Statistics
&
Probablities

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

Presentation Understand,

What is Statistics ?

Analysing
Interpretation
Collection
Organisation.

WHAT IS STATISTICS?

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

Private and Confidential

4

What is Statistics

Statistics is a way to get information from data.

Why Learn Statistics?

Make it Structured
To draw insights
DESCRIBE

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 ?

China — 1+2
US → 2+1f 1+1
India → 1+1
Japan
England
Mexico

US 1	China	US 2	Sweden	China
Canada	China	Japan	Mexico	3 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

→ Organised
Summarized
Easy to understand

Case 2

Problem

A parent changes school of their Son who is studying in 11th standard since his academic results are not good in 10th 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 11th standard since his academic results are not good in 10th Standard in his 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 ?

In XYZ
1. Coaching is good
2. Results are goods
3. Students not comp.
4. organized
5. Pass

Case 2

Problem

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

They change Student A from ABC school to XYZ school
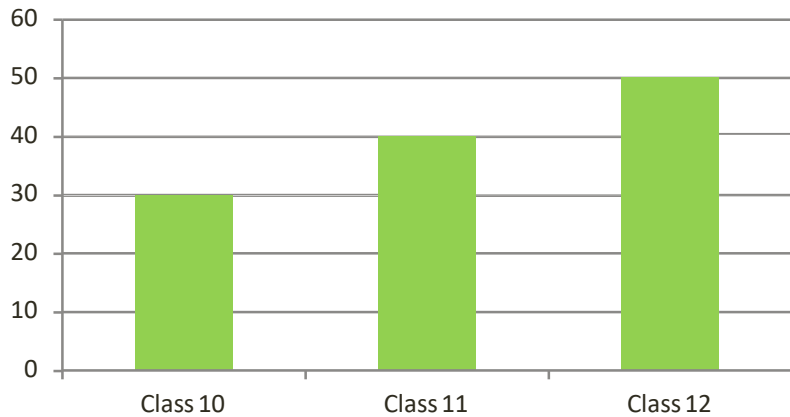
Results

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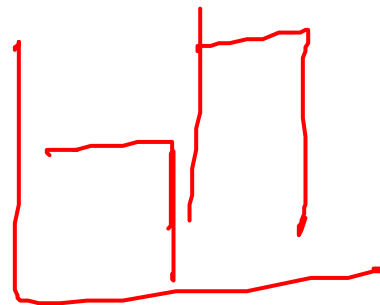2. Ranked 2nd in XYZ school

What's the conclusion: Has the student improved ?

Number of Students

No of Students in ABC School

No of Students in XYZ School

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

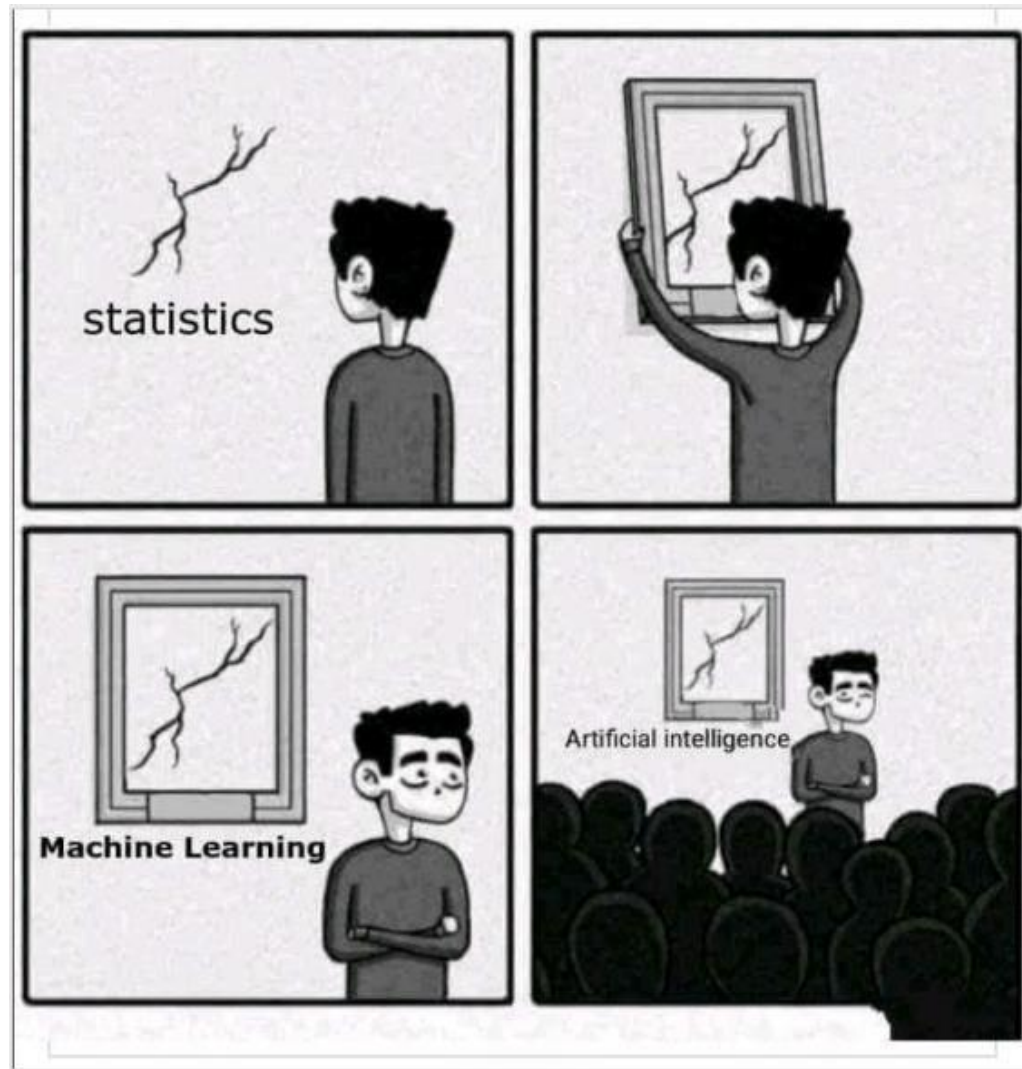Make reliable forecasts about a business activity

Improve business processes

Statistics is ...

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

What does it Tell?

