

07.01.2023

$X.$	$Y.$	$(X-\bar{x})$	$(Y-\bar{y})$
2	3	-4	-4
4	5	-2	-2
6	7	0	0
8	9	+2	+2
10	11	+4	+4

$$\bar{x} = \frac{2+4+6+8+10}{5}$$

$$= 6$$

$$\bar{y} = 7$$

$(X-\bar{x})^2$	$(Y-\bar{y})^2$	$(X-\bar{x})(Y-\bar{y})$
16	16	16
4	4	4
0	0	0
4	4	4
16	16	16
$\sum (X-\bar{x})^2 = 40$	$\sum (Y-\bar{y})^2 = 40$	

$$\sum (X-\bar{x})(Y-\bar{y})$$

$$= 40$$

$$\begin{aligned} \text{Cov}(x, y) &= \frac{1}{n-1} \sum (x - \bar{x})(y - \bar{y}) \\ &= \frac{1}{4} \times 40 = 10 \checkmark \end{aligned}$$

$$\begin{aligned} \text{Var}(x) &= \frac{1}{n-1} \sum (x - \bar{x})^2 \\ &= \frac{40}{4} = 10 \end{aligned}$$

$$\text{SD}(x) = \sqrt{10}$$

$$\begin{aligned} \text{Var}(y) &= \frac{1}{n-1} \sum (y - \bar{y})^2 \\ &= \frac{40}{4} = 1 \end{aligned}$$

$$\begin{aligned} \text{SD}(y) &= \sqrt{1} = 1 \\ r &= \frac{\text{Cov}(x, y)}{\text{SD}(x) \times \text{SD}(y)} \end{aligned}$$

$$r = \frac{10}{\sqrt{10} \sqrt{10}} = \frac{10}{10} = 1$$

$$-1 \leq r \leq +1$$

Degree of correlation

$r = +1 \Rightarrow$ Perfectly Positive Cor.

$$X + Y = 100$$

$$Y = 100 - X$$

$$r = -1 \checkmark$$

X	Y
0	100
1	99
:	:
100	0

1 unit } 1 unit

$$2x + 3y = 6 \checkmark$$

$$3y = 6 - 2x$$

$$y = 6 - 2x$$

$$y = 2 - \frac{2}{3}x$$

(r)

H.W

$$x = 4$$

x	y	(x - $\bar{x}$)	(y - $\bar{y}$)
0	2		
1	4/3		
3	0		
6	-2		
$\bar{x}$	$\bar{y}$		

Interpretation Part.

Selling Price of the house
depending on size of the
house. (in sq. ft)

Selling Price of the house
= y

Size of the house = x
 $y = f(x) \Rightarrow$ Bivariate.

