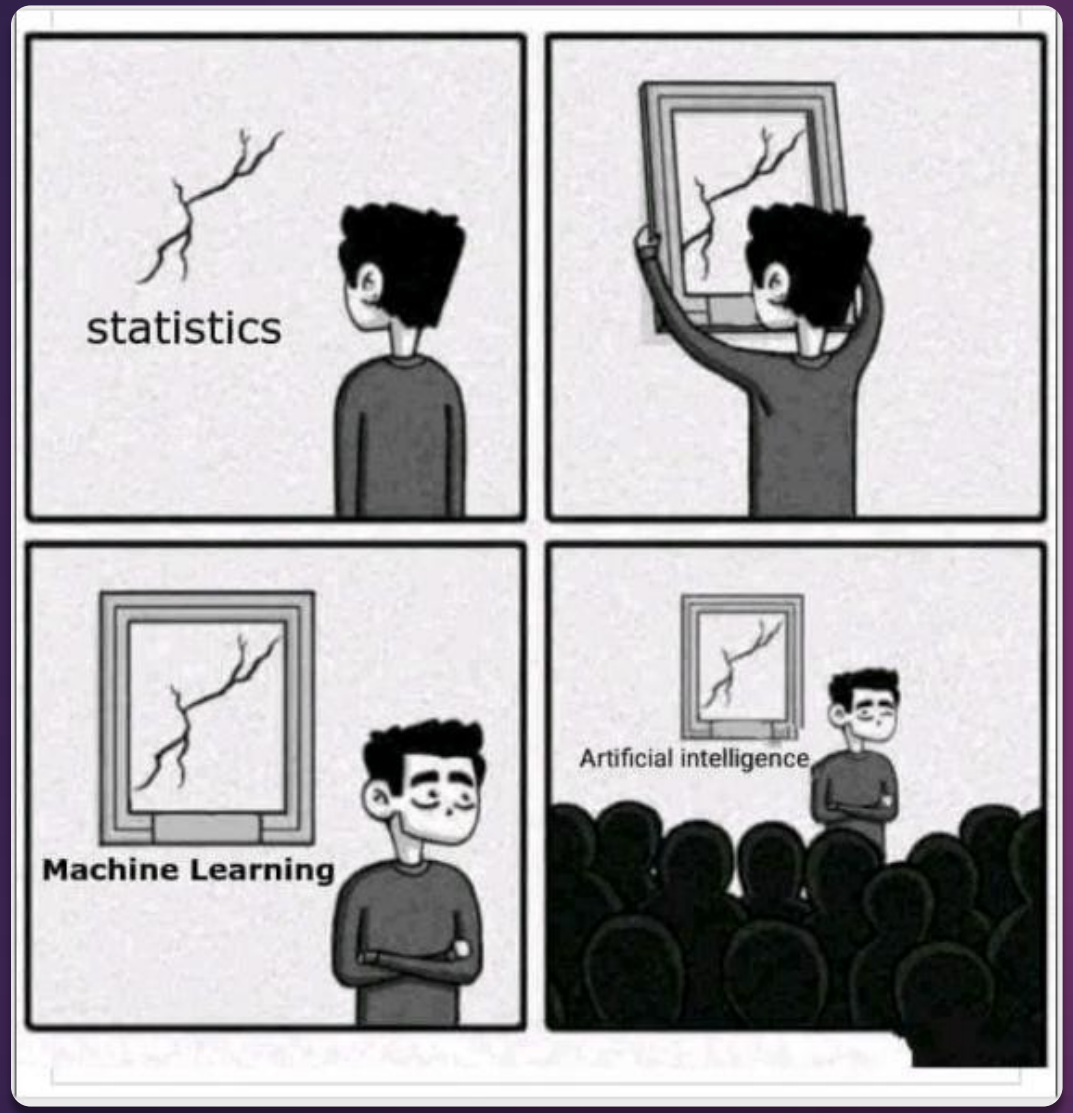
This is Statistics

COLLECTING DATA

ANALYSING DATA

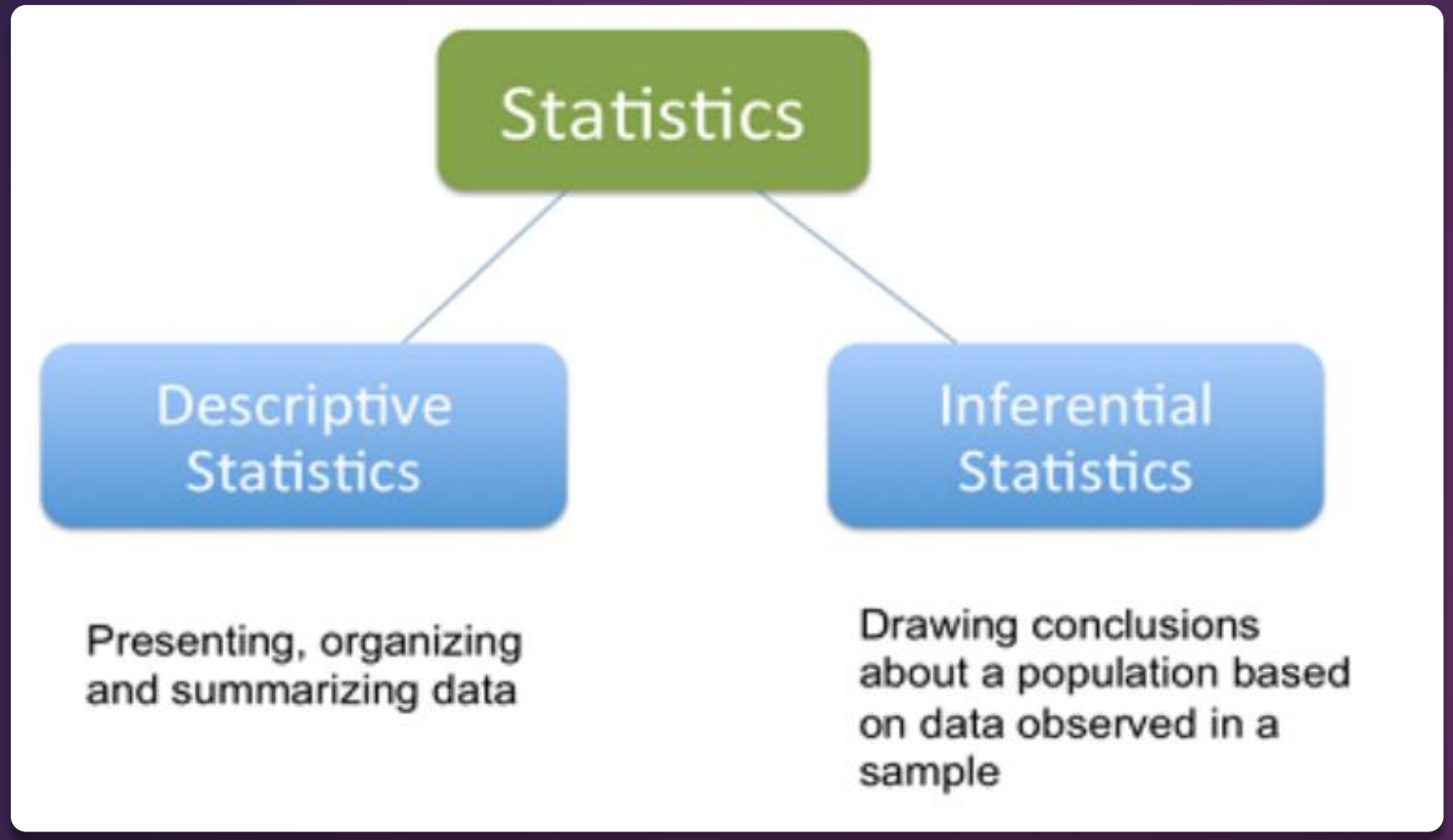
INTERPRETING DATA

PRESENTING DATA



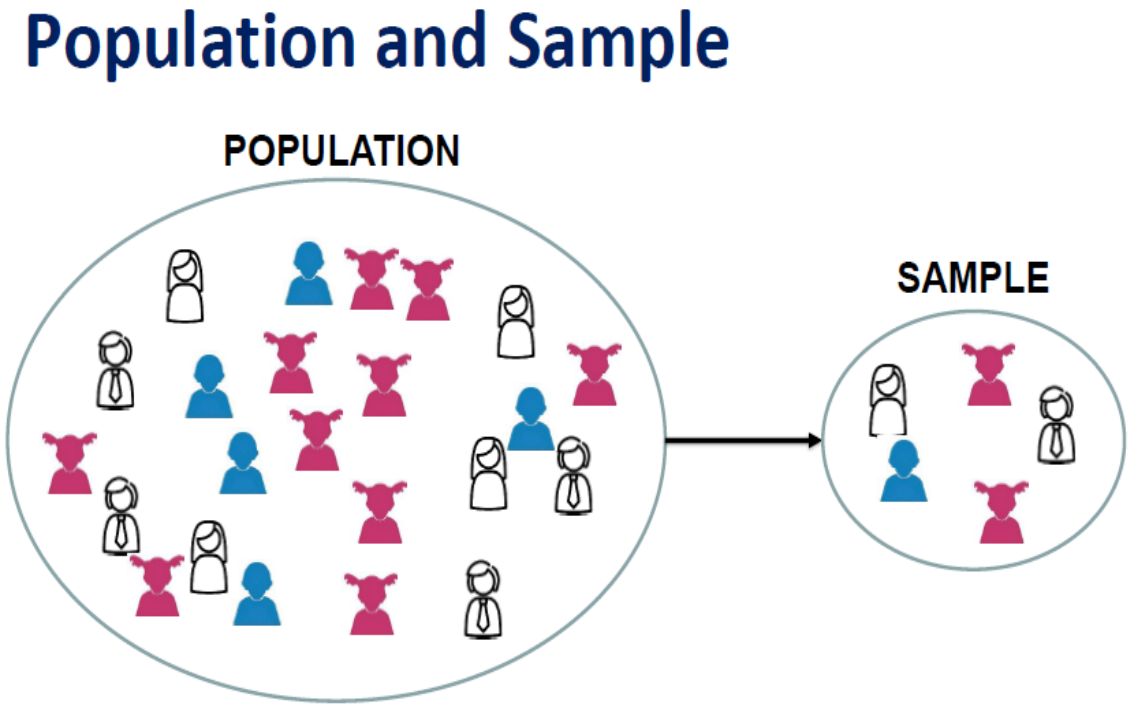
In a nutshell.

Classification of statistics



Population and Sample

- The entire group of individuals is called the **population**.
- Usually populations are so large that a researcher cannot examine the entire group.
