Statistics

Usecase

HDFC

doc → A → ATM

30Kms

C → ATM

doc → B → ATM

PRODUCT BASED COMPANY

X Statistician ⟶ 5 years

- DATA ANALYST
- Business Analyst
- DATA SCIENTIST
- PRODUCT MANAGERS

① DATA ANALYST

② DATA SCIENTIST

② Find the average size of the shark throughout the world?

blue whale

Amazon

③ Amazon Big Billion Day Sale {Intuit} ⟶ Which month should you select?

Statistics {life cycle of DATA Science Project}

HCL

Google → Domain Expert

SERVICE {TCS}

④

DATA SCIENCE

Requirement Gathering

Business Analyst

PRODUCT MANAGER

PROJECT MANAGER

DATA ANALYTICS TEAM

DATA ANALYST / DATA SCIENTIST

← PRODUCT MANAGER

← BA

Apple

PRODUCT BASED

Domain Knowledge

{Google}

Youtube, GPAY, Google Ads, GMAIL

SALES

Domain Expertise

Product Manager

① DATA ANALST

② DATA SCIENTIST

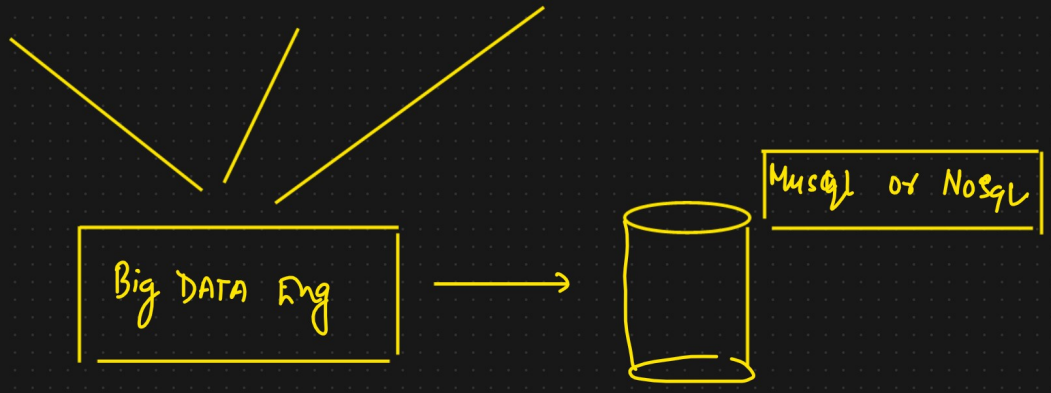
③ BIG DATA Engineers

④ Cloud Engineers

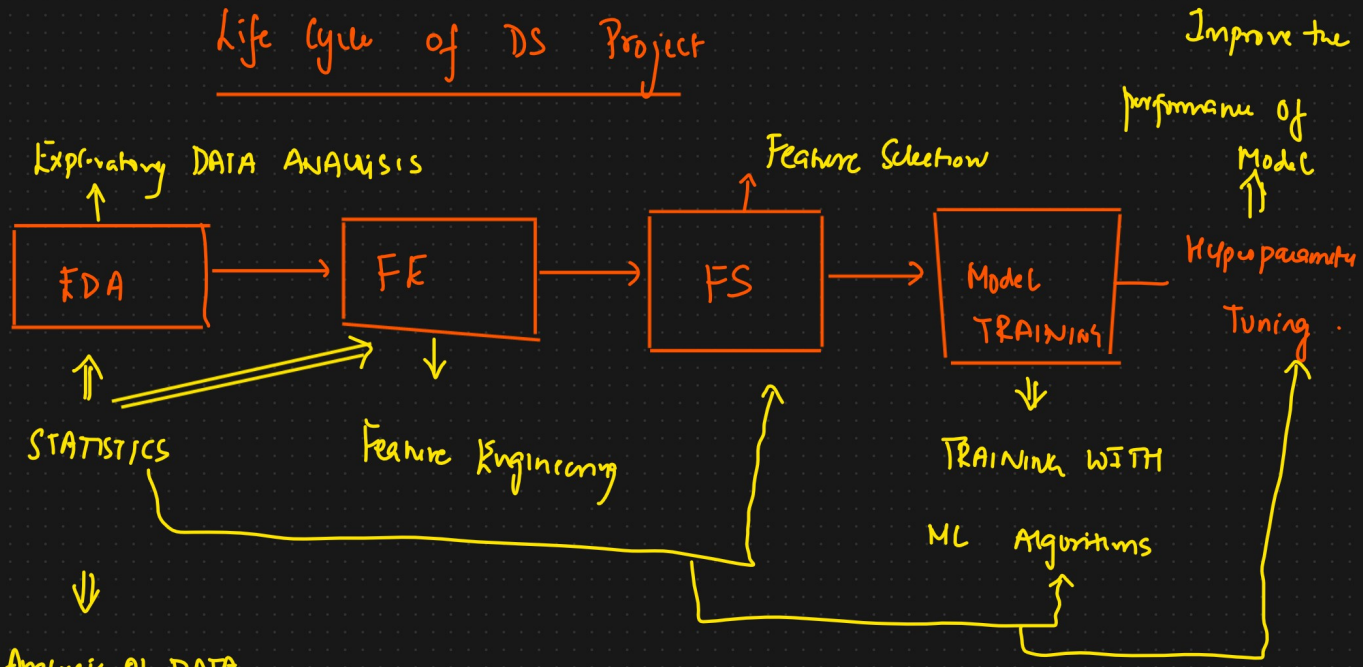
Internal DATABASE

3rd party API's

Web Scraping

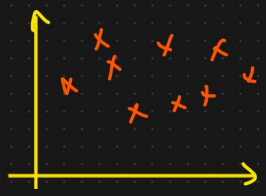


Life Cycle of DS Project



Analysis of DATA

DESCRIPTIVE STATS



⇒ Summarizing the DATA

⇒ DESCRIPTIVE STATS

Age = { 12, 13, 14, 18, 20, 25 } ⇒ Average Age ⇒ Measure of Central Tendency

⇓

DESCRIPTIVE STATS

Statistics : **Defn** : Statistics is the science of collecting, organising and analysing the data.

Data : "facts or pieces of information"

Eg: Ages of students in classroom

$\{24, 25, 32, 29, 28\}$ ⇒ Mean, Median, Mode

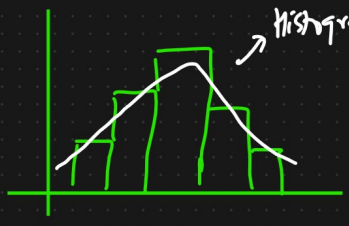
Standard deviation

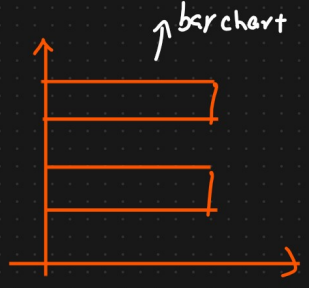
② Weights of students in classroom

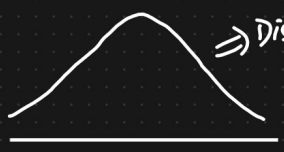
STATISTICS

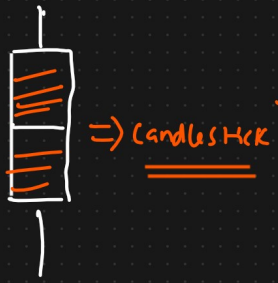
Descriptive Stats [EDA + FE]

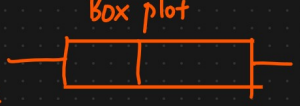
① It consists of organising and summarising the data.











Inferential Stats

② It consists of collecting sample data and making conclusion about population data using some experiments

Hypothesis Testing

University → 500 people

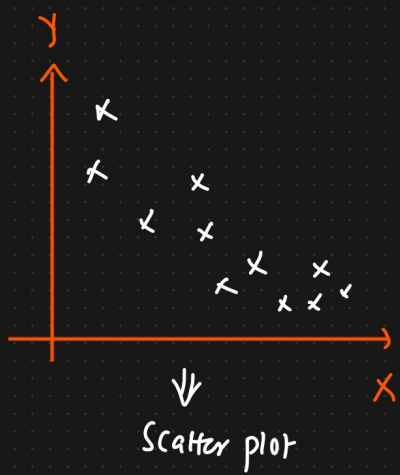
Class A → 60 people

Sample data ⇒ Age ⇒ Average age of the entire university

Yes

Hypothesis Testing

C.I ⇒ Confidence Interval



(height) x — (weight) y

x↑	y↑
x↓	y↓
x↑	y↓
x↓	y↑

p - value

1. Z test
2. t test
3. Chi square test
4. F test

Sample Data → Conclusions → Population Data

Conclusions ↓ Hypothesis Testing

Sample DATA Vs Population DATA

EXIT POLL ⇑

PARTY A will win

PARTY B will loose ✓

Eg: Let's say there are 20 classrooms in a university and you have collected the age of students in one classroom

Ages {21, 20, 18, 34, 17, 22, 24, 25, 26, 23, 22}

Weight { - - - - - - - - - - - }

Descriptive Stats : What is the average age of students in the classroom?

Relationship between Age & Gender?

Inferential Stats : Are the average age of the students in the classroom less than the average age of the students in the university?

⇓

Greater
⇓
Equal

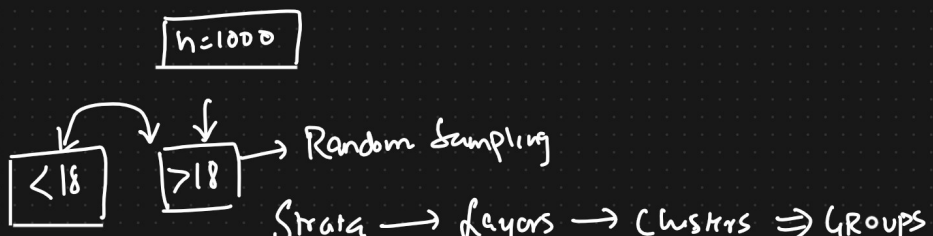
Hypothesis Testing
⇑
1000 Students

Class A ⇒ 50 girls → ⇓ 95%
50 boys → ⇓ 92%

Choose a sample
⇑
Sampling Techniques

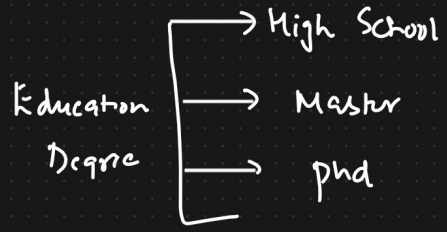
Population (N) Sample (n)

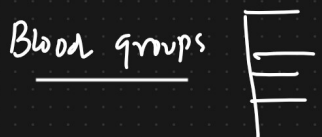
① Simple Random Sampling : Every member of the population (N) has an equal chance of being selected for your sample (n)

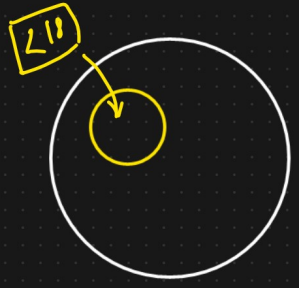


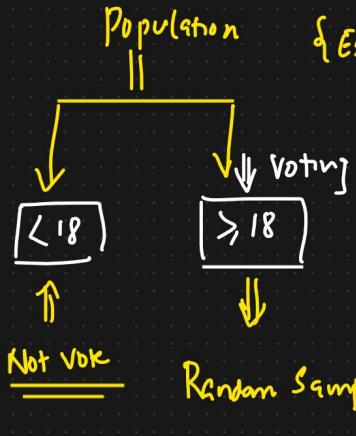
Strata → Layers → Clusters ⇒ GROUPS

② Stratified Sampling







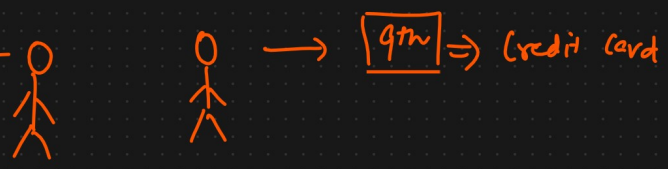


③ Systematic Sampling → {AIRPORT}

n^{th} person

Select every n^{th} individual out of Population (N)

{CREDIT CARD}



Select every n^{th} individual out of Population (N)

④ Convenience Sampling ÷ Only those who are interested in the survey will only participate

{DATA SCIENCE SURVEY → General AI Survey}

Interview job for a specific ⇓ {Fill the Form}

① Survey Regarding New Technology ⇒ Convenience Sampling

② RBI Survey ⇒ Women ⇒ Stratified + Random Sampling · Married Women

③ Credit Card ÷ Stratified + Random Sampling

① Variable : A variable is a property that can take any values

Eg: age = 14

age = 25

age = 100

Variables

Ages = [24, 25, 26, 27, 28, 29] ⇒ Collection

Two diffcunt type of variable

① Quantitative Variable ⟶ Measured Numerically {Mathematical Operation}.

