Day 1

Statistics

- Statistics: Statistics is the Science of Collecting, Organizing and analyzing data.

- Data: "facts or pieces of information"

1) Eg: Height of students in a classroom
→ [175 Cm, 150 Cm, 140 Cm, 130 Cm, 155 Cm]

2) Intelligence Quotient (IQ) of 5 randomly selected individuals (109, 89, 129, 101, 105, 106) → Data.

- Types of statistics:-

Statistics

Descriptive stats	Inferential stats
• It consists of organizing and summarizing of data.	• It consists of using data that you've measured to form conclusion.
• Eg: Pdf, Histogram, Box plot, Bar Chart, Pie Chart.	• Eg: Hypothesis Testing, P-value, Z test, t test, Anova, chisquare.

Eg: Lets say there are 20 maths classes at your university and you've collected the age of students in one class.

Ages [21, 20, 18, 34, 17, 22, 24, 25, 26, 23, 22]

- Descriptive Stats:- What is the average (mean, mode) age of student in your maths class?

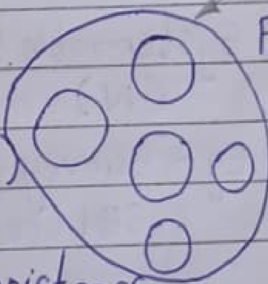
- Inferential Stats:- Are the ages of students in this maths classroom similar to what you would except in a normal maths class of this university?

Sample data	Inferences	Population data

Population and Sample Data : Inferential Statistics

Ex:-

1) Elections - Punjab (Exit Polls)
   [AAP, Congress]
   (Stratified + Random S)



Population of Punjab

2) Eg - 2015 → Data Scientist
   Jackets → Size ; 10K, 40K → Christmas
   10% Small, 20% XL, 40% L, 2-3% XXL
   Only 1-2% → waste.

Population (N)   Sample (n)

- Sampling Techniques

① Simple Random Sampling
② Stratified Sampling
③ Systematic Sampling
④ Convenience Sampling

① Simple Random Sampling: Every member of the Population (N) has an equal chance of being selected for your sample (n).

② Stratified Sampling:

Strata → Layers (Non Overlapping)
Clusters → groups

Gender → Male
Gender → Female

Age groups: 0-18, 18-35, 35-60

Blood Group
Tax Slabs
Courses

③ Systematic Sampling:
(N) → Select every n^{th} individual
Eg - Survey → Mall (luggage checking)
SBI Credit Card

④ Convenience Sampling: Only those people who are interested will only be participating.
Eg - Data Science - AI ; Health Care (Blind People)
Youtube Survey
RBI - House hold Survey - Female

- Variable

A variable is a property that can take on any value.

Eg - Height = 182, 150, 145, 160 [182, 150, 145, 160]
⇓
NO

- Two kinds of Variable :

① Quantitative Variable : Measured Numerically
[Add, Substract, ×, ÷]

② Qualitative Variable :
Eg - Gender - Male [Based on some characteristics
Female we can devide Categorical
variables]

[Quantitative → Qualitative Variable]
Eg :- IQ

0 - 10	10 - 50	50 - 100
Less IQ	Medium IQ	Good IQ

Quantitative

× 2.51 Discrete Variable
× 2.75 Eg: whole number
Eg: No. of Bank Accounts
[2, 3, 4, 5, 6, 7]
Eg: Total No. of Children in a family
Eg: Total no. of employees in a Company Eg: 10K.

Continuous Variable
Eg: Height = 172.5, 162.5 Cm, 163.5 Cm
Rainfall: 1.35, 1.25, 1.75
weight, Temp, Stock Price.

- Assignment Questions -

① What kind of variable Marital Status is ?
Ans Categorical

② What kind of variable Nile River Length is ?
Ans Continuous Quantitative

③ What kind of variable Movie duration is ?
Ans Continuous Quantitative

④ What kind of variable IQ is ?.

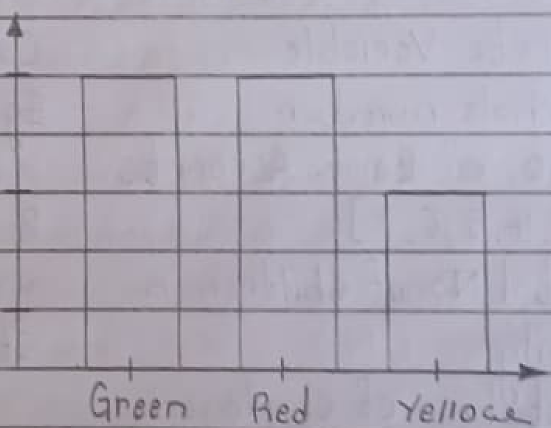
Ans. Continuous variable

- Frequency Distribution

Sample Dataset : Green, Red, Yellow, Green, Red, Yellow, Green, Red

Colors	Frequency
Green	3
Red	3
Yellow	2

① BAR GRAPH Frequency



BAR CHART

- Variable Measurement Scales

4 types of Measured Variable

① Nominal data (categorical data)

Eg: Colors, Gender, types of flowers

② Ordinal data:

Student (Marks)	Rank
100	1
96	2
57	4
85	3
44	5

Ordinal Data

Degree	Salary
PHd	1
B.E.	~~2~~ 3
Master	~~3~~ 2
BCA	4
12	5

③ Interval data:

- A variable measured on an interval scale gives information about more or betterness as ordinal scales do.
- Temperature using celsius or Fahrenheit is a good example.

④ Ratio data:

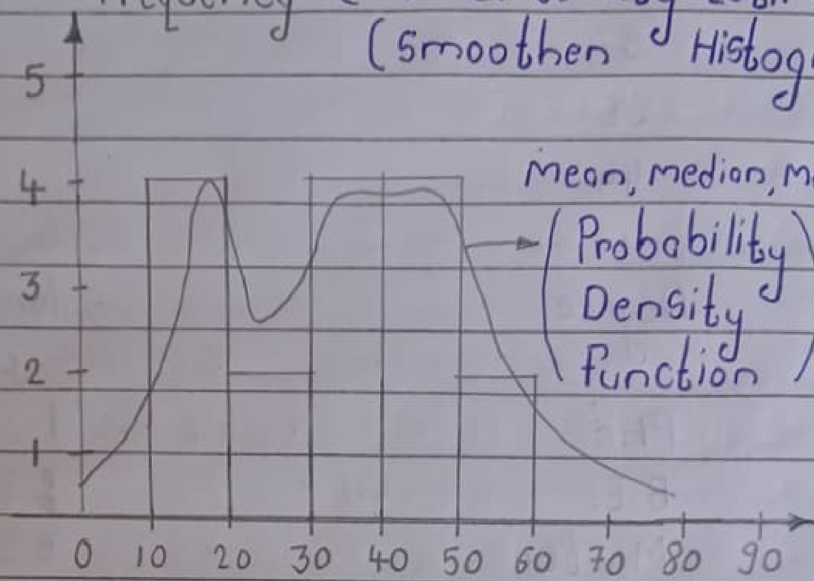
Something measured on ratio scale has the same properties that an interval scale has except, with a ratio scaling, there is an absolute zero point. Temperature measured in kelvin is an example.

• Histograms: Continuous

Ages = [10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 45, 50, 51]

Bins — 10

1) Frequency (Kernel density Estimator) (Smoothen Histogram)



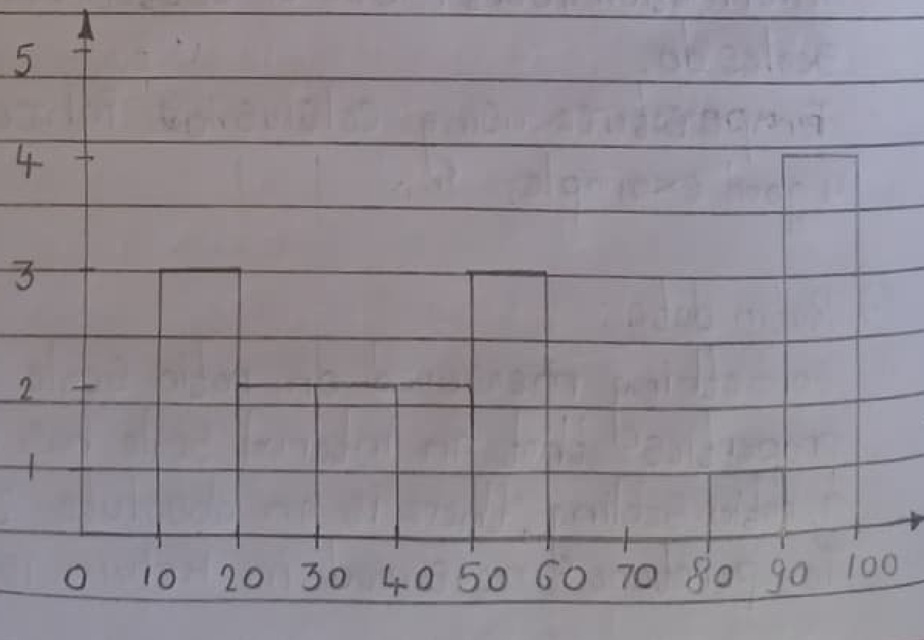
Mean, median, Mode

(Probability Density Function) PDF

0-50 → 0-5, 5-10, 10-15, 15-20, 20-25, 25-30, 30-35

Assignment

2) Eg: 10, 13, 18, 22, 27, 32, 38, 40, 45, 51, 56, 57, 88, 90, 92, 94, 99 (Bins-10)

