18.12.2022

Statistics.

Statistic + S

Sample.

Why.

To estimate population.

(1) Population. (2) Parameters (3) Sample.

(4) Statistic.

Big. 140 Crores

30 Persons

1 → 200

2 → 400

Time

Cost = f(T)

PW.

Parameter: ⇒ Characteristics of the pop?.

N = 1000

Avg. income of 1000 employees.

$x_1 + x_2 + \dots + x_{1000} / 1000$

N

population ⇒ mean → Parameter

Sample. → Small (N)

1000 → 100 employees

Avg income of 100 employees

Statistic → Characteristic of the Sample.

Parameter → Pop

N ⇒ Pop size
μ ⇒ Pop mean. Parameter
n ⇒ Sample size.
x̄ ⇒ Sample mean. Statistic

Set 1: 20 30 40 50 60 (10 10 10 10)

n = 5

mean = 40
median = 40

Why?

Set 2: 20 35 40 45 60 (15 5 5)

mean = 40
median = 40

5000 obj. → Stock.

10 yrs. data. daily stock.

1 - 365 → (2) 240 x 14 = 2400

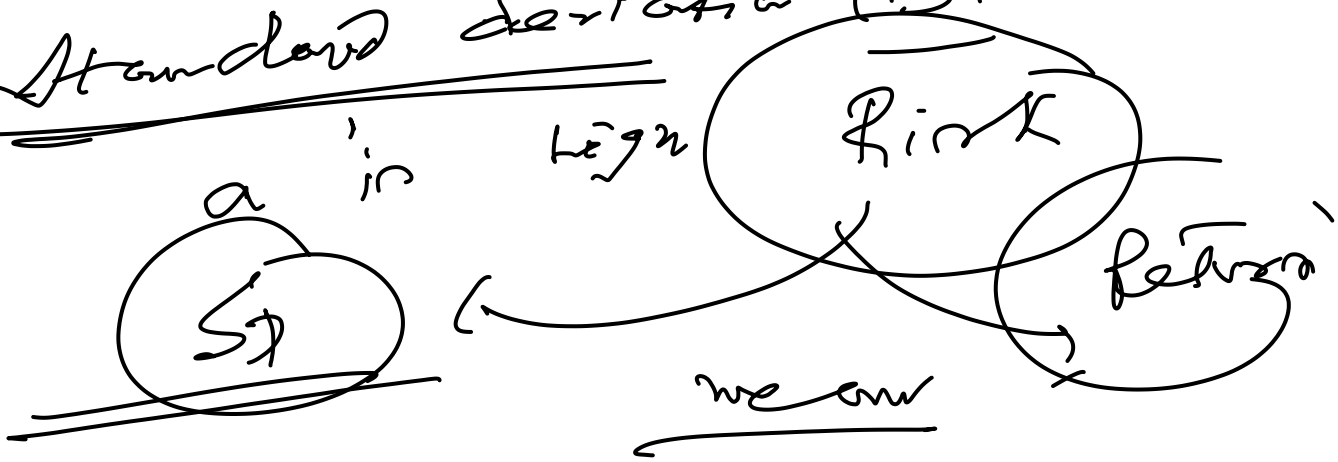
Stock 1 - 2400 days closing price

Avg return: $\frac{\text{Total}}{2400} =$

0 0 0 . . . 0

Standard deviation (S.d)

a in Sign



we can

(1) mean (2) S.d.

$$\checkmark \quad S.d. = \sqrt{\frac{1}{n}\sum (x_i - \bar{x})^2}$$

= Root - mean square deviation about mean

$(S.d)^2$ = variance

(X) → Cm

Sd(X) → Cm

Var(X) → Sq Cm

Balance sheet

C_1	C_2
x_1	c_{11}
x_2	c_{21}
x_3	c_{31}
x_4	c_{41}
x_5	c_{51}

Need for test
Give x_1 x_2 x_3 x_4 x_5

C_2
c_{12}
c_{22}
c_{32}
c_{42}
c_{52}

Automobile
Homogeneous

Step 1 mean
Step 2 S.D.
Step 3 Skewness
Step 4 Kurtosis

Coefficient of variation = $\frac{\text{mean}}{\text{S.D.}}$

[Departure from the symmetry]
[Peakedness of the frequency distribution]

Descriptive measures

Variance = Square of S.D.

