Statistics

1. (1) What is statistics?
   (2) Data
   (3) Types of statistics?
2. (1) Population and Sample
   (2) Types of Sampling Technique
3. (1) Variables
   (2) Kinds of Variables
   (3) Variables Measurement Scales
4. (1) Frequency Distribution
   (2) Bar Graph
   (3) Histogram
5. (1) Arithmetic Mean for Population and sample.
   (2) Measure of central Tendency
   (3) Median
6. (1) Mode
   (2) Measure of Dispersion
7. Percentiles and Quantiles
   Percentile
8. Five Number Summary
9. Box plot
10. Why Sample variance is divided by n-1?
11. Gaussian Distribution / Normal Distribution
12. Normalization
14. (1) Probability
    (2) Addition Rule ("or") Mutual Exclusive / Non Mutual Exclusive
15. (1) Multiplication Rule ("and") Indepent / Dependent
16. Permutation and Combination
17. P value, Hypothesis testing, Confidence Interval, Significance value.
18. Type 1 and Type 2 Error
    One Tail and two tail Test
19. Confidence Interval
21. One sample z-test
22. One sample t-test
23. Chi Square Test
24. Covariance.
25. Pearson correlation Coefficient
    Spearman rank correlation coefficient
26. P value and Significance value.
28. Log Normal Distribution
    Bernoulli Distribution
    Binomial Distribution
    Power Law {Pareto Distribution}
29. Central Limit Theorem
    Anova
30. One way anova

Day 1

What is statistics
{ For Bet

Data :-

Ex: (1) The
(2) Age

Types of
(1) Descrip
- It consis
summar

Ex: Class
now
percent
what ep t
in Depo
- Average
→ what is
of the s
{ Standard
→ what is th
passing o

Day 1

What is statistics?

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

{For Better Decision Making}

Data :- Facts or pieces of information that can be measured.

Ex: 1 The IQ of a class {98, 97, 60, 55, 75, 65}

2 Age of students of a class {30, 25, 24, 23, 27, 28}

Types of Statistics

1 Descriptive Statistics

- It consists of organizing and summarizing data.

2 Inferential statistics.

- It is a technique where in we used the data that we have measured to form conclusions.

Ex: Classroom of Maths student their are around 20 students and now I want to find Marks of 1st semester. Marks with respect to percentage are {84, 86, 78, 72, 75, 65, 80, 81, 92, 95, 96, 97, ...}

what type of question comes in Descriptive statistics?

- Average {Mean, Median, Mode}

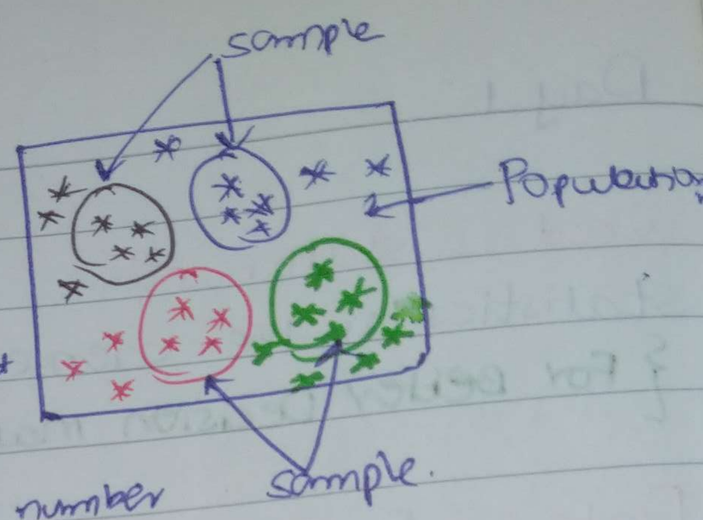
→ what is the average marks of the students in the class

{Standard Deviation} etc

→ what is the percentage of stundent passing out from class?

what type of question comes in Inferential Statistics?

→ Are the marks of the students of this classroom similar to the marks of the Maths classroom in the college?



Population (N) and Sample (n)

Ex :- Elections → ~~Exist Poll~~ → Exit Poll

→ In exit poll, we cannot ask every person that whom did you vote. so we take sample of data from ~~particular~~ different different region and maximum number of vote will crete exit poll.

Sampling Techniques.

① **Simple Random Sampling** :- Pick up any people Eg:- Exit Poll

→ When performing simple Random Sampling, every member of the Population (N) has an equal chance of being selected for your sample (n).

② **Stratified Sampling** :- Stratified sampling is a technique where the Population (N) is split into non-overlapping groups. (strata layer)

Ex :- ⊙ Gender → Male → survey
→ Female → survey

⊙ Age (0-10) (10-20) (20-30) (40-100)

⊙ Profession Doctor / Engineer

③ **Systematic Sampling** :- Here from the Population (N) we pick up any n^{th} individual.

Eg :- Mall → survey (covid) → 8th person → survey.

④ **Convenience Sampling / Voluntary Response Sampling** :- Only those people who are basically interested in this will basically do it.

