Page No.
Date

TOPS TECHNOLOGIES
AHMEDABAD

NAME ⇒ TANMAY SUTHAR

Course ⇒ Data Analysis

12 Dec. Batch

Teacher's Signature .....

What is Statistics

It is a Science of collecting, organizing, analyzing data for better decision making.

What is Data

Facts or pieces of information that can be measured.

Eg :-

No. of class

{ 89, 90, 50 ..... }

age of Students

{ 24, 22, 21 ..... }

Types of Stats

Descriptive Stats : It consist of orgnizing & Summarizing data

Inferential Stats : Using data, we can make conclusion using some techniques

Teacher's Signature .....

Page No.
Date

Eg:
Class of 20 students
1st Sem Maths & 86, 70, 90, 55 ... 3

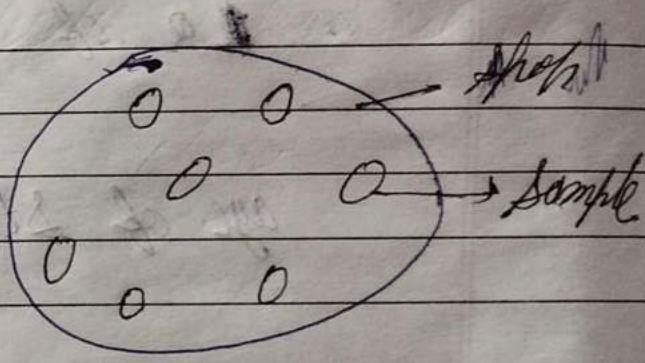
1 what is the avg of class (descriptive)
2 7th Sem (Inferential)

Sample and population

Population = N

↓

Whole dataset is known as population



Sample = n

Small subsets of data taken from population → Sample

Teacher's Signature .....

Sampling Techniques

1 Simple Random Sampling - Every member of population has an equal chance of getting selected in sample (n).

2 Stratified Sampling - Splitting data into non-overlapping groups

Age group — 0-20
— 20-40

Genders — M
— F

3 Systematic Sampling.

4 Convenience sampling - domain expert

1 Exit poll ?

2 . . . . .

(3) A method for selecting a sample from a population in a randomized manner.

Teacher's Signature .....

Page No.
Date

Variables

It is a property that can hold / store / take any value

Age : { 8, 10, 15, 20, 25 ... }

Marks : { 76, 80, 95 ... }

Types

1 qualitative : Categorical values
(based on some characteristics we can derive categorical values)

IQ : 0 - 10 → low →
10 - 50 Avg
50 - • good

2 quantitative : Numerical value (measurable numerically)

height : { 162, 159, 155 ... }

weight : { 59, 65, 79 ... }

- discrete (int) whole No
- continuous (float) decimal No

Teacher's Signature .....

1 No. of Students
2 No. of Rank ans

1 height
{ 165.2, 167.9 ...

2 weight
165.5 - 160.9 ... 3

1 Blood Pressure → continuous / discrete
2 Marital Statics → qualitative
3 Rivers length → cont
4 Day length → cont
5 genders → qualitative & category 3

Variable Measurement scales

marks
[90, 85, 80, 100] → [2, 3, 1]

1 ordinal: ordered [rank, order, Mates, graduates
2 Nominal: categorical values (colors, classes, degrees)
3 Interval: [No zero / absolute point] (orders as well as value matters)
4 Ratio: Zero means nothing

Teacher's Signature .....

Page No.
Date

Interval

[20°C : 210°C]

Ratio

[20kg : 40kg]

1 : 2

20°C : 210°C

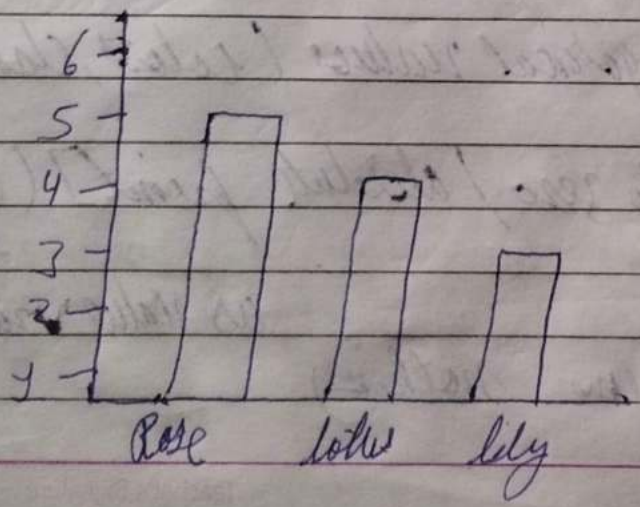
1°C : 2°C

Frequency ....

data : flowers

[Rose, lily, lotus, Rose, Rose, rose, rose, lotus, lily, lotus, lily lotus]

flowers	frequency	(cumulative
Rose	5	5
lily	3	8
lotus	4	12
	12	



bar graph / chart

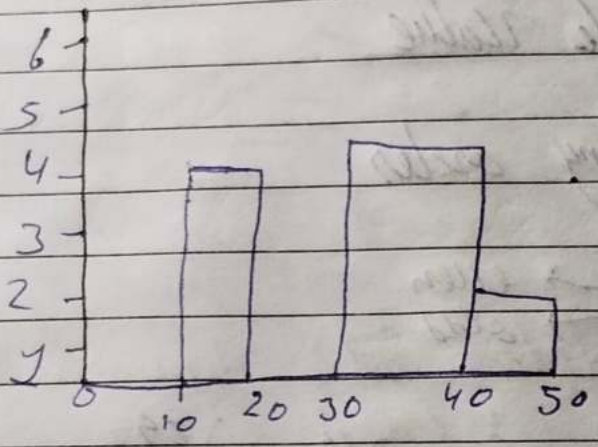
Teacher's Signature .....

Histogram

Marks = [12, 15, 12, 15, 21, 27, 28, 35, 34, 36, 39, 42, 45]

bins		
	(10 - 20)	4
	(20 - 30)	3
	(30 - 40)	4
	(40 - 50)	2

→ continuous



→ Measure of central Tedency

Avg → Mean

PoP

$$\mu = \frac{\sum x_i}{N} \qquad \bar{x} = \frac{\sum x_i}{n}$$

{2, 3, 5, 3, 2, 1, 3}

$$\frac{2+3+5+3+2+1+3}{7}$$

Teacher's Signature .....

Page No.
Date

⇒ Mean: It refers to the measure used to determine the center of the distribution/ determ of the data.

