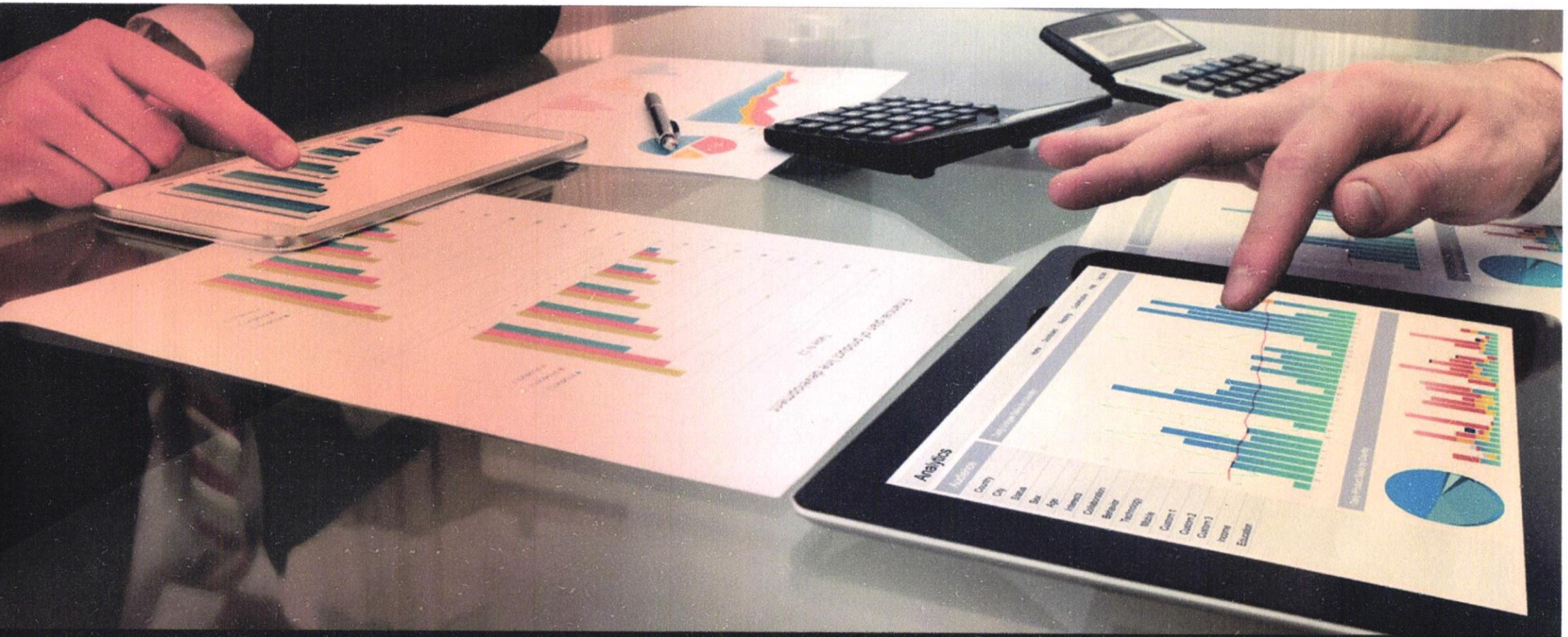
Cohort
July 2021

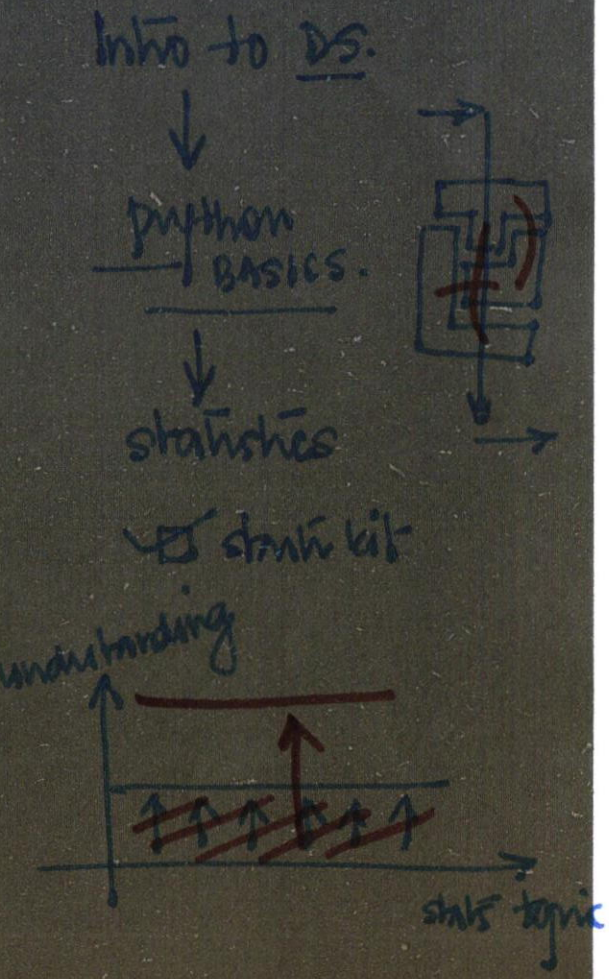


Introduction to Statistics

INSAID

Agenda

- **What is Statistics?**
- Importance of Statistics
- Applications of Statistics
- What is Data?
- Data On Tap
- Types of Data
- Data Quality Issues
- Population vs Sample
- Types of Statistics
- Descriptive Statistics



INSAID

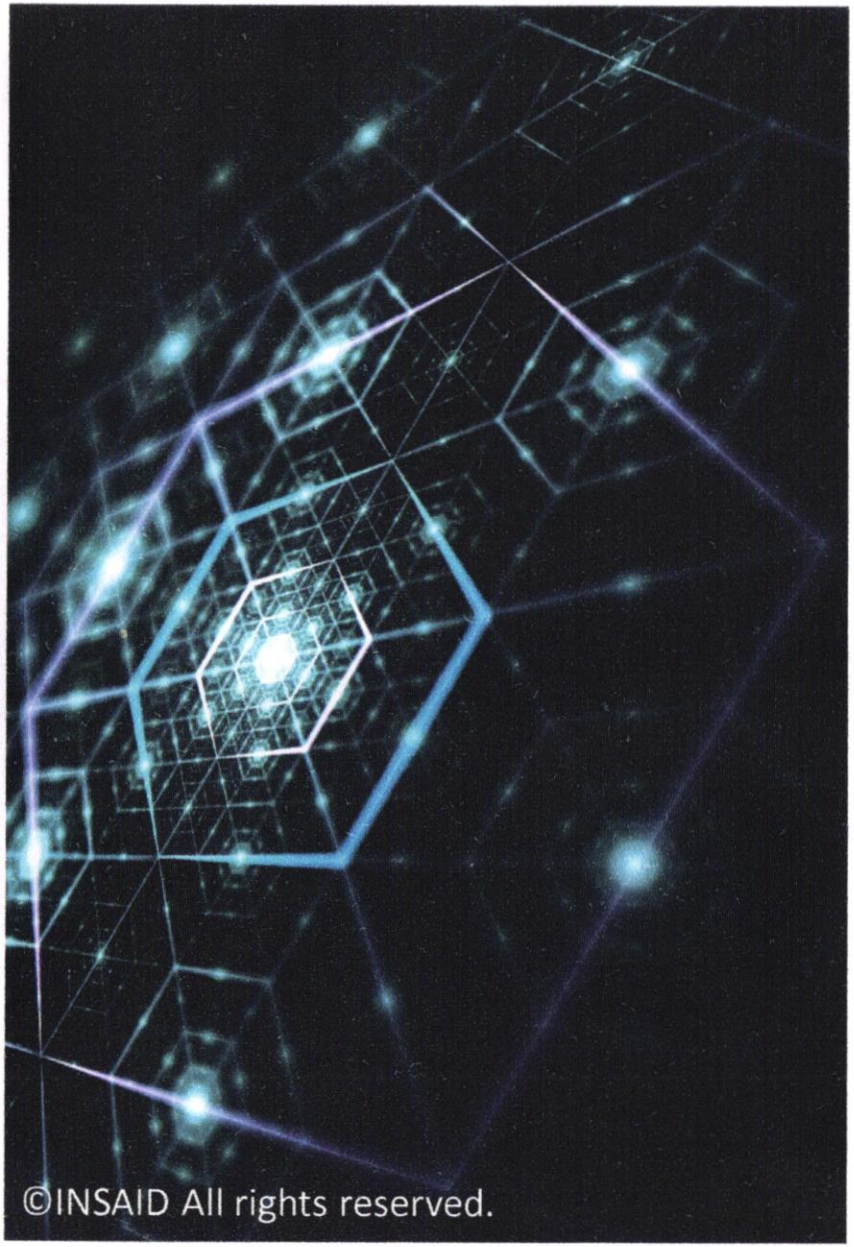
What is Statistics?

- A form of **mathematical analysis**.
- Science of **conducting studies** to gather, review, analyze and draw conclusions from data.

Importance of Statistics

- Business Growth
- Collect the relevant data
- Knowledge Discovery
- Pattern Recognition
- Forecasting

INSAID

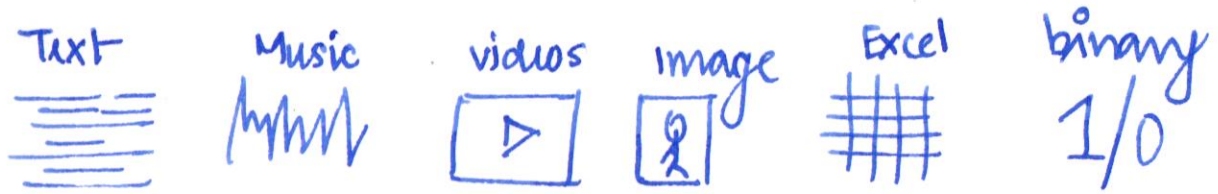


blood pressure 120/80

Applications of Statistics

- **Medicine:** WHO research on epidemic spread
  - #covid ts
  - vaccination rate
- **Business:** Sales Forecasting — Annual sales, Revenue.
- **Weather Forecasting:** Analysis over daily weather — Temp, Hum, Expected Rainfall
- **Stock Market:** Stock profit analysis — price, momentum.
- **Health:** Analysis of human genes — BP, pulse rate, sugar, ECG, CBC.
- & Many More...

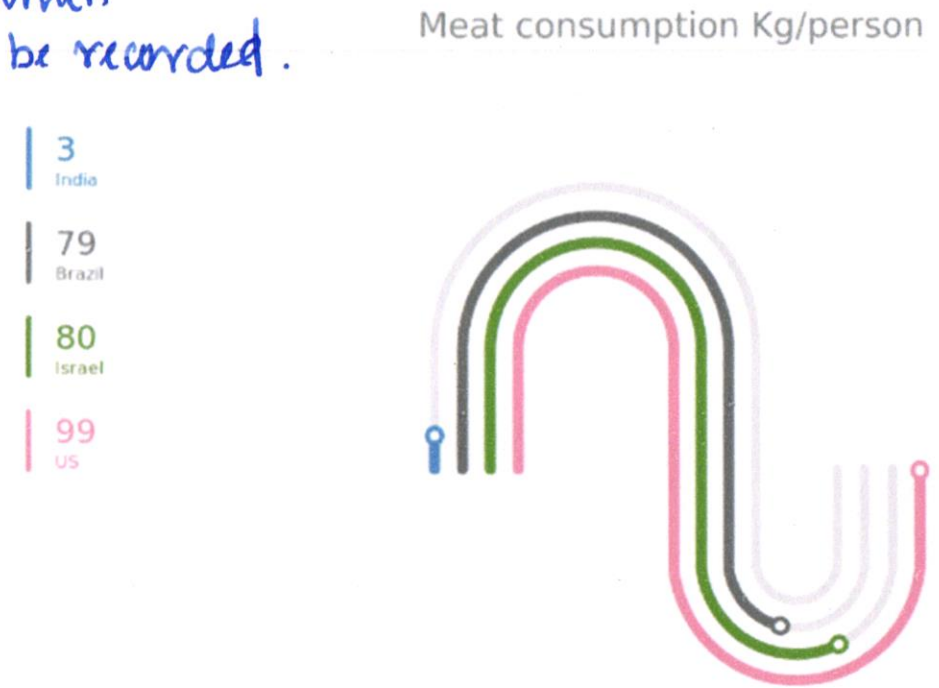
INSAID



What is Data?

anything which can be recorded.

- **Raw** or **isolated facts** from which the required information is produced.
- Can be **recorded** & have **implicit meaning**.
- **Example:** Meat consumption Kg/person



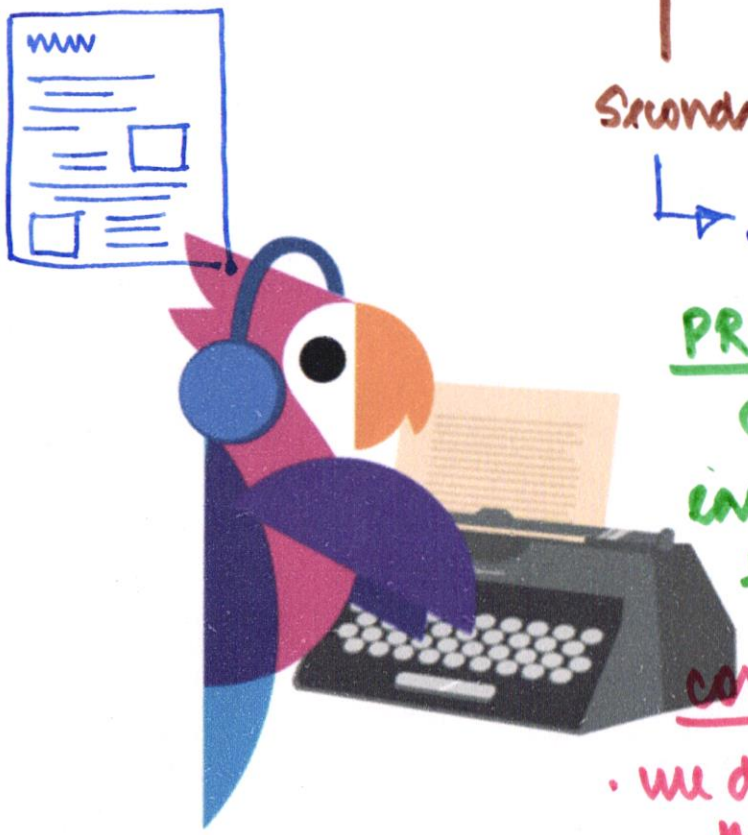
INSAID

Data On ~~Tap~~ Wire / Internet

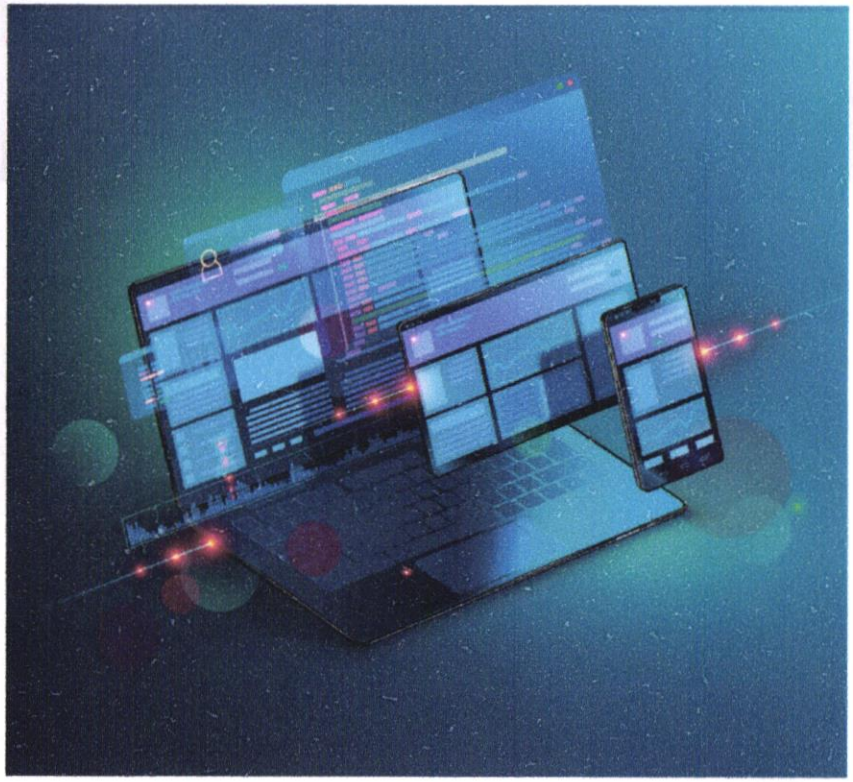
- Generally, data is available in two forms:
  - Primary Form
  - Secondary Form

