2) Inferential Stats:

- Z-test
- t-test
- ANOVA → f-test
- CHISQUARE
- Hypothesis testing {P-values}
- Confidence interaval.
- 2-table, t-table.

# What is statistics:

↳ Statistics is the science of collecting, organising & analysing the data.

↳ {Better Decision Making}

- Data: facts or prices of information that can be measured.

(1) Descriptive stats:

↳ It consist of Organising & Summarizing data.

(2) Inferential stats:

↳ Techniques where in we use the data that we have measured to form conclusions.

# Sampling Techniques :

- ~~Pollution~~
- population (N)
- Sample (n)

1) Simple random Sampling :

↳ Every member of the population (N) has an equal chance of being selected for sample (n).

2) Stratified Sampling :

↳ Where the population (N) is split into non-overlaping groups (Strata).

E.g : Gender : 1) Male ← Survey
2) Female ←

3) Systematic Sampling :

↳ We pick every $n^{th}$ individual from the poppulation (N).

eg : Sorueg in a Mall
↳ pick every $5^{th}$ person from the row.

4) Convienience Sampling :

↳ For eg : Doing a survey :

↳ pick people who have knowledge of data science.

# DATA

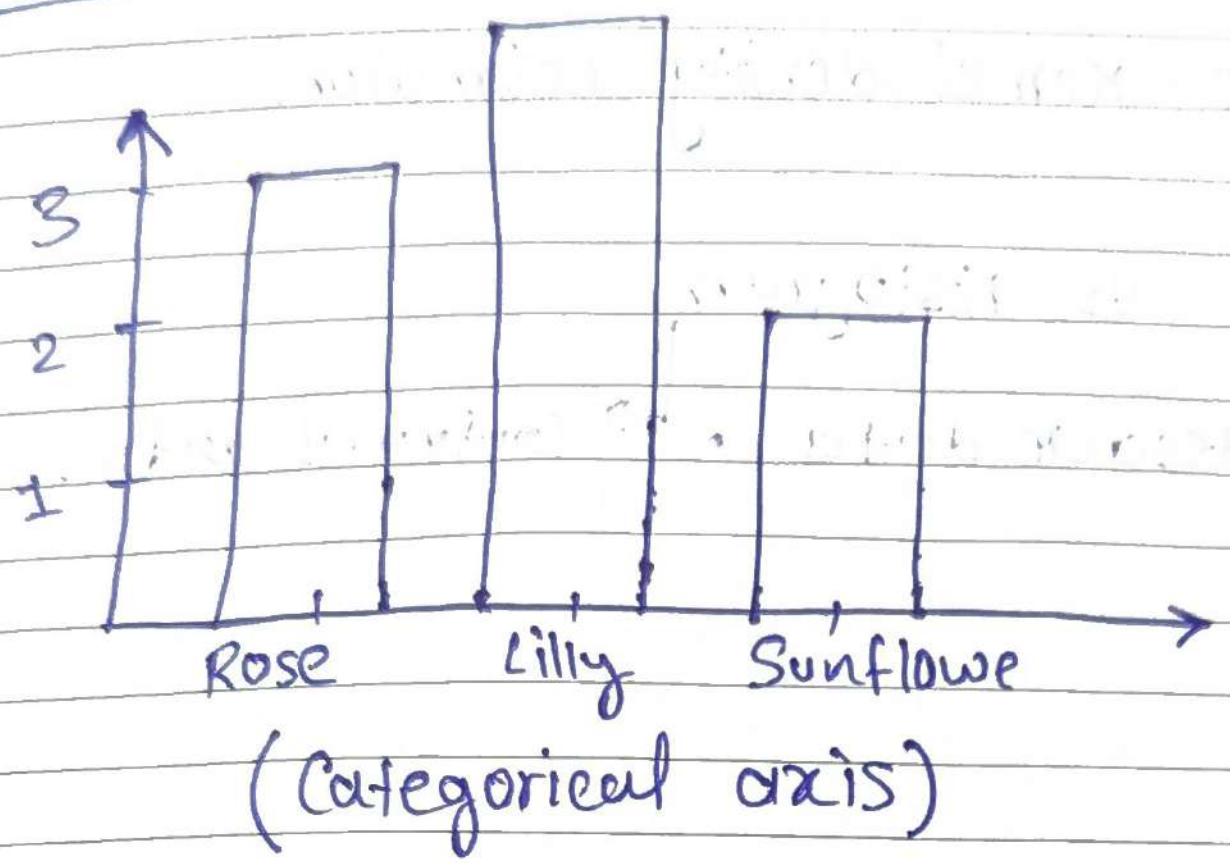
- Numerical
  - Continuous
  - Discrete
- Categorical
  - Ordinal
  - Nominal

# Frequency Distributions :

Sample dataset : Rose, Lilly, Sunflower, Rose, Lilly, Sunflower, Rose, Lily, Lily.

| Flower | Frequency | Cumulative frequency |
|---|---|---|
| Rose | 3 | 3 |
| Lilly | 4 | 7 (3+4) |
| Sunflower | 2 | 9 (7+2) |

## (1) Bar Graph :



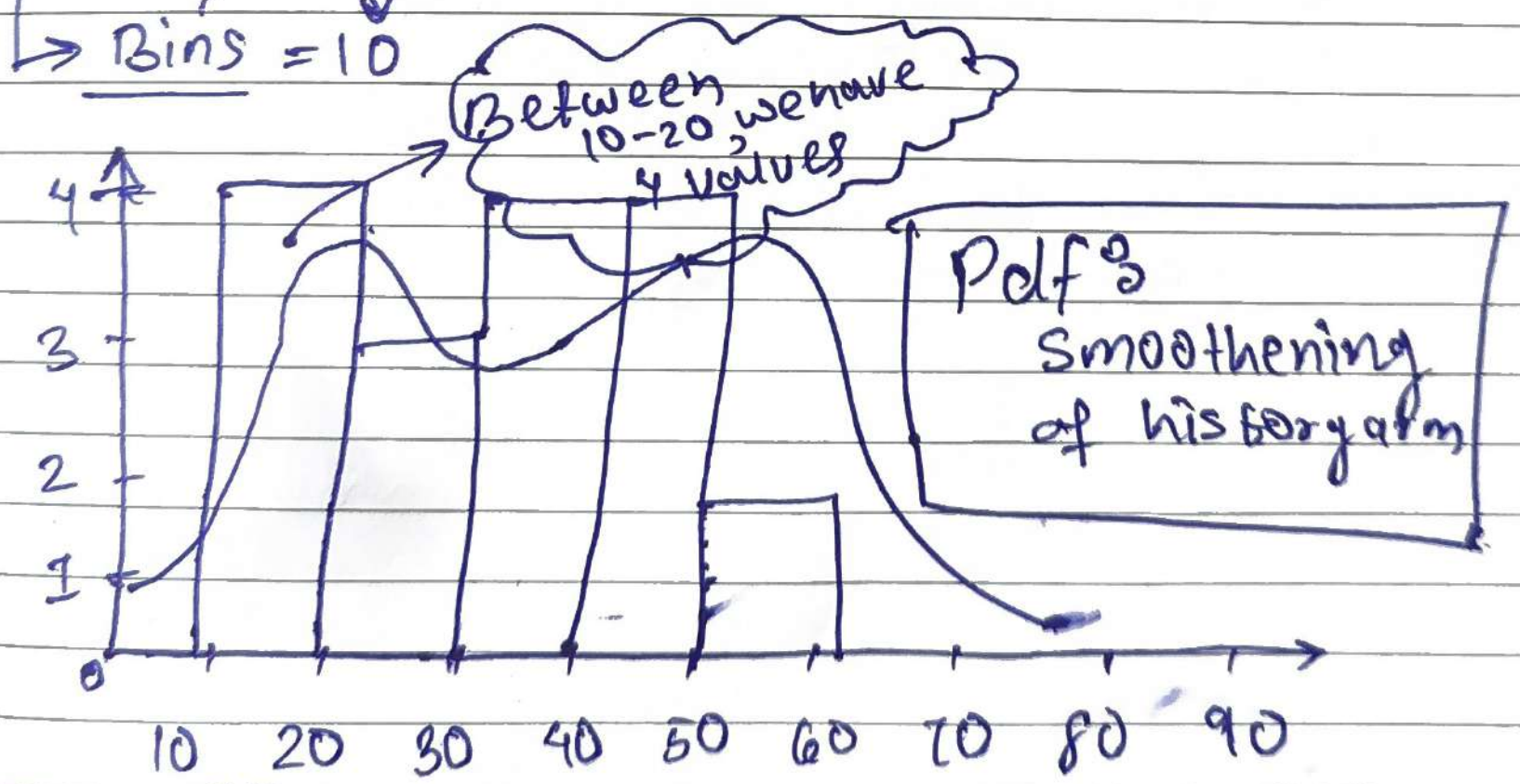

(Categorical axis)

## (2) Histograms :

(continuous data)