$\{1, 2, 2, 2, 3, 3, 4, 5, 5, 6, 100\}$ → outliers

$\frac{32}{10} \Rightarrow 3.2 \quad \rightarrow \quad \frac{132}{11} \Rightarrow 12$

⇒ Median: middle value

↳ ascending order

↳ data → even
→ odd

⇒ Even $\frac{\left(\frac{n}{2}\right)^{th} + \left(\left(\frac{n}{2}\right)^{th} + 1\right)^{th}}{2}$

dataset: $\{11, 12, 13, 14, 15, 16\}$

$n = 6$

$\frac{\left(\frac{6}{2}\right)^{th} + \left(\frac{6}{2}\right)^{th} + 1}{2}$

Teacher's Signature .....

$$\frac{3^{th} + (3+1)^{th}}{2}$$

$$\frac{3^{th} + 4^{th}}{2} \Rightarrow \frac{13 + 14}{2}$$

$$\frac{27}{2}$$

$$\Rightarrow 13.5$$

median = 13.5

odd

$$\frac{(n+1)^{th}}{2}$$

{ 11, 12, (13), 14, 15 }

$n = 5$

$$\left(\frac{5+1}{2}\right)^{th} \Rightarrow \frac{6^{th}}{2} = 3 : 13$$

{ 11, 12, 13, 14, 15, 100 }

$$\frac{\left(\frac{n}{2}\right)^{th} + \left(\left(\frac{n}{2}\right)^{th} + 1\right)^{th}}{2}$$

$$\frac{\left(\frac{6}{2}\right)^{th} + \left(\left(\frac{6}{2}\right)^{th} + 1\right)^{th}}{2}$$

$$\frac{3^{th} + 4^{th}}{2} = \frac{13+14}{2} = 13.5$$

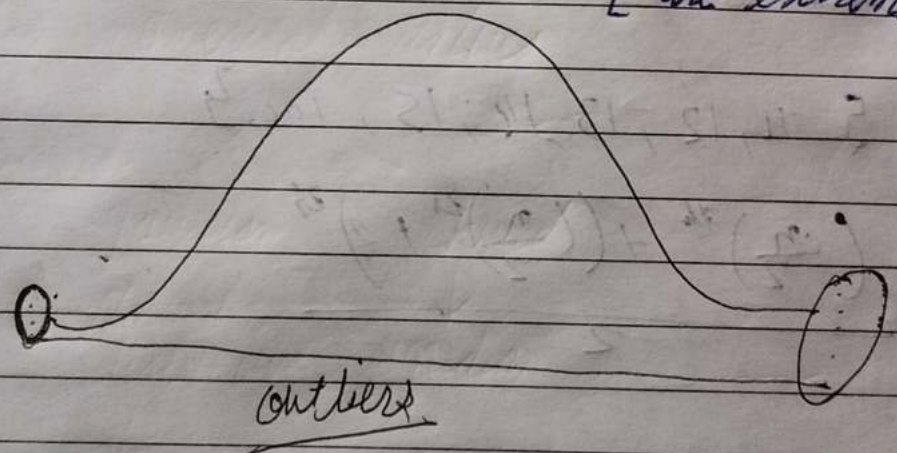
$\{21, 23, 25, 29, 32, 100\}$

$$\frac{25+29}{2} = \frac{54}{2} = 27$$

outlies :-

a data points who doesn't follow pattern or trend of the dataset then it is considered as outlies.

[are extreme points].



Teacher's Signature .....

Page No. 11
Date 16/12/24

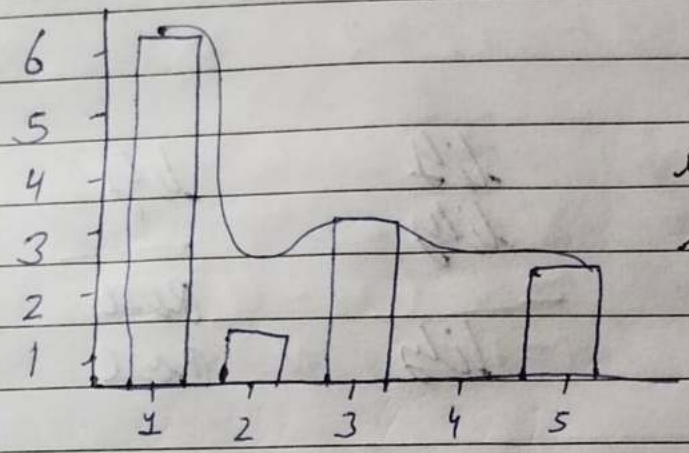
Mode

Most frequent value (repeated)

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

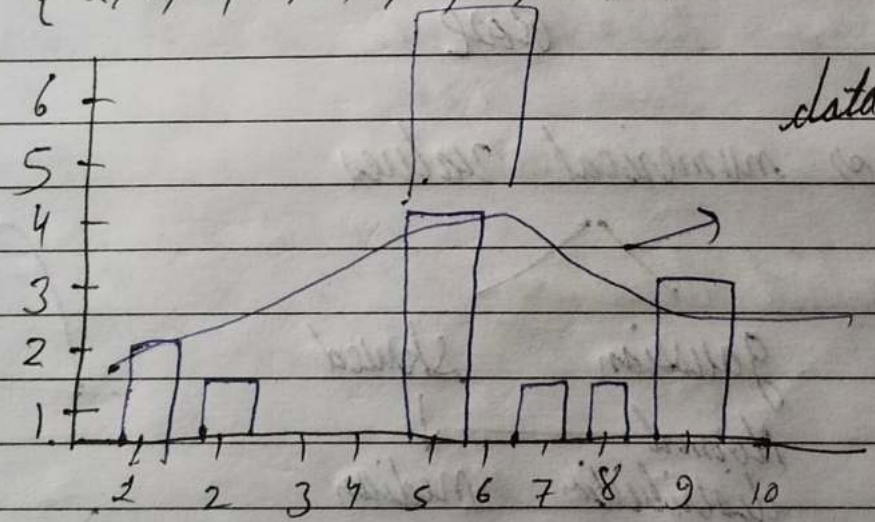
1 = 6 3 = 3

2 = 1 5 = 2



Kernal density estimation Kde plot

[2, 1, 2, 5, 9, 5, 7, 8, 9, 5, 9, 5]



data distribution Kde

Teacher's Signature .....