Classification

Statistics

Descriptive Statistics

Inferential Statistics

Presenting, organizing and summarizing data

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

Variables

Non-Numeric

Discrete — Catigorical

Variables

Discrete or Continuous

Numerical

Qualitative — Or Numeric → Quantitative → Number

Nominal   Ordinal   Discrete   Continuous

Order   1, 2, 3, 4 5   1, 1·1, 1·2, 1·3, 1·4
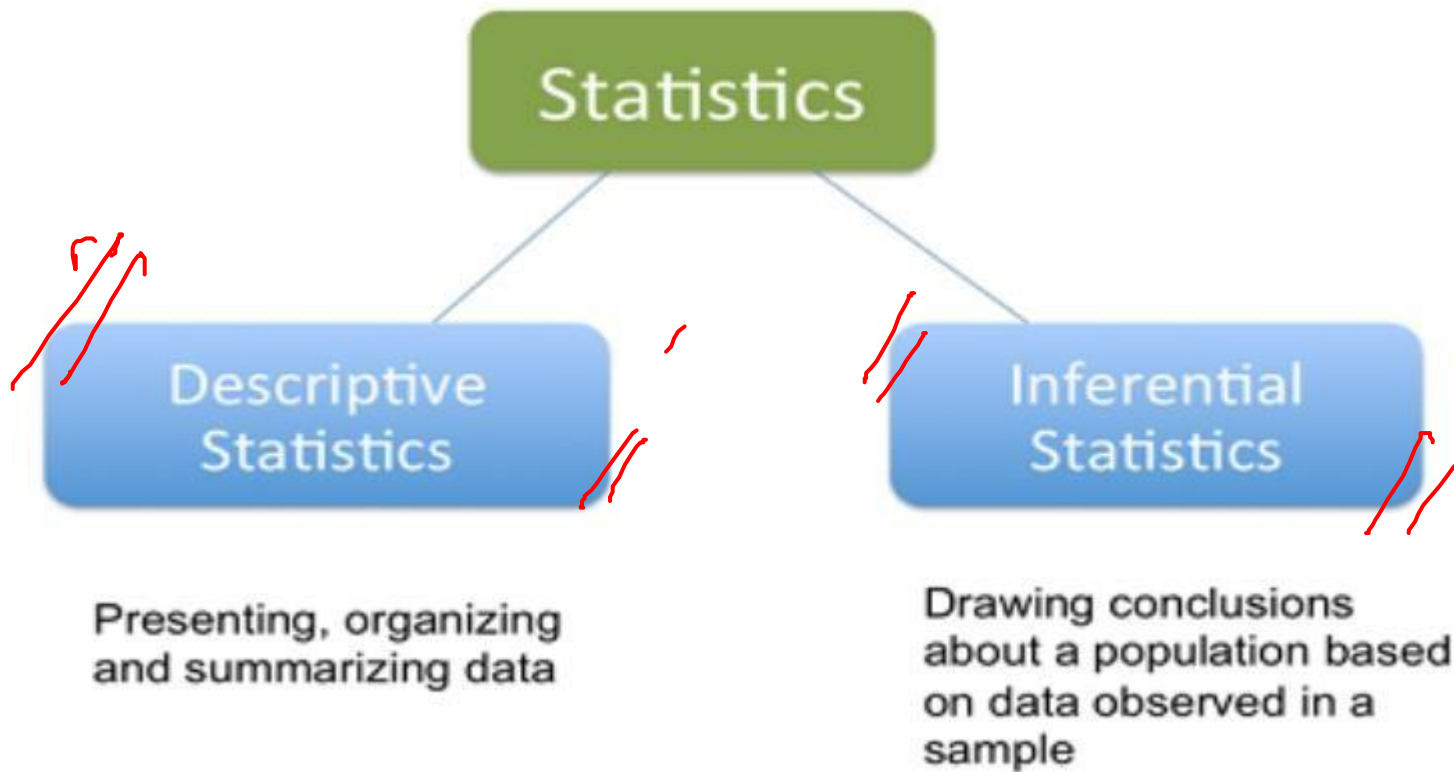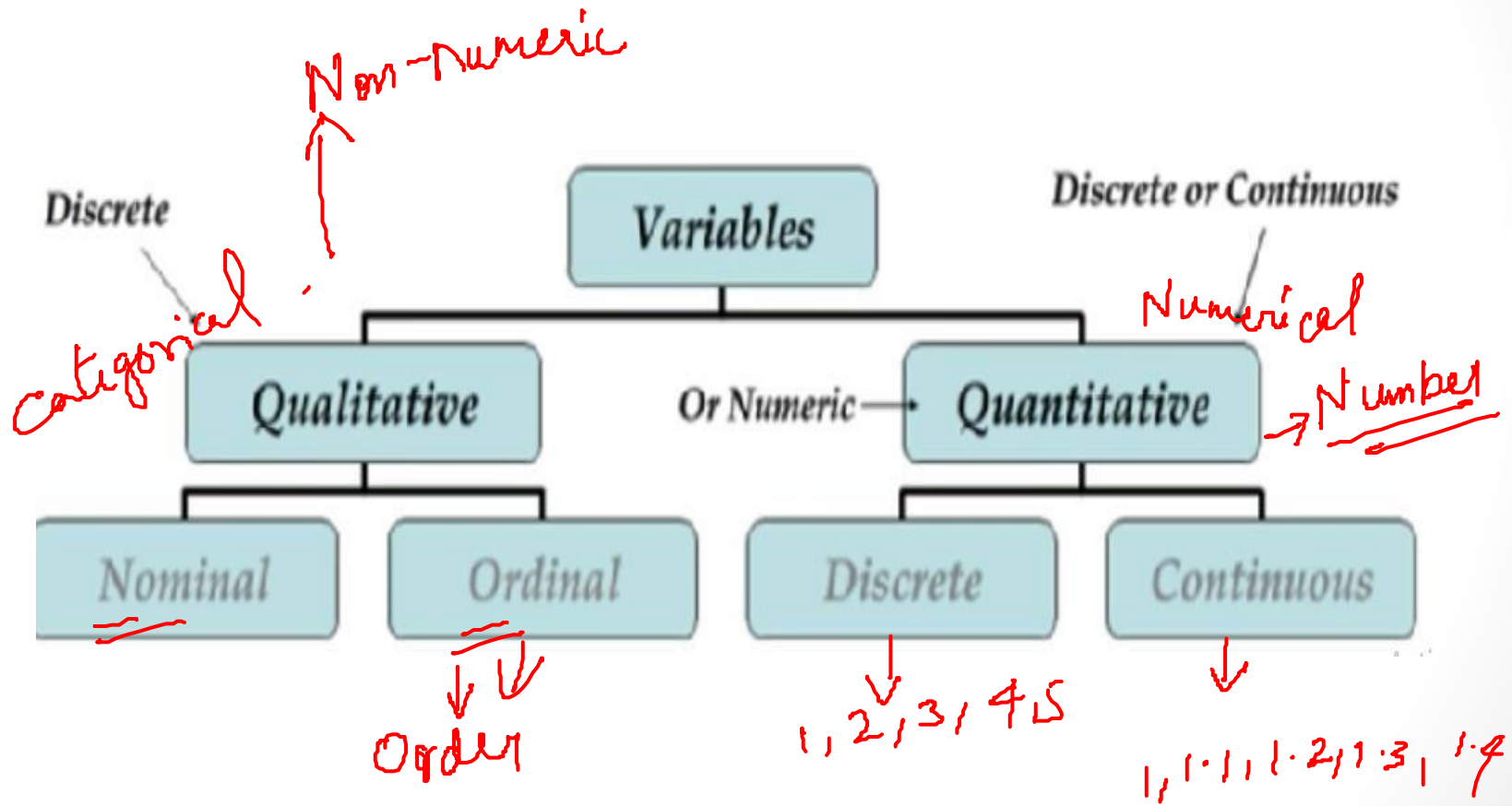
Categorical Data (Qualitative)