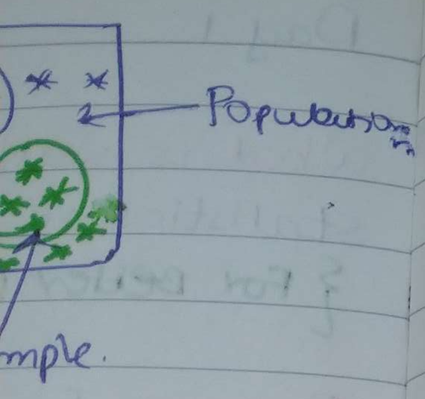
Eg :- Data Science Survey.

Note :- ① Exit Poll { Simple Random Sampling }

② RBI → Household → women { Stratified Sampling }

③ Drug → Test { To whom to ~~give~~ given based upon that we will selected or decide. }

ample



pou
er of the
r your sample(n).

ave where the
strata)
layer

ick up any

Only those
do it.

u deleted}
ude.

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

Eg:- Height = {150, 160, 170, 180} cm
Weight = {30, 40, 50, 60} kg

Kinds of variables :-

① Quantitative Variable,
- Measured Numerically.

Eg:- Age, Height
we can add, subtract etc

② Qualitative / Categorical Variable.
- Based on some characteristic we can derive categorical variable.

Eg:- Gender → Male, Female — we cannot add, subtract etc

Eg:- Blood Group, T-shirt Size, IQ.

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

↓

Quantitative Variable.

Discrete variable

Eg. Whole Number
No. of Bank Account 2, 3, 4.
Total Number of children in family
2, 3, 4.

Continuous variables

Eg:- Height = {162.3 cm, 170.1 cm}
weights = {100 kg, 92.3 kg}

Variables Measurement Scales.

① Nominal Data {Categorical Data} → split in classes → Eg: Gender, Color. etc

② Ordinal Data → Order of the data matters, but values does not.

Eg:- Students (Marks) Rank } Ordinal data

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

③ Interval Data → Orders matters, values also matter, natural zero is not present.

Eg:- Temperature (F) 70-80 80-90 90-100

HW Distance 10-20 20-30 30-40

④ Ratio :-

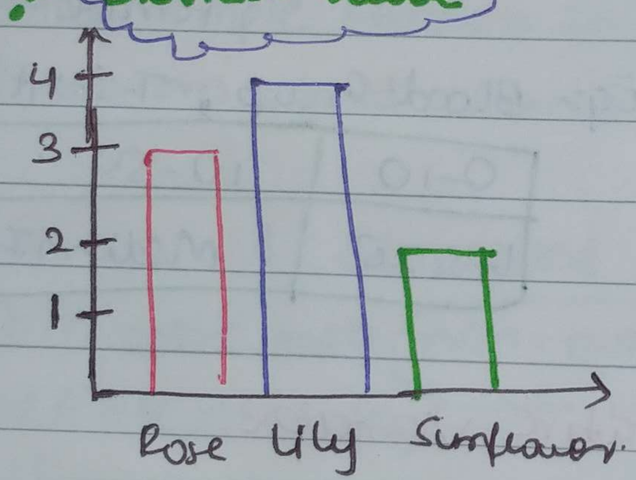
Frequency Distribution

Sample dataset = Rose, Lilly, Sunflower, Rose, Lilly, sunflower, Rose, Lilly, Lilly

Flower	Frequency	Cummulative Frequency
Rose	3	3
Lilly	4	7
Sunflower	2	9

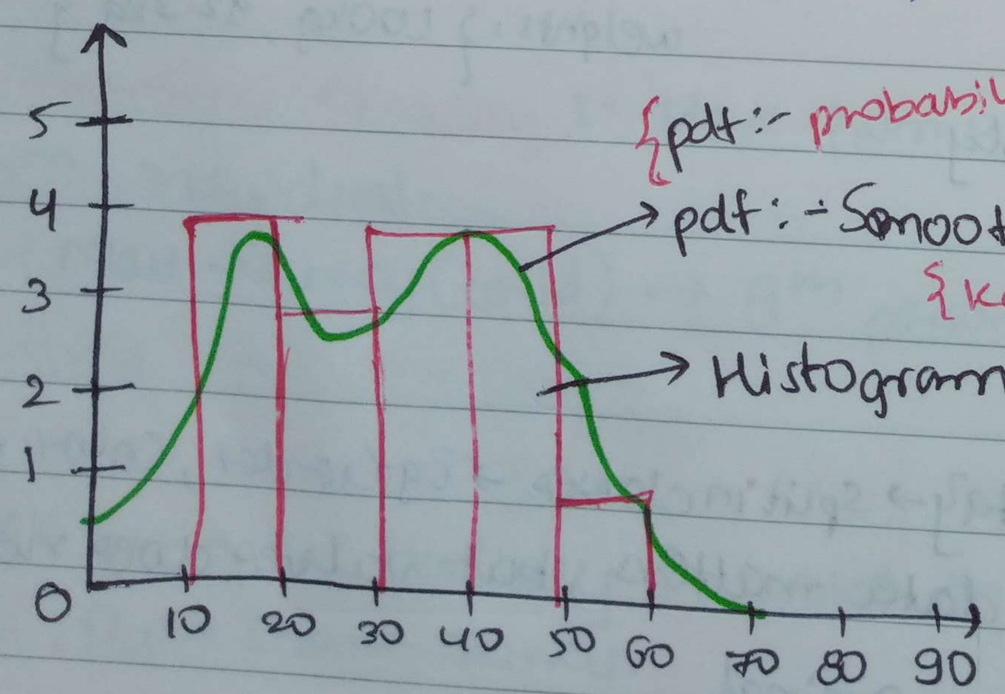
Bar Graph :- Discrete value

← {Summirazing the data}



Histogram :- Continuous value

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



{pdf :- probability density function}

{kde :- Kernel density Estimator}

Day 2

① Arithmetic Mean for Population and Sample

Mean (Average)