- Therefore, a **sample** is selected to represent the population in a research study



The goal is to use the results obtained from the sample to help answer questions about the population.

POPULATION

PARAMETERS:
A descriptive measure of population

E.g. Mean, variance or standard deviation of population

STATISTIC:
A descriptive measure of sample

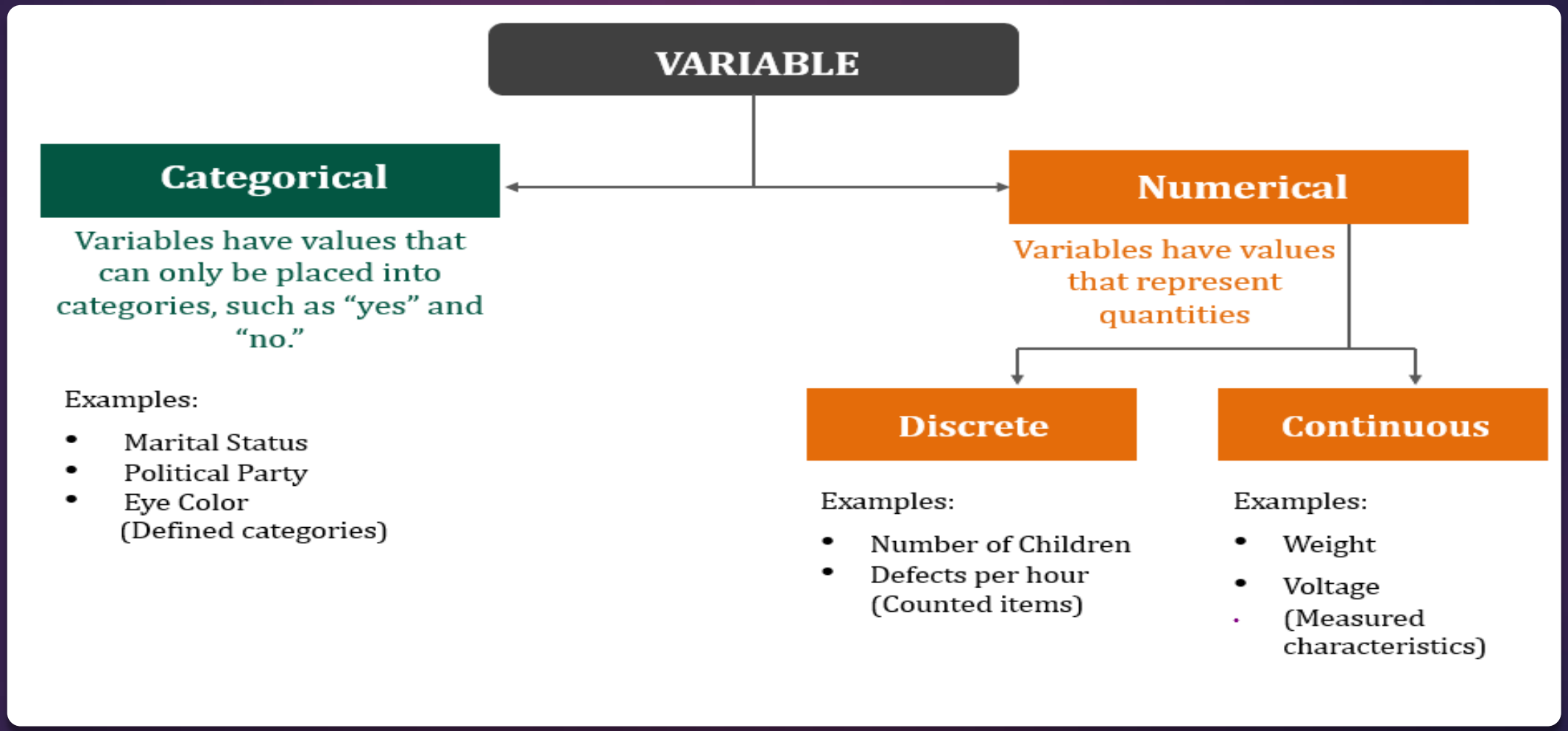
E.g. Mean, variance or standard deviation of a sample