Range is the measure.

SD = S. variance = S^2

$\sqrt{S^2} = (\pm)S$.

2nd Part: SPSS.

Description of the data.

In a class, there are 15 Students (sample size) composed of (gender), male & female. tion given the marks in three subjects, full marks 100, namely 1) Statistics 2) Research 3) Operations

Ques. (1) Avg marks for each student.

(2) How does Compute

frequency distribution of the marks of different subjects

(3) How can we explore data analysis ?

(4) How we Compare the Performance level of marks gender wise.

DODO

D ← Description
O ← ordering
D ← Distance
O ← Origin

Gender	Description	ordering	Dist	origin
Gender	✓	X	X	X
	✓	✓	X	✓
	✓	✓	✓	✓

Ø

Ans

A = 23

Nature of Data.

1) Nominal (D) — 2) ordering (D+O) — 3) Interval Scale (D+O+D) — 4) Ratio scale (D+O+D+O)

Categorical

grouping.

Non metric

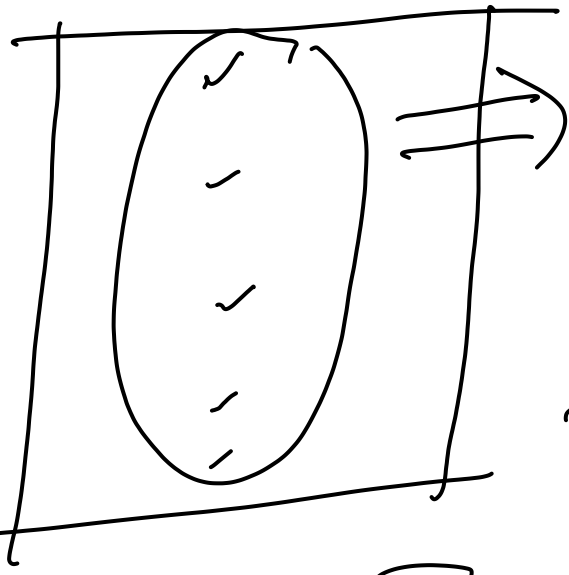
Non Categorical

Summary.

metric

DO | DQ → DQS

Categorical — Non-Categorical — Scale



Min value.

Maximum value

Range = Max - Min

Raw data.

Grouped data

Aggregate

Step. Frequency dist.

class

class notation

class Boundary

32-42

42-52

10-19
20-29
30-39
} ⇒ C.I. [Discrete gap]

10 19 20 29 30

Exclusive

Include

10-20
20-30
30-40
} ⇒ CB. 32-41 - (3)

41-42.

How we get freq dist. for CI ??

Range = max - min

Given information

Not known

1) (no of classes)
2) width (gap, difference between lower & upper values of a class)

Formula: (no of classes) × (width of the class) = Range

User will supply or according to his/her knowledge domain know the width of the class

Automatically will come no. of classes

1/2 = 0.5

Height
10-19
20-29
30-39

50-59
60-69
70-79

49.5 - 59.5
vg.

32 - 41.

42 - 51.

52 - 61.

62 - 71.

72 - 81.

82 - ✓

< 32 → 1

32 - 41 → 2

42 - 51 → 3

52 - 61 → 4

62 - 71 → 5

72 - 81 → 6

82 - - -) 7

gautam.