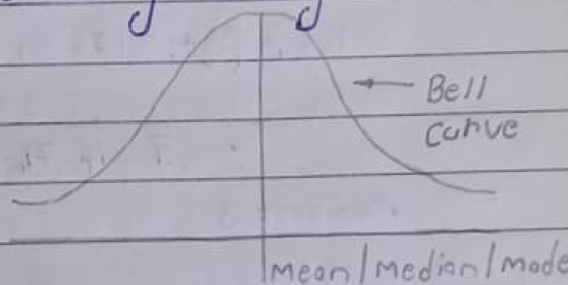


- Intermediate Stats

① Measure of Central Tendency
② Measure of Dispersion
③ Gaussian Distribution
④ Z-Score
⑤ Standard Normal Distribution
⑥ Central Limit Theorem

① Measure of Central Tendency — Central Position of dataset

- Mean
- Median — EDA & Feature Engineering
- Mode



Population (N) — Sample (n)

1) Mean

$x = [1,1,2,2,3,3,4,5,5,6]$ — Sample

$$\mu = \sum_{i=1}^{N} \frac{x_i}{N}$$

Population mean

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$

sample mean

$$= \frac{1+1+2+2+3+3+4+5+5+6}{10}$$

$$= \frac{32}{10} = 32$$

2) Median

1, 2, 2, 3, 4, 5

1, 2, 2, 3, 4, 5, 100

$$\bar{x} = \frac{1+2+2+3+4+5}{6} = \frac{17}{6} = 2.83$$

$$\bar{x} = \frac{1+2+2+3+4+5+100}{7}$$

$$= \frac{117}{7} = 16.71$$

Median

1, 2, 2, $\boxed{3}$, 4, 5, 100

odd or even

1, 2, $\boxed{2, 3}$, 4, 5

$$\frac{2+3}{2} = 2.5$$

$\bar{x} = 16.71$

$2.5 \approx 2.83$

median = 3

3) Mode : Highest frequency

1, 2, 2, 3, 3, 4, 4, 5, 6

1, 2, 2, 3, 3, 3, 4, 5, 6, 6, 7

$\Downarrow$ 3

$\Downarrow$ Error in Python 2.7

in new python

$\Downarrow$ [2, 3, 4]

Feature Engineering

NAN values $\Rightarrow$ Continuous values + outlier

Mean

Median

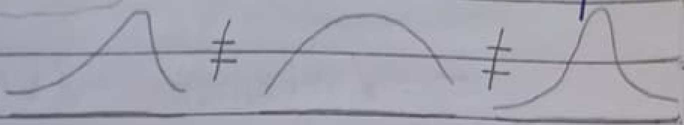
Categorical Variable

mode

② Measure of Dispersion : Spread → How data is spread

① Variance

② Standard deviation



① Variance   Benes Correction ; Degree of Freedom

Population Variance   Sample Variance

Population mean   Sample mean

$$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N} \qquad s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{\boxed{n-1}}$$

Eg :-

$x = [1, 2, 2, 3, 4, 5]$

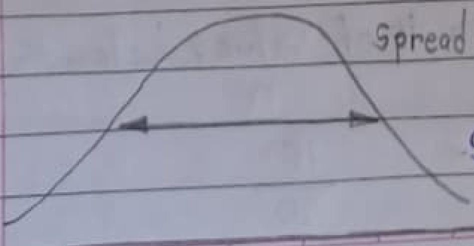
x	$\bar{x}$	$x - \bar{x}$	$(x-\bar{x})^2$	
1 1	2.83	-1.83	3.34	
2 2	2.83	-0.83	0.6889	$\frac{10.84}{5}$
2 3	2.83	-0.83	0.6889	
3 4	2.83	0.17	0.03	2.168
4 5	2.83	1.17	1.37	$n=6$
5	2.83	2.17	4.71	

$\bar{x} = 2.83$ 10.84

σ^2

Variance = 6.42 ; Spread ↑↑ σ^2 Variance ↑ Spread ↑

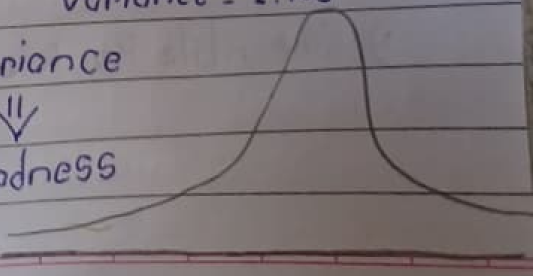
Variance = 2.168



Spread Variance

⇓

Spreadness



② Standard Deviation:

$\sigma = \sqrt{\text{variance}} = \sqrt{2.168} = 1.472$

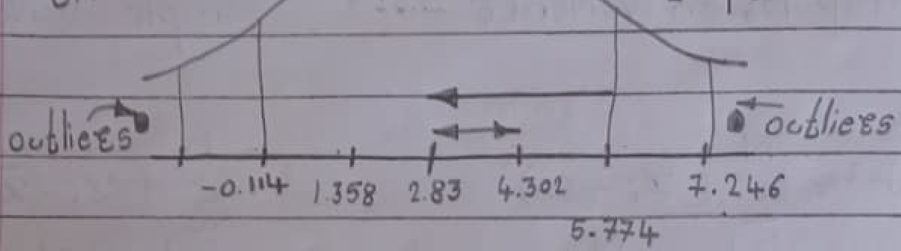
2.830 − 1.472 = 1.358

1.358 − 1.472 = 0.114

Gaussian Distribution 1, 2, 2, 3, 4, 5

⇓

Emperical Formula



2.83	
1.472	
4.302	
1.472	
5.774	
1.472	
7.246	

- Percentiles and Quartiles

Percentages: 1, 2, 3, 4, 5

% of numbers that are odd ?

% of odd = $\frac{3}{5}$ = 60 %

Percentile: (CAT, GATE, SAT)

Def^n:- A percentile is a value below which a certain percentage of observation lie

- 99 percentile means the person has got better marks than 99 % of the students.

Average ⇒ 5

Dataset :- 2, 2, 3, 4, $\boxed{5, 5}$, 5, 6, 7, 8, 8, 8, 8, 8, 9, 9, 10, 11, 11, 12

n = 20

What is the percentile ranking of 10 ?

1) Percentile Rank of x = $\frac{\#\text{ of values below } x}{n} \times 100$

$= \frac{16}{20}$

= 80 Percente

2) What value exists at percentile ranking of 25%?

$$\text{value} = \frac{\text{Percentile}}{100} \times (n+1) \quad \text{Demrag 2.2}$$

$$= \frac{25}{100} \times (21)$$

$$= 5.25 \rightarrow \text{Index}$$

value = 5 Quarties (25%)

- Five Number Summary

1. Minimum
2. First Quartile (25%) Q_1
3. Median
4. Third Quartile (75%) Q_3
5. Maximum

Inter Quarter Range (75% - 25%)
$Q_3 - Q_1$

Removing the outliers
[1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, ~~27~~]
(Lower Fence ⟷ Higher Fence)

Lower Fence = $Q_1 - 1.5(IQR)$
Higher Fence = $Q_3 + 1.5(IQR)$
$IQR = Q_3 - Q_1 = 7 - 3 = 4$

$(25\%)\,Q_1 = \frac{25}{100} \times (20) = 5^{th}$ index

$Q_1 = 3$

Lower Fence = $3 - 1.5(4)$
$= 3 - 6 = -3$

Higher Fence = $7 + 1.5(4)$
$= 7 + 6 = 13$

$(75\%)\,Q_3 = \frac{75}{100} \times 20 = 15^{th}$ index

$Q_3 = 7$

[-3 ⟷ 13]