Ex: Age, weight, height, rainfall (cm), temp, distance

② Qualitative Variables ⟶ Categorical Variables { Based on some characteristics they are grouped together }.

Eg: Gender, Types of flowers, Types of Movies

Quantitative variable

↓ Discrete Variable

↓ Continuous Variable.

Eg: Whole number → fixed

Eg: No of Bank Accounts

{1, 2, 3, 4, 5}

2.5 X

Eg: No of Children : Whole numbers

Pincode = fixed

Eg: Continuous → Decimal values

Eg: Height, weight, ages, Rainfall

Speed

Categorical variables.

Marital → Married

→ Not Married

Assessment

Gender ? Categorical

① What kind of variable is Marital Status ? Categorical variable

Ganga River length ? Continuous

Movie duration ? Continuous

Pincode ? Discrete

IQ ? Discrete

105.75, 90.5,

Pancard

AMLPA - -

↓

Categorical variables

Pincode

360059

360058

360057

Fixed

⇒ It is many ?

Categorical

↓

FE

does no. of Categories

Continuous

Discrete ←

Continuous

Whole number

Bank Account = { 2, 3, 4, 5 }

Pincode = { }

C

6

Categorical

	Cities	Gender	Pincode
5	—	M	—
	—	F	—
	—		—
	—		—
			—
			—
			—
			—

PAN ←

⇒ Categorical

Day 2 - Statistics

Agenda

1. Histograms ✓
2. Measure of Central Tendency ✓ } ⇐ 1.15 hrs
3. Measure of Dispersion ✓
4. Percentiles And Quartiles
5. 5 Number Summary (Box plot) }

① Histogram

Ages = { 0, 10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 43, 50, 51, 65, 68, 78, 90, 95, 100 }

1. Sort the Numbers
2. Bins → No. of groups
3. Bins size → Size of Bins

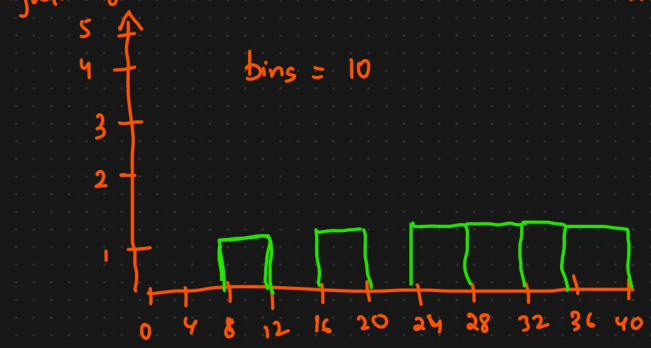
[10, 20, 25, 30, 35, 40]

min = 10

max = 40

bin size

$\frac{40}{10} = 4$ //



bins = 20

bins = 10

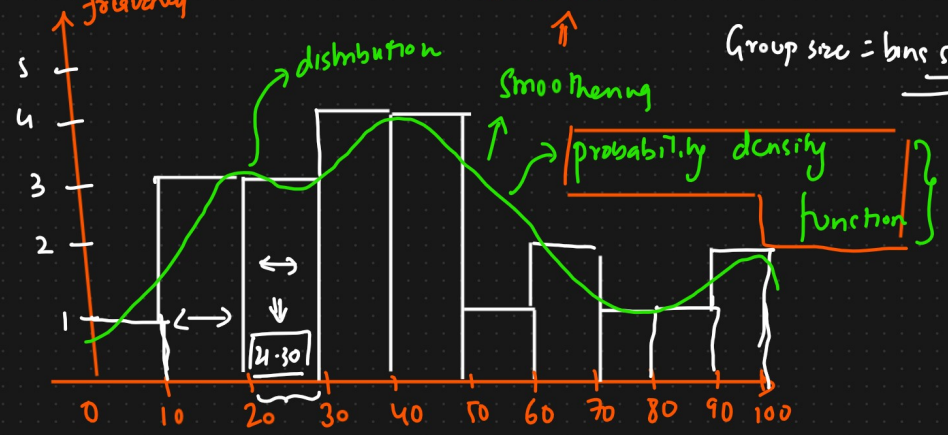
bin size = $\frac{100}{20} = 5$

bin size = $\frac{100}{10} = 10$

Group = bins

Group size = bins size

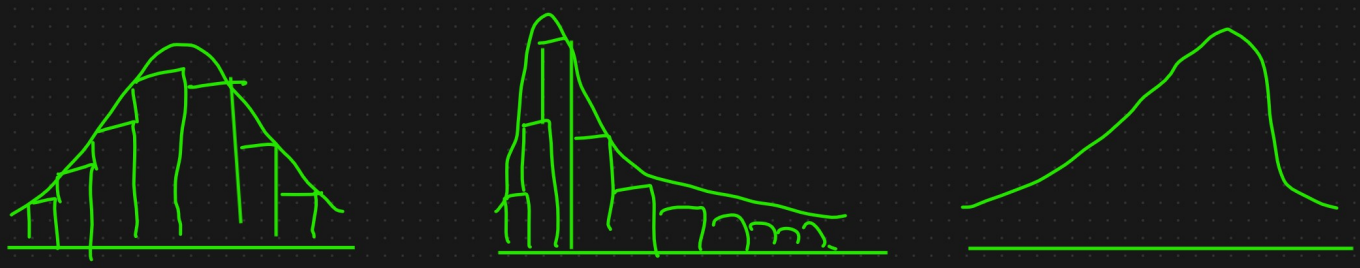
⇒ Kde



Bin size = $\frac{\text{Max} - 0}{\text{bins}}$

bins:

Ages = { 0, 10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 43, 50, 51, 65, 68, 78, 90, 95, 100 }



Assignment.

Weight = {30, 35, 38, 42, 46, 58, 59, 62, 63, 68, 75, 77, 80, 90, 95}

bins = 10

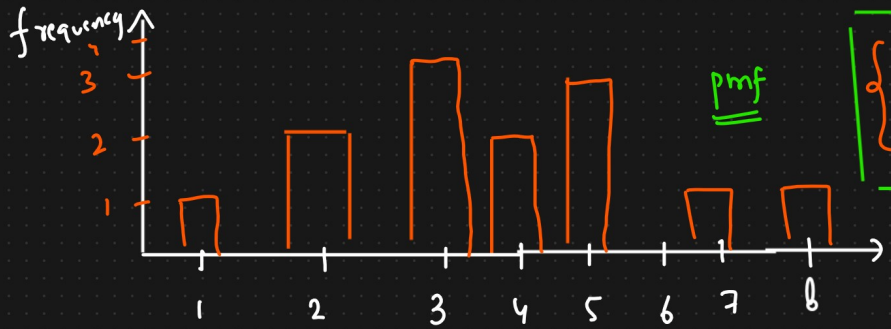
bin size = $\frac{95-30}{10} = \frac{65}{10} = 6.5$

{Continuous value}

② Discrete :

pdf

No. of Banks accounts = [2, 3, 5, 1, 4, 5, 3, 7, 8, 3, 2, 4, 5]



pmf

{Probability mass function}

⇓ pmf

pdf : probability density function → Continuous

pmf : probability mass function → discrete.

Measure of Central Tendency

① Mean ✓

② Median

③ Mode.

{ A measure of CT is a single value that attempts to describe a set of data identifying the central position

Mean $X = \{1, 2, 3, 4, 5\}$ Average / Mean $= \frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$

Population (N) $N \gg n$ Sample (n)

Population mean (μ) $= \sum_{i=1}^{N} \frac{x_i}{N}$ $n \not\approx N$

Sample mean ($\bar{x}$) $= \sum_{i=1}^{n} \frac{x_i}{n}$

$N = 6$ $N > n$ $n = 4$

Population Age $= \{24, 23, 2, 1, 28, 27\}$ Sample Age $= \{24, 2, 1, 27\}$

$\{24, 23, 28, 27\}$ ← Average

$\bar{x} \neq \mu$

13
54
$\bar{x}_1$

Population mean (μ) $= \frac{24+23+2+1+28+27}{6}$ Sample mean ($\bar{x}$) $= \frac{24+2+1+27}{4}$

$\mu = 17.5$ $\bar{x} = 13.5$

$\mu \gtrless \bar{x}$

$\bar{x} \gtrless \mu$

np.nan ← Null values

Practical Application (Feature Engineering)

Age	Salary	Family Size
–	–	–
–	–	–
→ NAN	–	–
–	–	–
–	NAN	–
–	–	NAN
–	NAN	–
→ NAN	–	–

← Loss of Info

Age: 29.6 → 38

NAN ⇒ Not A Number

Null → Val = 10/4

Mean 10/4 = 1, 2, 3, 4 NAN NAN → mean

Age	Salary
24 ✓	45
28 ✓	50
29 ✓	NAN
NAN ×	60
31 ✓	75
36 ✓	80
NAN ×	NAN

Salary: 62 → 85

Outliers ∈ [80] [200] ∈ ∈

④ Median

{1,2,3,4,5} = {1,2,3,4,5,100}

↓outlier

$\bar{x} = 3 \longrightarrow \bar{x} = 19.16$

$\frac{1+2+3+4+5+100}{6}$

$= \frac{115}{6} = 19.16$

Steps to find out median

① Sort the Numbers

② Find the central number { ① if the no. of elements are even we find the average of central elements

② if the no. of elements are odd we find the central elements.

Sorted

{0,1,2,3,4,[5,6],7,8,100,120}

Mean = 25.6

median = $\frac{5+6}{2} = 5.5$

median = 5

③ Mode : {Most frequent occuring elements}

{1,2,2,[3,3,3],4,5}

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

[2,3]

Dataset

Types of flower {Categorical variable}

Lily

Sunflower

Rose

NAN ← ROSE

Rose

Sunflower

Rose

NAN ← ROSE

[

FE

40% ⇒ Blood

Under 19

17, 18, 19, 16, 15, 32 → Outlier

(F) Measure of Dispersion

① Variance (σ^2) ← Spread of Data.

② Standard deviation (σ) ←

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

Variance

Population Variance (σ^2) { Degree of freedom

$$\sigma^2 = \sum_{i=0}^{N} \frac{(x_i - \mu)^2}{N}$$

{ Bessels Correction }

Sample Variance (s^2)

$$s^2 = \sum_{i=0}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$

⇓

Assignment ⇐

{1-1000} = ⇓ First one.

{1-100} = Second one. ⇓

$\{1,2,3,4,5,6,7,8,9,10\}$ (1-10) $=$ Variance

$\{1,2,3,4,50,60,70,100\}$ (1-100) $=$ Variance

Variance {verify}.