SAMPLE

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

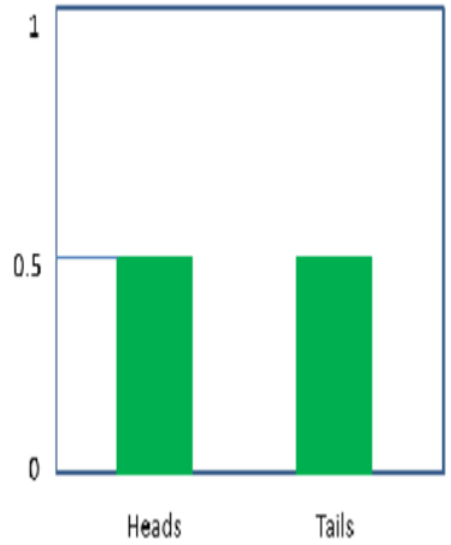
Classification of variable



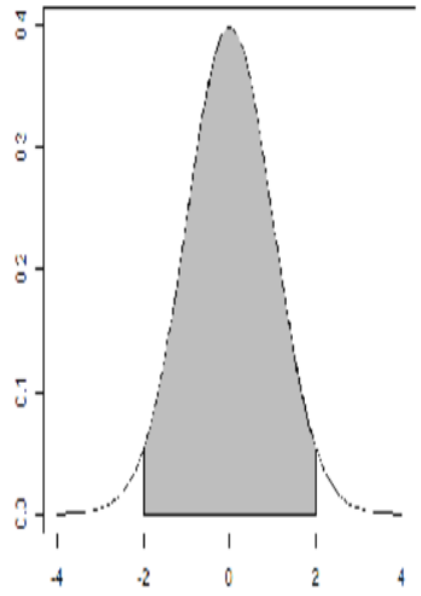
Discrete vs Continuous

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



Countable



Measurable

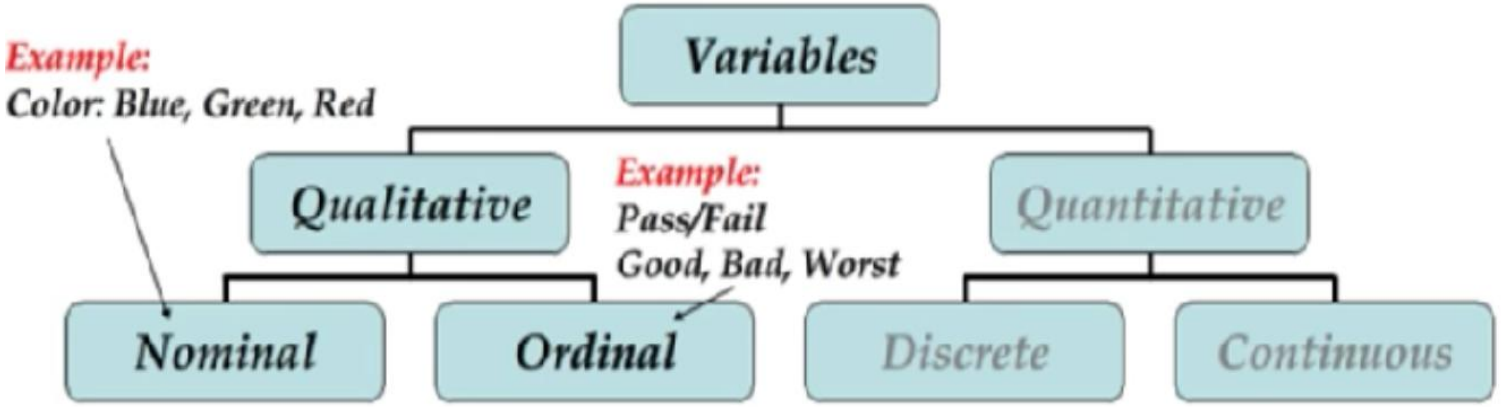
DEPENDENT AND INDEPENDENT VARIABLES

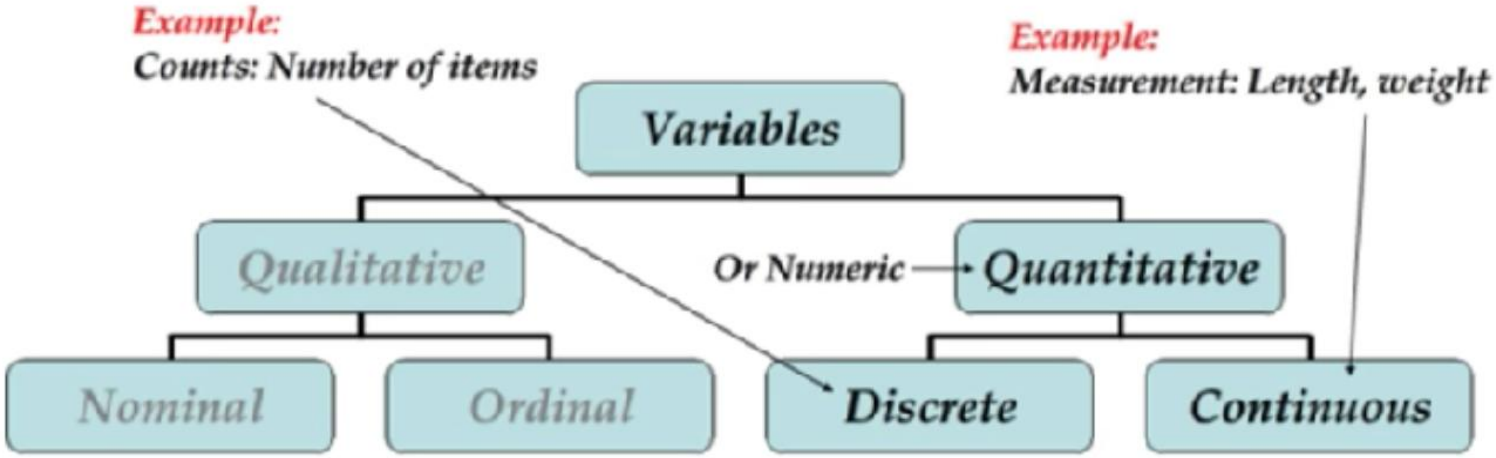
An independent variable, sometimes called an experimental or predictor variable, is a variable that is being manipulated in an experiment in order to observe the effect on a dependent variable.