Population (N)

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

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

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

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

$$\boxed{\mu = 3.2}$$

Sample (n)

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

$$\boxed{\bar{x} = 3.2}$$

② Measure of Central Tendency ① Mean ② Median ③ Mode

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

③ Median

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

$\boxed{\mu = 3.2}$ ← Mean

Outliers

Data = {1, 1, 2, 2, 3, 3, 4, 5, 5, 6, 100}

$\boxed{\mu = 12}$ ← Mean

Data = {1, 1, 2, 2, 3, 3, 4, 5, 5, 6, 100}

Median:- step ① Sort the data.

step ② No. of element : 11

step ③ Selectelle middle number

1, 1, 2, 2, 3, 3, 4, 5, 5, 6, 100

$\boxed{\text{Median} = 3}$

Outliers

Data Updated = {1, 1, 2, 2, 3, 3, 4, 5, 5, 6, 100, 112}

step ① : Sort the data

step ② : No. of element : 12

step ③ : Select middle 2 number

1, 1, 2, 2, 3, 3, 4, 5, 5, 6, 100, 112

$$\boxed{\text{Median} = \frac{3+4}{2} = \frac{7}{2} = 3.5}$$

Note:- Median work very well with the outlier.

④ Mode :- Most frequent Element

Data :- { 1, 2, 2, 3, 4, 5, 6, 6, 6, 7, 8, 100, 200 }

$\boxed{\text{Mode} = 6}$ → Measure of Central Tendency

Data :- { 1, 2, 2, 3, 4, 5, 6, 6, 6, 7, 8, 100, 100, 100, 100 } → Outliers

$\boxed{\text{Mode} = 100}$

Note :- Huge difference between both the mode due to outlier.

Where to use Mode? → works well with categorical data.

Measure of Dispersion :- Spread

① Variance ② Standard Deviation

Data = { 1, 1, 2, 2, 4 }

$\boxed{\mu = 2}$

Data = { 2, 2, 2, 2, 2 }

$\boxed{\mu = 2}$

Now how we will identify that the above two data are totally different from each other. For that we use Variance.

① Variance

Population variance (σ^2)

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

Sample variance (s^2)

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

→ HW page 10 explained.

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

② Population Standard Deviation

$$\boxed{\sigma = \sqrt{\text{variance}}}$$

Sample Standard Deviation

$$\boxed{s = \sqrt{\text{variance}}}$$

x	μ
1	2.83
2	2.83
2	2.8
3	2.8
4	2.8
5	2.8

$\boxed{\mu = 2.83}$

$\boxed{\sigma^2 = \frac{10.84}{6}}$ Variance.