$21 + 80 = 101$, $101/7$

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

$\mu = 3$

$\sigma^2 = \dfrac{[(1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2]}{5}$

$= \dfrac{4+1+0+1+4}{5} = \dfrac{10}{5} = 2$

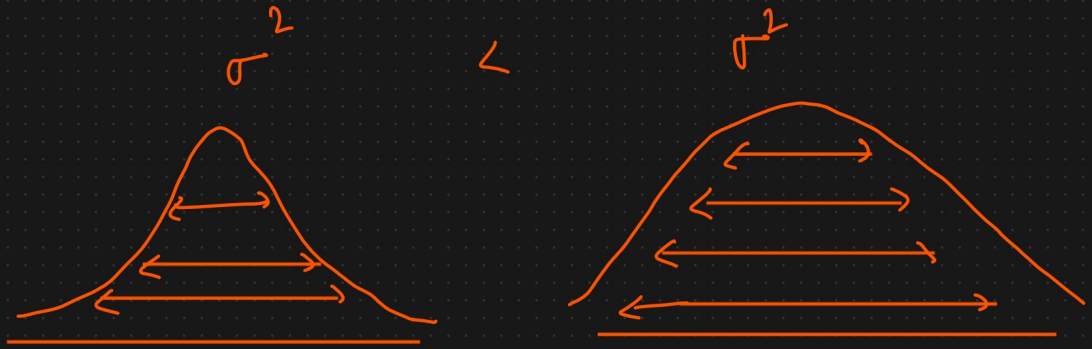
$\{1,2,3,4,5,6,80\}$

$\mu = 14.4$

$\sigma^2 = \dfrac{(1-14.4)^2 + (2-14.4)^2 + (3-14.4)^2 + (4-14.4)^2 + (5-14.4)^2 + (6-14.4)^2 + (80-14.4)^2}{7}$

Variance ⇑ Spread ⇑

$\sigma^2 = 719.10$



3.00
1.41
1.59

⑧ Standard deviation $\left(\sqrt{\sigma^2}\right) \Rightarrow \boxed{\sigma}$ $\boxed{4}$

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

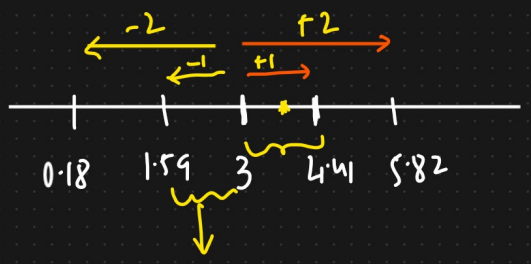
$\mu = 3$

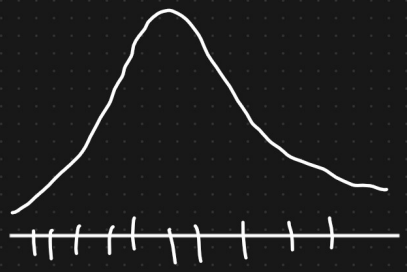
$\sigma^2 = 2$

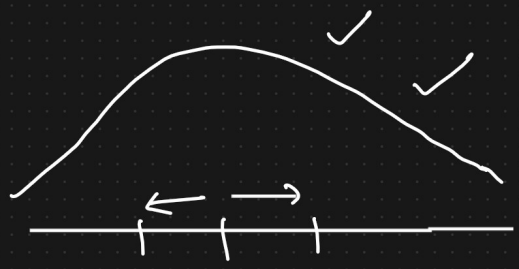
$\sigma = \sqrt{2} = 1.41$

3 ⇒ 0 Standard deviation

4.41
1.41
5.82







⑨ Percentiles And Quartiles

Percentage = $\{1, 2, 3, 4, 5, 6, 7, 8\}$

$$\text{Percentage of Even Number} = \frac{\text{No. of even Numbers}}{\text{Total no. of Number}} = \frac{4}{8} = 0.5 = 50\%$$

Percentiles : GaK, CAT, IELS, SAT, GRE, JEE, NEET ⇒ Percentiles

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

99 percentile = It means the person has got better marks than 99% of the entire students

Ascending order

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

Rank

What is the percentile rank of 10 ?

Yes

CAT Exam 99%

Next class

Percentile Rank of x = $\frac{\#\text{No. of value below } x}{n}$ = $\frac{16}{20}$ = 80 percentile.

=0.8

45 percentile

= $\frac{14}{20}$ = 70 percentile

(4) What is the value that exists at 25 percentile

95%

$$\text{Value} = \frac{\text{Percentile}}{100} * (n+1)$$

#n

$$= \frac{25}{100} \times 20 = 5^{\text{th}} \text{ Index}$$

Op = 5

$$= \frac{95}{100} \times 21$$

(5) 5 number Summary

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

Box plot

=> Remove the outliers.

⑤ Maximum

{1, 2, 2, 2, 3, 3.3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 24}

24 ↓ outlier

15 16

5.25

[Lower Fence ⟷ Higher Fence]

[-3.65 ⟷ 14.25]

Lower Fence = Q1 − 1.5 (IQR)

Higher Fence = Q3 + 1.5 (IQR)

IQR = Q3 − Q1 (75 − 25)

Inter Quartile Range (IQR)

$Q1 = \frac{25}{100} \times 21 = 5.25$ Index = 3

$Q3 = \frac{75}{100} \times 21 = 15.75$ Index: $\frac{8+7}{2} = 7.5$

Lower Fence = $3 - (1.5)(4.5) = -3.65$

Higher Fence = $7.5 + (1.5)(4.5) = 14.25$

{1, 2, 2, 2, 3, 3.3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 24}

-5

5 Number Summary.

① Minimum = 1 ✓

② Q1 = 3 ✓

③ Median = 5 ✓

④ Q3 = 7.5 ✓

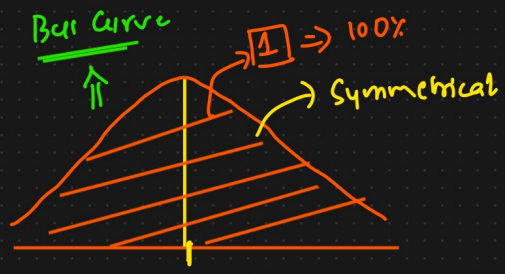
⑤ Maximum = 9 ✓

Box Plot

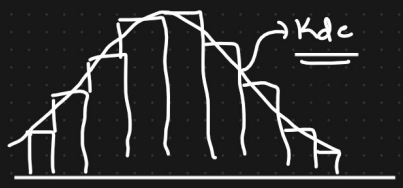
To Treat outliers

Day 3 - Stats

1. Normal Distribution ✓
2. Standard Normal Distribution ✓
3. Z-score ✓
4. Standardization And Normalization ✓
5. Gaussian / Normal Distribution



Kernel density estimator



Age, weight, height → Dismbution

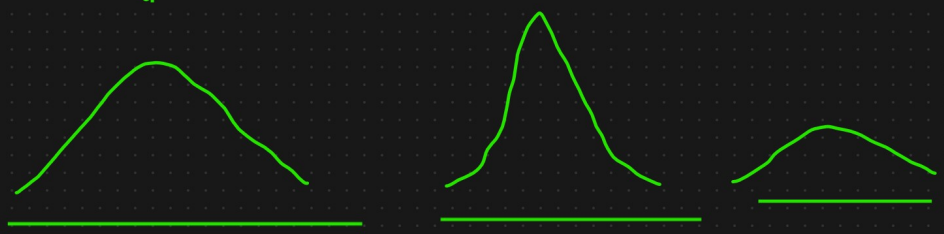
=> Doctors

Domain Expertise

IRIS DATASET ⇐

Petal length, Sepal length, petal width, Sepal width

Gaussian Distri



6. [Empirical Rule of Normal Distribution]

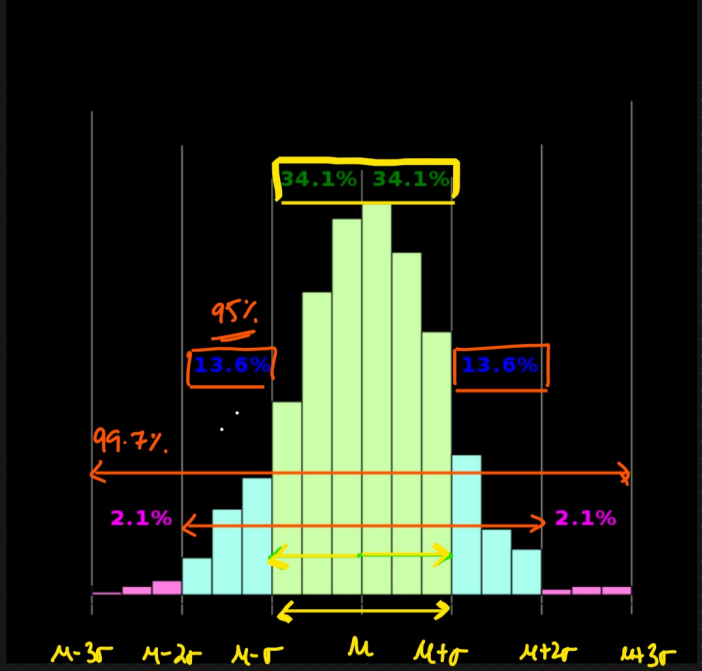
Empirical Rule

⇑

68 - 95 - 99.7 %

⇓

CLEAR



Gaussian / Normal Distribution

⇓

Assumptions of the data

Age = { } ⇒ Gaussian / Normal Dist

✓

Q-Q plot ⇒ Distribution is Gaussian or Not?

Standard Normal Distribution

$X \sim$ Gaussian Distribution (μ, σ)

⇓

$y \sim SND(\mu = 0, \sigma = 1)$.

⇓ $Z\text{-score} = \frac{x_i - \mu}{\sigma}$

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

⇓

$\mu = 3$

$\sigma = 1.41$

$Z\text{-score} = \frac{x_i - \mu}{\frac{\sigma}{\sqrt{n}}}$ $\quad n = 1$

⇒ Standard Error ⇒ Inferential stats.

$Z\text{-score} = \left| \frac{x_i - \mu}{\sigma} \right|$ ⇐ Simple

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

$\mu = 3 \quad \sigma = 1.41$ &

$= \frac{1-3}{1.414} = -1.414$

$: \frac{2-3}{1.414} =$

$\frac{4-3}{1.414} = \frac{1}{1.414} =$

$\rightarrow Y = \{-1.414, -0.707, 0, 0.707, 1.414\}$

Why? [Standardization] $\Rightarrow$ $\mu = 0$ & $\sigma = 1$

Age (years)	Weight (kg)	Height (cm)
24	72	150
26	78	160
32	84	165
33	92	170
34	87	150
28	83	180
29	80	175

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

0-1

Same Scale

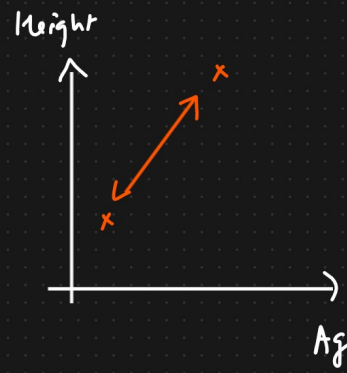
Machine Learning

Maths Equations

$\Downarrow$

Algorithm $\Rightarrow$ Mathematical Model

Mathematical Calculation Time $\uparrow\uparrow\uparrow$



Feature Scaling

✓ Normalization $\Rightarrow$

$\Downarrow$

$[0 - 1]$ $[-1 - 1]$

$[0 - 5]$

$[0 - 4]$

Standardization $\{$ Z-score $\}$

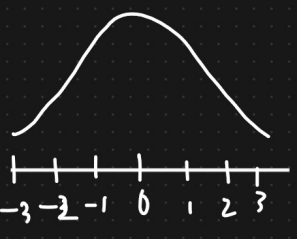
$\Downarrow$

$\mu = 0, \sigma = 1$

$[-3 \longleftrightarrow 3]$

$x_i - \mu$

$\Downarrow$



Normalization [lower scale ↔ Higher scale] → Images => 0-255

Standard

① Min Max Scaler [0-1]

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

$= \frac{1-1}{5-1} = 0$

$\frac{2-1}{5-1} = \frac{1}{4}$

$\frac{3-1}{5-1} = \frac{2}{4}$

$\frac{4-1}{5-1} = \frac{3}{4}$

x	y	y'
1	0	-1.414
2	0.25	-0.707
3	0.5	0
4	0.75	0.702
5	1	1.414

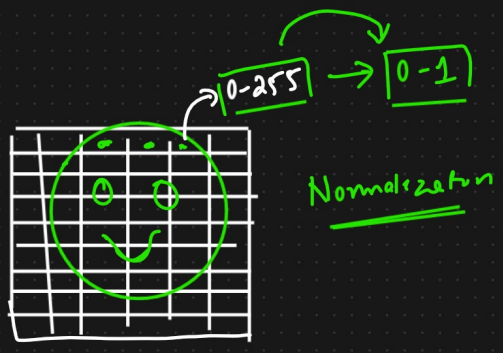
⇓ Apply ??