	SA(f)	Size (sq. ft)
1	y_1	x_1
2	y_2	x_2
...	...	...
10	y_{10}	x_{10}

Goodness fit
statistic

$$r = 0.665$$

(1) True correlation ✓

(2) Highly true " "
 $0.5 < r = 0.66 < 1$

(3) $r^2 = (0.66)^2$
 $= 0.4256$ ✓

Explained variation
 Coefficient of Det-

How much the variation of DV
 can be explained by independent
 variable

r^2

43.5% variation of SP. Can be explained by size of the house -

How much percentage not explained
 $= 1 - r^2 = 56.44\%$

Explained
 is more

unexplained ✓
 obj

$$r = 0.66$$

$$r^2 = 0.4356$$

$$1 - r^2 = 0.5644$$

$$U. Exp > Exp.$$

Multivariate

> 1 independent variables

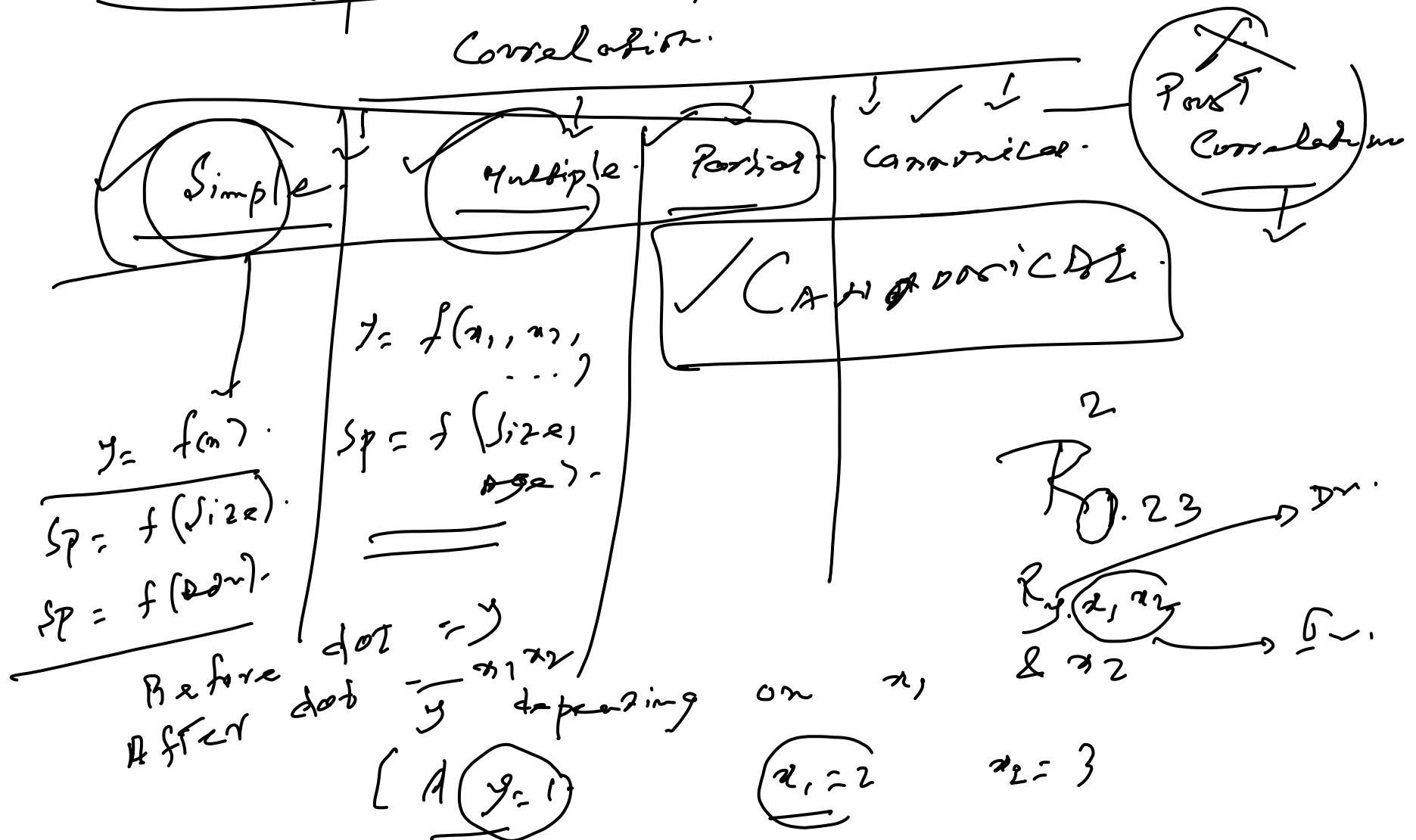
Age of the house

Location

$SP(\gamma)$	$S_{12}(a_1)$	$S_{12}(a_2)$
y_1	x_{11}	x_{21}
y_2	x_{12}	x_{22}
$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$

$$\begin{array}{l} Sp = f(\text{Size}, \text{Age}) \\ \hline y = f(x_1, x_2) \end{array}$$

Correlation.

