Statistics

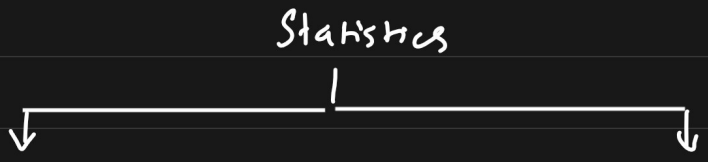
Statistics is the science of collecting, organizing and analyzing data.

Data : "facts or pieces of information"

Eg : Height of students in a classroom
→ {175cm, 152cm, 140cm, 130cm, 155cm}

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

Two Types



① Descriptive Stats

- It consists of organizing and summarizing of data.
- Eg : Pdf, Histogram, Box plot, Bar chart, Pie charts

② Inferential Stats

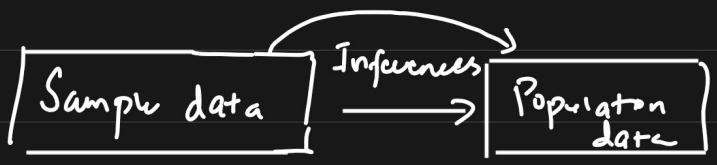
- It consists of using that you've measured to form conclusions
- Eg : Hypothesis Testing, p value, Z test, t-test, Anova, Chisquare

Eg : Lets say there are 20 maths classes at your university and you've collected the ages of students in one class.
Ages {21, 20, 18, 34, 17, 22, 24, 25, 26, 23, 22}

Descriptive stats : What is the average age of student in min = mode

your maths class?

Inferential question : Are the ages of students in this maths classroom similar to what you would expect in a normal maths class at this university?



Population And Sample Data. → Inferential Statistics → Results

Exit Polls ←

Elections : Punjab {AAP, Congress}

↙ Population of Punjab

1000 people ⇓

Eg: 2015 → Data Scientist

Jackets → Size ✗

10K, 40K → Christmas

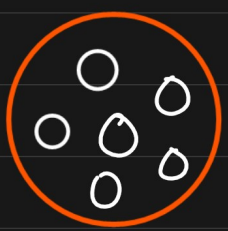
1-2% ⇒ Jackets

Population (N) ✓ Sample (n) ✓

Sampling Techniques

⇓

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 groups

Clusters ⇑ groups

Gender → Male, Female

Age groups → 0-18, 18-35, 35-60

Blood Groups

Tax Slabs

Courses

Education qualification

Thanos

Australia

↳ Snap

Customs

③ Systematic Sampling

(N) → Select every nth individual

6th

3-4th

Stratified

Eg: Survey → Mail

(SBI credit card)

④ Convenience Sampling : Only those people who are interested will only be participating.

Eg: Data Science → AI

Youtube survey →

Healthcare Disease

{ Blind people }

{ Economics } → DATA Science

→ RBI → House hold Survey → Female ← UPSC

Exit Poll : Stratified + Random Sampling

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, x, ÷ }

② Qualitative Variable.

↳ Eg: Gender → Male, Female { Based on some {characteristics} we can derive categorical variables }

{ Quantitative ⟶ Qualitative Variable. }

Eg: IQ

0-10 ⇓ Less IQ

10-50 ⇓ Medium IQ

50-100 ⇓ Good IQ

Quantitative

- Discrete variable
  - Eg: Whole number
  - Eg: No. of Bank accounts
  - {2, 3, 4, 5, 6, 7} 2.5, x 2.75 x
  - Eg: Total No. of children in a family
- Continuous Variable.
  - Eg: Height = 172.5, 162.5cm, 163.5cm.
  - Rainfall: 1.35, 1.25, 1.75, 2.25cm
  - Weight
  - Temp

Stock price.

Eg: 2, 3, 4, 5

2.5, 2.75

Eg: Total no. of Employees in a Company Eg: 10K,

Ass:

① What kind of variable Marital Status is? Categorical

② What kind of variable Nile River length is? Continuous quantitative

③ What kind of " Movie duration is? " "

④ What kind of variable IQ is? " "

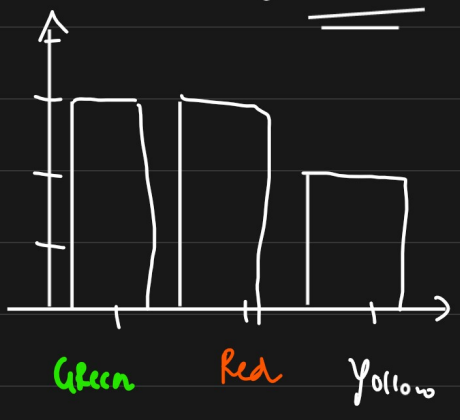
Frequency Distribution

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

Colors	Frequency
Green	3
Red	3
Yellow	2

① BAR GRAPH Frequency

Bar chart



⑧ Variable Measurement Scales

4 types of Measured Variable.

① Nominal data {Categorical data} → Ranking is not that important

Eg: Colors, Gender, Types of flowers

② Ordinal data

Percentiles

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

Ordinal Data.

PHd → ⇓ ↙ {NLP}

Degree		Salary
PHd	1	✓
B.G	3	✓
Master	2	✓
BCA	4	✓
12	5	✓

③ Interval data ✓⇓

Assignment ⇓

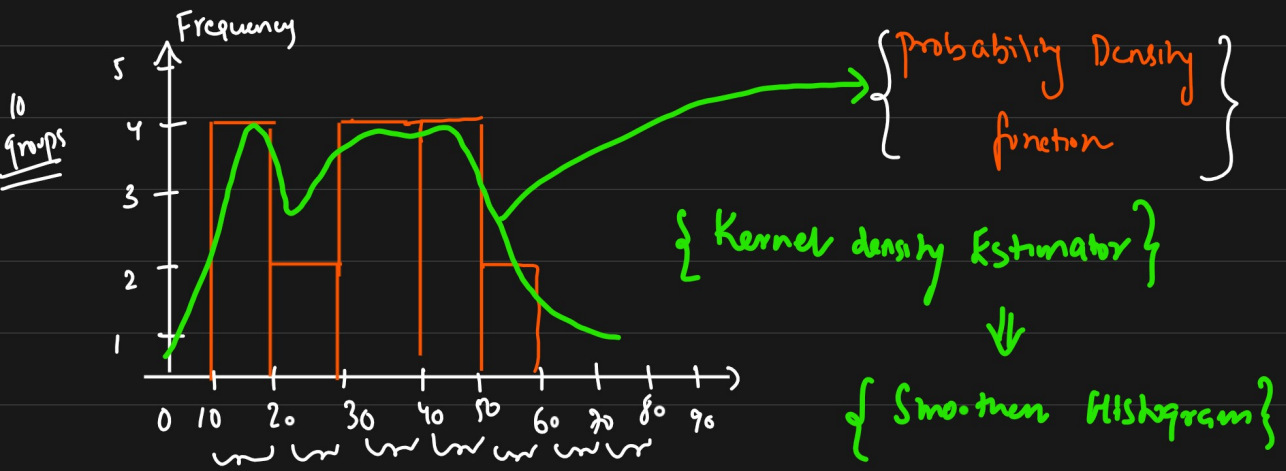
④ Ratio data ✓

⑧ Histograms ÷ Continuous

Ages = {10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 43, 50, 51}

Histogram → Bins = 10

Mean, Median, Mode.