Nominal Examples

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

does not have any order

Ordinal Examples

has order

- 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
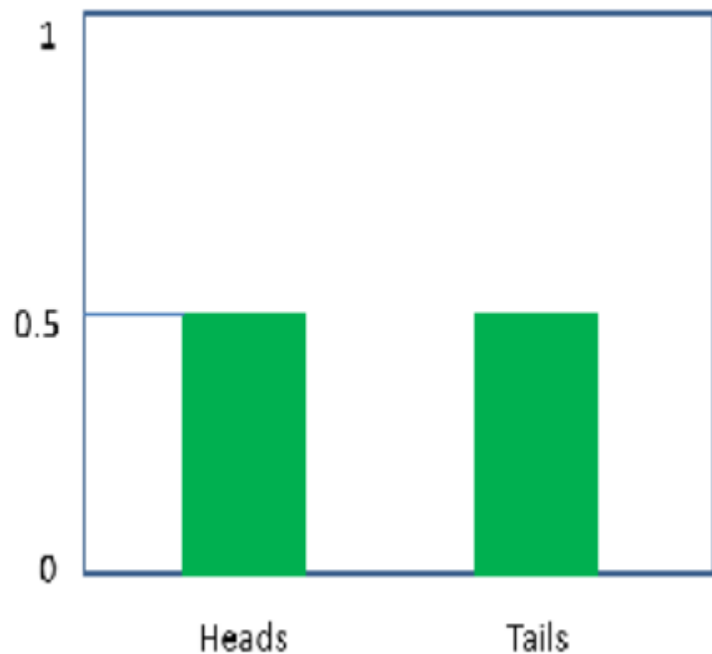
Countable
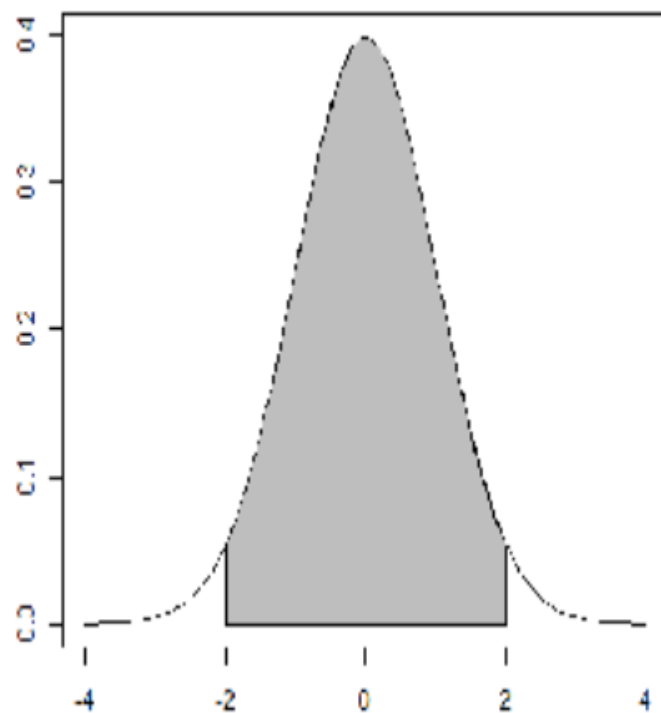
Measurable

Discrete or Continuous?

- Time between customer arrivals at a retail outlet

Discrete or Continuous?

- Time between customer arrivals at a retail outlet
  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 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

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 -

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

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 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

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 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
  GPA
  Units

- Categorical

  Gender
  Major
  Housing

Numerical or Categorical?

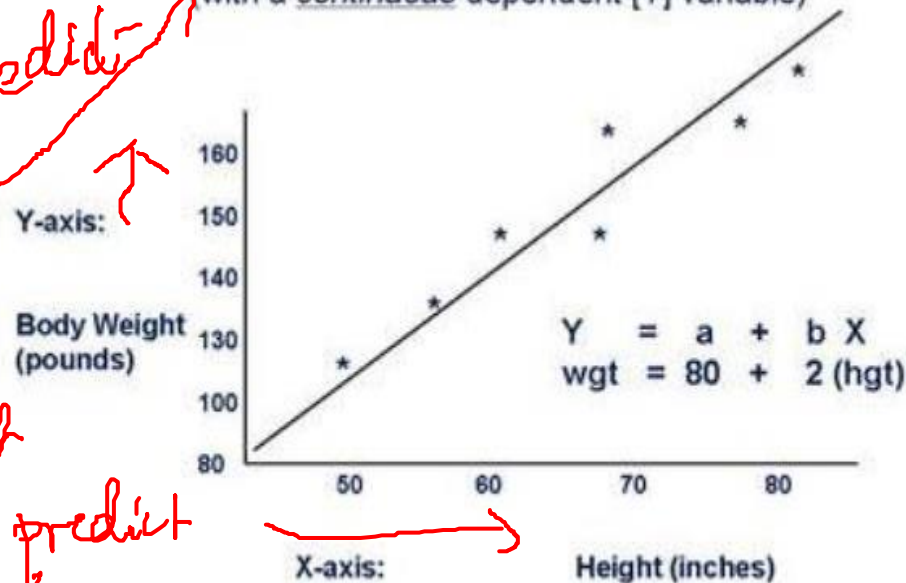| Age	Gender	Major	Units	Housing	GPA
18	Male	Psychology	16	Dorm	3.6
21	Male	Nursing	15	Parents	3.1
20	Female	Business	16	Apartment	2.8

- 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.

Simple Linear Regression
(with a *continuous* dependent [Y] variable)

Y-axis:

Body Weight (pounds)