INSAID

Primary Form

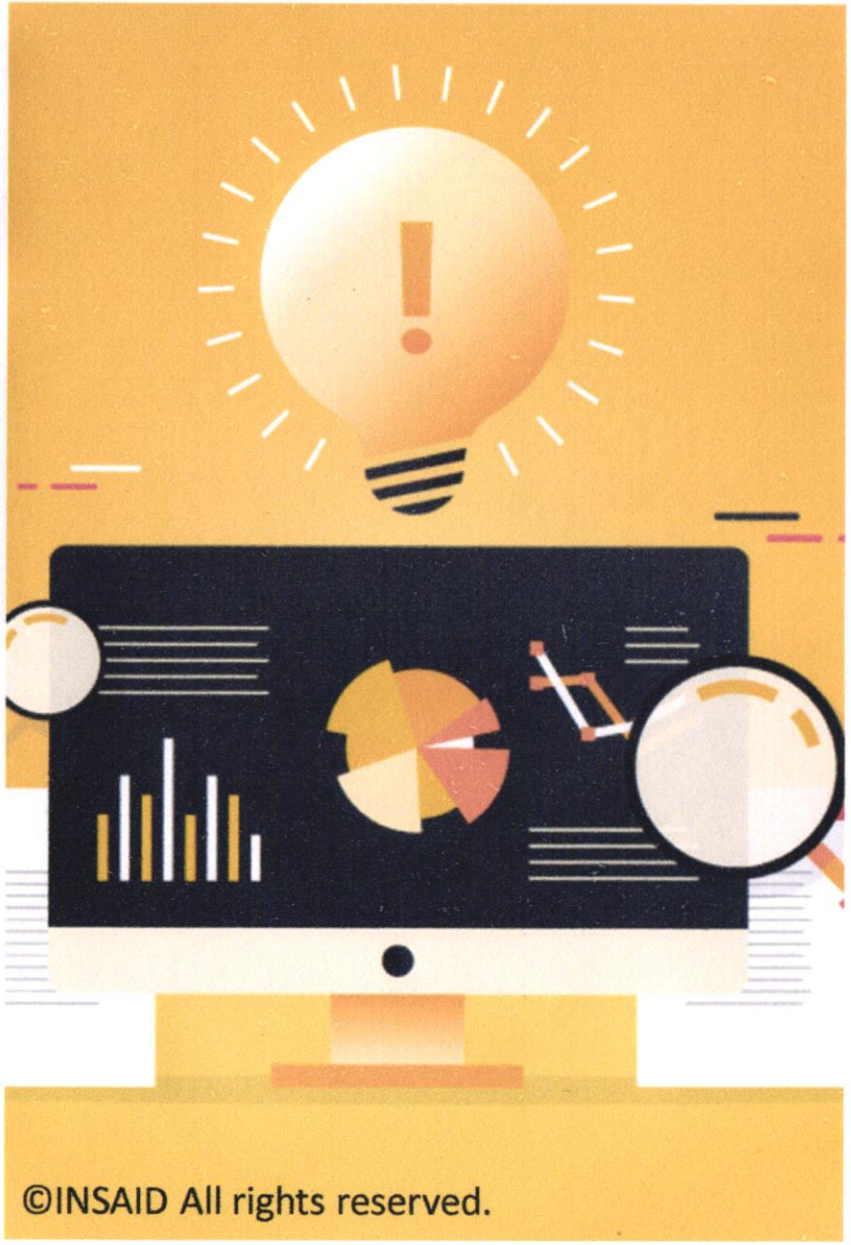


- Collected first hand by the researcher.
- It can be acquired in the form of -
  - Interviews, Observation, Questionaries, Case Studies, Group Discussions etc.

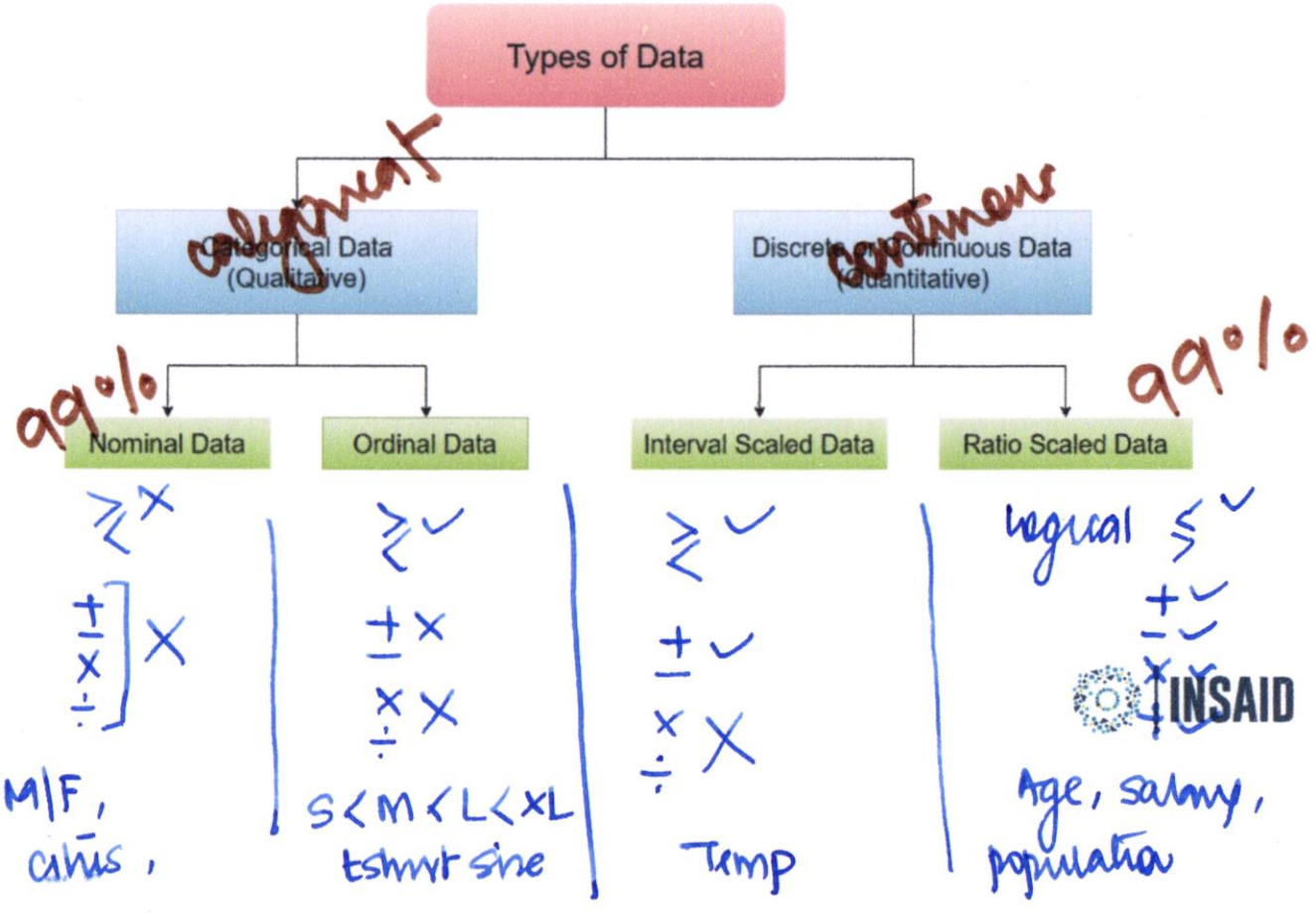


Secondary Form

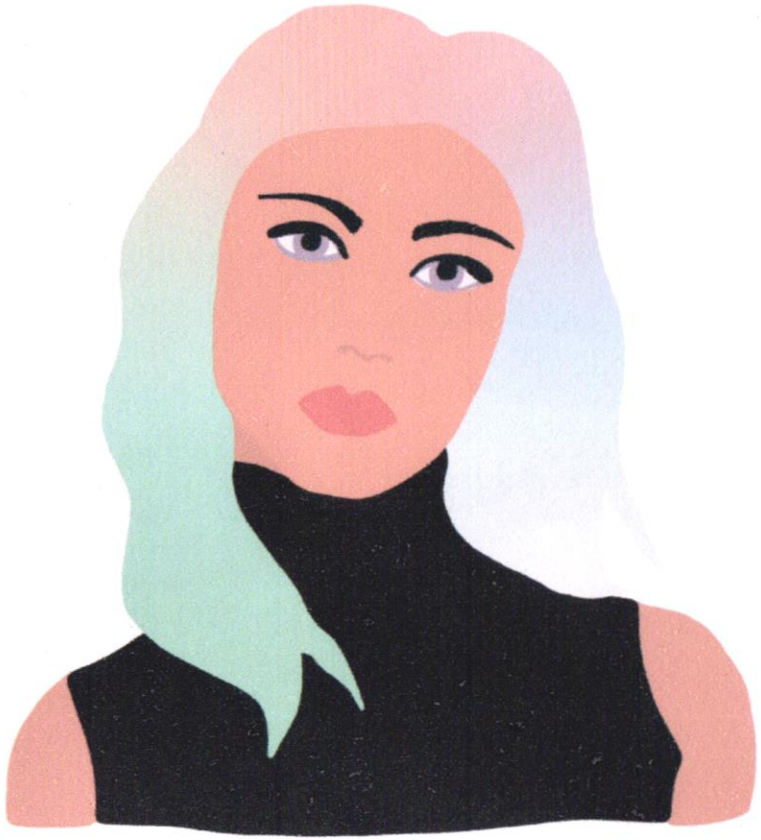
- Already collected by and readily available from other sources.
- It can be acquired in the form of -
  - Previous Research, Diaries, Letters, Web Info, Census Data etc.



Types of Data



Nominal Data



- Data related to names.
- The values are symbols or names.
- Also referred as Categorical Data.
- **Example:** Hair Color

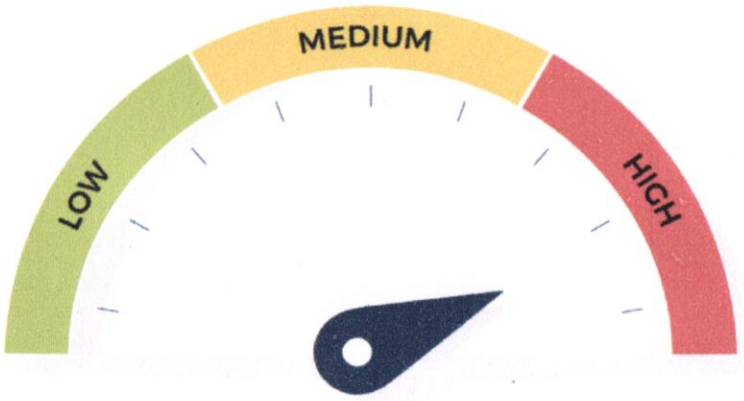
example of Nominal

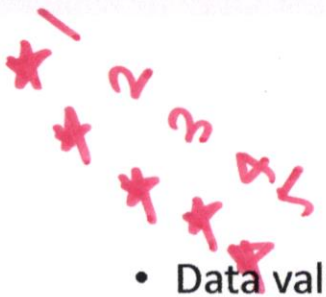
Binary Data

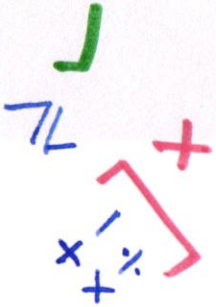
- Nominal data only with two categories.
- **Example:** Winner & Loser.
- **Note**: The **order** is **NOT** important such as Male/Female or Female/Male

INSAID

Ordinal Data





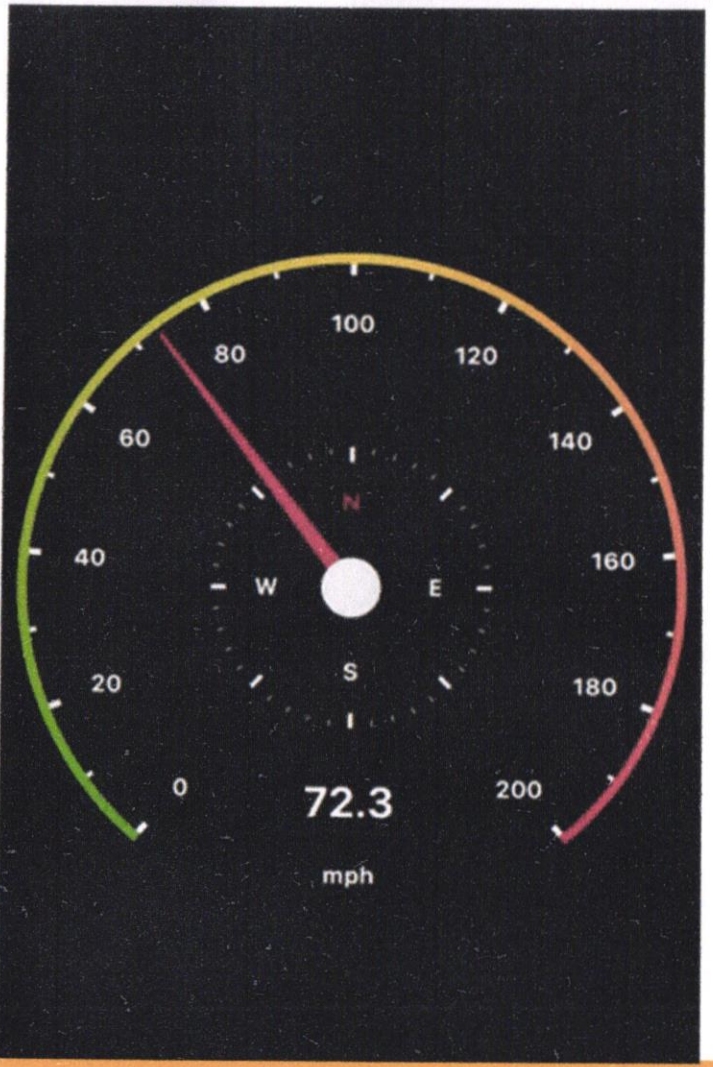


- Data values have meaningful order or ranking.
- Magnitude between successive values is not known.
- **Example:** Economic Status

INSAID

Numeric Data – Interval Scaled

- It is quantitative in nature with non-inherent zero-point.
- Can be **discreet** or **continuous** in nature.
- Measured on scaled of equal units.
- Can be Positive, Zero or Negative.
- Allows quantify the difference between values.
- **Example:** Temperature

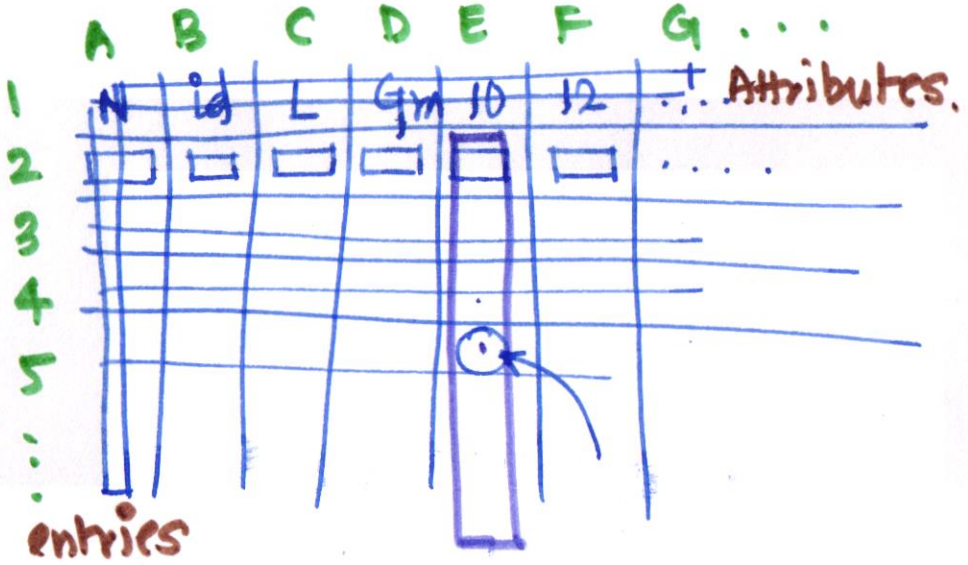


Numeric Data – Ratio Scaled

- It is quantitative in nature with true zero-point.
- Can be **discreet** or **continuous** in nature.
- Measured on scaled of equal units.
- Can be Positive, Zero or Negative.
- Allows quantify the difference between values.
- **Example:** Speedometer

INSAID

Data Quality Issues



- Duplicity
- Inconsistency
- Correctness
- Timeliness
- Missing Values

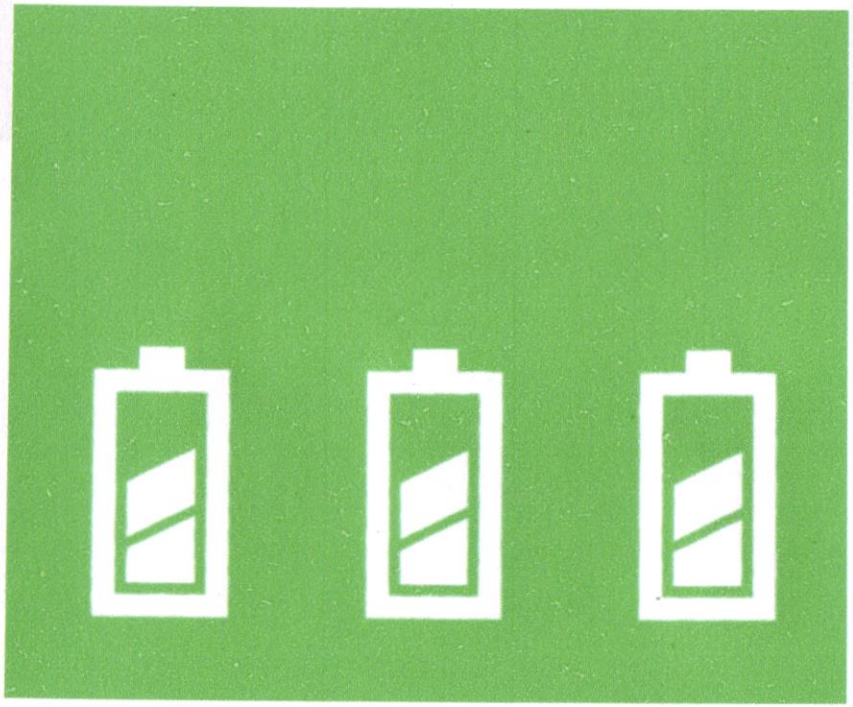
Exit polls.