0 - 80 → 0-5, 5-10, 10-15, 15-20, 20-25, 25-30, 30-35

Assignment

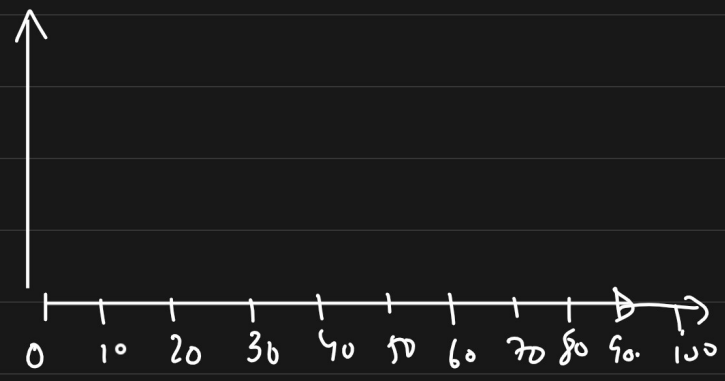
Eg: 10, 13, 18, 22, 27, 32, 38, 40, 45, 51, 56, 57, 88, 90, 92, 94, 99

bins = 10

0-10 10-20 20-30 30-40

40-50 50-60 60-70

70-80 80-90 90-100

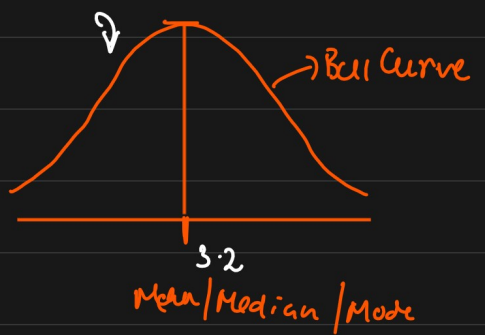


Intermediate Stats

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 the dataset

1. Mean ✓
2. Median ✓ {EDA & Feature Eng}.
3. Mode ✓



Population (N)

$X = \{1,1,2,2,3,3,4,5,5,6\}$

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

μ ↓ Population mean

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

$$= \frac{32}{10} = 3.2$$

Sample (n)

Sample → $\bar{x}$ ↓ Sample Mean

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

Median

1, 2, 2, 3, 4, 5

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

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

$\bar{x} = \frac{1+2+2+3+4+5+100}{7} = \frac{117}{7} = 16.71$

Median ✓

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

$\bar{x} = 16.71$ //

Median = 3

1, 2, 2, 3, 4, 5 → odd or even

1, 2, 2, 3, 4, 5

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

2.5 ≈ 2.83 //

Median

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

Error

[2, 3, 4]

Mode ÷ Highest frequency.

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

3

EDA

Feature Engineering

↳ NAN values ⇒

Continuous Values + Outlier

↳ Mean

Median

⇒ Categorical Variable.

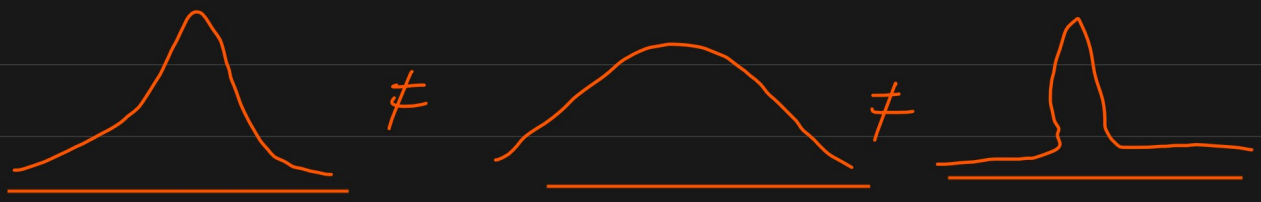
Mode

Agri — Libby, Sunflower, Rose, ..., MAN, MAN

Measure of Dispersion → {Dispersion}

① Variance
② Standard deviation

} ⇓ Spread ⇒ How the data is spread



① Variance { Bessels Correction, Degree of freedom }

Population Variance

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

(μ: Population mean)

Sample Variance

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

($\bar{x}$: Sample mean; $n-1$: Bessel's correction)

Eg:

$X = \{1, 2, 2, 3, 4, 5\}$

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

$n = 6$

$$\frac{10.84}{5} = 2.168$$

($n-1$)

$\mu = 2.83$ {let consider as an example} 10.84

Variance ↑↑

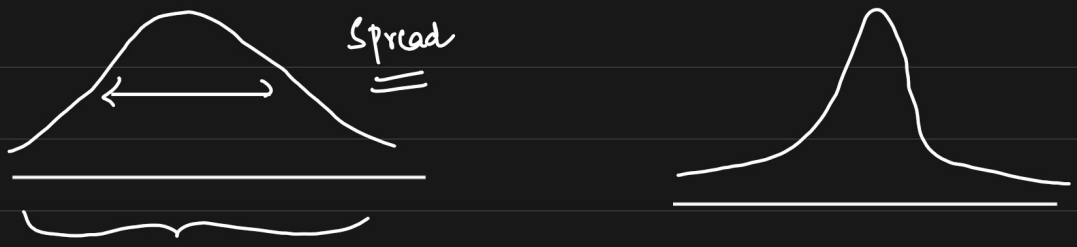
σ^2 ⇓ ✓

Variance = 6.42 Spread ↑↑

σ^2

Variance = 2.168

Spread ↑↑



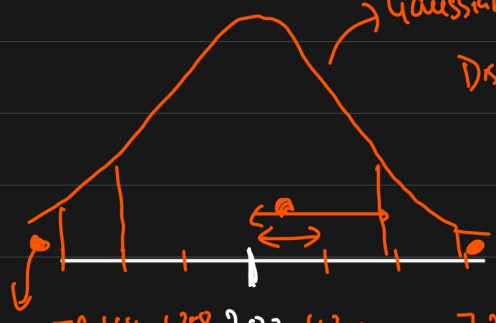
Standard deviation

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

Variance ⇓

Spreadness ↓

1, 2, 2, 3, 4, 5



2.83 1 0
− 1.472
1.358

1.358
1.472
0.114

1 1
2.83
1.472
4.302
1.472
5.774
1.472
7.246

Percentiles And Quartiles

Percentages : 1, 2, 3, 4, 5

% of the numbers that are odd?

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

Percentiles : {CAT, GATE, SAT} ⇒ 99%

Defn : A percentile is a value below which a certain percentage of observation lie

99 percentile mean the person has got better marks than 99% of the Students.

Dataset : 2, 2, 3, 4, 5, 5, 5, 6, 7, 8, 8, 8, 8, 8, 9, 9, 10, 11, 11, 12

Average = 5

What is the percentile ranking of 10 ? n = 20

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

$= \frac{16}{20} \times 100 = 80$ percentile

$= \frac{17}{20} = 85$

② What value exists at percentile ranking of 25% ?

$\text{Value} = \frac{\text{Percentile}}{100} \times (n+1)$ Doubt??

$= \frac{25}{100} \times (21) = 5.25 \rightarrow$ Index

Value = 5

Quartiles (25%)

Five Number Summary

1. Minimum
2. First Quartile (25%) Q1
3. Median
4. Third Quartile (75%) Q3
5. Maximum

Removing the outliers

Inter Quartile Range = (75% - 25%)
Q3 - Q1

{1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, ~~29~~}

[Lower Fence ⟷ Higher Fence]

Lower Fence = Q1 - 1.5(IQR)

Higher Fence = Q3 + 1.5(IQR)

(25%) $Q1 = \frac{25}{100} \times (20) = 5^{th}$ index

Q1 = 3

IQR = Q3 - Q1 = 7 - 3 = 4

(75%) $Q3 = \frac{75}{100} \times 20 = 15^{th}$ index

Q3 = 7

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

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

[-3 ⟷ 13]

[-3 ⟷ 13]