$$Y = a + b X$$
$$wgt = 80 + 2 (hgt)$$

160
150
140
130
100
80

50   60   70   80
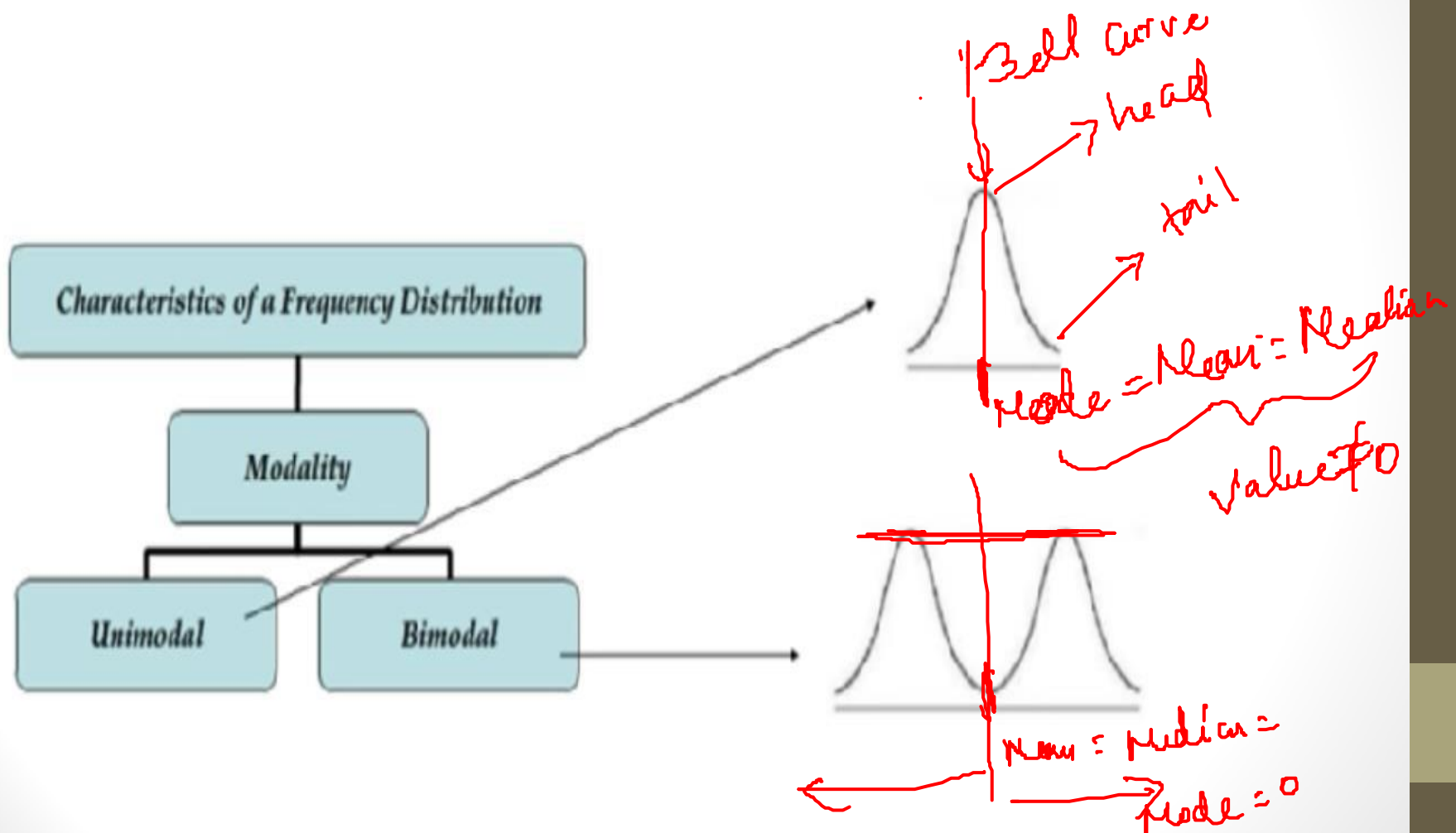
X-axis: Height (inches)

Variable to predict → Y

↓
col

X' → cols that I use to predict the 'Y'

