CONTENTS

- Probability
- Permutation & Combination
- Covariance
- Pearson Correlation
- Spearman Correlation
- Hypothesis Testing
- Point Estimate
- Chi Square Test

Saturday

SCHEDULE

09.00
10.00
11.00
12.00
13.00
14.00
15.00
16.00
17.00
18.00
19.00
20.00

→ Probability

" Probability is a measure of the likelihood of an event.

Eg : 1 Tossing a fair coin
$P(H) = 0.5$, $P(T) = 0.5$

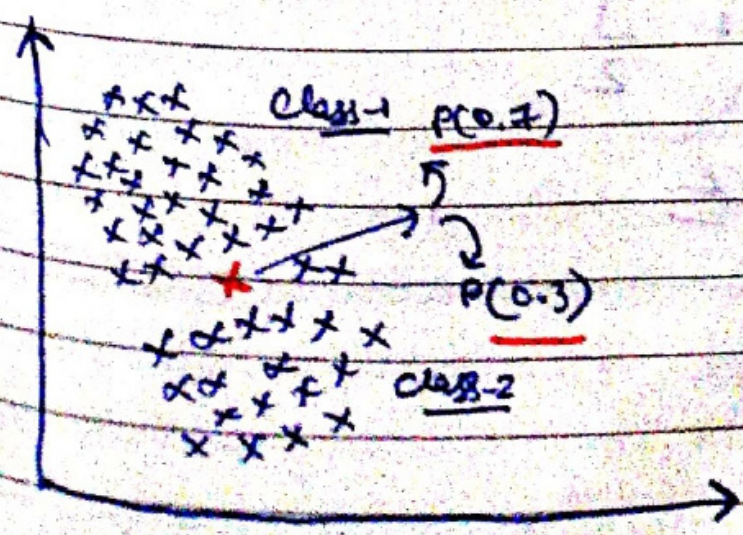
Eg : 2 Sholay coin : unfair coin
$P(H) = 1$

Eg : 3 Rolling a dice
$P(1) = \frac{1}{6}$, $P(2) = \frac{1}{6}$, $P(3) = \frac{1}{6}$

* Use of Probability :
In Machine Learning.

For Eg :- like In a classification problem it say there are some data points belongs to 2 different classes. Now if we add a new data point then we judge through some algorithm that what will be the probability to add that data point in which class.

Sunday 06



JUN

07 127-238 / Wk-19 2018 ■ MAY ■

Monday

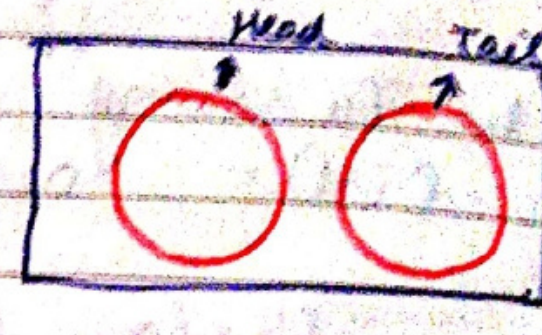
09.00 10.00 11.00 12.00 13.00 14.00 15.00 16.00 17.00 18.00 19.00 20.00

(1) Mutual Exclusive Events :-

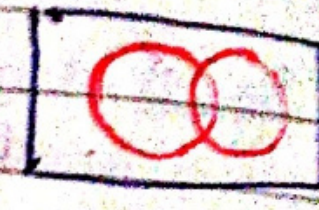
Two events are mutually exclusive if they cannot occur at same time.

eg:-

1. Tossing a coin
2. Rolling a dice



(2) Non-Mutual Exclusive Events



Two events can occur at the same time.

eg:- 1.) Picking a card from deck of cards, two events "Heart" and "King" can be selected

Q1 What is the Probability of coin landing on Heads or Tail.

Hint: Use Addition Rule for mutual exclusive events.

Sol $P(A \text{ or } B) = P(A) + P(B)$

$\Rightarrow \frac{1}{2} + \frac{1}{2}$

$\Rightarrow 1$ Ans

Q2 A Bag contains: 10 Red marbels, 6 Green and 3 red & green marbles.

(I) what is the prob. of choosing a marbel that is red or green?

Hint: use Addition Rule for non-mutual exclusive event.

Sol: $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

$P(A \cap B)$

$= \frac{13}{19} + \frac{9}{19} - \frac{3}{19}$

R → 10+3 = 13

G → 6+3 = 9

$\Rightarrow \frac{19}{19} = 1$ Ans

Q3 What is the prob. of choosing a 'heart' or 'Queen' from a deck of cards. [non-mutual]

Sol $P(H) = \frac{13}{52}$, $P(Q) = \frac{4}{52}$

$P(H \cap Q) = P(H) + P(Q) - P(H \text{ and } Q)$

$\Rightarrow \frac{13}{52} + \frac{4}{52} - \frac{1}{52}$

$= \frac{16}{52}$ Ans

Wednesday

09:00 → **Multiplication Rule**

10.00

11.00 (1) **Dependent Events :-**

12.00

13.00 Two events are dependent if they

14.00 affect one another.

15.00

16.00 **For eg :-**

17.00 In a bag of marble there are

18.00 6 Red Balls & 4 Green Balls, $P(R) = 6/10$

19.00 $P(G) = \frac{4}{10}$, after removing 1 Red ball

20.00 Prob. of green balls affect by $\boxed{P(G) = \frac{4}{9}}$

balls.