Remaining
1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, ~~27~~

5 Number Summary

Minimum = 1

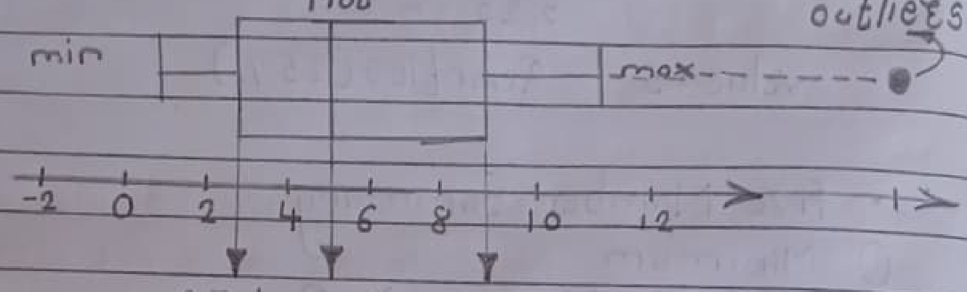
$Q_1 = 3$

Median = 5

$Q_3 = 7$

Maximum = 9

Box Plot



25 % median 75 %

IQR = (Inter Quartile Range)

- Distribution

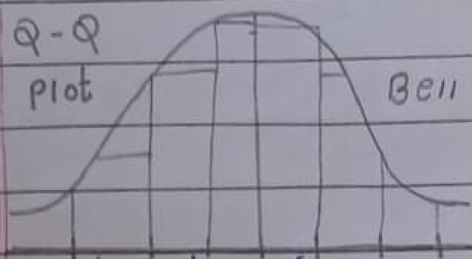
① Normal / Gaussian Distribution
② Standard Normal Distribution
③ Z - Score
④ Log Normal Distribution
⑤ Bernoullis Distribution
⑥ Binomial Distribution

Tossing Coin

p	p	p	p	p
q	q	q	q	q

① Gaussian / Normal Distribution 80-20%

Q-Q plot



Bell Curve

Properties (Power Law)

① Emperical Rule of Gaussian Distribution

Dataset → (IRIS Dataset) → Petal, Sepal length
weight of human Being, Height
68.2 - 95.4 - 99.7

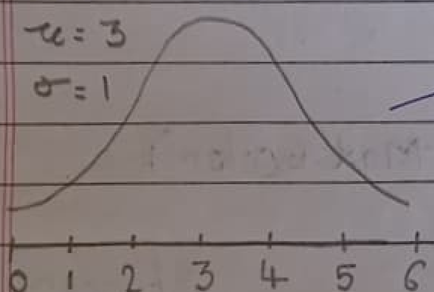
② Standard Normal Distribution

(1, 2, 3, 4, 5) $\mu = 3$

$\sigma = 1.414 \approx 1$

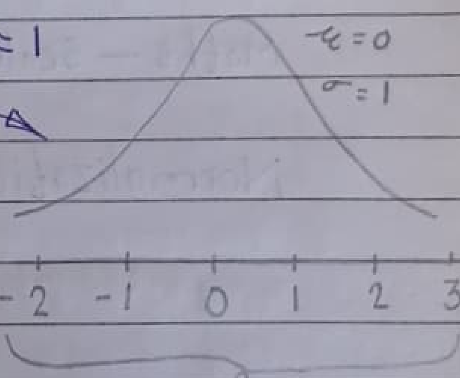
$\mu = 3$

$\sigma = 1$



$\mu = 0$

$\sigma = 1$



(1, 2, 3, 4, 5)

$$\left[\text{Z-score} = \frac{x - \mu}{\sigma}\right]$$

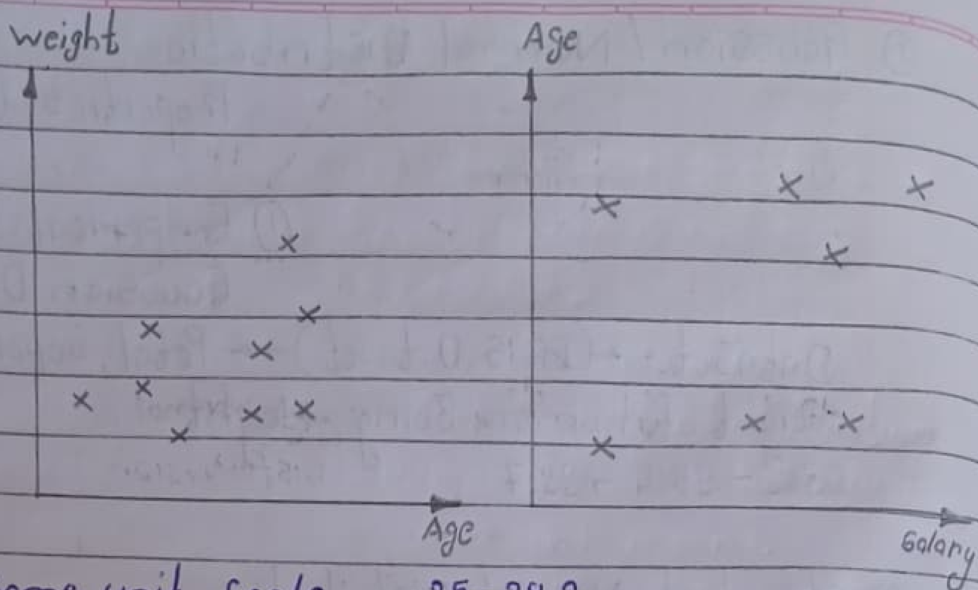
$\mu = 0$

$\sigma = 1$

$\frac{2-3}{1}$; $\frac{3-3}{1} = 0$; $\frac{1-3}{1} = -2$

Standardization Vs Normalization

Years	Kg	INR
Age	weight	Salary
25	75	25k
26	80	30k
28	85	40k
30	60	80k
32	70	



Same unit Scale $\frac{25-28.2}{2.56}$

Maths — Scale

Normalization (Min Max Scalar)

0 to 4
0 to 1

Standardization (ML)
Normalization (CNN)
Convolutional Neural Network

f_1	f'_1
2	0.14
5	0.571
6	0.71
8	1
1	0

$$x_{Nom} = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

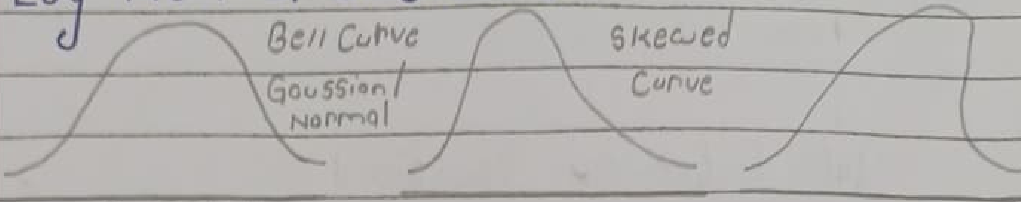
⇓

Min Max Scaler

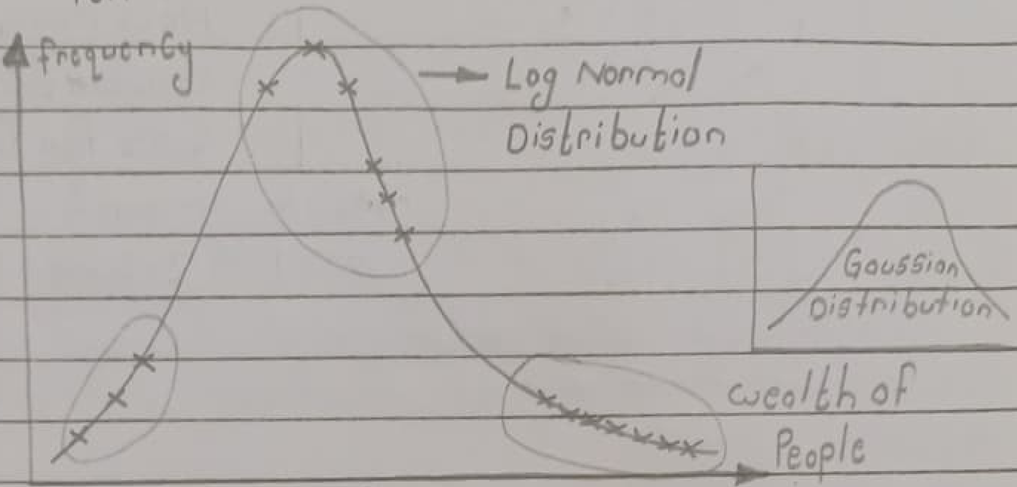
$= \frac{2-1}{8-1} = \frac{1}{7} = 0.142$ $\qquad \frac{8-1}{8-1} = 1$

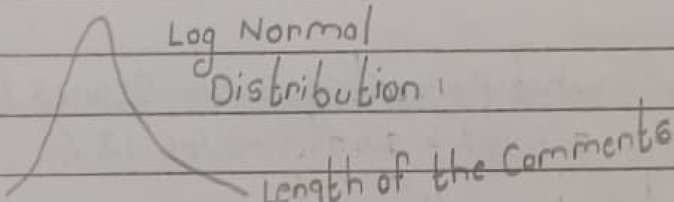
$= \frac{5-1}{8-1} = \frac{4}{7} = 0.571$ $\qquad \frac{1-1}{8-1} = 0$