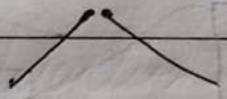
Page No.
Date

for categorical missing data

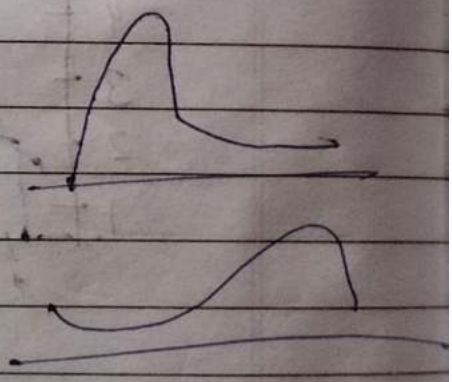
0-5% → mode
→
↳ new category "missing"
"unknown"
or
"random"

1S	species		
Rose	lily	lily	
lotus	lily	____	null
lily	____	Rose	na
____	lily	Rose:	
Rose	____		1S - (10)
lily	Rose		
Rose	lotus		
	Rose		

For numerical values

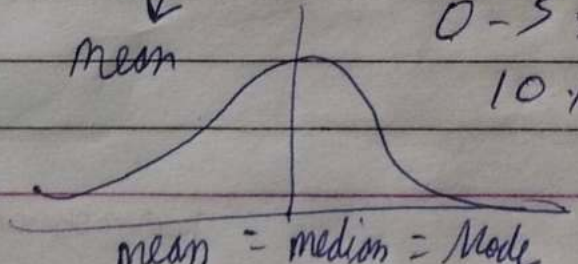


Gaussian
Normal distribution
↓
mean



Skewed
↓
Median

0-5%
10%



mean = median = Mode

Teacher's Signature .....

Measure of dispersion → spread

(1, 5)

[5, 1, 1, 1, 2, 3 → $\frac{5+5}{5}$ = 2

[2, 2, 2, 2, 2] = (2)

→ Variance

It means how far the numbers in a dataset are from the mean (avg).
(how each value differs from a dataset in mean.

High Variance → more spread (far from mean)

low Variance → closer to mean

PoP

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N} \qquad s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

Summation

S_1 — Bessel's correction

is the use of n-1 instead of n in the formula for the sample variance and sample standard deviat

Teacher's Signature .....

x	$x_i - \mu$	$(x_i - \mu)^2$
1	-1.83	3.34
2	-0.83	0.69
2	-0.83	0.69
3	0.17	0.02
4	1.17	1.36
5	2.17	4.70

108

$$\mu = \frac{1+2+2+3+4+5}{6} = 2.83$$

$$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N} = \frac{10.8}{6}$$

$$\sigma^2 = 1.8$$

$x_i \rightarrow$ every data point
$\mu \Rightarrow$ mean of PoP
$N \Rightarrow$ total PoP
$\Sigma =$ summation

Standard deviation

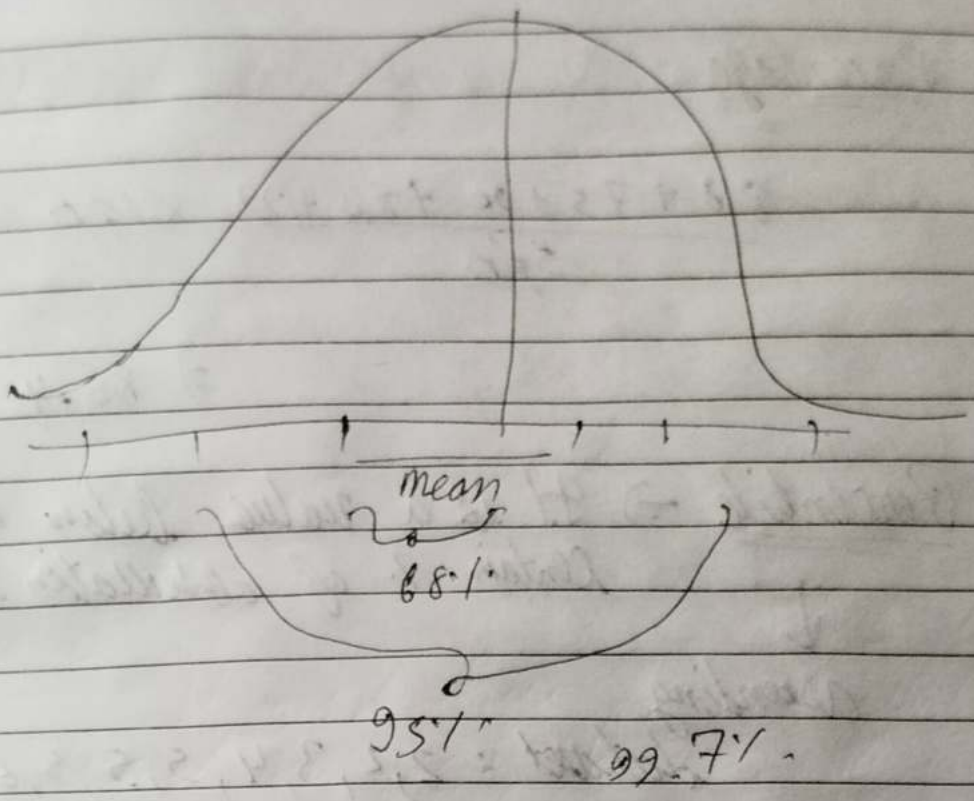
PoP

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

Sample

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

Teacher's Signature .....



$\mu \pm 1\sigma = 68\%$.
$\mu \pm 2\sigma = 95\%$.
$\mu \pm 3\sigma = 99.7\%$.

→ Square root of variance

→ It gives measure of spread that is in the same unit as the original data, making it easier to interpret.

Teacher's Signature .....

Page No.
Date

⇒ Percentage

$$\frac{88+75+90+80+99}{500} \times 100$$

$$\Rightarrow 86.4\%$$

⇒ Percentile ⇒ It is a value below which a certain % of observation lie.

↓
ascending

dataset = 2, 2, 3, 4, 5, 5, 5, 6, 7, 8, 8, 8, 8, 8, 9, 9, 10, 11, 11, 12

$n = 20$

Percentile rank of 10 = $\frac{\text{no. of values below } x}{n}$

$$\Rightarrow \frac{16}{20} = \frac{4}{5} \times 100$$

80% values are below 10 = 80

$$11 = \frac{17}{20} \times 100 = 85$$

85% value are below 11

Teacher's Signature .....

what value exists at percentile rank 25?

$$\text{value} = \left(\frac{\text{percentile}}{100} \times n\right) + 1$$

$$\left(\frac{25}{100} \times 20\right) + 1$$

$$5 + 1 = 6 \rightarrow \text{index value}$$

$$75 = \left(\frac{75}{100} \times 20\right) + 1$$

$$\therefore 15 + 1 = 16$$

Page No.
Date

five numbers Summary

1. Minimum
2. 25 percentile → first quantile

Median → 50 Percentile

75 percentile

Maximum

[1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 15, 27]

lower fence = $Q_1 - 1.5(IQR)$

higher fence = $Q_3 + 1.5(IQR)$

IQR → $Q_3 - Q_1$ → Inter quartile Range

$Q_2 = 3$

$Q_3 = 8$

$IQR = 8 - 3 = 5$

lower fence = $3 - 1.5(5)$

$= -4.5$

higher fence = $8 + 7.5$

$= 15.5$

Teacher's Signature .....