---

(2) **with Independent Events :-**

Each and every

~~three~~ events is ~~one~~ independent.

for eg :- Rolling a dice.

Q1 What is the probability of rolling "5" & "3" with a normal 6 side dice?

Hint: Use Multiplication Rule for Independent Event & Mutually exclusive

Sol $P(A \text{ and } B) = P(A) * P(B)$

$$\Rightarrow \frac{1}{6} * \frac{1}{6}$$

$$\Rightarrow \frac{1}{36} \text{ Ans}$$

09:00 → Permutation & Combination

10:00

11:00 • Permutation ⇒ ${}^{n}P_{r} = \frac{\lfloor n}{\lfloor n-r}$

12:00

13:00

14:00 ⇒

15:00 • Combination ⇒ ${}^{n}C_{r} = \frac{\lfloor n}{\lfloor r \lfloor n-r}$

16:00 (unique combination)

17:00 (not repeated)

18:00

19:00 For eg:-

20:00

Q1 In a bag 5 chocolates are available & you have to select only 3

Sol $\underline{5} \quad \underline{4} \quad \underline{3}$ ⇒ 5 × 4 × 3 ⇒ 60.

no. of options of selection = 60

OR

Q1 ${}^{n}P_{r}$ ⇒ $\boxed{{}^{5}P_{3}} = \frac{\lfloor 5}{\lfloor 5-3} \Rightarrow \frac{5 \times 4 \times 3 \times 2 \times 1}{2 \times 1}$

⇒ $\boxed{60}$

• In case of combination: (repetition not allow)

Q1 ${}^{n}C_{r} = {}^{5}C_{3} \Rightarrow \frac{5 \times 4 \times 3}{3 \times 2 \times 1} \Rightarrow \boxed{10}$

09:00 → **Covariance :** { Used in feature Selection }

10.00

11.00 Covariance is usually given by the curve

12.00 between 'x' & 'y'

13.00

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

14.00

15.00

16.00

17.00

18.00 var, $\sigma^2 = \frac{\Sigma (x_i - \bar{x})^2}{n-1}$ ⇒ var(x) = $\frac{\Sigma (x_i - \bar{x})(x_i - \bar{x})}{n-1}$

19.00

20.00

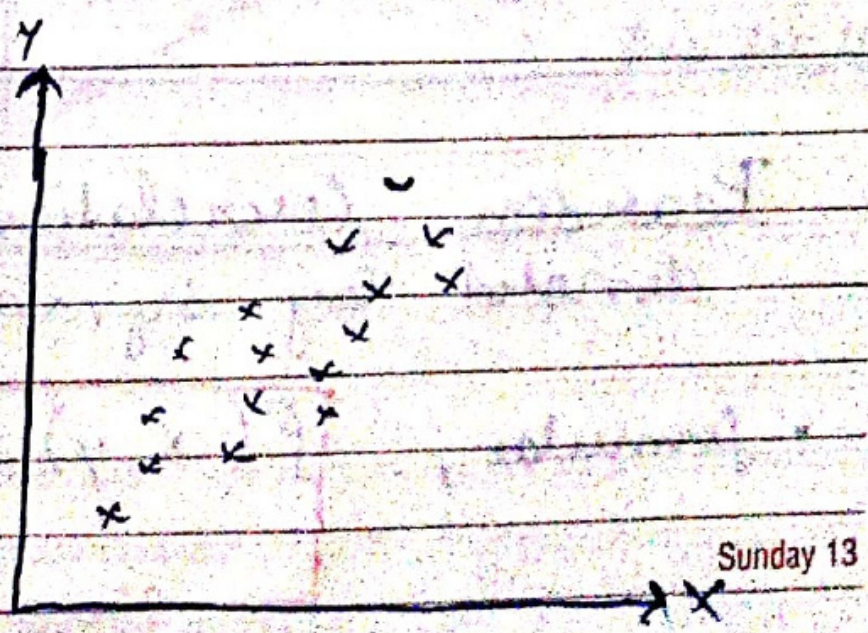
⇒ $var(x) = cov(x, x)$

⇒ $cov(x,x) = Var(x)$

* Interview Ques : Relationship b/w cov & var.

Eg :-

X	Y
Age	Weight
12	41
13	44
14	47
15	50
16	55
17	59
18	61



Sunday 13

⇒ Age ↑ ↔ weight ↑
Age ↓ ↔ weight ↓ } observation

Sol $\bar{x} = 15$

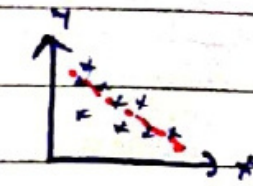
$\bar{y} = 51$

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

$\Rightarrow$ 24 → (+Ve) ↴

* ∴ Acc to final Ans X ∝ y

* If Ans is (-ve) then $X \propto \frac{1}{y}$



* If, Cov is '0', then NO relation b/w X & y

→ Pearson Correlation Coefficient (P) denoted by symbol Rho (ρ) [for linear Data]

• formulae, $$\rho(x,y) = \frac{Cov(x,y)}{[\sigma x \cdot \sigma y]}$$ → we are restricting -1 to +1 by using [σx . σy]