RNN ANN CNN

B.C PMD NASA

BF BF BF

Deep learning



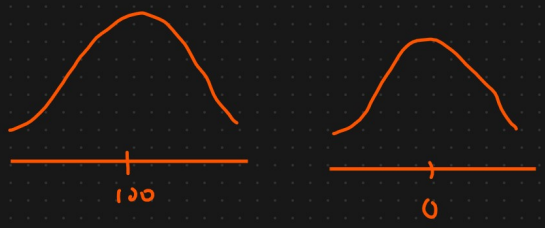
① Standardization

$$Z\text{-score} = \frac{x_i - \mu}{\sigma}$$

X → Normal Distribution (μ, σ)

⇓ Z-score.

y → SND $(\mu = 0, \sigma = 1)$



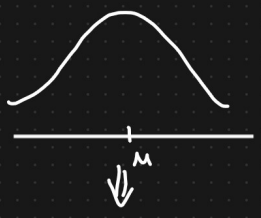
Why do we do this → Bring the features in the same scale

Normalization [0 - 1]

① Min Max Scaler ✓ ⇔ Standardization

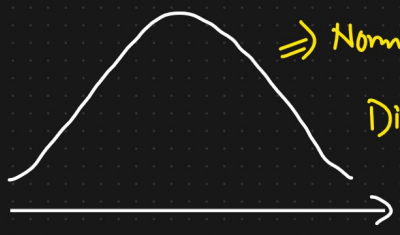
↓ ML

↓ ML



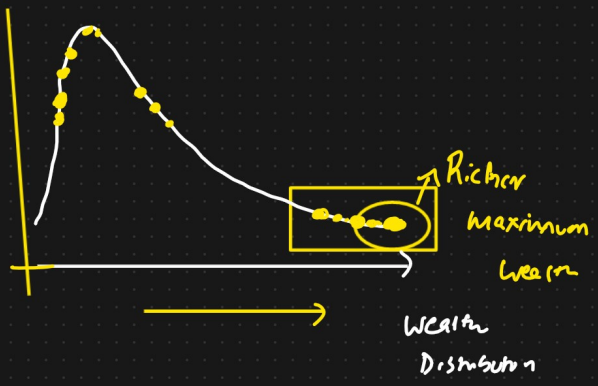
Min Max Scaler

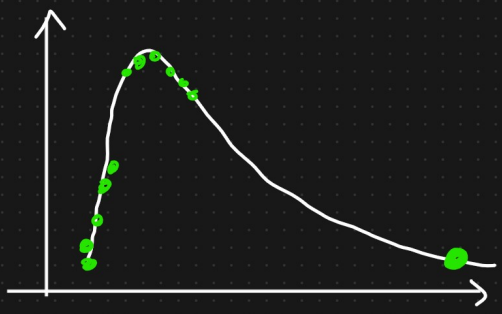
① Log Normal Distribution



⇒ Normal/Gaussian Distribution.

Log Normal Distribution



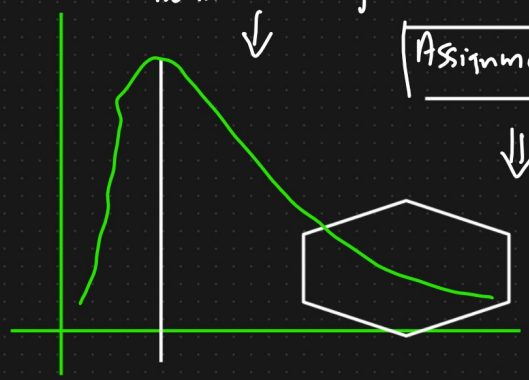


mean will be higher

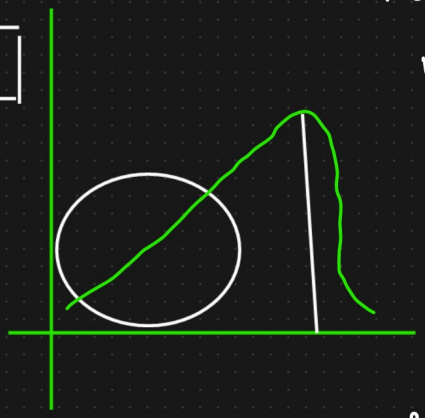
↓

Assignment

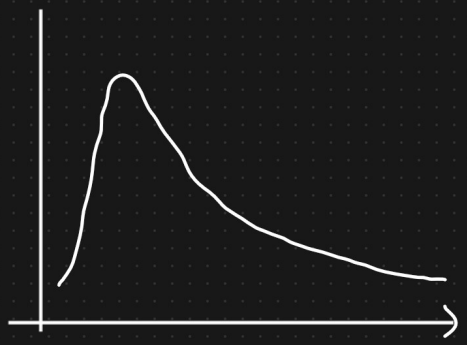
↓



Relation of mean, median, mode

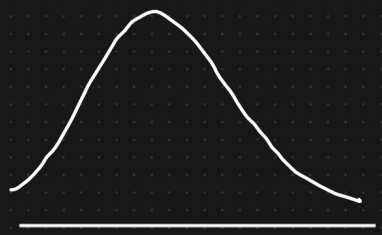


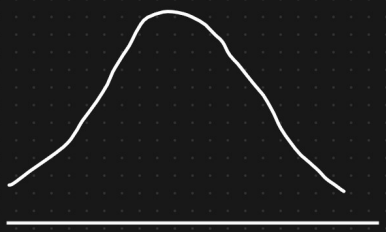
From Ascending order give the relation of mean, median & mode?



$X \sim$ Log Normal Distribution

$y = \ln(x)$



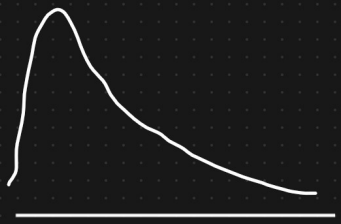


$X \sim ND(\mu, \sigma)$

antilog

$\Rightarrow \exp(x)$

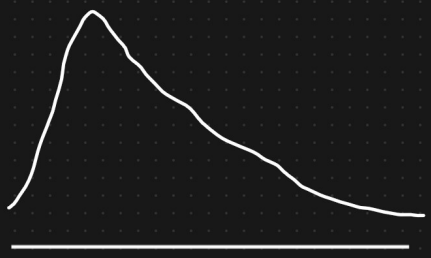
y



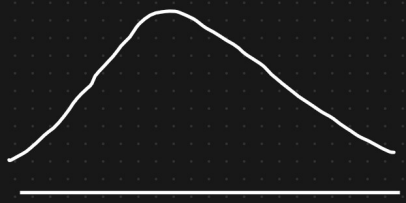
(¥)

Natural log

$\log_e$



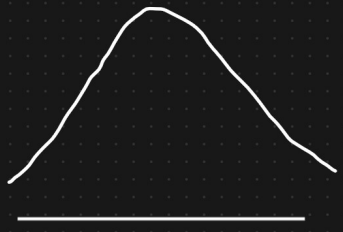
$\Rightarrow y = \ln(x)$



$X \sim$ Log Normal Distribution

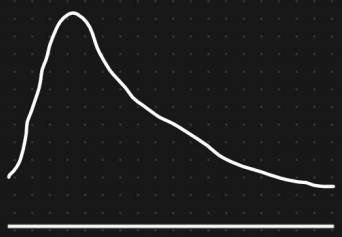
(μ, σ)

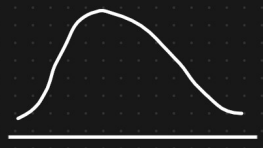
Inverse



y

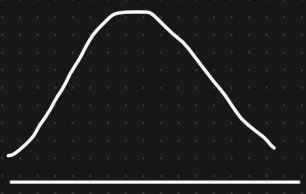
$\Rightarrow X = \exp(y)$

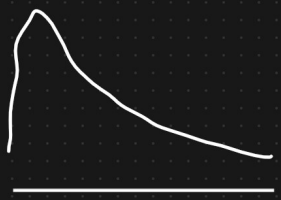


$X \sim f_{ND}$ $\Rightarrow$ $\log_c(x)$ $\Rightarrow$  $\Leftarrow$ QQ plot

$\Downarrow$

Tomorrow



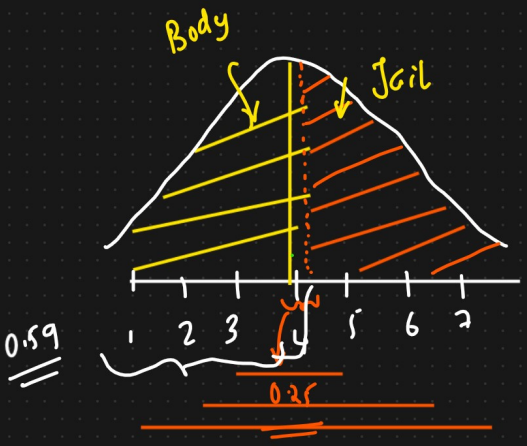
$\Rightarrow y = \exp(x)$ $\Rightarrow$ 

x

(#) $X = \{1, 2, 3, 4, 5, 6, 7\}$

$\mu = 4$

$\sigma = 1$



Question: What is the percentage of score that falls above 4.25?

fall below 3.75?

0.59 $\Rightarrow$ 59%

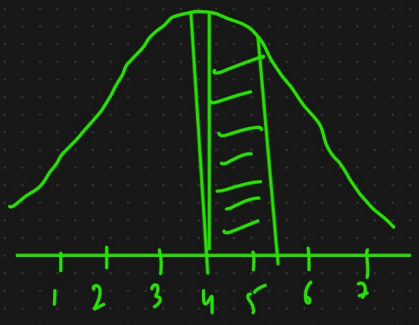
$1 - 0.59 = 0.41 \Rightarrow$ 41%

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

① Z-table (Area under the curve)

↳

$Z\text{-score} = \frac{3.75 - 4}{1} = -0.25 \approx 40\%$



4.75 & 5.25

(4) In India the average IQ is 100 with a Standard Deviation of 15. What is the percentage of population would you expect to have an IQ

Answers.

① Lower than 85 = 0.1587
② Higher than 85 = 0.8413
③ Between 85 and 100 = 0.3413

Assignment

Day 4 - STATS

1. Central limit Theorem. ✓
2. Probability. ✓
3. Permutation And combination ✓
4. Covariance, Pearson Correlation, Spearman Rank Correlation. ✓
5. Bernoullis Distribution
6. Binomial Distribution
7. Power law (Pareto Distribution).

① Central Limit Theorem

$n < 30$ ⇓⇓ ⇒ $n \geq 30$ ⇑⇑

Size of sample ↑ n → m ↑ No. of samples

The larger the value the better

Population data. ⇒ Gaussian / Normal Distribution

N — μ, σ

$S_1 \rightarrow \{ x_1\ x_2\ x_3 \cdots x_n \} \rightarrow \bar{x}_1 = \bar{s}_1$

$S_2 \rightarrow \{ x_3\ x_4 - x_1 \cdots x_n \} \rightarrow \bar{x}_2\ \bar{s}_2$

$S_3 \rightarrow \{ x_4\ x_1 \cdots x_n \} \rightarrow \bar{x}_3\ \bar{s}_3$

$\bar{x}_m\ \bar{s}_m$

Sampling with replacement

μ, σ ⇐ Log Normal Distribution =

⇐ Gaussian Distribution

μ, σ ⇐ Left skewed

10 different region $n \geq 30$

↳ Assumptions

Size of shark through out the world ?

② **Probability** : Probability is a measure of the likelihood of an event

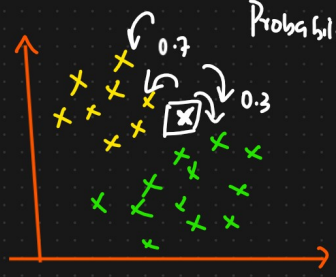
Eg: Tossing a fair coin $P(H) = 0.5$ $P(T) = 0.5$

SHOLAY ⟶ COIN $P(H) = 1$

⇓

unfair coin

Strong Basic ⇑ Probabilities

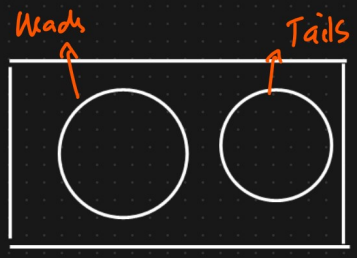


Rolling a Dice $P(1) = \frac{1}{6}$ $P(2) = \frac{1}{6}$ $P(3) = \frac{1}{6}$

① **Mutual Exclusive Event**

Two Events are mutually exclusive if they cannot occur at the same time

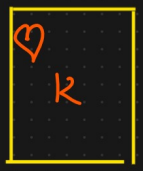
① Tossing a coin



② Rolling a dice

② **Non Mutual Exclusive Events**

Two events can occur at the same time.



Bag of Marbles

① Picking randomly a card from a deck of cards, two events "heart" and "king" can be selected.

Mutual Exclusive Event

① What is the probability of coin landing on heads or tails

⇓

Addition Rule for mutual exclusive events

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

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

② What is the probability of getting 1 or 6 or 3 while rolling a dice?

$$P(1 \text{ or } 6 \text{ or } 3) = P(1) + P(6) + P(3)$$

$$= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2} //.$$

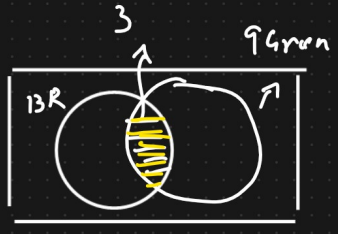
Non Mutual Exclusive Event

Bag of Marbles : 10 Red, 6 Green, 3 (R&G) (R/G)

③ When picking randomly from a bag of marbles what is the probability of choosing a marble that is red or green?