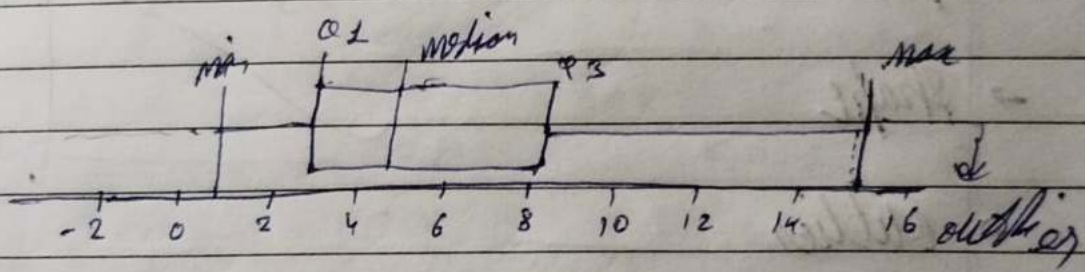
1 Min = 1

2 Q_1 = 3

3 Median = 5

4 Q_3 = 8

5 Max = 15



Teacher's Signature .....

Page No.
Date 20/12/24

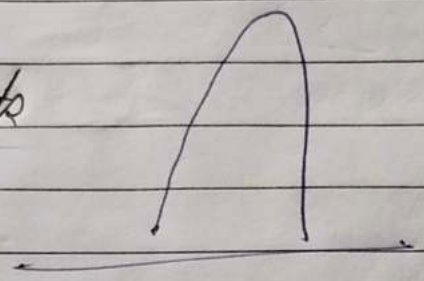
distribution

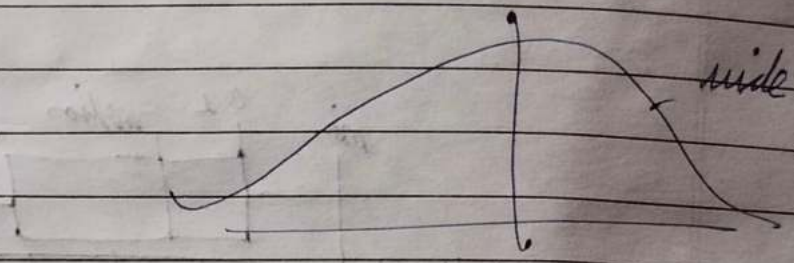
data distribution

- It refers to a way in which values or data pts are spread or arranged.
- It shows how often different values occur in data set & describes the overall pattern of the data.

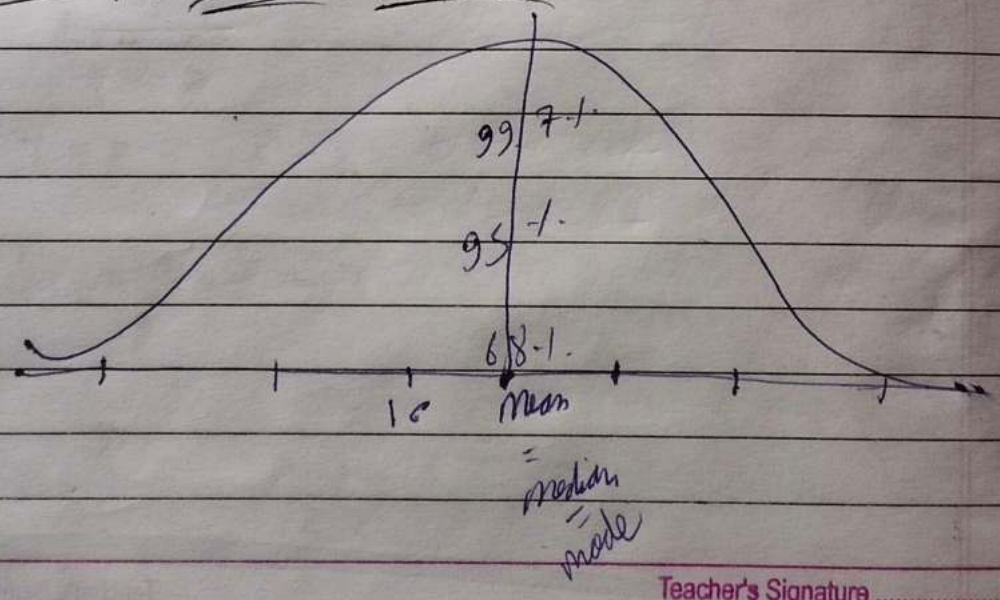
⇒ 4 key points

- center
- spread.
- Shape
- outlier





gaussian / Normal distribution



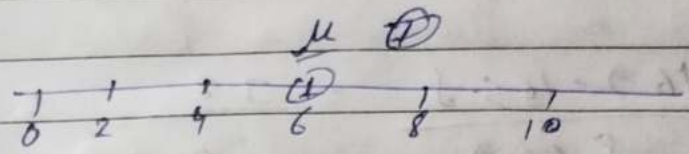
Teacher's Signature .....

$\mu \pm 1\sigma = 68\%$

$\mu \pm 2\sigma = 95\%$

$\mu \pm 3\sigma = 99.7\%$

} empirical Rule



$\sigma = 2.5$
7.5
1.5σ

Z-Score

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

$\mu = 0$ $\sigma = 1$

Standard Normal distribution

↓

$\mu = 0$ $\sigma = 1$

Height	Weight	$x_i - \mu_n$	
169	60	3.2	10.24
172	65	6.2	38.44
150	45	-15.8	249.64
168	70	2.2	4.84
170	71	4.2	17.64
829			320.8

Teacher's Signature .....

Page No.
Date

$M_x = \frac{829}{5} = 165.8$

$\sigma^2 = \frac{320.8}{5} = 64.16$

$\sigma_x = 8.009$

$z_1 = \frac{169 - 165.8}{8.009} = \frac{3.2}{8.009} = 0.399 \Rightarrow 0.4$

$z_2 = \frac{6.2}{8.009} = 0.7$

$z_3 = \frac{-15.8}{8.009} = -1.9$

$z_4 = \frac{2.2}{8.009} = 0.27$

$z_5 = \frac{4.2}{8.009} = 0.52$

~~$\Rightarrow -0.01$~~

$\Rightarrow -0.01$

Teacher's Signature .....