Variance

S.D = Fro

Percentil

Percentag

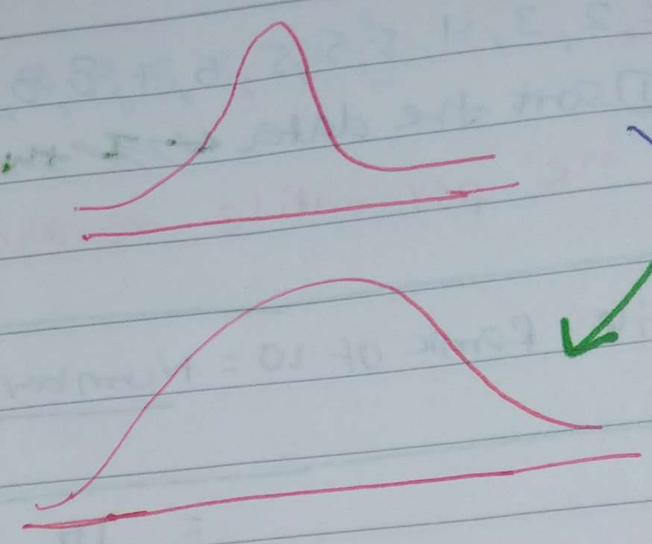
% of

%

Percenti

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

$\boxed{\Sigma = 10.84}$

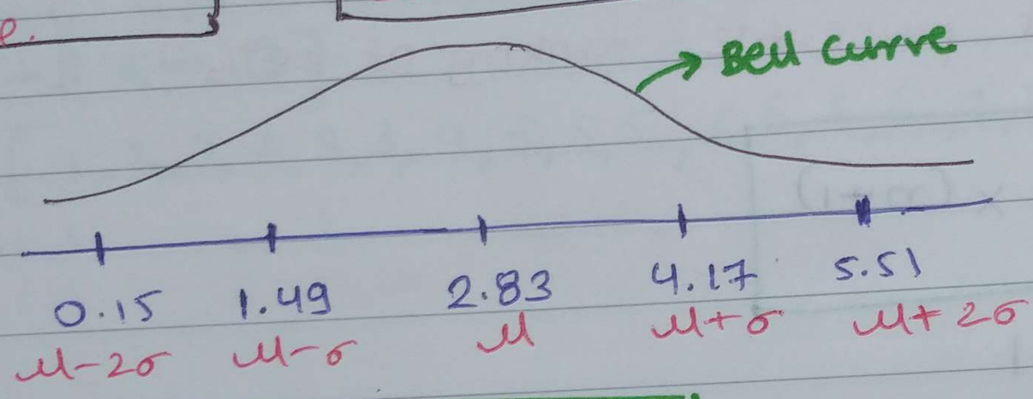


Variance is more?

$\boxed{\mu = 2.83}$

$\boxed{\sigma^2 = \frac{10.84}{6} = 1.81}$ Variance.

$\boxed{s = \sqrt{\text{variance}} = \sqrt{1.81} = 1.345}$ Standard Deviation



Variance = Spreadness of data

S.D = From mean how far the data is

Percentiles and Quartiles:

Percentage :- 1, 2, 3, 4, 5

% of the numbers that are odd?

$$\% = \frac{\text{Number of numbers that are odd}}{\text{Total Number}} = \frac{3}{5} = 0.6 = 60\%$$

Percentile :- A percentile is a value below which a certain percentage of observation lie.

Data = {2, 2, 3, 4, 5, 5, 5, 6, 7, 8, 8, 8, 8, 8, 9, 9, 10, 11, 11, 12}

Note :- ① Sort the data

What is the percentile ranking of 10?

$$\text{Percentile Rank of 10} = \frac{\text{Number of value below 10}}{n} \times 100$$

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

$$\text{Percentile Rank of 10} = 80\%$$

What value exists at percentile ranking of 25%?

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

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

$$= 5.21 \rightarrow \text{Index value.}$$

~~Index element 5 = 5~~ Index element at 5 is 5

~~Index element 6 = 5~~ Index element at 6 is 5

$$\therefore \text{Value} = \frac{5+5}{2}$$

$$\text{Value} = 5 \rightarrow 25\%$$

Five Number summary {Removing the Outlier}

① Minimum

② First Quantile (Q_1) 25%

③ Median

④ Third Quantile (Q_3) 75%

⑤ Maximum

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

Step ① · Sort the Data. ~~[illegible]~~.

[Lower fence ⟷ Higher fence]

Lower fence = $Q_1 - 1.5(IQR)$

Higher fence = $Q_3 + 1.5(IQR)$

$IQR = Q_3 - Q_1$

$Q_1 = 3$

$Q_3 = 7$

Lower fence = $Q_1 - 1.5(IQR) = 3 - 1.5(4) = 3 - 6 = -3$

Higher fence = $Q_3 + 1.5(IQR) = 7 - 1.5(4) = 7 + 6 = 13$

[$-3 \longleftrightarrow 13$]

Data = [1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9]

Minimum = 1

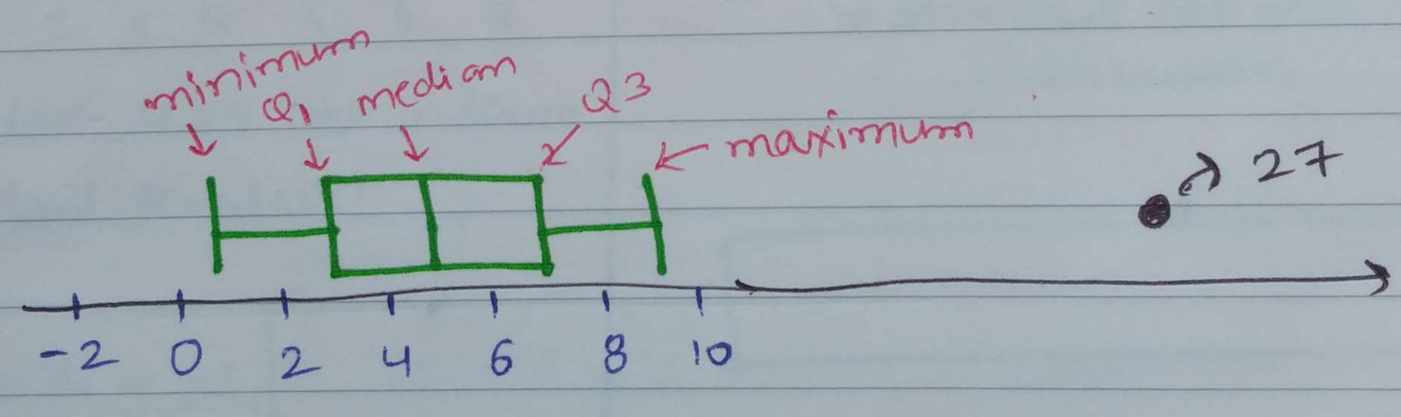
$Q_1 = 3$

$Q_3 = 7$

Median = 5

Maximum = 9

BOX PLOT



Note:- Box Plot is used to see the outlier in visulazation.

why Sample Variance is divided by $n-1$?