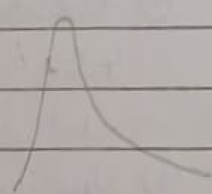
③ Log Normal Distribution



Emperical
Formula





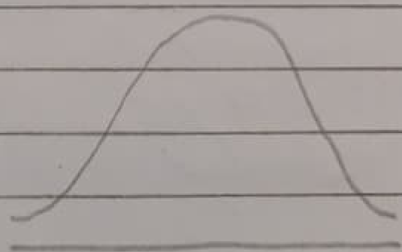


x = Log Normal Distribution

$y = \ln(x)$

$x = \exp(y) \to e^{y}$

x	$y = \ln(x)$
25	—
30	—
40	—
45	—



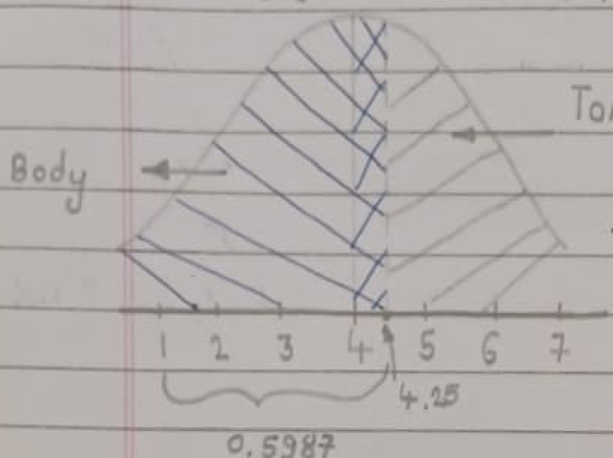
④ Bernoulli's Distribution

Day 2

① Z - Score $= \dfrac{x_i - \mu}{\sigma}$

Stats Interview Question

1-0.5+x $\{1,2,3,4,5,6,7\}$



How many standard deviation $\boxed{1}$ 4.25 fall from the mean?

$\mu = 4$

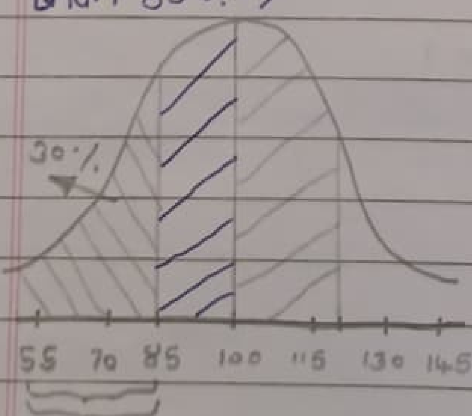
$\sigma = 1 \Rightarrow$ Z-Score $= \dfrac{x_i - \mu}{\sigma}$

$= \dfrac{4.25 - 4}{1}$

$= 0.25$

① Question:- what percentages of scores fall above 4.25? $\Rightarrow 1 - 0.5987 = 0.4013 \Rightarrow 40.13\%$

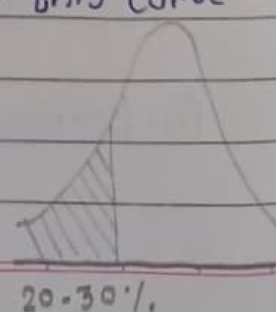
② Question:- In India the average IQ is 100, with a standard deviation of 15. what is the percentage of the population would you expect to have an IQ lower than 85? =>



Z - Score $= \dfrac{85 - 100}{15}$

$= \dfrac{-15}{15} = -1$

① Area under this Curve

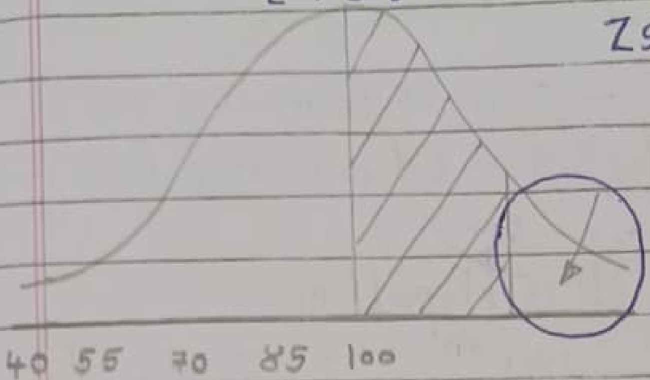


$0.5 - 0.15866 = 0.34143 \Rightarrow 34.14\%$

[Greather 100 less than 125]

$$Z score = \frac{125-100}{15} = \frac{\cancel{25}^{5}}{\cancel{15}_{3}}$$

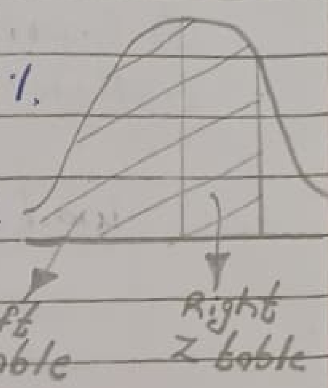
$$= 1.666$$



Ans = 0.4515

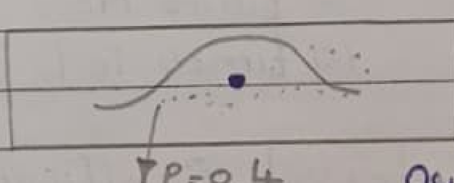
= 45.15 %

$0.5 - 0.4515 = 0.0485 = 4.8\%$



- P value, Hypothesis Testing, Confidence Interval

out of all 100 touches the no. of touches is



space bar key 80

out of all 100 touches, the no. of items times is 40 times.

- Hypothesis Testing, C.I, Significance value Together

Coin → Test whether the coin is a fair coin or not by performing 100 tosses.

$P(H) = 0.5$

$P(T) = 0.5$

Criminal is → Count
SHOLAY
$P(H) = 100\%$ $P(T) = 0\%$

Domain Expertise
Health Care
Covid vaccine Test
Error

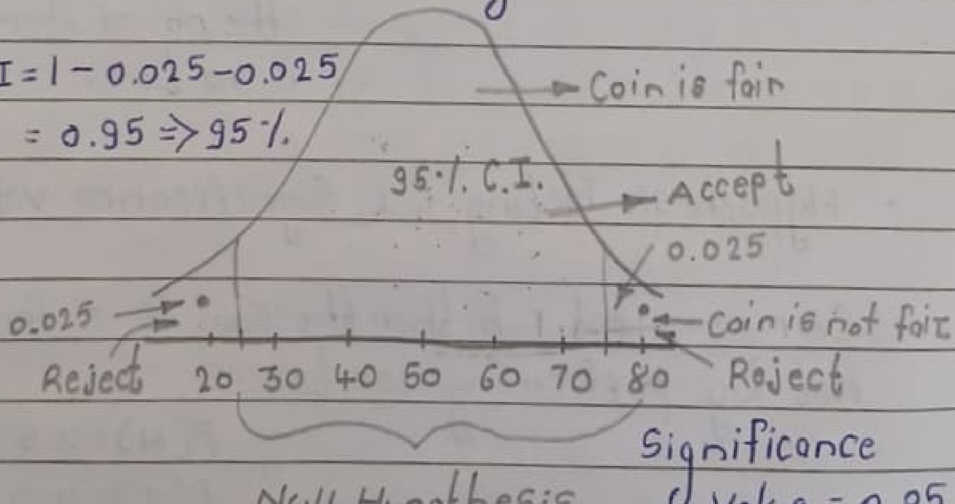
- Hypothesis Testing

① Null Hypothesis – Coin is fair – (H_0)
② Alternate Hypothesis – Coin is not fair – (H_1)
③ Experiments
④ Reject or Accept Null Hypothesis

100 tosses →
- 50 times Heads, 50 times Tails } Coin is fair
- 60 times Head, 40 times Tail } Coin is fair
- 30 times Head, 70 times Tail }