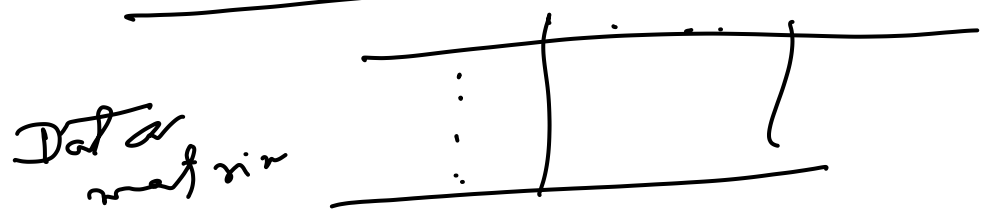
GAUTAM. BANDYOPADHYAY.
CDMS. NITDGP. AC. IN

31.12.2022

EXPLORE.

EDA.

Application Explore Data Analysis.



1) Row indicates no of cases i.e. Sample size.

Column indicates no of variables / Attributes / features.

Gbl: 15 rows

Roll no. sub1 sub2 sub3 Gender

(15 × 5) = 75 cells value.

1) Descriptive

2) Individual. Group analysis

✓ Stem & Leaf.

(3) Box-plot. EDA.

Stem & Leaf.

32 - 41. — (3)

32 32 32 32 34 40

Box plot

Five Point Summary

(1) Maximum.
(2) Q_3 (3) Q_2 (4) Q_1.
(5) Minimum.

(6) Box plot.

Total motor Bank.

60 months.

min Q_1 | max Q_3 Q_2

Gap 1. => Q_3 & Q_2.
Gap 2. => Q_2 & Q_1.
Gap 3. => Q_1 & min.
Gap 4. =>

Gap 1. > Gap 2.

Assignment :-

(1) Collect from moneycontrol.com.
Stock prices of Reliance industries.
from: 2018 Jan to 2022 Nov.
59 months.

(2) Calculate. Rate of return. (ROR)

$\frac{P_n - P_{n-1}}{P_{n-1}}$

Jan ✓

RoR:

Feb. ✓ → $\frac{Feb - Jan}{Jan}$

59 ✓

$\log \frac{P_n}{P_{n-2}}$

$\frac{P_2}{P_1}$

RoR

Jan	2	
Feb	4	$(4-2)/2$
Mar	6	$(6-4)/4$

$\log \frac{4}{2}$

$\log$ returns are used in R.

Step (3). Explore the data set.

- mean, median, S.d, variance, min, max, Range, inter-quartile range, Skewness & Kurtosis.

Step (4) Calculate Stem & Leaf.

Step (5) Box plot. (SPSS. Excel.)

Test.

1) How many data lies between mean & Q_3. ? => Gap1

2) How many data lies between Q_3 & Q_2 ? => Gap2

3) How many data lies between Q_2 & Q_1 ? => Gap3

4) How many data lies between Q_1 & min ? => Gap4

Next Part.

Two stocks from same sector.

50 months

	Max = Q3	Q3 − Q2	Q2 − Q1	Q1 − min
Stock 1	$20	f_2	f_3	f_4 ...
Stock 2	50	–	–	– –

upward

Downward.

classify based on

binary.

Based multi layers

Binary

Good

Poor

Stock

Good.

Avg.

Poor.

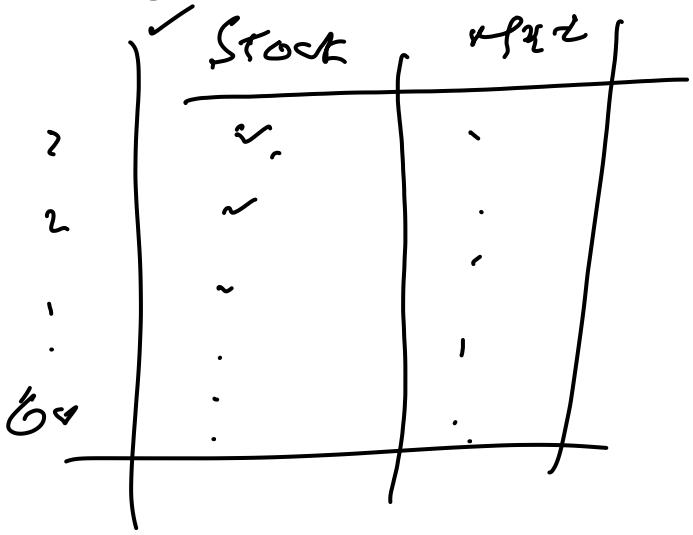
3 stage

Stock.