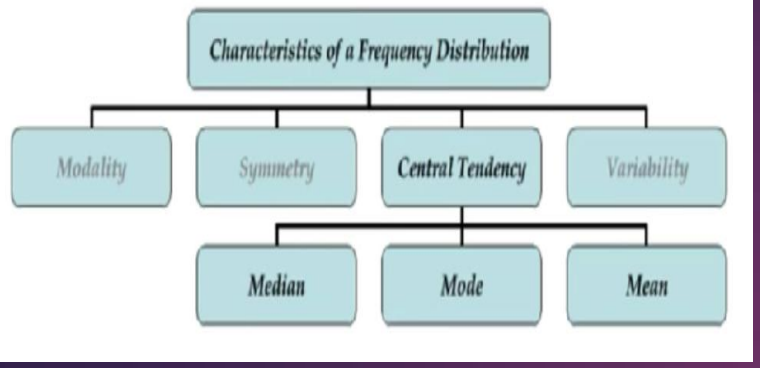
Dependent variable – y axis
Independent variable – x axis

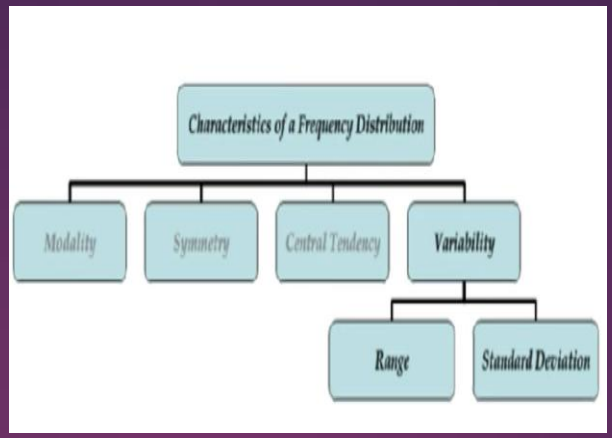
Dependent variable is also called as Target or Outcome variable

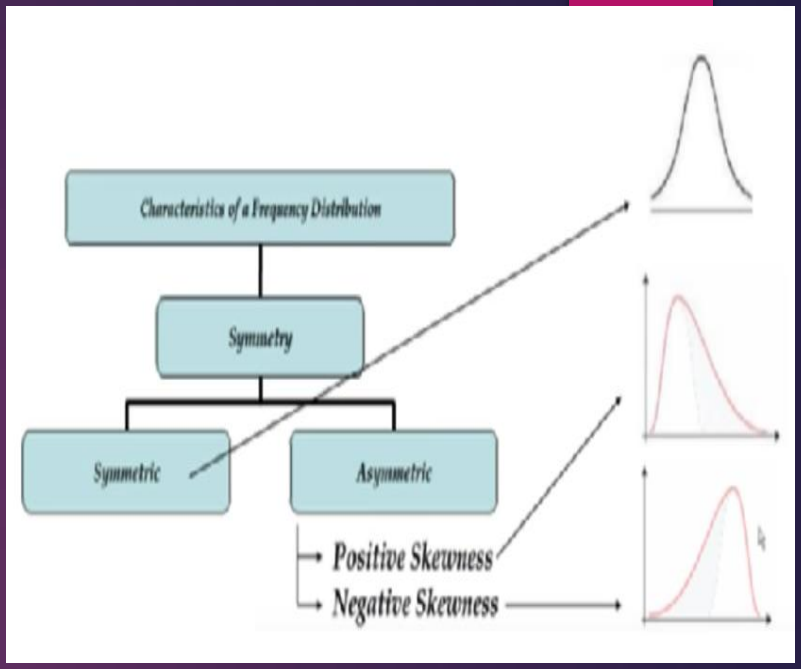
Data types of variable



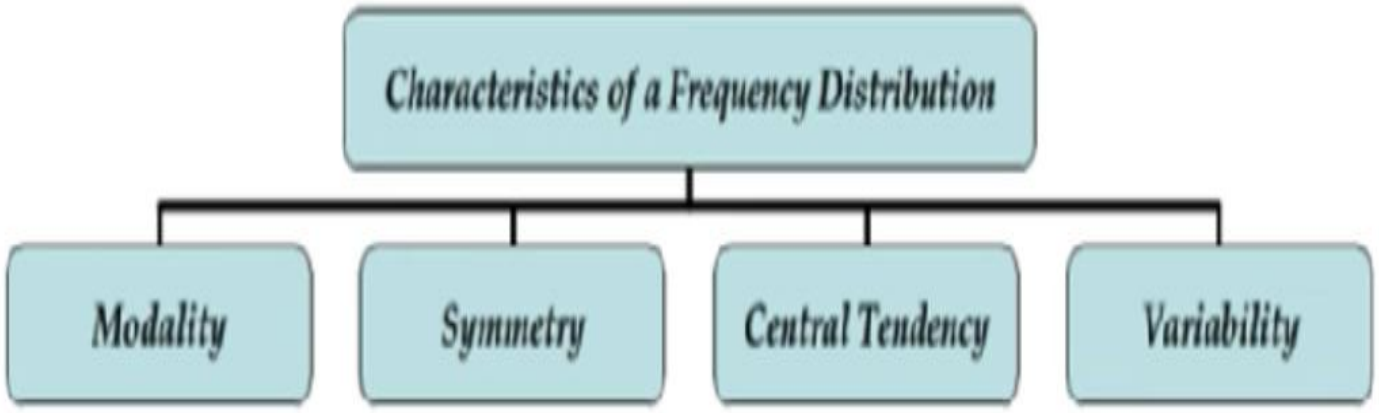


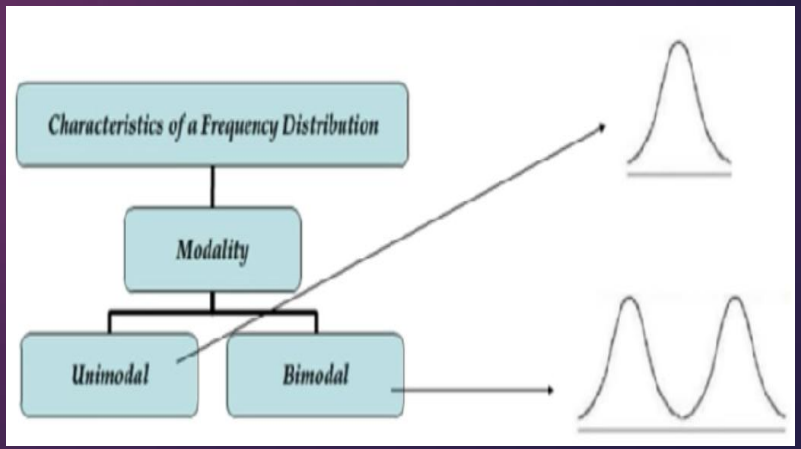






Summarizing Data:





Measures of Central Tendency

The goal of measures of central tendency is to come up with the one single number that best describes a distribution of scores.

There are three basic measures of central tendency, and choosing one over another depends on two different things.