* W.r.to covariance there is no such restriction that how much +ve value it can have or -ve value it can have.
But with the help of Pearson Correlation we are trying to restrict the values between [-1 to +1]

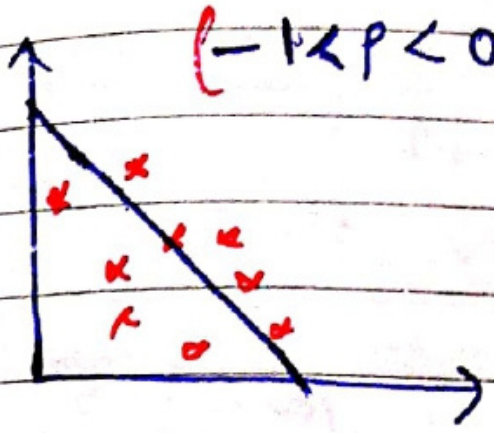
Tuesday

09:00
10.00
11.00
12.00
13.00
14.00
15.00
16.00
17.00
18.00
19.00
20.00

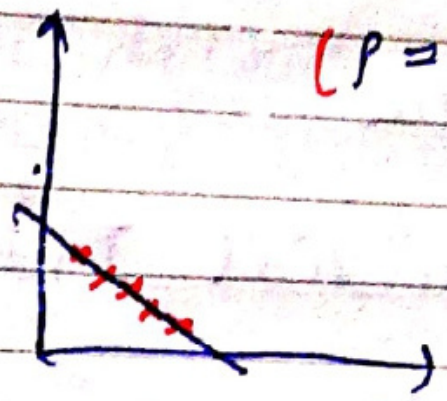
* "More the value towards +1, More positive correlated it is."
* "More the value towards −1, More negative correlated it is."

Eg (at Wikipedia)