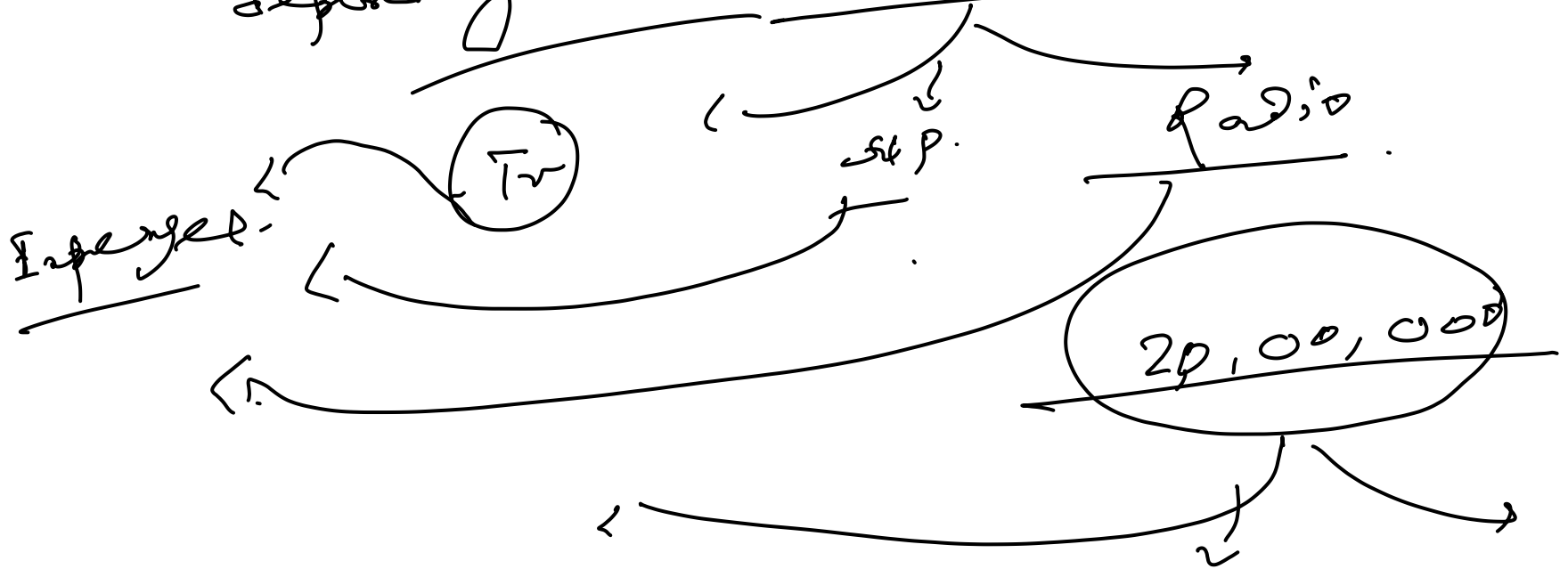


$$R_{123}^2 = 0.91$$

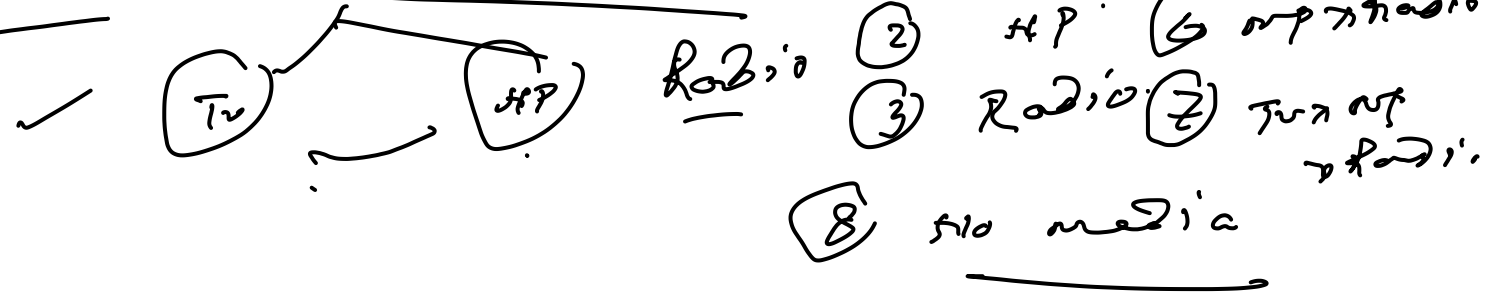
9j.

R²

Selling of any product
depending on ~~Arrangement~~



Selling of any product depending on media.



Sell ~~setting~~ of any prodn in Rs = (Y)

TV Expenses = x_1
 HP " = x_2
 Radio " = x_3

Product $\swarrow$ @ 40 Mkt.

for each Mkt. 200 ~~200~~ 2nd Month data.

1st Mkt.

	Month	y	x_1	x_2	x_3
1	1	✓	0	0	0
2	2	✓	✓	✓	✓
⋮	⋮				
⋮	⋮				
200					

2nd Mkt.

40th yr.

200 rows
4 Col.

$$200 \times 4 = 800 \times 40 = 32000$$

Synergy Effort

$$\left\{ \begin{array}{l} \frac{2,00,000}{2,00,000} \Rightarrow \text{TV} \quad \text{Radio} \\ \frac{2,00,000}{1,00,000} \Rightarrow \text{TV} \quad \text{Sally} \\ \frac{1,00,000}{1,00,000} \Rightarrow \text{TV} \quad \text{John} \end{array} \right.$$

Three independent variables

Association between x_1 , x_2 & x_3 then the
 effect of x_3 is nil.
 $\Downarrow$ Partial Correlation

$r_{12.3}$

$$\underline{\underline{Sally}} = f(\text{TV, Radio, Sally})$$

Simple
multiple
regression

$$\begin{array}{c} y \text{ (Dep)} \quad x_1 \quad x_2 \quad x_3 \\ \quad \quad \quad (Irr) \quad (Irr) \quad (Irr) \end{array} \quad R^2_{1.234} -$$

$$y \text{ (Dep)} \quad x_1 \quad x_2 \quad x_3$$

Selly of any Product = $f(\text{TV, RR, Radio})$

$$y = f(x_1, x_2, x_3)$$

Q1. Is there any linear relationship?

Q2. If linear relation exist, how strong is it?

Q3. Data how we check data set is homogeneous?

Q4. All independent variables are important or not?

Q5. Is there any effect of Pinner?

Q6. Is there any relationship among

the independent variables.
Explore the data. Descriptive.
visualize.

Graphics

Explore the model.

Explore Supervised model (Linear)

① Regression.

↓
Multiple Regression.

Ridge, Lasso

& Elastic
net

② Survival
Regression

Censored
Based

③ Time Series.

↓
Regression
ARIMA.

Explore Regression (non-linear)

↓
Binary
Logistic

↓
Multi
Logistic

↓
Discriminant
Analysis (Linear)