- Confidence Interval, Significance values

$CI = 1 - 0.025 - 0.025$
$= 0.95 \Rightarrow 95\%$

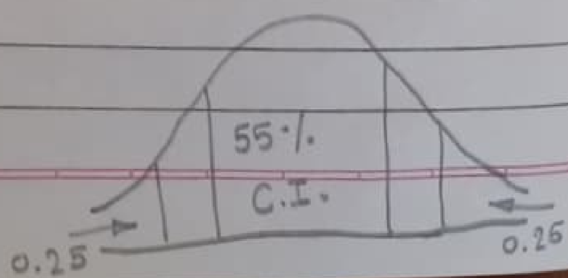


Significance Value = 0.05

$\alpha = 0.45$

Medical $\alpha \uparrow\uparrow$

$\frac{0.45}{2} = 0.225$



Real World Project

Hypothesis Testing

Location A — ATM machine

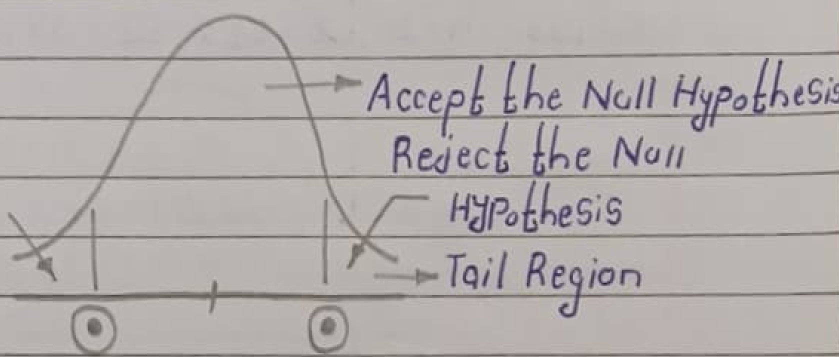
↓ 10 km

Location B — ATM should be opened here or not

Bank A

Data Analyst

- Confidence Interval



- Point Estimator

The value of any statistics that estimates the value of a parameter is called Point Estimator.

Estimate

Sample mean $\boxed{\bar{x}} \xrightarrow{\text{Estimate}} \boxed{\mu}$ Population mean

2.8, 2.9, 2.5, 3.5 → 3

Inferential

(Sample mean → Population)

• Confidence Interval

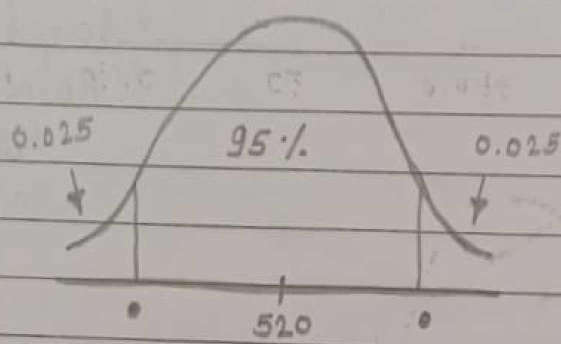
t test

Point Estimate $\pm$ Margin of Error $\Rightarrow$ Population.

Q. On 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% CI about the mean?

Ans :-

$\sigma = 100$, $n = 25$, $\bar{x} = 520$, $CI = 95\%$, $\alpha = 0.05$ ← $1 - 0.95$



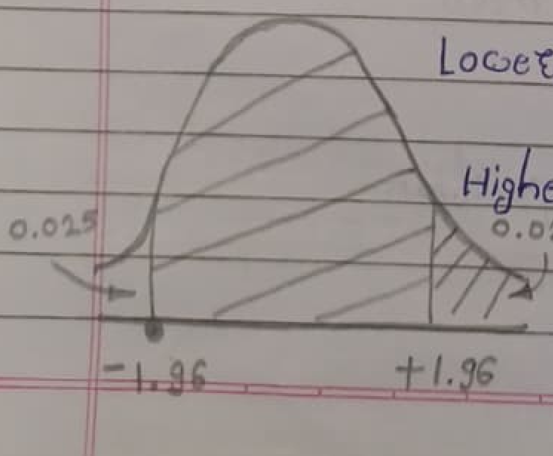
① Population Std is given {Z score} → z table

Point Estimator $\pm$ Margin of Error → C.I.

$$\bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$ ← standard Error

Lower Fence C.I. $= \bar{x} - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

Higher Fence C.I. $= \bar{x} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$



$1 - 0.025$
$= 0.9750$

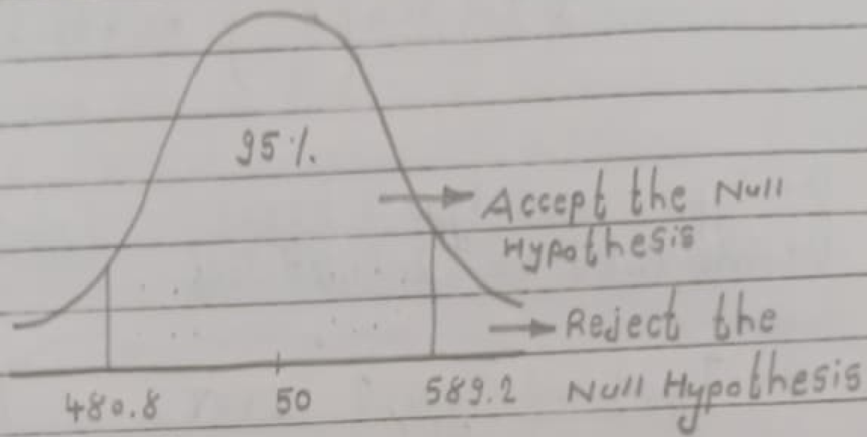
$Z_{\frac{0.05}{2}} \Rightarrow Z_{0.025} = 1.96$

$$\text{Lower Fence} = 520 - (1.96) \times \frac{100}{\sqrt{25}}$$

$$= 520 - (1.96) \times 20$$

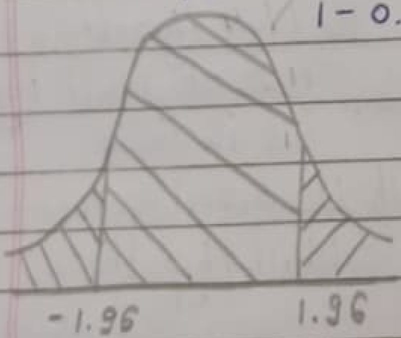
$$= 480.8$$

$$\text{Higher Fence} = 520 + (1.96) \times 20 = 559.2$$



$$Z_{\alpha/2} = Z_{0.025}$$

$$1 - 0.025 = 0.9750$$



$$\langle 488.64, 581.36$$

- 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% CI about the mean?

Ans $\bar{x} = 520$, $s = 80$, $\alpha = 0.05$, $n = 25$

t-test → t-table (Because Population sd is not given)

PAGE No. DATE

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

$t_{0.025}$

- Degree of freedom = $n-1 = 25-1 = 24$

$$\bar{x} \pm 2.064\left(\frac{80}{5}\right) \Rightarrow 486.976 \leftrightarrow 553.024$$

1). Type 1 and Type 2 Error
2). One Tailed Vs 2 Tailed Test

1) Type 1 and Type 2 Error

Reality check

$H_0 \Rightarrow$ Coin is fair
$H_1 \Rightarrow$ Coin is not fair

① Null Hypothesis is True or Null Hypothesis is False

Null Hypothesis	
$H_0 \rightarrow$ The Criminal is not guilty $H_1 \rightarrow$ -11- is guilty	Decision [Experiments] Null Hypothesis is True or False

Outcome 1 :-
We reject the Null Hypothesis in reality if it is False → Yes

Outcome 2 :-
We reject the Null Hypothesis when in reality it is true. → No → Type 1 Error ×

Outcome 3 :-
We accept the Null Hypothesis when in reality it is false. → Type 2 Error ×

Confusion matrix
(stock market is going
to Crush)

Outcome 4 :- We accept the Null Hypothesis when in reality it is True.

2) 1 Tail and 2 Tail Test

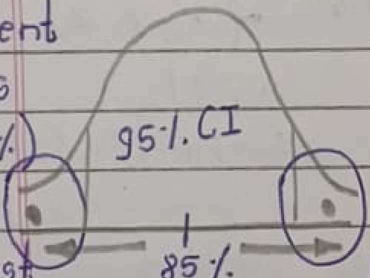
Eg:- Colleges in Karnataka has an 85% placement rate. A new College was recently Opened and it was found that a sample of 150 students had a placement rate of 88% with standard deviation of 4%. Does this College has a different Placement rate?