-ve 3

Remaining

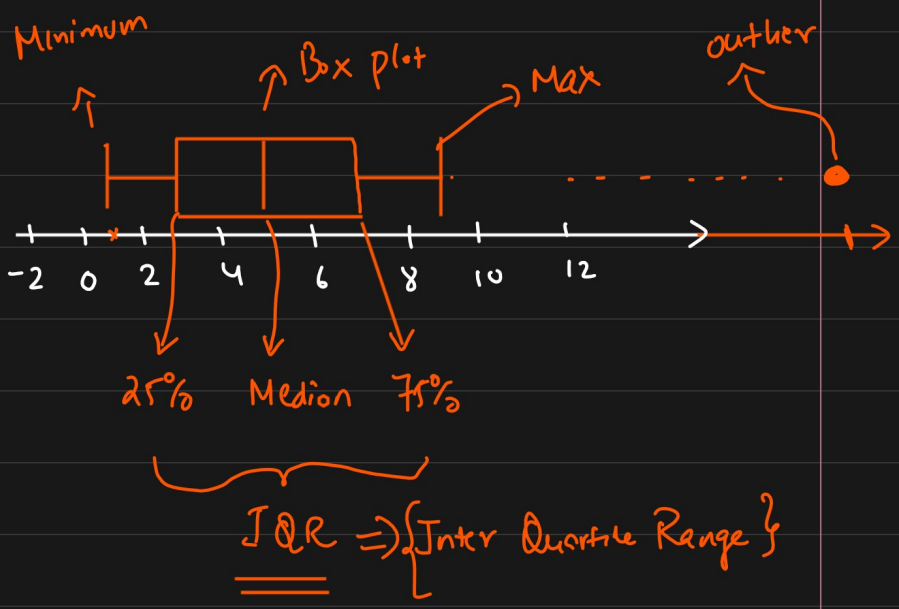
$\frac{5+5}{2} = 5$

1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, ~~20~~

5 Number Summary

Minimum = 1
Q1 = 3
Median = 5
Q3 = 7
Max = 9

Box plot → Outliers

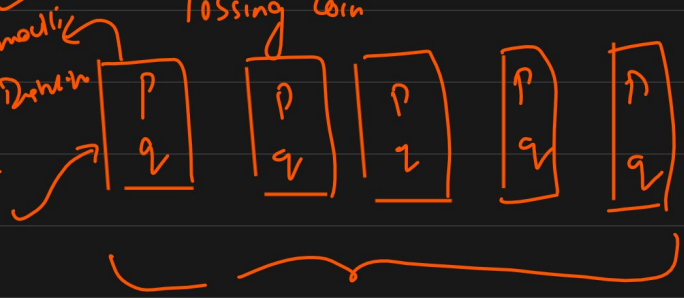


IQR ⇒ {Inter Quartile Range}

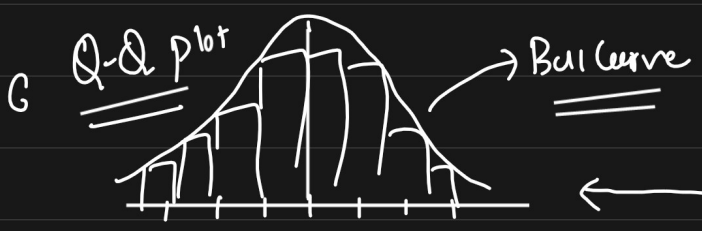
Use of Box plot

Distributions

1. Normal / Gaussian Distribution ✓
2. Standard Normal Distribution ✓
3. Z - Score ✓
4. Log Normal Distr ✓
5. Bernaullis Distribution ✓
6. Binomial Distribution }



1) Gaussian / Normal Distribution



Properties { Power Law }

1) Empirical Rule of Gaussian Distribution ⇓ 80-20%

↳ DATASET → IRIS Dataset } → Petal, Sepal length

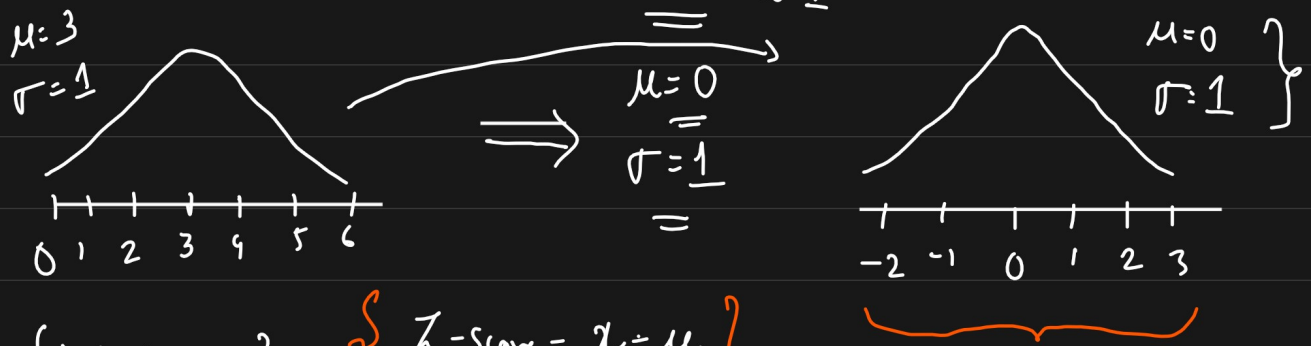
2) Weight of human beign 3) Height → Doctor ⇓ Domain expertise }

68.2, - 95.4 - 99.7 ↙ Outliers

Standard Normal Distribution

{1,2,3,4,5} $\mu = 3$

$\sigma = 1.414 \approx 1$



{1,2,3,4,5} $\left\{ Z\text{-score} = \frac{x - \mu}{\sigma} \right\}$

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

$= -2$

Why ?? $\mu = 0$, $\sigma = 1$

Standardization Vs Normalization

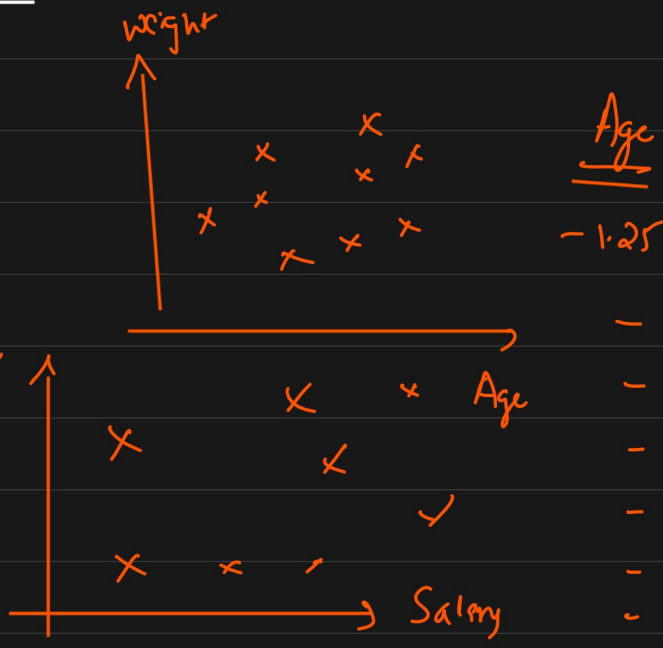
Age (years)	Weight (kg)	Salary (INR)
25	75	25K
26	80	30K
28	85	40K
30	60	80K
32	70	

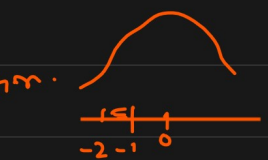
Different unit

Same unit scale ??

$\frac{25 - 28.2}{2 \text{ to } 6}$

Maths → Scale





ML Scaler

Normalization [Min Max Scaler] { Standardization }

&

0 to 4
0 to 1 } { Min Max Scaler }

Normalization

⇓

CNN ← Convolutional Neural N/w.

(0-1)

Normalization

$$\left\{ x_{Norm} = \frac{x_i - x_{min}}{x_{max} - x_{min}} \right\}$$

⇓

Min Max Scaler = 0 to 1

f1
2
5
6
8 ✓
1 ✓

f1'
0.14
0.571
0.71
1
0

$\frac{8-1}{8-1} = 1$

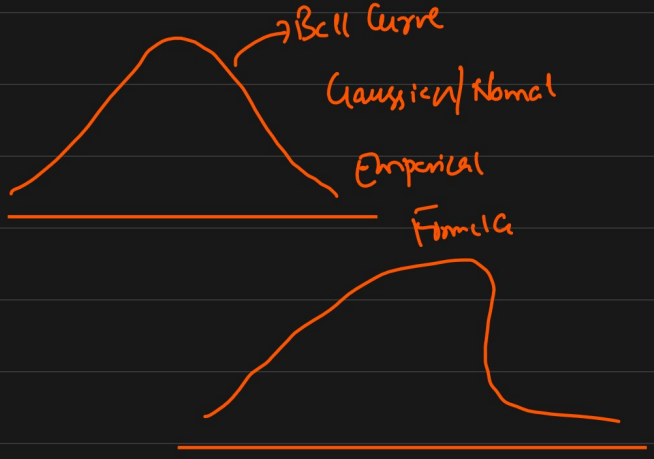
$\frac{1-1}{8-1} = 0$

$\frac{6-1}{8-1} = \frac{5}{7}$

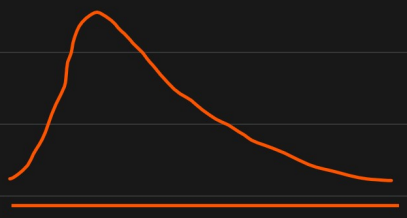
$= \frac{2-1}{8-1} = \frac{1}{7} = 0.142$