a) $(-1 < \rho < 0)$

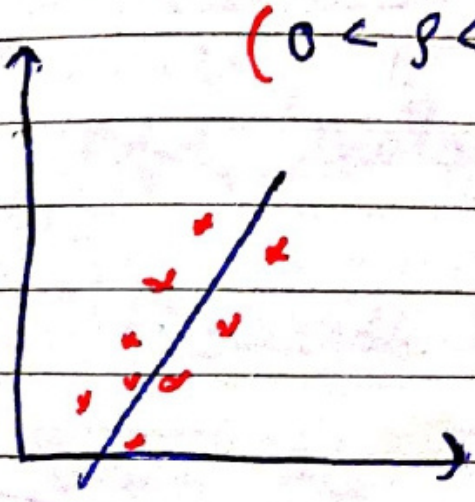


b) $(\rho = -1)$

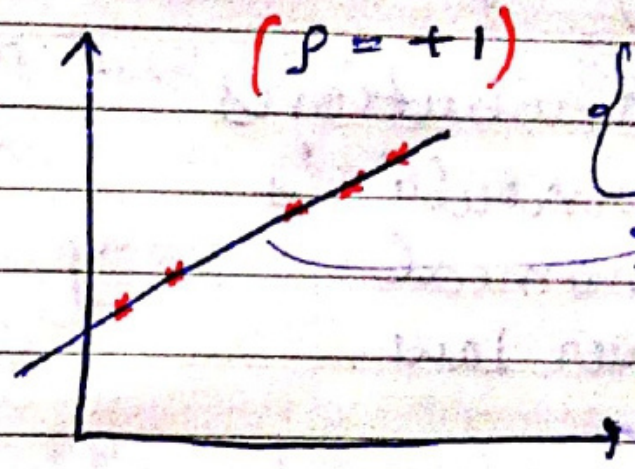


$\{x \uparrow \quad y \downarrow$
$x \downarrow \quad y \uparrow\}$

c) $(0 < \rho < +1)$



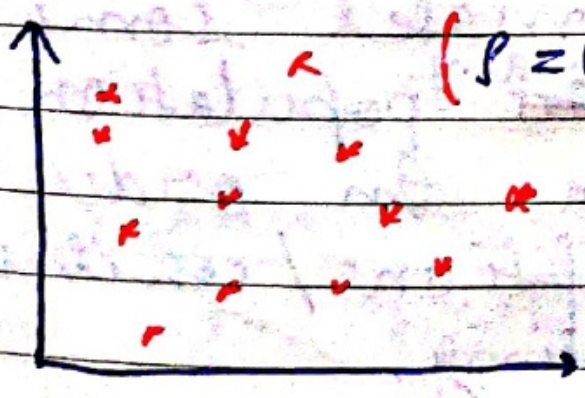
d) $(\rho = +1)$



$\{x \uparrow \quad y \uparrow$
$x \downarrow \quad y \downarrow\}$

→ Best fit line

e) $(\rho = 0)$



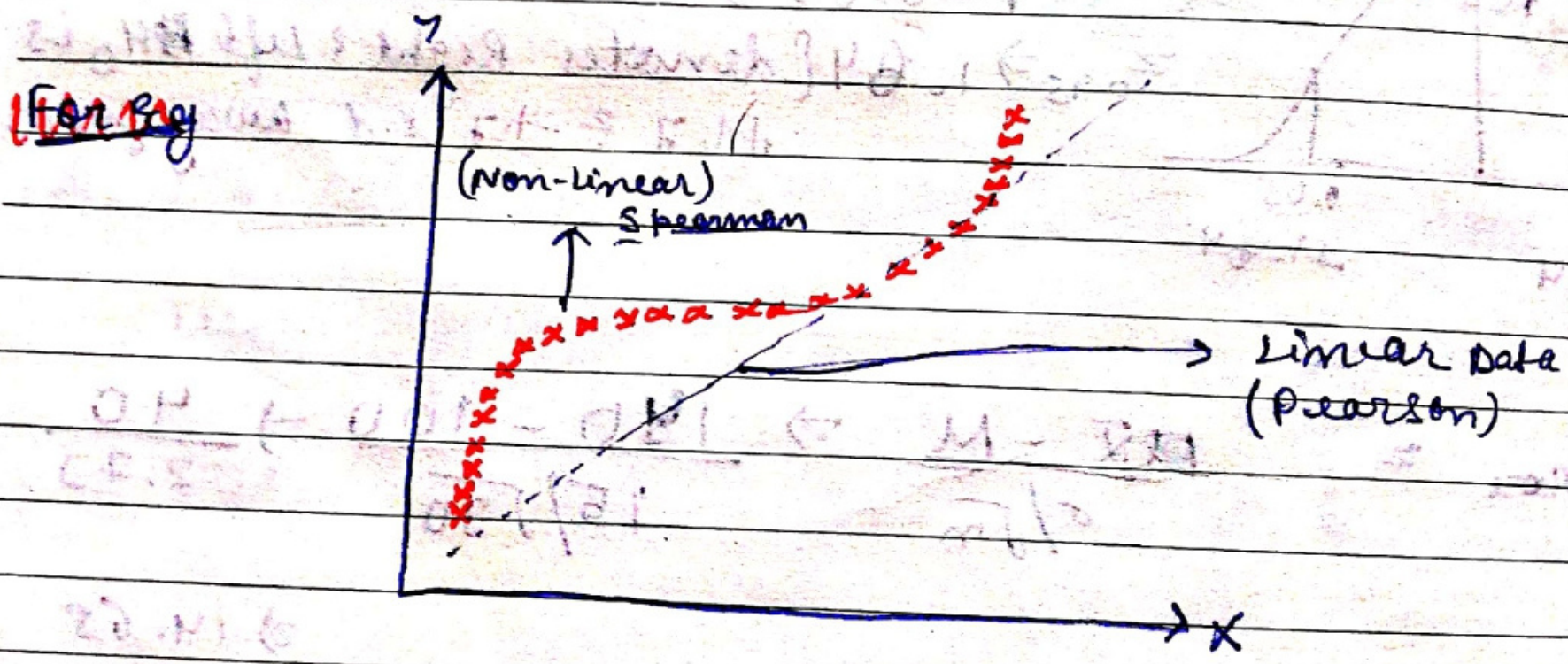
[No Relation]

→ Spearman Rank Correlation :- {Non-Linear}

Disadvantage with Pearson Correlation is that it only holds Linear Data for entire correlation.

So for Non Linear Data points we have to use Spearman Rank Correlation spearman assigns Rank, (in ascending Order) ↵ formulae :-

$$R_s = \frac{\text{cov}(R(x), R(y))}{\sigma(R(x)) * \sigma(R(y))}$$



- Pearson Corr : 0.88 ($0 < \rho < 1$)
- Spearman Corr : 1 (Rank)

SCHEDULE

Thursday

09.00
10.00
11.00
12.00
13.00
14.00
15.00
16.00
17.00
18.00
19.00
20.00

→ Hypothesis Testing

In order to validate assumption regarding population data through Sample data we use Hypothesis Testing.

→ 2 types of Hypothesis Testing :-

- Steps of Hypothesis Testing

① Null Hypothesis :- The default one (right one)
- Eg1:- Person is not a criminal until he is proven by court.

Eg2 Q-[Coin is fair or Not]
To prove it should be P(H) = 0.5 & P(T) = 0.5]
- In case of fair coin is Null Hypothesis. (except Shotay coin)

② Alternate hypothesis
- from eg-1 Person has commited a crime.
- from eg-2 coin is not fair

③ Perform Experiments :- Perform experiments to prove some conclusion.
- for eg. performing experiments tossing a coin, 100 times
  a) First time got 50 time Head : coin is fair in this
  b) Second time got 60 times Head : can say Coin is fair
  c) third time got 70 times Head : " " " " "

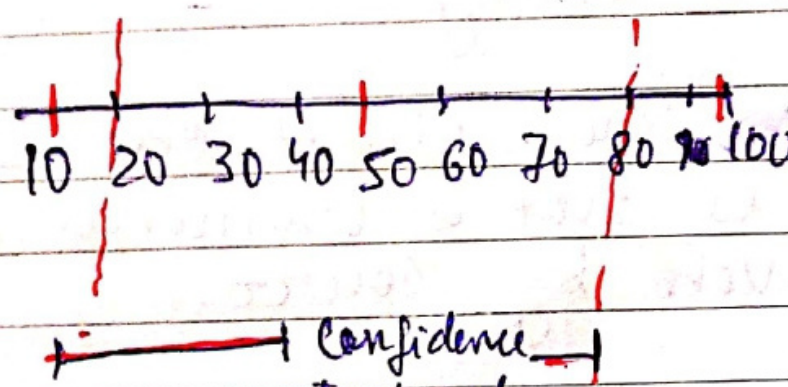
JUN

Domain expert

SCHEDULE

09.00
10.00
11.00
12.00
13.00
14.00
15.00
16.00
17.00
18.00
19.00
20.00

So, conclusion is based on the range between 10 to 100. So if the range (C-I) is fixed that b/w 20 - 80 if condition lies in this interval then we can say coin is fair, outer this range coin will not be fair. This range is called Confidence interval.



