29/10/2025

Introduction

* It is used to check the quality of the data. (Descriptive Statistics)

* To make a statement or a conclusion we use Inferential Statistics.

What is Statistics:-

It is the science of collecting, Organizing and Analyzing data. (For better decision making)

What is Data?

Data means facts or Piece of information that also can be measured.

Ex: The IQ of a class students

{ 98, 97, 68, 57, 110 }

We can measure Avg, Min, Max

Descriptive Statistics:-

It consists of organizing & Summerizing Data.

Inferential Statistics:-

* Techniques where we used the Data that we have measured to form Conclusion.

* To make a statement / Conclusion on a Descriptive statistics we use inferential statistics.

---

1) Are the avg mark of the Java students is same as python class students in Besant?

→ Inferential Statistics

2) What is the average marks of SQL students?

→ Descriptive Statistics

Population (N) and Sample (n) :-

* The Entire group of the data we call it as Population (N).

Eg: All People in India

* A Subset of a population we call it as a Sample.

Eg: One lakh people from diffeient region of India.

* Populations are Larger than Sample.

* Sample should be representative of the population.

* Samples allow for easier, faster & Less costly data collection. ☺

Types of Sampling Techniques:-

1) Simple Random Sampling

Every member of a population has an equal chance of being selected for our sample.

Eg:- Avg mileage of any 120 cc bikes.
- Ratio of maried people in banglore

2) Stratified Sampling:-

Where the population is split into non overlaping groups.

Eg: 1) The person is Eligible for voting or not.

3) Systematic Sampling:-

From the population, Every nth sample we will collect.

Eg: 1) While doing surveying in the mall on modernization, collecting information of every fifth person who is coming out from the Mall.

4) Convenience Sampling:-

The sample is collected based on our convenience from the particular domain experts.

× ☺ ×

30/10/2025

Note:- Sampling techniques Selection always Depends on Problem Statement.

Variable:

A Vaiable is a property that can take an any value.

Two kinds of Vaiables:-

1) Quantitative (Numerical) Vaiables
2) Qualitative (Categorical) Vaiables

① Quantative Vaiable:-

A Value can be measured and we can perform mathematical operation like (A, S, M, D)

Ex: mpg, weight, height ....

Numerical → Continuous (float) → ∞
0-1
→ Discrete (int) → 1

Discrete → Numerical: age (10, 20, 25)
→ Categorical: gear (4, 5, 6)

② Qualitative / Categorical Variable:-

Non-Measurable Data and Based on some characteristics we can derive Categorical Variables.

Ex: gender — male, female, other

DT — IT, Non-IT

Blood — A+, B+, A-, B-, O+, O-

Variable Measurement Scales:-

4 Types of Measured Variables

1) Nominal Data
2) Ordinal Data
3) Interval Data
4) Ratio Data

1) Nominal Data:-

The categorical Data which are having different classes.

2) Ordinal Data:-

Orders of the Data matters but values Doesn't.

Marks [50, 40, 100, 91, 35] ③ ④ ① ② ⑤ — ① - 100

Marks [40, 80, 75, 60, 90] ⑤ ② ③ ④ ① — ① - 90

} Rank is imp, Value is not

3> Interval Data:-

Order Matters and Value also matters but natural zero is not present

4> Ratio Data:-

The ratio data can measured, ordered, equi distant and have meaningful zero.

Eg: Sal, height, weight, ...

Descriptive Statistics:-

I Measure of Central tendency:-

1> Mean
2> Median
3> Mode

1> Mean:

$$\text{Population Mean } (\mu) = \frac{\sum_{i=1}^{N} x_i}{N}$$

$$\text{Sample Mean } (\bar{x}) = \frac{\sum_{i=1}^{n} x_i}{n}$$

2) Median

- Sort the value Either asc↑ or des↓ order.
- Choose the Mid value.
- If you get 2 mid values take avg of those 2 values.

[1, 2, 2, 3, 4, 5]

↓

Mean = 2.8

Median = 2.5

[1, 2, 2, 3, 4, 5, 100]

↓

Mean = 16.7

Median = 3

100 = outlier

↓

out of Boundary

- Mean will be affected by outliers whereas median won't affect by outliers.
- Used for Null value Imputation

3) Mode

Most Repetative value

[1, 2, 3, 1, 2, 2, 3, 4, 5]

Mode = 2

∴

31/10/2025

	Person 1	Person 2
M	7 am	8
T	7.30 am	11
W	8 am	9
T	7.15 am	7
F	7.30 am	10
S	?	?

Manager ask Team lead at what time the persons will come?

Person 1 → 7-8 — 7.15 (15min)
7.30 (30min)
7.45 (15min)

Person 2 → 9-10 — 7 (2hr)
11 (1hr)

- If variance is high (8-11) Perdiction accuracy Low
- I variance is Low (7-8) Prediction accuracy high.

(II) Measure of Dispersion:-

1) Variance
2) Standard Deviation
3) Range

1) Variance:-

- Population Variance (σ^2)
- Sample Variance (s^2)

$$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N}$$

$$S^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$

$n-1$ = degree of freedom

$\_ + \_ + \_ = 57 \rightarrow n=3$

$(20 + 30) + (7) = 57$

$n-1$

$3-1=2$

no freedom

Degree of freedom

$\{1, 2, 2, 3, 4, 5\}$ km

Population Variance $\sigma^2 = \sum_{i=1}^{6} \frac{(x(i) - \mu)^2}{6}$

$$\mu = \frac{1+2+2+3+4+5}{6} = 2.8$$

$$\sigma^2 = \frac{(1-2.8)^2 + (2-2.8)^2 + (2-2.8)^2 + (3-2.8)^2 + (4-2.8)^2 + (5-2.8)^2}{6}$$

$$\sigma^2 = \frac{10.83}{6} =$$

$\sigma^2 = 1.8 \text{ km}^2$ — Variance

$\sigma = 1.34$ km — Standard Deviation

Standard Deviation:-

$$SD = \sqrt{\text{Variance}}$$

Population SD $\sigma = \sqrt{\sigma^2}$

$\sigma = \sqrt{1.8}$

$\sigma = 1.34$

Sample Variance (S^2)

$$S^2 = \sum_{i=1}^{6} \frac{(x_i - \bar{x})}{n-1} \qquad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$S^2 = \frac{(1-2.8)^2 + (2-2.8)^2 + (2-2.8)^2 + (3-2.8)^2 + (4-2.8)^2 + (5-2.8)^2}{5}$$

$\bar{x} = 2.8$

$$S^2 = 2.16 \text{ km}^2$$

Standard Deviation of Sample Variance

$$S = \sqrt{S^2}$$

$$S = \sqrt{2.16}$$

$$S = 1.46$$

3> **Range:-**

Range = Maximum − Minimum

= 5 − 1

Range = 4

×

Percentile & Quantile:-

Percentile → It is a value below which a certain percent of observation will lie.

