Date

Statistics

Statistics is the discipline that concerns the collections organization, analysis, interpretation, and presention of data

population Data
sample Data

Statistics

Confidence Interval

Descriptive
(It consist of organi and Symmary Data)

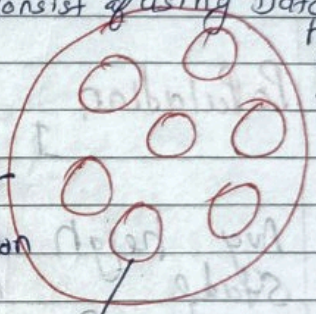
1) Analysis data, Summazing data, orgnalize data in the farm of num s graph

2) BARPlot, Histrogram, PIE chart, PDF, CDF ND

3) measure of Center Tendenay
mean, Median, Mode

4) measure of Variances
SD, Varianes

Inferential stats
(It consist of using Data you have measure from Population



Sample

Party 1 ⇒ ∘
Party 2 ⇒ ″
Party 3 ⇒ ×

take Sample
↳ Inferenes and Conclusion
↓
Population

1) Z Test
2) T Test
3) Chi square

(14)

Date

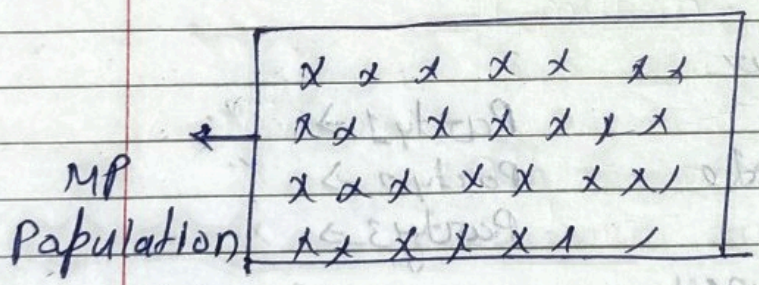
Population vs Sample

1) Population → Population mean → mean, median, mode
2) Sample → Sample mean → mean, median, mode
3) Random Variable → Discrete, Continous, Data

Population

I find out

Avg height of all the people of a state
- MP
- UP

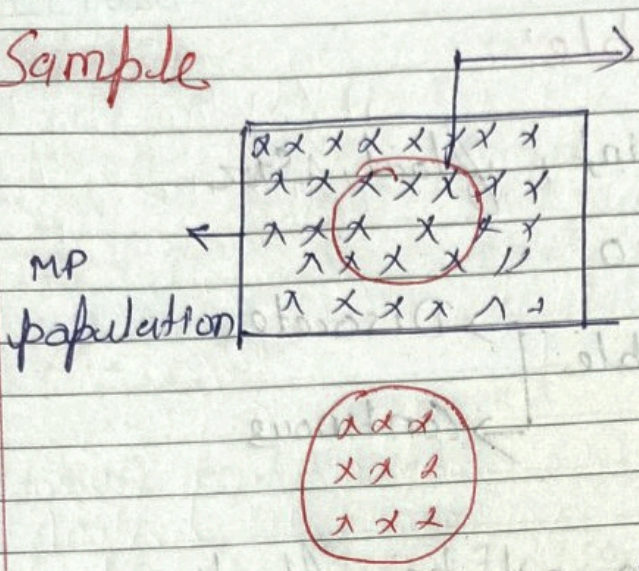


μ = mean = Avg height of all the people

$$\text{population mean} = \frac{1}{1M} \sum_{i=1}^{1M} x_i$$

(13)

Sample

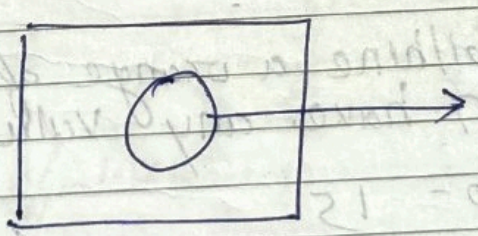


$$\text{Sample mean} = \frac{1}{10000} \sum_{i=1}^{10000} x_i$$

Eg Sample → Exit Polls

→ Party will win the election from the state

→ News channel Exit Pole → Sample Data



option which Party is win

Population mean > Sample mean

① Simple Random Sampling

(17)

Date

Random Variable:

Variable = int, float, str

x = 10

Random Variable → Discrete
Random Variable → Continous

Discrete:-

whole no, not be floating no.