"Defining some Range is called Confidence Interval" declared by Domain expert.

* We fail to reject the Null hypothesis [within Confidence Interval]
* We reject the Null hypothesis [outside confidence interval]

eg-2 Conclusions

→ Person is Criminal or not [murder case]

① Null Hyp. :- Person is not Criminal

② Alternate Hyp:- Person is criminal (declared)

③ Experiments/Proof: DNA, finger print, eye-witness
↳ Judge → make decisions.

Monday

09.00
10.00
11.00
12.00
13.00
14.00
15.00
16.00
17.00
18.00
19.00
20.00

→ **Point estimate:-** The value of any statistics that estimates the value of a parameter is called Point estimate.

like we are estimating μ (Population mean) on the basis of $\bar{x}$ (sample mean) in Inferential stats.

Parameter ↓

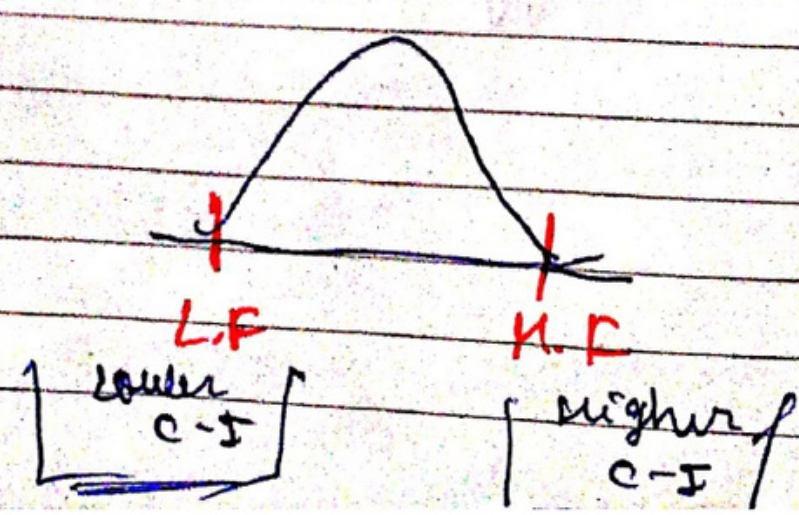
Statistic $\bar{x}$ ⟶ μ ← .

Point estimate ↑

For eg: Value of stat. ($\bar{x}$) that estimates the value of a parameter (μ) is called Point estimator.

⇒ Point Estimate ($\bar{x}$) ± Margin of Error (≥ or ≤ than Population mean) ⇒ Parameter μ (Population Mean)

- Lower fence = Point Estimate − Margin of Error
- Higher fence = Point Estimate + Margin of Error



* use always Z test, when population s.d is given
* use always t-test, when sample s.d is given

■ MAY ■ 2018 142-223 / Wk 21

22 Tuesday

Margin of error ⇒ $Z_{\alpha/2} \left(\frac{\sigma}{\sqrt{n}} \right)$

σ → Population s.d.

$\frac{\sigma}{\sqrt{n}}$ ⇒ standard error

25 145-220 / Wk 21

Friday

2018 ■ MAY ■

If in Ques. σ is not given & Sample s.d is give not population

then use formulae

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

, degree of freedom $= n-1 \Rightarrow 25-1 \Rightarrow 24$.

Q6 On the quant test of Amazon exam, a sample of 25 test takers has a mean of 480 with a population s.d of 85. Construct a 90% of Confidential Interval about the mean.