Mean

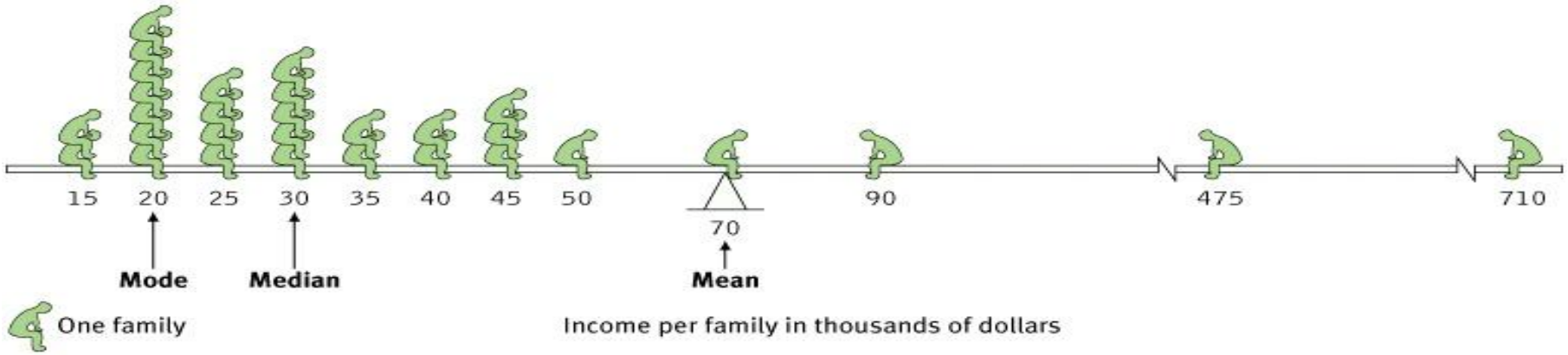
Median

Mode

Mean

- The arithmetic average of some data is average score or value and computed simply by adding together all scores and dividing by the number of scores. It uses information from every single score.

$$\text{For a population: } \mu = \frac{\Sigma X}{N} \qquad \text{For a Sample: } \bar{X} = \frac{\Sigma X}{n}$$



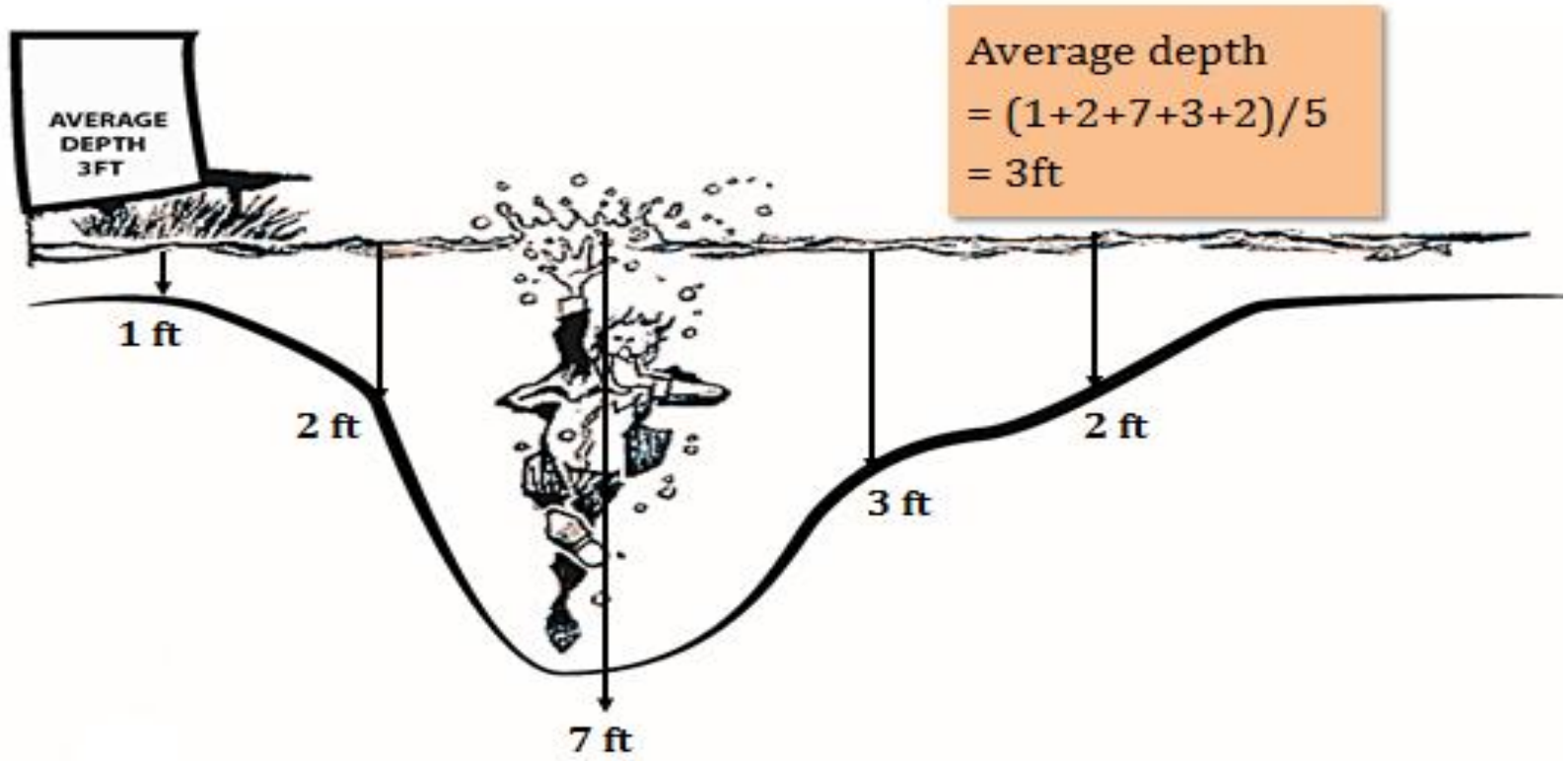
Alan went for a trek. On the way, he had to cross a stream. As Alan did not know swimming, he started exploring alternate routes to cross over.

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.

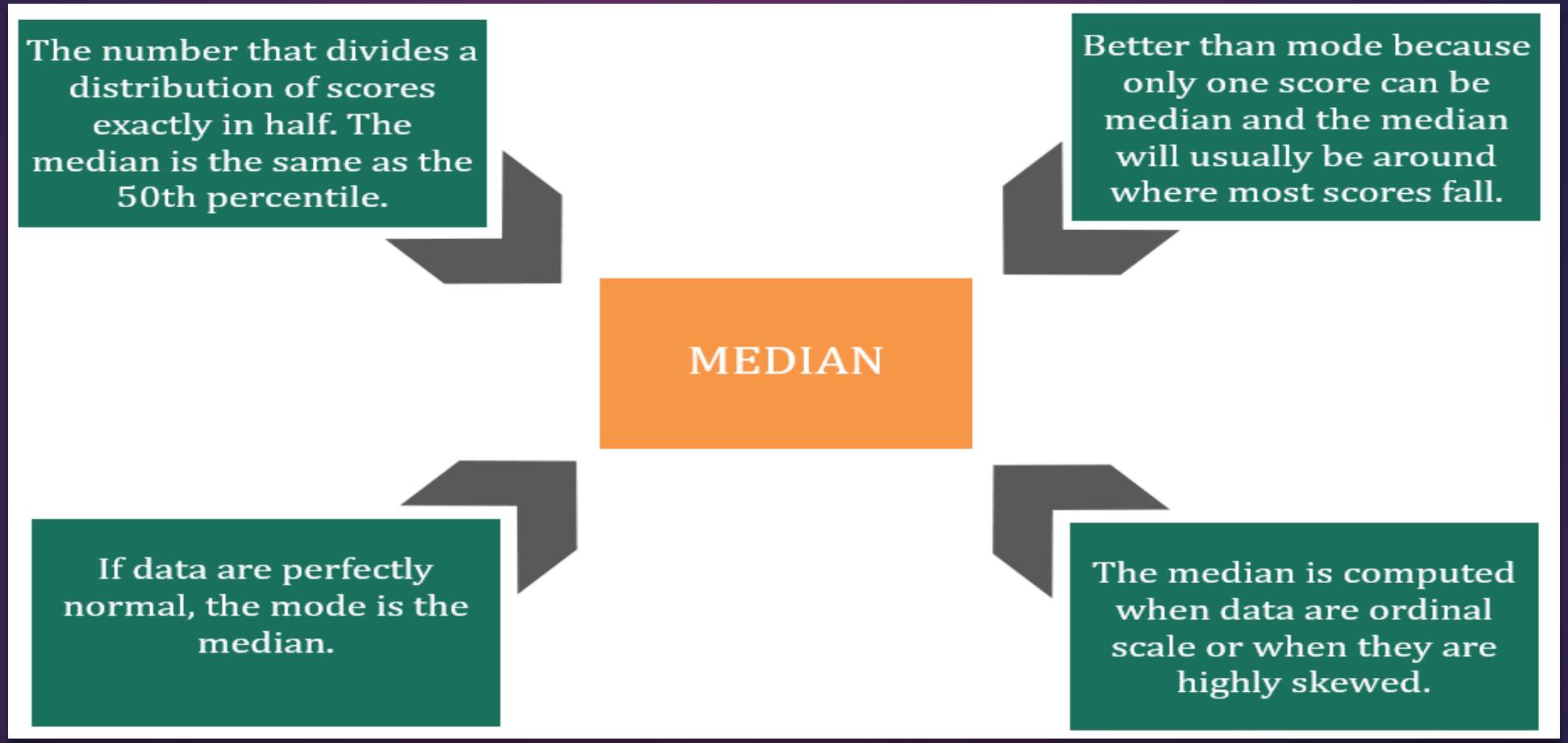
Alan never reached the other end and drowned in the stream.

