Stats - 5th DAY

① Covariance, Pearson And Spearman Rank Correlation } {Completed}

② Infuential Stats

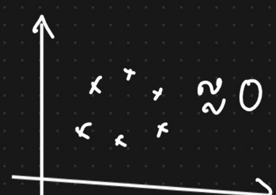
i) P-value

ii) Hypothesis Testing

iii) Confidence Interval

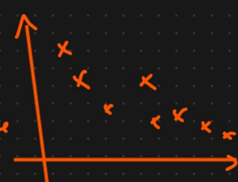
iv) Significance Value

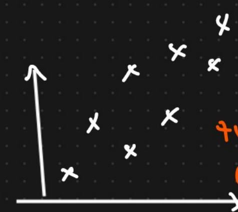
Covariance



≈ 0

-ve Covariance





+ve Covariance

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

$\underbrace{Var(x) \Leftrightarrow Cov(X,X)}$

Disadvantage

① +ve → ∞ -ve → ∞

0 = No Correlation

⇓

② Pearson Correlation Coefficient $\overbrace{[-1 \text{ to } 1]}$

+1 {more positive correlated}

{-1 more negative correlated}

$$\boxed{\rho_{x,y} = \frac{Cov(x,y)}{\sigma_x \cdot \sigma_y}}$$

→ $\boxed{-1 \text{ to } 1}$

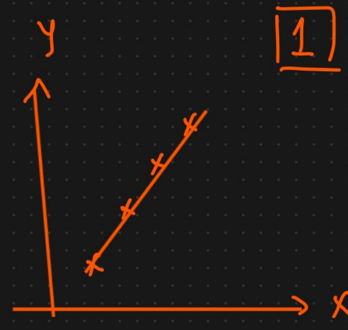
$\boxed{X \quad Y}$ = 0.8 ✓

$\boxed{X \quad Z}$ = 0.9 ✓

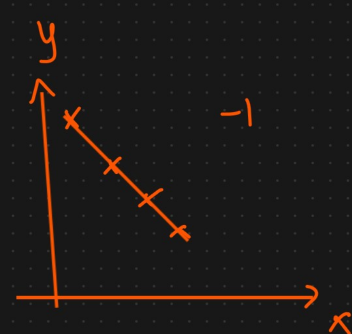
↳ 1 +ve correlated

$\boxed{X\uparrow Y\uparrow \quad X\downarrow Y\downarrow}$ ↑

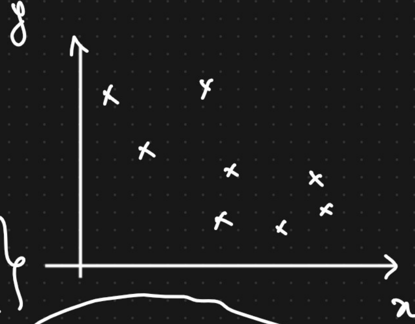
Between {0 to 1}



$\boxed{1}$



-1



{ -0.9, -0.92 }

Between

{ -1 to 0 }

[-1 to +1]

$\boxed{X\uparrow \; Y\downarrow \quad X\downarrow \; Y\uparrow}$

Age ⟷ Working → 0.9
Age ⟷ Career Gap → 0.9

Machine Model

O/P feature

Age	Working Exp	Career Gap ✓	Salary
24	2	1	45000
32	10	3	50000
25	4	0	35000
26	5	0	40000

$\boxed{\text{Feature Selection}}$ ✓

Age ⟷ Working Exp = 0.9

{ Age ↑ Work Exp ↑
Age ↓ Work Exp ↓ }

Age ⟷ Career Gap = 0.9

Important features for your work

100 features ≈ Most Important features.

Disadvantage

① Pearson Correlation works well with Linear Data
⇓
Non Linear Data.

② Spearman Rank Correlation

	X	y	R_x	R_y
x↑ y↑	1	2	4	4
x↓ y↓	3	4	3	3
	7	5	2	2
	0	7	5	1
	8	1	1	5

$$\boxed{r_s = \frac{Cov(R(x), R(y))}{\sigma R(x)\ \sigma R(y)}}$$

⇓

Non linear Data.