Sol

- lower C-I = Point Estimate ($\bar{x}$) − Margin of error

$\Rightarrow$ $480 - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

C-I = 90
α = 0.1

$\Rightarrow$ $480 - Z_{0.1/2} \cdot \frac{85}{5} \; 17$

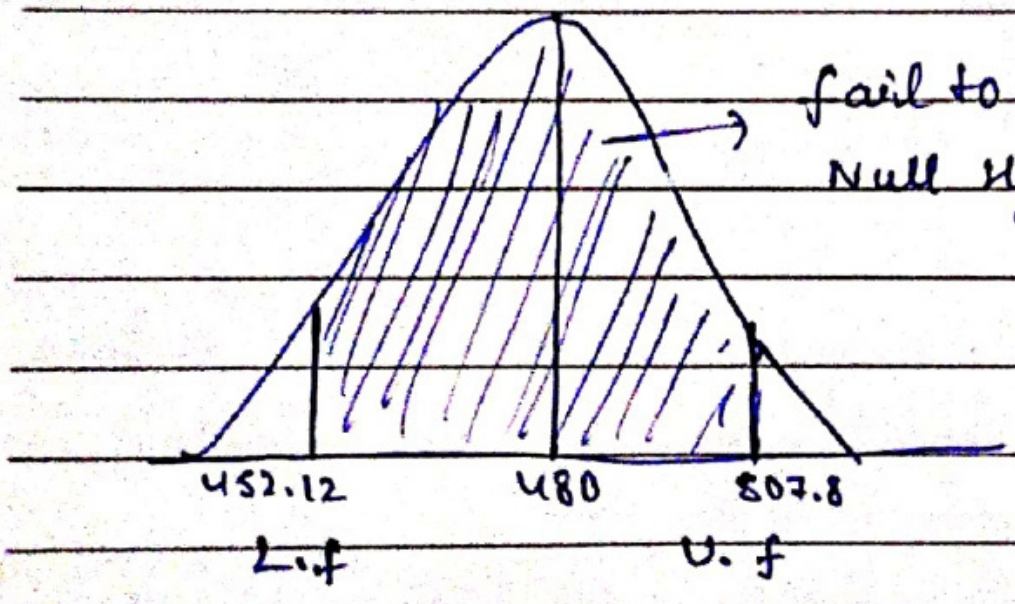
$\Rightarrow$ $480 - Z_{0.05} \times 17$

$\Rightarrow$ $480 - 1.64 \times 17$

$\Rightarrow$ 452.12

- Upper fence $\Rightarrow$ $480 + 1.64 \times 17$

$\Rightarrow$ 507.88



Range [452.12 − 507.8]

28 148-217 / Wk-22 2018 ■ MAY ■

Monday

Hypothesis Testing

Z-test (if $n \geq 30$ OR Population std is given)

→

t-test (if $n \leq 30$ & 's' is given)

→ Hypothesis Testing Prob.

Two tail test

Q) Bull A factory has a machine that fill 80 ml of Baby medicines in a bottle. An employee believe that avg. amount of baby medicine is not 80 ml. Using 40 sample he measures the avg amount dispersed by a machine to be 78 ml with a S.d of 2.5

(a) State Null Hypothesis

(b) At 95% C-I, is there enough evidence to support Machine is working properly or not.

Sol:-

Step (I) Null hypothesis, $\mu = 80$ (machine working properly)

Alternate H_0, $\mu \neq 80$ (machine not working properly)