Eg: {1, 1, 2, 3, 4, 5, 5, 6, 7, 7, 8}

How much % of data will come below 6?

$$\text{Percentile rank of } x = \frac{\#\text{ of value below } x}{N} \times 100$$

$$= \frac{7}{11} \times 100$$

$$= 63\%$$

63% of observation data is < 6

Quantile:-

It will helps to find the value which is present in the given Percentile rank.

Eg: {1, 1, 2, 3, 4, 5, 5, 6, 7, 7, 8}

Which value is present at 25%?

$$\text{Value} = \frac{\text{Percentile}}{100} \times n + 1$$

$$= \frac{25}{100} \times 2$$

Value = 3rd index

$$90\% = \frac{90}{100} \times 12$$

$$= 10.8 - \text{index}$$

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

$$\therefore = 7$$

Five Number Summary :-

1) Minimum
2) First Quartile (Q_1) 25%
3) Median (Q_2) 50%
4) Third Quartile (Q_3) 75%
5) Maximum

Note :- Choose these 5 numbers after removing the Outlier from the data by finding boundary values.

[Lower fence Upper Fence]

$$LF = Q_1 - 1.5(IQR)$$

$$UP = Q_3 + 1.5(IQR)$$

IQR (Inter Quartile Range)

$$IQR = Q_3 - Q_1$$

{ 1, 1, 2, 3, 4, 4, 4, 5, 5, 6, 7, 7, 8, 8, 9, 28, 36 } = 17

Quartile (Q1) $= \frac{25}{100} \times (17+1)$

$Q_1 = 4.5 \rightarrow 4 \rightarrow 3$

$(Q_3) = \frac{75}{100} \times 18$

$Q_3 = 13.5 \rightarrow 13 \rightarrow 8$

$IQR = Q_3 - Q_1 = 8-3$

IQR = 5

$LF = 3 - 1.5(5)$

LF = -4.5

$UF = 8 - 1.5(5)$

UF = 15.5

∴ [-4.5 (Lower Fence) ⟷ (Upper Fence) 15.5]

Anything out of the range of -4.5 & 15.5 is a outlier.

∴ In { 1, 1, 2, 3, 4, 4, 5, 5, 6, 7, 7, 8, 8, 9, 28, 36 }

28 & 36 are outliers

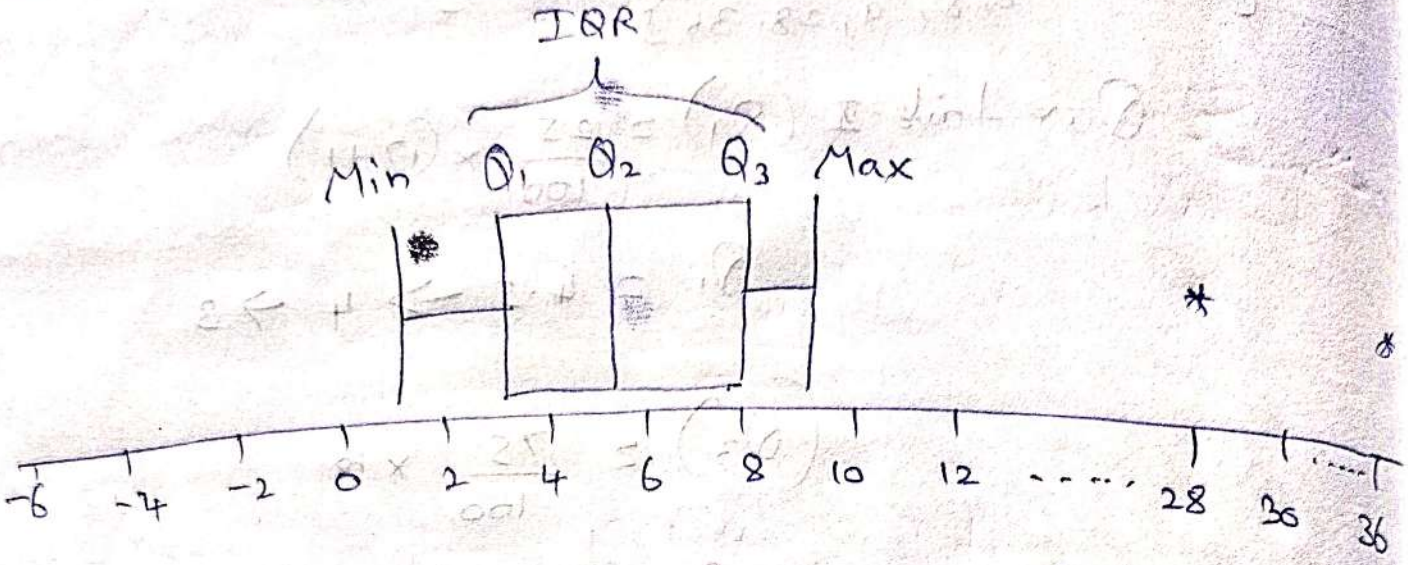
∴ Minimum = 1

$Q_1 = 3$

Median $(Q_2) = 5$

$Q_3 = 8$

Box Plot of Five no Summary:-



This Graph is used to find the Outliers

3/11/2025

Different types of Distribution:-

1) To understand data Patterns
2) To Summerize the data Easily
3) To Calculate Probabilities.
4) To Make Prediction & Decision
5) To choose right statistical test