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

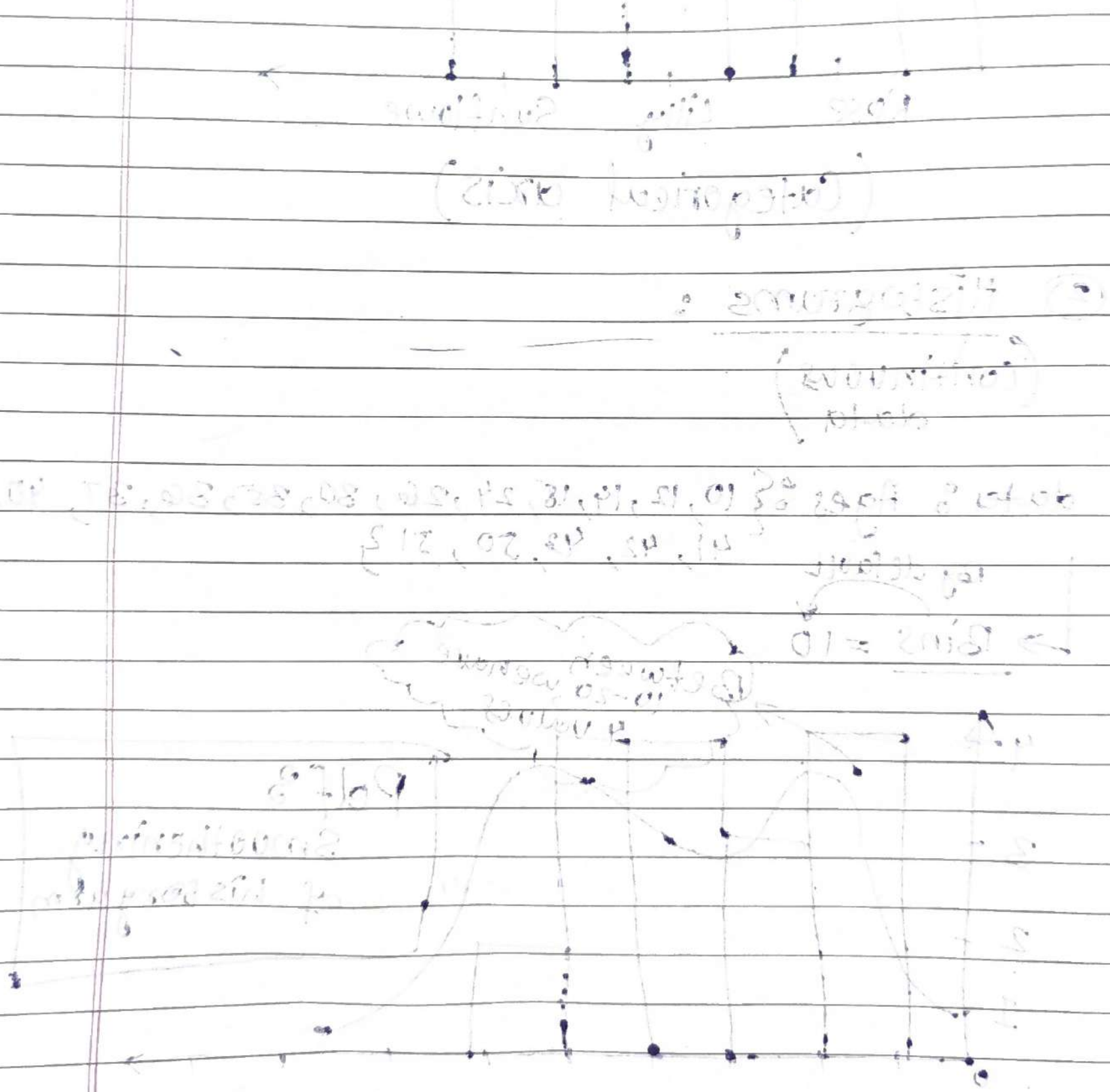
By default → Bins = 10



- Pdf - Probability density function.
- Kde - Kernel density estimator.

Bar vs histogram

- → descrete data
- → continuous data.



# Measure of central tendency :

1) Arithmatic mean for population & sample

Mean ⟶ Average

Population (N) Sample (n)

$X = \{1,1,2,2,3,3,4,5,6,6\}$ $\bar{X} = \sum_{i=1}^{n} \frac{x_i}{n}$

$\mu = \sum_{i=1}^{N} \frac{x_i}{N}$ $= 3.2$

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

$= 3.2$

⟹ Central Tendency

① Mean ② Meelian ③ Mode.

↳ Refers to the measure used to determine the centre of the distribution of data.

Data : {1,1,2,2,3,3,4,5,5,6}

Mean : 3.2

new Data : {11, 100}

Mean : $\frac{32+100}{11} = 12$ //

Old Mean : 3.2

~~old~~

New Mean : 12

- By just adding a new data point which is far away from the other datap.
- we called it outliers :
- worst impact on the distribution.

Now, let's take median.

⇒ Median :

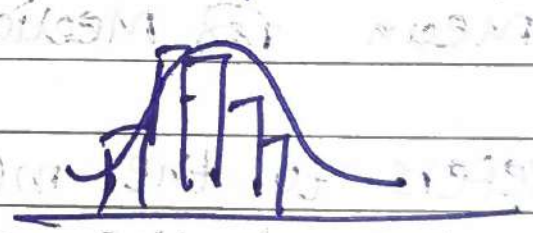
↳ Sort the numbers in accending

★ ↳ {1,1,2,2,3,3,4,5,5,6,100}

Total number of elements or datapoints : (11) odd

middle element : 3

Distribution :



↳ if, even then take the mean of the two middle datapoints.

So, our median is : 3

- It means if we have outliers in our data, then mean will fluctuate very much so, we consider Median, it handles it well.

=> Mode : Most frequent element.

Data : {1, 2, 3, 3, 4, 6, 6, 6}

Mode : 3

Outlier data : {1, 2, 3, 3, 6, 6, 6, 100, 100, 100, 100}

Mode : 100

So, everytime when outliers are there we will use Median as a central tendency.

# Mode is used when we have missing values we can replace it with mode.

↳ Categorical data.

# Measure of Dispersion :
↳ Spread

(1) Variance

(2) Standard Deviation

# What is Spread ?

data-1 : $\{1, 1, 2, 2, 4\}$, $\mu_1 = 2$

data-2 : $\{2, 2, 2, 2, 2\}$, $\mu_2 = 2$

Both the mean are same, but data points are different.

Question : if the means are same then what makes the dist different.

1] Variance

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

For population

For Sample :

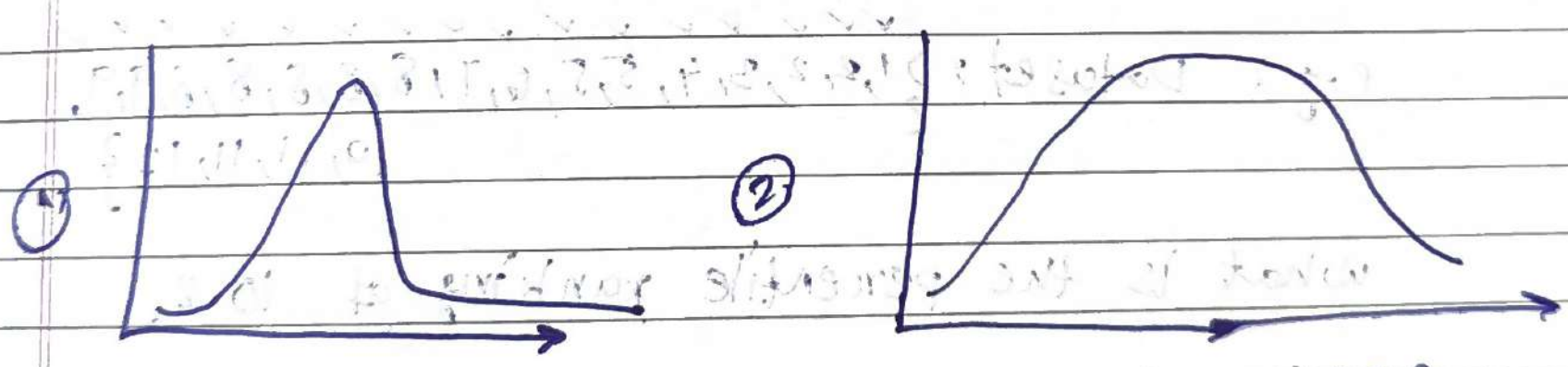
$$S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

- The difference is in denominator,

Population : $N$

Sample : $n-1$

Why ?




Less variance

more variance
high variance

(2) Standard Deviation

$$\sigma = \sqrt{var}$$

# Percentiles & Quartiles {Find Outliers}

Data: 1, 2, 3, 4, 5

% of the odd numbers?

$$\% \text{ of odd} = \frac{\text{Amount of odd no.}}{\text{Total no. of odd}}$$

$$= \frac{3}{5} = 0.6 = 60\%$$

- **Percentiles:**

↳ A percentile is a value below which a certain percentage of observation lie.

eg: Dataset: {1, 2, 2, 3, 4, 5, 5, 6, 7, 8, 8, 8, 8, 8, 9, 9, 10, 11, 11, 12}.

What is the percentile ranking of 10?

$$\text{Percentile Rank of } x = \frac{\text{no. of values below } x}{n} \times 100$$

$$= \frac{16}{20} \times 100 = \boxed{80\%\text{ile}}$$

★ it means 80% of the entire population lies below 10.

- What value exists at percentile ranking of 25% ?