Step II

$\mu = 80$ ml

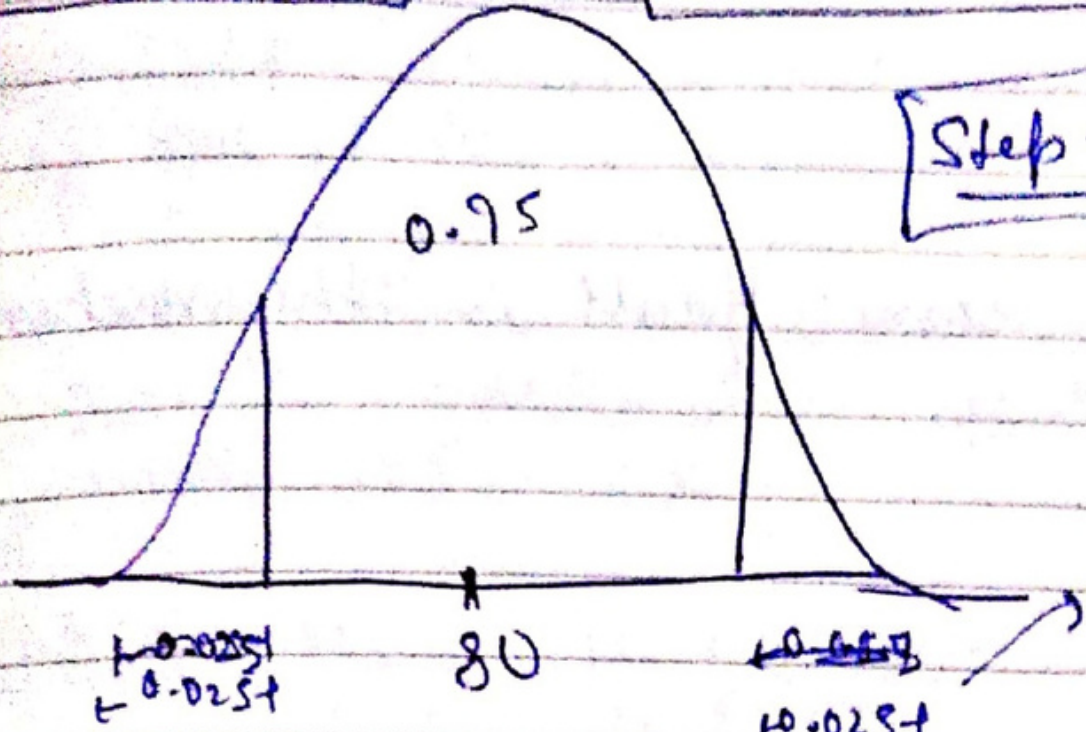
$n = 40$

$\bar{X} = 78$ ml

$s = 2.5$

→ C.I = 0.95

→ α, (S.V) = 1 − 0.95 → 0.05



Step - III Deciding b/w z & t

value from z table

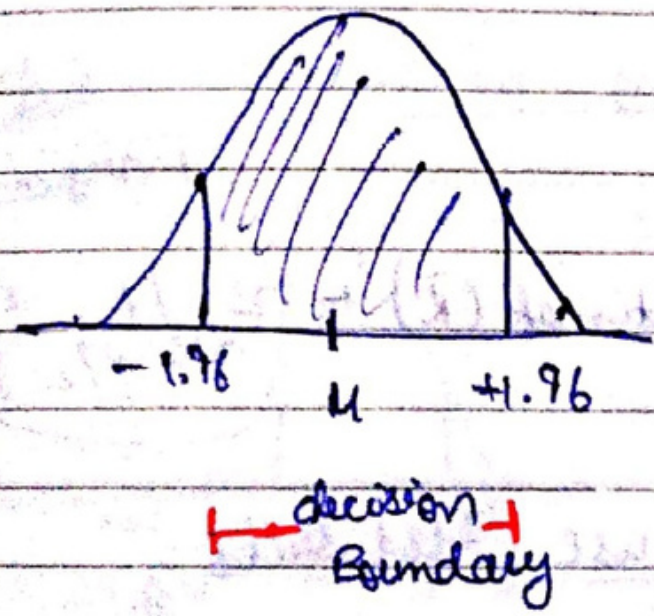
$1 - 0.025$

$\Rightarrow z_{0.975}$

$\Rightarrow -1.96$

Z value $[-1.96 \text{ — } 1.96]$

decision Boundary



Step IV

→ Now Calculate / test statistics (z-test)

$$z = \frac{\bar{x} - \mu}{s/\sqrt{n}} \Rightarrow \frac{78 - 80}{\frac{2.5}{\sqrt{40}}}$$

$s/\sqrt{n}$ → standard error

$\Rightarrow -5.05$

→ Conclusion

My Decision Rule:- If $z = -5.05$ is less than -1.96 or greater than $+1.96$, then we Reject Null H_0 with 95% of Confidence Interval.

⇒ There is some fault in the machine.

Q9 A factory manufactures cars with a warranty of 5 years or more on the engine & transmission. An engineer believes that the engine or transmission will malfunction in less than 5 years. He test a sample of 40 cars and finds the average time to be 4.8 years with a standard deviation of 0.50 State the H_0 & H_1 at a 2% of significance level, is there enough evidence to support the idea that the warranty should be revised?

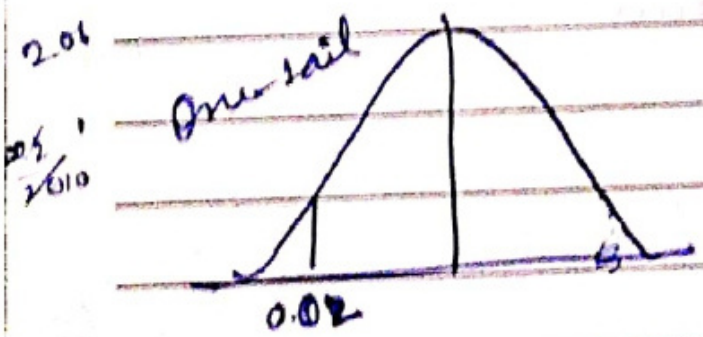
Step-1 H_0: $\mu \geq 5$

H_1: $\mu < 5$

Step2: $\alpha = 0.02$, CI $\Rightarrow$ 0.98

$\sigma = 0.5$ (population s.d)

Step3: perform z-test (one-tail) left-tail



$z_{0.02} \Rightarrow 2.05$

{decision Boundary denotes -2.05 s.d away from mean}

$$z_{score} \Rightarrow \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \Rightarrow \frac{4.8 - 5}{0.5/\sqrt{40}} \Rightarrow \frac{-0.2}{0.079} \Rightarrow -2.53$$

Step4: Conclusion

$-2.53 < -2.05$

$\Rightarrow$ Reject null Hypothesis, H_0

$\Rightarrow$ Engine or transmission malfunction is less than 5 years.

"With the help of z-score whatever area under the curve we are getting is p-value"

Q9 Q9 with the help of P-value

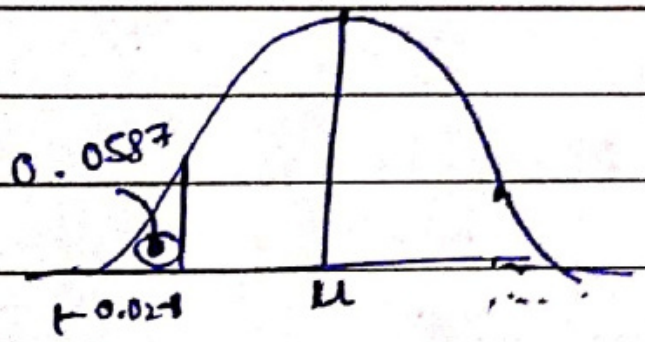
Sol: Step 1: $H_0: \mu = 5$

$H_1: \mu \neq 5$

Step 2: $\sigma = 0.5$ {one-tail test}

$\alpha = 0.02$

Step 3: Apply z-test & find p-value



$$Z_{score} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \Rightarrow \frac{4.8 - 5}{0.5/\sqrt{40}} = -2.53$$

Area under the curve, $Z_{-2.53} = 0.0587$

So Here pvalue is $\boxed{0.0587}$

Step 4: $\therefore$ $0.0587\ (P) < S.V\ (\alpha)$ {Conclusion}

$\Rightarrow$ We reject H_0

$\Rightarrow$ Engine malfunction in less than five years.