⇓

Non mutual Exclusive



Addition Rule for Non Mutual Exclusive Event

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

$$= \frac{13}{19} + \frac{9}{19} - \frac{3}{19} = \frac{19}{19} = \frac{1}{=}.$$

Deck of cards → What is the probability of choosing ♡ or Queen

$$P(\heartsuit \text{ or Queen}) = P(\heartsuit) + P(\text{Queen}) - P(\heartsuit \text{ and Queen})$$

$$= \frac{13}{52} + \frac{4}{52} - \frac{1}{52} = \boxed{\frac{16}{52}}$$ //

⑭ Multiplication Rule

⑱ Dependents Events : Two events are dependent if they affect one another

Bag of marbles { o o o ~~o~~ / o o o }

$\Rightarrow P(W) = \frac{4}{7}$ → $P(Y) = \frac{3}{6}$

↑ white marble

⑱ What is the probability of rolling a "5" and then a "3" with a normal 6 sided dice?

{ Independent Dependent }

Ans) $P(1) = \frac{1}{6}$ $P(2) = \frac{1}{6}$ $P(3) = \frac{1}{6}$ $P(4) = \frac{1}{6}$

Multiplication Rule for Independent Events

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

$$= \frac{1}{6} * \frac{1}{6} = \boxed{\frac{1}{36}}$$

P(A or B) ⇒ → Mutual Exclusive
→ Non Mutual Exclusive

P(K or Q) ↑ common

P(A or B) = P(A) + P(B) − P(A and B) → Non Mutual Exclusive.

P(A or B) = P(A) + P(B) [Mutual Exclusive]

Dependent and Independent Events

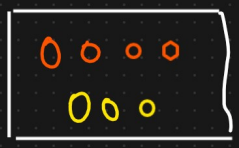
Tossing a coin ⇐ Sample {H & T}

P(H) = 0.5 P(T) = 0.5

Event A ↓ Event B ↓

P(A and B) = P(A) * P(B)

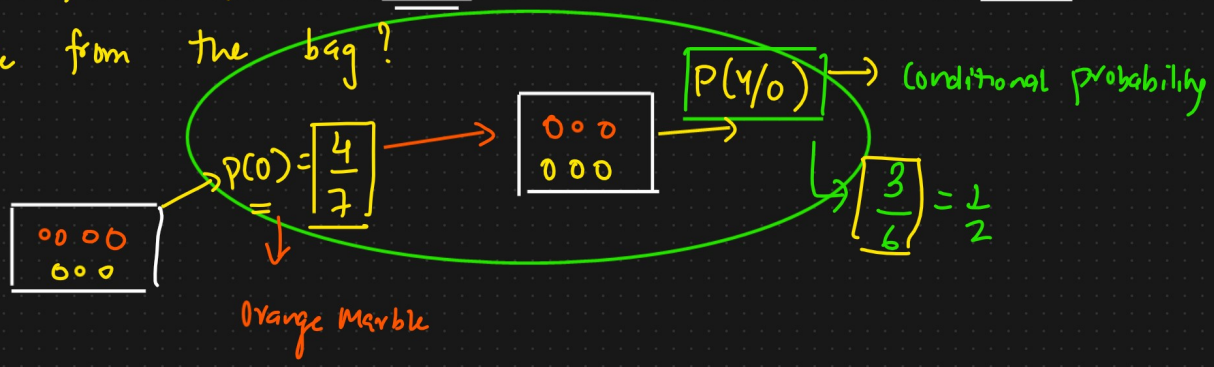
②



⇒ Dependent Events

Probability of drawing a "Orange" and then drawing a "yellow" marble from the bag?

Ans)



Naive Bayes

P(O and Y) = P(O) * P(Y/O)

= 4/7 * 3/6 = 4 2 / 14 7 = 2/7

⊛ Permutation

School of Children

{ Dairy Milk, Kit Kat, Milky Bar, Sneakers, 5 star }

5 × 4 × 3

= 60 ways ⇒ Permutation

With permutation, order matters

{ DM KK MB } { } { } { }

{ KK DM MB } ✓ { } { } { }

{ } { } { } { } { }

Possible Arrangements.

$n = 5$

n = Total No. of objects

r = # of selection

$$^{n}P_{r} = \frac{n!}{(n-r)!} = \frac{5!}{(5-3)!}$$

$$= \frac{5 \times 4 \times 3 \times 2!}{2!} = 60$$

⊛ Combination

Repetition will not occur

{ DM KK MB }

Unique Combination

X { MB KK DM } ←

$$^{n}C_{r} = \frac{n!}{r!(n-r)!} = \frac{5!}{3!(2)!} = \frac{5 \times \overset{2}{\cancel{4}} \times \cancel{3 \times 2 \times 1}}{\cancel{3!} \times 2}$$

$= 10//$

DREAM 11 ⇐

⑧ Covariance ✓

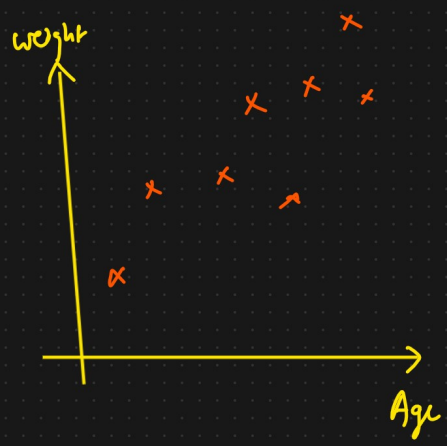
{Feature Selection}

X	Y
Age	Weight
12	40
13	45
15	48
17	60
18	62
$\bar{x} = 15$	$\bar{y} = 51$

Age ↑	Weight ↑
Age ↓	Weight ↓

⇑

Quantify the relationship x & y using mathematical question



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

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

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

⇓

$\boxed{2y} \Rightarrow$ +ve

$Cov(x,x)$ ⇐

$$Cov(x,x) = Var(x) \Leftarrow$$

+ve Covariance

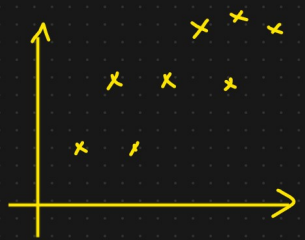
x ↑	y ↑
x ↓	y ↓

-ve Covariance

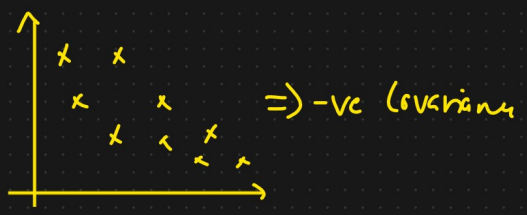
x ↑	y ↓
x ↓	y ↑

Covariance = 0

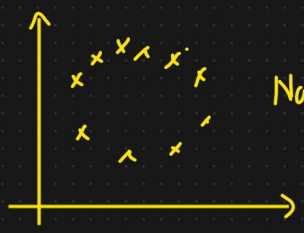
No relation with x & y



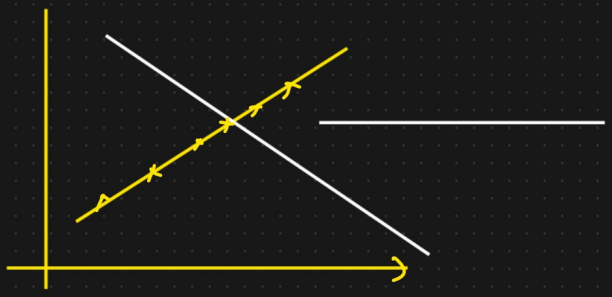
⇒ +ve.



=) -ve Covariance



No relation.



X	Y
10	4
8	6
7	8
6	10
7.25	7

$Cov(x,y) =$ -ve

$= [(10-7.25)(4-7) + (8-7.25)*(6-7) + (7-7.25)*(8-7) + (6-7.25)*(10-7)$

$= -3.25$

x↑	y↓
x↓	y↑

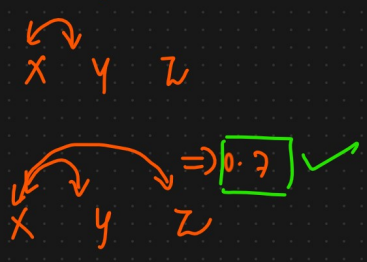
Scale

Covariance = +ve, -ve

⑧ Pearson Correlation Coefficient (-1 to 1)

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

-ve, +ve



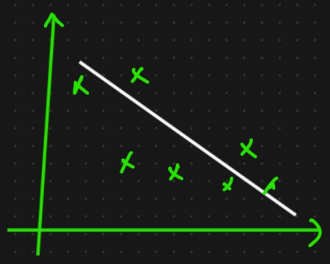
0.5

More the value towards +1

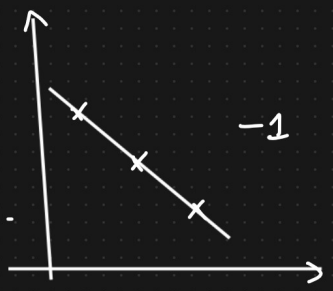
More +ve correlated it is

-1

negative correlated



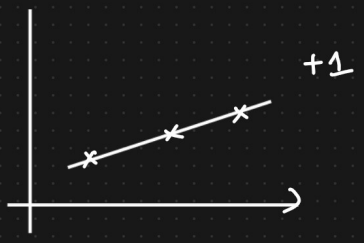
-1 to 0



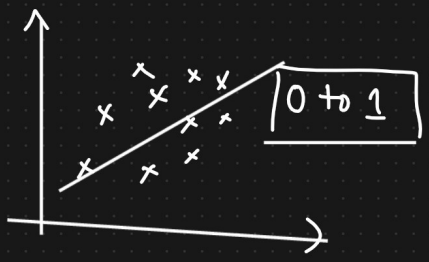
-1

$x\downarrow\ y\uparrow$

$x\uparrow\ y\downarrow$



+1



0 to 1

(K) Spearman Rank Correlation

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

Ascending order

X	y
10	4
8	6
7	8
6	10

R(x)	R(y)
4	1
3	2
2	3
1	4

Ascending

Spearman Rank Correlation

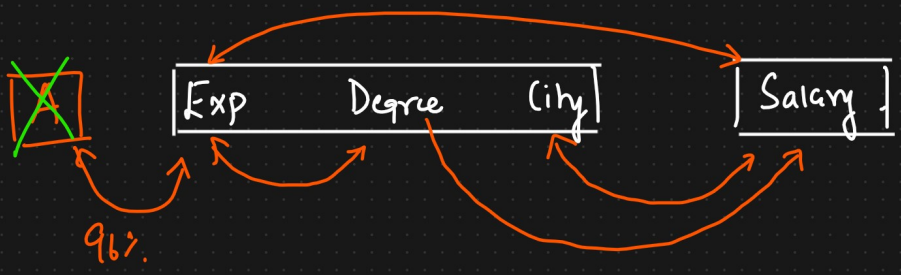
Why this Correlation will be used ?