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

$s^2 \rightarrow$ Large, $n-1 \rightarrow$ small

Denominator :- Small

$\therefore s^2 =$ Large

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

$s^2 \rightarrow$ smaller, $n \rightarrow$ large

Denominator : Large

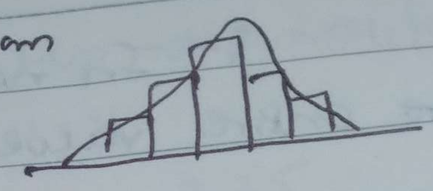
$\therefore s^2 =$ Small

Because we are underestimating true variance.

$$\therefore \boxed{s^2 \approx \sigma^2}$$

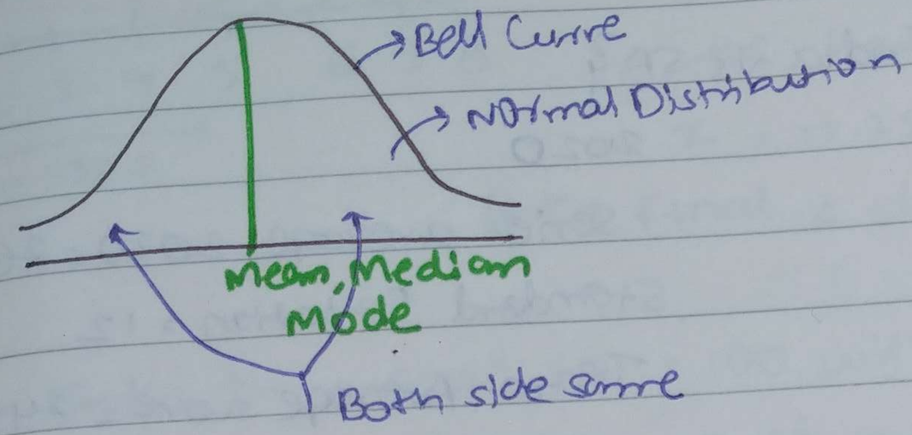
Day 3

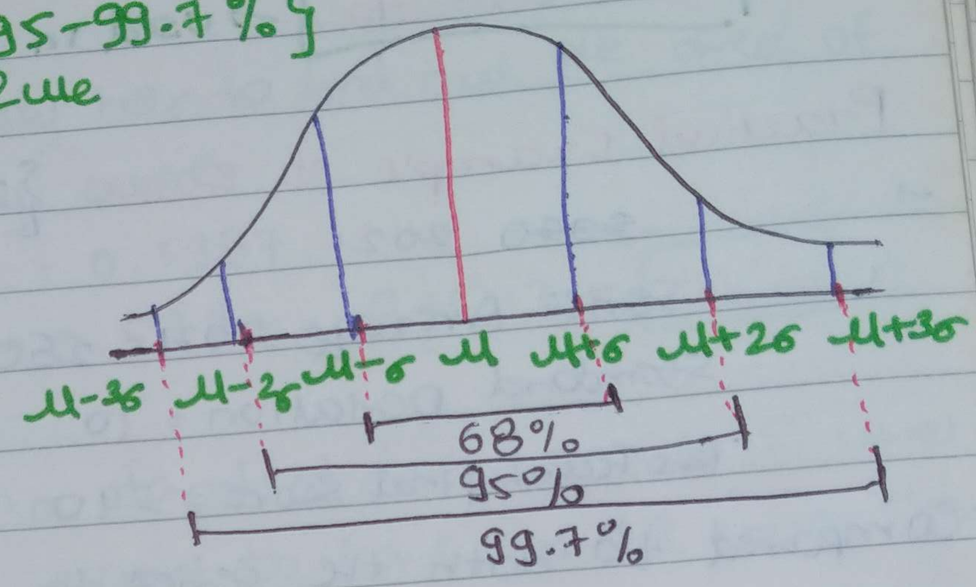
① Distribution of Dataset → Histogram



Ages = {24, 26, 27, 28, 30, 32, ... }

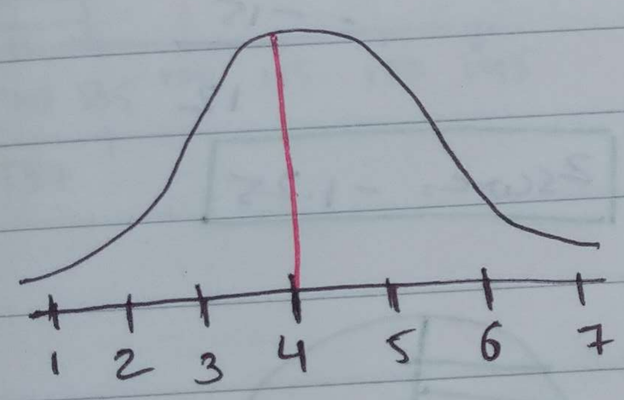
Gaussian / Normal Distribution {68-95-99.7 % Rule}





Eg:- Height → Normally Distributed → Domain Expert → Doctor.
weight, Iris Dataset

Eg: $\mu = 4$, $\sigma = 1$ →



where does 4.75 lies in terms of standard Deviation?

$$\boxed{z_{score} = \frac{x_i - \mu}{\sigma}}$$

$$= \frac{4.75 - 4}{1}$$

$= 0.75$ sd to right

where does 3.75 lies?

$$= \frac{x_i - \mu}{\sigma} = \frac{3.75 - 4}{1} = -0.25 \text{ sd to left}$$

Now, what happen if we apply z-score to every element.