Population vs Sample

- **Population:** Entire group that you want to draw conclusions about.
- **Sample:** Specific group that you will collect data from.
- The size of the **sample** is always less than the total size of the **population**

INSAID

Example: Population vs Sample



- **Company** A **wants** to **buy 50,000 batteries**.
- Only buys if no more that 1% are defective.
- **Testing** all batteries is **expensive**.
- Instead select few samples (500) and test them.
- The results can show estimate of the percentage of defective batteries in population.

INSAID

Random variable

Those **variables** whose value is unknown until the experiment is done.

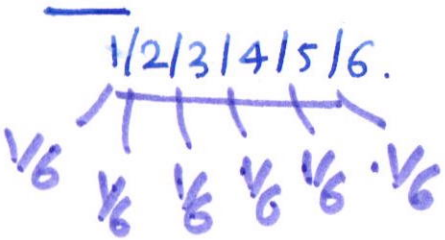
but we know some thing about it.

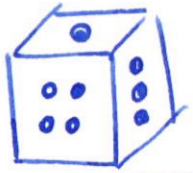
Probability

chances of occurring

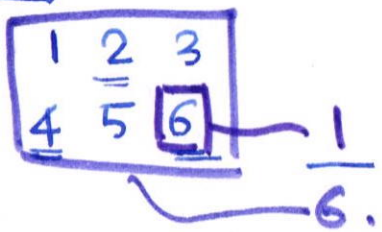
$$P(H) = \frac{\#\text{fav. outcomes}}{\text{Total \# outcomes}} = \frac{1}{2}$$

unbiased



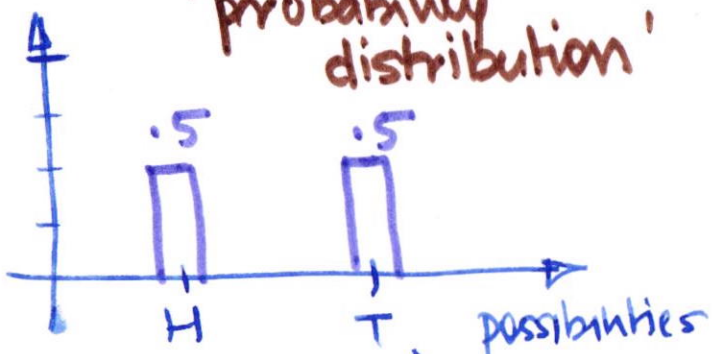


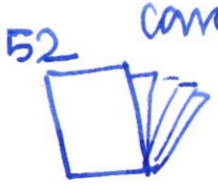
$$P(6) = \frac{1}{6}.$$

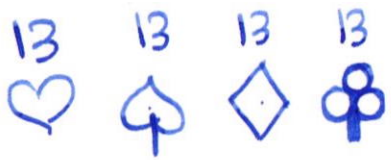


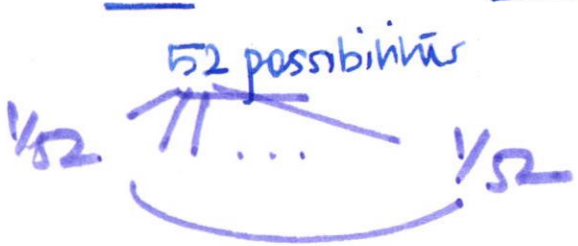
$$P(\text{even}) = \frac{3}{6} = \frac{1}{2}$$

probabilities







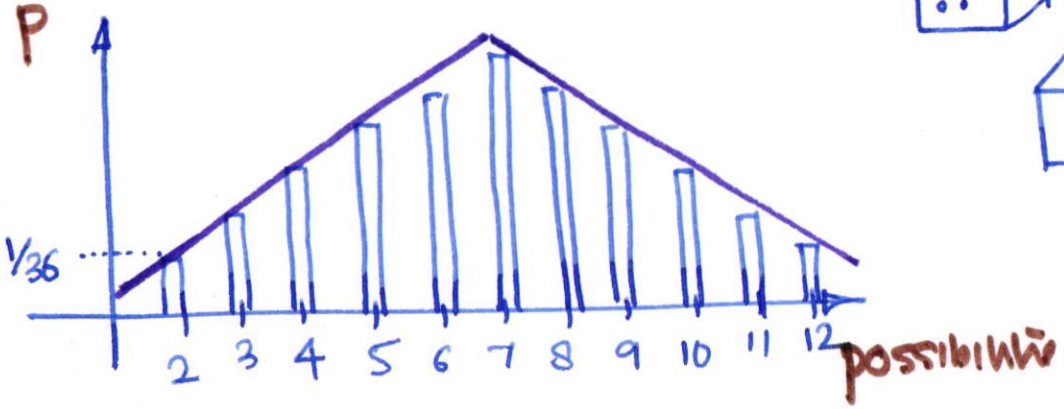


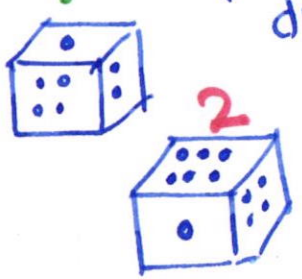
$$P(H) = \frac{13}{52} = \frac{1}{4}$$

$$P(\text{king}) = \frac{4}{52} = \frac{1}{13}$$

Random variable

Prob dist.





pair of dice.

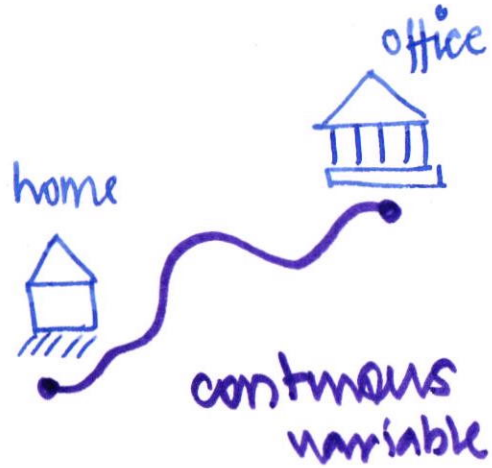
sum of the outputs.

[2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]

prob. distributions can be different

outcome 2

outcome 1	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	
4	5	6	7	8		
5	6	7	8			
6	7	8				12



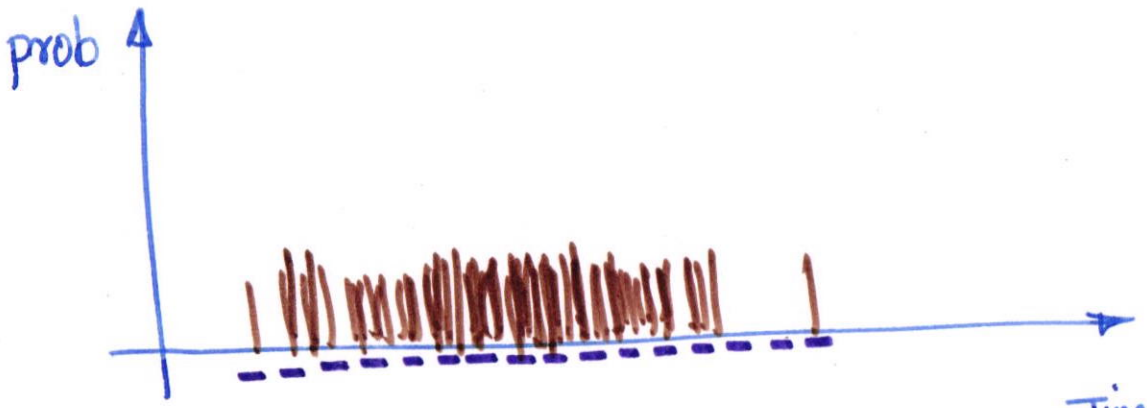
distance
Home - office = _____ RV ✗

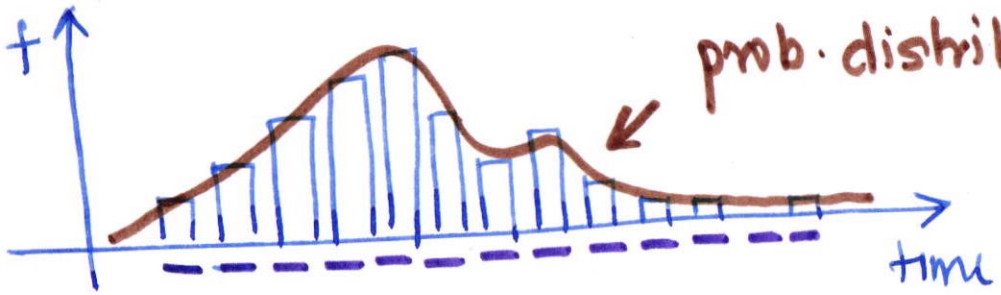
Time = _____

traffic, speed, mode of trans, weather, 'barrat'

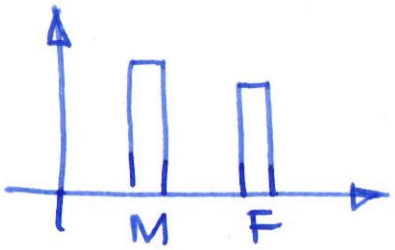
Time.
[20.0222.56.
21.3121.37,
51.01 15.34
. . . . .
. . . . .
. . . . .
. . . .] bucketing

20:00 - 20:15
20:16 - 20:30

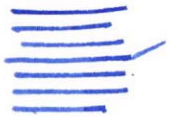




Exit Polls.



130 voters.



Sample

Representative

state.

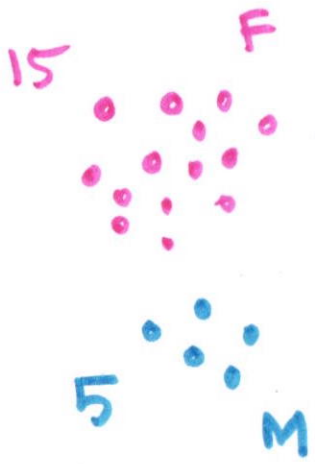
language

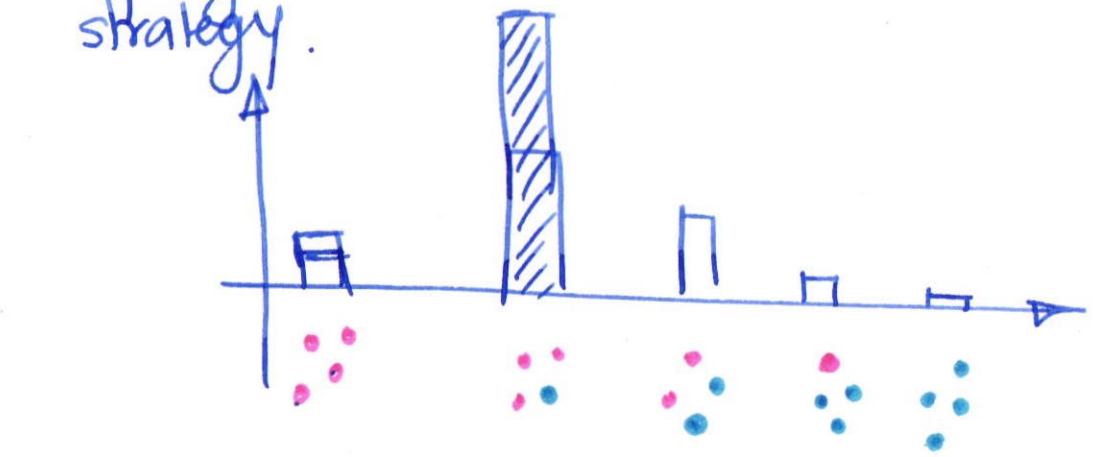
region : ...

.... demographic

prob. dist of SAMPLE $\approx$ prob. dist of POPULATION

sampling strategy.





4 people at RANDOM.

Randomly with each person being chosen equally likely.

summarise data

'conclude'

Types of Statistics

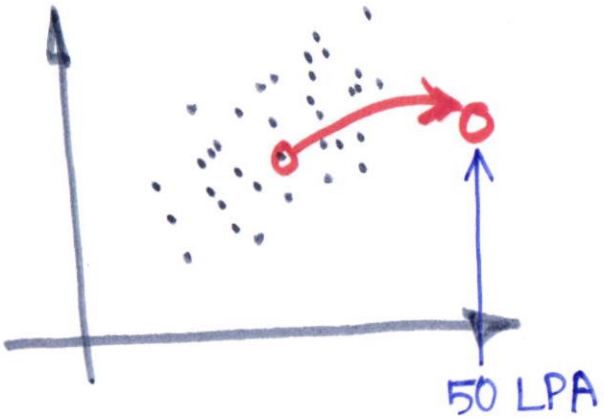
- Descriptive Statistics
- Inferential Statistics

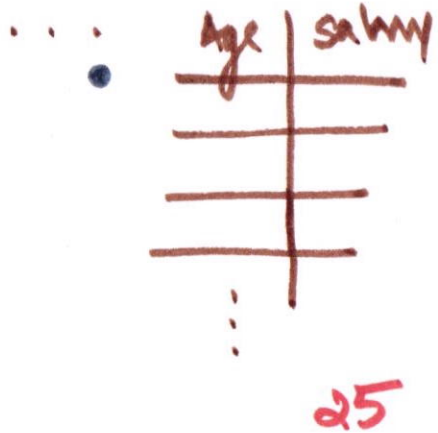
central tendency
mean mode median

measure of spread
SD, var, R, IQR

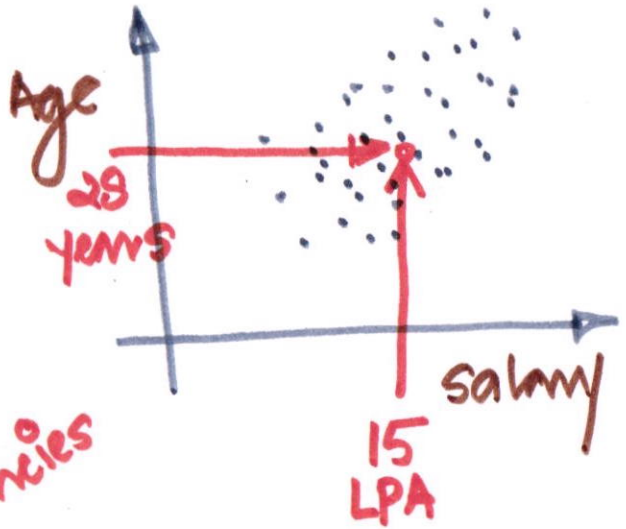
measure of shapes
skew, kurtosis

INSAID





scatter



central tendencies

$$\text{Sum} \sum_{i=1}^{n} \text{Variable}_i$$

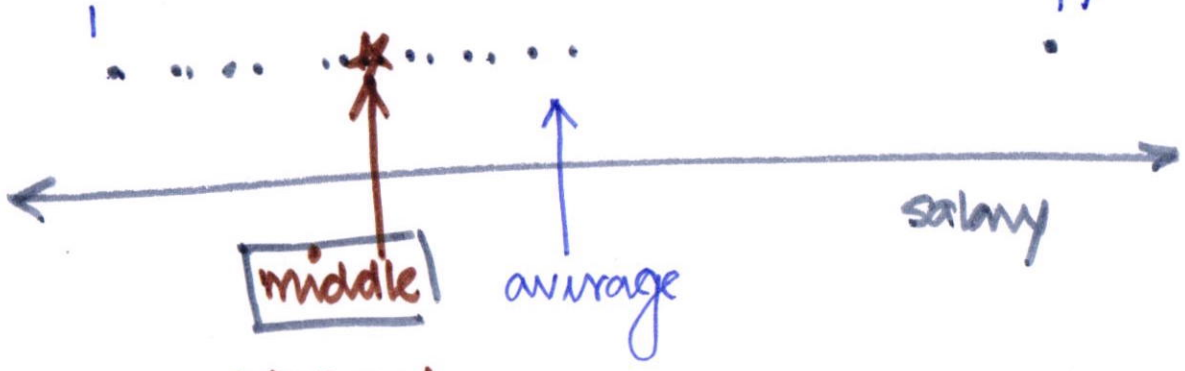
central

Average ↔ mean

$$\frac{\sum_{i=1}^{n} \text{salary}_i}{n}$$

problem –

if there is an outlier, Avg is not representative

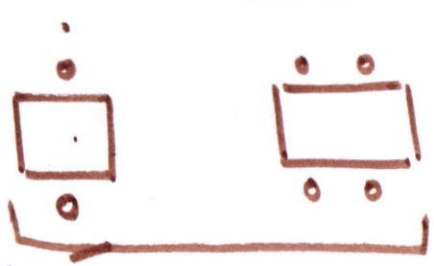


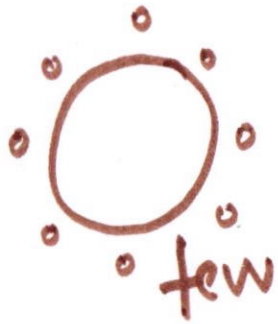
MEDIAN

value of 50 percentile person.

better representation than average, if there is an outlier.