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

$$= \frac{25}{100} \times 21 = 5.25$$

(25% → Index position)

↳ So, 5.25 lies b/w 5, 5 in the dataset.
taking avg $\frac{5+5}{2} = \boxed{5}$ → 25%

↳ 75% ?.

$$= \frac{75}{100} \times 21 = 15.75^{th} \text{ position.}$$

↳ [9] → 75%

# Five Number Summary :

1. Minimum
2. First quartile ($Q_1$)
3. Median
4. Third quartile ($Q_3$)
5. Maximum

⇒ Removing the outlier

{ 1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 27 }

outlier ← [ lower fence ⟷ higher fence ] → outlier

← [ lower bound ⟷ Upper Bound ] →

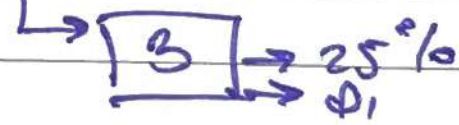
lower fence = ~~1.5 IQR - Q1~~

$= Q1 - 1.5 \times (IQR)$

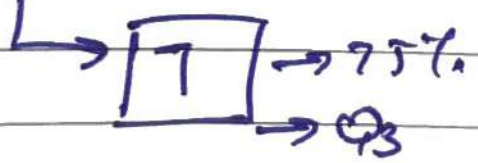
upper fence = $Q_3 = 1.5 \times (IQR)$

⇒ IQR = Inter Quartile Range. $= Q_3 - Q_1$

$Q_1 = 25\% = \frac{25}{100} \times (19+1) =$ 5th elm




$Q_3 = 75\% = 15^{th}$ elm




IQR = 7 - 3 = 4 //

lower fence = ~~1.5(Q~~

= Q1 - 1.5 (IQR)

= 3 - 1.5(4)

= -3

upper fence = Q3 + 1.5 (IQR)

= 7 + 1.5(4)

= 13

any value < -3
(outlier)

any value > 13
(outlier)

outlier we got : 27

remove : 27

Remaining Data:

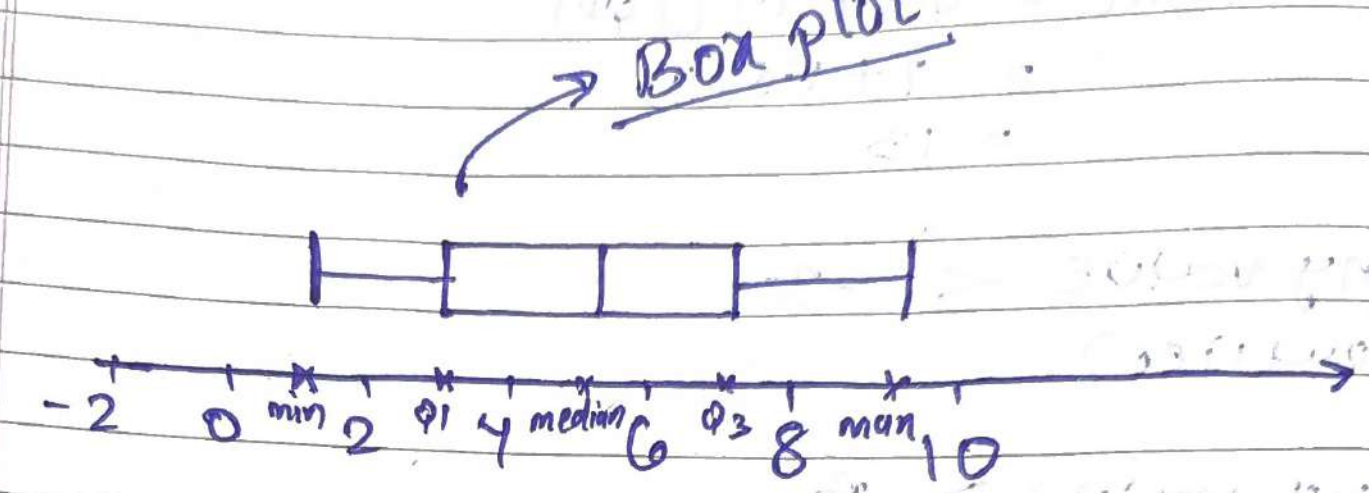
1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9

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

} Five number Sumary

Now,

∴ Drawing Boxplot



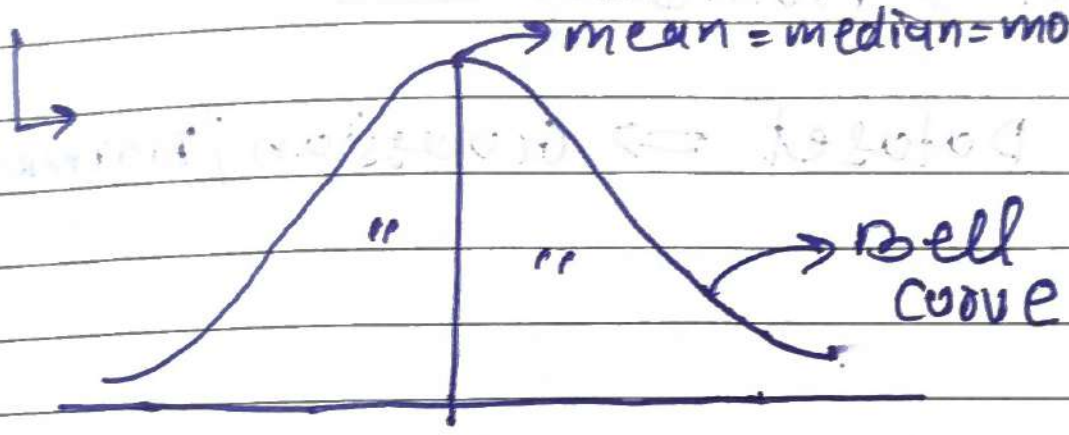

∴ ⇒ Boxplot

↳ can be used to determine outlier.

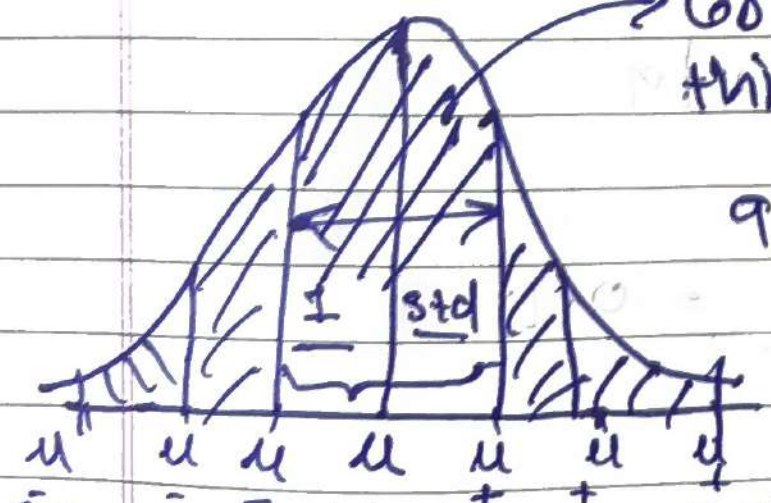
## Distributions :-

- Normal
- Standard normal
  - Z-score
  - log normal
  - Bernoullis
  - Binomial.

## (1) Gaussian / Normal Distribution




- Symmetrical
- Mean = median = mode
- Best or perfect for ML models
- we can derive very important things from this distribution.




→ 68% of datapoints comes in this region (1-std)

95% of entire data lies btw 2-std.

99.7% of entire data lies btw 3-std.

=> Empirical formula :

68 - 95 - 99.7 % Rule

For eg: Height => Normally distributed
↳ Domain Expertise

2) Weight => Normally ——

3) IRIS Dataset => Gaussian / Normal

# Z-score

↳ How much std away it is from the mean.

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

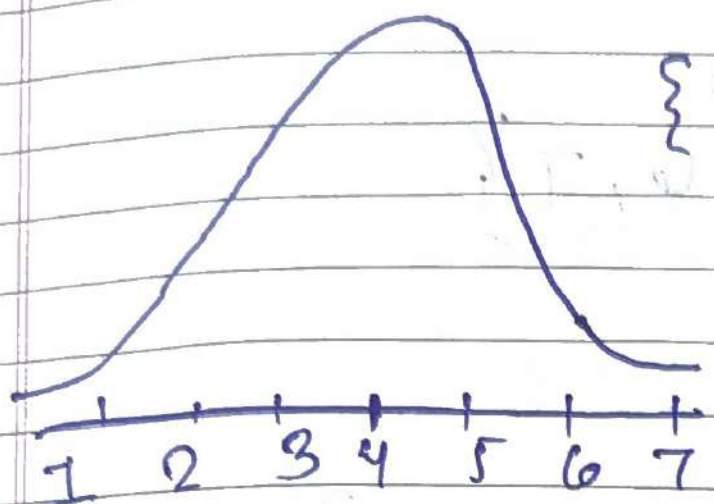
$\mu = 4$
$x_i = 4.75$ } for this data
$\sigma = 1$

$$Z\text{-score} = \frac{4.75 - 4}{1} = 0.75$$

4.75 lies int 0.75 std to the right.

Page No.
Date

Now,




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

$\mu = 4$

$\sigma = 1$

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

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

$$z(1) = \frac{1-4}{1} = -3 \qquad \{-3,-2,-1,0,1,2,3\}$$

$$z(2) = \frac{2-4}{1} = -2$$

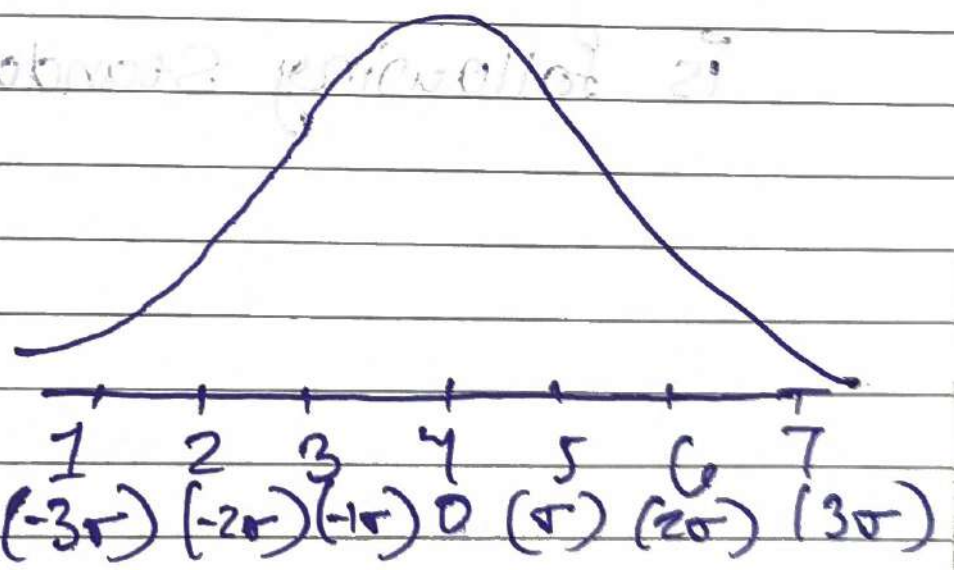
$$z(3) = \frac{3-4}{1} = -1$$

$z(4) = 0$

$z(5) = 1$

$z(6) = 2$

$z(7) = 3$



Initial data

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

⇓ Z-score

{-3, -2, -1, 0, 1, 2, 3}

★ This distribution is called

[STANDARD NORMAL DISTRIBUTION]

any dat satisfying ($\mu = 0$ & $\sigma = 1$)

is following Standard Normal Distribution.