Spearman ⟶ Non linear Data + linear

Pearson ⟶ linear Data

X Non linear Data.

→ Inferential Stats

{Hypothesis Testing}

① P value

⇓

Defn

Out of all 100 touches I am touching 2 times

Defn

P value is the probability of the NULL Hypothesis to be True?

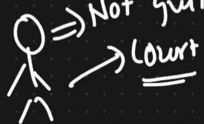
p-value = 0.02

p-value

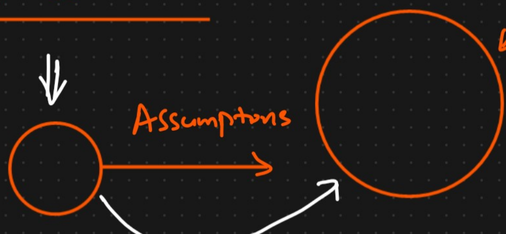
Out of all 100 touches we touch 2 times

Hypothesis Testing ⇒ Inferential Stats

{Not guilty until proved}



⇓



Steps of hypothesis Testing

100

[5D, 60]

① Null Hypothesis – Coin is fair → (H_0)

② Alternate Hypothesis – Coin is not fair → (H_1).

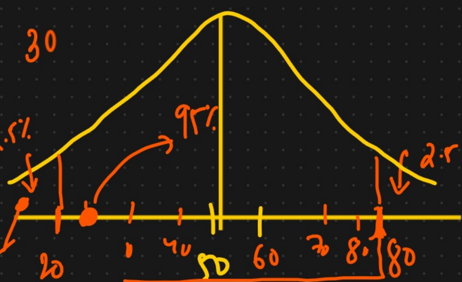
③ Perform Experiments –

{C·I} → 95% → {Domain Expert}

Toss the coin 100 times

30

100 − 95 = 5%



10 → Null Hypothesis is rejected }

30 → Null Hypothesis is accepted }

Significance value $\alpha = 0.05$ ⇒ 95% C·I

① → Coin is fair or not

⇓

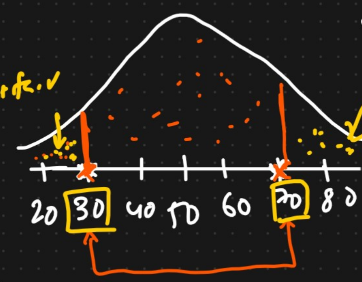
Experiment Toss

$\mu = 50$

$\sigma = 10$

100 times

30

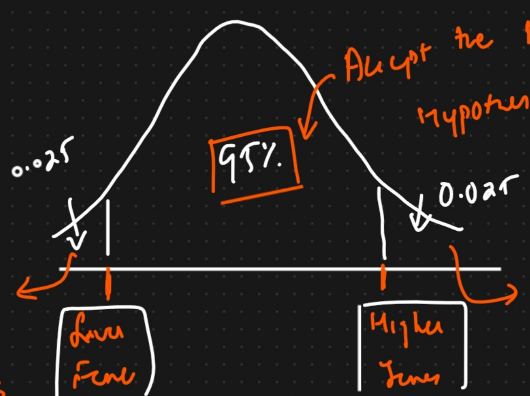


{ 70 Tail
30 Head

C·I = 95%

{ Significance value = 1 − C·I
& p value = 1 − 95% = 0.05

C·I



{ Confidence Interval }

C·I = 95% ⇒

Significance value $\alpha = 1 - 0.95$

$= 0.05$

Point Estimate : The value of any statistics that estimates the value of a parameter is called Point Estimate

⇓

Point Estimate ← $\bar{x}$ ⟶ μ → Parameter

20 + 20   40

$\bar{x}$ ⟶ μ

⇓

Sample Mean

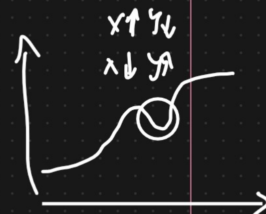
30 + $\boxed{10}$

40

45

$\boxed{1.96}$

$\boxed{\text{Point Estimate} \pm \text{Margin of Error}}$ = Parameter.



$$\left\{\begin{array}{l} \text{Lower Fence} = \text{Point Estimate} - \text{Margin of Error} \\ \text{Higher Fence} = \text{Point Estimate} + \text{Margin of Error} \end{array}\right\} \text{C.I}$$

Assignment

Q) In the Quant test of CAT exam, the population standard deviation is known to be 100. A sample of 25 test takers has a mean of 520. Construct a 95% C.I about the mean?