0.95

95%

↓ +ve, -ve } Good ↑

0.2 | 0.01



+ve Correlated

+ve Correlated

+ve Correlated

Inferential Statistics

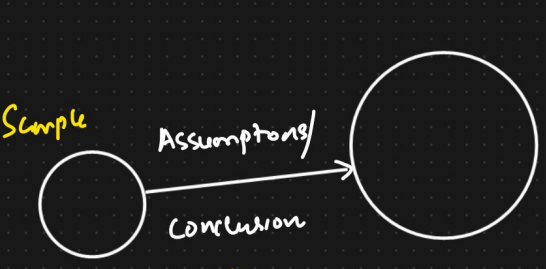
1. Hypothesis Testing
2. p-value
3. Confidence Interval
4. Significance Value

Z test
t test
Chi square test
Annova test (F-test)

3 Distributions
1. Bernoulli's
2. Binomial
3. Power Law

→ TRANSFORMATION

Inferential Stat



⇓ Hypothesis Testing

Steps of Hypothesis Testing

Experiment → [Coin is fair or not]

P(H) = 0.5 P(T) = 0.5

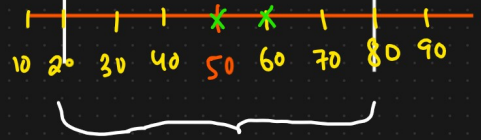
1. Null Hypothesis : Coin is fair ⇒ Accepted
2. Alternate Hypothesis : Coin is not fair
3. Perform Experiments

10 times
100 times
75
60 40
70 30
80 20

10 → Null Hypothesis is Rejected
→ Alternate Hypothesis is Accepted

⇓
50 times Head ⇒ Fair
60 times Head

C.I = [20-80]
⇓
Coin is fair



C.I ⇒ Confidence Interval

70 times ⇒ Domain Expert
⇓
Confidence Interval

↳ We fail to Reject the Null Hypothesis [within C.I]
↳ We Reject the Null Hypothesis [outside C.I]

⇒ Conclusions

② Person is criminal or not {Murder Case}

① Null Hypothesis : Person is not Criminal

② Alternate Hypothesis : Person is criminal

③ Experiment / Proof : DNA, finger print, weapons, eye witness, footage

⇓

Judge ⇒

Conclusions

Vaccines ⇒ Medical ⇒ critical

Confidence Interval : (CI) ⇒ Domain Experiment

Significance Value

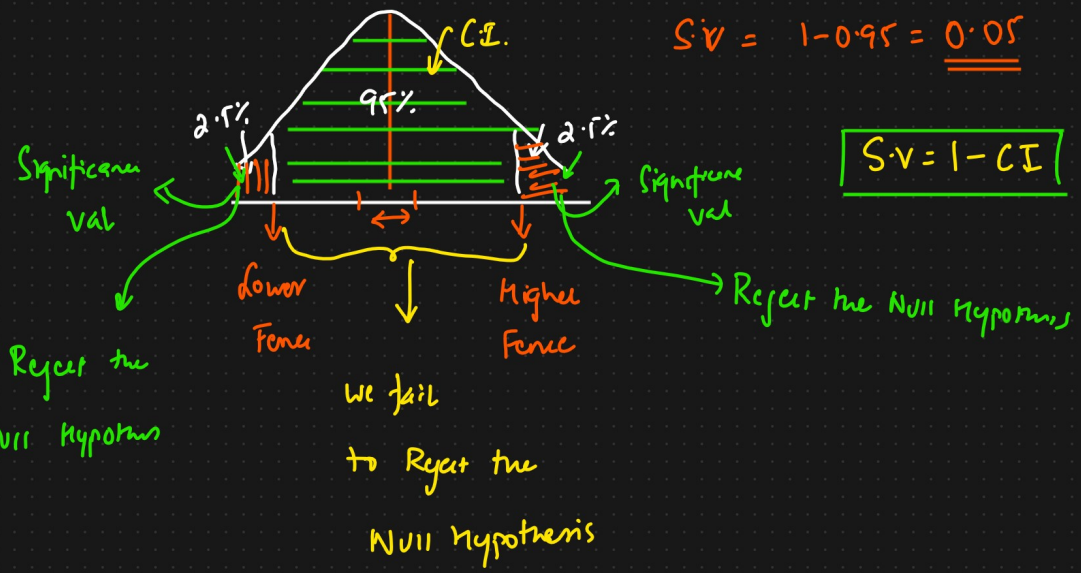
95%

C.I = 95%

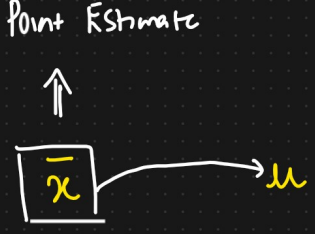
S.V = 1 - C.I

S.V = 1 - 0.95 = 0.05

S.V = 1 - CI

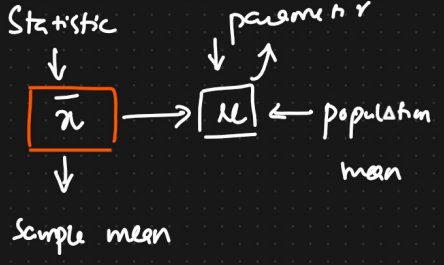


Point Estimate : The value of any statistic that estimates the value of a parameter is called Point Estimate



$\bar{x} > \mu$

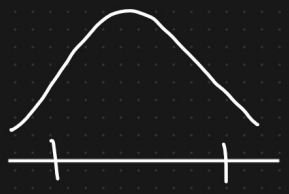
$\bar{x} \leq \mu$



Point Estimate ± Margin of Error = Parameter ⇒ population mean

Lower C.I : Point Estimate − Margin of Error

Higher C.I : Point Estimate + Margin of Error



Margin of Error = $Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ ⇒ Standard Error

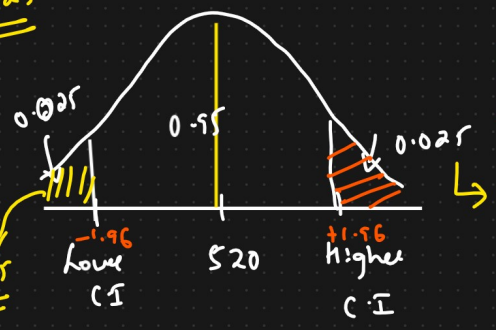
σ → population sd

α = Significance Value.

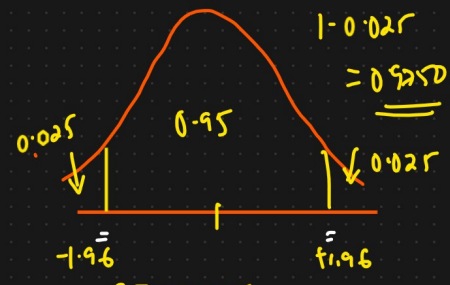
(*) On the quant test of CAT Exam, a sample of 25 test takers has a mean of 520 with a population standard deviation of 100. Construct a 95% C.I about the mean?

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

$\frac{5}{2} = 2.5 \Rightarrow 0.025$



1 − 0.025 = 0.975



C.I = 0.95

S.V = 1 − 0.95 = 0.05

Lower C.I = Point Estimate − Margin of Error

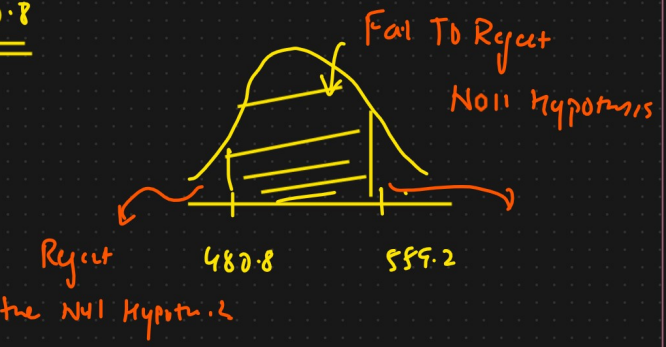
$= 520 - Z_{0.05/2} \frac{\sigma}{\sqrt{n}}$

$= 520 - Z_{0.025} \frac{100}{\sqrt{25}}$ 20.

$Z_{0.05/2} \Rightarrow Z_{0.025}$

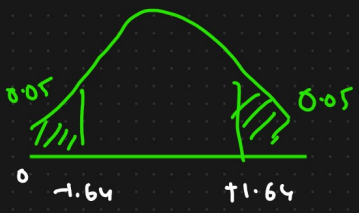
$= 520 - 1.96 \times 20 = 480.8$

Higher C.I $= 520 + 1.96 \times 20 = 559.2$



(1) $\bar{x} = 480$ $\sigma = 85$ $n = 25$ C.I = 90% Significance

$= 1 - 0.90 = 0.10$



Lower C.I $= 480 - Z_{0.10/2}\left[\frac{85}{5}\right]$

$= 480 - Z_{0.05}\left[\frac{85}{5}\right]$

$= 480 - 1.64\left[17\right]$

$= 480 - 27.8 = 452.12$

Higher CI $= 480 + 1.64[17]$

$= 480 + 27.8$

$= 507.8$

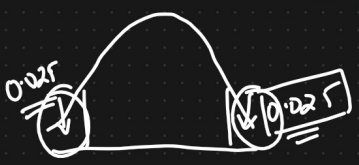
$[452.12 \leftrightarrow 507.8]$.

(2) On the quant test of CAT exam, a sample of 25 test takers has a mean of 520, with a sample standard deviation of 80. Construct 95% C.I about the mean?

Ans) $\bar{x} = 520$ $S = 80$ $n = 25$ C.I = 95% S.V $= 1 - 0.95 = 0.05$

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

t test



0.05

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

Lower C.I = $520 - t_{0.05/2}\left(\frac{80}{\sqrt{16}}\right)$.

$t_{0.025}$

$= 520 - 2.064 * 16$

Lower C.I = 486.976

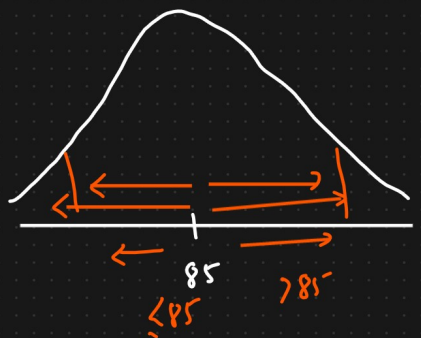
Higher C.I = 553.024

① 1 Tail and 2 Tail Test

Ⓐ Colleges in Town A has 85% placement rate. A new college was recently opened and it was found that a sample of 150 students had a placement rate of 88%. with a standard deviation of 4%.

Does this college has a different placement rate with 95% C.I?

Two Tail.



↓
greater than 85%
lesser than 85%

1 tail Right side

① Z-test
② t-test

Hypothesis Testing

(4) A factory has a machine that fills 80ml of Baby medicines 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 & Alternate Hypotheses

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

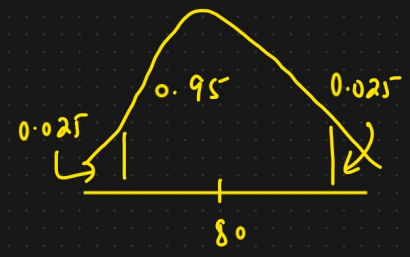
$\mu = 80ml \quad n = 40 \quad \bar{x} = 78 \quad s = 2.5$

Step 1

Ans) Null Hypothesis $\mu = 80 \rightarrow$

Alternate Hypothesis $\mu \neq 80 \rightarrow$

Step 2 : C.I = 0.95 S.V(α) = 1 − 0.95 = 0.05



Step 3 :

$n = 40$

$s = 2.5$

① n ≥ 30 or population sd } → Z test

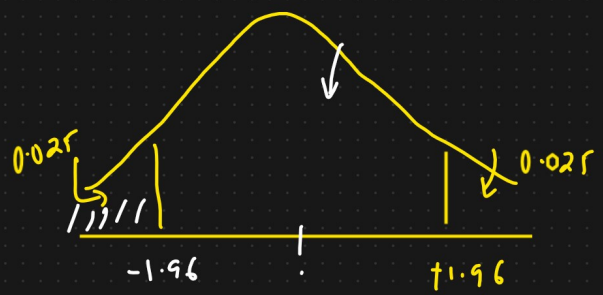
② n < 30 and sample std } → t test

Z test

Let perform the Experiment

Decision Boundary

$1 - 0.025 = 0.975$



(iii) Calculate test statistics (z-test)

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \rightarrow \text{Standard Error} = \frac{78 - 80}{\frac{2.5}{\sqrt{40}}} = -5.05$$

(iv) Conclusions

Decision Rule: If $z = -5.05$ is less than -1.96 or greater $+1.96$, Reject the Null Hypothesis with 95% C.I

Reject the Null Hypothesis { There is some fault in the machine.

(v) A complaint was registered, the boys in a Government School are underfed. Average weight of the boys of age 10 is 32 kgs with S.D = 9kgs. A sample of 25 boys were selected from the Government School and the average weight was found to be 29.5 kgs? With CI: 95% Check it is True or False.

Ans) Conditions for z-test $n=25$ $\mu=32$ $\sigma=9$ $\bar{x}=29.5$

① We Know the population sd. OR

② We do not Know the population sd but our sample is large $n \geq 30$

Conditions For T test

1. We donot know the population std
2. Our sample size is small $n < 30$
3. Sample std is given.

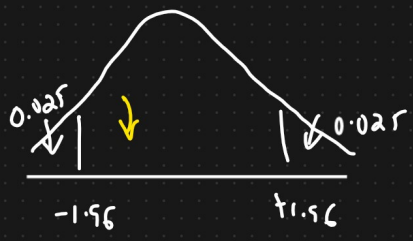
Step 1

1. $H_0 : \mu = 32$

   $H_1 : \mu \neq 32$

2. $C.I = 0.95 \qquad \alpha = 1 - 0.95 = 0.05$

3. Z test



$$Z\text{-score} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{29.5 - 32}{9/\sqrt{25}} = -1.39$$

Conclusion $= -1.39 > -1.96$ Accept the Null Hypothesis 95% C.I

We fail to Reject the Null Hypothesis

The Boys are fed well.