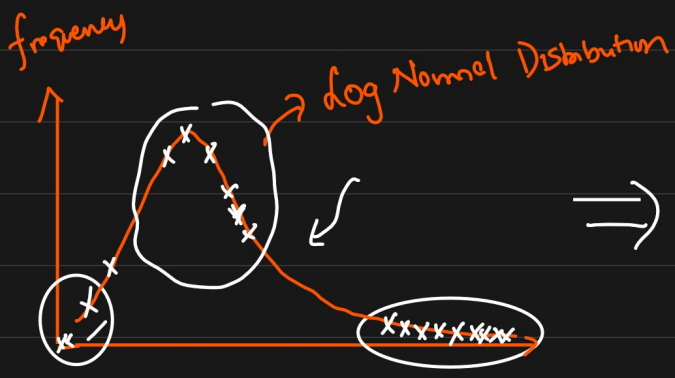
$\frac{5-1}{8-1} = \frac{4}{7} = 0.571$

Log Normal Distribution



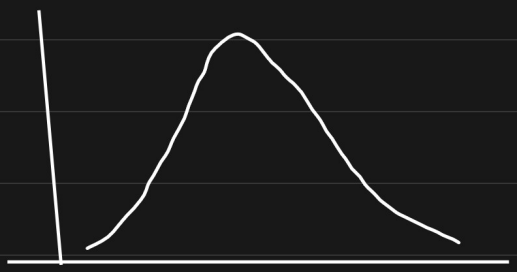
Skewed Curve

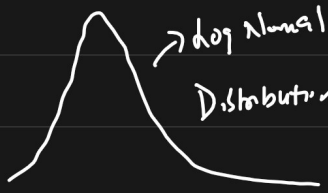




Gaussian Distribution

Normal Distribution



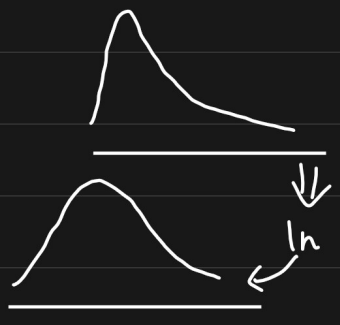


X = Log Normal Distributed

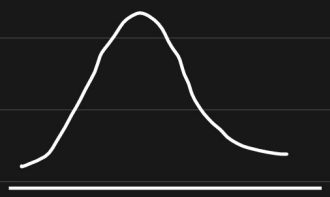
$\{Y = \ln(x)\}$ Gaussian Distribution

$\Downarrow \Downarrow \Downarrow$

$\{X = \exp(y)\} \rightarrow e^{y}$



X	$Y = \ln(x)$
25	=
30	=
40	=
35	=



(*) Bernoulli's Distribution.

Day 2 - Stats

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

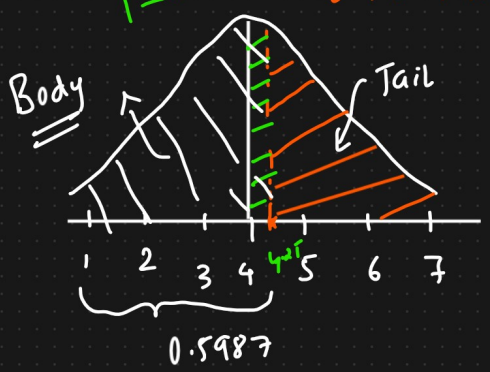
Stats Interview Question

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

$\mu = 4$

$\sigma = 1$

How many Standard deviation 4.25 fall from the mean??



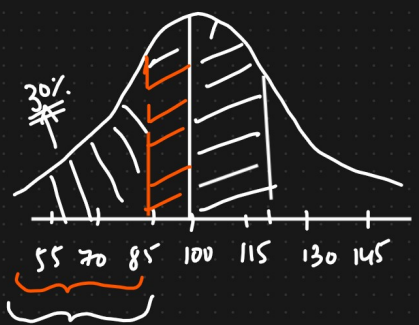
$$Z\text{-score} = \frac{x_i - \mu}{\sigma} = \frac{4.25 - 4}{1} = 0.25$$

Question : What percentage of scores fall above 4.25?

$1 - 0.59871 = 0.4013 \Rightarrow 40.13\%$

② 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?

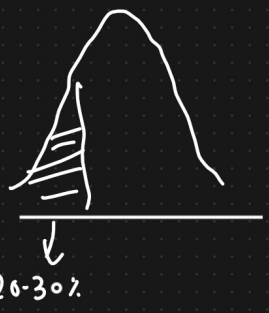
Ans)



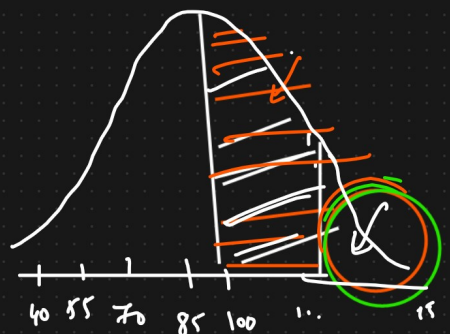
$$Z\text{-score} = \frac{85 - 100}{15} = \frac{-15}{15} = -1$$

① Area under this curve

$0.5 - 0.15866 = 0.34143 \Rightarrow 34.134\%$



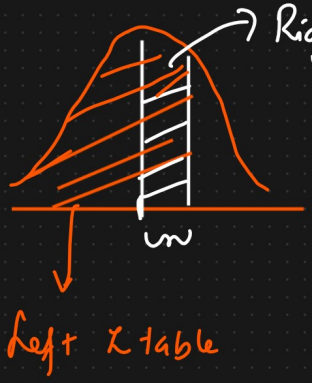
{ Greater than or equal 100 less than or equal 125 }



$$Z\text{score} = \frac{125 - 100}{15} = \frac{25}{15} = 1.66$$

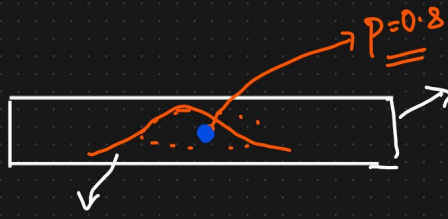
Ans = 0.4515 => 45.15%.

0.5 − 0.4515 = 0.0485 => 4.8%



P value, Hypothesis Testing, Confidence Interval

Out of all 100 touches, the no. of touches is 80



Out of all 100 touches, the no of times 40 times.

Hypothesis Testing, C.I, Significance value Together

Fair Coin

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

P(H) = 0.5 ↓ P(T) = 0.5 ↓

Hypothesis Testing

Criminal is → Court SIMOLAY P(H) = 100% P(T) = 0%

1. Null Hypothesis — Coin is fair → (H_0)
2. Alternate Hypothesis — Coin is not fair → (H_1)
3. Experiments
4. Reject or Accept the Null Hypothesis