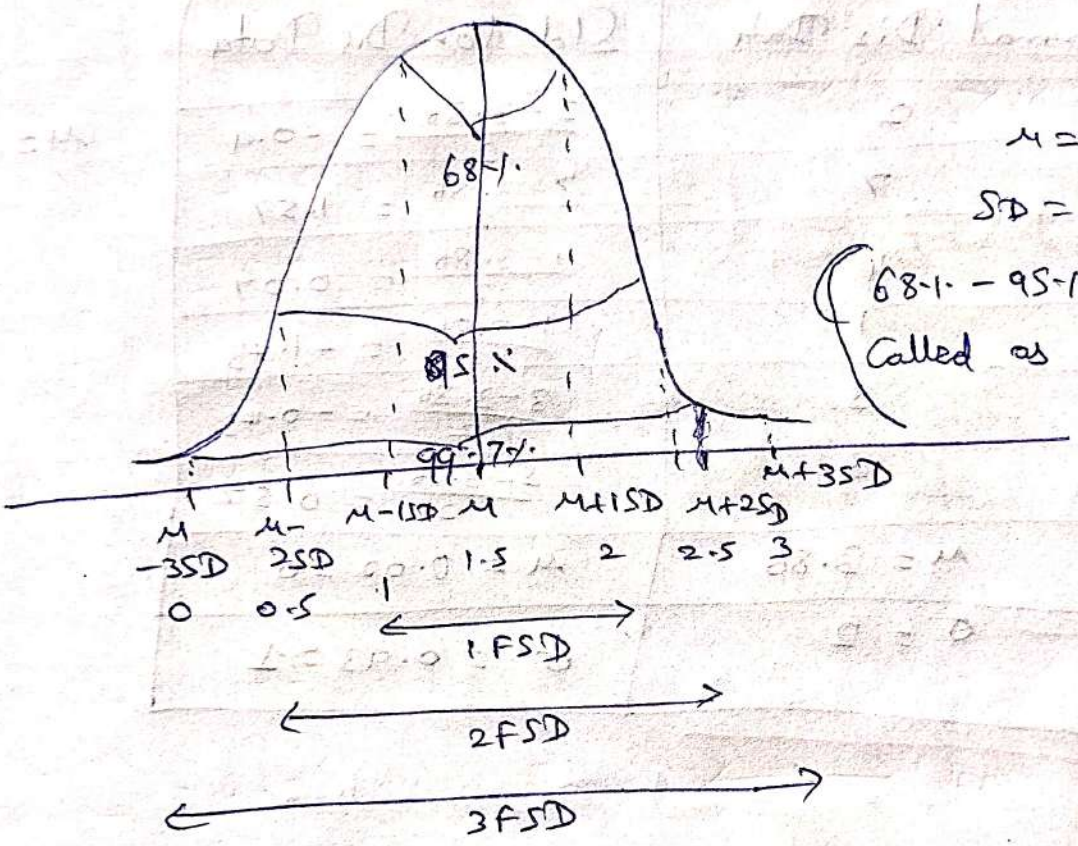
There are 2 Category of Distribution

1) Continuous Distribution (Numerical)
2) Descrete Distribution (Categorical)

1) Normal Distribution
2) Standard Normal Distribution → Continuous Dist (Numerical)
3) Bernoulli Distribution
4) Binomial Distribution → Descrete Dist (Categorical)
5) Poision Distribution

---

1) Normal Distribution:-



μ = 1.5

SD = 0.5

(68% - 95% - 99.7% Called as Confidence Intervals)

Mean = Median = Mode

Emphirical Rules:-

- 68% of data will Present in 1st SD
- 95% of data will Present in 2nd SD
- 99% of data will Present in 3rd SD

---

2) Standard Normal Distribution:-

Always mean $\mu = 0$

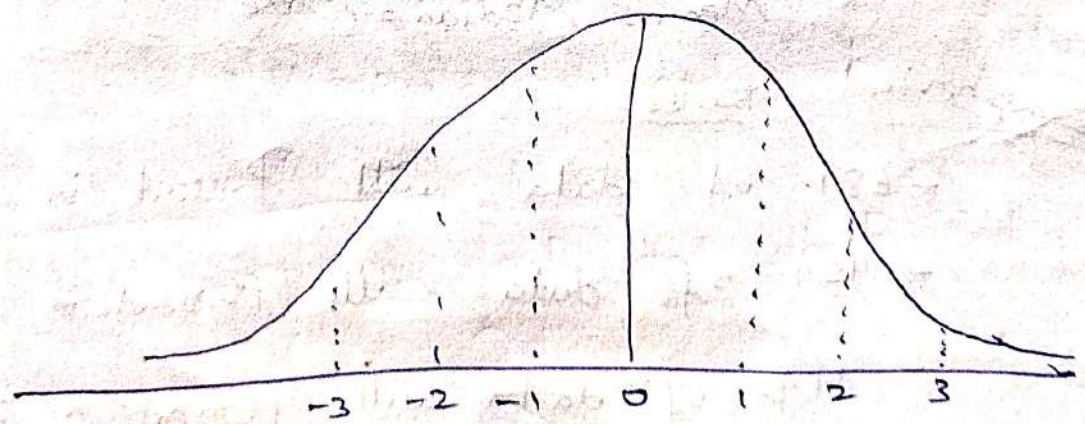
Always SD $= 1$

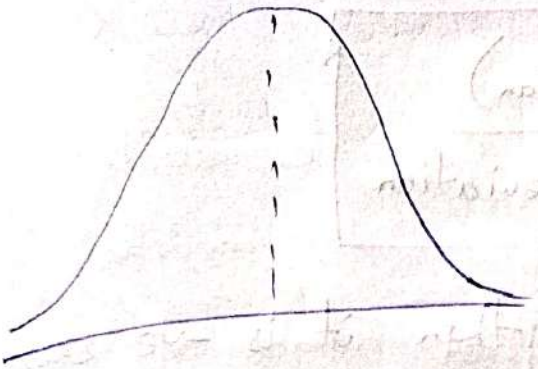
To convert Normal → Standard Normal

$$Z\,Score = \frac{x_i - \mu}{\sigma}$$

Normal Dis Data	Std Nor Dis Data
2	$\frac{2-3.86}{2} = -0.9$
7	$\frac{7-3.86}{2} = 1.57$
4	$\frac{4-3.86}{2} = 0.07$
1	$\frac{1-3.86}{2} = -1.43$
3	$\frac{3-3.86}{2} = -0.43$
5	$\frac{5-3.86}{2} = 0.57$
$\mu = 3.86$	$\mu = 0.02 = 0$
$\sigma = 2$	$\sigma = 0.93 = 1$