Summarizing Data
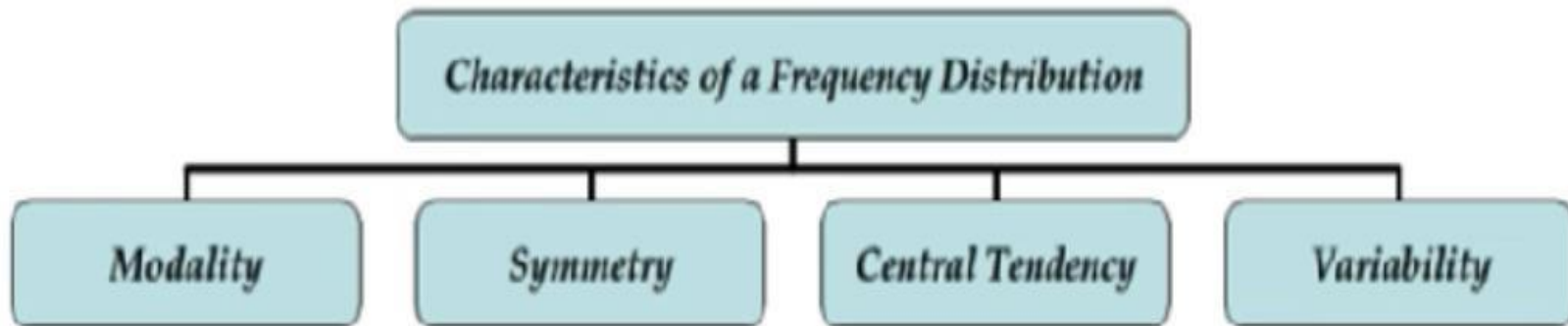
Modality

Symmetry
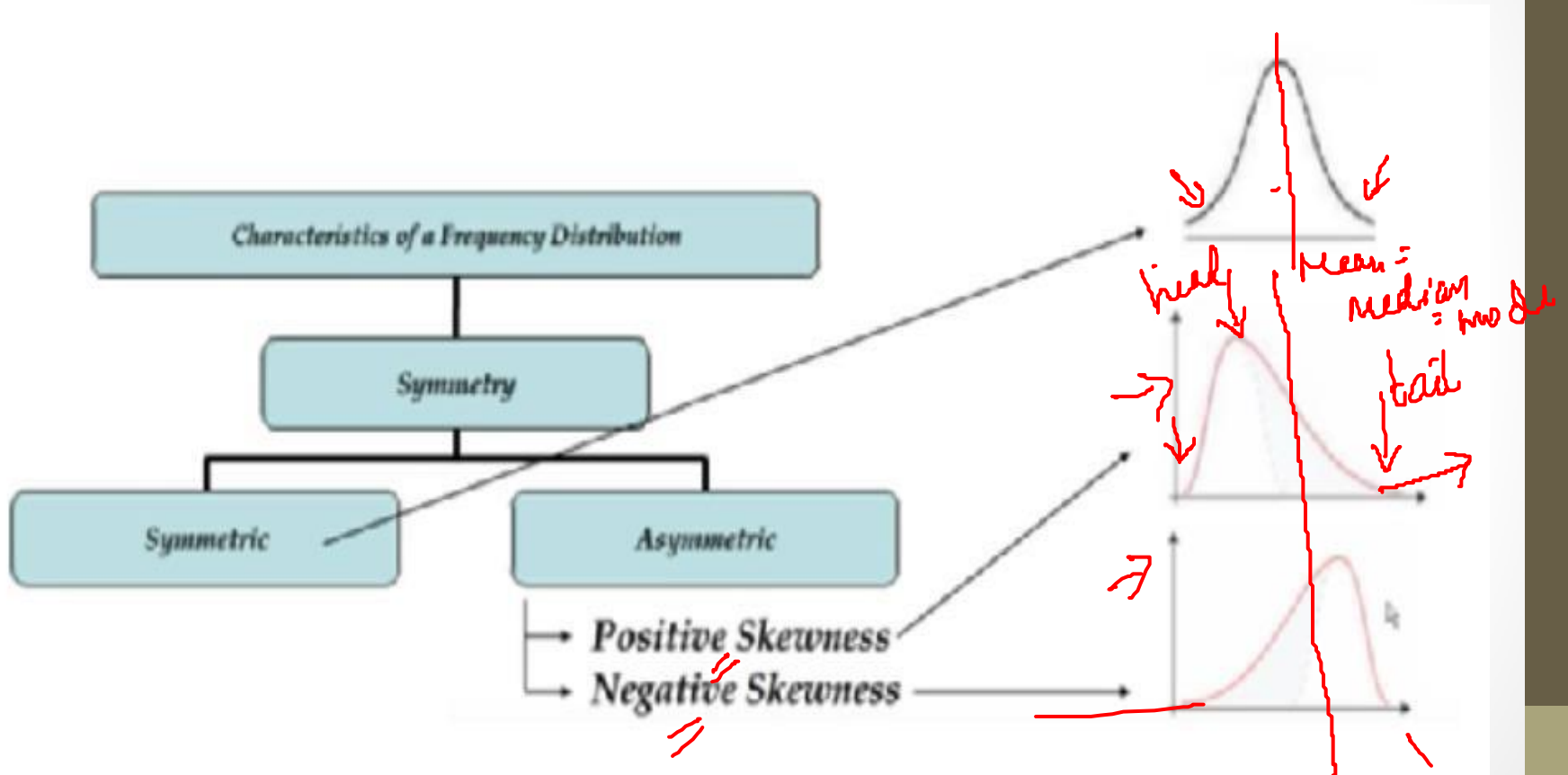
Central Tendency
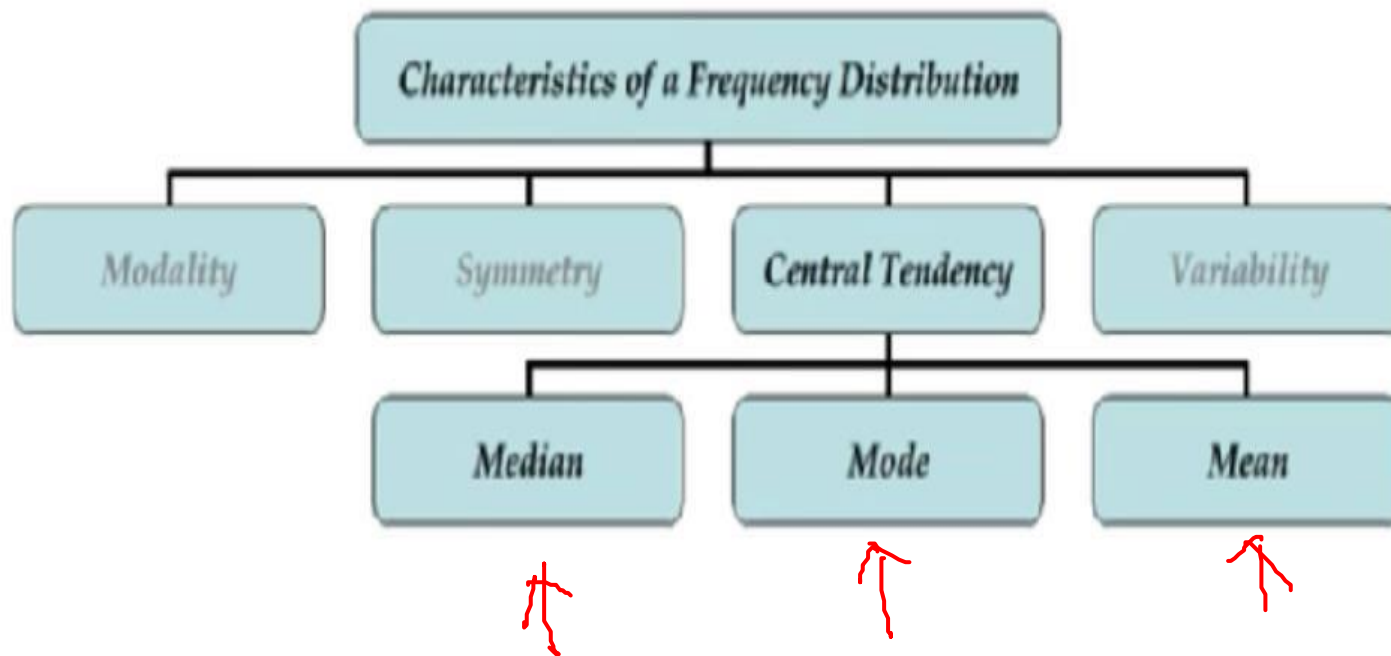
Variability
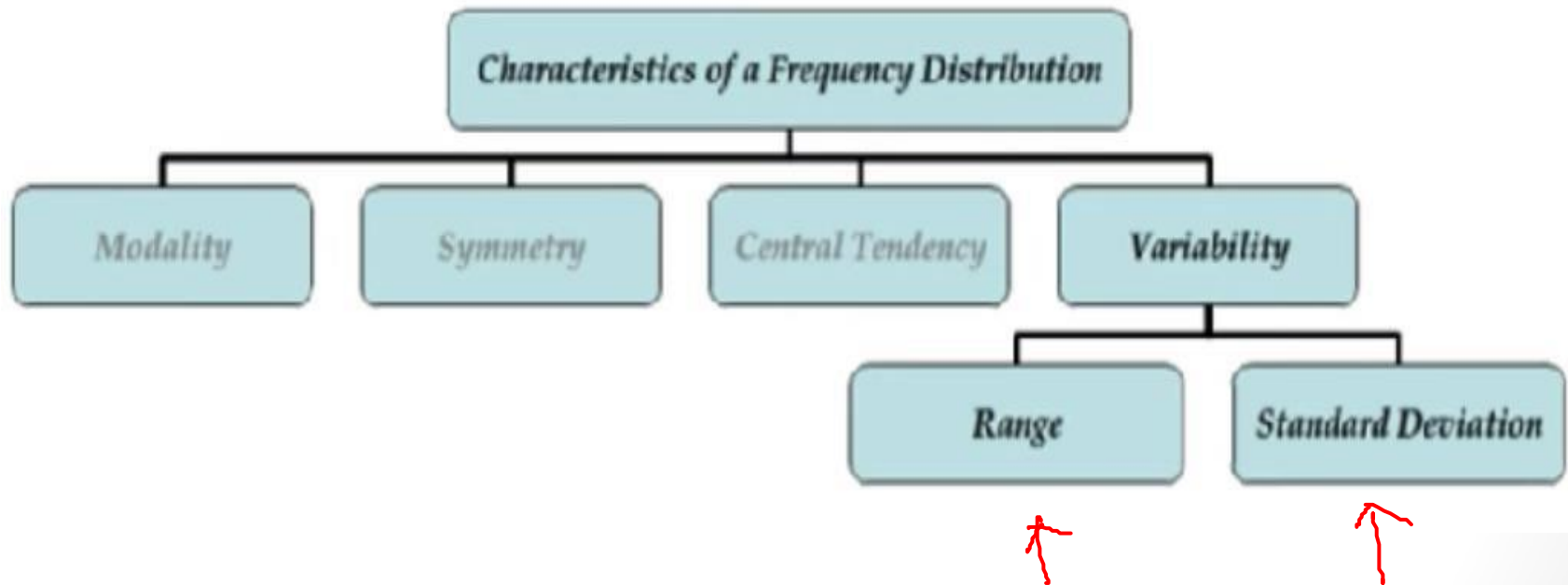
Central Tendency

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

$$\text{Mean, } \mu = \frac{\sum x}{n} \Rightarrow \frac{x_1 + x_2 + x_3 + x_n}{n}$$

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.