Eg.

① no of bank account

2, 3, 4, 5, ✗ (2.5)

② Population of state

1 million ✗ (1.5) Person

Continous:-

withine a range of value we can have any value

10 - 15

10.1, 10.2, 10.3, 14.9, 15, 14.9

↳ ~~decimal~~ int num is also

Eg: 5.11
5.6

(15)

Date

Measure of Center Tendency

1) Mean
2) Median
3) Mode

{Center Tendency refer to the measure used to determine the "center" of the distribution of data}

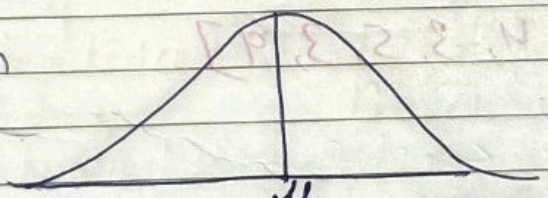
Mean - {Average}

Sample of Height - {168, 170, 150, 160, 182, 140, 175}
x_1 x_2 x_3 x_4 x_5 x_6 x_7

{1, 2, 3, 4, 5}

$$\mu = \sum_{i=1}^{n} x_i \cdot \frac{1}{n} = \frac{\text{Sum}}{7} = 63.5$$

M = Population
n = Sample



measure of Central tendency

Median:

Disadvantage of mean

eg [1, 2, 3, 4, 5] = $\frac{15}{5} = 3$

[1, 2, 3, 4, 5, 50] = $\frac{65}{5} = 13$

$\mu = 3$ → OR $\mu = 3$

$\mu = 13$ → New mean = 13

Very Different

(17)

Date

Median:-

the no of sorted order

1) find the shard no
2) find the centre element

[1, 2, 3, 4, 5, 50]

even no = $\frac{3+4}{2} = 3.5$

[1, 2, 3, 4, 5]

odd no = 3

Mode:-

maximum ure no is selation

[1, 2, 3, 4, 3, 5, 3, 9]

M = 3

Eg Avg

23
24
27
32
35
21
—
—
—

Fill missing value
↳ mean
median
mode

(17)