Empirical - Unsupervised model -
↓
do not have the output.
↓
input are given

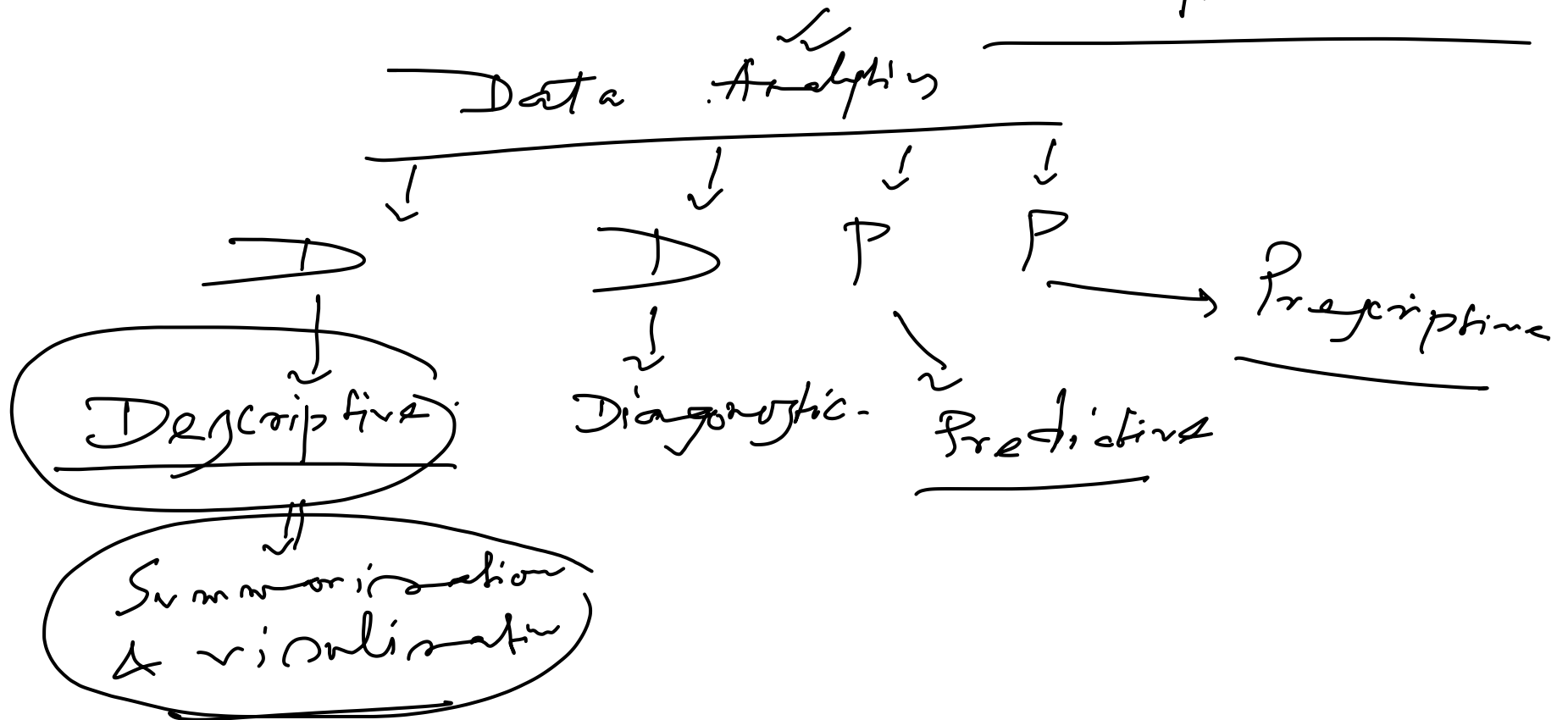
- {
- (1) Cluster Analysis.
 - (2) PCA
 - (3) EFA
 - (4) CFA.

- (1) Probability. (2) Distribution
(3) Testing of Hyp. (4) Anova
parametric

Spss, Jemari, R-Studio, Python.

Soft copy of book

07.01.2023



Set 1. 20 30 40 50 60

✓ $\left. \begin{array}{l} \text{mean} = 40 \\ \text{median} = 40 \end{array} \right\}$

$x = 51$ } $x = 5000$

✓ Set 2 20 35 40 45 60

$\left. \begin{array}{l} \text{mean} = 40 \\ \text{median} = 40 \end{array} \right\}$

$x = 5$

5

2 4 6 8 10

$\left. \begin{array}{l} \text{mean} = 6 \\ (\bar{x} = 6) \end{array} \right\}$

x	$x - \bar{x}$	$(x - \bar{x})^2$
2	-4	16
4	-2	4
6	0	0
8	+2	4
10	+4	16
		$\Sigma (x - \bar{x})^2 = 40$

$$s(x) = \frac{\sum (x - \bar{x})^2}{n-1} = \frac{\sum (x - \bar{x})^2}{4} = 10$$

$$s.d. = \sqrt{s(x)} = \sqrt{10}$$

Indices (P_1) (P_2) $(\bar{x}_1) = \bar{x}_2$

$\begin{pmatrix} 20 \\ 0.8 \end{pmatrix}$

$\{ \}$

$\{ \}$

$\bar{x}_1 = 0$

Coefficient of variation (cm)

$\frac{S.D.}{\text{mean}} \approx 100\%$

(St_1)

(St_2)

Data.

Five Point Summary.

Descriptive (4) Point

① Max ② Q_3 ③ Q_2 ④ Q_1 ⑤ min.
 ① Mean ② SD. ③ Skewness.

④ Kurtosis.