Market → Sensex

Go wants → Stock Price (Closing), Market Price.

Go wants



Step (1): Find ROR for individual stock & market.

Step (2): Find AROR for stock & market

Step (3): If AROR for stock > AROR for Mkt ⇒ Good.

Otherwise Poor.

Step (4): SD(ROR) for stock / SD(ROR) for Mkt.

If SR(ROR) for stock < SD(ROR) for Mkt ⇒ Good.

Otherwise Poor.

In stock MKT ARoR.

⇒ Return.

SD (RoR) ⇒ Risk.

obj: To maximize return & minimize risk.

Coefficient of variation

(CV) = $\frac{SD}{mean} \times 100\%$.

⇓

used for comparision between same items. (homogeneous same type

Good Poor. (Two)

Good Average Poor (Three)

Max Q_3 Q_2 Q_1 min.

Method 1.

max Q_3 Q_2. → Good.

60 months   mean

30	max – Q3	40	=> Good
10	Q3 – Q2		
10	Q2 – Q1	20	
10	Q1 – min		

60 $\frac{40}{60}$

$2/3 = 66\%$

60

Modelling.

	Good	Poor	
Good	30	20	40
Poor	5	15	20

Objects – Past

Model – Future / Past

Good – Good (correct)

Good – Poor (wrong)

Poor – Good (wrong)

Poor – Poor (correct)

60

No – No

YES – YES

	NO	YES	
NO	50	10	60
YES	10	30	40
			100

Past

Testing of Hyp.

Buyer.

Seller

(2)

Reject the good lot. = Error

Accept a bad lot = Error

Accept the good lot

reject the bad lot.

	Accept	Reject
Good.	Correct	Error
Poor.	Error	Correct

Obj How to min. error

Describing. Stock Prices.

AROR.

SD [ROR].

Mkt Price

AROR

SD (ROR).

Case 1: AROR for stock > AROR of NIFTY
& SD(ROR) for stock < SD(ROR) for NIFTY
⇒ Good.

Case 2: AROR for stock > AROR of NIFTY
& SD(ROR) for stock > SD(ROR) for NIFTY
⇒ Average.

Case 3: AROR for stock < AROR of NIFTY
& SD(ROR) for stock > AROR of NIFTY
⇒ Poor.

Data mining: used for comparison between among stocks.

OLAP:
= On line Analytical Processing.

OLAP.

Two types of variables.

(1) Summary variable [Scale]

(2) Grouping [Categorical]

Sub1, Sub2, Sub3 ✓

Gender

		Marks	Grade
Male / Female	Sub1	✓	✓
	Sub2	✓	✓
	Sub3		

01.01.2023

Gender → (2) ✓

Background → (3) ✓

Locality → (2) ✓ UG PG

Education

Hobby → (5) ✓

Total Combination = $2 \times 3 \times 2 \times 5$

= (60)

60 Reports

1971.

Product → Comp

Brand	B_1	B_2	B_3 (3)
Price	P_1	P_2	P_3 (3)
Color	C_1	C_2	C_3 (3)
Goodwill	Y	N	(2)
Warranty	Y	N	(2)

Q.

$3 \times 3 \times 3 \times 2 \times 2 = 108$

108 minimum reports.