Measure of Dispersion

Date

Covariance:-

one of the very imp. topic Data preprocessing

I have two Random Variable

Quantify a relation

Size	price
1200 sqr	1000k
1500 sqr	3000k
2000 sqr	5000k

Size ↑↓ price

find the relation b/n Two Variable

S↑ P↑

S↓ P↓

Covariance:-

$$Cov(Size, Price) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)$$

x y

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

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x) \times (x_i - \mu_x)$$

$$\boxed{Cov(x, x) = Var(x)}$$

If

x↑ y↑ = +ve

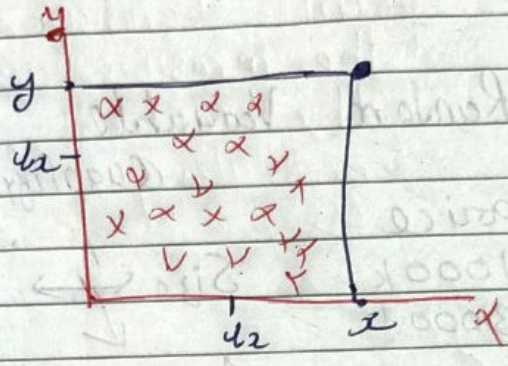
x↓ y↓ = +ve

x↑ y↓ = -ve

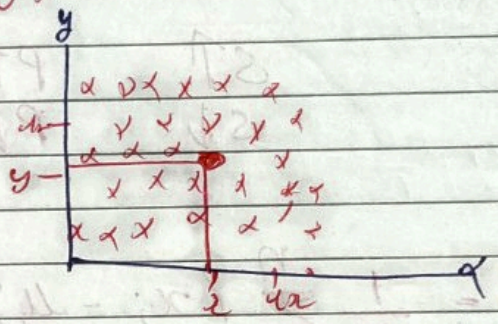
x↓ y↑ = -ve

Date

$Cov(x,y) = (+ve) * (+ve) = +ve$



$Cov(x,y) = (-ve) * (-ve) = +ve$



$Cov(x,y) = (-ve) * (+ve) = -ve$

$Cov(x,y) = (+ve) * (-ve) = -ve$

$x \uparrow \; y \uparrow$ = +ve How much +ve

$x \uparrow \; y \downarrow$ = -ve How much -ve

Standard deviation = $\sqrt{Var}$

19

Date

Pearson Correlation

Co-Variance = $Cov(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y - \mu_y)$

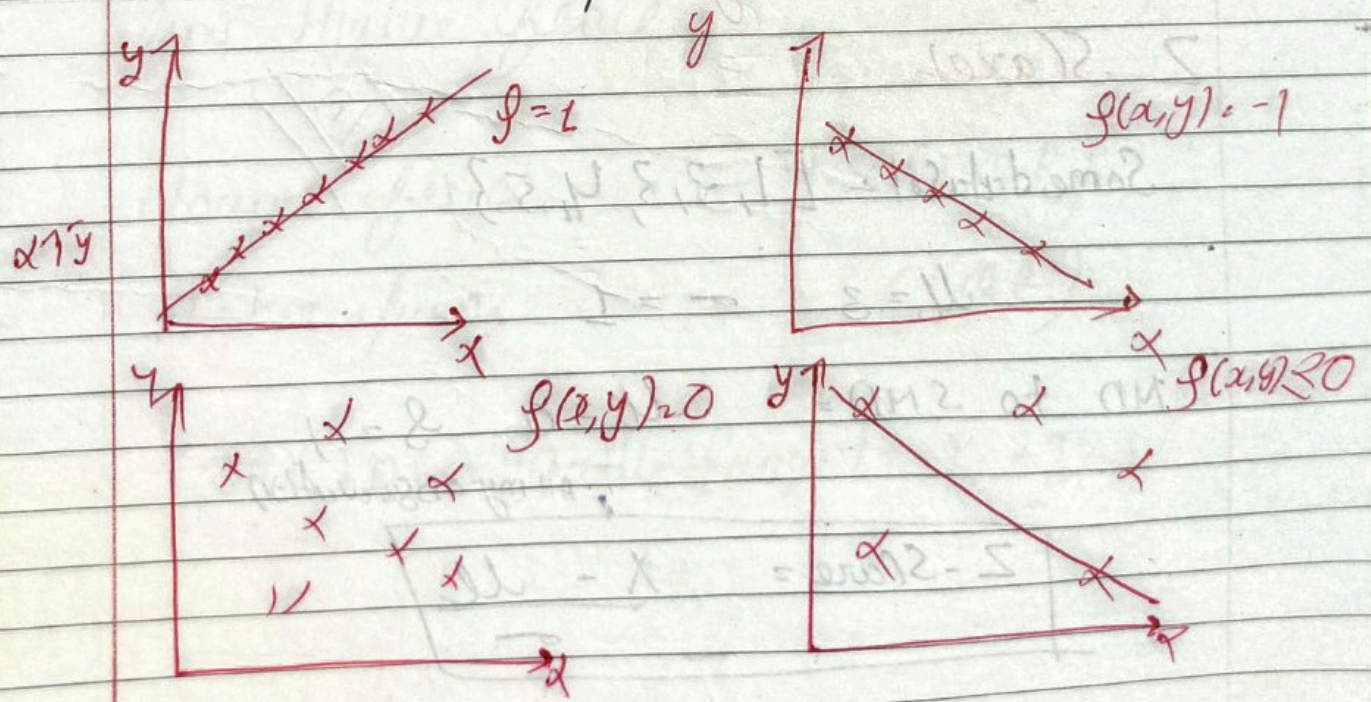
$x\uparrow\ y\uparrow$ = +ve
$x\downarrow\ y\downarrow$ = +ve
$x\uparrow\ y\uparrow$ = -ve