Normalization

$$x = \frac{x - x_{min}}{x_{max} - x_{min}} \longrightarrow \text{Min Max Scaler}$$

Spread distribution

positively skewed distribution

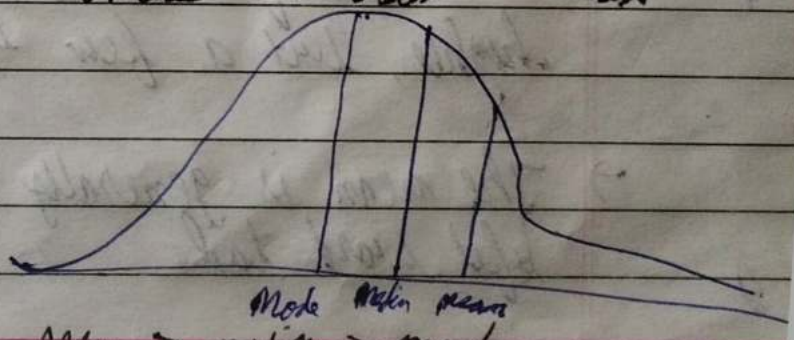
What?

In positively skewed distribution, most values are concentrated on the lower end with a long tail extending to the right. A few high values pulls the average to the right of the median.

Skewness → A distortion or asymmetry that deviates from the symmetrical bell curve

→ also called right-skewed or right tailed distribution.

Mode < Median < Mean



Mean > median > mode

Teacher's Signature .....

When?

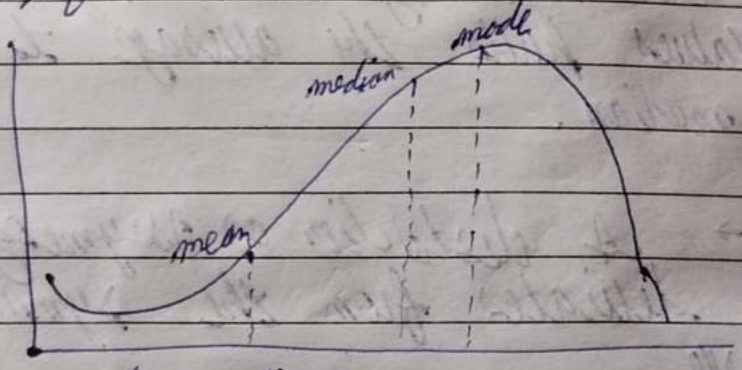
useful for data with some but significant high values such as income levels where few individuals earn much more than the rest.

→ The mean is greater than the median due to outlier on the higher end.

Negatively skewed distribution

What?

most values are clustered at higher end, with a few values creating a long tail to the left.



When?

used for datasets where values are typically higher, but a few lower values exist (retirement age)

→ The mean is generally less than median due to left word tail.

Teacher's Signature .....

→ Also called left skewed or left tailed.

why?

It helps detect cases where data points are generally higher but occasionaly much lower.

Exponential distribution (Continuous)

λ = constant rate

It describes the time b/w events in a process where events occur independently at a constant rate λ.

$$f(x) = \lambda e^{-\lambda x} \qquad x \geq 0$$

Bernoulli distribution (discrete)

It models a single experiment with two possible outcomes.

Success ($x=1$) , failure ($x=0$)

Binomial distribution (discrete)

Binomial dist. extends Bernoulli to n independent trials.

Page No.
Date

$$P(X=K) \Rightarrow \binom{n}{k} p^K \cdot (1-p)^{n-K}$$

(so)

X = random variable (no. of success out of a trial)

n = total no. of trial (100)

K = No. of success ($0 \leq K \leq n$)

P = P(success in 1 trial) 0.5

$$\binom{n}{K} = \frac{n!}{(n-K)!\,K!}$$

$n = 5$ $K = 3$ $P = 0.5$

$$P(3) = \frac{5!}{2!\,3!} \times 0.5^{(3)} \times 0.5^{(2)}$$

Uniform distribution (continous)

$[a, b]$

$$f(x) = \frac{1}{b-a} \quad a < x < b$$

The probability of any value within the rang $[a, b]$ is same

Teacher's Signature .....

* Uniform distribution (discrete)

all outcomes are equally likely

$P(x) = \frac{1}{n}$ → total no.

* Confidence Interval

$\bar{x} = 50$ → point estimate → μ

40-60 → 20K

2 lakh

5 -50+5

45-55 → 50K

40 - 50 60 . +10-10

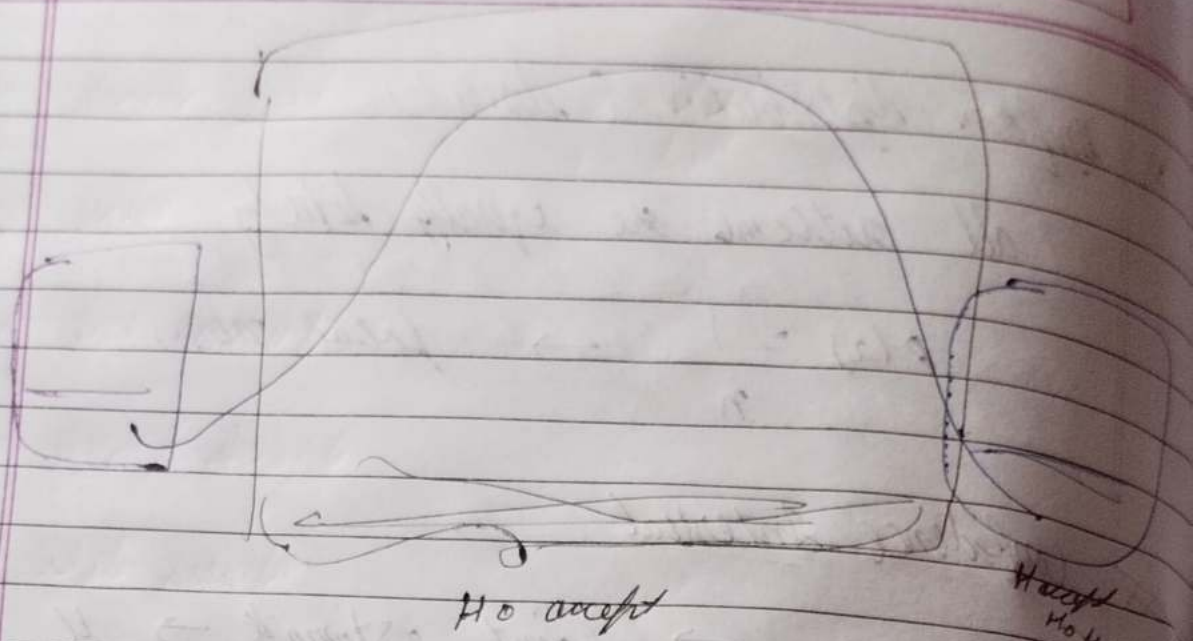
±5 ±10 Margin of error

50 → point estimation

$\frac{45-55}{40-60}$ } Confidence interval

It is a range of values within which we expect a particular population parameter to fall.