OLAP: Summary ✓

Conjoint

Grouping variables

Summ

On line analytical processing

mean $\bar{x} = \frac{1}{n}\sum x_i$

S.d. $\sigma = \sqrt{\frac{1}{n-1}\sum (x_i - \bar{x})^2}$ } Univariate. Single variable

{ Demand of cold drinks (y) depending on Temp (x) } Bivariate.

y (o) x

y Demand of cold drinks depending on Temp (x_1) and Profit on economic ——— (x_2) Multivariate

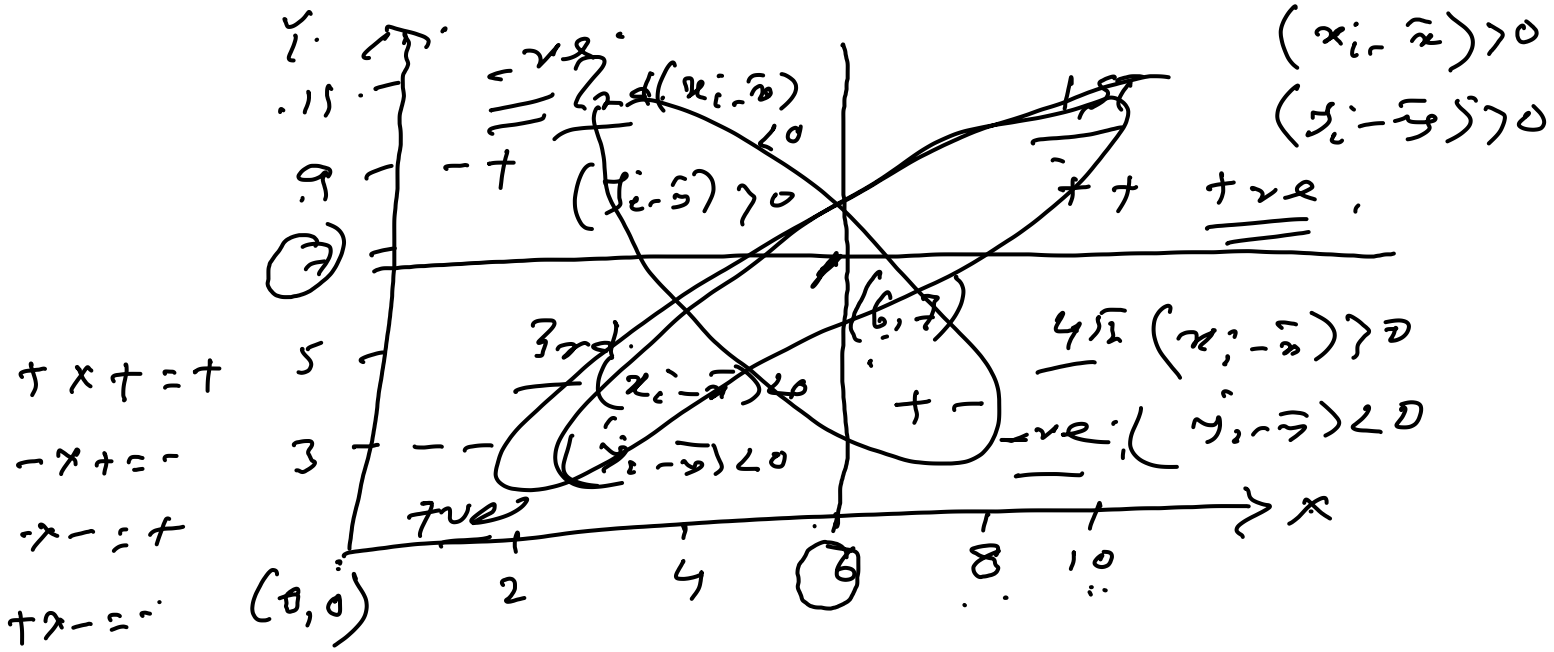
Dr.	Ir	
1	1	Bivariate.
1	>1	Multivariate
>1	>1	Multivariate

Stock Mkt.