100 tosses →
- 50 times Head
- 50 times Tails

} Coin is fair

60 times Head
40 times Tail

} Coin is fair

30 times Head
70 times Tail

Domain Expertise

Confidence Interval, Significance Values

Health Care

↓

Covid vaccine Test

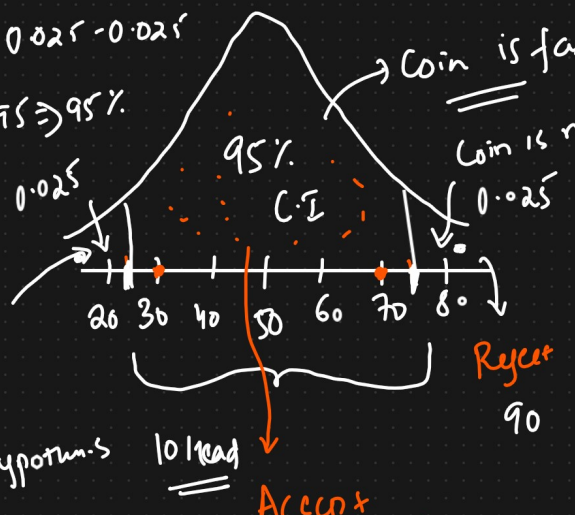
Error ↓↓

↓

Significance value = 0.05

C.I = 1 − 0.025 − 0.025

= 0.95 ⇒ 95%



Null Hypothesis

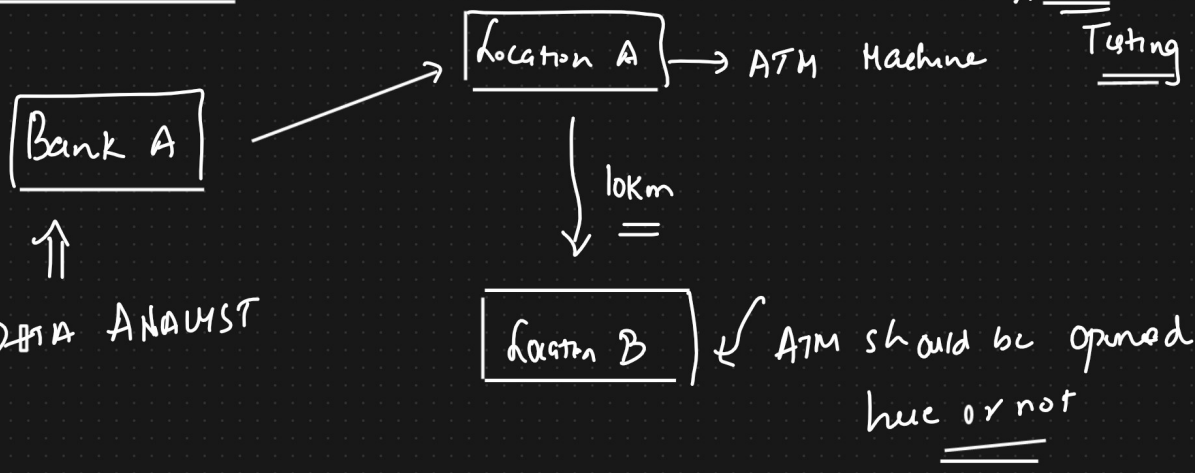
$\alpha = 0.45$

Medical $\alpha \uparrow\uparrow$.

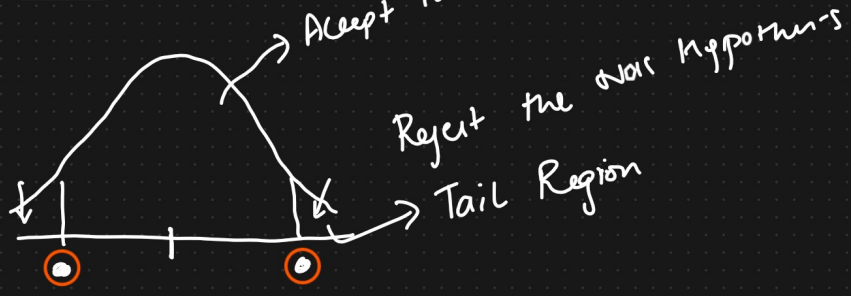
$\frac{0.45}{2} = 0.225$

Real World Project

Hypothesis Testing

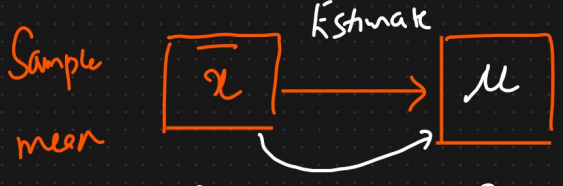


① Confidence Interval



Point Estimate

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



Influential

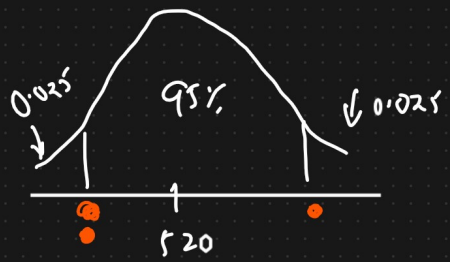
{ Sample Mean → Population }

Confidence Interval

t test Point Estimate ± Margin of Error ⇒ 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 80 (⇓) 95% C.I about the mean?

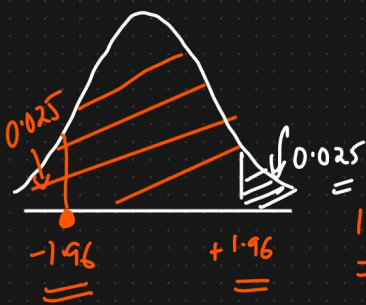
Ans) $\sigma = 100$ $n = 25$ $\bar{x} = 520$ C.I = 95% $\alpha = 0.05$ (1 − 0.95)



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

Point Estimate ± Margin of Error → C.I

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



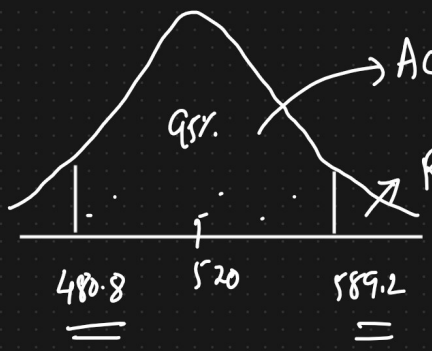
1 − 0.025 = 0.9750

Lower fence C.I = $\bar{x} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ → $Z_{\frac{0.05}{2}}$ ⇒ $Z_{0.025} = 1.96$

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

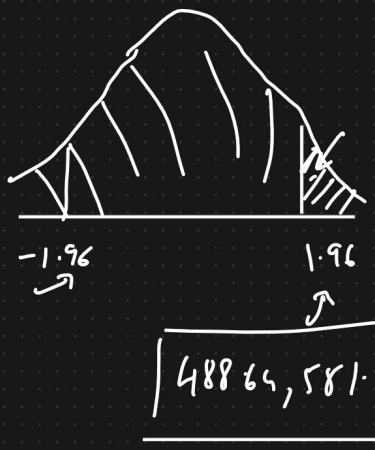
Lower fence = $520 - (1.96) * \frac{100}{\sqrt{25}} = 520 - (1.96) \times 20 = 480.8$

Higher fence = $520 + (1.96) \times 20 = 559.2$.



$$\frac{\alpha}{2} = 0.025$$

$$1 - 0.025 = 0.9750$$



488 to 552

(4) 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) $\bar{x} = 520 \quad S = 80 \quad \alpha = 0.05 \quad n = 25$

t-test ⇒ t-table

{Because population sd is not given}

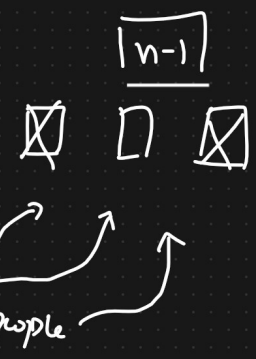
$$\bar{x} \pm t_{\alpha/2}\left(\frac{S}{\sqrt{n}}\right) \longrightarrow \text{Standard Error}$$

$t_{0.025}$

t-test

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

3 | n-1



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

① Type 1 and Type 2 Error.
② One Tailed vs 2 Tailed Test

Type 1 and Type 2 Error

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

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 ✗

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

[Cancer → Not cancer]
↓ Fire

Reality Check

① Null Hypothesis is True or Null Hypothesis is False

Decision of Experiments?

Null Hypothesis is True or False.

Null Hypothesis
$H_0 \rightarrow$ The Criminial is not guilty
$H_1 \rightarrow$ " " is guilty

{→ Stock market is going to crash?

Confusion Matrix

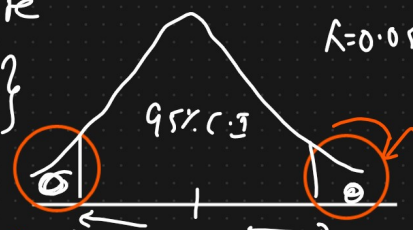
② 1 Tail and 2 Tail Test

Eg: Colleges is 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 a standard deviation of 4%. Does this College has a diffucnt placement rate?