Why did Alan drown?

The reason is… Average!



Median: Arrange the data in an ascending order and find the mid point using (n+1)/2



Calculating the median

If you have an odd number of scores, pick the middle score.

- 1 4 6 7 12 14 18
- Median is 7

If you have an even number of scores, take the average of the middle two.

- 1 4 6 7 8 12 14 16
- Median is (7+8)/2 = 7.5

Median is less susceptible to the influence of outliers

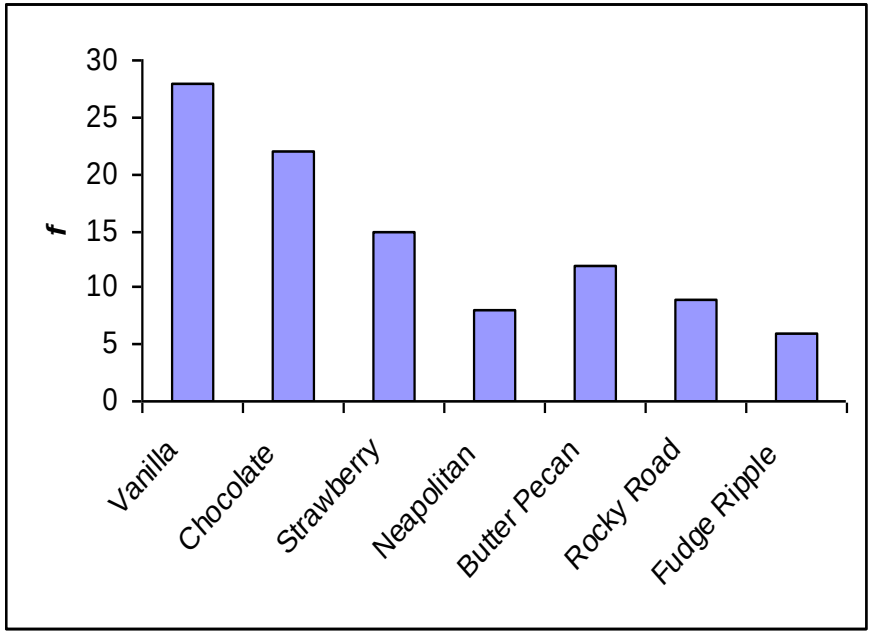
- Average deal size in pipeline = $100,000
- Deal #10 is of significantly higher value than all the other deals and impacts the average calculation
- Median = $55,000 more realistic measure

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

Mode:

- The most common observation in a group of scores is that distributions can be unimodal, bimodal, or multimodal.
- Only if the data is categorical (measured on the nominal scale) the mode can be calculated.
- The most frequently occurring score (mode) is Vanilla.

Flavor	f
Vanilla	28
Chocolate	22
Strawberry	15
Neapolitan	8
Butter Pecan	12
Rocky Road	9
Fudge Ripple	6



Central Tendency

- 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

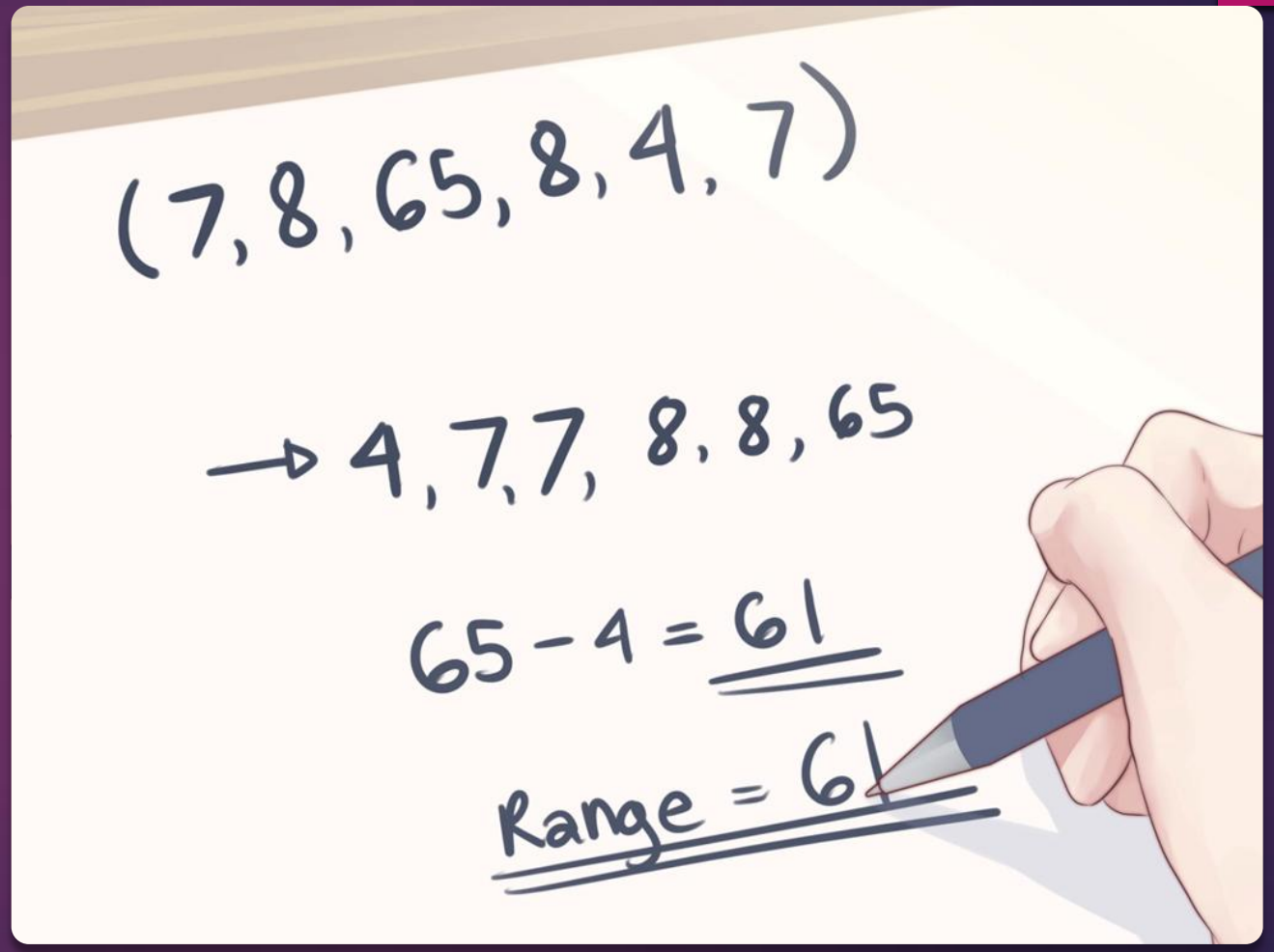
Mode
= Value with highest frequency
= 43.4