$n=5$

$\Sigma x = 30 \quad \bar{x} = 30/5 = 6$

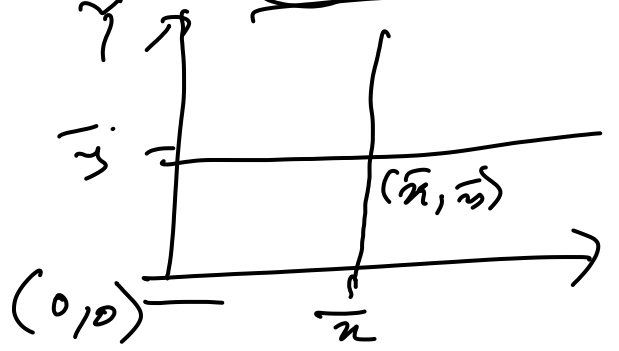
X:	2	4	6	8	10	$\bar{x} = 6$
Y:	3	5	7	9	11	$\bar{y} = 7$



$$\frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{n-1}$$

$$= \text{Covariance}(x, y)$$

Covariance is the transformation of the origin $(0,0)$ to $(\bar{x},\bar{y})$ i.e, xy plane into $(x-\bar{x})(y-\bar{y})$



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

(X): ...

(Y): ...

Cov(XY) = +ve ⇒ +ve association

= −ve ⇒ −ve "

= 0 ⇒ no association

$-\infty < X < +\infty$ $\quad -\infty < Y < +\infty$

$\therefore -\infty < Cov(XY) < +\infty$

$-\infty < Cov(XY) < 0 \Rightarrow$ −ve association.

$0 < Cov(XY) < \infty \Rightarrow$ +ve association.

X (Temp in °C) — Y (Rs in ₹)

Cov(YX) = 0 Rs °C

⇒ not unit free.

$Cov(XY) \neq 0$.

Degree of association

Sign ~~among~~ Four

(1) very large value.

(2) not unit free.

(3) Degree of Association

Q.4.

x :	2	4	6	8	10.	$\bar{x} = 6$
$(x - \bar{x})$:	−4	−2	0	+2	+4.	
$(x - \bar{x})^2$	16	4	0	4	16	

$$\sum (x - \bar{x})^2 = 40$$

σ^2 = variance. $= \frac{1}{n-1} \sum (x - \bar{x})^2$.

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

σ = S.D. $= \sqrt{10} = 3.16$ (approx)

X: 2 4 6 8 10 ~ $\bar{x}$, σ_x

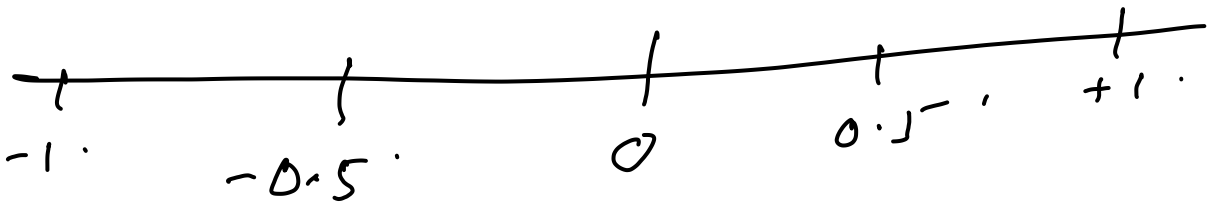
Y: 3 5 7 9 11 $\bar{y}$, σ_y

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

$$\gamma_{x,y} = \frac{Cov(x,y)}{\sigma_x \sigma_y}$$

Limits of γ: $-1 \le \gamma \le +1$

(1) unit free. (2) $-1 \le \gamma \le +1$



(1) $\gamma = -1 \Rightarrow$ PNC.

(2) $-1 < \gamma < -0.5 \Rightarrow$ SNC.

(3) $-0.5 \le \gamma < 0 \Rightarrow$ WNC

(4) $\gamma = 0 \Rightarrow$ ZC

(5) $0 < \gamma < 0.5 \Rightarrow$ WPC.