} Box plot
 x_1, x_2, x_3, x_4
 $\downarrow$
min

Gap 1

Max - Q_3 (x_1)

Gap 2

$Q_3 - Q_2$ (x_2)

Gap 3

$Q_2 - Q_1$ (x_3)

Gap 4

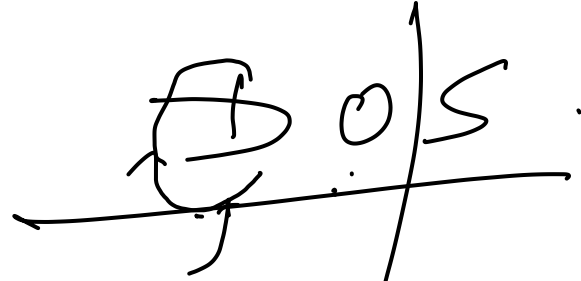
$Q_1 - \text{min}$ (x_4)

10th

$\checkmark$
 C_1

C_2

(mean)
 (S.D.)
 (Skewness)
 (Kurtosis)



Age	20
20-25	21
20-31	23
	25

1-5 1-7 1-9

ordinal.

Sampling.
2.

Summary variable

1 = -
2 = -

Scale

15 ✓

100

Roll no ✓

Sub 1 ✓

Sub 2 ✓

Sub 3 ✓

Gen deri ✓

24 ✓

Male.

Inc. Exp.

Aug. Ad. Pr. Exp.

Female

Data mining Tool

↓ OLAP
on line analytical processing (D)

✓ Diagnostic (or Imp) (D)

- (1) Prob. for life
- (2) Distribution
- (3) Sampling
- (4) Estimation
- (5) Testing of Hyp.

Predictive
Model

Future = $f(\text{Past})$
↓
Predict
Known

Explore the data → Stochastic

Explore the Model

Exploration

Supervised
↓
output → input
↓ ↓
D I

no outputs

unsupervised

Reinforcement

Supervised.

① Linear.

② Non-linear.

Linear. →

Regression

Ridge

Lasso, Elastic

Non-linear. →

Binomial logistic
multi logistic

$D \approx$

$P \approx$

$P \approx$

⑥ - Prescrit.

Reg.

Survival Reg.

Time Series

Excel. →

Software
SPSS →

Tamale

→ R / 12 Studio / Python
Environ.

CI

	15
32-41	3
42-51	3
52-61	2

Frequency

32-41

3

Stem-leaf

Maths

0 1
0
0 1
0
0

outlier

Symmetric

2 4 6 8 10

$$\bar{x} = 6$$

$$s_x = 0$$

normal

Above

6

(2)

Below

6

(2)

(30)

60

2

4

6

8

10

$$\bar{x} = \frac{30 + 60}{6} = 15$$

2

4

6

8

10

(15)

(60)

Bivariate: Two variables:-

$$\bar{x} = 6$$

$$\bar{y} = 7$$

$$\rho_{xy} =$$

$$\phi_{xy}$$

$$\bar{x} \quad \bar{y}$$

(X):

2

4

6

8

10

(Y)

3

5

7

9

11

Bivariate:

Which univariate
& Bivariate measures

(1)

$\bar{x}, \bar{y}, \rho_{xy}, \phi_{xy}$

(2)

Covariance (xy)

$\bar{x} = 6$
 $\bar{y} = 7$

x	y	$(x - \bar{x})$	$(y - \bar{y})$	$(x - \bar{x})^2$	$(y - \bar{y})^2$	$(x - \bar{x})(y - \bar{y})$
2	3	-4	-4	16	16	-16
4	5	-2	-2	4	4	-4
6	7	0	0	0	0	0
8	9	+2	+2	4	4	4
10	11	+4	+4	16	16	16
				$\sum (x - \bar{x})^2 = 40$	$\sum (y - \bar{y})^2 = 40$	$\sum (x - \bar{x})(y - \bar{y}) = 40$

$$Cov(x, y) = \frac{1}{n-1} \sum (x - \bar{x})(y - \bar{y})$$

$$= \frac{40}{4} = 10$$

$$\Delta x = \sqrt{10}$$

$$\Delta y = \sqrt{10}$$

Size
Square:

X unit	Y unit (17) \$

$$-\infty < X < +\infty$$

$$-\infty < Y < +\infty$$

$$-\infty < \text{Cov}(XY) < +\infty$$

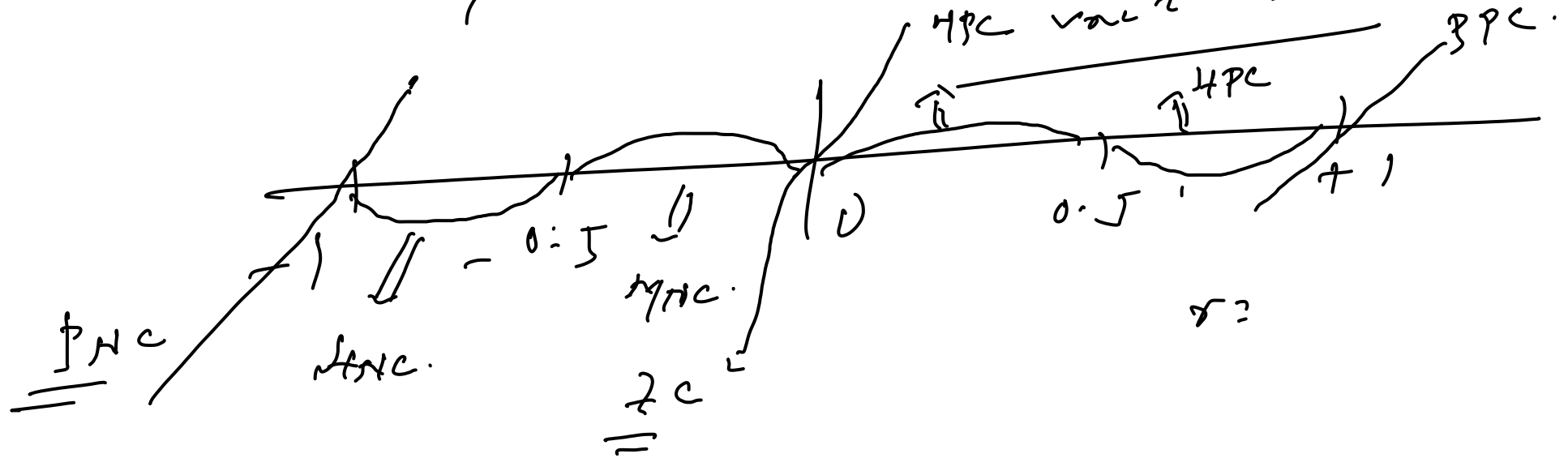
$$\text{Cov}(XY) = 800$$

X is free
not unit free

$$\gamma = \frac{\text{Cov}(XY)}{\sigma_X \sigma_Y}$$

$$-1 \leq \gamma \leq +1$$

HPC unit free



r^2 = Coefficient of Determination
 = Explained variation

$1-r^2$

→ How much variation

r^2 // of the div explained by
independent variables / s.

← Bivariate → Multivariate

Simple

✓ sp

① - ①

size

① - ②

Age

Age

{ } { }

r^2

$r^2 = 0.66$

34

$r^2 = 0.42$
95

✓ $\hat{Y} = f(x_1, x_2, \dots, x_p)$

unsupervised

Diabetes.

1,13,000

(181) no variables

① Dropping of less important variables

② checking of outliers

③ Normalization

④ checking linearity

(By Scatter Diagram)

value of $\bar{Y}$
Variable - $\frac{\max - \min}{t-1}$
max - min
t = 2, 3, \dots

⑤ Homogeneity

⑥ low multi-collinearity

⑦ low auto-correlation

Low Corr. among the x_i 's

$\tilde{W} \tilde{W}^T H + P^T$

DDPP

$\tilde{W} \tilde{W}^T H$

Reg. $\rightarrow$ what is Reg?
why Reg?
How? $\Rightarrow$ SP - ...
 (12)
 $p =$ Perplexity
 $p =$ Patience
 $p =$ Branch
 $p =$ Perf
 (2) = (5)

Prediction Technique using

12.02.2023.

Probabilities -

(22 Imp)

Small Case.

Age group.	No. of Covid patients	No. of death
(G ₁) < 30	5000	5
(G ₂) 30 - 50	3000	15
(G ₃) 50 - above	2000	40
	10000 = N.	60

Total no of deaths = 60 out of 10000

$$\therefore P(D) = \frac{60}{10000} = \frac{6}{1000} = 0.006 \checkmark$$

$$= \frac{FHC}{THC}$$

$$P(G_1) = \frac{5000}{10000} = \frac{1}{2} = 0.5 \quad P(G_2) = \frac{3000}{10000} = \frac{3}{10} = 0.3$$

$$Q \quad P(G_3) = \frac{2000}{10000} = \frac{2}{10} = \frac{1}{5} = 0.2$$

Information. are known

$$(1) P(D) = 0.006$$

$$(2) P(G_1) = 0.5 \quad (3) P(G_2) = 0.3$$

$$(4) P(G_3) = 0.2$$

Prob.

Unconditional.

(when no
condition is
given)

$$P(D) = 0.006$$

Conditional

(when condition
is given)

$\Downarrow$

$$\begin{cases} P(D/G_1) = \frac{5}{5000} = \frac{1}{1000} = 0.001 \\ P(D/G_2) = \frac{15}{3000} = \frac{1}{200} = 0.005 \\ P(D/G_3) = \frac{40}{2000} = \frac{1}{50} = 0.02 \end{cases}$$

How we get

Not known

$$P(G_1 / D) = ?$$

$$P(G_2 / D) = ?$$

$$P(G_3 / D) = ?$$

Prob.

Prior
(which are known)

Posterior
(which are not known
calculated or estimated
on the basis
of prior
Prob.)

Bayes Th.

TABLE Showing THE CALCULATION OF
POSTERIOR PROBABILITY.