$\bar{x} ?_2$

↓

$\boxed{80\%}$

Ans) $\sigma = 100$ $\quad n = 25$ $\quad \bar{x} = 520$

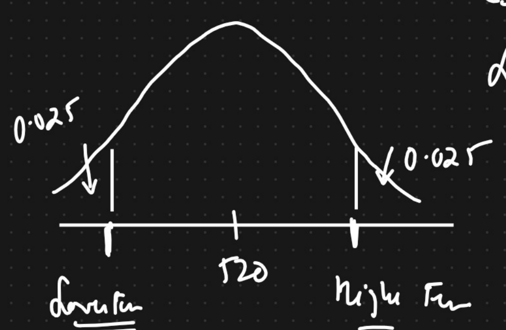
Z-score table

Z test

C.I 95%

$\alpha = 1 - \text{C.I}$

$= 1 - 0.95 = \boxed{0.05}$



α = Significance value

$1 - \text{C.I} = 0.05$

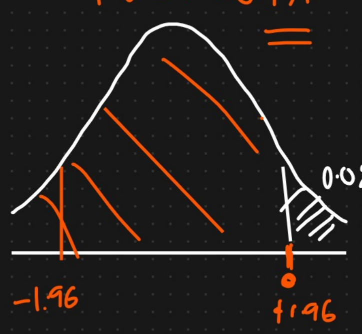
Point Estimate $\pm$ Margin of Error

↓

$$\bar{x} \pm Z_{\alpha/2} \boxed{\frac{\sigma}{\sqrt{n}}} \Rightarrow \text{Standard Error}$$

n = 25

$1 - 0.025 = 0.975$



$$Z_{\frac{0.05}{2}} = Z_{0.025} = 1.96$$

$\alpha = 0.05/2$ $\quad Z_{0.05/2} = \boxed{Z_{0.025}}$

Lower Fence $= \bar{x} - Z_{\alpha/2} \dfrac{\sigma}{\sqrt{n}}$

$= 520 - 1.96 \dfrac{100}{\sqrt{25}}$

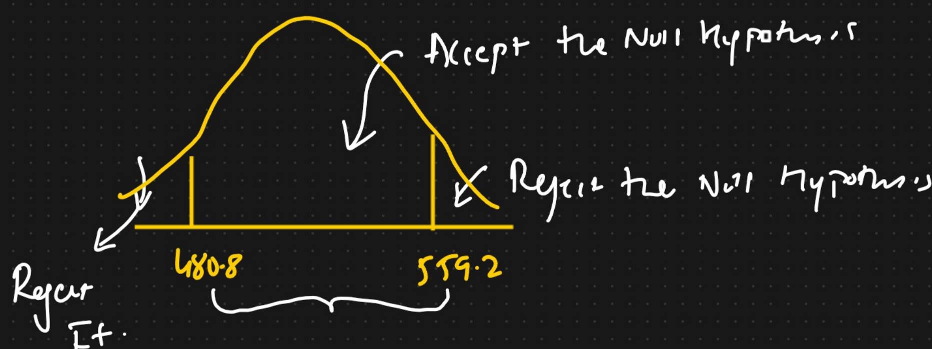
$= 520 - 1.96 \times 20$

$= \underline{\underline{480.8}}$

Higher Fence $= \bar{x} + Z_{\alpha/2} \dfrac{\sigma}{\sqrt{n}}$

$= 520 + 1.96 \times 20$

$= \underline{\underline{559.2}}$



(*) On the Quant test of CAT Exam, a Sample of 25 test takers has a mean of 520 with a sample Standard deviation of 80. Construct 95% C.I about the mean?