(6) $0.5 \le \gamma < 1 \Rightarrow$ SPC

(7) $\gamma = +1 \Rightarrow$ PPC

Difference between Covariance matrix & Correlation matrix

Four variables. Then,

x_1	x_2	x_3
2	3	4
4	5	9
6	7	10
8	9	
10	11	

100×4

	x_1	x_2	x_3	x_4
1				
2				
100				

Covariance matrix.

	x_1	x_2	x_3	x_4
x_1	$\sigma(x_1)$	c_{12}	c_{13}	c_{14}
x_2	c_{21}	$\sigma(x_2)$	c_{23}	c_{24}
x_3	c_{31}	c_{32}	$\sigma(x_3)$	c_{34}
x_4	c_{41}	c_{42}	c_{43}	$\sigma(x_4)$

4×4

$C_{ij} = C_{ji}$

$C_{12} = C_{21}$.

$C_{13} = C_{31}$

$C_{23} = C_{32}$

Covariance matrix is symmetric.

6

100×4.

A

A^T. 4×100

4×100 100×4. $i = 1, 2, 3, 4$

$A^T \times A$. 4×4. $j = 1, 2, 3, 5$.

When there are 4 variables, we have to compute 4 variances.

and 6 Covariances.

∴ Total variance & Covariance
$= 4 + 6 = 10$.

Variances $= 4 = (4C_1) = 4$.

Covariances $= 6 = (4r_2) = 6$.

Combination & Permutation.

A B C D.

2 Cinema Tickets.

AB AC AD
BC BD CD

6 ways

Name of the Cinema - Popatlal

BA CA DA
CB DB DC

6 ways.

$= (12)$

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

Permutation.

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

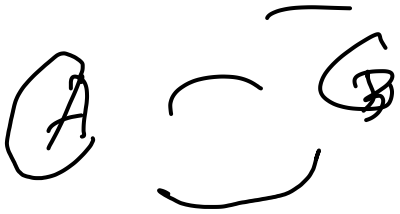
$(a+b)^2 = a^2 + 2ab + b^2$ - (Combination

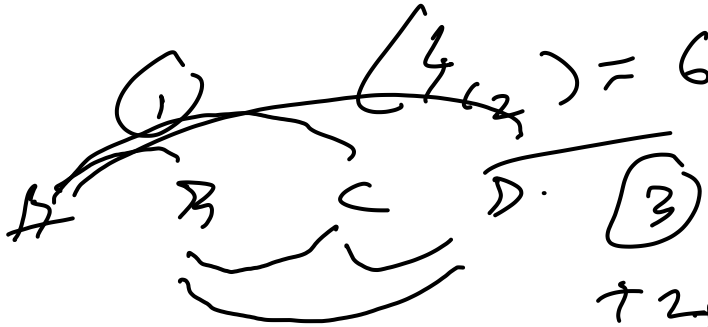
$a+b$

$a+b$.

$a^2 + ab + ba + b^2$.

$ab = ba$





$(4_{C_2}) = 6$

$+2+1$

$= 8$

$AB = BA$.

AB BA.

// // // //

$r_{12} = \frac{C_{12}}{r_1 r_2}$

A B C D.

B A C D.

(4_{C_2})

$=$

$$\begin{pmatrix} 1 & r_{12} & r_{13} & r_{14} \\ r_{21} & 1 & r_{23} & r_{24} \\ r_{31} & r_{32} & + & r_{34} \\ r_{41} & r_{42} & r_{43} & \end{pmatrix}$$

$r_{13} = \frac{C_{13}}{r_1 r_3 x_1}$

	x_1	x_2	x_3	x_4
x_1	r_{11}/r_{11}	C_{12}	C_{13}	C_{14}
x_2	C_{21}	r_{22}/r_{22}	C_{23}	C_{24}
x_3	C_{31}	C_{32}	r_{33}/r_{33}	C_{34}
x_4	C_{41}	C_{42}	C_{43}	r_{44}/r_{44}.