↳ 85% $\alpha = 0.05 = 95\%$ CI

(Placement rate less than 85%)

(Placement rate greather than 85%)

2 Tail Test



1 Tail Test   1 Tail Test

1. Z test Hypothesis Testing
2. T test Hypothesis Testing
3. Significance Value and P value
4. ANOVA Test
5. CHI SQUARE Test
6. Practical

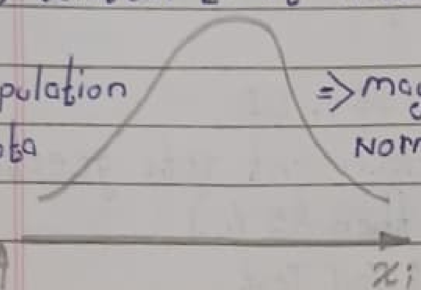
Day 3

① Centeal limit Theoeem
② Inferential Statistics
- z test {z table}
- t test {t table}
- z test propotion population
- chi Squaee (Categorical Test)
- ANNOVA (F Test)

① Centeol Limit Theoeem

$\boxed{n \geq 30}$

{ Population data



⇒ may be Gaussian / Normal Distr → $[x_1, x_2, x_3 \dots x_{30}]$

Sample 1 $\bar{x}_1$

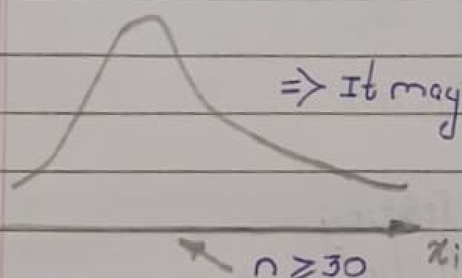
Sample 2 $[x_1, x_2, x_3, x_4 \dots x_{30}] \rightarrow \bar{x}_2$

⋮

Sample n $\bar{x}_3$

⋮

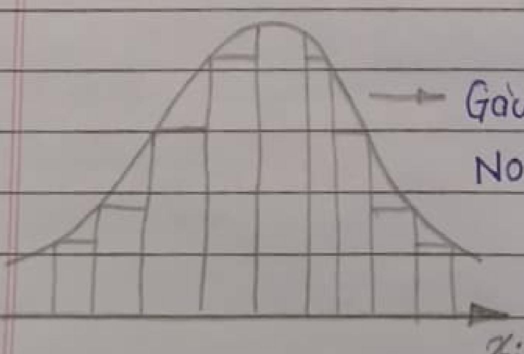
$\bar{x}_m$



⇒ It may not

$n \geq 30$

Sample mean distribution



→ Gaussian Distribution
Normal Distribution

(2) Inferential Statistics {Data Analyst, Data Scientist}

- 100 K ⇒ T-shirt → NO → Sample data → XL, L, Small
- iNeuron → Meetup → Hitesh → 300-400 people → T-shirts → ordered

$$\left[\begin{array}{ll} 20\%\ L & 10\%\ XXL \\ 10\%\ XL & 60\%\ \text{Medium} \end{array}\right] \begin{array}{l} 500 \\ \text{t-shirts} \end{array}$$

Next Event

- ATM $\mu, 5$
- Measure the size of $\bar{x}$ entire sharks C.I. [ ]
- Amazon delivery {Percentile, Quartiles}

- Hypothesis Testing

1) A factory has a machine that fills 80 ml of baby medicine in a bottle. An employee believes the average amount of baby medicine is not 80 ml, using 40 samples, he measures the average amount dispersed by the machine to be 78 ml with a standard deviation of 2.5

(a) State Null and Alternate hypothesis

(b) At a 95% CI, is there enough evidence to support machine is not working properly.

Ans Step 1 :- Given - $n = 40$, $\bar{x} = 78$, $s = 2.5$

$H_0 = \mu = 80$ {Null Hypothesis}

$H_1 = \mu \neq 80$ {Alternate Hypothesis}

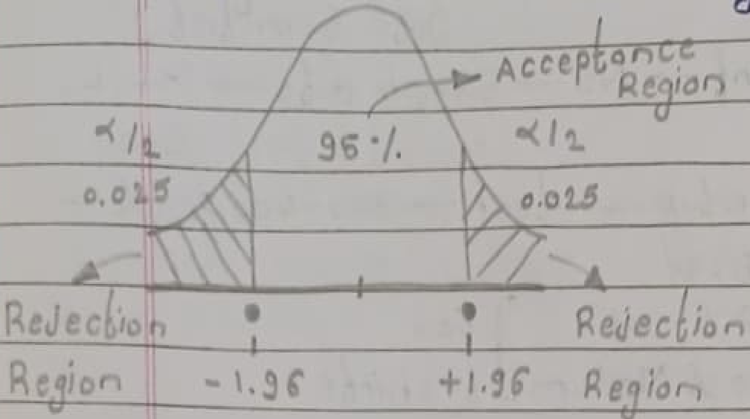
Step 2 :-

$\alpha = 0.05$ $(1 - 0.95)$

C.I = 95%

why Z test?	why t test?
① $n \geq 30$	① $n < 30$
② Population std or sample std	② sample std

Step 3 :- Decision Boundary



$$\boxed{Z = \frac{x_i - \mu}{\sigma/\sqrt{n}}}$$

n = 1

Step 4 :- Calculate Test Statistics

$$Z = \frac{\bar{x} - \mu}{\underbrace{S}_{\text{Sample standard deviation}}/\sqrt{n}}$$

$S/\sqrt{n}$ → standard Error

$$= \frac{78 - 80}{2.5/\sqrt{40}} = \frac{-2 \times \sqrt{40}}{2.5} = \frac{-2}{2.5} \times 6.32$$

$$= -5.05$$

Step 5 :- State the Results

Decision Rule :- If $Z = -5.05$ is less than -1.96 or greater than 1.96, then reject the null hypothesis with 95% C.I

Reject Ho Null Hypothesis {There is some fault in the machine}

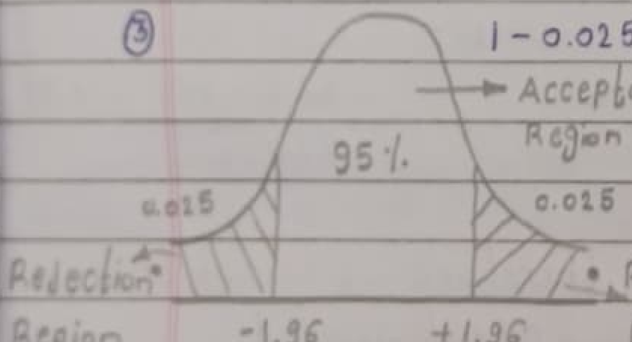
2) In the population the average IQ is 100 with a standard deviation of 15. A team of Scientists wants to test new medication to see if it has a +ve or -ve effect, or no effect at all. A sample of 30 participants who have taken the medication has a mean of 140. Did the medication affect Intelligence? C.I = 95%.

Ans: $\sigma = 15,\ n = 30,\ \bar{x} = 140$

① $H_0 = \mu = 100$

$H_1 : \mu \neq 100$

② $\alpha = 0.05$; C.I = 95%.

③ 

If z is less than −1.96 or greater than 1.96, reject the Null Hypothesis.

④ $$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{140 - 100}{15/\sqrt{30}} = 14.60$$

⑤ $14.60 > 1.96$, Reject the Null Hypothesis

Python{ztest, P value}

PAGE No.
DATE

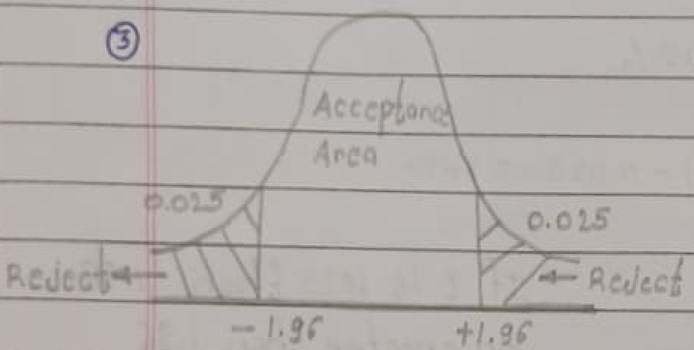
3) A Complain was registered, the boys in the Muncipal Primary School are underfed. Average weight of boys of age 10 is 32 Kgs, with S.D = 9 kgs. A Sample of 25 boys was selected from the muncipal school and the average weight was found to be 29.5 kgs ? with C.I = 95 %, check whether it is True or False ?

Ans $\mu = 32$ kgs, $\sigma = 9$ kgs, $n = 25$, $\bar{x} = 29.5$, $\alpha = 0.05$ 1-0.95

① $H_0 = \mu = 32$ ② $\alpha = 0.05$

$H_1 = \mu < 32$

③



④ $$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

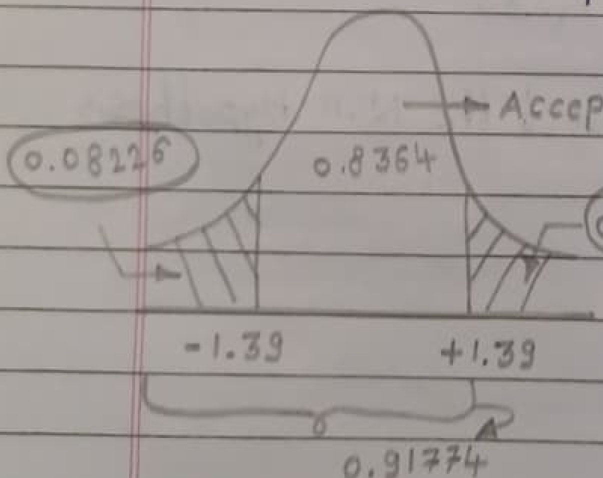
$$= \frac{29.5 - 32}{9 / \sqrt{25}} = -1.39$$

⑤ Conclusion: -1.39 therefore we accept the Null Hypothesis. So, the boys all not underfed.