$\alpha = 0.05 = 95\%$ C.I

↳ 85%

{ placement rate less than 85% } ⇓ 1 tail



{ placement rate greater than 85% } 2 Tail Test

1 tail Test

Saturday

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

10 min probability

Sunday

EDA → 3-4 projects
FE →

→ Machine Learning

① Central Limit Theorem

② Infcuential Statistics

*) Z test {Z table} [5-6 problems]

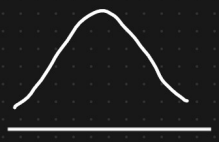
*) t test {t table}

*) Z test propotion population.

*) Chi square (categorical Tcst)

*) ANNOVA (F Test)

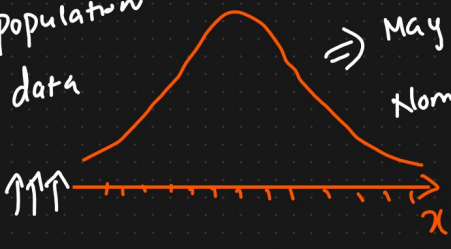
Infcuntial



① Central limit Theorem

n>30 ✓ Sample mean distribution

{ Population data ⇒ May be Gaussian Normal Distr



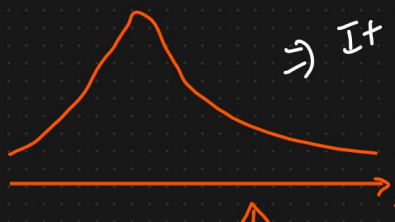
$\rightarrow [x_1, x_2, x_3, x_4, x_5 \cdots x_{30}] \rightarrow \bar{x}_1$ Sample 1

Sample 2 $[x_1, x_2, x_3\ x_4, x_5 \cdots x_{30}] \rightarrow \bar{x}_2$

$\rightarrow \bar{x}_3$

$\rightarrow \bar{x}_m$

⇒ It may not



Sample m

10,

n>30

Sample mean distribution

Gaussian Distribution

Normal Distribution

② Influential Statistics {Data Analyst, Data Scientist}

① 100K ⇒ T-shirt ⇒ No ⇒ Sample data ⇒ XL, L, Small

② iNeuron → Meetup → Hitesh ⇒ 300-400 people → T-shirts

{ 20% L, 10% XL, 10% XXL, 60% Medium } 500 t-shirts

Ordered: L, M, XL, XXL

Next Event : Z test $\bar{x}$ $(\mu, 5)$ → C.I [ ]

③ ATM ④ Measure the size of entire sharks

⑤ Amazon delivery { Percentile, Quartiles } ⇒

⑥ Hypothesis Testing

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

(a) State Null and Alternate Hypothesis

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

Ans) Step 1

$n = 40$ $\bar{x} = 78$ $S = 2.5$

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

$H_1 : \mu \neq 80$ {Alternate hypothesis}

Why Z test?

Step 2:

$\alpha = 0.05$ C.I = 95%

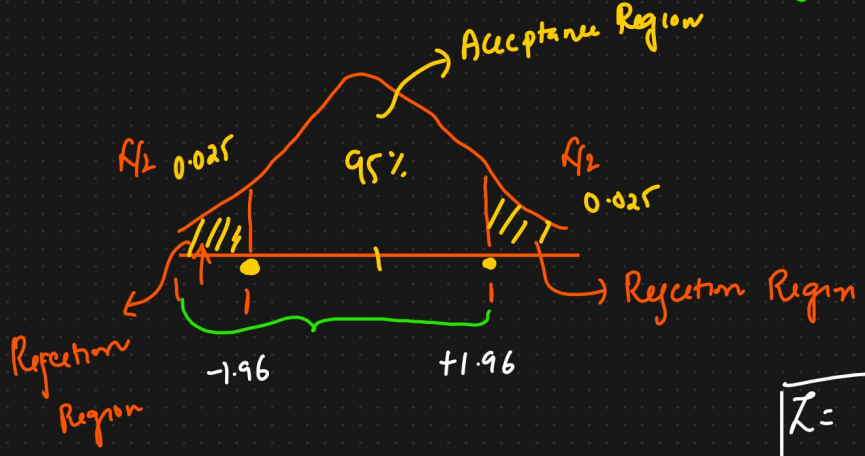
$n < 30$ { ① $n > 30$ $n \leq 30$ ② population std or sample std }

Step 3: Decision Boundary

$1 - 0.025 = 0.9750$

Why t test

{ ① Sample std = ② $n < 30$ }



$n = 1$ }

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

④ Calculate Test Statistics

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

$s/\sqrt{n}$ ⇒ Standard Error

Sample standard deviation

$$= \frac{78 - 80}{2.5 / \sqrt{40}} = \frac{-2 \times \sqrt{40}}{2.5} = \frac{-2}{2.5} \times 6.32 = -5.05$$

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 a 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? { 95% ⇓ C.I }

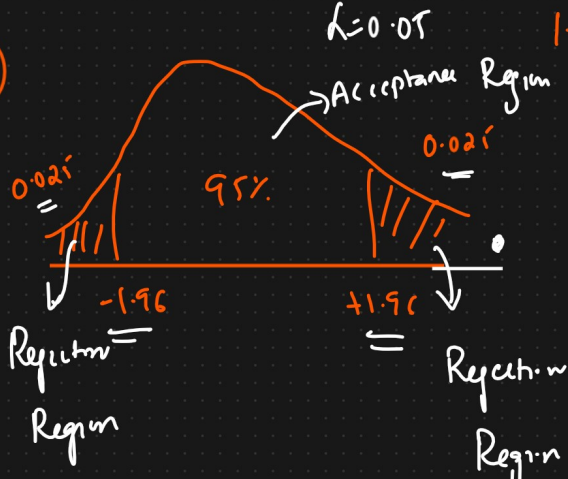
Ans) $\sigma = 15 \quad n = 30 \quad \bar{x} = 140$

(1) $H_0 : \mu = 100$

$H_1 : \mu \neq 100$

(2) $\alpha = 0.05 \quad C.I = 95\%$

(3)



If Z is less than -1.96 or greater than 1.96, reject the NULL Hypothesis

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

$14.60 > 1.96$ Reject the NULL Hypothesis

(*) A Complain was registered, the boys in the Municipal Primary School are underfed. Average weight of boys of age 10 is 32kgs with S.D = 9kgs. A sample of 25 boys was selected from the municipal school and the average weight was found to be 29.5 kgs? With C.I = 95% Check whether it is True or False?

Ans) $\mu = 32\text{kgs}$ $\sigma = 9\text{kg}$ $n = 25$ $\bar{x} = 29.5$ $\alpha = 0.05$

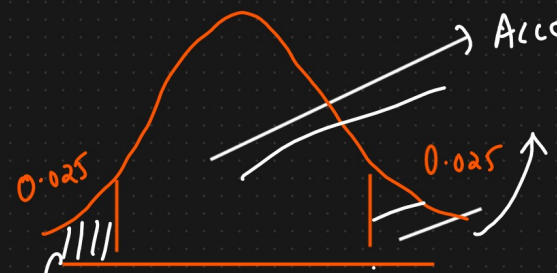
1) $H_0 = \mu = 32$ }

$H_1 = \mu < 32$

② $\alpha = 0.05$

$1 - 0.95 = 0.05$

③



④ $Z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$

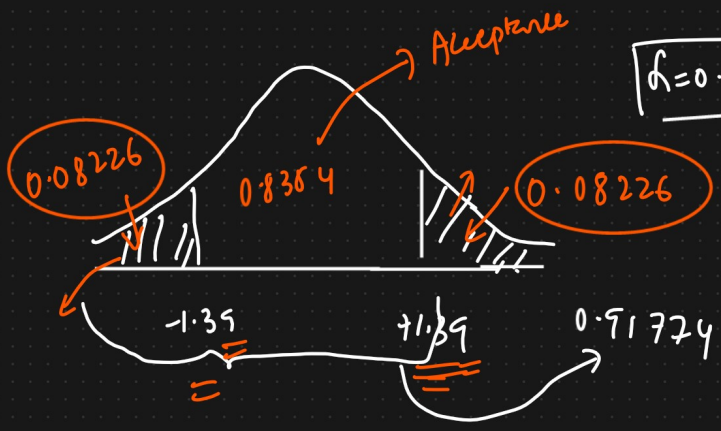
$= \dfrac{29.5 - 32}{9/\sqrt{25}} = -1.39$

Conclusion : $-1.39 > -1.96$ therefore we accept the Null Hypothesis

So, the boys are not underfed.