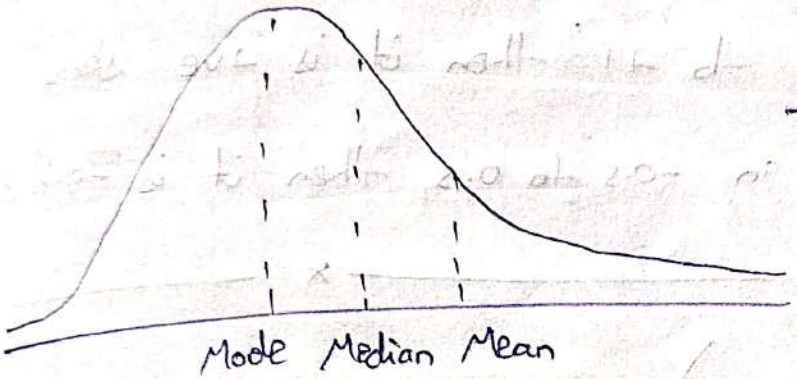
$\mu = 3.86$



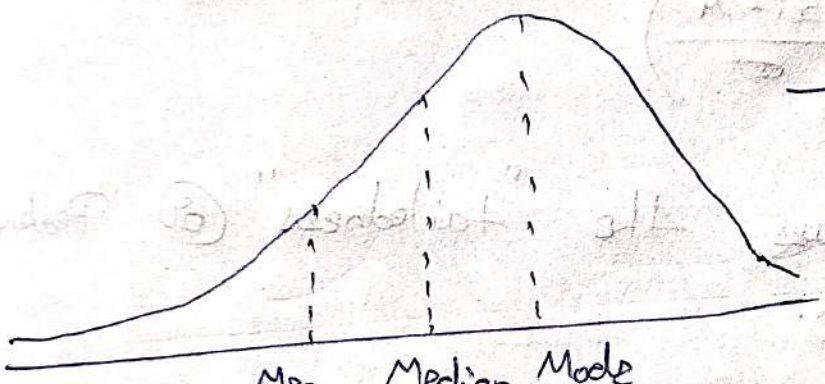


→ Normal Distribution

Mean = Median = Mode



→ * Skewed Distribution
* +ve Skew Data
* Right Skew
* Mean > Median



→ * Skewed Distribution
* -ve Skew
* Left Skew
* Mean < Median

4/01/2025

Positive Skew (Right skew):-

Tail on the right side is longer. Most data are on the Left.

Negative Skew (Left Skew):-

- Tail on the left side is longer.
- Most data are on the right.

Zero Skew (Symmetric):-

The data is evenly distributed around the mean (like a normal distribution).

$$\text{Skewness} = \frac{3(\text{Mean} - \text{Median})}{\text{Standard Deviation}}$$

- If value near to -1 then it is -ve skew
- If value near to +1 then it is +ve skew
- If value lies in -0.5 to 0.5 then it is Zero skew

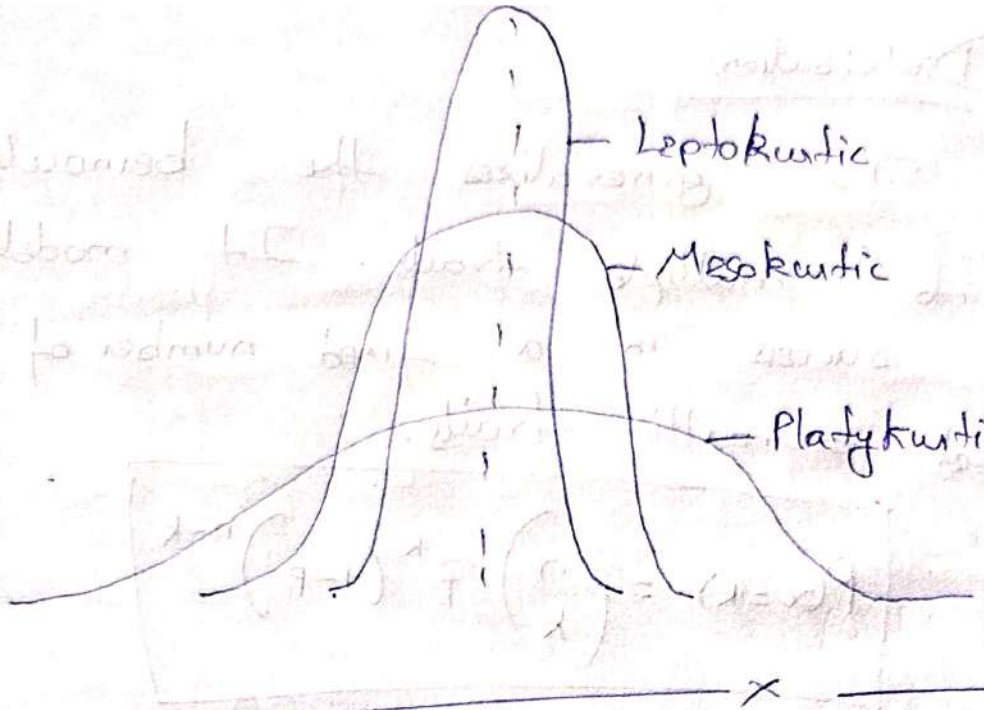
Kurtosis :-

$$K = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \mu}{\sigma} \right)$$

It measures the "tailedness" or "Peakness" of data distribution.

Types of kurtosis

1) Mesokurtic ($K=3$)
   - * Normal Distribution
   - * No outliers
   - * Moderate tail and Peak
2) Lepokurtic ($K > 3$)
   - * Heavy tails and sharp Peak
   - * More Outliers
3) Platy kurtic :- ($K < 3$) :-
   - * Light tails and flat Peak
   - * Fewer outliers



Bernoulli Distribution:-

It is the simplest form of a discrete probability distribution and models a random Experiment with Exactly 2 outcomes.

Success denoted by = P

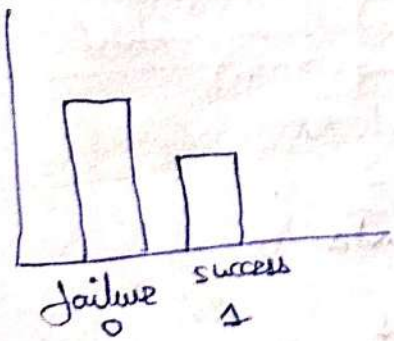
Failure denoted by = 1-P

Total Probability = 1

The rang is (0-1)

$P(T) = \frac{1}{2} = 0.5$

$P(H) = \frac{1}{2} = 0.5$



Binomial Distribution:

It generalises the bernoulli distribution to multiple trails. It models the number of success in a fixed number of independent and identical Bernoulli trails.

$$P(x=k) = \binom{n}{k} p^k (1-p)^{n-k}$$

k = no. of success n = no of trails

P = probability of success

Poison Distribution:-

It is used to model the number of Events that occur in a fixed time interval. (or) Space & occur independently the parameter λ represent the avg no of Event in the interval.

$$P(x=k) = \frac{x^k e^{-\lambda}}{k!}$$

λ - Average no of Events

k - No of arguments

Inferential Statistics

Probability:

It measures likelyhood of an event.