- Significance Value =>

P value

$\alpha = 0.05$



1 - 0.08226 - 0.08226

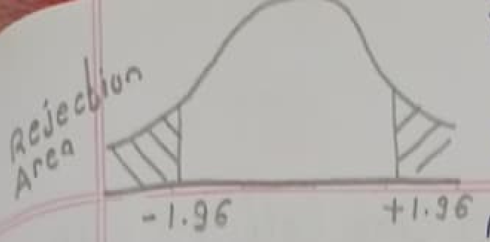
0.08226 + 0.08226 = 0.1645

0.1645 > 0.05 significance value

⇓

P > 0.05 → Accept the Null Hypothesis

1 - 0.91774 = 0.08226



Z test

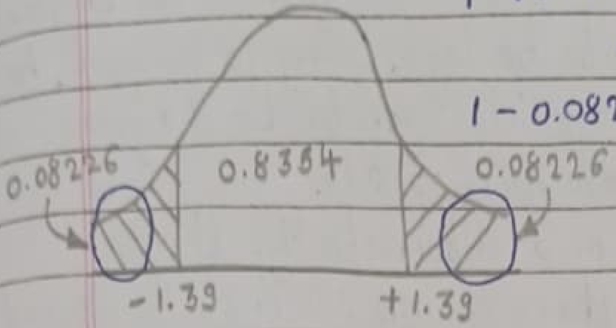
$\Downarrow$

P value

Here

P value = 0.08226 + 0.08226

= 0.16



1 - 0.08226 - 0.08226

Domain

↓

0.1645 > Significan Value

$\Downarrow$

Accept the Null Hypothesis.

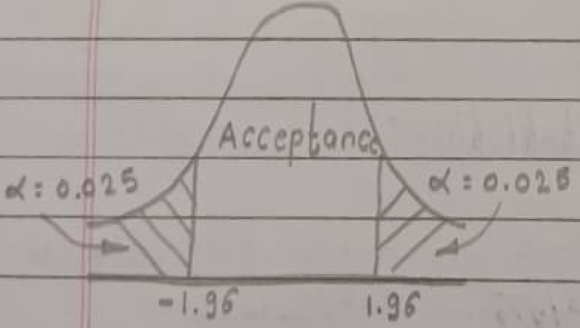
4) The average weight of all residents in town xyz is 168 Lbs. A nutnotionist believes the true mean to be different. She measured the weight of 36 individuals and found the mean to be 169.5 Lbs with a standard deviation of 3.9

(a) At 95% CI is there enough evidence to discard the Null Hypothesis?

Ans $H_0 : \mu = 168$ $n = 36$ $\bar{x} = 169.5$ $s = 3.9$

$H_1 : \mu \neq 168$ $C = 0.95$ $\alpha = 1 - C.I. = 0.05$

z-test, t-test



$$z = \frac{\bar{x} - \mu}{s/\sqrt{n}} = 2.31$$

2.31 > 1.96 ; Reject the Null Hypothesis

5) A Company manufactures bike battries with an average life span of 2 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 there enough evidence to discard the Ho ?

Ans

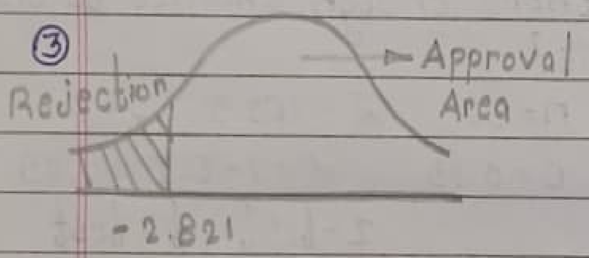
(1) $H_0 : \mu \geq 2$, $n = 10$, $\bar{x} = 1.8$, $s = 0.15$ {Sample std is given}

$H_1 : \mu < 2$ $\quad$ < 30

t-test

(2) $\alpha = 0.01$, $\alpha = 1 - C.I = 1 - 0.99 = 0.01$

(3)



Degree of freedom $= n-1$
$= 9$

(4) Calculate t-test statistics :

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{1.8 - 2}{0.15/\sqrt{10}} = \frac{-0.2}{0.15/3.1622} = -4.216$$

(5) Conclusion

$-4.216 < -2.821$; Reject the Null Hypothesis

$\Downarrow$

Z test with Proportions

PAGE No. | DATE

6) A test Company believes that the percentage of residents in town xyz. That owns a cell phone is 70 %. A marketing manager believes that this value to be different. He Conducts a Survey of 200 individuals and found that 130 responded yes to owning a cell phone.

ⓐ State Null and Alternate Hypothesis ?

ⓑ At a 95% C.I, is there enough evidence to reject the Null Hypothesis ?

Ans ① H_0 $P_0 = 0.70$, $n = 200$, $x = 130$

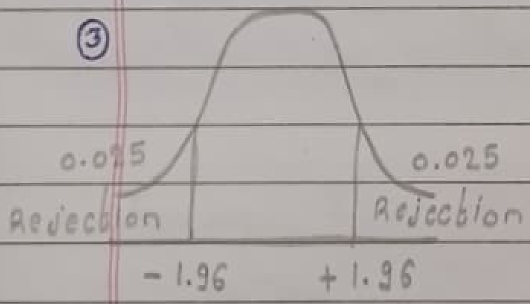
H_1 $P_0 \neq 0.70$

$$\hat{P} = \frac{x}{n} = \frac{130}{200} = \frac{13}{20} = 0.65$$

$$q_0 = 1 - P_0$$

② $\alpha = 0.05$, $C.I = 95\%$

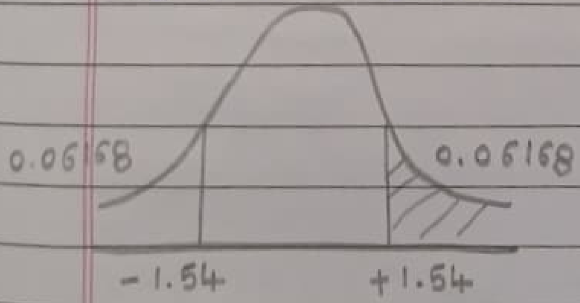
③



$$Z_{test} = \frac{\hat{P} - P_0}{\sqrt{\frac{P_0 q_0}{n}}}$$

$$= \frac{0.65 - 0.70}{\sqrt{\frac{0.7 \times 0.3}{200}}} \approx -1.54$$

At 95% C.I there is $-1.54 > -1.96$, So we accept the Null Hypothesis



P value

$2 \times 0.06160 > 0.05$

Accept Null Hypothesis

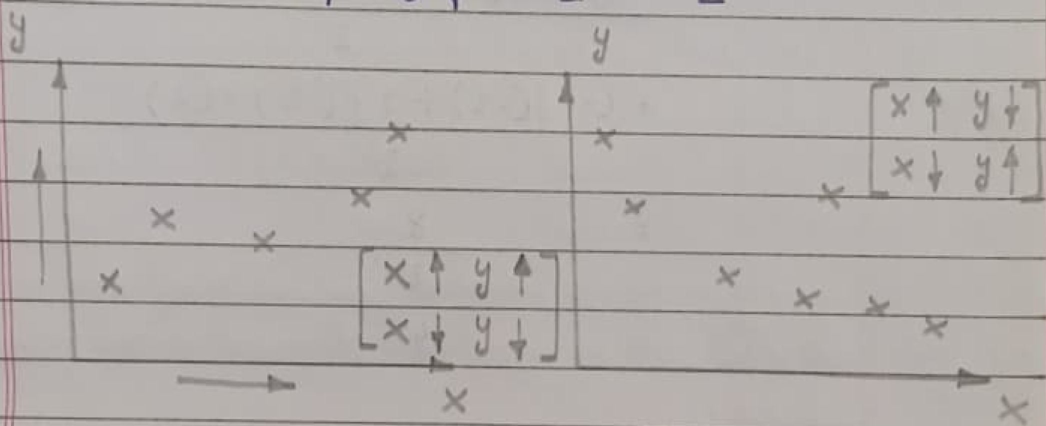
$1 - 0.93872 = 0.06168$

Day 4

① Covariance
② Pearson Correlation Coefficient
③ Spearmean Rank Correlation Coefficient
④ CHI SQUARE TEST
⑤ ANNOVA (F-Test)