Confidence Interval = point estimate ± margin of error



✱ Hypothesis testing

10×D → 100 - test cam

10 → 1000

A statistical hypothesis test is a method of statistical inference used to decide whether the data at hand is sufficient to support a particular hypothesis.

Hypothesis testing allows us to make probabilistic statements about population parameters.

Teacher's Signature .....

Null hypothesis : H_0

The null hypothesis assumes that there is no significant relationship or effect b/w the variables [in simpler terms → it says nothing new is happening].

It serves as a starting point for HTE represents 'status quo' or the assumption of no effect until proven otherwise.

The purpose of HT is to gather evidence to reject or fail to reject hypothesis in favour of alternate hypothesis, which claim there is significant effect or relationship.

Alternate hypothesis H_0 or H_1

It is a statement, that contradicts the NH & claims there is significance effect or relationship.

Rejection Region method

1 H_0 & H_a (1 - 0.05) (0.5)

2 α → value

↳ Significance level → 0.5 / 0.1

3 assumptions ND → 27, 30

6, Teacher's Signature .....

4 decide test — z - test, t - test

5 Value

6 Test conduct

7 Reject / Accept

8 State results

Q) 50 → units per day

$\sigma = 5$

Training →

30 emp → 53 units per day.

1 $\mu = 50 \rightarrow H_0$ $\quad H_a: \mu > 50$ (one tailed test)

2 $\alpha = 0.05$ | 5%

3 Data Normal, Random, $n = 30$

4 z - test

5 z - score $= \frac{x_1 - \mu}{\sigma}$ prop.

Teacher's Signature .....

$$\frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{53 - 50}{5/\sqrt{30}}$$

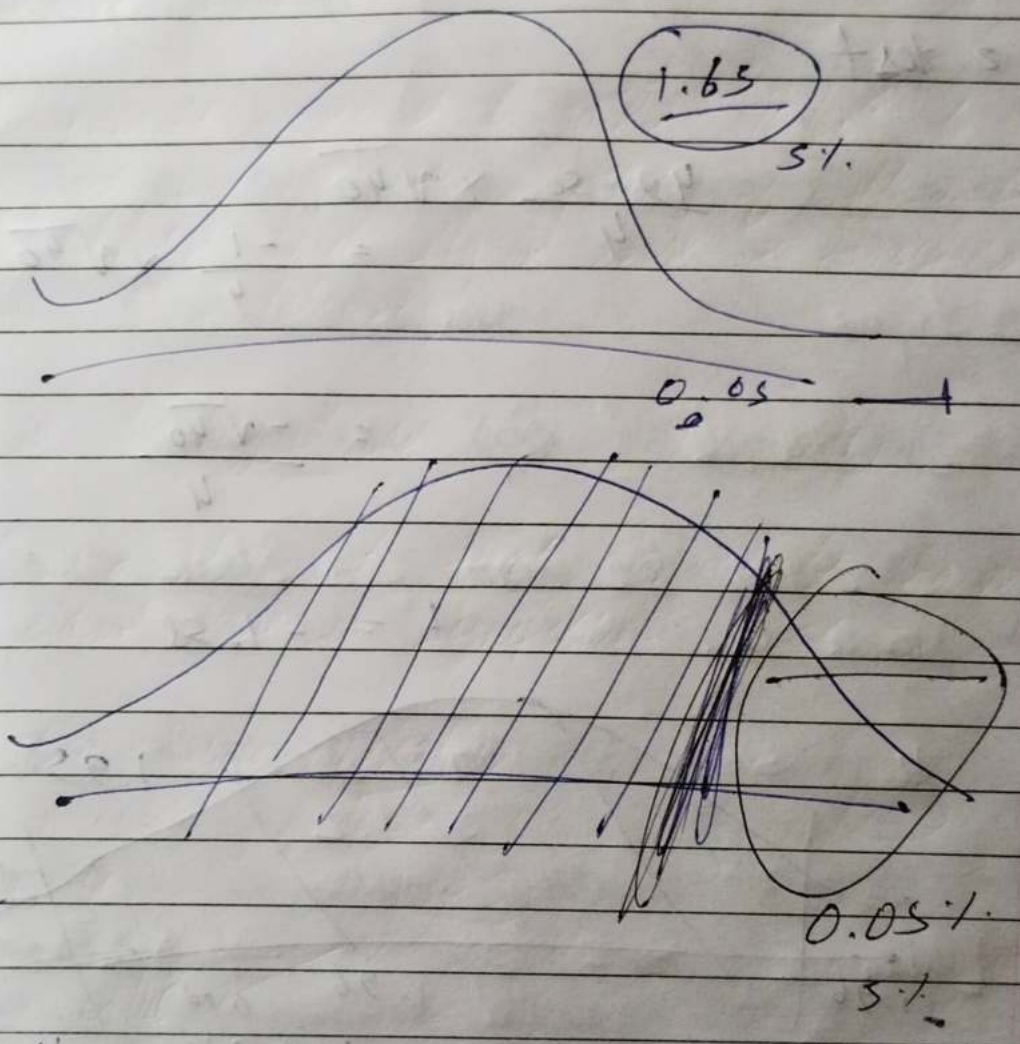
$$= \frac{3}{5} \times \sqrt{30}$$

$\alpha = 5\%$

$$= \frac{9}{25} \times 30$$

$$z = 3.28$$

(6)



7 rejection of H_0

8 $\mu > 50$

Teacher's Signature .....