Measure of dispersion: Describes the data spread or how far the measurements are from the centre

1. Range
2. Variance/Standard Deviation
3. Interquartile range (IQR)

Range:

Max-Min



Standard deviation:

This is the most useful and most commonly used of the measures of variability.
The standard deviation looks to find the average distance that scores are away from the mean.

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

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

$$\sqrt{\frac{\Sigma(X-\overline{X})^2}{(n-1)}}$$

=square root
Σ=sum (sigma)
X=score for each point in data

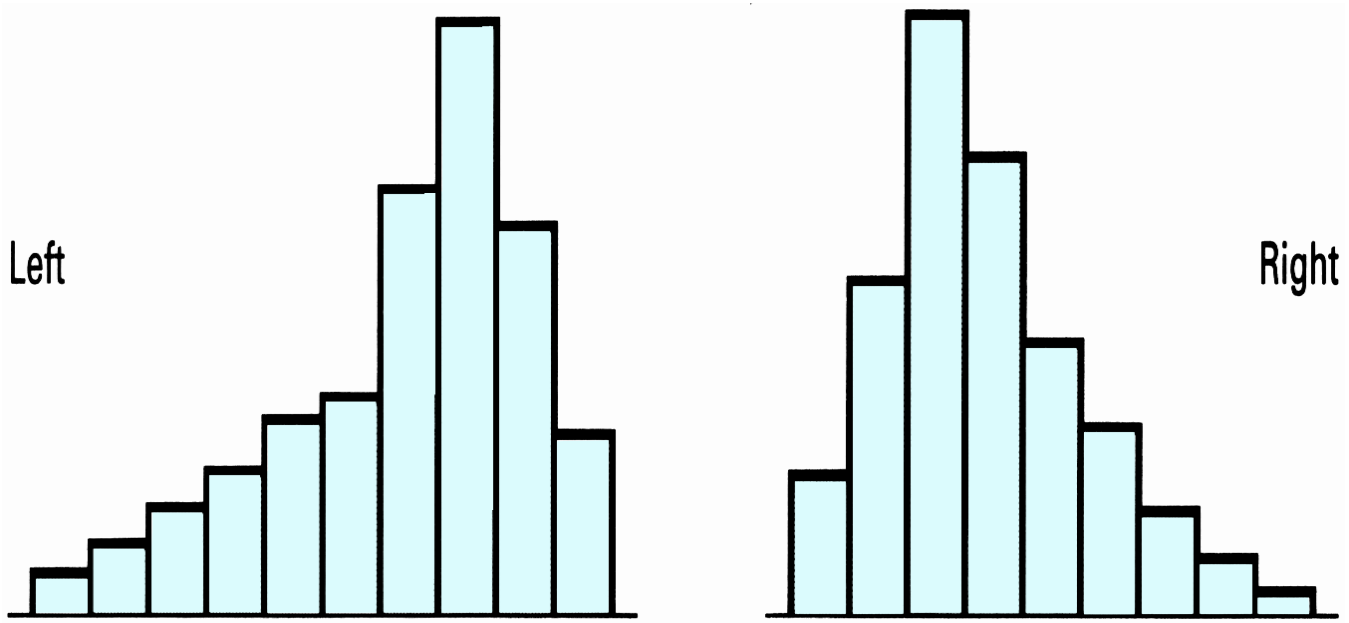
$\overline{X}$=mean of scores for the variable
n=sample size (number of observations or cases)

Other measures:

- Skewness
- Coefficient of Variation
- Box plot
- Scattered plot

Skewness

- Skewness is the lack of symmetry of the data
- Left – Negative Skew
- Right – Positive Skew



Coefficient of Variation

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

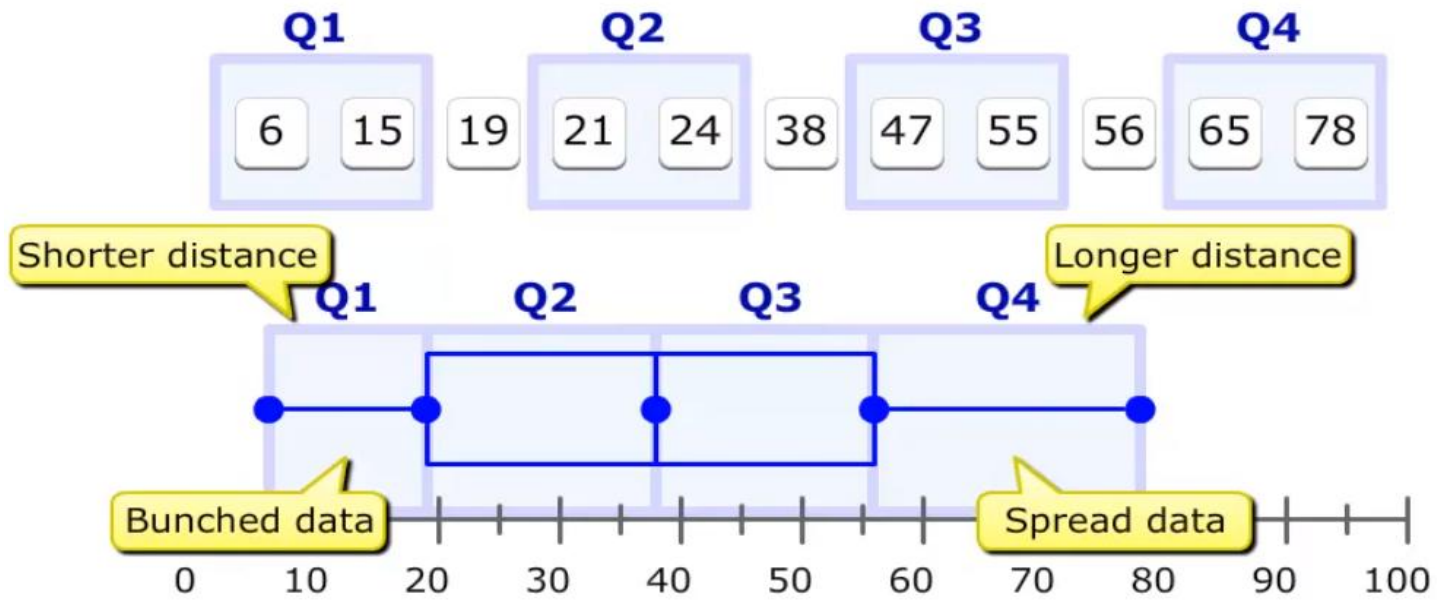
- Measure of Relative Variation
- Always a %
- Shows Variation Relative to Mean
- Used to Compare 2 or More Groups