} find the Direction of relationship

Pearson Cov = $\rho(x,y) = \frac{Cov(x,y)}{\sigma x, \sigma y}$

$\rho(x,y) = -1 \leq \rho \leq 1$

$\rho(x,y) = \frac{Cov(x,y)}{\sigma x\, \sigma y}$ $\quad -1 \leq \rho \leq 1$

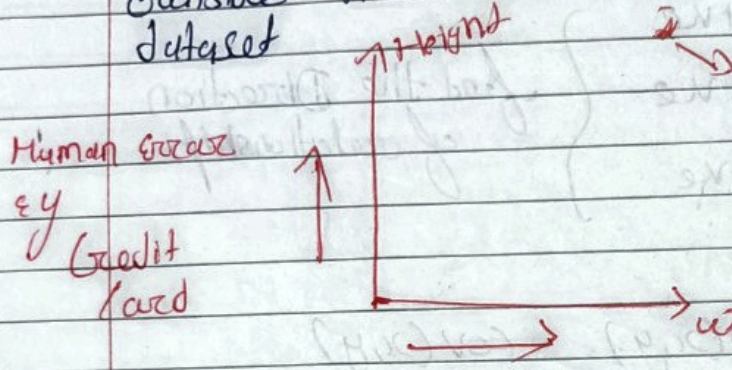


(20)

Date

Outliers

An outlier is a data point in a dataset that is distant from all other observations. A data point that lies outside the overall distribution of the dataset



Identify outlier

1) Z-Score
2) Interquartile Range

Z-Score

Some dataset = [1, 2, 3, 4, 5]

$\mu = 3 \quad \sigma = 1$

ND to SND = $\mu = 0, \quad \sigma = 1$

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

X = my observation

(21)

Date

Interquartile Range

Find Number Summary

1) Minimum
2) First Quartile (Q_1)
3) Median
4) Third Quartile (Q_3)
5) Max

Removing the outlier

{1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 6, 7, 8, 8, 9, 27}

find lower fence

find Higher fence

lower fence = $Q_1 - 1.5(IQR)$

Higher fence = $Q_3 + 1.5(IQR)$

IQR (Interquartile Range) = $Q_3 - Q_1$

(12)

Date

$Q_1 = 25^{th}$ = $\frac{25}{100} \times 19+1$

$\frac{25}{100} \times 20 = 5^{th}$ Index

$\boxed{Q_1 = 5}$

$Q_3 = \frac{75}{100} \times 20 = 15$ Index

$\boxed{Q_3 = 7}$

$\boxed{IQR = Q_3 - Q_1 = 7 - 5 = 4}$

lower fence = $Q_1 - 1.5\ (IQR)$

$5 - 1.5 \times 4$

$3 - 6 = -3$

Higher fence = $Q_3 + 1.5\ (IQR)$

$7 + 1.5 \times 4$

$7 + 6 = 13$

(23)

Date

[lower fence — Higher fence]

[-3 — 13]