Prior				
No of Group (i)	(1)	(2)	(3)	(4) (2) x (3)
G_i	$P(G_i)$	$P(D/G_i)$	$P(G_i)P(D/G_i)$	$P(G_i/D)$
$i=1,2,3$				
G_1	$\frac{1}{2} = 0.5$	$P(D/G_1) = 0.001$	0.0005	$\frac{0.0005}{0.006} = 0.083$
G_2	$\frac{3}{10} = 0.3$	$P(D/G_2) = 0.005$	0.0015	$\frac{0.0015}{0.006} = 0.25$
G_3	$\frac{1}{5} = 0.2$	$P(D/G_3) = 0.02$	0.004	$\frac{0.004}{0.006} = 0.67$

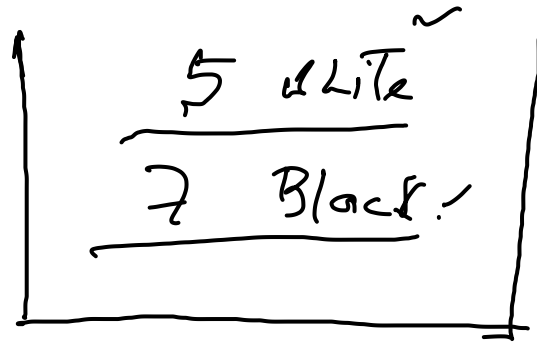
(C) $\sum P(G_i) = 1$ $\sum P(D/G_i) = 0.026$ $\sum P(G_i)P(D/G_i) = 0.006$ $\sum P(G_i/D) = 1$

$\sqrt{10000} = \frac{60}{10000}$ $\frac{515}{10000} + \frac{40}{10000} = 0.026$ $= \frac{0.006}{0.026} = P(D)$

Bayes' Theorem

$$P(G_i/D) = \frac{P(G_i)P(D/G_i)}{\sum P(G_i)P(D/G_i)}$$

unknown



Box

Assumption All Balls are

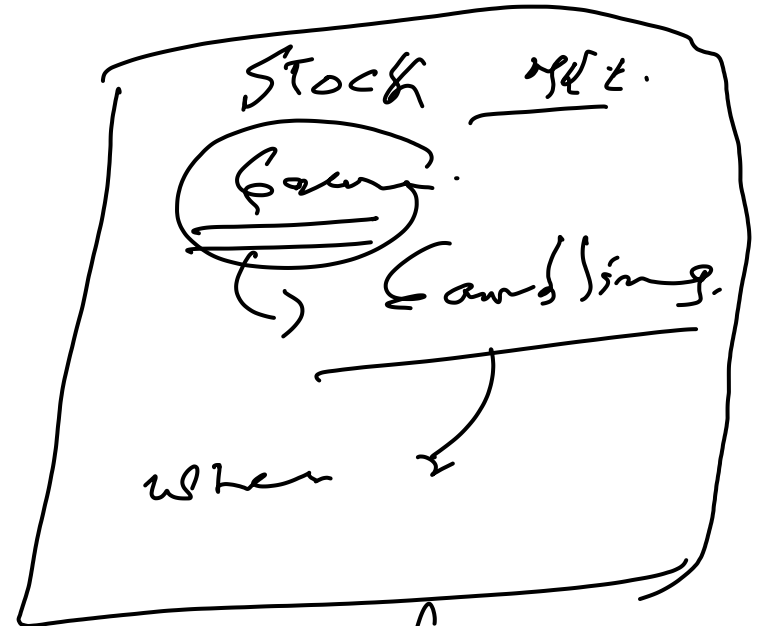
(1) equal wt.
(2) shape of the balls

Total balls = 12

Some condition is given -

1 White

1 black



are uniform

3 balls.

there is a gain of Rs 10
" there is a loss of Rs 7.

Since 3 balls are drawn
What are the possibilities?

units: (R 30)

- $\left\{ \begin{array}{l} \textcircled{1} \cdot \text{All 3 balls are} \\ \textcircled{2} \cdot 2 \text{ w + 1 B.} \\ \textcircled{3} \cdot 1 \text{ w + 2 B.} \\ \textcircled{4} \cdot 3 \text{ B.} \end{array} \right.$

$$\binom{20-21}{=13}$$

symmetrically
Exclusive.

$$= 10 - 2 \times 7 = -4$$

$$= -21$$

$$3 \times 2 \times 1$$

$$n_4 = -21$$

$$p_4 = \frac{35}{220}$$

$$\binom{12}{3}$$

$$= \frac{12 \times 11 \times 10}{3! \times 6}$$

$$= \frac{270}{6}$$

$$n_1 = +30$$

$$p_1 = \frac{10}{220}$$

$$n_2 = +13$$

$$p_2 = \frac{70}{220}$$

$$n_3 = -4$$

$$p_3 = \frac{105}{220}$$

$$h(3w)$$

$$= \frac{\binom{5}{3}}{\binom{12}{3}}$$

dat. X be a r.v. (random variable) which can take the values x_1, x_2, x_3 & x_4 with probability p_1, p_2, p_3 & p_4 respectively.

$\left\{ \begin{array}{l} X: \\ p: \end{array} \right.$	$x_1 = 30$	$x_2 = 13$	$x_3 = -4$	$x_4 = -21$
	$p_1 = \frac{10}{220}$	$p_2 = \frac{70}{220}$	$p_3 = \frac{105}{220}$	$p_4 = \frac{35}{220}$

$p_1 + p_2 + p_3 + p_4 = 1$

Discrete Probability Distribution
which is Countable.