Ans) $\bar{x} = 520$ $\quad s = 80$ $\quad C.I = 95\%$ $\quad \alpha = 0.05$

[Z-test] $\Downarrow$ $\rightarrow$ [t-test]

$\bar{x} \pm t_{\alpha/2} \left(\dfrac{s}{\sqrt{n}} \right)$

Population std X given

C.I

t test

$\boxed{n-1}$

① Degree of freedom $= n-1 = 25-1 = \boxed{24}$

$t_{\alpha/2} = t_{0.05/2}$

$\underline{\underline{t_{0.025}}}$

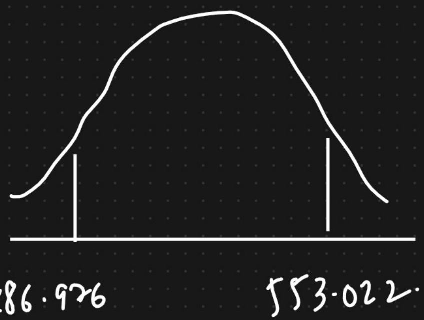
Lower fence $= 520 - 2.064 \times \left(\dfrac{80}{\sqrt{25}} \right)$

$= 520 - 2.064 \left(\dfrac{80}{5} \right)$

$= 486.976$

Higher Fence $= 520 + 2.064 * \left(\frac{80}{\sqrt{25}}\right)$

$= 553.022$

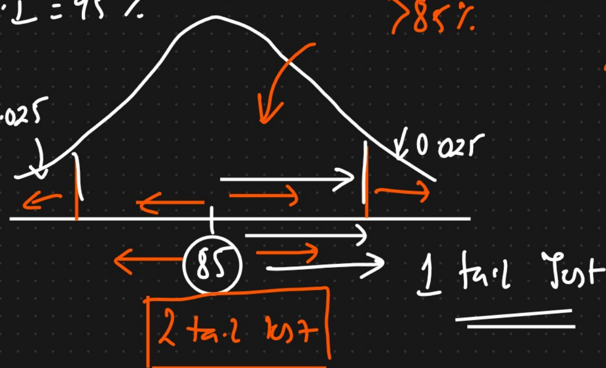


✱ One Tail & 2 Tail

① Colleges in Town A has 85% placement rate. A new college was recently opened and it was found that a sample of 150 students had a placement rate of 88%. with a standard deviation of 4%. Does this college have a different placement rate? with C.I = 95%

>85%

1 tail or 2 tail test

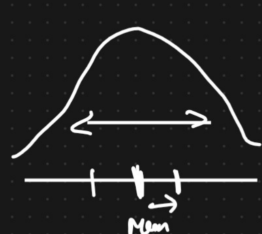


>85%

Measure of Dispersion

Spread of the Data

$\mu = 10 \quad \sigma = 5$



5 10 15

Manager ⇒

Loc A

$\boxed{A}$ ← ATM ⟵⟶ $\boxed{B}$ ← ATM

Loc ⇓

⇑ ⇓

Average Transaction

↓

C·I

× [– –] ×

Size of the shark

⇓

$\boxed{C \cdot I}$ ←

z test ⇐ σ $\bar{x}$ n

t - test ⇐ s $\bar{x}$ n

Steps

Assignment

L, XL

HR ⇒ Sample data

⇓ ⇑

100000 Employees

100k Employees

Brilliant Data Analyst

⇑

$\boxed{C \cdot I}$ ⇒ Confidence Interval

500 data ⇒ 300 XL 200 L

$\boxed{C \cdot I}$ XL = L = $\boxed{95\%}$

How many XL & L T-shirts you need to order

XL

[ ]

$\boxed{\nu}$

Degree of freedom → $n-1$ ← t test

[jayant@ineuron.ai , Bharath@ineuron.ai] $\Rightarrow$ Assignment

Krish.naik@ineuron.ai