Now,
Practical Application :

Dataset :

| Age | Salary | weight |
|---|---|---|
| 24 | 40K | 70 kg |
| 25 | 80K | 80 kg |
| 26 | 60K | 55 kg |
| 27 | 70K | 45 kg |

→ Z-score

↳ Standard Normal distribution.

This process is called

* **Standardization.**

↳ convers distribution into standard normal distribution
and the parameter is $\{\mu = 0 \& \sigma = 1\}$

# Normalisation :

↳ (0 to 1)

we range our datapoint between 0 to 1.

⇒ MinMax Scaler :

we can range in any lower & upper value like from (-1 to +1)
(0 to 1)
(-3 to +3)

# Probability :

↳ is the measure of the likelyhood of an event.

Eg: ~~Flipping~~ Rolling a Dice $\{1,2,3,4,5,6\}$

$$P(6) = \frac{1}{6} = \frac{\text{number of ways an even can occur}}{\text{number possible outcomes.}}$$

# Addition Rule :

↳ (Probability; "or")

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

Eg: Rolling dice : $\{1, 2, 3, 4, 5, 6\}$

we can get only one output after rolling only single event is occuring at the same time.

Non mutual exclusive

eg : Deck of cards {Queen, ♡}

Page No.
Date

Q] Probability of A or B / head or tail in tossing a coin.

⇒ mutually exclusive

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

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

⇒ Roll a Dice

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

$= \frac{1}{6} + \frac{1}{6} + \frac{1}{6}$

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

# For Non-Mutual Exclusive

Q] You are picking a ~~deck~~ card from decks
Prob of queen or a heart.

⇒ Non-mutually exclusive

$P(\spadesuit) = \frac{4}{52}$

$P(\heartsuit) = \frac{13}{52}$

$P(\spadesuit \text{ and } \heartsuit) = \frac{1}{52}$

Addition Rule for Non-mutual exclusive events.

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

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

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

$P(\spadesuit \text{ or } \heartsuit) =$ 

## (3) Multiplication Rule :

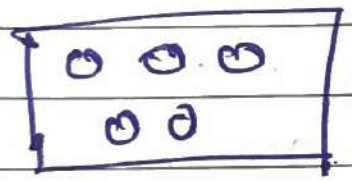
### 1) Independent events

Eg : Rolling a dice : {1, 2, 3, 4, 5, 6}.

1 is not depend on 2
any event is not depend on any other event.

### 2) Dependant Event

eg :
we have a bag with 3 red balls & 2 green.



$P(red) = \frac{3}{5}$

picking a red ball.

eg : but after picking red ball the total number of red balls becom 2 and total balls become 4.

Now, $P(Green) = \frac{2}{4} = \frac{1}{2}$

- 2nd green event is depend on the 1st event

Q] Rolling a dice, 1st event got "5" 2nd event got "4"

Probability of 5 & 4 ? (Independent event)

P(5 and 4) = P(5) x P(4)

Multiplication Rule.

$P(5 \text{ and } 4) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$ //

Q] What is the probability of drawing a queen and then a Aces from a deck of Cards?

(Dependent event)

$P(A \text{ and } B) = P(A) \times P(B/A)$ → Naive Bayes

Conditional probability

$P(B/A)$ → Probability of B after A event occured

$P(Q \text{ and } A) = P(Q) \times P(Q/A)$

$= \frac{4}{52} \times \frac{4}{51}$ //

## (3) Multiplication Rule :

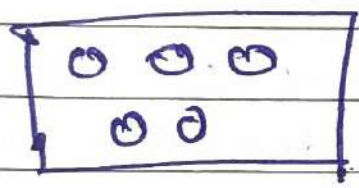
### 1) Independent events

$\xi$ : Rolling a dice : $\{1, 2, 3, 4, 5, 6\}$.

1 is not depend on 2
any event is not depend on any other event.

### 2) Dependant Event

eg :
we have a bag with 3 red balls & 2 green.



$P(\text{red}) = \frac{3}{5}$
↳ picking a red ball.

eg : but after picking red ball the total number of red balls becom 2 and total balls become 4.

Now, $P(\text{GREEN}) = \frac{2}{4} = \frac{1}{2}$

• 2nd event is depend on the 1st event

Q] Rolling a dice, 1st event got "5" 2nd event got "4"

Probability of 5 & 4 ? (Independent event)

P(5 and 4) = P(5) x P(4).

Multiplication Rule.

$P(5 \text{ and } 4) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$ //

Q] What is the probability of drawing a queen and then a Aces from a deck of Cards?

(Dependent event)

$P(A \text{ and } B) = P(A) \times P(B/A)$ → Naive Bayes

→ Conditional probability

$P(B/A)$ → Probability of B after A event occured

$P(Q \text{ and } A) = P(Q) \times P(Q/A)$

$= \frac{4}{52} \times \frac{4}{51}$ //

# Permutation & Combination

## 1) Permutation:

School trip ⟹ (chocolate factory)

⟶ Dairy milk, 5-star, Milky bar, Eclairs, Gems, Silk.

Student ⟷ Assignment.

ask student to write 3 names of chocolates they like.

5-1 :

(1) (2) (3)

How many options this student have.

(1) at first place student have total 6 choices

6

(2) Now 5 choices left and so on.

6 5 4

So, Total permutation:

$6 \times 5 \times 4 = 120$ //

Permutation Formula : ${}^{n}P_{r} = \frac{n!}{(n-r)!}$

(n → Total no. of choice; r → How many no. asked.)

$= \frac{6!}{(6-3)!} = \frac{6 \times 5 \times 4 \times 3!}{3!}$

$= 120$ //

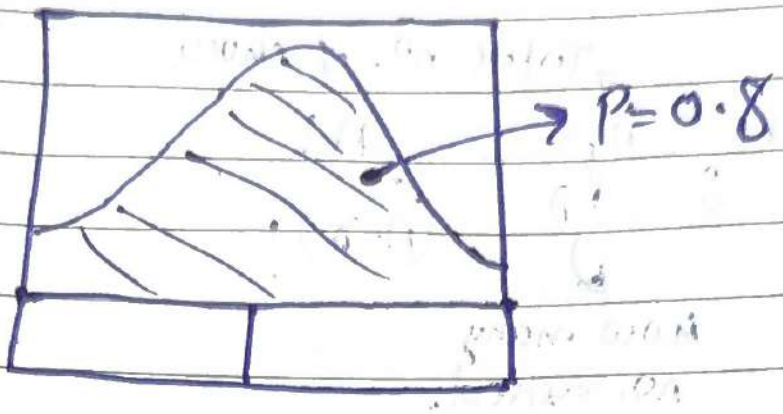
=> Combination :

${}^{n}C_{r} = \frac{n!}{r!(n-r)!} = \frac{6!}{3!(3!)} = \frac{6 \times 5 \times 4 \times 3!}{3!(3 \times 2 \times 1)}$

$= 20$ //

# P-value

## Mouse Pad:




The shaded region is the region where we use the mousepad most

$P = 0.8$

↳ means every 100 times I touch the mouse pad I will touch here 80 time.

Coin ⟹ whether it is fair or not by performing 100 tosses.

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

↳ for a fair coin.

# Inferential STATS

# Hypothesis Testing

(1) Null Hypothesis : Coin is fair. (default).

(2) Alternat hypothesis : Coin is unfair (opposite of Null hypothesis).

(3) Experiment.

(4) Reject or accept Null Hypothesis.

So, after deciding Null & Alternative hypo — we will perform experiment

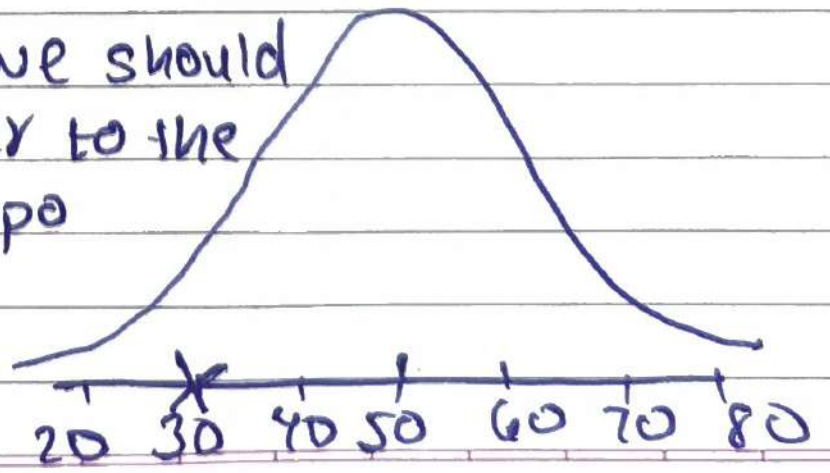
we get result : 30 times head
70 times tail

Concusion ?

# ↳ Significance value ($\alpha$)

$\alpha = 0.05$ → considered maximum.