Eg: Dice = {1, 2, 3, 4, 5, 6}

$$P(x) = \frac{\text{No of favourable outcomes}}{\text{total no. of outcomes}}$$

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

$$P(2,4,5) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2} = 0.5$$

Toss the 2 Coins: {HH, HT, TH, TT}

What is the P of getting only 1H = $\frac{2}{4} = \frac{1}{2}$

Both tail = $\frac{1}{4}$

There are two rules in Probability:-

1) Addition Rule (OR)
2) Multiplication Rule (AND)

Addition Rule:-

1) Mutually Exclusive Event
2) Non - Mutual Exclusive Event

5/11/2025

1> Mutual Exclusive Event:-

The different events won't occur at the same time.

Ex: If you toss the coin what is the probability of landing on head or tail.

$$P(A \text{ or } B) = P(A) + P(B)$$

$$P(H \text{ OR } T) = P(H) + P(T)$$

$$= \frac{1}{2} + \frac{1}{2}$$

$$P(H \text{ or } T) = 1$$

2> Non-Mutual Exclusive Event:-

Multiple Events can occur at the same time.

Ex: Picking the cards from deck cards.

What is the Probability of getting Jack (or) H

$$P(J \text{ OR } H) = P(J) + P(H) - P(J \cap H)$$

$$= \frac{4}{52} + \frac{13}{52} - \frac{1}{52}$$

$$= \frac{16}{52}$$

$$P(J \text{ OR } H) = 0.31$$

$\therefore$ 31 %

Multiplication Rule :-

1) Independent Event :-

Here all the values have the same preority after n number trails also.

(or) 1 Event don't depend on another event.

Eg: 1st toss of Coin

$P(H) = \frac{1}{2}$ ← nth toss of coin

7th toss of Coin $P(H) = \frac{1}{2}$

$P(H) = \frac{1}{2}$

Eg: What is the probability of Dice rolling and getting a 5 and then 4 ?

$P(A \text{ and } B) = P(A) * P(B)$

$P(5 \text{ and } 4) = P(5) \times P(4)$

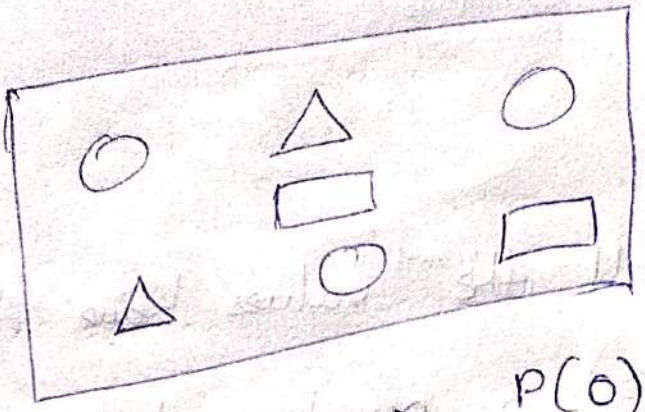
$= \frac{1}{6} \times \frac{1}{6}$

$= \frac{1}{36}$

$P(5 \text{ and } 4) = 0.027 \rightarrow 2.7\%$

2) Dependent Event :-

Present Event is depend on the previous Event.



$= 7$

1st time $\rightarrow$ $P(\bigcirc) = \frac{3}{7} = 0.43$

2nd time $\rightarrow$ $P(\triangle) = \frac{2}{6} = 0.3$

3rd time $\rightarrow$ $P(\square) = \frac{2}{5} = 0.4$

4th time $\rightarrow$ $P(\bigcirc) = \frac{2}{4} = 0.5$

Ex: From a deck of Cards what is the probability of getting a king and then 8?

$$P(A \text{ and } B) = P(A) \times P(B/A)$$

$$P(K \text{ and } 8) = P(K) \times P(8/K)$$

$$= \frac{4}{52} \times \frac{4}{51}$$

$P(K \text{ and } 8) = 0.07 \rightarrow 7\%$ chance

Permutation & Combination

1) Permutation:-

$$^nP_r = \frac{n!}{(n-r)!}$$

{Dosa, Idly, Vada, Puri}

DI	ID	VD	PD
DV	IV	VI	PI
DP	IP	VP	PV

$\therefore {}^nP_r = \frac{4!}{(4-2)!} = 12$,,

n → no of items

r → no of ways

Permutation refers to the different way in which a set of items can be arranged in order and in permutation the orders of the items matters but not items.

---

2) Combination:-

{Dosa, Idly, Vada, Puri}

DI
DV
DP
IV
IP
VP
} 6

$$^nC_r = \frac{n!}{(n-r)!\,r!}$$

$$= \frac{4!}{(4-2)!\,2!}$$

$$= \frac{4 \times 3 \times 2 \times 1}{(2\times1)\cdot(2\times1)}$$

$$= 6$$

It refers to the different way of selecting item from a set where the order of selection doesn't matters but items should not repeat.

6/11/2025

Hypothesis Testing

H_0 / H_N ⟶ Null Hypothesis (True Statement)

H_1 / H_A ⟶ Alternative Hypothesis

H_0 - The avg height of Indian is 5.7

H_{A1} - No, the avg height of Indian is not 5.7

H_0 - True Statement

H_{A1} - It will helps to find either we have to accept (or) reject H_0.

P - Value

If P value $< = \alpha$ we can reject H_0

If $P > \alpha$ we can accept H_0

α - Significance value

$\alpha = 1 - CI$

CI (Confidence Interval) { 68% 95% 99.7% }

If CI is 68% → $\alpha = 0.32$

If CI is 95% → $\alpha = 0.05$

If CI is 99.7% → $\alpha = 0.003$

P. Value formula:-

$$z = \frac{\hat{p} - p_0}{p_0(1-p_0)/n}$$

$\hat{p}$ = Sample Propotion

p_0 = Assumed Population Propotion in the H_0

n = Sample Size

$P = 0.35$, $\alpha = 0.05$ → accept

$P = 0.03$, $\alpha = 0.05$ → reject

- Hypothesis Testing is a farmeawork for making jferences about data and models in Machine Leaning.