Q10 A company manufactures bikes batteries with an average life span of 2 year or more years. An engineer believes this value to be less. Using 10 samples, he measures the average life span to be 1.8 years. with a standard deviation of 0.15.

a) State the Null and Alternate hypothesis.

b) At a 99% C.I, is their enough evidence to discard the H_0 ?

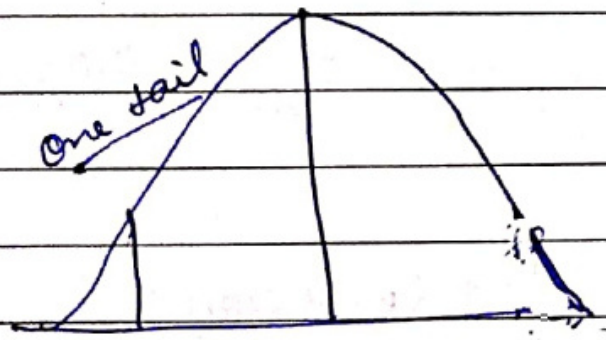
Sol Step 1: $H_0 \Rightarrow \mu_0 \geq 2$

$H_1 \Rightarrow \mu_0 < 2$

Step 2: Sample var, (s) = 0.15 { one-tail }, left tail

$\bar{x} = 1.8$, $n = 10$ (coz eng. believes)

Step 3:- perform t-test test statistics :-



S.L.V = 0.01

$t_{-0.01} = 2.82$

Decision boundary [-2.82]

t-test ∴ degree of freedom = n-1 ⇒ 9

$t_{score} = \frac{\bar{x} - \mu}{s/\sqrt{n}} \Rightarrow \frac{1.8 - 2}{0.15/\sqrt{10}} \Rightarrow \frac{-0.2}{0.047} \Rightarrow -4.255$

Step 4: Conclusion

Here (Z score) $-4.255 < -2.82$ (decision Bound.)

⇒ Reject H_0

⇒ Life span of Battery is less than 2 years

ACCURATE

Building Careers Accurately!

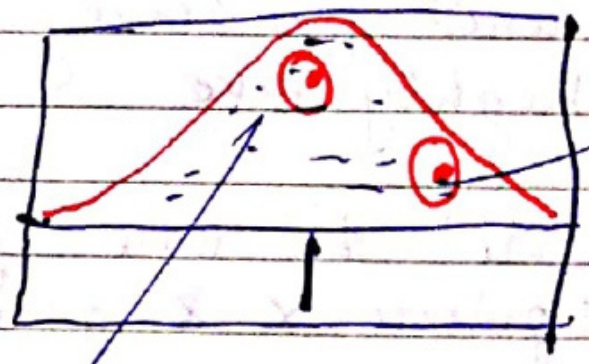
Tuesday

SCHEDULE

09.00
10.00
11.00
12.00
13.00
14.00
15.00
16.00
17.00
18.00
19.00
20.00

→ P-value :- P value ≠ Significance Value (2)
check derived from C-I

Foreg:- In case of our Mousepad



Here the value p=0.02 that indicates out of all the hundred ~~toucg~~ touches we do randomly 2 touches will go over here.

High pvalue.
Let say 0.80
So we can say out of all hundred touches we do 80 touches.

* → P value use to find Hypothesis.

Step 1:- first find Z score., By using Z score find area under curve using Z-table.

Step 2:- Compare the p value with S.V.
(the coming value from z table is pvalue)

* If, P value < α value :- Reject null Hypothesis
| P value > α value :- Accept Ho

Note * if test, is 1 tail value is area under curve 0.05
* if test, is 2 tail value is 0.05 + 0.05 ⇒ 0.1

→ (χ^2) Chi Square Test. (always applied on categories)

* Chi Square test claims about population proportions.

* It is a non parametric Test that is performed on categorical data

* find degree of freedom. ordinal data / Nominal data.

* If $\chi^2 >$ ... Decision Boundary
  ↳ Reject H_0

* there is no 1 tail or 2 tail.

Q1 In a 2000 U.S census the age of individuals in a small town found to be the following.

<18	18-35	>35
20%	30%	50%

In 2010, ages of $n = 500$ individuals were sampled. Below are the results

(out of 500)

<18	18-35	>35
121	288	91

Using $\alpha = 0.05$ would you conclude the population distribution of ages

has changed in the last 10 years?

Sol:

	≤18	18-35	>35
observed	121	288	91
expected (out of 500)	100	150	250

Step1: H_0 : The data meets the expected distrib.
H_1 : The data does not meet " "

Step2: $\alpha = 0.05$, CI = 95%.

Step3: D.o.f {categories}
D.f = category - 1
⇒ 3 - 1 ⇒ 2

Step4: Decision Boundary

* (check in chi square table, compare it with df & α value)

∴ [5.991]

Step5: Chi Square Test Statistics

$$\chi^2 = \sum \frac{(\text{Observed} - \text{expected})^2}{\text{expected}} \text{ OR } \sum \frac{(f_o - f_e)^2}{f_e}$$

$$\Rightarrow \frac{(121-100)^2}{100} + \frac{(288-150)^2}{150} + \frac{(91-250)^2}{250}$$

$\chi^2 \Rightarrow 232.494$