Average depth
3 feet

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

$$\overline{5}.7 - 6.9$$

$$5.7 - 10 \qquad \overline{5}.06$$
$$5.8 - 20 \qquad 6.270$$
$$5.9 - 30$$

Why did Alan Drown?

Why did Alan Drown?

AVERAGE DEPTH 3FT

1 ft

2 ft

7 ft

3 ft

2 ft

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

Beware of Averages!!!
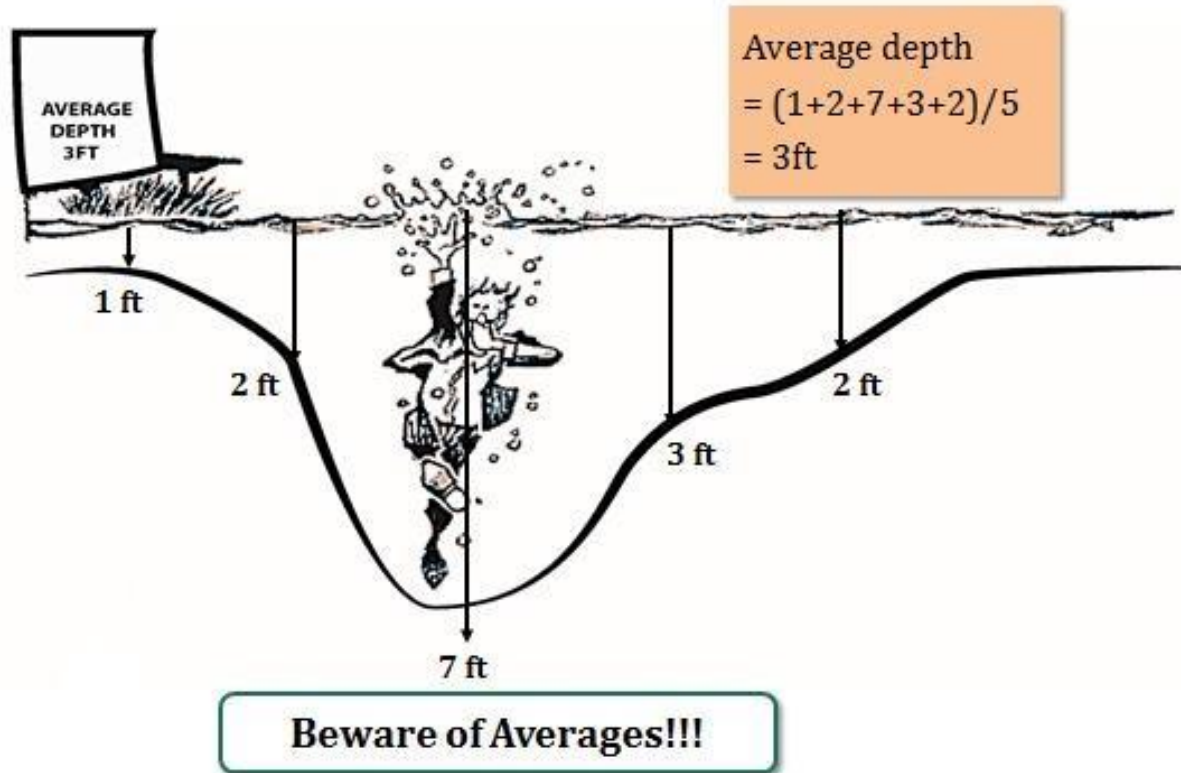
The "Hotshot" Sales Executive

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

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
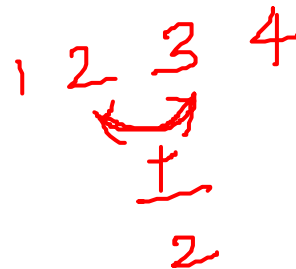
- Average deal size in pipeline
  = $100,000

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

- Median = $55,000 more realistic measure

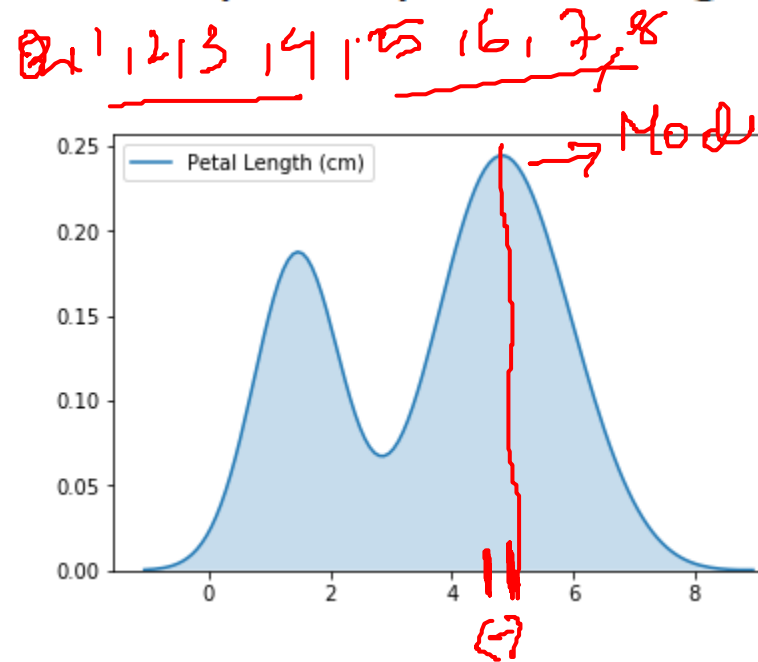| Deal #	Deal Value	Deal Status
1	70,000	Open
2	50,000	Closed
3	55,000	Closed
4	60,000	Closed
5	55,000	Closed
6	50,000	Closed
7	50,000	Closed
8	60,000	Closed
9	50,000	Closed
10	5,00,000	Open