- It helps in model evaluation, feature selection, assumption validation and ensuing the fast robosness & reliability of conclusions drawn from models.

Type I and Type II Error:-

R Ho is True, D Ho is True ✓

R Ho is True, D Ho is False (Type I Error)

R Ho is False, D Ho is True (Type II Error)

R Ho is False, D Ho is False ✓

R - Reality D - Decision

Type I - If you failed to accept Ho

Type II - If you failed to reject Ho (Most dangerou

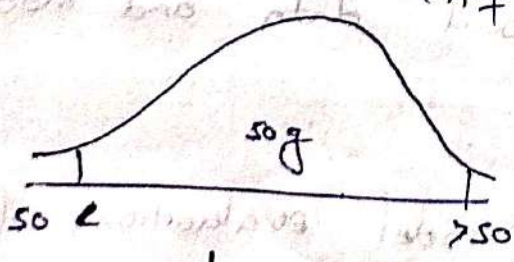
One tail test and two tail test

Ho → The chips packet weight is 50g

HA → The chips packet weight is not 50g.

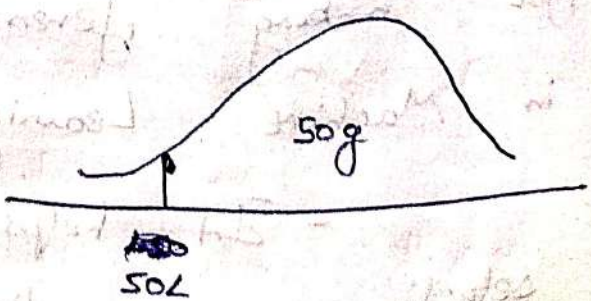
Ho = 50g

HA ≠ 50g



↓

Ho (Two tail)



HA (One tail)

Z test & T test:

The avg age of an college stu is 24 years with SD = 1.5, Sample of 36 Students, The mean is 25 years with 95% confidence Interval. Do the age will vary or not?

Sol:

Given: Population mean $\mu = 24$

Population SD $\sigma = 1.5$

Sample mean $\bar{x} = 25$

Confidence Interval = 95%.

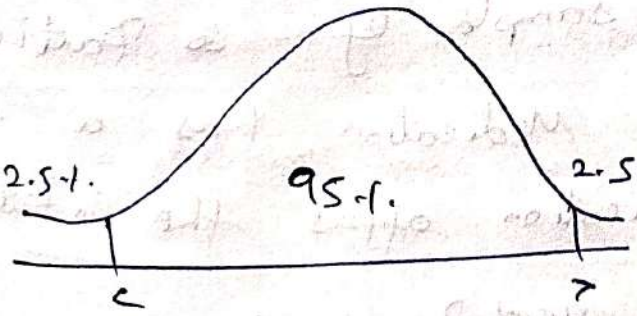
$\alpha = 1 - 95\% = 0.05$

If they given Population SD — "Z test"

If they given Sample SD — "T test"

Null Hypothesis $H_0 \rightarrow$ Avg age is 24

Alternate Hypothesis $H_A \rightarrow$ Avg age is not 24



$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

$$= \frac{25-24}{\frac{1.5}{\sqrt{36}}} = \frac{1}{(1.5/6)} = 4$$

$\therefore \boxed{Z = 4}$

In Z table value for 4 with α of 0.05

$\therefore$ 0.99997

$\boxed{AUC = 1}$

$\therefore 1 - 0.99997 = 0.00003$

$$= \frac{0.00003}{2}$$

$\boxed{P = 0.000015}$ $\boxed{\alpha = 0.05}$

$\therefore \boxed{P < \alpha}$

$\therefore$ So we can Reject Ho

In the population the avg IQ is 100 w SD = 15. Researcher want to test a new medication see if there is +ve @ -ve Effect on intelligence @ no Effect at all a sample of 36 Participants who have taken the Medication has a mean of 140 did the medication affect the intelligence & of with a CI 95%?

$M = 100 \quad n = 30 \quad \bar{x} = 140$

$CI = 95\% \quad \alpha = 0.05$

$$Z = \frac{\bar{x} - M}{\sigma / \sqrt{n}} = \frac{140 - 100}{15 / \sqrt{30}} = 14$$

Z table check value for 14 with α 0.05

$Z = 14 \rightarrow 1$

$P = 1 - 1$

$P = \frac{0}{2} = 0$

$P < \alpha$ so we reject

H_0 = The avg IQ is 100

H_A = The avg IQ is not 100

$M = 130 \quad n = 30 \quad \bar{x} = 140$

$CI = 95\% \quad \alpha = 0.05$

$$Z = \frac{140 - 130}{15 / \sqrt{30}}$$

$$= \frac{10}{2.7}$$

$Z = 3.66 \quad \alpha = 0.05$

7/11/2025

Ex: $\mu = 100$ $n = 30$ $\bar{x} = 140$ $S = 20$ $\alpha = 0.05$

↓

Sample SD, so t dist

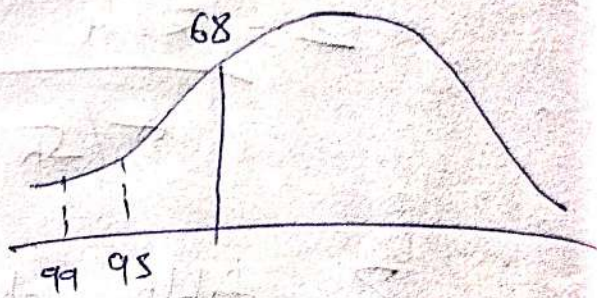
$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}}$$

$$= \frac{140 - 100}{\frac{20}{\sqrt{30}}}$$

$$= \frac{40}{20/\sqrt{30}}$$

$$= \frac{40}{3.6514}$$

$$t = 10.95$$



DOF = n - 1

= 30 - 1

= 29

$t >$ t table value with DOF = 29 & $\alpha = 0.05$

$H_0 \rightarrow$ Avg IQ = 100

IQ ≠ 100 } Two tail

∴ Table value = 2.045

∴ t value > Table value

10.95 > 2.045

So we can reject //

∴

Ex: Credit Card launch:

$n = 140 \quad \bar{x} = 1990 \quad \sigma = 2500 \quad S = 2833$

$CI = 95\% \quad \alpha = 0.05$

We have to calculate the range interval.