Tables.





few

time
continuous

mode

group	# people
1	4
2	3
3	1
4	6
5	7
6	4
7	3

most frequent

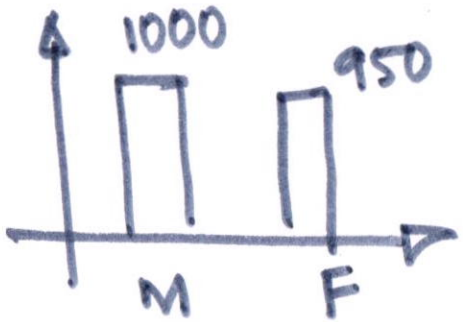
4.25

MODE

value which is most frequent.

MODE

categorical variables



mode (Gnaw) = M

Descriptive Statistics

- Measures of Central Tendency
- Measures of Spread
- Measures of Shape

INSAID

Measures of Central Tendency

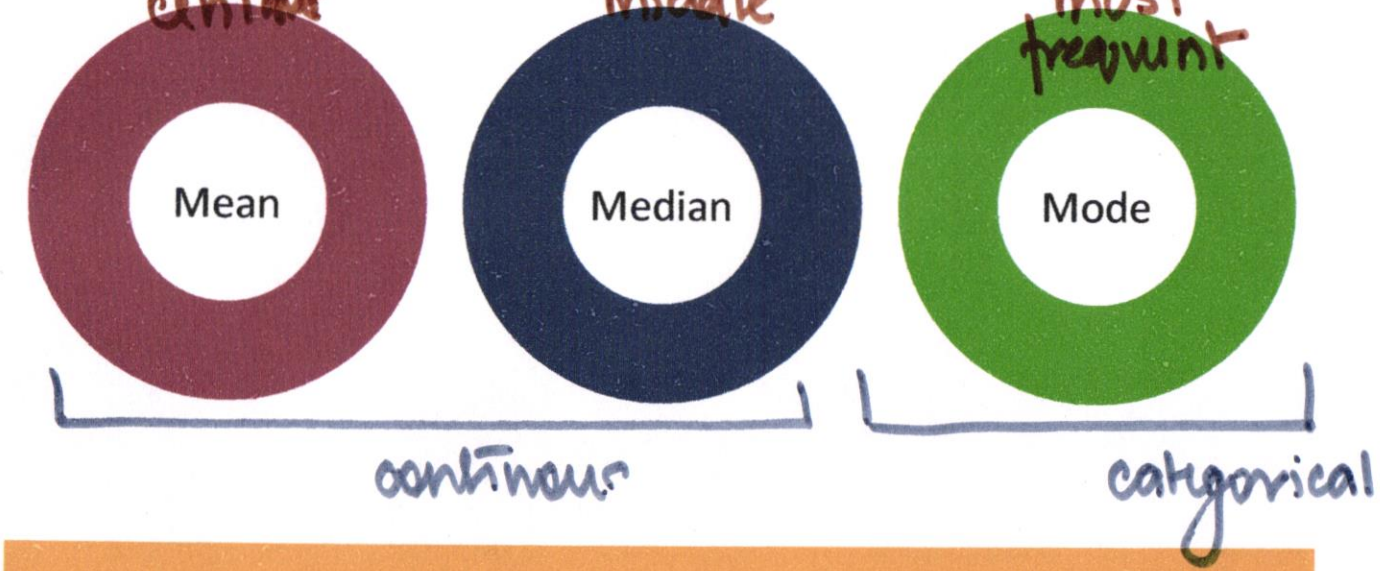
$\frac{\sum_{i=1}^{n} x_i}{n}$ centre

value at 50th percentile

Middle

most frequent

most frequent



INSAID

Rs. 5 Lakhs

Rs. 6 Lakhs

Rs. 7 Lakhs

Rs. 8 Lakhs

Mean

- It is the **average** of a set of data.
- **Example:** Average salary of employees
- Mean = (5 + 6 + 7 + 8) ÷ 4 = 6.5 Lakhs

INSAID

Rs. 5 Lakhs

Rs. 6 Lakhs

Rs. 7 Lakhs

Rs. 25 Lakhs

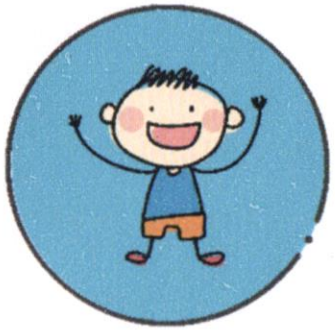
Mean (Outliers Effect)

- Mean provides a misleading balance point because of an outlier.
- Mean = (5 + 6 + 7 + 25) ÷ 4 = 10.7 Lakhs

INSAID

30 Kg

35 Kg



38 Kg

40 Kg

42Kg

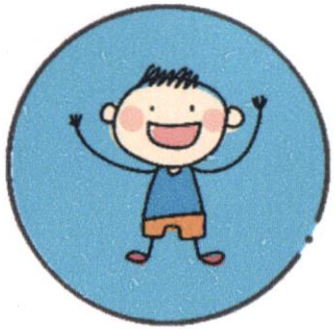
Median

- It is the **middle value** in a set of ordered data values.
- **Example:** Weight of kids.

INSAID

30 Kg

35 Kg



38 Kg

40 Kg

75 Kg

Median (Outliers Effect)

- Median is a positional average and refers to the middle value in a distribution.
- **Example:** Weight of kids.

INSAID

Mode = 2

Mode = 3

Mode

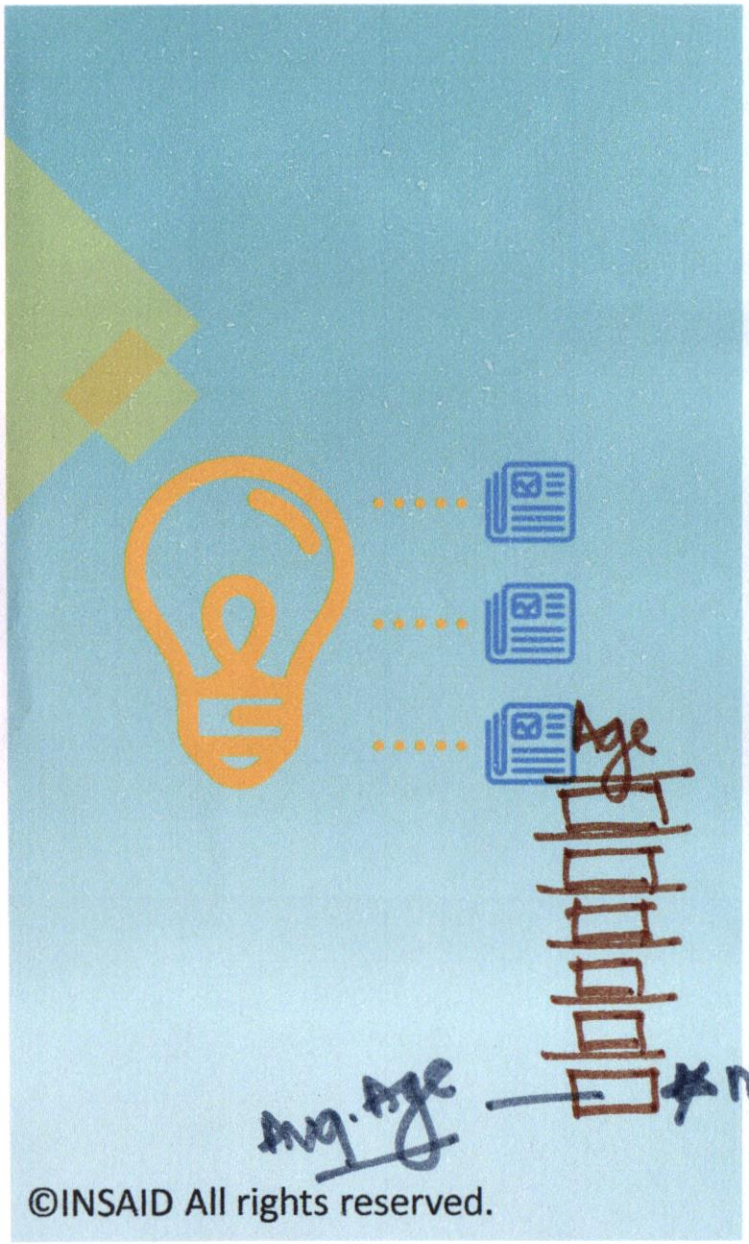
- It is the most frequently occurring value in the dataset.
- **Example:** In above diagram, Male is occurring the most.

INSAID

Pizza Table Example

- Deciding the seating arrangement of a restaurant by figuring out the most frequently occurring group size.
- A sample of 20 groups is selected at random:
- People = {2, 4, 1, 2, 3, 2, 4, 2, 3, 6, 8, 4, 2, 1, 7, 4, 2, 4, 4, 3}
- There are 2 modes, each occurring six times – 2, 4
- Basis the above, the manager will decide on 2 seater and 4 seater tables being kept in the patio.

INSAID



hope - the centre represents the group.

Central Tendency Use Cases

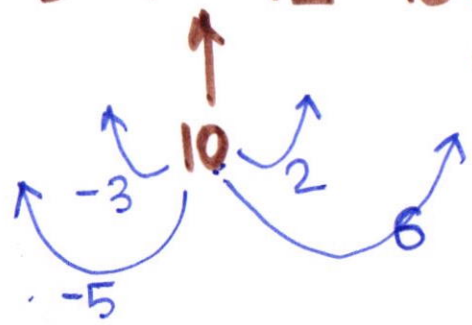
- Understand data spread around the center in Data Analysis
- Replace missing values in Data Analysis
- Showcase performance of players across their career
- Study salary spread across the company
- & Many More...

INSAID

We represent the data with MEAN

A B C D E

5 7 12 16.



$$\text{average} = \frac{5+7+12+16}{4}$$

$$= 10 = \mu \text{ (mu)}$$

on an average how much are we wrong, in saying centre is representati -ve

$$\text{average mistake} = \frac{(-3)+(-5)+(2)+(6)}{4}$$

$$= 0$$

$$\text{avg error} = \frac{\sum_{i=1}^{n} e_i}{n}$$ ×

$$\text{avg. abs. Error} = \frac{\sum_{i=1}^{n} |e_i|}{n}$$

$$= \frac{3+5+2+6}{4}$$

$$= 4$$

$e_i = \mu - x_i$

mean $- x_i$

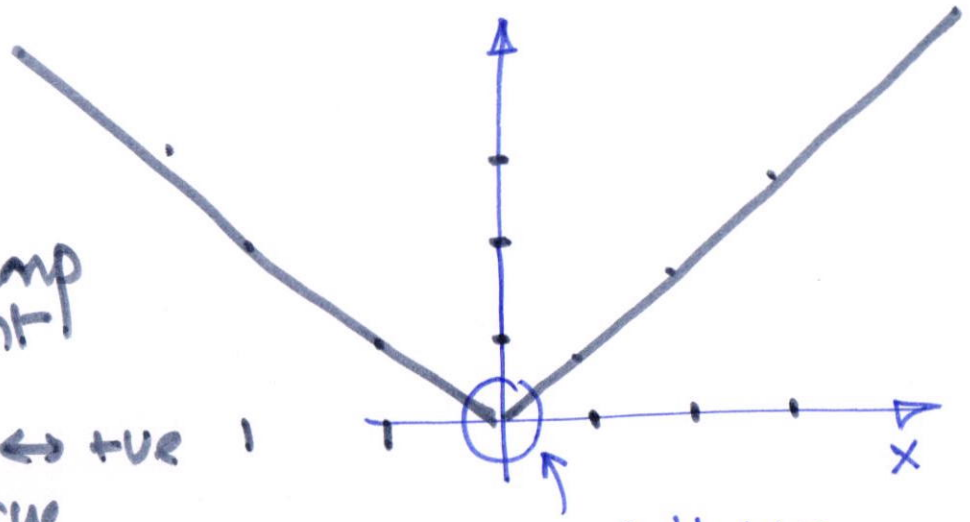
variance σ^2

$$\frac{\sum_{i=1}^{n} (e_i)^2}{n}$$

$= 18.5$ $(\text{lak Rs})^2$

- ☑ no sharp point
- ☑ −ve ↔ +ve issue solved

$y = |x|$



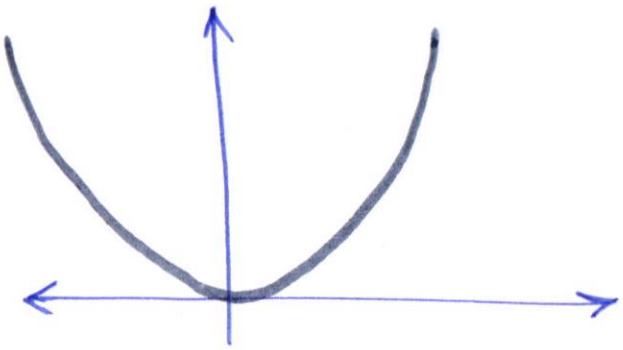
we don't like sharp point

- ☑ lost the meaning

standard deviation σ (sigma)

$$\sqrt{\frac{\sum_{i=1}^{n} (e_i)^2}{n}}$$

$= 4.3$ Lakh Rs.