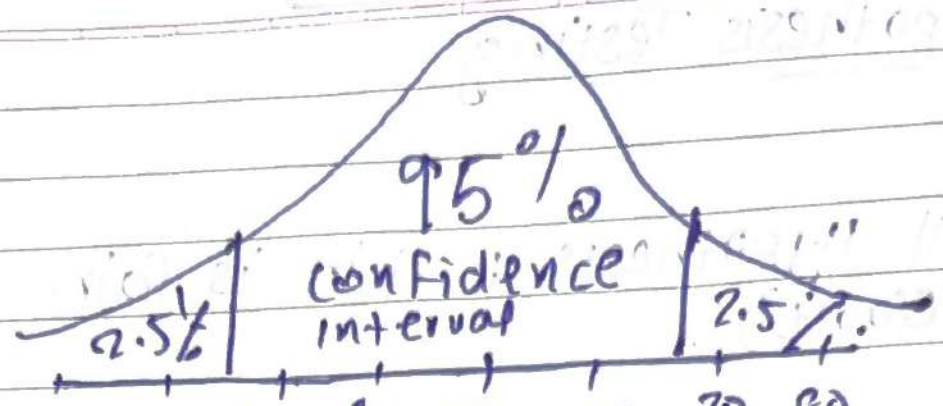
To make concusion we should have the result nearer to the meen, to proof null hypo correct.




How much = significance near to mean? value($\alpha$)

$\alpha = 0.05$

5%.




C·I (confidence interval) $= 100 - \alpha$

$= 95\%$

Now,

30-H } if the experiment btw 30 & 70
70-T } inside the C·I then its fair.

→ it's inside the C·I

[coin is fair]

Null-hypothesis

- we accept null hypothesis

if it's 10-Head

- we reject null hypothesis.

# Type-1 & Type-2 Errors

→ Null hypothesis ($H_0$) ⇒ coin is fair
→ Alternate hypothesis ($H_1$) ⇒ not fair.

Reality check:
- → Null hypothesis is True
- → Null hypothesis is False

Outcome:

1] We reject the null hypothesis, when in reality it is false:

→ So, it's a good thing isn't it?
yes.

[This is good one]

Outcome:

2] We reject Null ($H_0$) in reality its true:

→ So, its bad thing, because WE said ($H_0$) coin is unfair, but it's fair

[this is Type-1 error]

Outcome-3 : We accept (Ho)
(coin is fair)
but in reality its false
↳ error!

[Type-II error]

eg: email spam ---

eg: Death sentence ---

Outcome: 4 : We accept Ho
(coin is fair)
but in reality True
↳ good case.

# Confusion Matrix :

| | P | N |
|---|---|---|
| T | TP | TN |
| F | FP | FN |

T = True
F = False
P = positive
N = Negative.

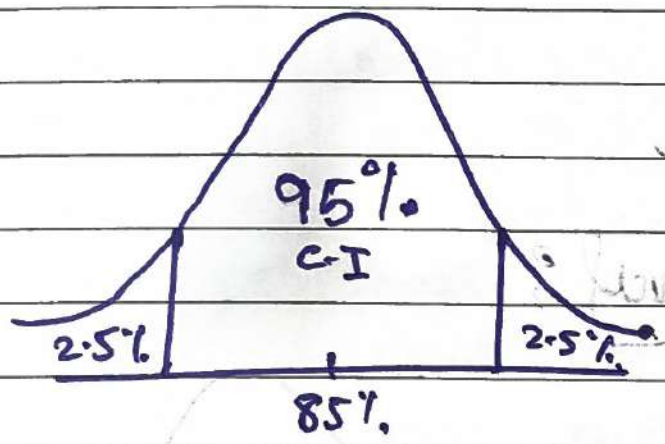
we said Fair, actual it's unfair. (Type-II error)

we said unfair, actual it's fair

# 1-tail & 2-tail tests

E.g: Colleges in Karnataka have 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 4%. Does this college has a different placement rate? ($\alpha = 0.05$)

2-tailed test :

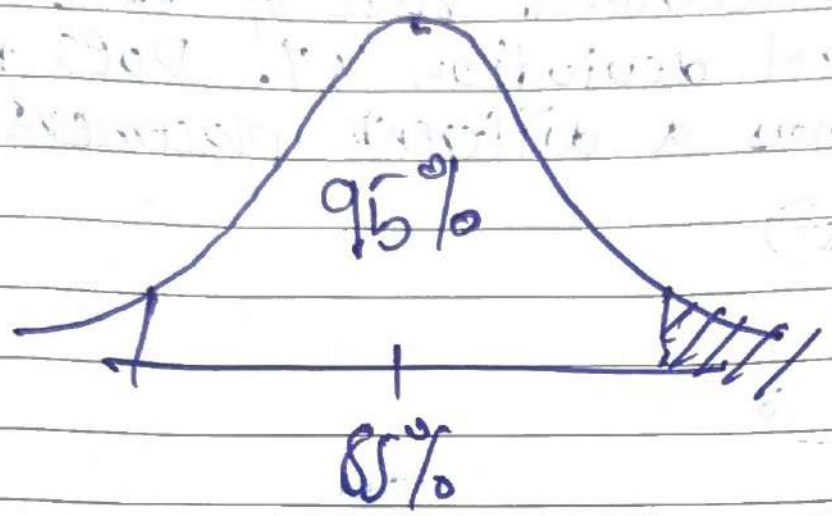



We have to check whether 88% is lying on the graph.
It may lie below and above 85%
That's why it is two-tailed test

let's change the question:

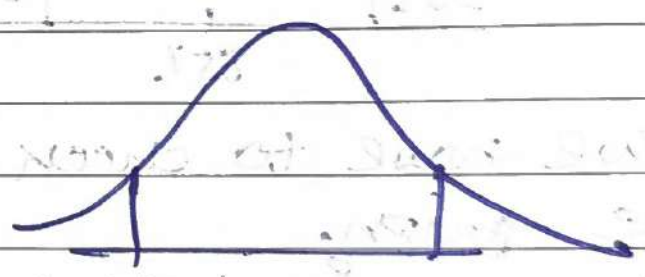
Does this college have a placement rate greater than 85%?

we have to check only above (greater)




1-tailed-test

# Confidence Interval:



⇒ point estimate

The value of any statistics that estimates the value of a ~~placement~~ parameter.

$\alpha = 0.05$

Inferential stats

↳ based on sample data we take decision or estimate about the population.

Sample → population

Mean $\bar{x}$ → 2.9

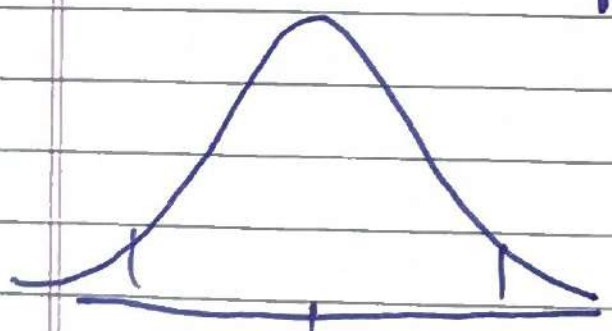
Mean $\mu$ → 3

⇒ C.I

Point estimate ± Margin of error

$\sigma = 100$
$n = 25$
$\alpha = 0.05$
$\bar{x} = 520$

e.g.: On the quant of CAT exam the std is known to be 100. A sample of 25 test takers has a mean of 520 score. Construct a 95% CI about the mean.




(1) whenever population std is given, we apply a z test.

z test → score

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

↳ Standard error.

upper bound → $\bar{x} + Z_{\frac{0.05}{2}} \frac{100}{\sqrt{25}}$

lower bound → $\bar{x} - Z_{\frac{0.05}{2}} \frac{100}{\sqrt{25}}$

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

↳ check value on Z-table.

$1 - 0.025$

$= 0.975$

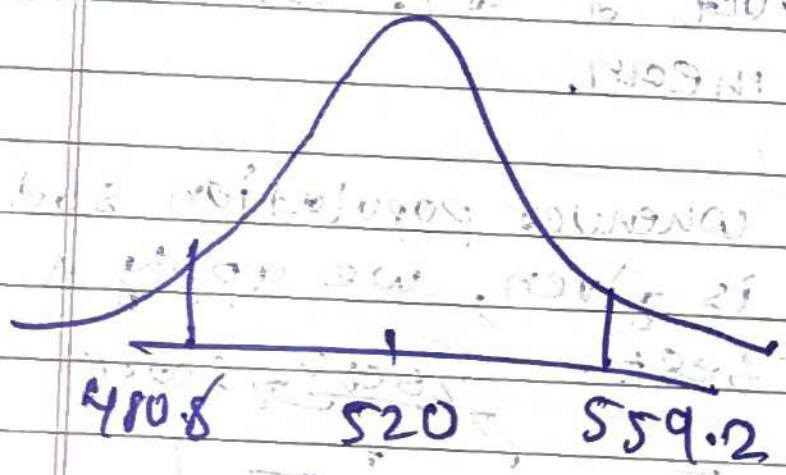
↳ Z-table

1.96

Upper $= 520 + 1.96(20)$

$= 559.2$

lower $= 520 - 1.96(20) = 480.8$



9] On the quant ~~that~~ test of CAT exam, a sample of 25 test taken has a mean of 520 with a std of 80. Construct 95% confidence interval about the mean.

⇒ No population standard dev given

↳ Here, we use t-test.