Default always the CI will be 95%.

$$CI = \bar{x} \pm Z_{\alpha} \frac{\sigma}{\sqrt{n}}$$

Lower bound

$$= \bar{x} - Z_{\alpha} \frac{\sigma}{\sqrt{n}}$$

$$= 1990 - \frac{Z_{0.05}}{2} \frac{2500}{\sqrt{140}}$$

$$= 1990 - Z_{0.025} (211.29)$$

$$= 1990 - (1.96)(211.29)$$

$$= 1990 - (414.12)$$

$$= 1575$$

Upper bound

$$\bar{x} + Z_{\alpha} \frac{\sigma}{\sqrt{n}}$$

$$= 1990 + \frac{Z_{0.05}}{2} \left(\frac{2500}{\sqrt{140}} \right)$$

$$= 1990 + (1.96)(211.29)$$

$$= 1990 + 414.12$$

$$= 2404$$

∴ The avg balance they are going to maintain after fullfledged launch is [1576 2404]

Ex: On a quant test of a CAT Exam of a sample of 25 test takers as a sample mean of 520 with sample standard deviation of 80, construct 95% confidence interval about the mean.

Sol:

$n = 25$ $\bar{x} = 520$ $S = 80$ CI = 95 $\alpha = 0.05$

$CI = \bar{x} \pm t_{\alpha} \frac{S}{\sqrt{n}}$ DOF = $n - 1 = 24$

Lower Band

$CI = 520 - t_{\frac{0.05}{2}} \frac{80}{\sqrt{25}}$

$= 520 - t_{0.025}(16)$

$= 520 - (2.064)(16)$

$CI = 486.976$

Upper Band

$CI = 520 + t_{0.025} \frac{80}{\sqrt{25}}$

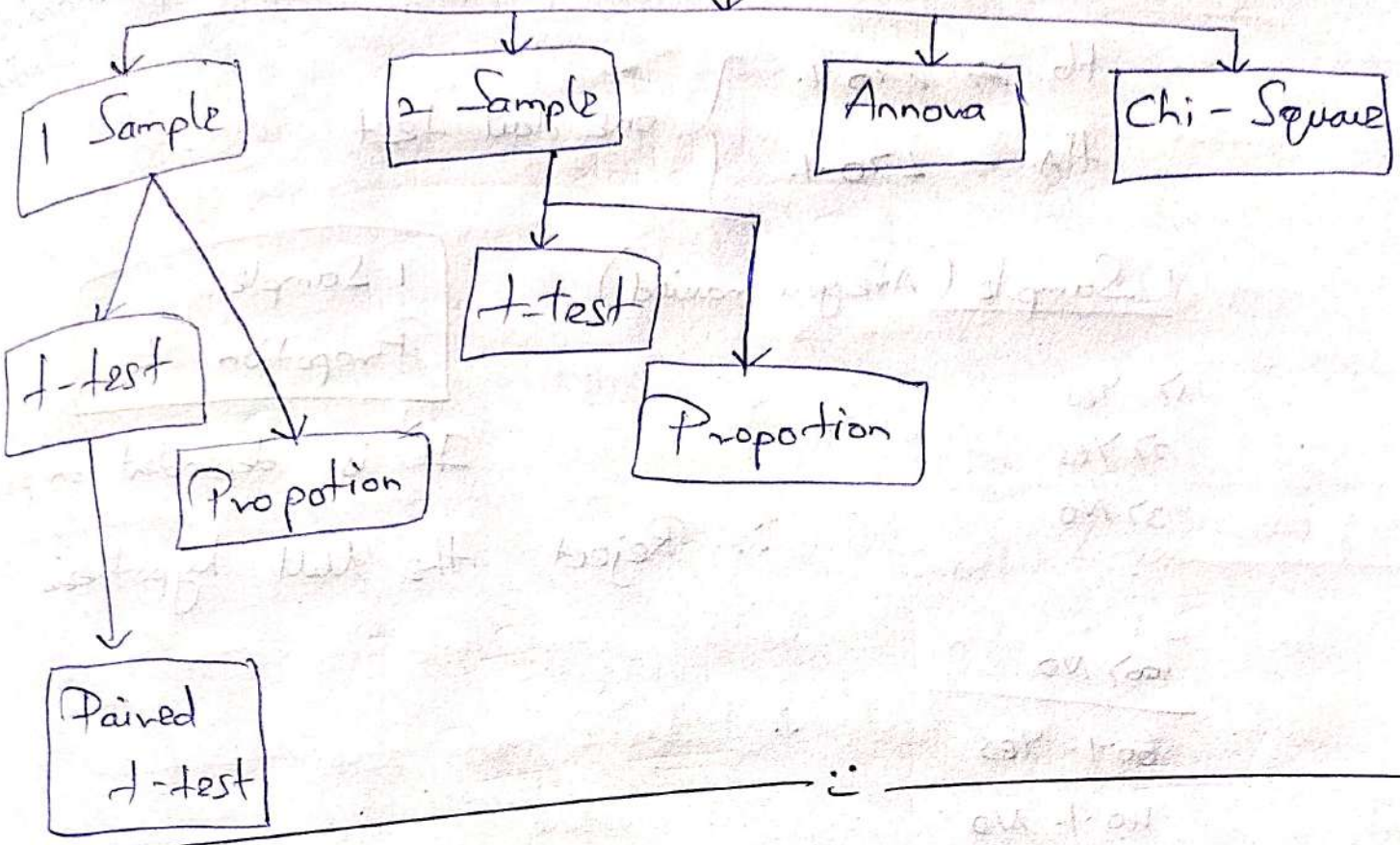
$= 520 + t_{0.025}(16)$

$= 520 + 2.064(16)$

$CI = 553$

$\therefore$ Rang [487 553]

Statistical tests



∴

10/11/2025

Ho - The avg sal of IT Emp in BLR is 27 K

HA - The avg sal of IT Emp in BLR is not 27 K.

1 Sample t-test → depends on population

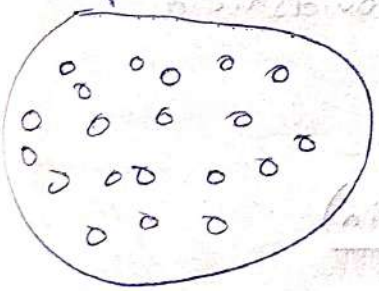
Ho → sal = 27 K

HA → sal ≠ 27 K

Sal < 27 K >

∴ Two tail test

Population



Sample 1

1) 28 K

2) 30 K

3) 18 K

⋮

(000) 35 K

Avg = 32 K

Reject Null Hypothesis

∴ The avg sal of IT Emp IN BLR is 32 K

H_0 - More than 70% of People are married in India

H_A - No more than 70% of People are ~~not~~ married in India.