Information

• Last 60 months stock (return) given.

✓ Step ①. Find five number summary

	Φ_1	Φ_2	Φ_3	max.
min				

Step ② (a) no of 60 month data. how fall in
many data Φ_{111} min - $\Phi_1 = f_1$

(b)	How many data	$\Phi_1 - \Phi_2 = f_2$
(c)	" "	$\Phi_2 - \Phi_3 = f_3$
(d)	" "	$\Phi_3 - \text{max} = f_4$
		$f_1 + f_2 + f_3 + f_4 = \underline{\underline{60}}$

min = ✓

We take the mid value of ϕ_1 & min.
$$\frac{\min + \phi_1}{2} = x_1$$

Mid value of ϕ_1 & ϕ_2
$$\frac{\phi_1 + \phi_2}{2} = x_2$$

Mid " " ϕ_2 & ϕ_3
$$\frac{\phi_2 + \phi_3}{2} = x_3$$

Mid " " ϕ_3 & max
$$\frac{\phi_3 + \max}{2} = x_4$$

X: $x_1 = \checkmark$ $x_2 = \checkmark$ $x_3 = \checkmark$ $x_4 = \checkmark$
f: $f_1 = \frac{f_1}{60}$ $f_2 = \frac{f_2}{60}$ $f_3 = \frac{f_3}{60}$ $f_4 = \frac{f_4}{60}$

Expected Return $\begin{cases} +ve. (\text{Gain}) \\ -ve. (\text{Loss}) \end{cases}$

$\sqrt{\text{Variance (Risk)}} = \text{Square root of}$
variance = Standard deviation

$$= \text{Risk}$$

$$X: \left. \begin{matrix} x_1 R & x_2 R & x_3 R & x_4 \\ p_1 & p_2 & p_3 & p_4 \end{matrix} \right\} \text{Intensities}$$

$$E(X) = x_1 p_1 + x_2 p_2 + x_3 p_3 + x_4 p_4$$

$$= 30 \times \frac{10}{220} + 13 \times \frac{70}{220} - 4 \times \frac{105}{220} - 21 \times \frac{35}{220}$$

$$= \frac{1}{220} [300 + 910 - 420 - 735]$$

$$= \frac{1}{220} [1210 - 1155] = \frac{55}{220} = \frac{1}{4} = 0.25$$

$$V(X) = E(X^2) - (E(X))^2$$

$$E(X) = 0.25$$

$$E(X^2) = x_1^2 p_1 + x_2^2 p_2 + x_3^2 p_3 + x_4^2 p_4$$

$$= (30)^2 \times \frac{10}{220} + (13)^2 \times \frac{70}{220} + 16 \times \frac{105}{220} + 441 \times \frac{35}{220}$$

$$= \frac{\quad}{\quad} \text{ st. w.}$$

$$E(R^2) - (E(R))^2 = \sigma(R)$$

$$\sqrt{\sigma(R)} = SD(R)$$

Stock-1.

Open

closing

Return

R_{it}

60 months

month 1

2

.

.

.

60



120 months



We apply Two formula for the
 calculation of ROR [Rate of
 return]

ROR.

When the variation
 of two consecutive
 return is small

$$\frac{R_n - R_{n-1}}{R_{n-1}}$$

$$\frac{\text{Current} - \text{Previous}}{\text{Previous}}$$

When the variation
 of two consecutive
 return is high

$$\ln \frac{R_n}{R_{n-1}}$$

natural logarithm

$$ROR = \ln \frac{R_n}{R_{n-1}}$$

Base of log is
 e.

$$R_n > R_{n-1} \quad ROR = +ve$$

ROR.

$$R_2 = 5 \quad Q_{2-5} = 3$$

R_1

$$R_2 \rightarrow \begin{aligned} & \ln R_2/R_1 \\ & \ln R_3/R_2 \end{aligned}$$

(59)

Data

$$\frac{R_2}{R_{2-1}} = \frac{5}{3} = 1.66$$
$$\ln 1.66 \approx 0.507$$

60

$$R_{60} \rightarrow \ln R_{60}/R_{59}$$

NYSE 100	BSE 100
----------	---------

(1) calculate ROR for each stock for last 60 months / n months.
 $n \geq 60$

(2) Calculate AROR
= Average rate of return

On the basis of AROR values.
We can discard same stock
(which will give no AROR)

out of 100. $\rightarrow$ 60 are remaining.

We Calculate S.d. (ROR) on 60 stocks

Ans Ans
$$S.d. (ROR) = \sqrt{\frac{1}{K-1} \sum (ROR - AROR)^2}$$

Collect Sample 30 }
2023 Feb

2016 Jan

2020 year

2020 Feb

plan ①

plan ②

30 stocks

2023 Jan

NSE 100

✓ ROR, AROR, $SD(ROR)$

ROR, AROR,
 $SD(ROR)$

✓ Five point Summary

$CV(ROR)$
Coefficient of variation.
 $= \frac{SD(ROR)}{AROR} \times 100\%$

min :
Q1 :
Q2
Q3
Max

Port folio

Prob. Dist-
 $E(x)$ $SD(x)$



Money Control. Gov, Yahoo Finance