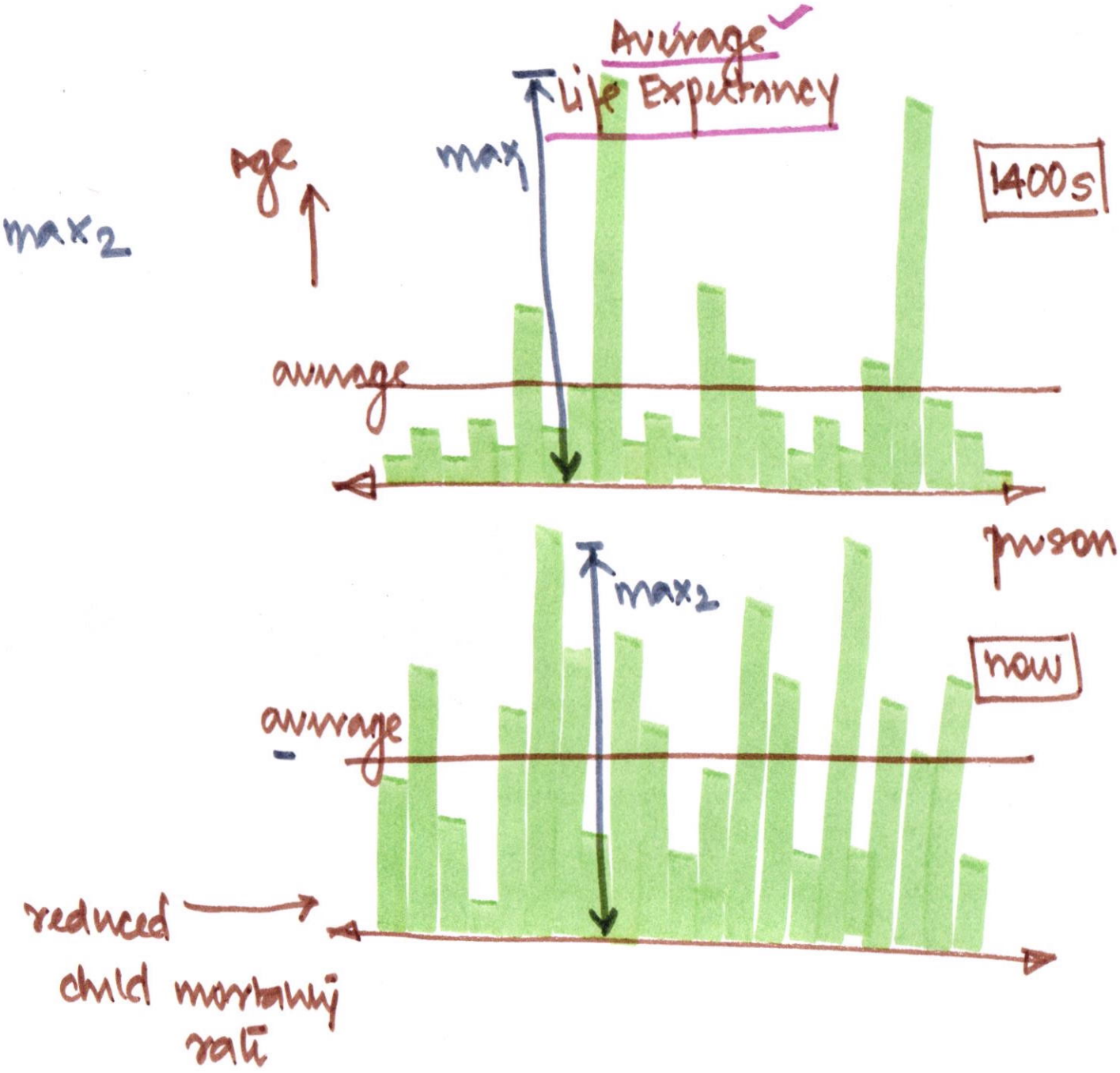


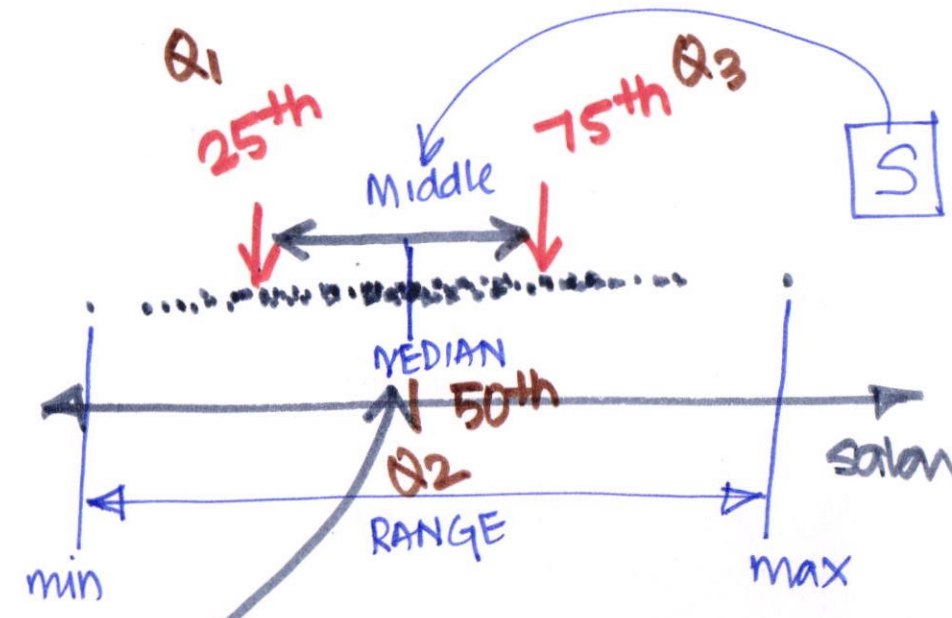
Age outlier?

YES NO

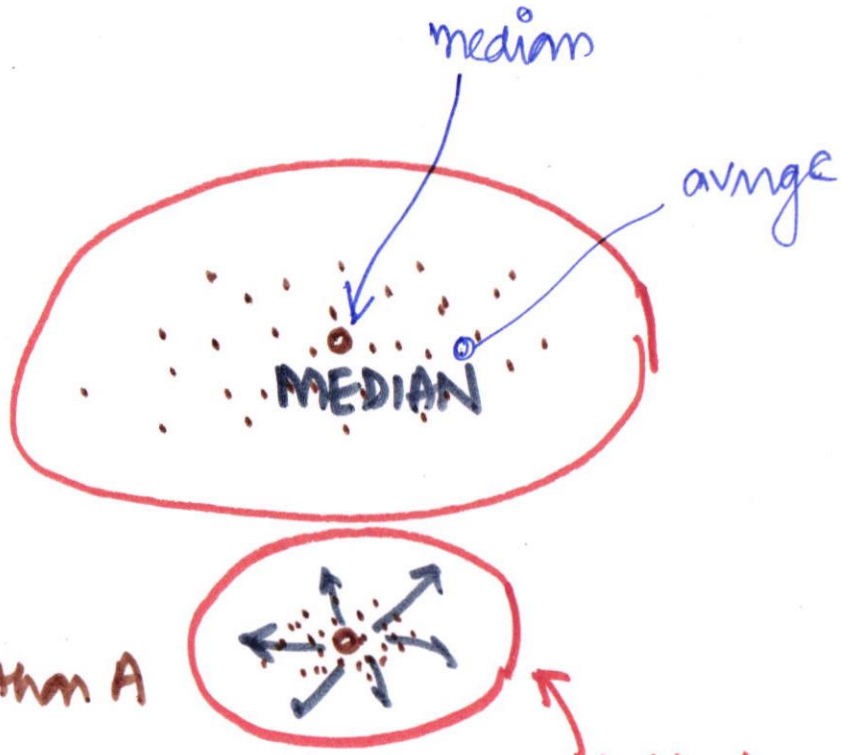
$max_1 > max_2$

AVERAGE





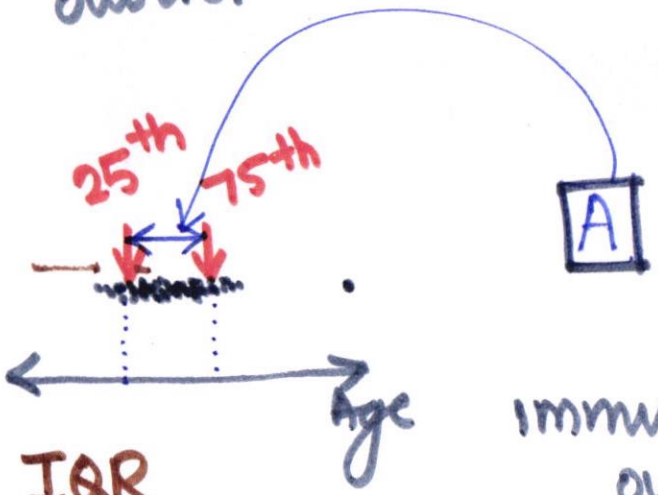
immuned to outlier



S is bigger than A

|||

salary has more spread than A



IQR

inter quartile range

immuned to outlier

Invisible line of demarkation is around the Median.

radius of demarkation ∝ spread

IQR

it point under study ≡ far enough from MIDDLE

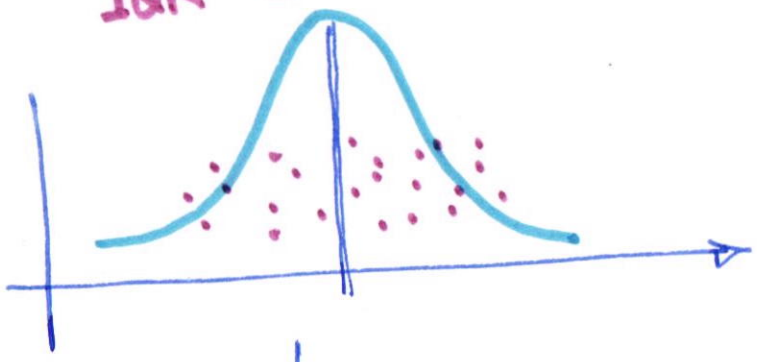
Then we can call it outlier.

C2

✗ Average salary = ~~8~~ 7 LPA

✓ median = 6.9 LPA

IQR =



no outlier

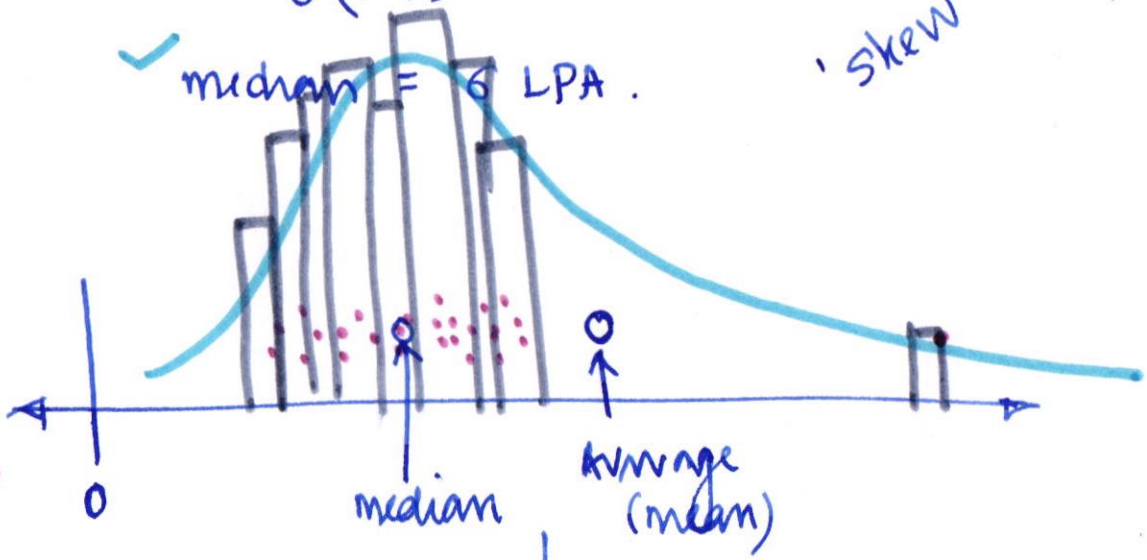
vs

C1

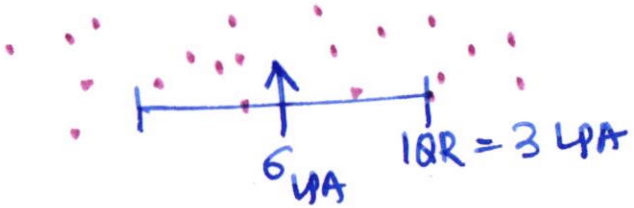
✗ Average salary = 10. LPA (CTC)

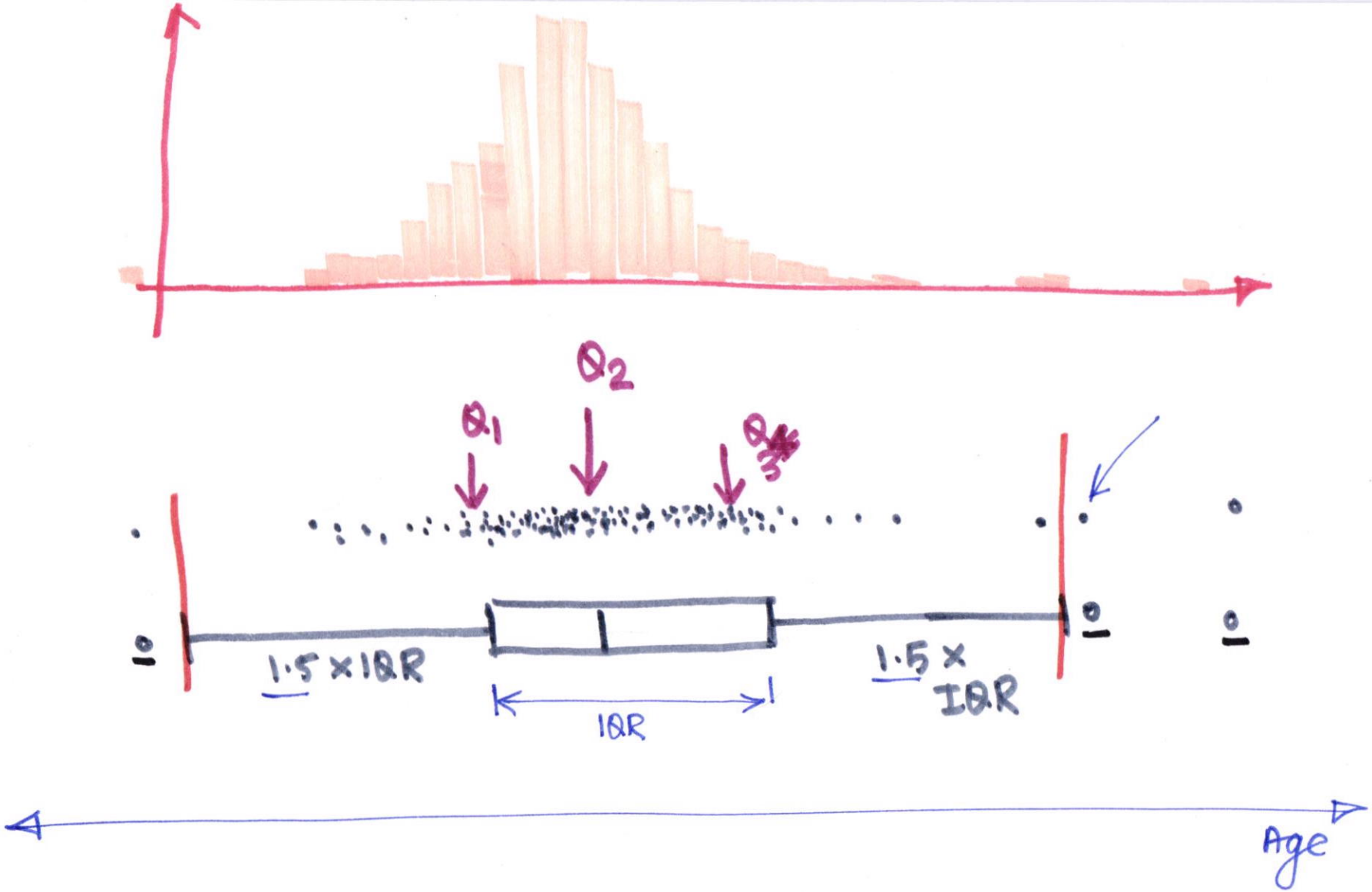
✓ median = 6 LPA.

'skew'



There is outlier




Q2
Q1
Q3
1·5 × IQR
IQR
1·5 × IQR
Age

Measures of Spread

Positional

- Range
- Quantiles & Quartiles
- Interquartile Range

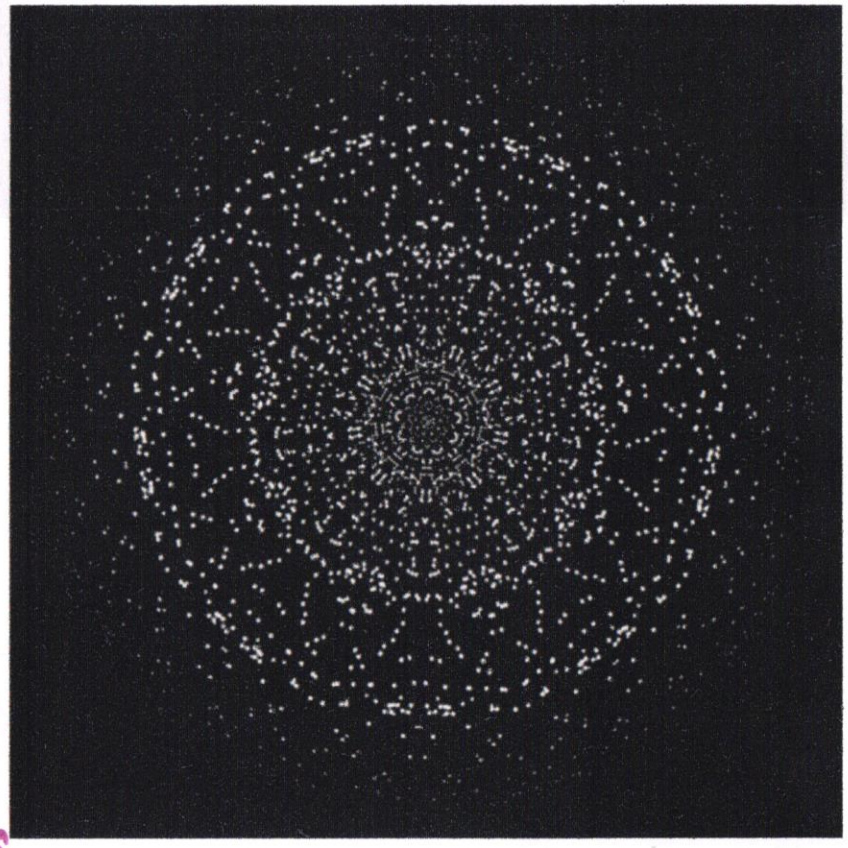
— Relm of Median ≡ feeling, that we might have outliers

- Standard Deviation
- Variance

— centre mean

reality

but, affected by outliers



INSAID

Range

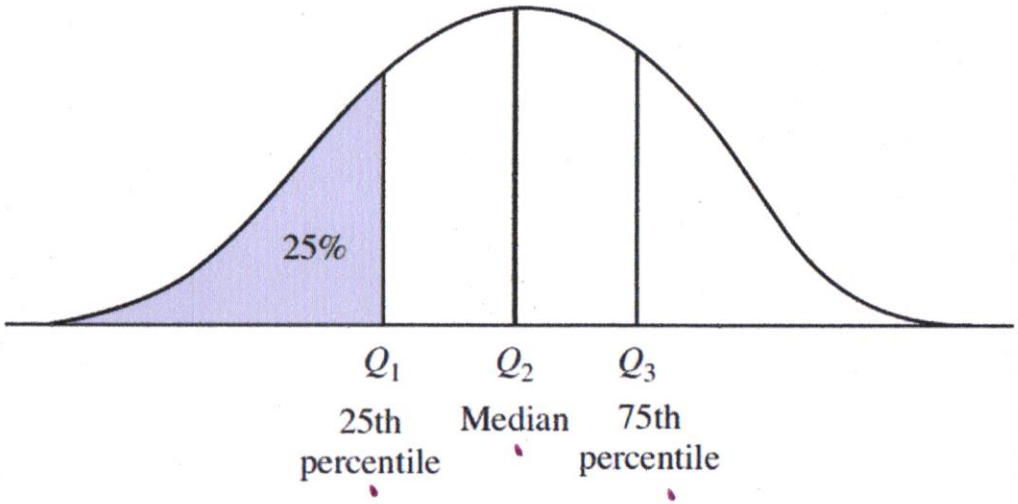
- Difference between the largest and smallest value.
- **Example:** Plant manager are asked to record the plant production output during weekdays.