Point estimate ± Margin of error

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

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

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

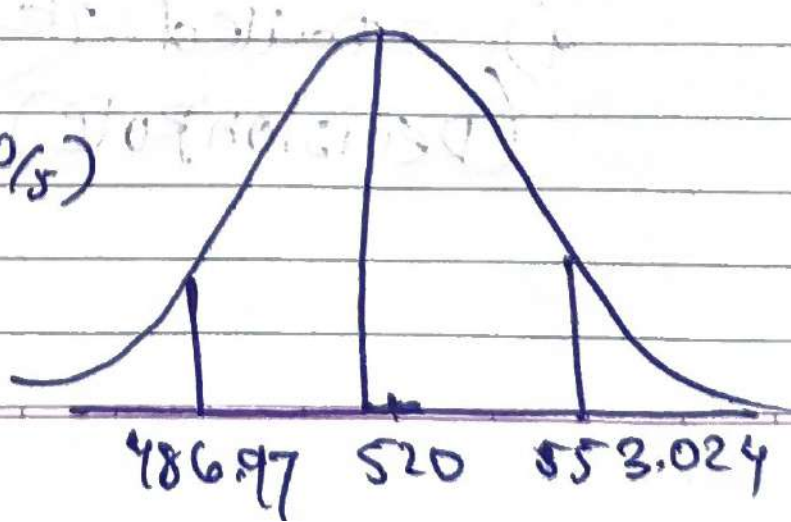
t-table ⇒ $t_{\frac{0.05}{2}} = 2.064$ ↳ from t-table.

Upper = $520 + 2.064\,(80/5)$

$= 553.024$

Lower = $520 - 2.064\,(80/5)$

$= 486.97$



(1) One Sample (z-test)

(1) population std is given

(2) Sample size $n \geq 30$

* In the population, the average IQ is 100 with a S.d of 15. Researchers wants to test a new medication to see if there is positive or negative effect on intelligence, or no effect at all. A sample of 30 participants who have taken the medication has a mean of 140. Did the medication affects the intelligence. ($\alpha = 0.05$). C.I = 95%

⟹ 1) Define NOH ($H_0$)

↳ $\boxed{\mu = 100 = H_0}$

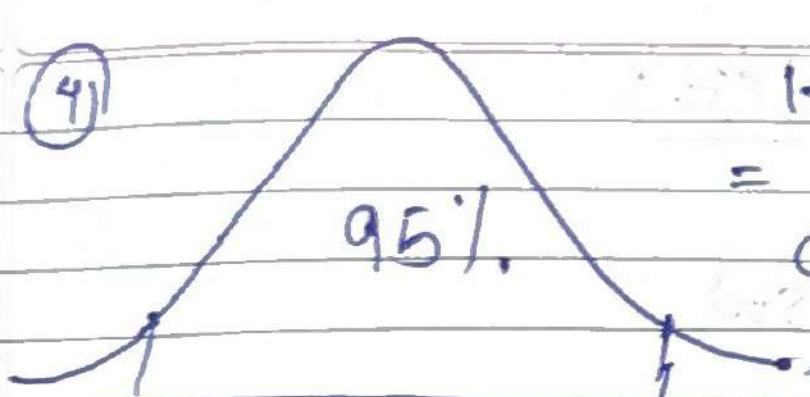
2) ($H_1$)

↳ $\boxed{\mu \neq 100 = H_1}$

3) $\alpha = 0.05$

4) 2 Tailed Test (Decision Role)

(4)




$1 - 0.025$

$= 0.975$

↳ z-table = 1.96

(5) Calculate Test Statistics

$$Z = \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)}$$ → Real formula

Standard error

[before, $n = 1$, so, $\frac{\sigma}{1} = \sigma$]

$$= \frac{140 - 100}{\frac{15}{\sqrt{30}}} = \frac{40}{15} \times \sqrt{30}$$

$= 14.60$

State our Decision,

$14.60 > 1.96$

if Z is less than $-1.96$ or greater than $1.96$, we reject null hypothesis.

Did the medication improved the intelligence?

↳ Yes!

## (2) One Sample t-test :

Z-test ⇒ popⁿ std.

t-test ⇒ unknown popⁿ std.

g] Population avg IQ = 100

n = 300 S = 20

$\bar{x}$ = 140

did the medical effect the intelligence?

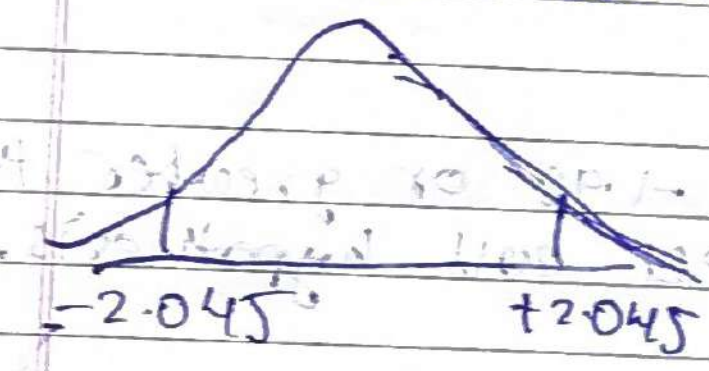
⇒ ① $H_0$ : $\mu = 100$

② $H_1$ : $\mu \neq 100$

③ degree of freedom

$n - 1 = 30 - 1 = 29$

④ state decision



(5) T-test

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = 10.96_{11}$$

$t = 10.96 > 2.045$

↳ we reject $H_0$

P ≤ significance value

$H_0$ X

$H_1$ ✓

# CHI SQUARE TEST :

↳ Chi square test claims that about population proportions It is a non parametric test that is performed on categorical (nominal or ordinal) data.

9] In the 2000 Indian Census, the age of the individual in a small town were found to be the following!

| less than 18 | 18-35 | >35 |
|---|---|---|
| 20% | 30% | 50% |

?.

In 2010, age 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?

⇒ What is non-parametric test?

↳ A statistical that does not rely on any assumptions about the population distribution,

use's

↳ when data is not normally distributed
when small sample size.
ordinal & nominal data.

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

| <18 | 18-35 | >35 | n=500 |
|---|---|---|---|
| 121 | 288 | 91 | observed |
| 500x0.2 | 500x0.3 | 500 x 05 | expected |

| <18 | 18-35 | >35 | |
|---|---|---|---|
| 121 | 288 | 91 | observed |
| 100 | 150 | 250 | expected |

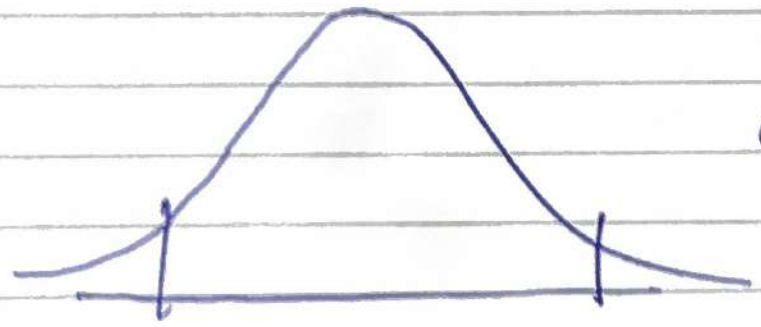
(1) $H_0$ = The data meets the distribution 2000 census

(2) $H_1$ = doesn't meet .....

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

(3) Degree of freedom $= n-1 = 3-1 = 2$

(4) Decision Boundary



Chisquare = $\chi^2$

2tailed test

if $\chi^2$ is greater than 5.99 reject $H_0$.

(5) Calculate Test Statistics.

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

$f_o$ = observed.
$f_e$ = expected.

$$= \frac{(121-100)^2}{100} + \frac{(288-150)^2}{150}$$
$$+ \frac{(914-250)^2}{250}.$$

$$= 232.94$$

$\chi^2 = 232.94 > 5.99$

{Reject the null ($H_0$)

## (2) Covariance :

P-value < significance value

[ $H_0$ X ]

| [X] | [Y] |
|---|---|
| Weight | Height |
| 50 | 160 |
| 60 | 170 |
| 70 | 180 |
| 75 | 180 |

[ $x \propto y$ ]

↳ According to above.

| No. of hours Study | playing |
|---|---|
| 2 | 0 |
| 3 | 4 |
| 4 | 3 |

$$x \propto \frac{1}{y}$$

Covariance

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

Sample

Page No.
Date

$Cov(x,y) \Rightarrow +ve$

↓

$+ve \rightarrow \boxed{x \propto y}$

directly proportional.