2 avg - 50g 40g →
s - 4g $\bar{x}$ M = 49

1 H_0: M = 50g

2 $\alpha = 0.05$

3 n ⇒ 30 z test

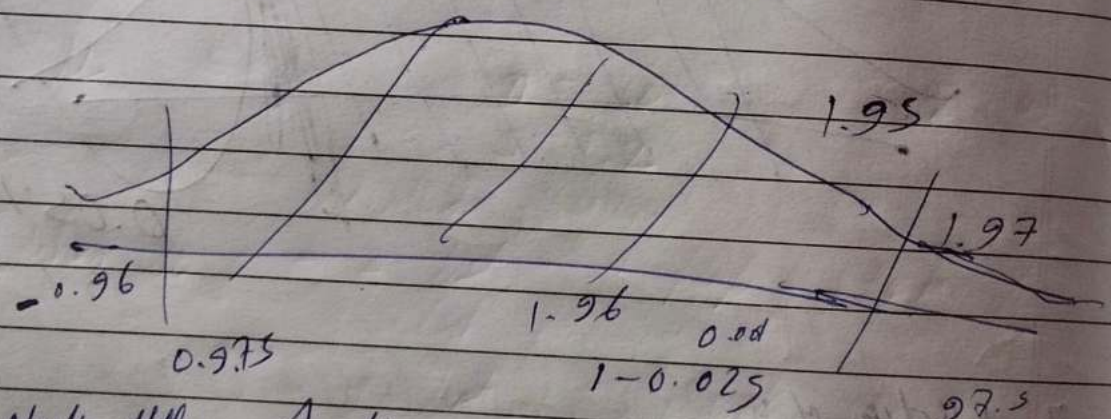
4 z-test

5 $\frac{49 - 50}{4} \times \sqrt{40}$

$= \frac{-1}{4} \times \sqrt{40}$

$= \frac{-\sqrt{40}}{4}$

6 $1 = -1.58$



7 Null Hypo Accept
M = 50g

Teacher's Signature .....

2 Errors

	Type 1	Type 2
	H_0 true	H_0 false
reject H_0	Type-1	correct
accept H_0	correct	Type 2

Type-1 False +ve

↳ H_0 reject → H_0 true (correct)

rejecting H_0 when H_0 is actually correct

Type-2 → false -ve

accept H_0 when H_0 is actually incorrect

$P \geq \underset{\alpha}{0.05}$ → Null accept.

$P < 0.05$ → Null reject.

$P < 0.01$ → strong evidence

$0.01 \leq P < 0.05$ → Moderate evi.

$0.05 \leq P < 0.1$ → weak evidence.

$P \geq 0.1$ → No evi

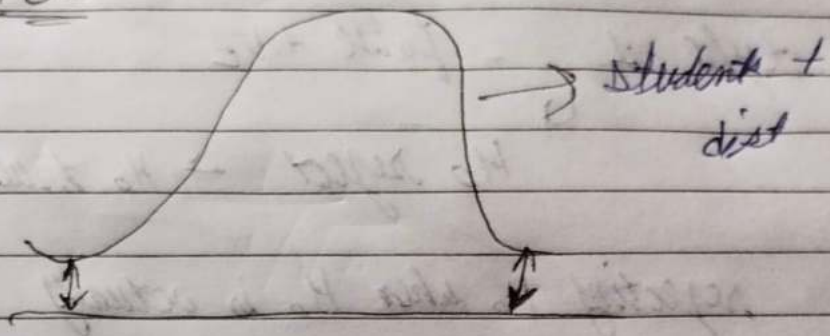
Teacher's Signature .....

$P < \alpha$ → H_0 reject

$P \geq \alpha$ → H_0 accept

P value → It is a measure of the strength of the evidence against the null hypothesis.

T - Test



3 types

One Sample t-test

Compares the mean of a single sample to a known μ.

$\mu = 50$ g $n = 30$ $\bar{x} = 49.7$ g

$s = 1.2$

[when we σ → z-test (pop → standard dev

s → t-test] (sample - std)

Teacher's Signature .....

8. $\mu = 50g$ $n = 25$ $\text{std} = 1.2kg$

$\bar{x} = 49.71g.$

$H_0 = \mu = 50$ $H_a = \mu \neq 50.$

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{49.7 - 50}{1.2/\sqrt{25}} = \frac{0.3}{1.2} \times 5$$

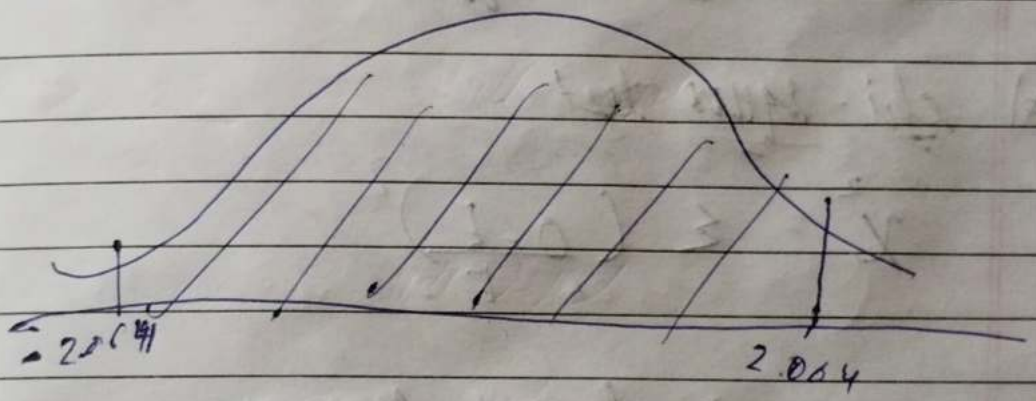
$$= \frac{-1.5}{1.2}$$

$$= -1.25$$

df = degree of freedom

$= n - 1$ $\therefore df = 24$

t critical $= 2.064$



H_0 accept (Sig)

Teacher's Signature .....

Page No.
Date

2 Independent Two Sample t-test

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$ → Standard Error

3 Paired t-test

$$t = \frac{\bar{d}}{sd/\sqrt{n}}$$

$\bar{d}$ = z → Mean difference

sd = std → diff

4 chi-square test

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

O → observed frequency

E → Expected frequency

$$= \frac{\text{grand total}}{\text{grand total}}$$

Teacher's Signature .....

$df = (r-1) \times (c-1)$

r = no. of rows

c = no. of columns

	Satisfied	not Satisfied	Total
high school	50	70	120
college	90	60	150
P.G	20	10	30
	160	140	300

Ef	S	nT	total
HS	$\frac{120 \times 160}{300} = 64$	$\frac{120 \times 140}{300} = 56$	
C	$\frac{150 \times 160}{300} = 80$	$\frac{150 \times 140}{300} = 70$	
PG	$\frac{160 \times 30}{300} = 16$	$\frac{140 \times 30}{300} = 14$	

S	64	56	120
C	80	70	150
PG	16	14	30

Teacher's Signature .....