(1) A factory manufactures cars with a warranty of 5 years or more on the engine and transmission. An engineer believes that the engine or transmission will malfunction in less than 5 years. He tests a sample of 40 cars and finds the average time to be 4.8 years with a standard deviation of 0.50. (1) State the null & alternate hypotheses

(2) At a 2% significance level, is there enough evidence to support the idea that the warranty should be revised?

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

$H_1 : \mu < 5$

Step 2: $\alpha = 0.02$ C.I = 0.98

(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% }

Inferential Statistics

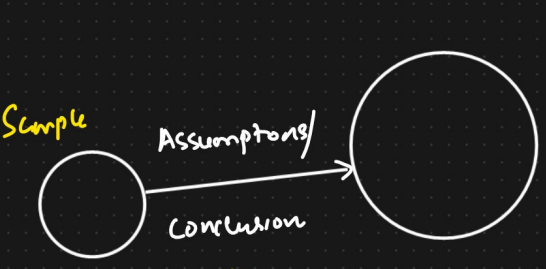
1. Hypothesis Testing
2. p-value
3. Confidence Interval
4. Significance Value

Z test
t test
Chi square test
Annova test (F-test)

3 Distributions
1. Bernoulli's
2. Binomial
3. Power Law

→ TRANSFORMATION

Inferential Stat



⇓ Hypothesis Testing

Steps of Hypothesis Testing

Experiment → [Coin is fair or not]

P(H) = 0.5 P(T) = 0.5

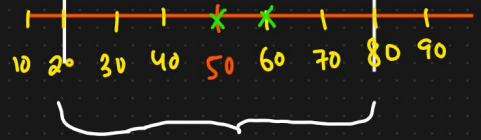
1. Null Hypothesis : Coin is fair ⇒ Accepted
2. Alternate Hypothesis : Coin is not fair
3. Perform Experiments

10 times
100 times
75
60 40
70 30
80 20

10 → Null Hypothesis is Rejected
→ Alternate Hypothesis is Accepted

⇓
50 times Head ⇒ Fair
60 times Head

C.I = [20-80]
⇓
Coin is fair



C.I ⇒ Confidence Interval

70 times ⇒ Domain Expert
⇓
Confidence Interval

↳ We fail to Reject the Null Hypothesis [within C.I]
↳ We Reject the Null Hypothesis [outside C.I]

⇒ Conclusions

② Person is criminal or not {Murder Case}

① Null Hypothesis : Person is not Criminal

② Alternate Hypothesis : Person is criminal

③ Experiment / Proof : DNA, finger print, weapons, eye witness, footage

⇓

Judge ⇒

Conclusions

Vaccines ⇒ Medical ⇒ critical

Confidence Interval : (CI) ⇒ Domain Experiment

Significance Value

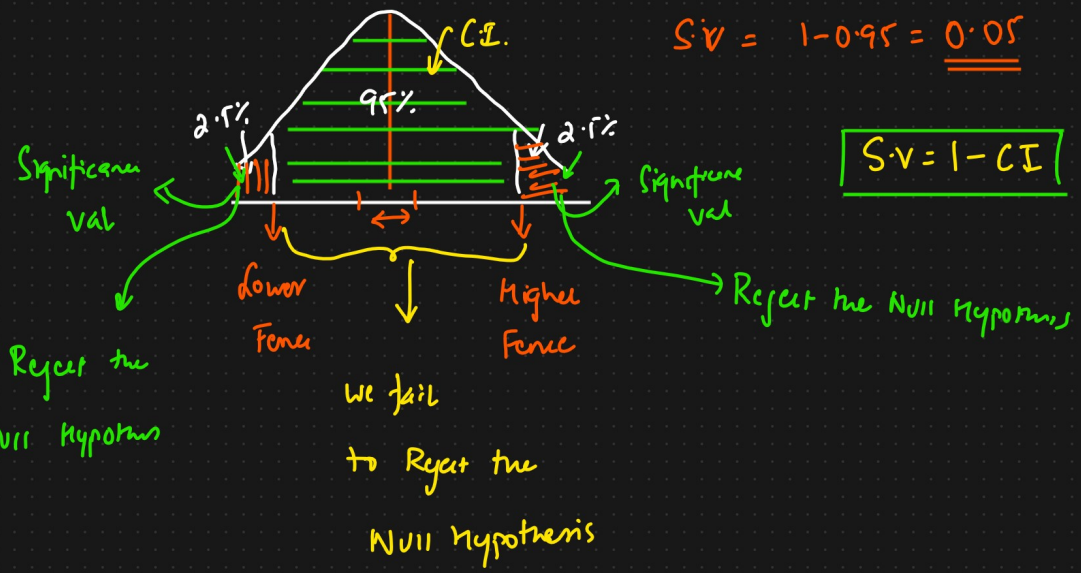
95%

C.I = 95%

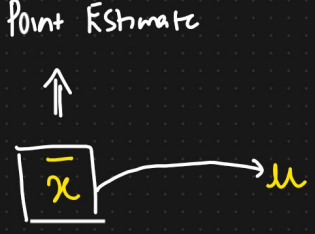
S.V = 1 - C.I

S.V = 1 - 0.95 = 0.05

S.V = 1 - CI

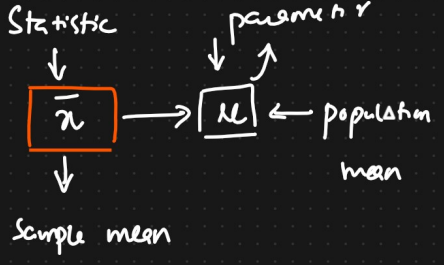


Point Estimate : The value of any statistic that estimates the value of a parameter is called Point Estimate



$\bar{x} > \mu$

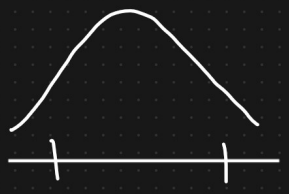
$\bar{x} \leq \mu$



Point Estimate ± Margin of Error = Parameter ⇒ population mean

Lower C.I : Point Estimate − Margin of Error

Higher C.I : Point Estimate + Margin of Error



Margin of Error = $Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ ⇒ Standard Error

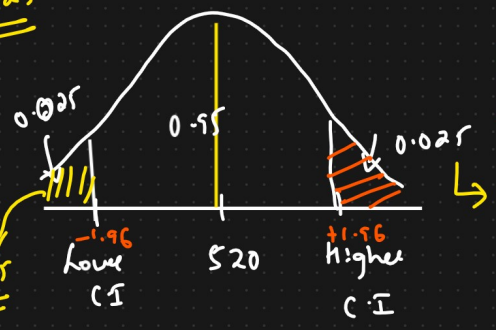
σ → population sd

α = Significance Value.

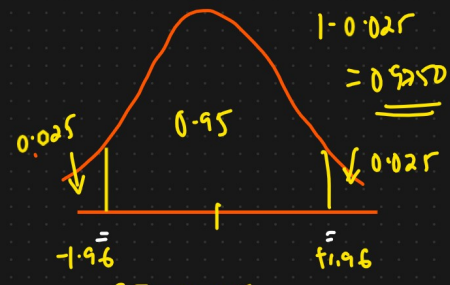
(*) On the quant test of CAT Exam, a sample of 25 test takers has a mean of 520 with a population standard deviation of 100. Construct a 95% C.I about the mean?

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

$\frac{5}{2} = 2.5 \Rightarrow 0.025$



1 − 0.025 = 0.975



C.I = 0.95

S.V = 1 − 0.95 = 0.05

Lower C.I = Point Estimate − Margin of Error

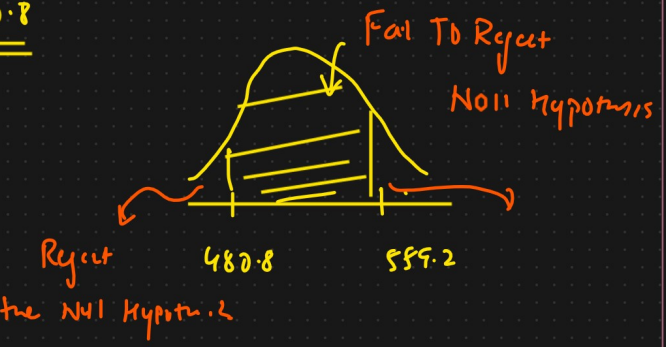
$= 520 - Z_{0.05/2} \frac{\sigma}{\sqrt{n}}$

$= 520 - Z_{0.025} \frac{100}{\sqrt{25}}$ 20.

$Z_{0.05/2} \Rightarrow Z_{0.025}$

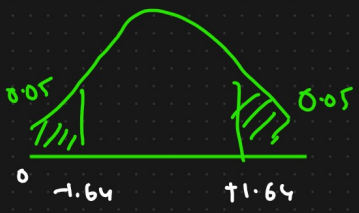
$= 520 - 1.96 \times 20 = 480.8$

Higher C.I $= 520 + 1.96 \times 20 = 559.2$



(1) $\bar{x} = 480$ $\sigma = 85$ $n = 25$ C.I = 90% Significance

$= 1 - 0.90 = 0.10$



Lower C.I $= 480 - Z_{0.10/2}\left[\frac{85}{5}\right]$

$= 480 - Z_{0.05}\left[\frac{85}{5}\right]$

$= 480 - 1.64\left[17\right]$

$= 480 - 27.8 = 452.12$

Higher CI $= 480 + 1.64[17]$

$= 480 + 27.8$

$= 507.8$

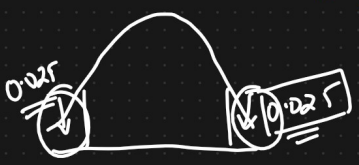
$[452.12 \leftrightarrow 507.8]$.

(2) On the quant test of CAT exam, a sample of 25 test takers has a mean of 520, with a sample standard deviation of 80. Construct 95% C.I about the mean?

Ans) $\bar{x} = 520$ $S = 80$ $n = 25$ C.I = 95% S.V $= 1 - 0.95 = 0.05$

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

t test



0.05

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

Lower C.I = $520 - t_{0.05/2}\left(\frac{80}{\sqrt{16}}\right)$.

$t_{0.025}$

$= 520 - 2.064 * 16$

Lower C.I = 486.976

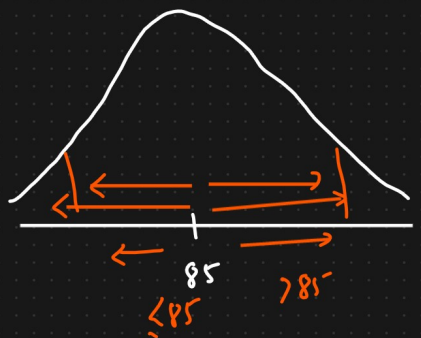
Higher C.I = 553.024

① 1 Tail and 2 Tail Test

Ⓐ Colleges in Town A has 85% placement rate. A new college was recently opened and it was found that a sample of 150 students had a placement rate of 88%. with a standard deviation of 4%.

Does this college has a different placement rate with 95% C.I?

Two Tail.