Significance value. ⇒

P value



$\alpha = 0.05$ ⇒ p value

$1 - 0.08226 - 0.08226 =$

$1 - 0.91774 = 0.08226$

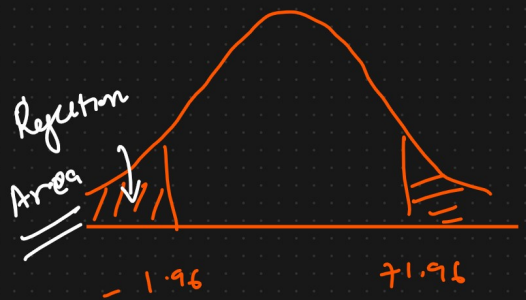
p value

$0.08226 + 0.08226 = 0.1645$

Significance value

$0.1645 > 0.05$

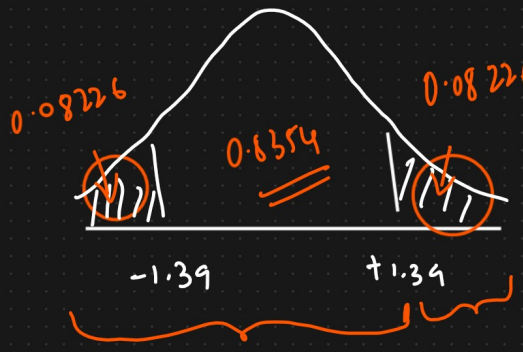
$p \geq 0.05 \Rightarrow$ Accept the Null Hypothesis



Z test

⇓

p value



New p value $= 0.08226 + 0.08226$

$= 0.16$

$1 - 0.08226 - 0.08226$

Domain

⇓

$0.1645 >$ Significance value

⇓

Accept the Null Hypothesis.

$1 - 0.91774 = 0.08226$

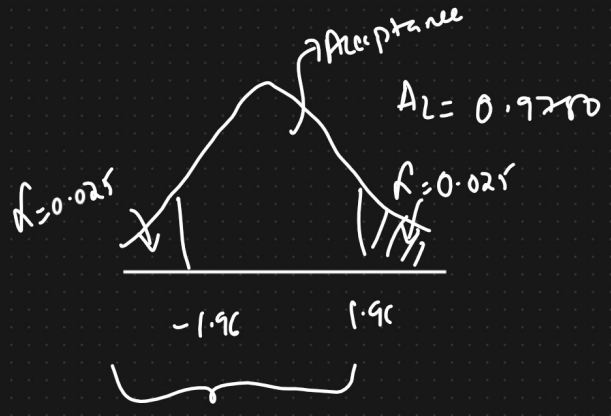
(#) The average weight of all residents in town XYZ is 168 lbs. A nutritionist believes the true mean to be different. She measured the weight of 36 individuals and found the mean to be 169.5 lbs with a standard deviation of 3.9.

(#) Assignment

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

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

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



Z-test , t-test

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

$2.31 > 1.96$ Reject the Null Hypothesis

(#) A company manufactures bike batteries 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

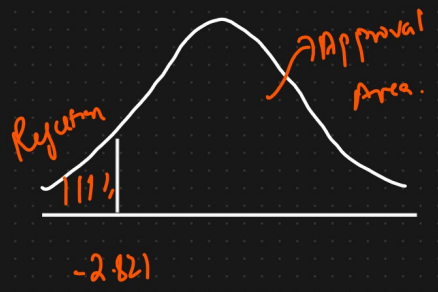
b) At a 99% C.I, is there enough evidence to discard the H_0?

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

$H_1 : \mu < 2$ $\quad n < 30$ $\quad$ t-test ??

② $\alpha = 0.01$ $\quad \alpha = 1 - C.I = 1 - 0.99 = 0.01$

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

$-4.216 < -2.821$ $\quad$ Reject the Null Hypothesis.

Z test with proportions

⑥ A tech 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

(a) State the Null and Alternate Hypothesis?

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

$n = 200$ $\quad x = 130$

(a) $H_0 \quad p_0 = 0.70.$

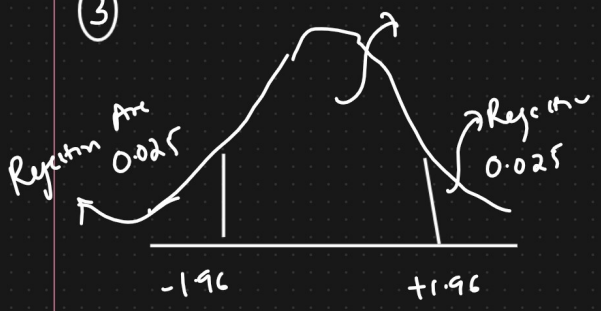
$H_1 \quad p_0 \neq 0.70$

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

$q_0 = 1 - p_0$

② $\alpha = 0.05 \quad C.I = 95\%$

③

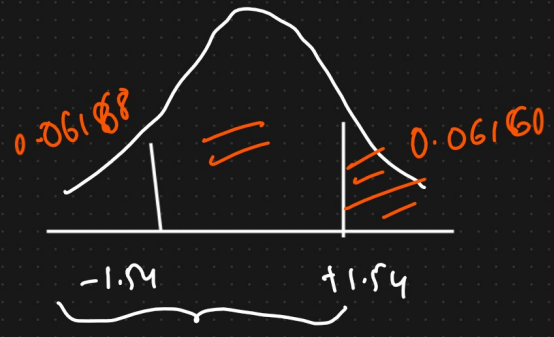


$$Z_{test} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$$

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

At 95% C.I there is

$-1.54 > -1.96$, So we accept the Null Hypothesis



pvalue

$2 \times 0.06168 > 0.05$

Accept Null Hypothesis.

$1 - 0.93832 = 0.06168$

(4) A Car company believes that the percentage of residents in City ABC that owns a vehicle is 60% or less. A Sales manager disagrees with this. He conducts a hypothesis testing surveying 250 residents and found that 170 responded yes. to owning a vehicle.

(a) State the Null & Alternate Hypothesis

(b) At 10% significance level, is there enough evidence to support the idea that vehicle ownership in city ABC is 60% or less?

pvalue = .0.014

Statistics

{ 11:30 - 12pm }

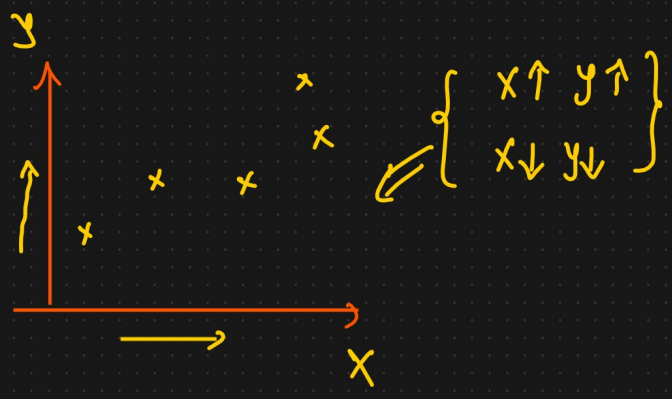
1. Covariance
2. Pearson Correlation Coefficient
3. Spearman Rank Correlation Coefficient.
4. CHI SQUARE TEST
5. ANNOVA (F-Test)

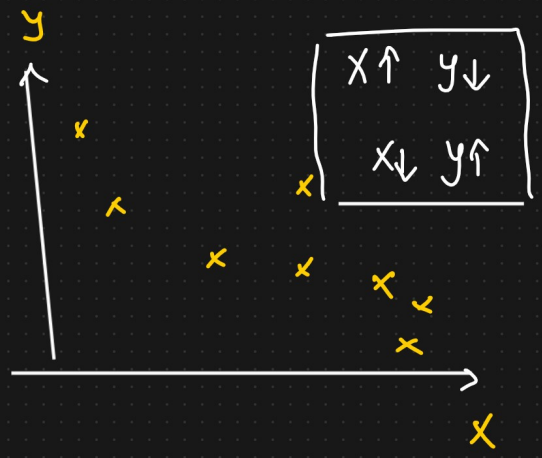
} practicals

Covariance

{ quantify the relationship between X & Y }

X	Y
X↑	Y↑
X↑	Y↓
X↓	Y↑
X↓	Y↓





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

$\Uparrow$ Cov(X, Y) $\Downarrow$

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

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

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

$$\Downarrow$$

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

+ve

X ↑	Y ↑
X ↓	Y ↓

$\Rightarrow$ Positively Correlation

X	Y
2	3
4	5
6	7

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

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

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

$$= \frac{(-2)(-2) + 0 + (2)\times(2)}{2} = \frac{8}{2} = 4$$

X ↑	Y ↓
X ↓	Y ↑

$\Rightarrow$ -ve Correlation $\Rightarrow$ -ve value.