Page No.
Date

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

$$S.S = \frac{(50-64)^2}{64} = \frac{196}{64} = 3.06$$

$$S.N.S = \frac{(70-56)^2}{56} = \frac{196}{56} = 3.5$$

$$C.S = \frac{(90-80)^2}{80} = \frac{100}{80} = 1.25$$

$$C.NS = \frac{(60-70)^2}{70} = \frac{100}{70} = 1.42$$

$$P.W.S = \frac{(20-16)^2}{16} = \frac{16}{16} = 1$$

$$P.WNS = \frac{(10-14)^2}{14} = \frac{16}{14} = 1.14$$

$$\chi^2 = 3.06 + 3.5 + 1.25 + 1.42 + 1 + 1.14$$
$$= 11.37$$

$$df = (3-1) \times (2-1)$$

$$= 2$$

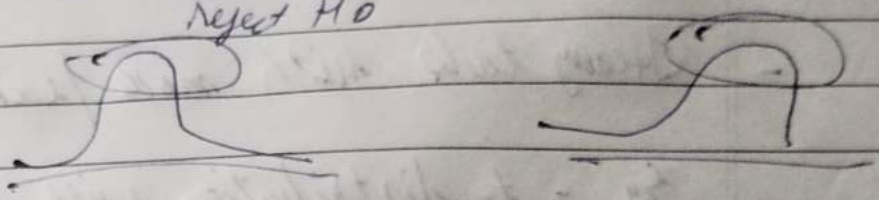
$$\alpha = 0.05$$

Teacher's Signature .....

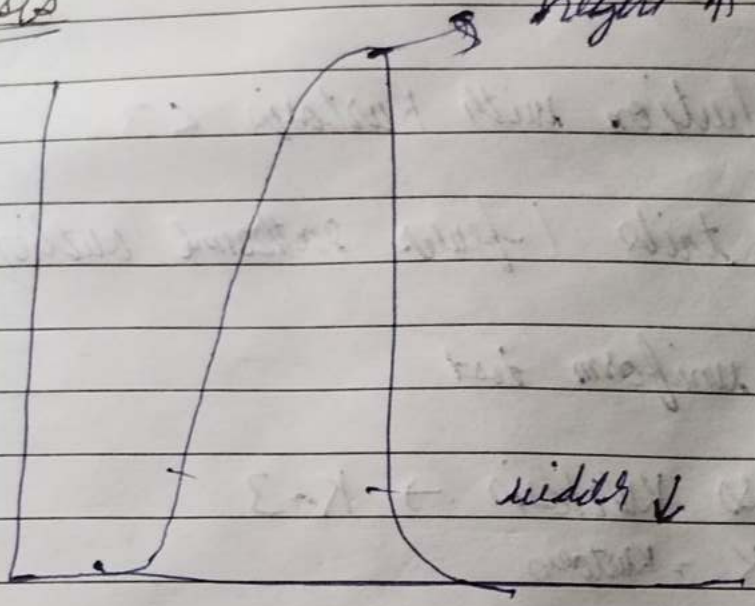
$\chi^2_{critical} = 5.991$

$\boxed{\chi^2 = 11.37}$

Reject H_0



Kurtosis



- Kurtosis measures "tailedness" of a distribution or how extreme the outliers are

1. Mesokurtic

- Tails are similar to ND

- distribution with kurtosis ≈ 3

- eg: Standard Normal distribution

Teacher's Signature .....

2 Leptokurtic

- A distribution with kurtosis > 3
- Heavy tails with more (extreme outliers)

Eg - t-distribution with very small df

3 Platykurtic

- distribution with kurtosis < 3
- Light tails (fewer extreme outliers)
- Eg : uniform dist

Excess kurtosis → K-3

K = Kurtosis

EK > 0 → Leptokurtic

EK < 0 → Platykurtic

$$K = \frac{n \cdot \sum (x_i - \bar{x})^4}{\left(\sum (x_i - \bar{x})^2\right)^2} \cdot \frac{1}{n}$$

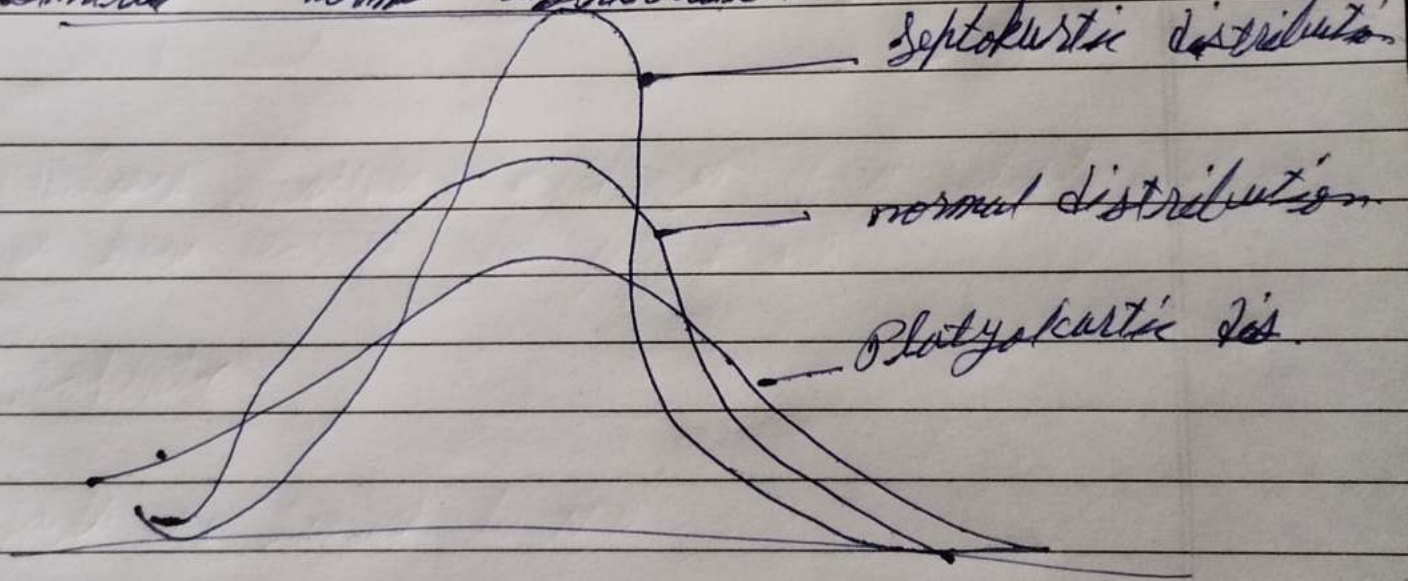
n = no. of observation

x_i = each data point

$\bar{x}$ = Mean

Teacher's Signature .....

- High Kurtosis (Leptokurtic)
  - More extreme outliers
  - Higher likelihood of rare, extreme values.
  - eg financial returns during market crisis/crashes)
- Lower Kurtosis (Platykurtic)
  - fewer extreme outliers
  - data is evenly spread.
  - Kurtosis near 3 (mesokurtic)
  - Similar to normal distribution



Teacher's Signature .....