↓
greater than 85%
lesser than 85%

1 tail Right side

① Z-test
② t-test

Hypothesis Testing

(4) A factory has a machine that fills 80ml of Baby medicines 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 & Alternate Hypotheses

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

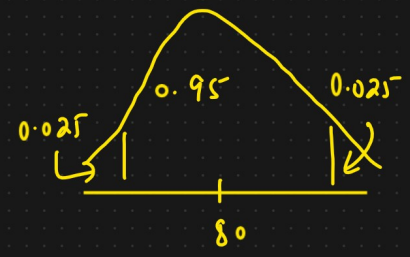
$\mu = 80ml \quad n = 40 \quad \bar{x} = 78 \quad s = 2.5$

Step 1

Ans) Null Hypothesis $\mu = 80 \rightarrow$

Alternate Hypothesis $\mu \neq 80 \rightarrow$

Step 2 : C.I = 0.95 S.V(α) = 1 − 0.95 = 0.05



Step 3 :

$n = 40$

$s = 2.5$

① n ≥ 30 or population sd } → Z test

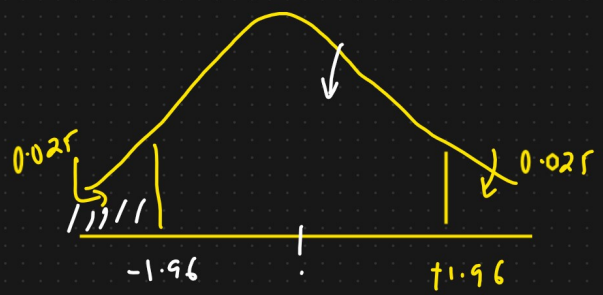
② n < 30 and sample std } → t test

Z test

Let perform the Experiment

Decision Boundary

$1 - 0.025 = 0.975$



(iii) Calculate test statistics (z-test)

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \rightarrow \text{Standard Error} = \frac{78 - 80}{\frac{2.5}{\sqrt{40}}} = -5.05$$

(iv) Conclusions

Decision Rule: If $z = -5.05$ is less than -1.96 or greater $+1.96$, Reject the Null Hypothesis with 95% C.I

Reject the Null Hypothesis { There is some fault in the machine.

(v) A complaint was registered, the boys in a Government School are underfed. Average weight of the boys of age 10 is 32 kgs with S.D = 9kgs. A sample of 25 boys were selected from the Government School and the average weight was found to be 29.5 kgs? With CI: 95% Check it is True or False.

Ans) Conditions for z-test $n=25$ $\mu=32$ $\sigma=9$ $\bar{x}=29.5$

① We Know the population sd. OR

② We do not Know the population sd but our sample is large $n \geq 30$

Conditions For T test

1. We donot know the population std
2. Our sample size is small $n < 30$
3. Sample std is given.

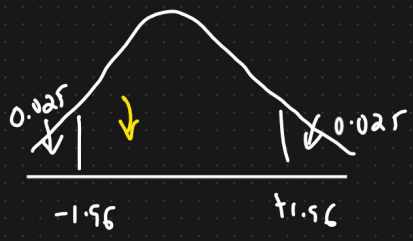
Step 1

1. $H_0 : \mu = 32$

   $H_1 : \mu \neq 32$

2. $C.I = 0.95 \qquad \alpha = 1 - 0.95 = 0.05$

3. Z test



$$Z\text{-score} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{29.5 - 32}{9/\sqrt{25}} = -1.39$$

Conclusion $= -1.39 > -1.96$ Accept the Null Hypothesis 95% C.I

We fail to Reject the Null Hypothesis

The Boys are fed well.

(1) A factory manufactures cars with a warranty of 5 years or more on the engine and transmission. An engineer believes that the engine or transmission will malfunction in less than 5 years. He tests a sample of 40 cars and finds the average time to be 4.8 years with a standard deviation of 0.50. (1) State the null & alternate hypotheses

(2) At a 2% significance level, is there enough evidence to support the idea that the warranty should be revised?

Z-score — = -2.5298

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

$H_1 : \mu < 5$

Step 2: $\alpha = 0.02$ C.I = 0.98

(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% }

Z-score = 14.96

⑥ A factory manufactures cars with a warranty of 5 years (or more) on the engine and transmission. An engineer believes that the engine or transmission will malfunction in less than 5 years. He tests a sample of 40 cars and finds the average time to be 4.8 years with a standard deviation of 0.50. ① State the null & alternate hypotheses

② At a 2% significance level, is there enough evidence to support the idea that the warranty should be revised?

Ans) $n = 40$ $\bar{x} = 4.8$ years $s = 0.50$

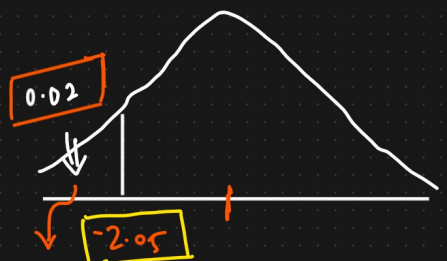
Step 1

$H_0: \mu \geq 5$ {Null Hypothesis}

$H_1: \mu < 5$ {Alternate Hypothesis}

Step 2: $\alpha = 0.02$ $C.I = 1 - 0.02 = 0.98 = 98\%$

Step 3:



Reject the Null Hypothesis

Step 4:

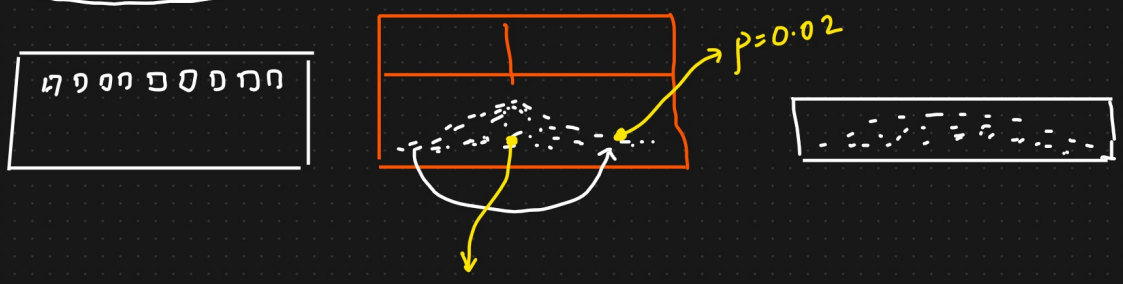
$$\text{Z-score} = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{4.8 - 5}{0.50/\sqrt{40}} = -2.5298$$

Conclusion: $-2.52 < -2.05$ Reject the Null Hypothesis

Warranty needs to be revised.

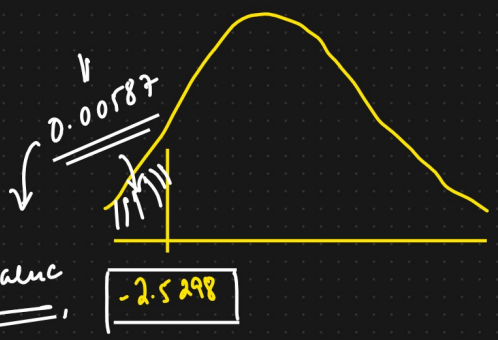
P-value ≠ {Significance value}. → C.I

Out of all 100 random touches



p=0.080

$\alpha = 0.02$



P-value < α [Yes]

Reject the Null Hypothesis

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

(a) Null & Alternate Hypothesis }

(b) 95%. Is there enough evidence to discard the null hypothesis?

Ans) $\bar{x} = 169.5$ $S = 3.9$ $n = 36$ $\mu = 168$ C.I = 0.95

Step 1

H_0 $\mu = 168$

H_1 $\mu \neq 168$.

Step 2 : C.I = 0.95 $\alpha = 0.05$

Step 3

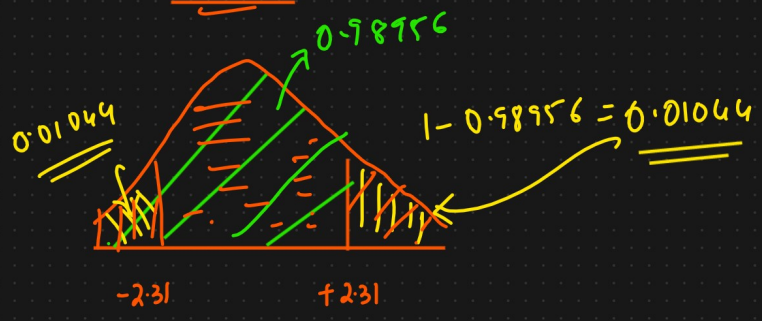
Step 4 : $Z\text{-score} = \frac{169.5 - 168}{3.9/\sqrt{36}} = \approx \boxed{2.31}$ $\Rightarrow 2.3076$

$2.31 > 1.96$ { Reject the Null Hypothesis }

2 tail test

A-value

2 Tail Test



p.value = 0.01044 + 0.01044 = 0.02088

$0.02088 < 0.05$ { Reject the Null Hypothesis }.

(*) 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 1.8years. 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) ① $H_0 : \mu \geq 2$

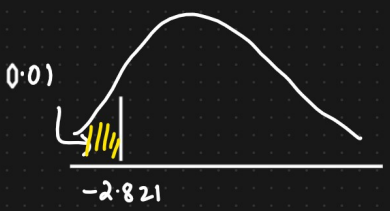
$H_1 : \mu < 2$.

$n < 30$ Sample standard deviation

② Step 2

$C.I = 0.99 \quad \alpha = 0.01$

③ Step 3



Degree of freedom $= n - 1$

$= 10 - 1 = 9$

④ Calculate t test statistics

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

⑤ $-4.216 < -2.82$ { Reject the Null Hypothesis }.

The average life of the battery is less than 2 years.

⑥ Z test with proportions

⑦ 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 find that 130 responded Yes. owning a cell phone?

⑧ State Null And Alternate Hypothesis?

⑨ At a 95% C.I, is there enough evidence to reject the Null Hypothesis?

Ans)

Step 1

Null Hypothesis : $P_0 = 0.70$ ✓

Alternate Hypothesis : $P_0 \neq 0.70$ ✓

$n = 200$ $x = 130$

$\hat{p} = \frac{130}{200} = 0.65$

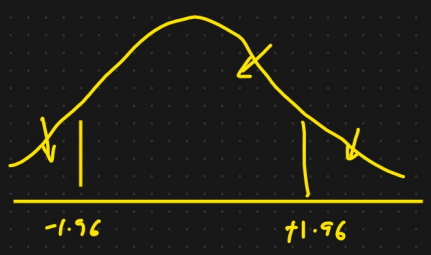
$q_0 = 1 - P_0 = 0.30$

Step 2 : C.I = 0.95 $\alpha = 0.05$

Step 3 :

Step 4 : Z test with propoportion

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

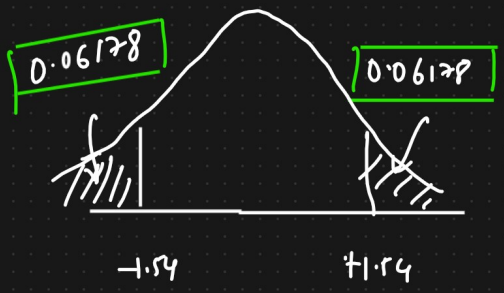


$$= \frac{0.65 - 0.70}{\sqrt{\frac{0.7 \times 0.3}{200}}} \approx -1.54$$

Conclusion

$-1.54 > -1.96$ Fail to Reject the Null Hypothesis

Ratio



P-value = 0.06178 + 0.06178

= 0.12356

Pvalue > Significance value Fail To Reject Null Hypothesis.

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

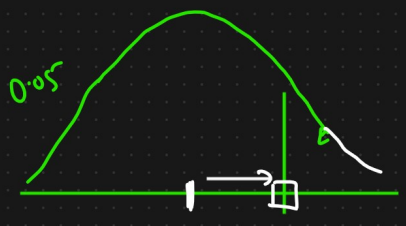
(b) At 10% significance level, is there enough evidence to support the idea that vehicle ownership in city ABC is 60% or less?

Ans) $H_0 : p_0 \leq 0.60$

$H_1 : p_0 > 0.60$

$\hat{p} = \frac{170}{250} = 0.68$

$q_0 = 0.40$



Z-score $= \frac{0.68 - 0.60}{\sqrt{\frac{0.6 \times 0.4}{250}}}$

$= \frac{0.08}{0.0309} = 2.588$

Reject the Null Hypothesis.

(K) Chi Square test

(1) Chi Square test claims about population proportions.

ORDINAL DATA

NOMINAL DATA

It is a non parametric test that is performed on categorical data.

(2) In the 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.

<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

	<18	18-35	>35
Observed	121	288	91
Expected	100	150	250

Step 1: Null Hypothesis H_0: The data meets the expected distribution

H_1: The data does not meet the " "

Step 2: $\alpha = 0.05$ C.I = 95%

Step 3: Degree of freedom {categories}.

$df = C - 1 = 3 - 1 = \boxed{2}$

↳ No. of categories.

$\alpha = 0.05$

Step 4: Decision Banking = $\boxed{5.991}$ {chi square table}

Step 5: Chi square Test Statistics

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

$$\boxed{\chi^2 = 232.494}$$

Conclusion

$\chi^2 > 5.99$ Reject H_0.

⇓ ⇓ ⇓

ANOVA TEST [F Test] → Assignment

3 Distribution

① Binomial
② Bernoulli
③ Poisson Raw of pattern
④ Poisson.