- Range = Max – Min
- Range = 35 – 15
- Range = 20

INSAID

Quantiles & Quartiles

- **Quantiles**: Data points that split data distribution into equal-size consecutive sets.
- **Quartiles**: Each part split by quantiles.
- **Example:** 4 - Quantile distribution

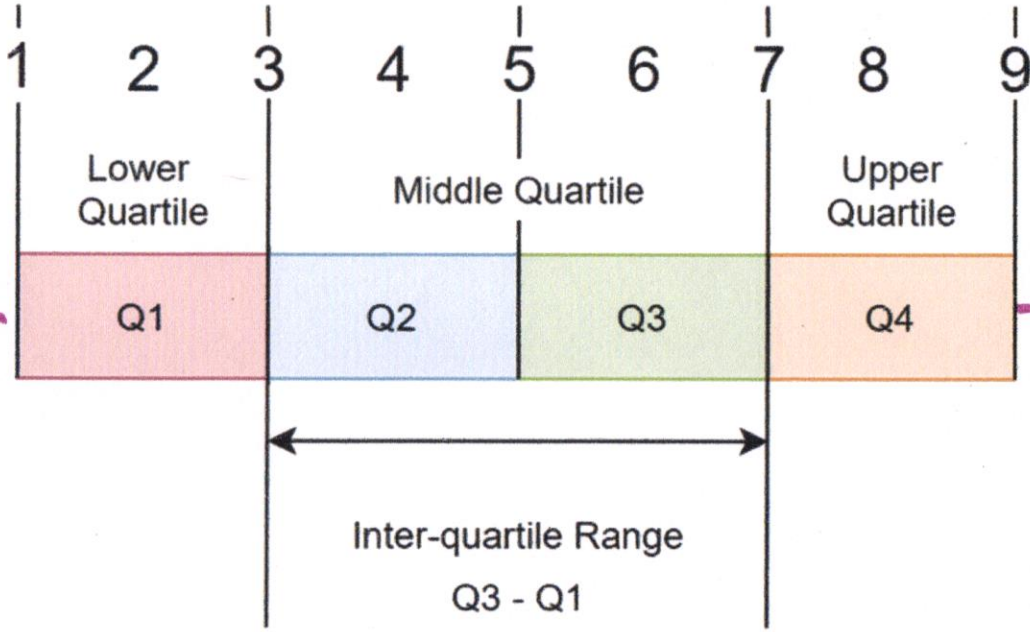


(4 – Quantile Distribution)

INSAID

Interquartile Range

- The distance between the first and third quartiles.
- Gives the range covered by the middle half of the data.

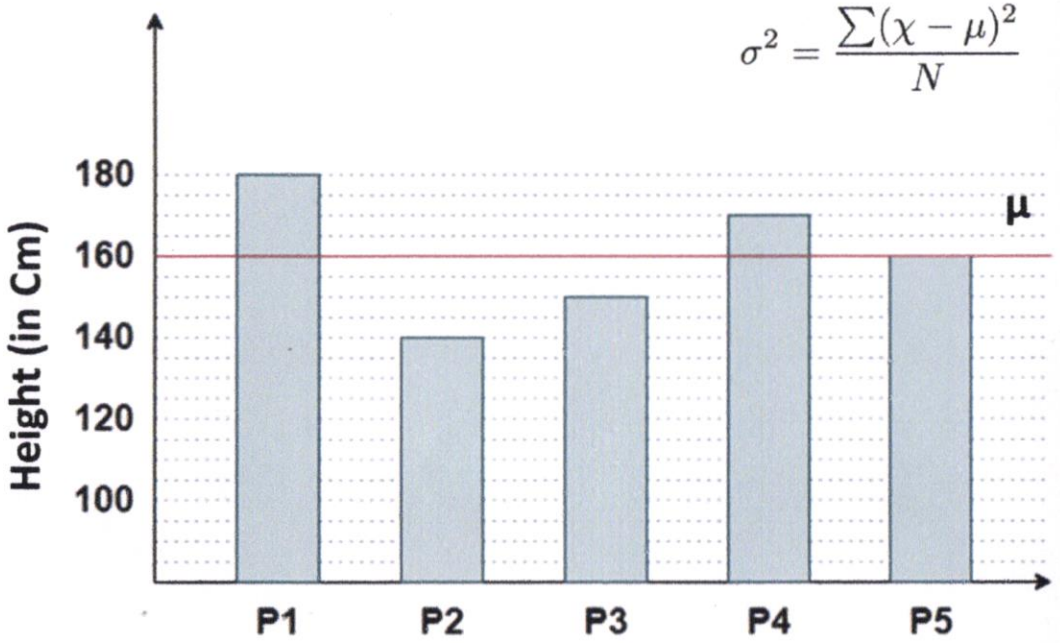


INSAID

Variance

- It is the average of the squared distances from the mean.
- **Pros:** Adds weight to outliers, the numbers that are far from the mean.
- **Cons:** It is not easily interpreted.
- **Example:** Height of students.

$$\sigma^2 = \frac{\sum(x-\mu)^2}{N}$$



Solution:

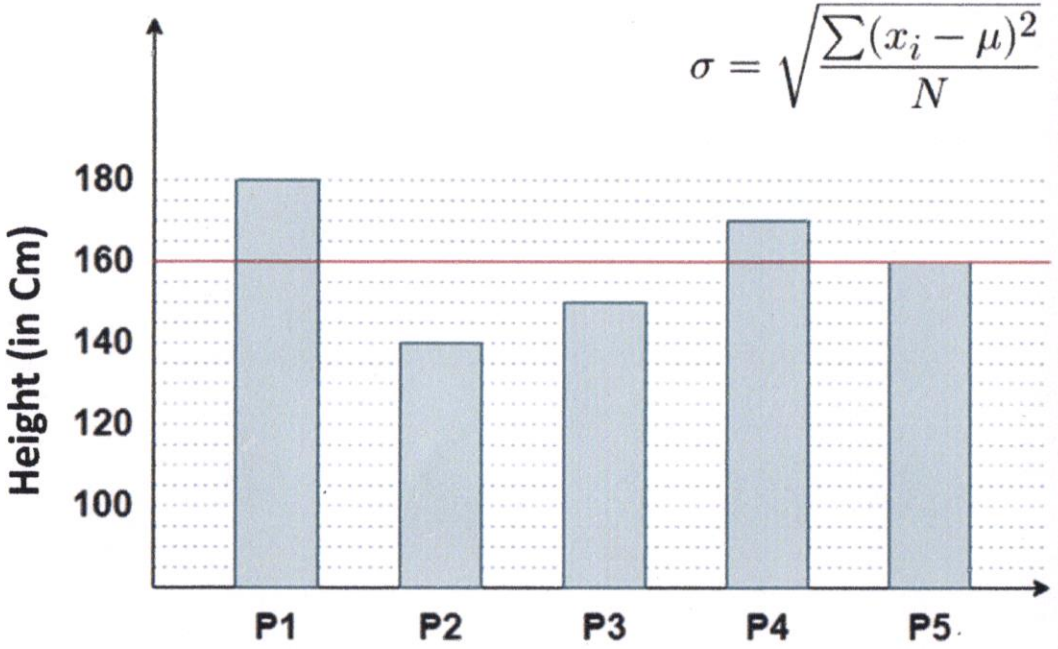
- $\sigma^2 = (400 + 400 + 100 + 100 + 0) \div 5$
- $\sigma^2 = 200$

INSAID

Standard Deviation

- It is the square root of variance.
- **Pros:** Easy to interpret.
- **Cons:** Highly affected by the outliers.
- **Example:** Height of students.

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



Solution:

- $\sigma = \sqrt{[(400 + 400 + 100 + 100 + 0) \div 5]}$
- $\sigma = 14.1$

INSAID

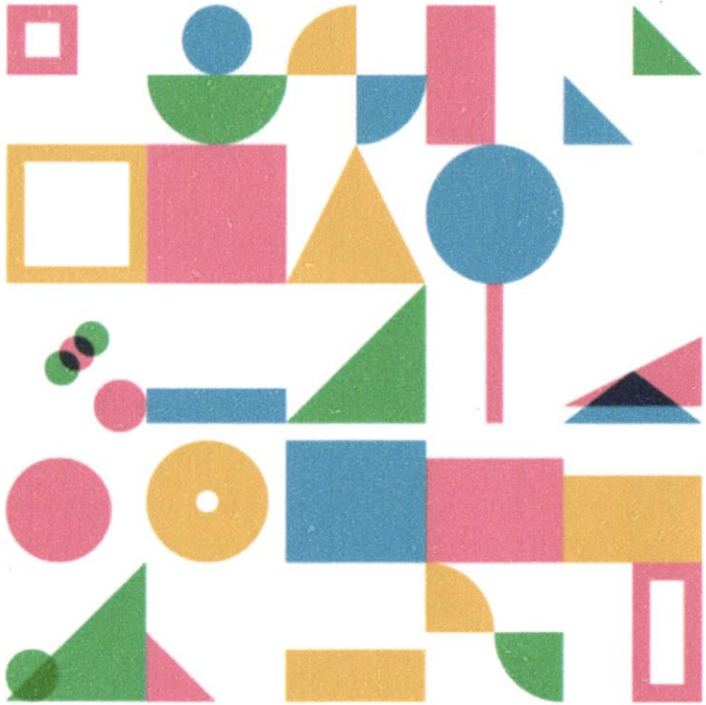
Measures of Spread Use Cases

- Gives us an idea of how well the mean represents the data.
- It help in figuring out the outliers for better results.
- Standardized the data and interpret with respect to mean.
- Monitor control systems for not normal behavior.
- & Many More...

INSAID

Measures of Shape

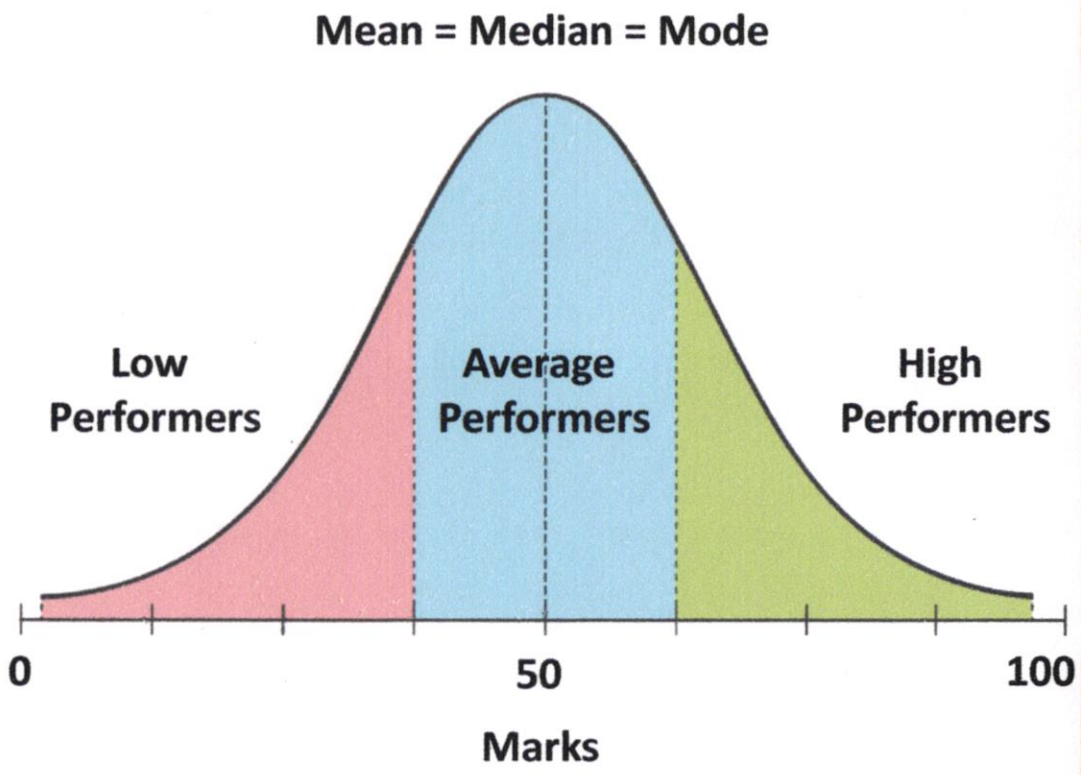
check it out Normal distribution



- Symmetricity
- Skewness
- Kurtosis

Symmetricity

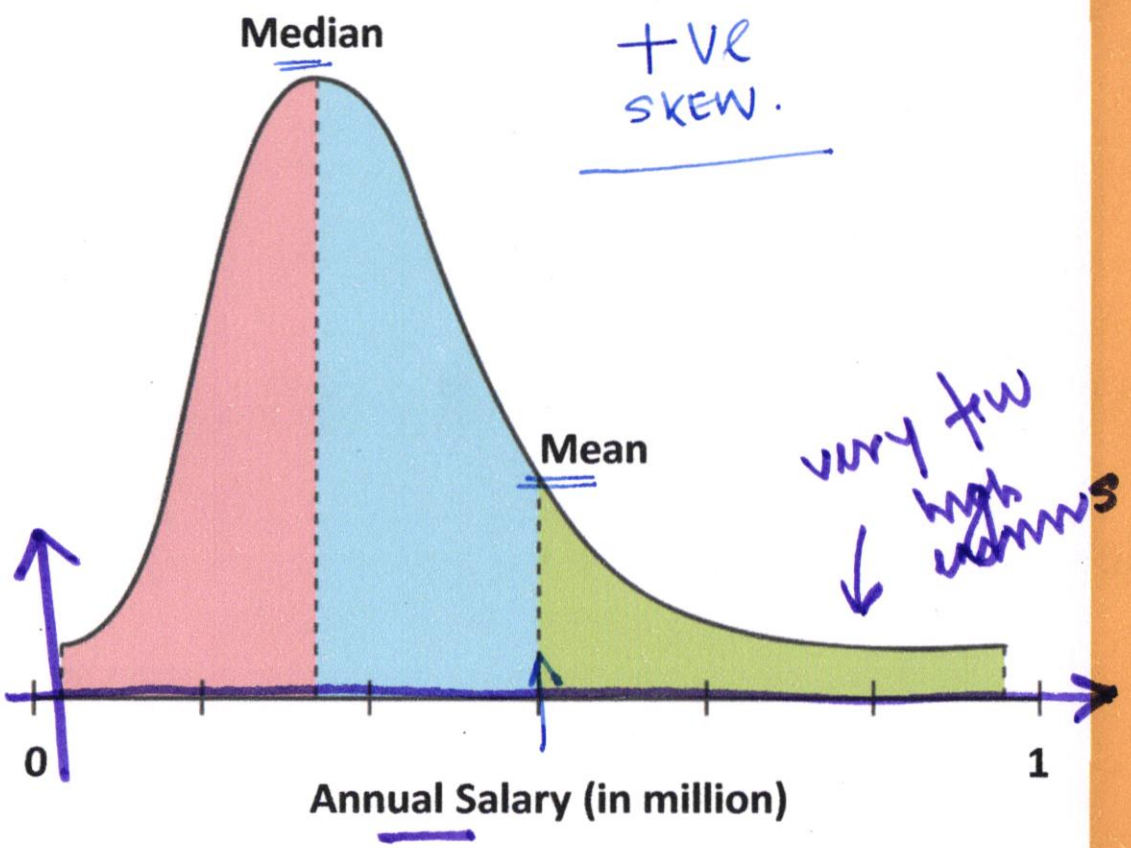
- Also known as normal distribution.
- The mean, median and mode lie at same point.
- **Example:** Distribution of marks received by 100 students in a math's test.