Mean, median, mode
Imputation
↳ Impute
Na, Nan, Null, NaT

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
= (7th + 8th 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.
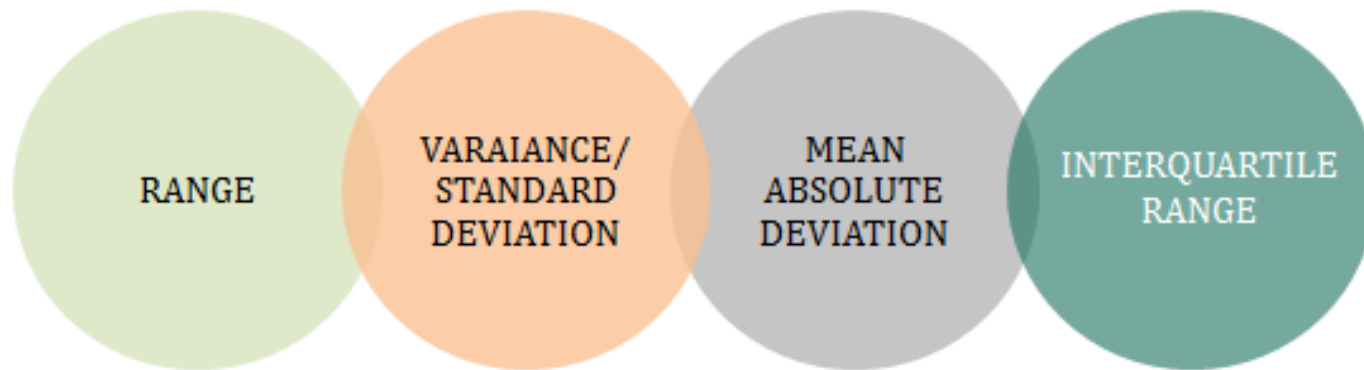
Spread of Data - Range

$$Range = Max - Min$$

Spread of Data - SD and Variance

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

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

μ - mean

$$\rightarrow \frac{x_1 + x_2 + x_3 + x_4}{n}$$

$$\Rightarrow \frac{10}{5}$$

$$\Rightarrow \sqrt{2} = 1.4$$

x | $x - \mu$ | Deviate
-3 | -2
-3 | -1
2 | -3 | 0
3 | -3 | 2
4 | -3
5 | -3
3

0
4

Deviation of each value form the mean

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

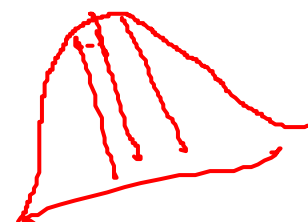| Points scored per game	3	6	7	10	11	13	30
Frequency, f	2	1	2	3	1	1	1

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

[handwritten: 2 ... 6 s4 ... >:12=2]

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

[handwritten:]

$$3 \times 2 + 6 \times 1 + 7 \times 2 + 10 \times 3 + 11 \times 1 + 13 \times 1$$
$$+ 30 \times 1$$
$$= 110$$

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 → 1.787

Player 2 = 3.30823887354653 → 3.308

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

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

→ Percentile

R = (P/100)*n

↓ Rank

└→ Values

P – The percentile you are interested

n – Number of values

CAT – Common Apt-Test

99%

98%

85 percent 85%

← of people below your score –

EXAMPLE FOR PERCENTILE

If the scores of a set of students in a math test
are 20 , 30 , 15 and 75 what is the percentile rank of the
score 30?

20 , 30, 15, 17

EXAMPLE FOR PERCENTILE

If the scores of a set of students in a math test are 20 , 30 , 15 and 75 what is the percentile rank of the score 30?

Arrange the numbers in ascending order and give the rank ranging from 1 to the lowest to 4 to the highest.

NUMBER	15	20	30	75
RANK	1	2	3	4

Example for percentile

Use the formula now,

$3 = (P/100)4$

$75 = P$
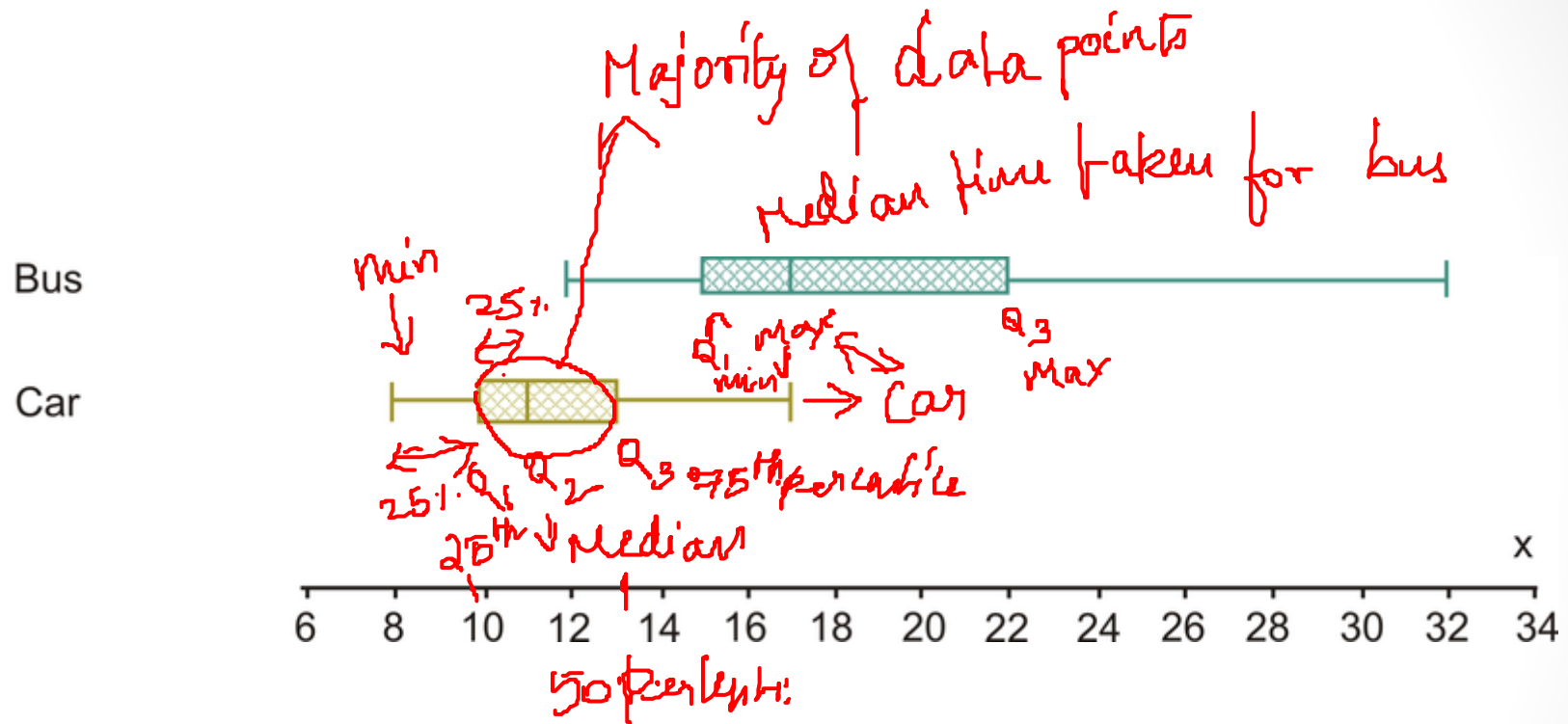
Therefore, the score 30 has 75th percentile

BOX plot

A data scientist conducted a survey of times it takes for him to reach to the office from his home. He drove through Car and recoded the times and went through bus and recorded the time

