Stats

1. Covariance ✓
2. Pearson Correlation Coefficient ✓
3. Spearman Rank Correlation Coefficient ✓ → Homework ?!
4. Log Normal Distribution ✓
5. Hypothesis Testing → Null Hypothesis, Alternate Hypothesis → Influential (P value)
6.

1. Covariance { Feature Selection }

X	y
12	20
13	25
14	30
28	40
25	50

→ DATASET

① Relationship between X & Y → features

→ X↑ y↑ X↓ y↑

X↑ y↓ → X↓ y↓

② Numbers we can actually quantify the relationship →

Quantify → CoVariance, Pearson Correlation, Spearman Rank Correlation

① Covariance

n = sample

Var := $\frac{}{n-1}$ ⇒

X	y
12	2
13	3
14	4
15	5
$\bar{x}$ 13.5	3.5

$$Cov(x,y) = \sum_{i=1}^{n} \frac{(x_i - \bar{x}) * (y_i - \bar{y})}{n-1}$$

$\bar{x}$ = mean of x

$\bar{y}$ = mean of y

n = Sample

$= (12-13.5) * (2-3.5) + (13-13.5) * (3-3.5) + (14-13.5) * (4-3.5) + (15-13.5) * (5-3.5)$

$= (-1.5) * (-1.5) + (-0.5) * (-0.5) +$

$= 1.6667$ // → +ve

X ↑	y ↑	✓
X ↓	y ↓	✓

> 100 < 100

1000

10,000

> 50 < 100

X	y
5	2
4	3
3	4
2	5

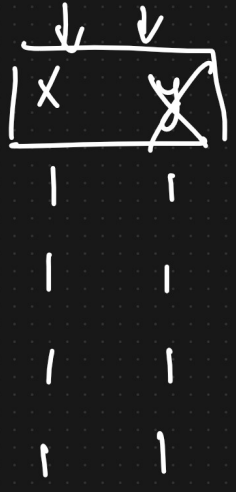
Cov (x,y) = −ve ⇒

X ↑	y ↓	✓
X ↓	y ↑	✓

Cov(x,y) = 0 ??

⇓

Does this mean??



Compare

Cov(x,y) = 0

Limiting the Quantified result

① +ve or −ve { No limit }

⇓

{ Limit } How much +ve / How much −ve } → Value

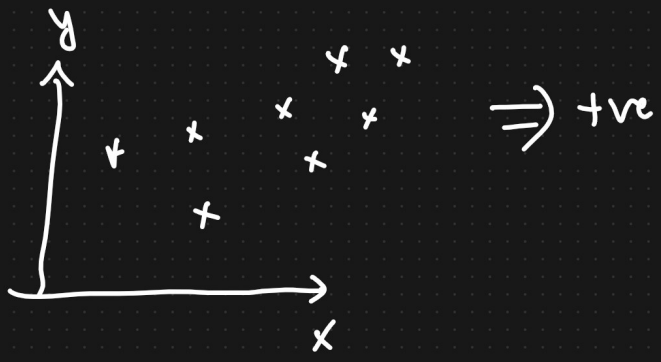
② Pearson Rank Correlation Coefficient

$$Cov(x,y) = \sum_{i=1}^{n} \frac{(x-\bar{x}) * (y-\bar{y})}{n-1}$$

Variance

$$Cov(x,x) = \sum_{i=1}^{n} \frac{(x-\bar{x}) * (x-\bar{x})}{n-1} = \sum_{i=1}^{n} \frac{(x-\bar{x})^2}{n-1}$$

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



$\Rightarrow$ +ve

$\Rightarrow$ negative.