$Cov(x,y)$

↓

$-ve \rightarrow \boxed{x \propto \frac{1}{y}}$

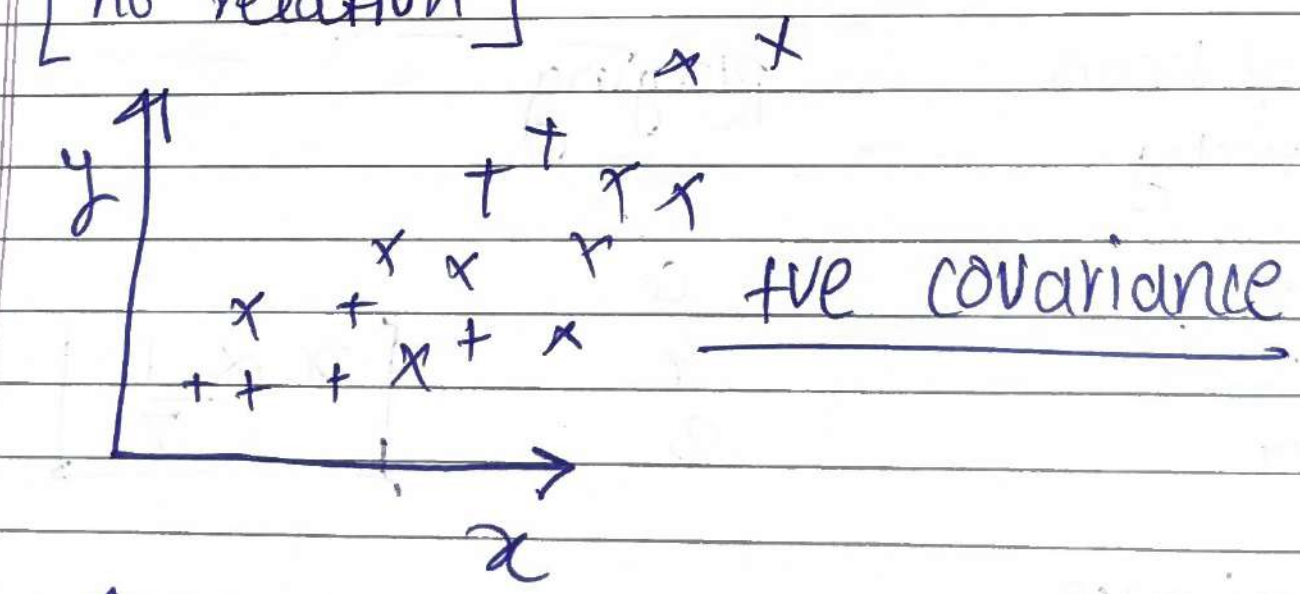
$Cov(x,y)$

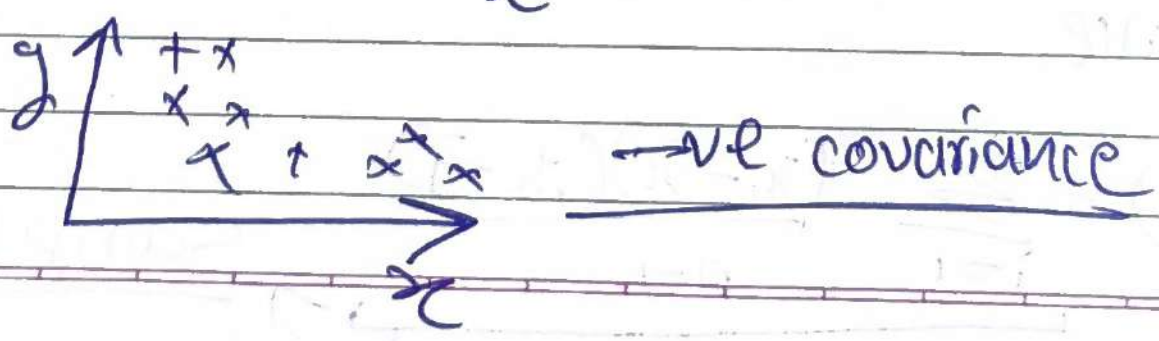
↓

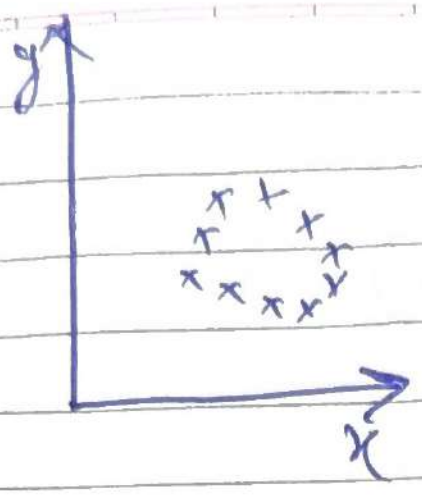
0

→ if $x$ is increasing no effect on $y$.

[no relation]






Covariance = 0

No relation

Disadvantage of Covariance

(1) There is no fixed range
no limit to magnitude
such as : (-200, +10, +1000)

we have to restrict this value within some range.

(2) Pearson Correlation :

↳ (-1 to +1)

more negative side (more -ve correlation)

more on +1 side (more +ve correlation)

- is good at satisfying linear properties
- for linear Spearman correlation.

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

## (3) Spearman Correlation :

$$\text{Spear}(x,y) = \frac{\text{Cov}(R(x), R(y))}{R_x \times R_y}$$

(Rank)

- For Non-linear Data.

What is Rank?

↳

| Weight (y) | ~~Age~~ Height (x) | R(x) | R(y) |
|---|---|---|---|
| 75 | 170 | 2 | 2 |
| 62 | 160 | 3 | 3 |
| 60 | 150 | 4 | 4 |
| 55 | 145 | 5 | 5 |
| ~~40~~ | ~~130~~ | | |
| 85 | 180 | 1 | 1 |

$P \leq 0.05 \longrightarrow$ Reject the $H_0$

# P-value & Significance

Q] avg weight of all residence in a city is 168 pounds with a std 3.9. we take a sample of 36 individuals and the the mean is 169.5 pounds.

C.I = 95%.

↳ Z-test

⇒ $\mu = 168$

$\sigma = 3.9$

$\bar{x} = 169.5$

$n = 36$

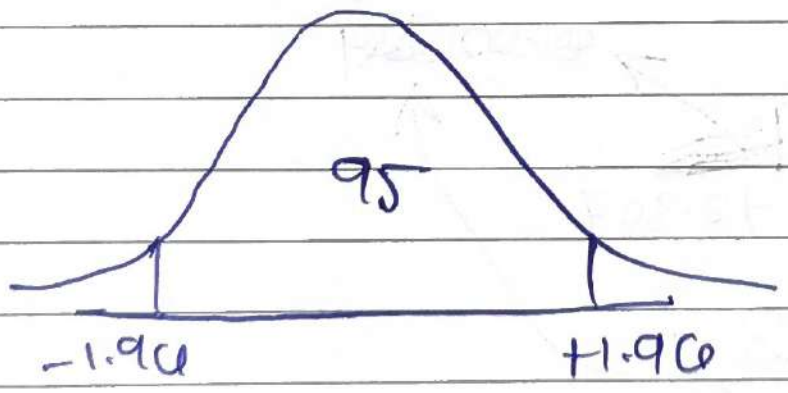
$\alpha = 0.05$

(1) $H_0 = \mu = 168$

(2) $H_1 = \mu \neq 168$

(2) $\alpha = 0.05$

(3) Decision



Page No.
Date

(3) Z-test:

$$Z = \frac{x - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{169.5 - 168}{\frac{3.9}{\sqrt{36}}}$$

$$= \frac{1.5}{3.9} \times 6 = 2.307$$

$Z = 2.307 > 1.96$

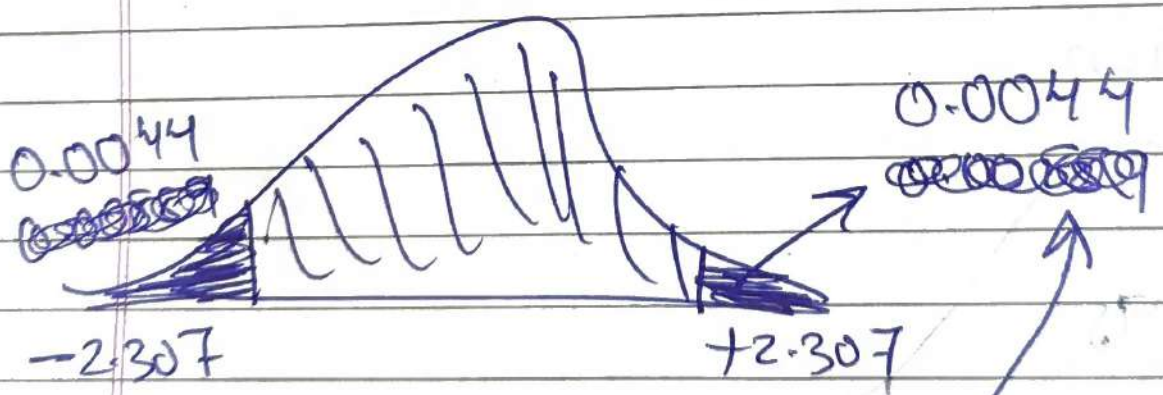
↳ True

↳ reject $H_0$.

2.307 ⇒ z-table ⇒ 0.99111

↓

Area under the curve




$+0.99111 = 63/2$

P-value = 0.0044 + 0.0044

= 0.0088

$0.0088 < 0.05$

↳ $\times H_0$ ✓

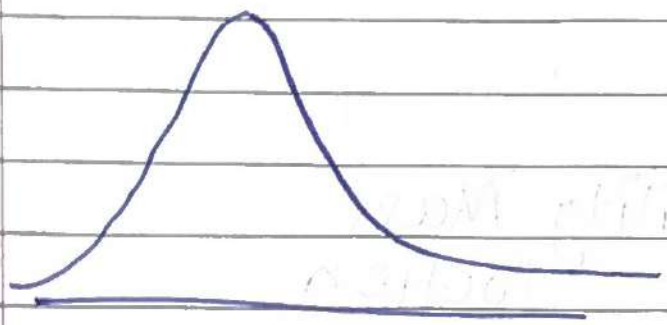
# P-value > significance

↳ ✗ $H_0$

P-value < significance

↳ ✓ $H_0$

# log Normal distribution.



{ y ≈ lognormal Dist. }

⇓

log(y) ⇒ Normal dist

if y belongs ⇔ lognormal dist

it will satisfy log(y) ⇒ will converted into normal distribution.