BUS (min)	12	14	16	16	17	18	22	25	32
CAR (min)	8	9	10	10	11	11	12	14	17
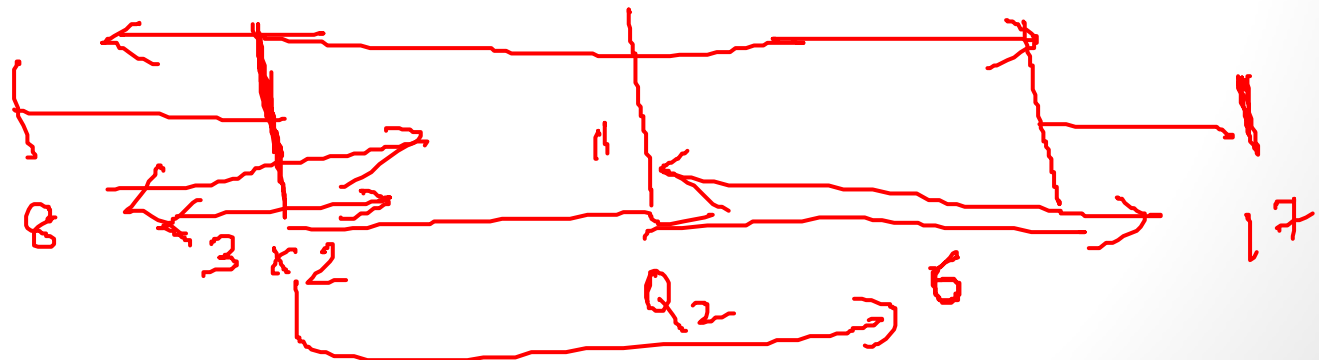
BOX plot

Inter Quartile Range

Quartile

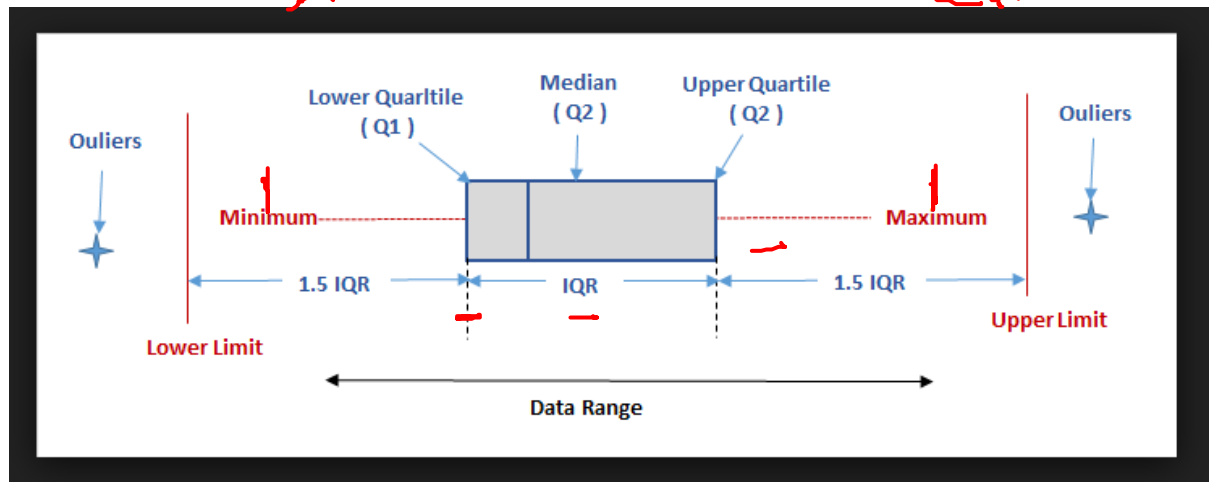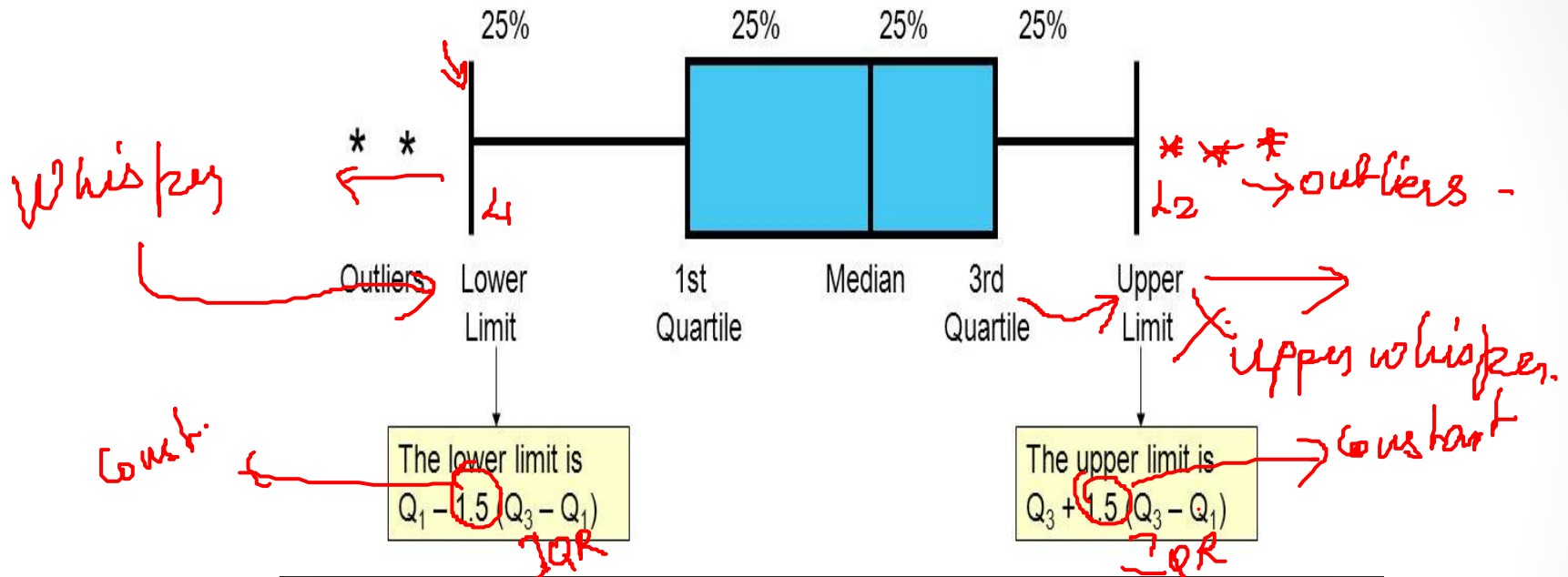Dividing data into ¼ – 4 parts

Q1 – First Quartile – 25th percentile

Q2 – Second Quartile – 50th percentile (Median)

Q3 – Third Quartile – 75th percentile

IQR (Inter Quartile Range) = Q3 – Q1

Box-and-Whisker Plots to find outliers

Box-and-Whisker Plots to find outliers