No covariance

≈ 0

-1 to 1

Pearson Correlation Coefficient

$$\rho_{(x,y)} = \frac{Cov(x,y)}{\sigma_x \sigma_y} \Rightarrow [-1 \text{ to } 1]$$

-1: Negative correlated

1: positive correlated

X	Y
2	3
3	4
4	5
5	6

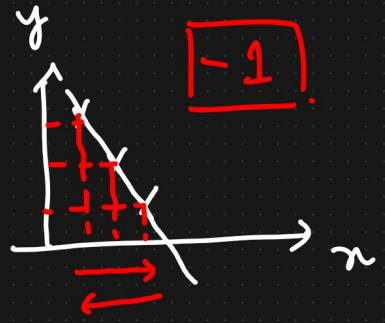
$\bar{x} = 3.5 \quad \bar{y} = 4.5$

$\frac{Cov(x,y)}{\sigma_x \sigma_y}$ ✓ = [-1 to 1] ✓

[1]

[x y] + 0.75

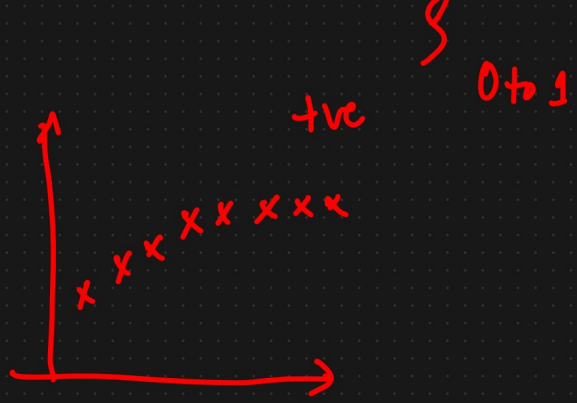
[x z] ✓ + 0.8 ✓

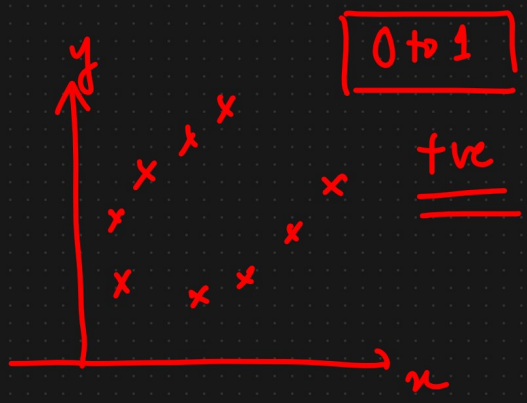


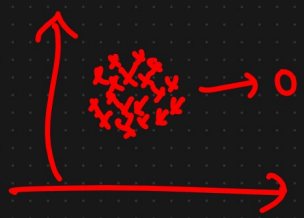
Perfectly negatively correlated

x ↑ y ↓

x ↓ y ↑







Spearman Rank Correlation

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

Feature Important ↓↓ → Domain knowledge

X	y	Rx	Ry
1	2	4	4
5	6	1	2
3	4	2	3
2	8	3	1

500 features Feature Selection

① Independent features
② Dependent features → In Dependent features

No. of study	No. of play	Time to sleep	o/p P/F
→ 7	2	7	P
→ 1	8	7	F

② Log Normal Distribution

Haunted — Run of the mouse

500 → 499 highly Dependent

<70%

X↑ y↓

X Y Z A → O/p

Domain Expert

Yes →

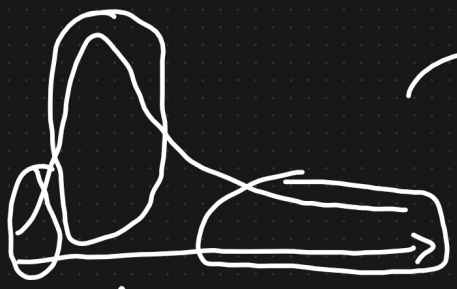
90% 99%

Same feature

⑰ Log Normal Distribution

{ 68 - 95 - 99.7 }

→ Normal Distribution



→ Log Normal Distribution

Some people
- doing comment
- Small comment
- Medium

Eg: { People writing comments In My channel }

Length of the comment

Eg: Wealth distribution

⇒ Yes ⇒ log(X) ⇒ Feature Transformation

→ Normally distributed data.