→ Normal Distribution

$D_1 = \{1, 2, 3, 4, 5, 6, 7\}$ → Initially data.

$D_2 = \{-3, -2, -1, 0, 1, 2, 3\}$ → After applying zscore for every element

↳ standard Normal Distribution ($\mu = 0, \sigma = 1$)

$$\boxed{y \sim SND(\mu = 0, \sigma = 1)}$$

☆ Now, why we do this?

Age (years)	Salary (Rs)	Weight (kg)

→ Converting the data into same scale and the process is known Standardization.

Normalization :- In normalization we have an option i want to shift ~~are~~ entire value between 0 to 1. or (-1 to 1)

$\boxed{\text{MinMaxScalar}}$ → used in Image classification

Practical Example of Zscore {India vs SA}

→ ~~2020~~ 2021

Series Average 2021 = 250

Standard Deviation = 10

~~Ra's~~ Team final score = 240

2020

Series average 2020 = 260

Standard Deviation = 12

Team Average score = 245

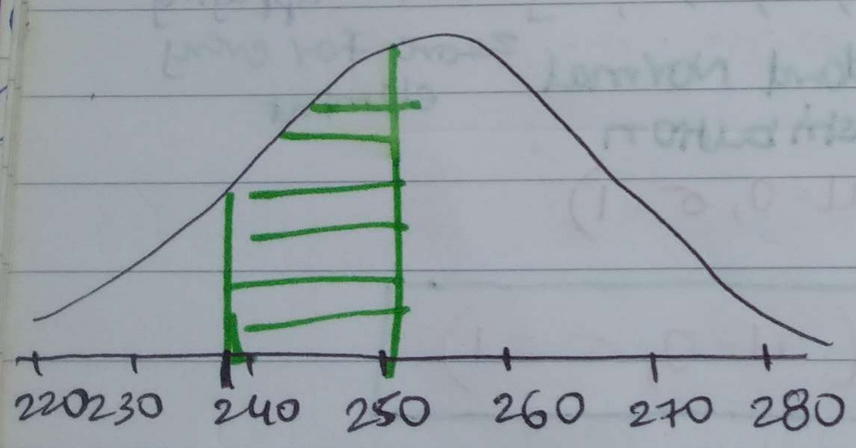
Compared to both the series in which year team perform better?