① • Covariance

$x\uparrow$	$y\uparrow$	x	y	[Quantity the relationship between x & Y]
$x\uparrow$	$y\downarrow$	–	–	
$x\downarrow$	$y\uparrow$	–	–	
$x\downarrow$	$y\downarrow$	–	–	



$$Cov_{x,y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N-1} \qquad Var(x) = \frac{\sum (x_i - \bar{x})^2}{N-1}$$

$$Cov(x,x) = \frac{\sum (x_i - \bar{x})^2}{N-1}$$

$$= \frac{\sum (x_i - \bar{x}) \times (x_i - \bar{x})}{N-1}$$

$$Var(x) = \sum_{i=1}^{n} \frac{(x-\bar{x})^2}{n-1} \Rightarrow \sum_{i=1} \frac{(x-\bar{x}) \times (x-\bar{x})}{n-1}$$

$$\Downarrow$$

$$cov(x,x) = \sum_{i=1}^{n} \frac{(x-\bar{x}) \times (x-\bar{x})}{n-1}$$

$x\uparrow\ y\uparrow$ Positively
$x\downarrow\ y\downarrow$ Correlation

x	y
2	3
4	5
6	7

$\bar{x} = 4, \bar{y} = 5$

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

$$= \frac{(2-4)(3-5)+(4-4)(5-5)+(6-4)(7-5)}{2}$$

$$= \frac{(-2)(-2)+0+(2)\times(2)}{2}$$

$$= \frac{8}{2}$$

$$= 4$$

$x\uparrow\ y\downarrow$ => -ve Correlation => -ve value
$x\downarrow\ y\uparrow$

$Cov(x,y) = 500$
$Cov(y,z) = 600$

Disadvantage Covariance

Cov(x,y) => the value of -ve value => Limit -350 +500 -300 -400 +1000

Relationship [-1 to 1]

② • Pearson Correlation Coefficient

[−1 to 1]

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

The more the value towards 1 more +ve it is Corelated

Dataset :- 1000 features

x y z A B C — Independent Features

O/P — Dependent Features

$x, y \Rightarrow 99\%$

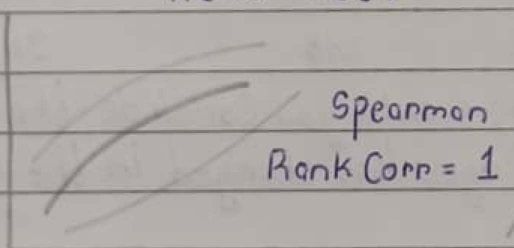
90% 0.9

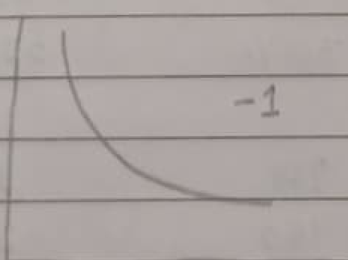
−ve Correlation keep it

③ • Spearman Rank Correlation

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

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





④ • CHI Square

The chi Square test claims about Population proportions.

It is a non parametric test that is performed on Categorical (nominal or ordinal) data.

1) • In the 2000 U.S Census, the ages of individuals in a Small town where found to be the following.

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

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

<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?

Ans

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

$n=500$

95 C.I

observed	121	288	91
Expected	100	150	250

① H_0 = the data meets the expected distribution
H_1 = the data do not meet the expected dis

② stats Alpha :- $\alpha = 0.05$

③ Calculate the degree of freedom

$$df = n-1 = 3-1 = 2 \Rightarrow 3 \text{ Categories}$$

(2) Ans:- 6.3 (Accept H_0)

PAGE NO.
DATE

(4) Decision   chi Square Test

If x^2 is greater 5.99 than, Reject Ho

(5) Calculate chi Square test

$$x^2 = \sum \frac{(f_0 - f_e)^2}{f_e} = \frac{(121-100)^2}{100} + \frac{(288-150)^2}{150} + \frac{(91-250)^2}{250}$$

$x^2 = 232.494$

232.494 > 5.991, Reject the Null Hypothesis.

2) A School Principal would like to know which days of the week students are most likely to be absent. The principal expect the students will be absent equally during the 5-day School week. The principal selects the random sample of 100 teachers asking them which day of the week they had the highest number of students absents, Occurs with equal frequencies (use 95 C.I)

	Monday	Tuesday	Wed	Thur	Fri
Observed	23	16	14	19	28
Expected	20	20	20	20	20

Day 5

① ANOVA (F-Test)
② EDA - (Solve Some Example)

ANOVA :- Analysis of Variance
ANOVA is a statistical method used to compare the means of 2 or more group.

ANOVA :-
① Factors (variables) Medicine
② Levels {Dosage}
Anxiety reducing

factor :- Dosage
Levels :- 0mg, 50mg, 100 mg

0 mg	50 mg	100 mg
9	7	4
8	6	3
7	6	2
8	7	3
8	8	

- Types of ANOVA

① One way ANOVA :- One factor with atleast 2 levels, levels are independent.

② Repeated Measures ANOVA :- One factor with atleast 2 levels, but levels are dependent.

Factor Running kms
Levels

Day 1	Day 2	Day 3
6	18	5

- Factorial ANOVA :-
  Two or more factor each of which with atleast 2 levels, levels can be either independent, dependent, or both (mixed)

	Day 1	Day 2	Day 3
Mean	9	7	4
	8	6	3
	7	5	2
Women	8	7	3
	8	8	4
	9	7	3

One way ANOVA (F-test) ⇒ Inferential stats
⇓
Comparing means of 2 or more groups.

- Researchers want to test a new anxiety medication. They split participants into 3 Conditions (0mg, 50mg, 100mg), then ask them to rate their anxiety level on scale of 1-10. Are there any difference between the 3 Conditions using $\alpha = 0.05$

0 mg	50 mg	100 mg
9	7	4
8	6	3
7	6	2
8	7	3
8	8	4
9	7	3
8	6	2

① $H_0 = \mu_{0mg} = \mu_{50mg} = \mu_{100mg}$
H_1 = not all μ's are equal

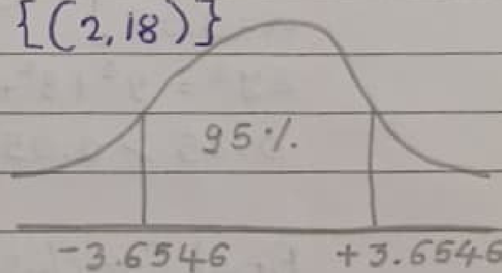
② state α and C.I.
$\alpha = 0.05$; C.I = 95%.

③ Calculate degree of freedom

$N = 21 \quad n = 7$
$a = 3 \rightarrow$ (No. of levels)

$df_{Between} = a - 1 = 3 - 1 = 2$
$df_{with} = N - 1 = 21 - 3 = 18$
$df_{Total} = N - 1 = 21 - 1 = 20$

④ state Decision Rule



$df_{Between} = a - 1 = 3 - 1 = 2$
$df_{within} = N - a = 21 - 3 = 18$

If F test is greater than 3.5546, Reject the Null Hypothesis.
If F test is less than −3.5546 −〃−

⑤ Calculate F Test statistics

$$F_{Test} = \frac{MS_{between}}{MS_{within}} = \frac{49.34}{0.57}$$

	SS	df	MS	Ftest
Bet^n	98.67	2	49.34	86.56
within	10.29	18	0.57	
Total	108.96	20		

$N = 21, n = 7$

$$SS_{between} = \frac{\Sigma(\Sigma a_i)^2}{n} - \frac{T^2}{N}$$

$\Sigma(\Sigma a_i)^2 = (9+8+7+8+8+9+8)^2 + (7+6+6+7+8+7+6)^2 + (4+3+2+3+4+3+2)^2$

$= 57^2 + 47^2 + 21^2$

1) $SS_{Between} = \frac{57^2+47^2+21^2}{7} = \frac{125^2}{21} = 98.67$

2) $SS_{within} = \Sigma y^2 - \frac{\Sigma(\Sigma a_i)^2}{n}$

$\left.\begin{matrix} P\ 0.75 \\ \alpha = 0.05 \end{matrix}\right\} = \Sigma y^2 - \left[\frac{57^2+47^2+21^2}{7}\right] = 10.29$

$\Sigma y^2 = 9^2+8^2+7^2+8^2+8^2+9^2 \text{----} + 2^2 = 853$

$0.75 > 0.05 \Rightarrow$ Accept

Final Conclusion

$86.56 > 35.546$; So we reject the Null Hypothesis

$\left.\begin{matrix} H_0 = \mu = \text{Some value} \\ H_1 = \mu \neq \text{Some value} \end{matrix}\right\} \longrightarrow 95\%$ C.I

$H_0 = \mu_{virg} = \mu_{setosa} = \mu_{---}$

H_1

	Pvalue	Reject the Null
0.0118	$0.0228 < 0.05$	Hypothesis
0.0118	$1 - 0.025 = 0.975$	
0.0228		