what outlier = 27

Min = 1

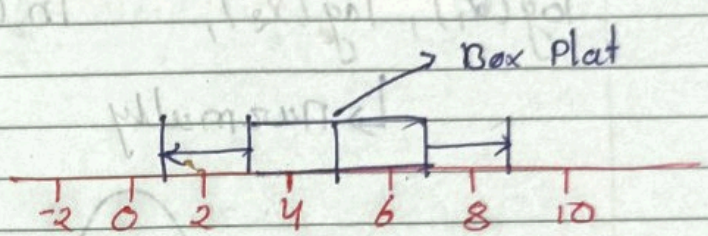
$q_1 = 3$

median = 5

$q_3 = 7$

max = 9

Box Plot

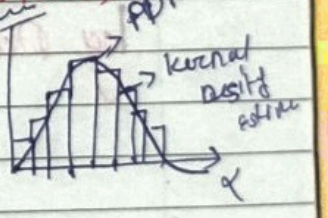


TYPE of Distributions

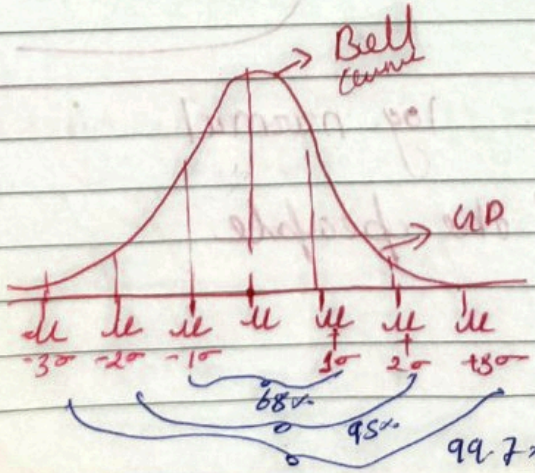
Gaussion Distribution / Normal Distribution

$X \sim N(\mu, \sigma)$

- X ↓ Random Variable
- μ ↓ mean
- σ ↓ SD



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



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

$$\sigma = \sqrt{Var}$$

① Symmetrical

② 68 - 95 - 99.7 %.
Empirical Rule

(24) Central Limit Theorem ⇒ Sample distribution

n ≥ 30

Date

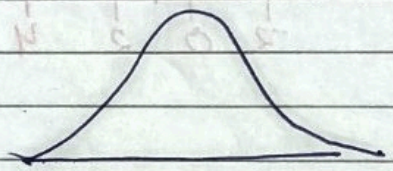
log Normal Distribution

X is log normal Distribution if ln(x) is normally Distributed

$x = \{x_1, x_2, x_3 \text{ ---- } x_n\}$

$\log(x_1), \log(x_2), \ln(x_3)$

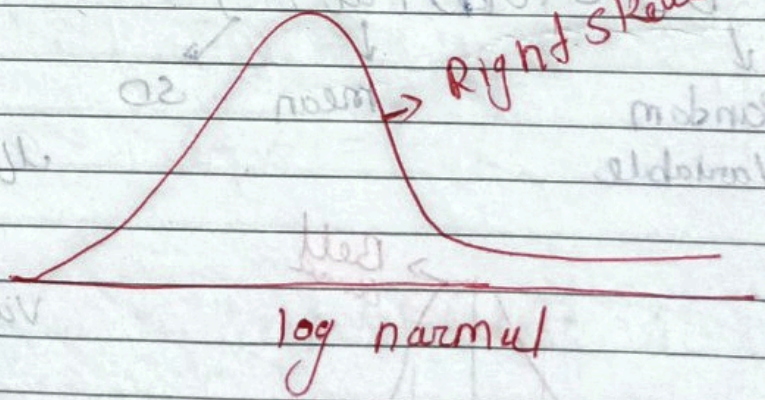
↳ normally



X is log normal Distribution if

ln(x) is N(μ, σ)

log Normal Distribution

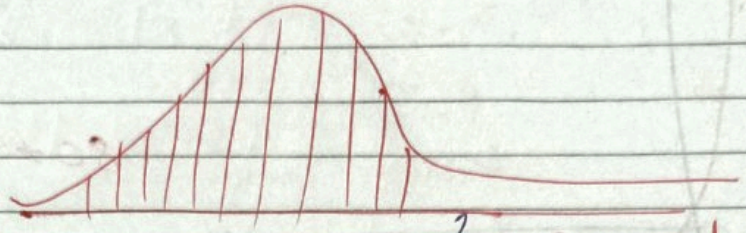


eg :- Income of the people

(2)