INSAID

Skewness - Positive

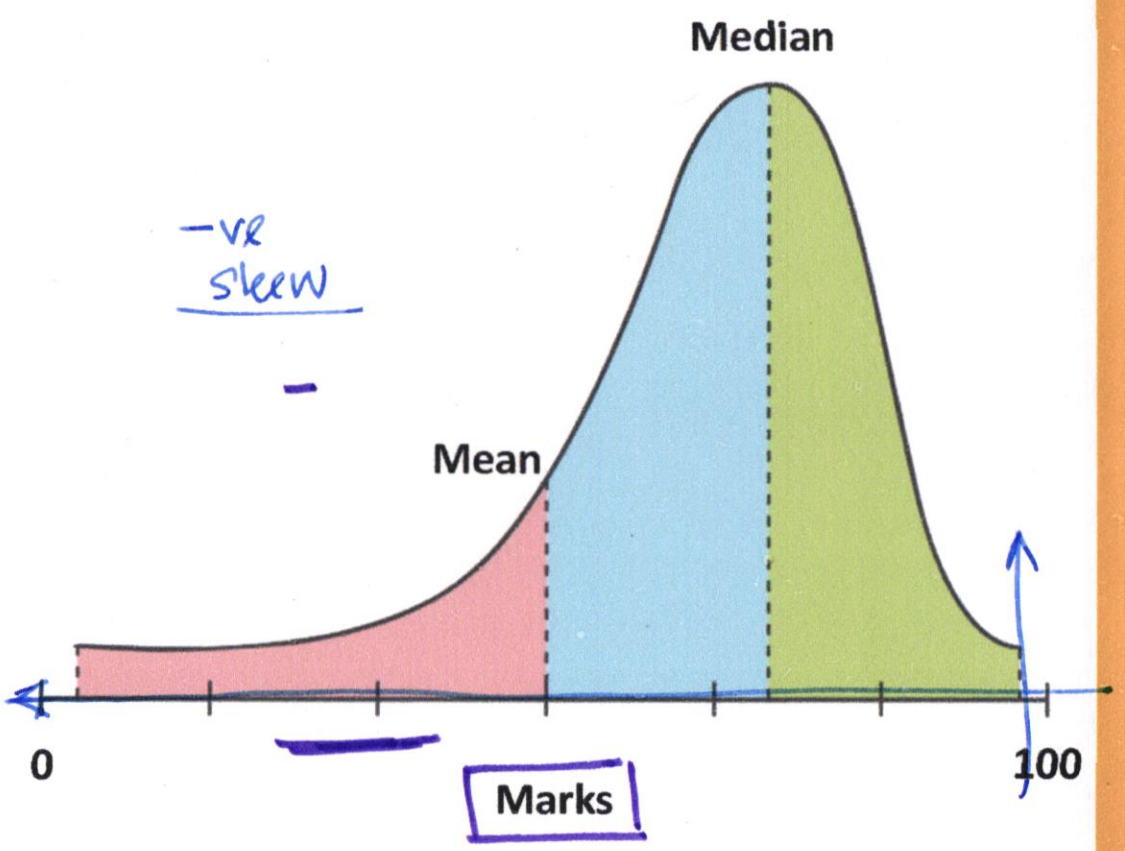
- Also known as right skewed distribution.
- Mean > Median
- **Example:** Income Distribution in a sample population



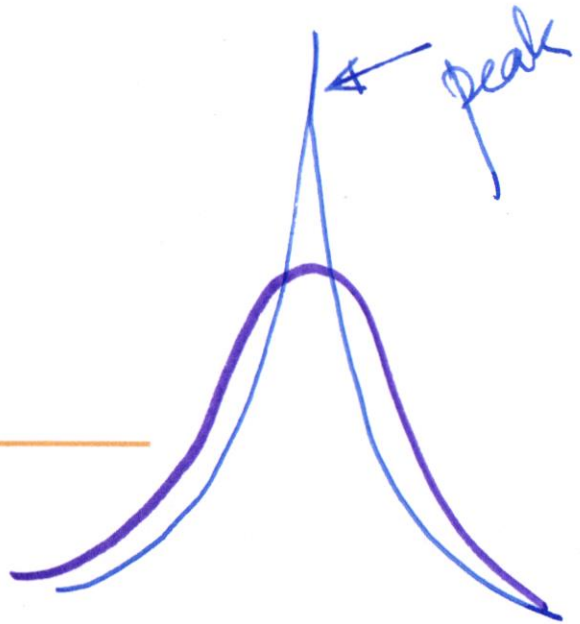
INSAID

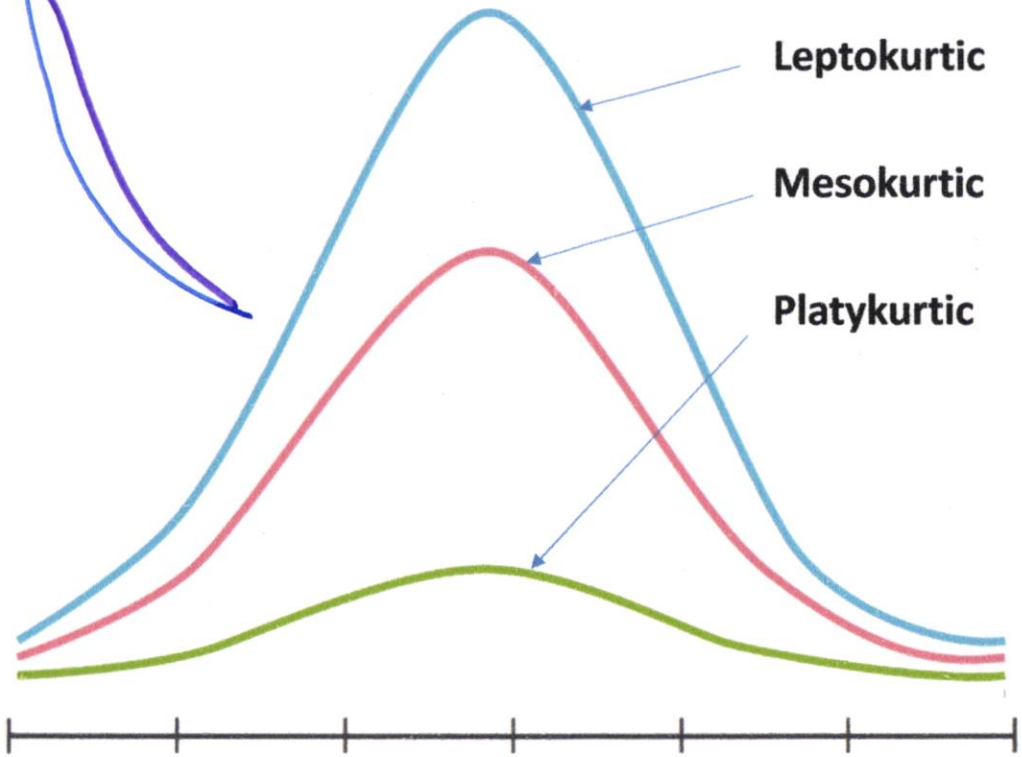
Skewness - Negative

- Also known as left skewed distribution.
- Mean < Median
- **Example:** Distribution of scores on a very easy test.



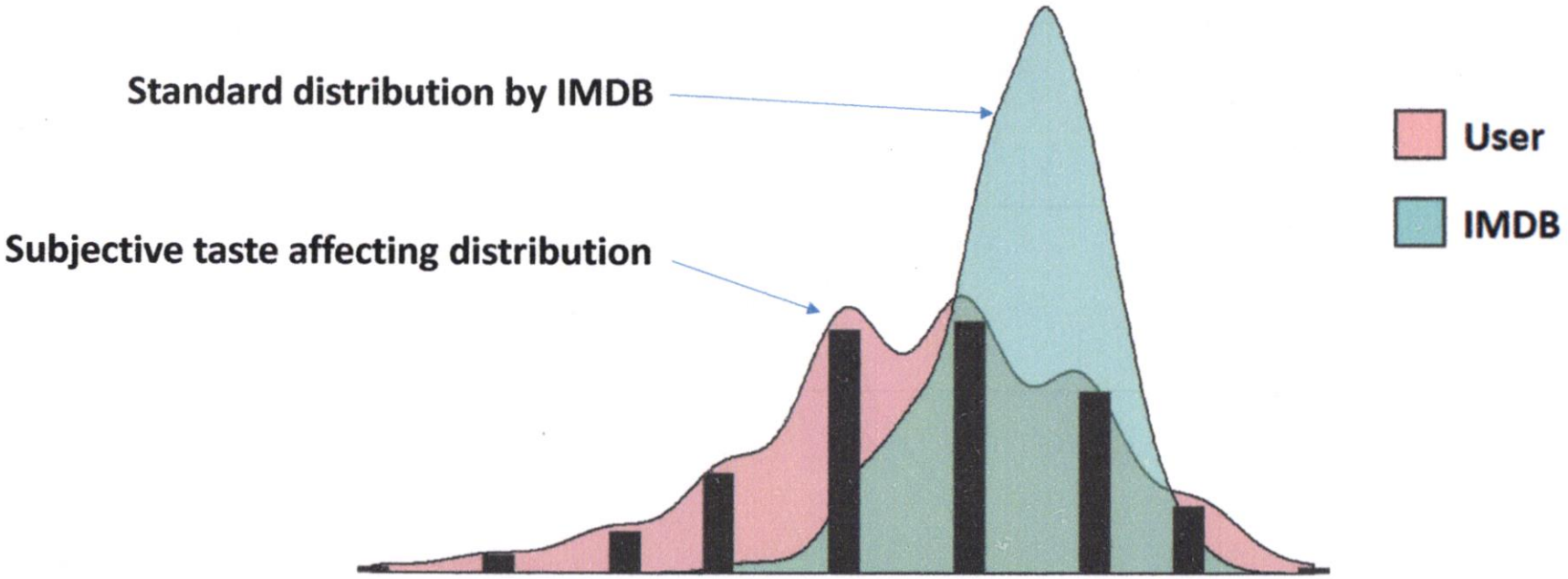
Kurtosis





- A statistical measure.
- Defines heaviness of distribution tails.
- Helps in differing the tails with normal distribution.

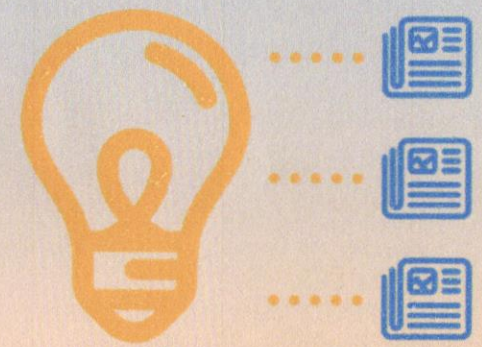
INSAID



Kurtosis Example

- Comparison of 400+ movie ratings by IMDB and a single user.

INSAID



Measures of Shape Use Cases

- It help in figuring out the outliers (tails) for better results.
- Study effect of outliers on the overall data.

Assumption - Data is Normal

check, correction

INSAID

Plots

- Histograms
- Bar Charts
- Scatter Plots
- Pie Charts
- Box Plots
- & Many More...

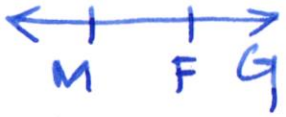
Types of variables

categorical [nominal, ordinal]

continuous [interval, ratio]

axis will also be categorical

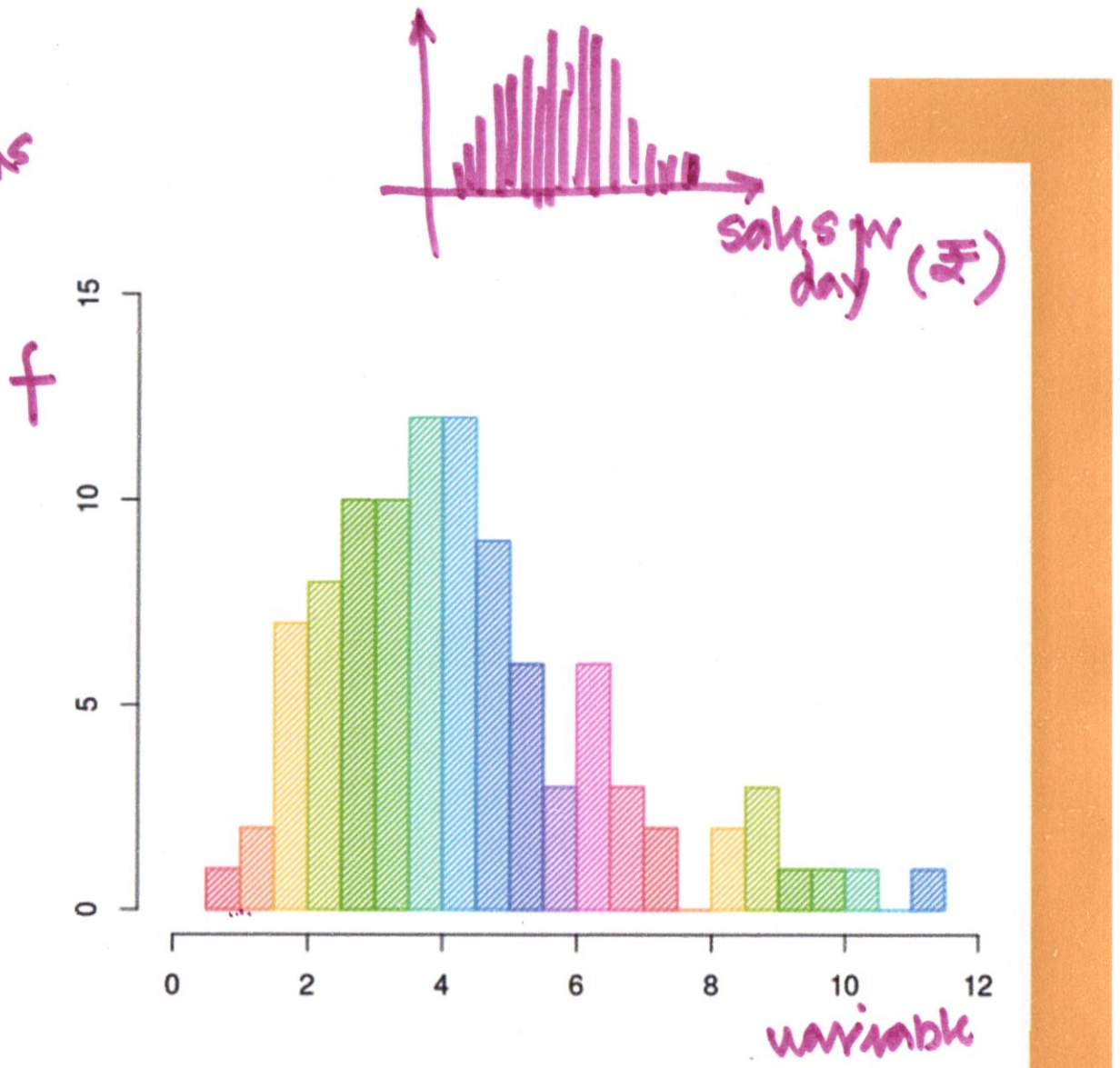
axis will also be continuous



INSAID

Histogram

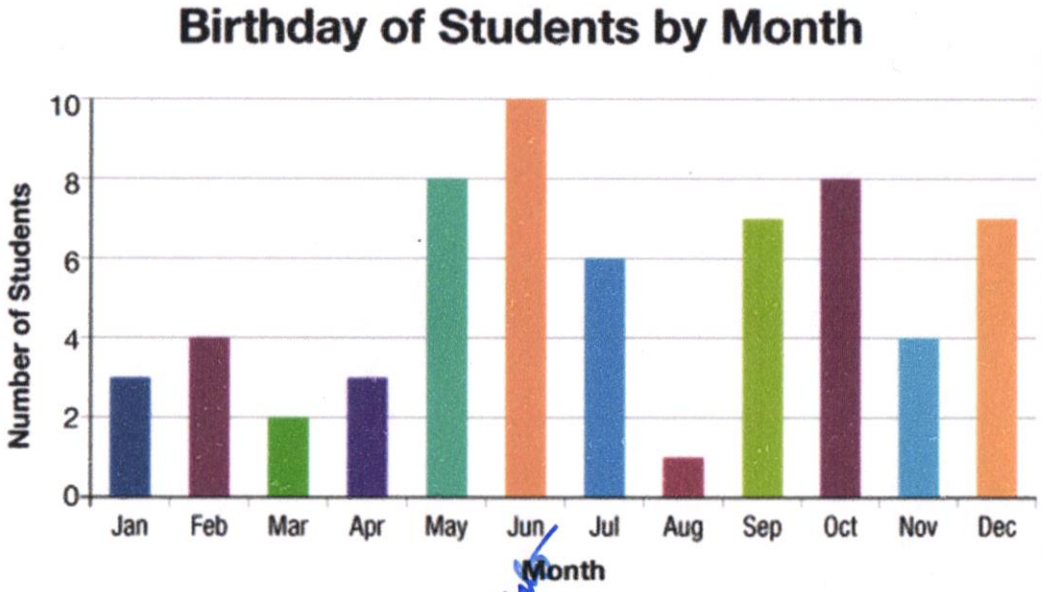
- A graphical display of data using bars of different heights.
- Each bar groups numbers into ranges.
- Taller bars show that more data falls in that range.
- Displays the shape and spread of continuous sample data.