$H_0 \Rightarrow > 70\%$
$H_A - < 70\%$
} one tail test

1 Sample (Are you married)

1) Yes
2) Yes
3) No
⋮
100) No

60% Yes
40% No

1 Sample Propotion test

It is dependent on popul

∴ Reject the Null hypothesis

---

2 Sample t-test :- (Independent on Population) :-

H_0 : Covidshield is better than Co-vaxin

H_A : No, Covaxin is better than Covidshield

How many hrs it take to react

Sample 1 (covidshield)

1) 2hr
2) 3hr
⋮
50) 1hr

3.5hr

Sample 2 (co-vaxin)

1) 4hr
2) 5hr
⋮
50) 2hr

7.2hr

∴ We have to accept H_0, because we got 3.5 hrs reaction time for covidshield & 7.2 hrs for co-vaxin

---

2-Sample Proportion : Independent on Population:-

H_0 – New Beauty treatment is better than old one

H_A – Old Beauty treatment is better than new one

Are you satisfied with the treatment

New (S_1)	Old (S_2)
1) Yes	1) No
2) No	2) Yes
⋮	⋮
50) Yes	50) No
50% Yes	80% Yes
50% No	20% No

∴ We can reject H_0 Bcz 80% of people like old

---

1 Sample Paired t-test

H_0 → By Joining new weightloss program you can see significant difference in your weight.

H_A → By Joining new weightloss program, you can't see significant difference in your weight.

Sample I

Before	After
1) 70	1) 65 ✓
2) 80	2) 70 ✓
⋮	⋮
10) 100	10) 192 ✗

We can accept the Ho, bcz majority got weight lost.

11/11/2025

Ho - Your batch students cant able to score > 90 M

HA - No my batch students can able to score > 90 M

B_1	B_2	B_3	B_4	B_5
63	75	92	47	81
62	88	98	58	75
58	83	88	63	78
73	65	69	70	85
80	73	95	68	70
65	78	(92)	59	80

ANNOVA

∴ If one sample proved means we can reject the Ho.

∴ We reject the null Hypothesis

In the 2000 Indian Censeus the age of the individual in a small town where found to be following.

In the year 2000

Less than 18	18 – 35	>35
20 %	30 %	50 %

In 2010 Age of n=500 individuals were sampled below are the results.

In year 2010

Less than 18	18 – 35	>35
121	288	91

Using $\alpha = 0.05$ would you calculate the population distribution of ages has changed in the last 10 years?

Year	<18	18-35	>35	
2000	20 %	30 %	50 %	
2010	100	150	250	→ ~~observed~~ Expected
	~~100~~ 121	288	91	→ observed

H_0 — 2010 sensus ratio same as 2000

H_A — No, 2010 sensus is not same as 2000

Chi-Square test:-

χ^2 > chi-square table value with Degree of freedom means you can reject else accept.

Degree of freedom = 3-1 = (2) - DOF

$\alpha = 0.05$

$\therefore \chi^2 > 5.991$ then reject H_0

$$\chi^2 = \sum_{i=1}^{n} \frac{(fo - fe)^2}{fe}$$

fo - observed value

fe - expected value

$$\chi^2 = \frac{(121-100)^2}{100} + \frac{(288-150)^2}{150} + \frac{(91-250)^2}{250}$$

$$\chi^2 = 232.494 > 5.991$$

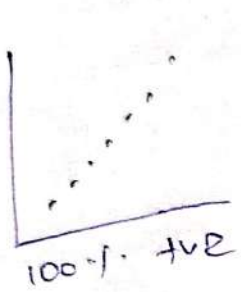
$\therefore$ So we can reject Null Hypothesis H_0

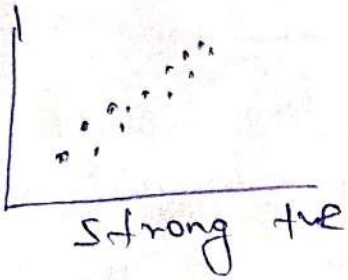
Correlation:-

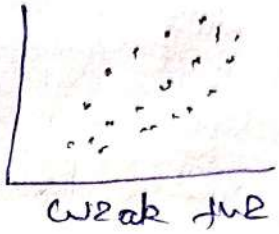
To check the relationship b/w two numerical features/column we use co-relation

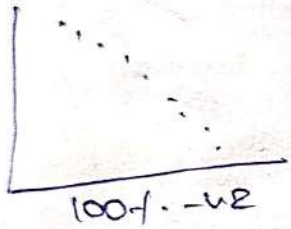
There are 3 types of Correlation:-

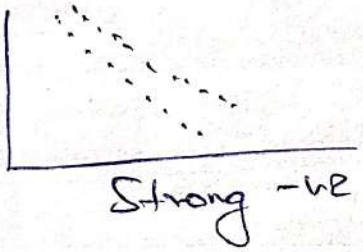
1) +ve Co-ve → x↑ y↑ → Exp ↑, Sal ↑
2) -ve Co-ve → x↑ y↓ → Weight ↑, MPG ↓
3) No Co-ve → x↑ y↑↓ → weight, Sal

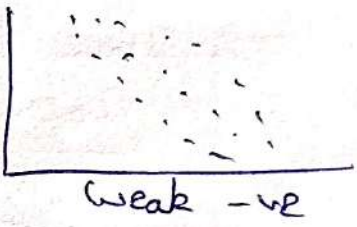


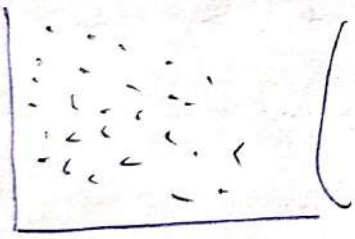












(No co-relation)

There are two types of co-relation formula

1) Pearson- Correlation
2) Spearman- Co-relation

- The co-relation value ranges from -1 to +1
- If co-relation value near to +1 is a +ve-Corr
- If co-relation value near to -1 is -ve Corr
- If value near to Zero - No-Corr.