2021

$$Zscore = \frac{x_i - \mu}{\sigma} = \frac{240-250}{10}$$

$$= \frac{-10}{10}$$

$$\therefore \boxed{Zscore = -1}$$

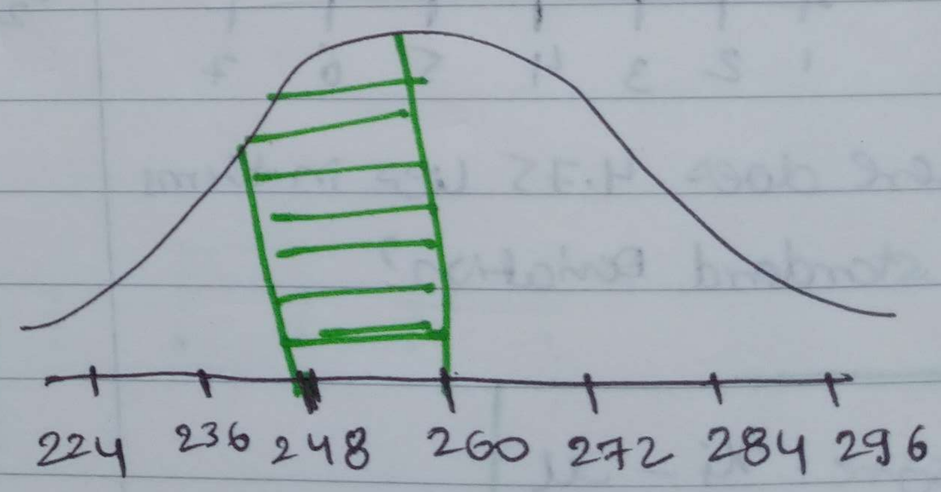


2020

$$Zscore = \frac{x_i - \mu}{\sigma} = \frac{245-260}{12}$$

$$= \frac{-15}{12}$$

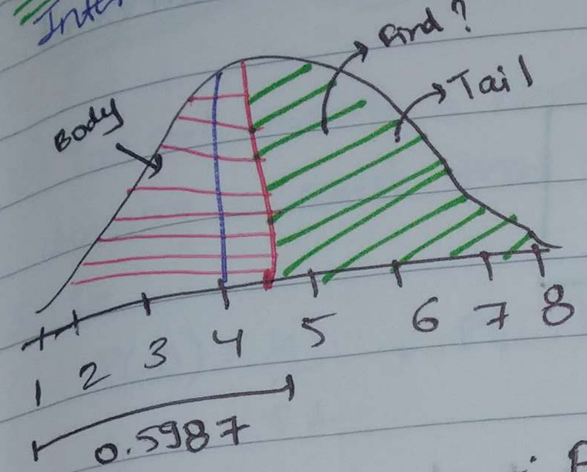
$$\boxed{Zscore = -1.25}$$



Team perform well in the year 2021

IMP IMP
Interview:-

Question:- what percentage of scores fall about 4.25? $\mu = 4$, $\sigma = 1$



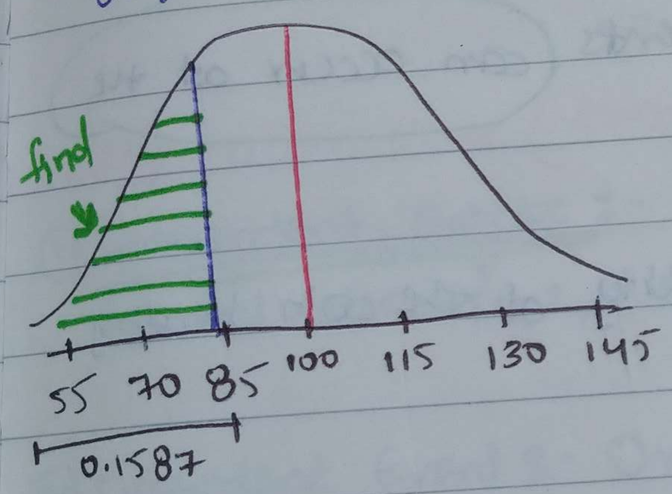
$$z = \frac{x_i - \mu}{\sigma} = \frac{4.25 - 4}{1} = 0.25$$

Z score actually help to find out the area of the body curve.

$z = 0.25 = 0.5987$ = Body

$\therefore$ Area Final $= 1 - 0.5987 = 0.4013 = \boxed{40\%}$

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



$$z = \frac{x_i - \mu}{\sigma} = \frac{85 - 100}{15} = \frac{-15}{15} = -1$$

$z = -1 = 0.1587$

$\therefore$ Area Final $= 0.1587 = \boxed{15\%}$

Day 4

Probability:- Probability is a measure of the likelihood of an event.

Eg:- Roll a dice = {1, 2, 3, 4, 5, 6}

$$P(6) = \frac{\text{Number of way an event can occur}}{\text{Number of possible outcome}} = \frac{1}{6}$$

Addition Rule (Probability, "or")

① Mutual Exclusive Event :- Two event are mutual exclusive if they cannot occur at the same time.

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

Tossing a coin = {H, T}

② Non Mutual Exclusive Event :- Multiple events can occur at the same time. Eg:- Deck of card {♠, ♡}

Question) If I Toss a coin, what is the probability of the coin landing on heads or Tails?

Sol:- Mutual Exclusive {Addition Rule}

$$P(A \text{ or } B) = P(A) + P(B) = 1/2 + 1/2 = 1$$

Q2) Roll a Dice → P(1 or 3 or 6)

$$P(1 \text{ or } 3 \text{ or } 6) = P(1) + P(3) + P(6)$$
$$= 1/6 + 1/6 + 1/6$$
$$= 3/6$$
$$= 1/2$$

l of an event.

clusive if they

cur at the

oin landing

Non Mutual Exclusive Event {Addition Rule}

Q) You are picking a card randomly from a deck. what is the probability of choosing a card that is Queen or a heart?

Sol:-

$P(Q) = \frac{4}{52}$, $P(\heartsuit) = \frac{13}{52}$, $P(Q \text{ and } \heartsuit) = \frac{1}{52}$

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

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

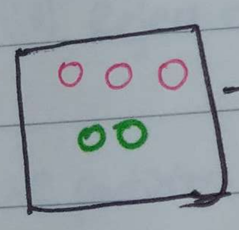
$$= \frac{16}{52}$$

Multiplication Rule (Probability, "and")

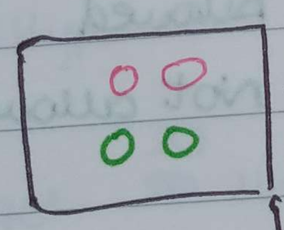
① Independent Events :- ~~Each~~ Each and Every are independent

Eg:- Roll a dice = {1, 2, 3, 4, 5, 6}

② Dependent Event :- One event affect the another event.



$\rightarrow P(R) = \frac{3}{5}$ $\Rightarrow$



$\rightarrow P(G) = \frac{2}{4}$

① what is the probability of rolling a "5" and then a "4" in a dice?

Sol:- Independent Event {Multiplication Rule}

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

$\therefore P(5 \text{ and } 4) = P(5) * P(4)$

$= 1/6 * 1/6$

$= \frac{1}{36}$

what is the probability of drawing a Queen and then a Aces from a deck of cards?

Dependent Event {Multiplication Rule}

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

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

$= \frac{4}{52} * \frac{4}{51}$

$= 0.0060$

Permutation and Combination

① Permutation :- All the possible arrangement of the element in any order

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

EX :- School trip {Chocolate factory} → Dairy, Sister, Milky Bar, Eclairs, Gems, Silk.