Date

Eg: Product Comments (Amazon products)



larger Description very very large

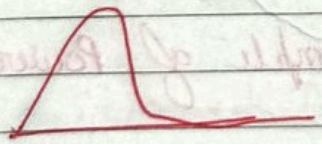
Eg of ND

1) Height

2) Weight

Eg of log ND

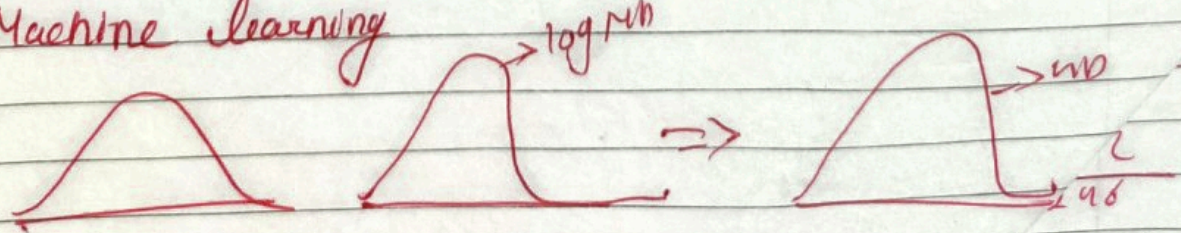
(1)



(1) wealth Distribution

(2)

Machine learning

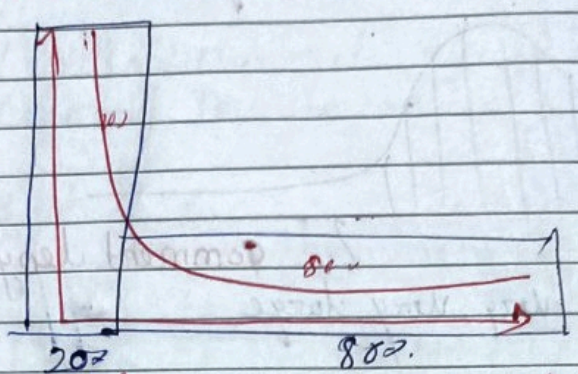


Data Transformation

(20)

Date

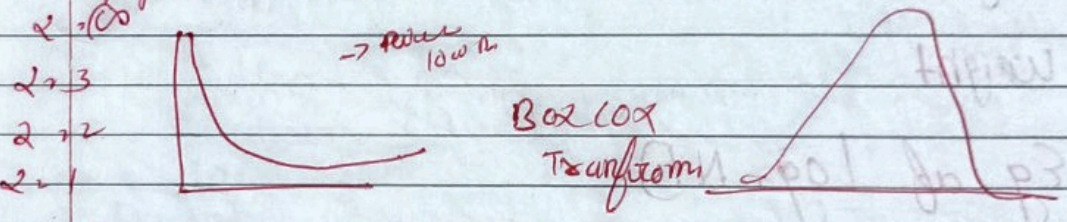
Power Law Distibution



80+20% Rule

Eg 20% of Team is responsable for winning 80% of match

→ 80% of weight distributed 20% of the tail people



Pareto Distibution

- This is a Example of Power law Distibution
- it is not normal Distibution

27

Date

Eg: $x = \{1, 2, 2, 3, 4, 5\}$

x	$\bar{x}$	$(x-\bar{x})^2$	$(x-\bar{x})^2$
1	2.83	-1.83	3.34
2	2.83	-.83	0.6889
2	2.83	-0.83	0.6889
3	2.83	0.17	0.03
4	2.83	1.17	1.37
5	2.83	2.17	4.71
			10.84

$\bar{x} = 2.83$

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

$= \frac{17}{6} = 2.83$

$=$

$\text{Var} = s^2 = \frac{10.84}{5} = 2.168$

↓ Sample Var

$SD = \sqrt{\text{Var}} = \sqrt{2.168}$

Standard Deviation = 1.472

Dataset → $\bar{x} = 2.83$ SD = 1.472

2.83
1.472
1.358

1.472
283
4.302
1.472
5.774
1.472
7.246