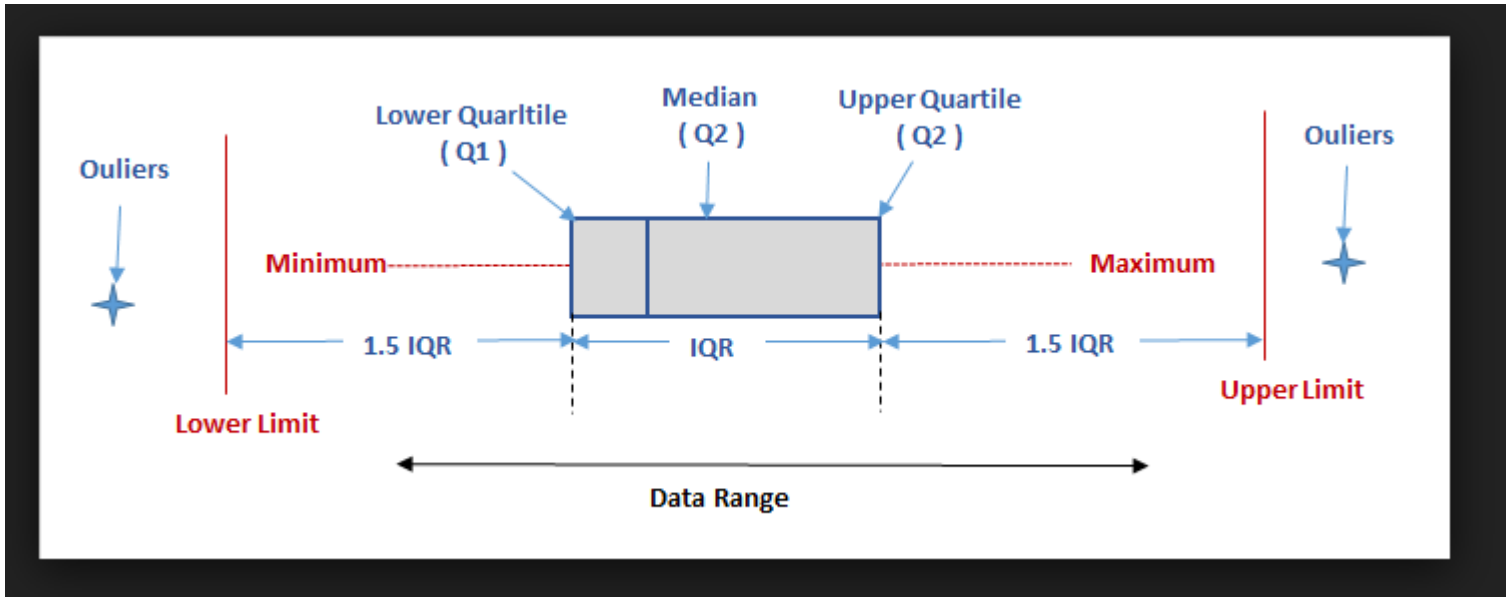
Case Study

- 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 by dropping a player whose CV is greater than 85%?

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

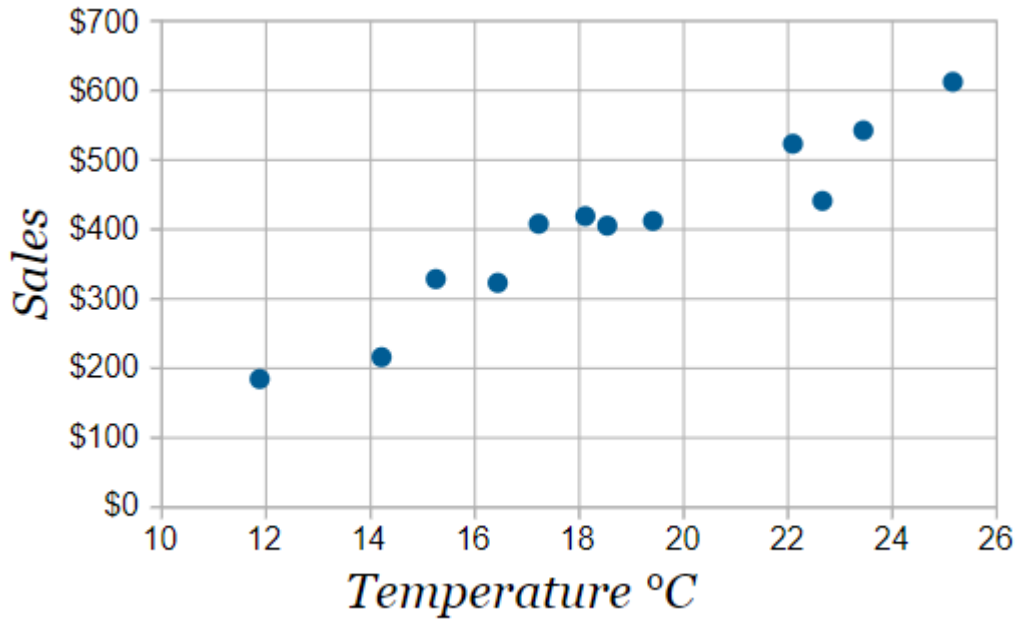
Box Plot


Q1
Q2
Q3
Q4
6
15
19
21
24
38
47
55
56
65
78
Shorter distance
Longer distance
Q1
Q2
Q3
Q4
Bunched data
Spread data
0
10
20
30
40
50
60
70
80
90
100


Ouliers
Lower Quarltile
(Q1)
Median
(Q2)
Upper Quartile
(Q2)
Ouliers
Minimum
Maximum
1.5 IQR
IQR
1.5 IQR
Upper Limit
Lower Limit
Data Range

Scatter plot

- Shows relationship between 2 columns



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

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

How do I go about choosing a best player?

Look for their total score?

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

sum doesn't help ☹

Let’s try mean

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

Mean doesn't help too☹

Let's try median

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

Median doesn't help either☹

Let’s try Range

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

Range gives an idea about how far the scores are spread

Standard deviation

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

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

S.No	x	x^2	x-mean	Abs(x-mean)	(x-mean)^2
1	5	25	0.3333333333	0.3333333333	0.1111111111
2	7	49	2.333333333	2.333333333	5.444444444
3	4	16	-0.6666666667	0.6666666667	0.4444444444
4	2	4	-2.666666667	2.666666667	7.111111111
5	6	36	1.333333333	1.333333333	1.777777778
6	2	4	-2.666666667	2.666666667	7.111111111
7	8	64	3.333333333	3.333333333	11.11111111
8	5	25	0.3333333333	0.3333333333	0.1111111111
9	3	9	-1.666666667	1.666666667	2.777777778
SUM	42	232	0	15.33333333	36
Average	4.666666667	25.77777778			

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

Well, Standard deviation helps ☺

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 = ?

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

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 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?????????