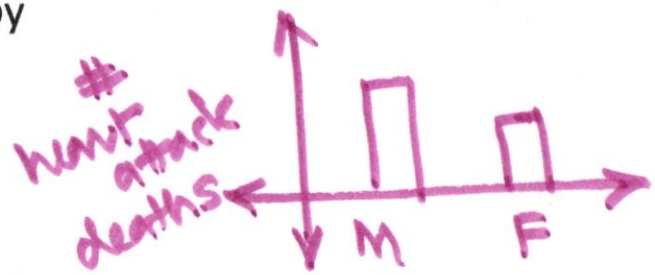


INSAID

Bar Chart

- Used to show how something changes over time or to compare items.
- Generally, X-axis contains categories and Y-axis contains continuous values.
- **Example:** Birthday of Students by Month





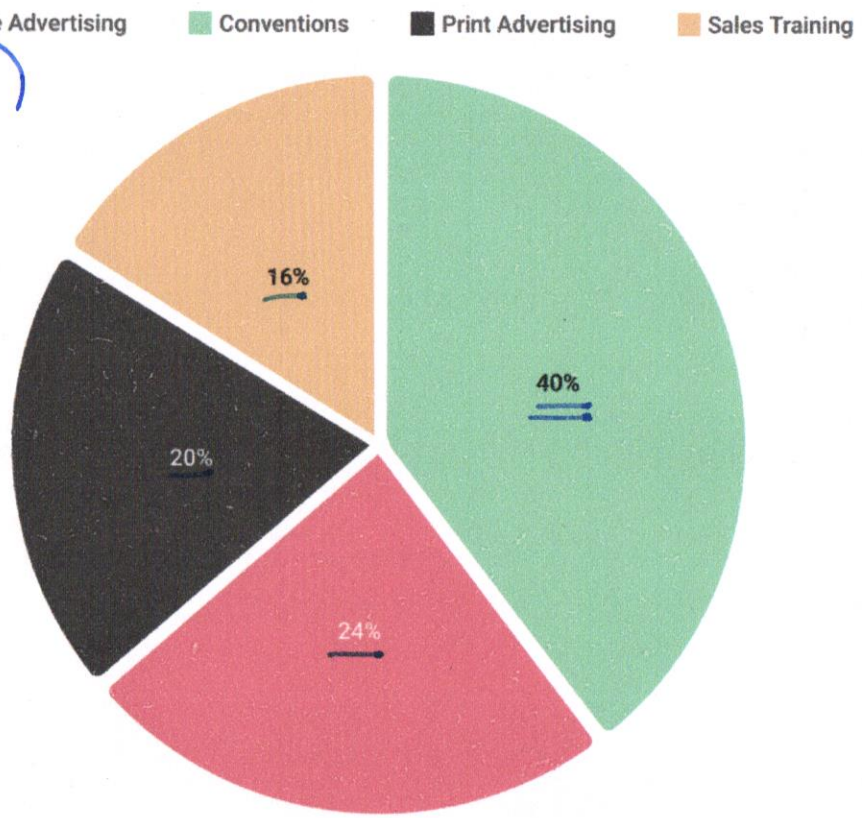
INSAID

Pie Chart

categorical data

frequency of category (%)

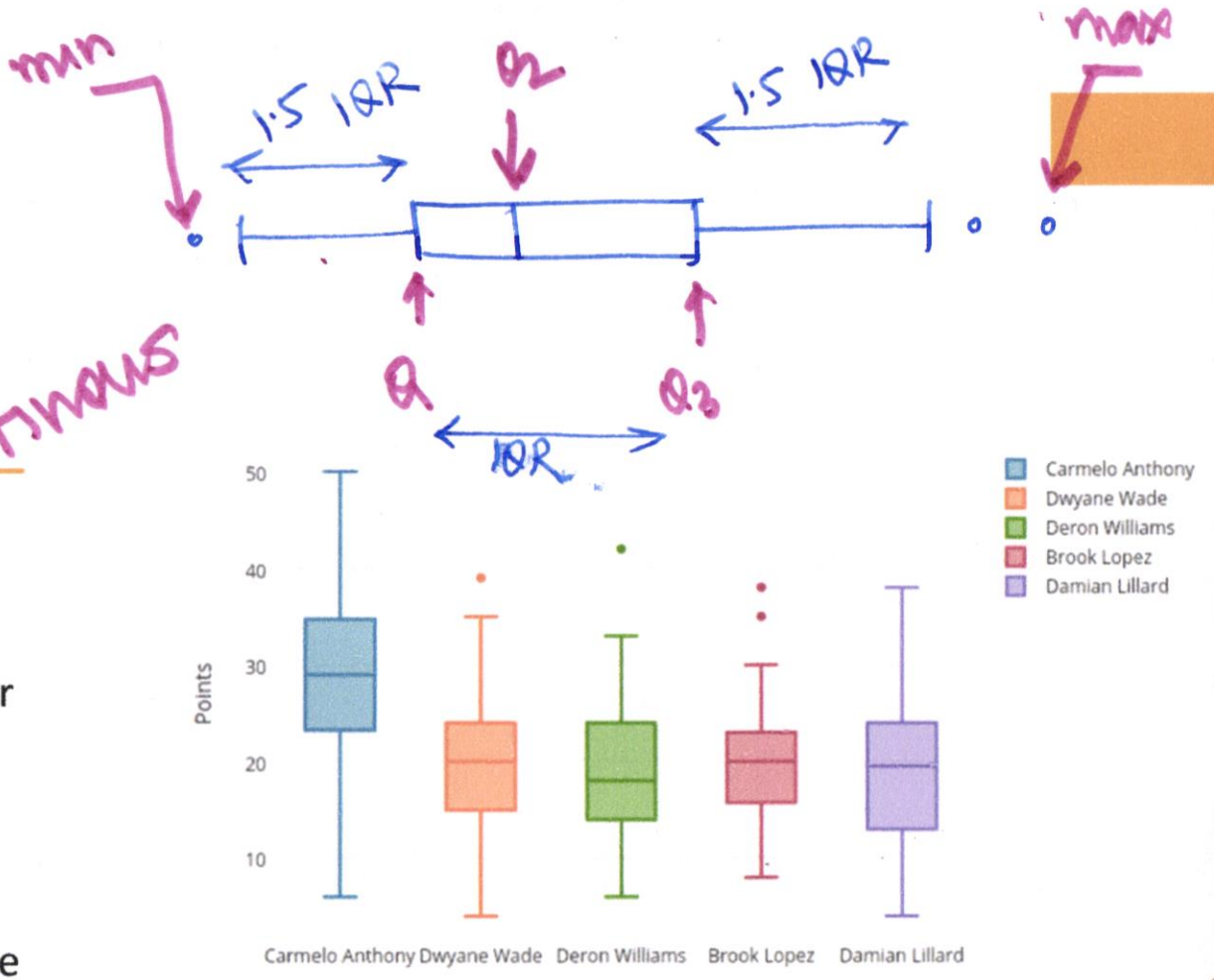
- A circular statistical graph.
- It is divided into slices to illustrate numerical proportion.
- The arc length of each slice represents proportion to the quantity.
- **Example:** Marketing Budget



INSAID

Box Plot

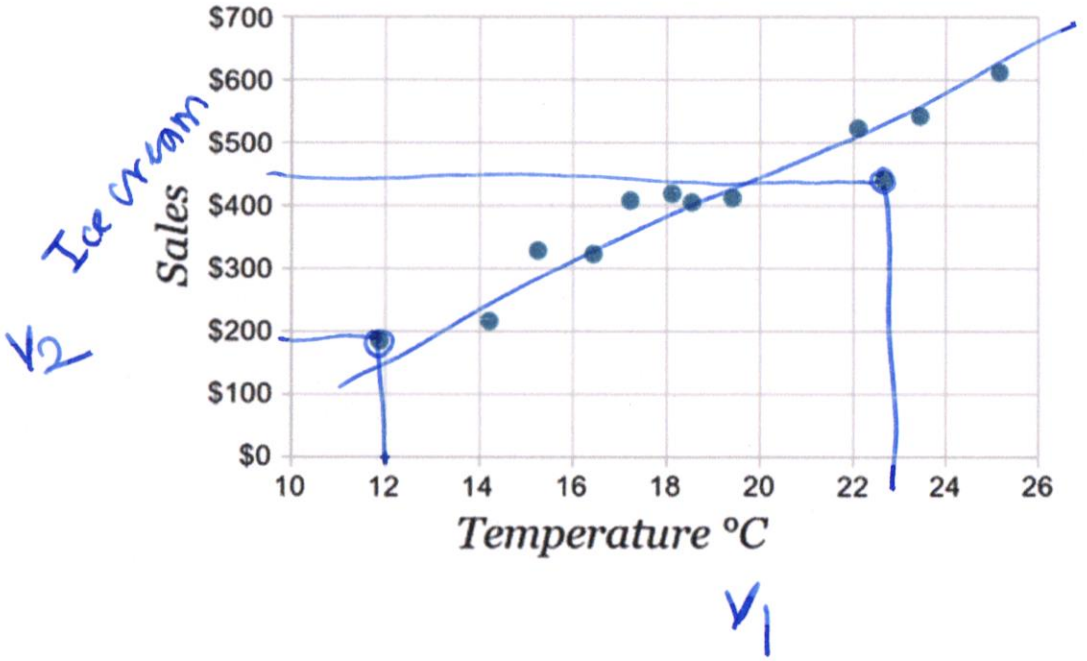
- A method for graphically depicting groups of numerical data through their quartiles.
- Gives five point summary, Minmum, Lower quartile, Median, Upper quartile and Maximum.
- **Example:** Marketing Budget



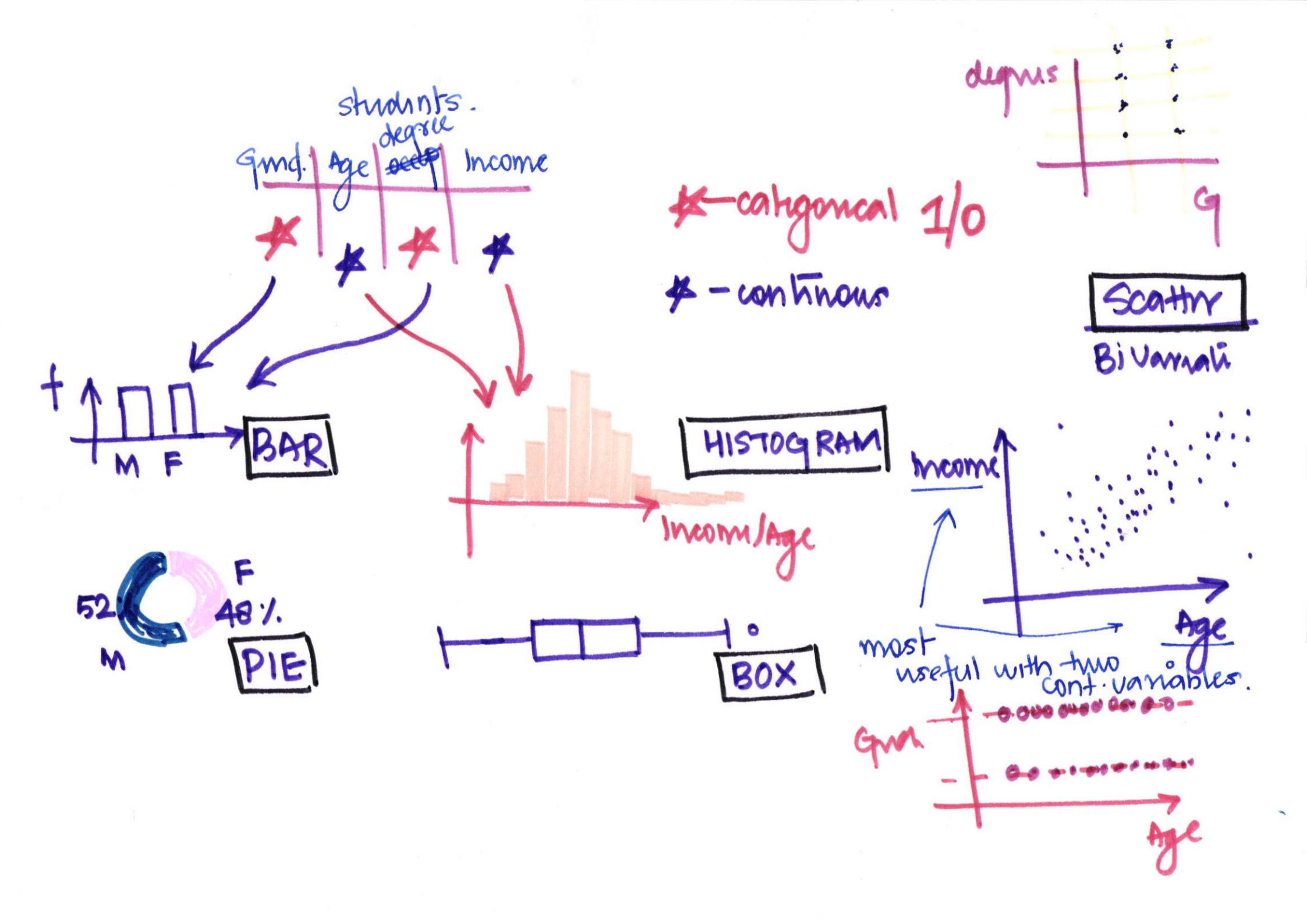
Scatter Plot

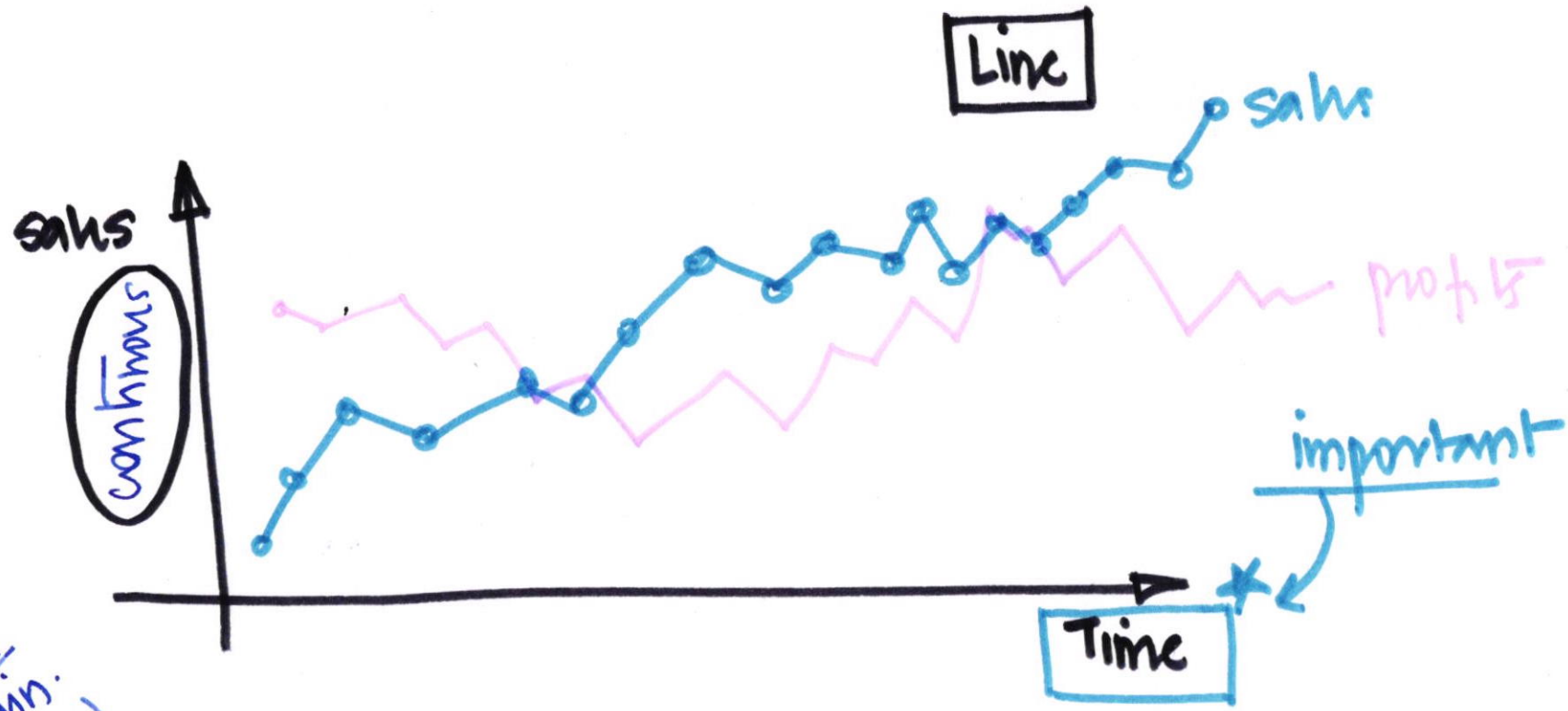
Bi variate Plot

- This graphd uses dots to represent values for two different numeric variables.
- Generally, they are used to observe relationships between variables.
- **Example:** Ice-cream sales



INSAID


students.
Gend.
Age
degree
Income
degree
G
categorical 1/0
continuous
Scatter
Bivariate
BAR
M F
HISTOGRAM
Income/Age
Income
Age
most useful with two cont. variables.
52%
48%
F
M
PIE
BOX
Gen
Age



~~dot fi~~
amount rain.
(mm)