# Bernoullis Dist :

↳ 2 outcomes

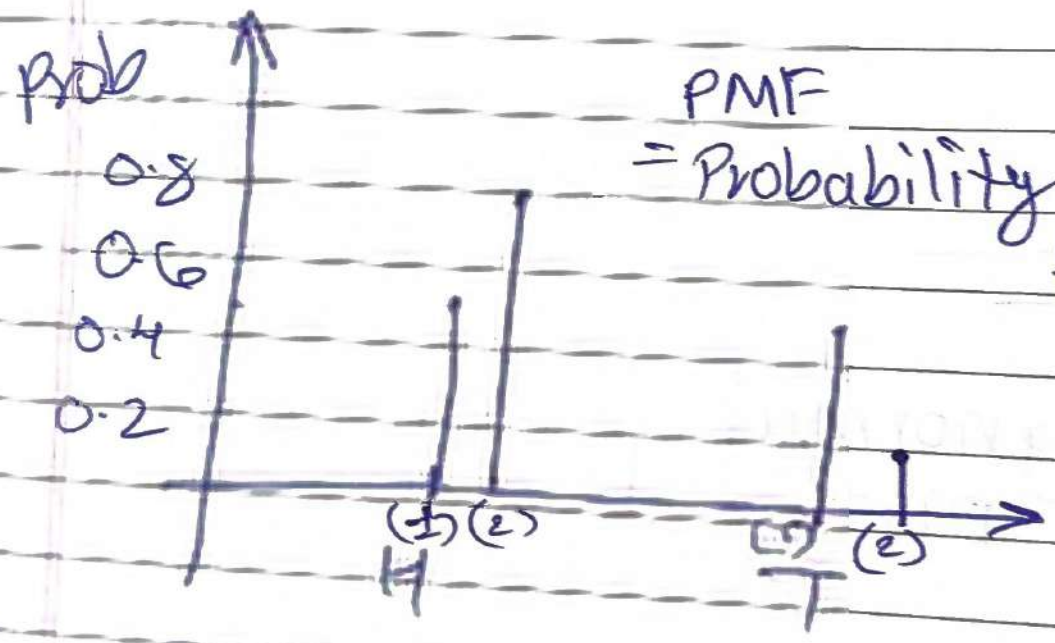
↳ (0 or 1)

e.g. Tossing a coin

$$P(H) = 0.5 = P$$

$$q = 1 - P$$

$$\boxed{q = 0.5}$$




PMF
= Probability Mass foction

Pdf ⟹ for continuous variables.
PMF ⟹ for categorical variables

# Binomial Distribution :

Every trial → Bernulli dist

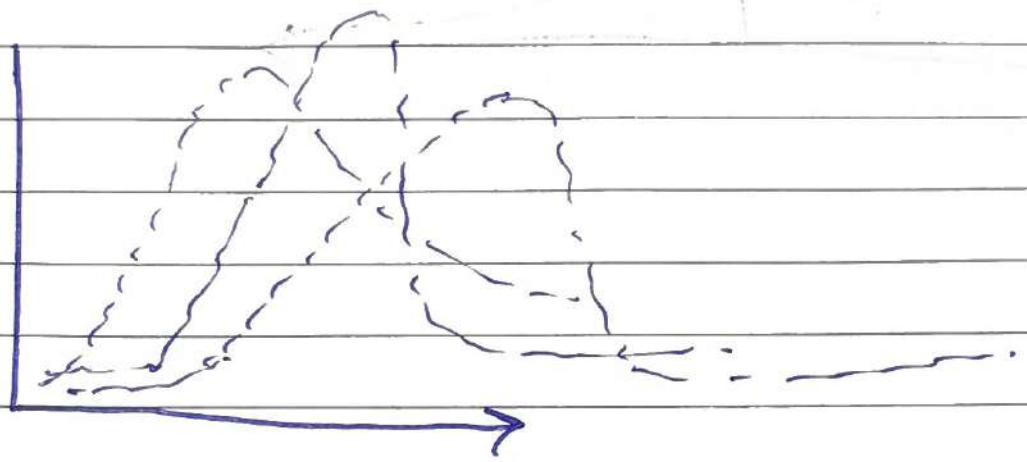
multiple trials combined

combination Bernullis distribution.

B(n, P)

↓ n0. of trials

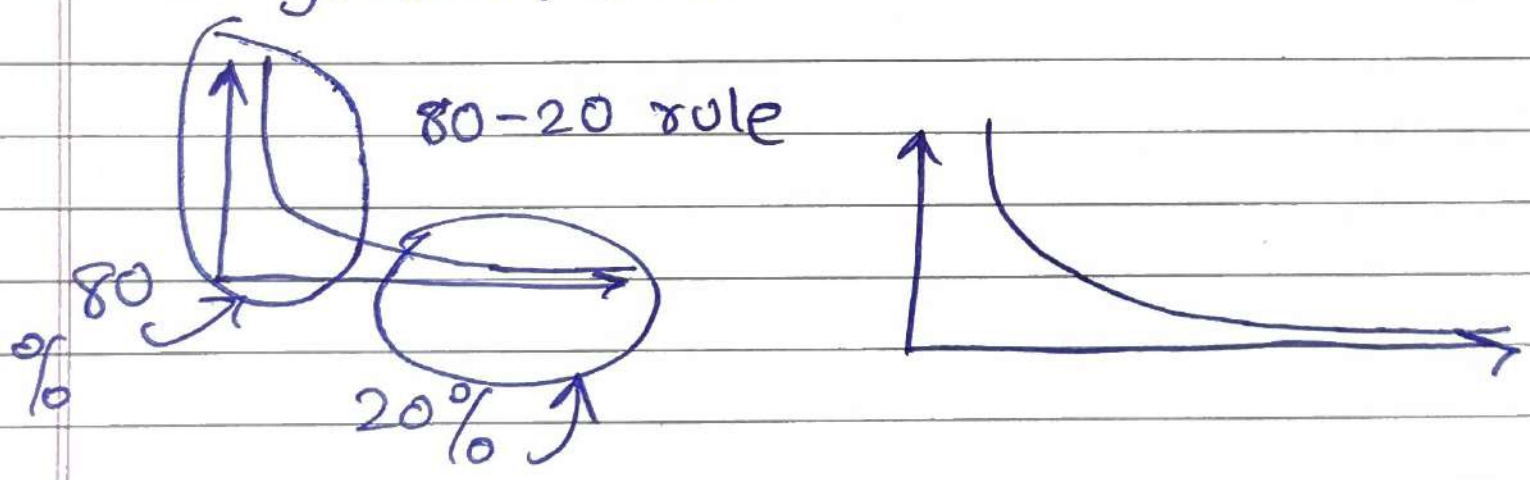
↳ Succes prob for eeach trial



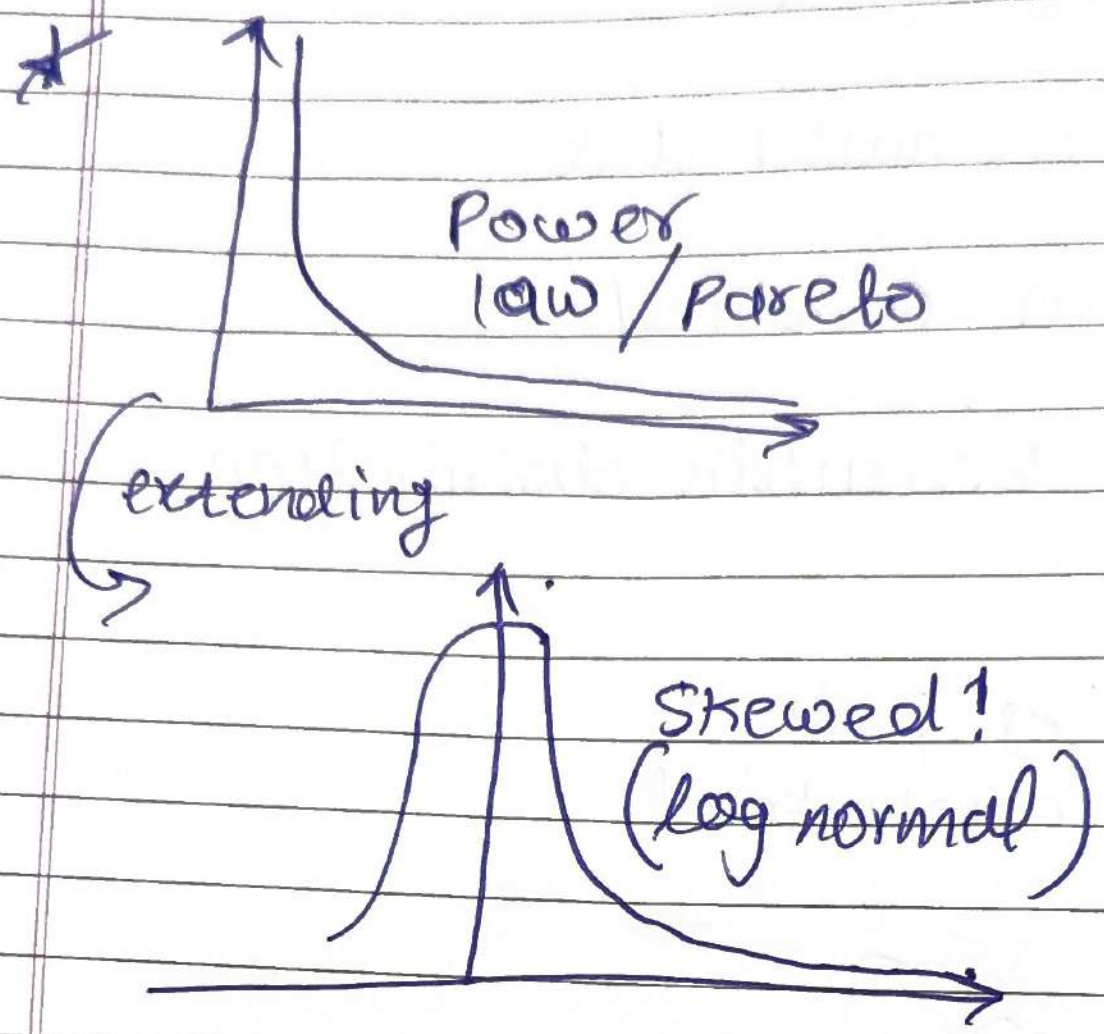
# Pareto dist:

↳ Power law

Non-guassion dist.




eg: 80% of wealth is distributed in 20%. of people.


Power
law / Pareto
extending
Skewed!
(log normal)