Disadvantage Covariance

$Cov(x,y)$ $\Rightarrow$ +ve value or -ve value $\Rightarrow$

$Cov(x,y) = 500$

$Cov(y,z) = 600$

Limit

+500 $\quad$ -300

-400 $\quad$ +1000

$\quad$ -400

∞

$\Downarrow$

Relationship $[-1$ to $1]$

$\Uparrow$

② Pearson Correlation Coefficient

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

[-1 to 1]

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

Dataset : 1000 features

Independent features

X | ~~X~~ | Z | A | B | C → O/P (Dependent feature)

+ve correlated

X, Y ⇒ 99%

90% 0.9

-ve Correlation ⇓ Keep it

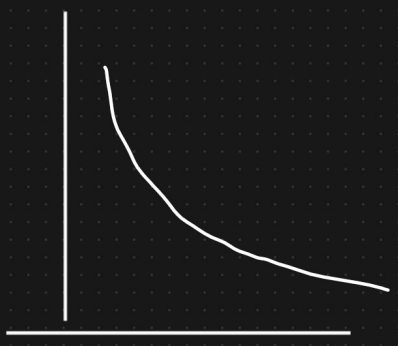
⑧ Spearman Rank Correlation

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

Marks

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

Spearman Rank Corr = 1



-1

⑦ 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.

Ex) In the 2000 U.S Census, the ages of individuals in a small town were 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

<18	18-35	>35
121	288	91

Using $\alpha = 0.05$, would you conclude the the population distribution of ages has changed in the last 10 years?

Ans)

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

95% C.I

n=500

Observed	121	288	91
Expected	100	150	250

① H_0 = the data meets the expected distribution

H_1 = the data donot meet the expected distr

② State Alpha : $\alpha = 0.05$

③ Calculate the degree of freedom

$df = n - 1 = 3 - 1 = 2$ ⇒ 3 Categories.

④ Decision

Chi Square Table.

If χ^2 is greater ⇓ 5.99 than, Reject H_0

⑤ Calculate Chi square Test

$$\chi^2 = \sum \frac{(fo - fe)^2}{fe} = \frac{(121-100)^2}{100} + \frac{(288-150)^2}{150} + \frac{(91-250)^2}{250}$$

$$\chi^2 = 232.494$$

$232.494 > 5.991$ Reject the Null Hypothesis.

⑥ 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 a random sample of 100 teachers asking them which day of the week they had the highest number of

Student absenses. The Observed and expected results are shown in the table below. Based on these results, do the days for the highest number of absences occurs with equal frequencies (use. 95% C.I)

	Monday	Tuesday	Wednes	Thursday	FRIDAY
Observed	23	16	14	19	28
Expected	20	20	20	20	20.

Ans = 6.3 {Accept the Null Hypothesis}

{Practicals + EDA + Feature Engineering}

Statistics

① ANOVA (F-Test) → 1 hour }
② KDA → { Solve Some Examples. }

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 { Gender }

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

factor : Dosage

Levels : 0mg, 50mg, 100mg

Types of ANOVA

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 Levels	Running kms	KARTIK Day 1	Day 2	Day 3	Day 4
1		6	10	2	10 kms
2		–	–	–	–

Eg. Study hours of KARTIK

1 Factor: Day 1 ⟷ Day 2 ⟷ Day 3

10 hrs | 5 hrs | 2 hrs

Gym

Day 1 ⟷ Day 2 ⟷ Day 3

2 hrs (5 Eggs) | 1.5 hr (2 Eggs) | 1 hr

④ Factorial ANOVA

⑤ Factorial ANOVA : Two or more factor (each of which with atleast 2 levels), levels can be either independent, dependent or both (mixed)

Eg

factor ↓	Day 1	Day 2	Day 3
Men	9	7	4
	8	6	3
	7	5	2
Women	8	7	3
	8	8	4
	9	7	3

(factor → Day 1, Day 2, Day 3)

One Way ANOVA (F-test) ⇒ Inferential Stats

⇓

Comparing means of 2 or more groups

1) 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 differences between the 3 conditions using $\alpha = 0.05$?

0mg	50mg	100mg
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 the Degree of freedom

Statistics

$N = 21$ ⇓ $n = 7$

$a = 3$ → {No. of levels}

→ $df_{Between} = a - 1 = 3 - 1 = 2$

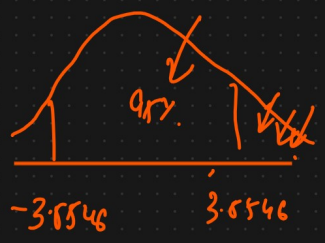
→ $df_{within} = N - a = 21 - 3 = 18$

→ $df_{Total} = N - 1 = 21 - 1 = 20$

④ State Decision Rule

$df_{Between} = a - 1 = 3 - 1 = 2$ $\{(2, 18)\}$

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



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	F Test
Between	98.67 →	2	49.34	86.56
Within	10.29	18	0.57	
Total	108.96	20		

$N = 21 \quad n = 7$

$SS_{between} = \left[\frac{\sum(\sum a_i)^2}{n}\right] - \frac{T^2}{N}$

$T^2 = [57 + 47 + 21]^2$

$= (125)^2$

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

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

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

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

$p\ 0.78$

$\alpha = 0.05$

$= \sum y^2 - \left[\frac{57^2 + 47^2 + 21^2}{7}\right] = 10.29$

$\sum y^2 \downarrow 853$

$\sum y^2 = 9^2 + 8^2 + 7^2 + 8^2 + 8^2 + 9^2 + \ldots + 2^2 = 853$

$0.78 > 0.05$

⇓

Accept

Final Conclusion

$86.56 > 3.5546$ so we reject the Null hypothesis

$H_0 = \mu = \text{Some value}$

$H_1 : \mu \neq \text{Some value}$ → 95% C.I

Virginica

Petal width

→ $H_0 = \mu_{virgin} = \mu_{stressed} = \mu_{\dots}$

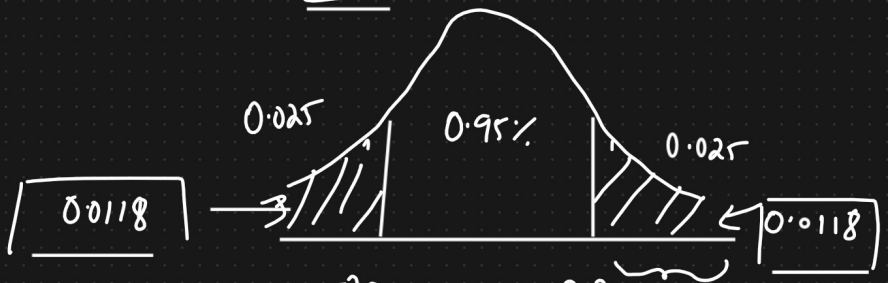
$H_1 =$ $\neq$ p value $\neq$ ⇒ Reject the Null Hypothesis

0.0118
0.0118
0.0228

0.0228 < 0.05 $\quad 1 - 0.025 = 0.975$

Z test $\quad \alpha =$

Z test statistics



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

$Z = 2.304$

$2.304 > 1.96$ Reject