$$X \sim \text{Log Normal Distribution}(\mu, \sigma)$$

$$\Downarrow$$

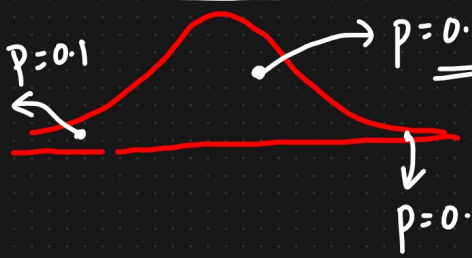
Normal Distribution $\quad \ln(X)$

Inferential Statistics

① P value {probability value}

Eg

← Space bar



100 times
⇓
80 times
5 times

Hypothesis Testing

Tossing a fair coin

100 times

$Pr(H) = \frac{1}{2}$ $Pr(T) = \frac{1}{2}$

⇓

50 times Head ✓

80 time Tail

↳ Null Hypothesis (H0) → True → Treats Equals or Same

↳ The coin is fair

Alternate Hypothesis : The coin is not fair

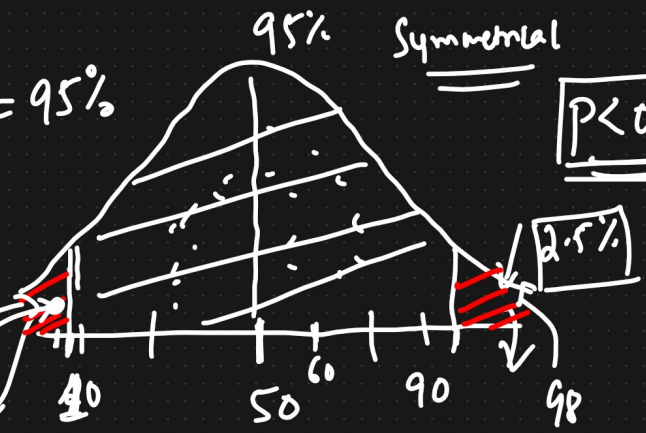
$\alpha = 1 - 0.95 = 0.05$

0.05

Confidence Interval = 95%

⇓

Significance Value



$\alpha = 0.05$

$\alpha = 0.05$

Symmetrical

Experiment

⇓

P<0.05

⇓

95 → 55

60 98

↳ Domain Expert

9 Heads

0.65

Reject the Null Hypothesis
Accept the Alternate Hypothesis
} No. it is not a fair coin

Eg: Person has committed a crime

Null Hyp (H0) = He is innocent

Altern (H0) = He is not innocent

↳ Experiment ⇒ Collecting Evidences → Finger print, DNA,

↳ Hypothesis Testing

$\alpha =$ [ ] → Judge

70% ⇒ is not Innocent

Fair
Rolling a Dice

$\frac{1}{6}$

100

8 16 16 16 16

α 0.1

Influential Stats

Sample data —Assumption→ Population Data.

Experiment
Hypothesis

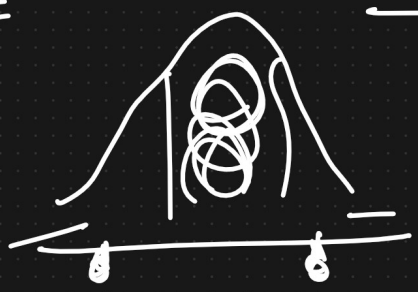
Average size of all the sharks in the world ??

C.I

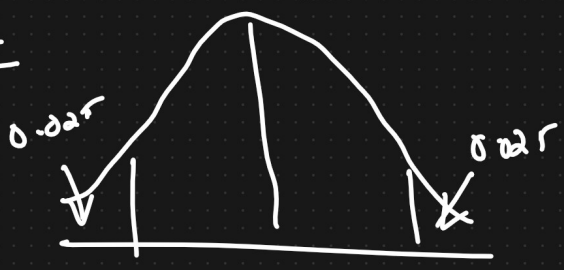
90% C.I

$\alpha = 10 \Rightarrow 0.1$

1



$P < 0.1$



$1 - 0.95 \Rightarrow 0.05$

C.I = 85

$\alpha = 0.15$