~~Student one 3~~ Task :- whichever chocolate you see and write only 3 entry.

Student → $\underline{6} \times \underline{5} \times \underline{4} = 120$ ⋮ → { Dairy, Gems, Milky / Milky, Gems, Dairy }

Using formula,

$n = 6$ ... option $r = 3$ ... to write

$${}^{n}P_{r} = \frac{n!}{(n-r)!} = \frac{6!}{3!} = \frac{6 \times 5 \times 4 \times 3!}{3!} = 120$$

② Combination :- Only unique ~~enter~~ entry no swaping or rearranged allowed

{ Dairy, Gems, Milky } → Allowed
{ Milky, Gems, Dairy } → Not allowed

$$\boxed{{}^{n}C_{r} = \frac{n!}{r!(n-r)!}} = \frac{6!}{3!(3!)} = \frac{6 \times 5 \times 4 \times 3!}{3! \times 3!} = 20$$

P-value

P = 0.01
1 time.

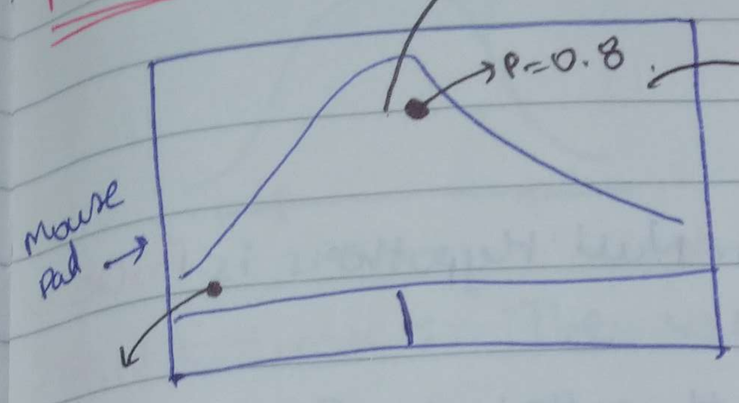
Hypothesis

Coin → Te
To check
i.e. 50

Hypothes
(i) Null
(ii) Alte
(iii) Exp
(iv) Reje

we perfo
should ha
be avai
∴ 100%

P-value



Every 100 times I touch the mouse pad 80 times I touch this specific region.

Hypothesis Testing, Confidence Interval, Significance Value;

Coin → Test whether this coin is a fair coin or not by perform 100 tosses.

To check whether coin is fair the pobability should be $P(H) = 0.5$ $P(T) = 0.5$

i.e. 50 times head ~~or~~ and 50 times tail.

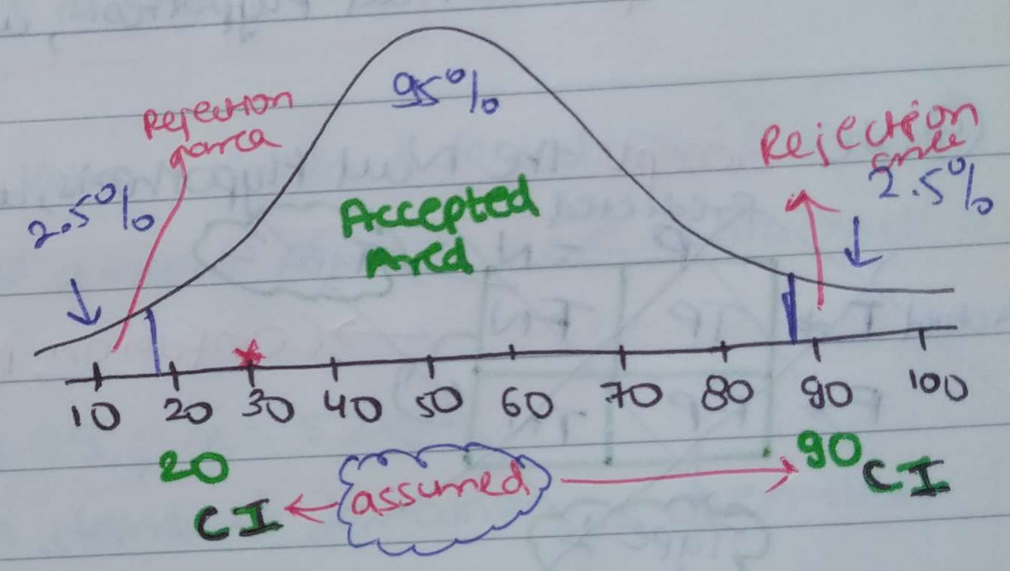
Hypothesis Testing

(i) Null Hypothesis :- Coin is fair

(ii) Alternate Hypothesis :- Coin is unfair

(iii) Experiment .... z test, t test ...

(iv) Reject or accept the Null Hypothesis



We perform 100 tosses we get 30 times head. Now ~~we~~ over experiment should be ~~near~~ near to the mean. How do we defined that how far it can be away from the mean? Significance Value. $\alpha = 0.05 = 5\%$ {Domain Expert}

$\therefore 100\% - 5\% \rightarrow 95\% \rightarrow$ Confidence Interval

Day 5.

Type 1 and Type 2 Error

Null Hypothesis (H_0) = Coin is fair

Alternate Hypothesis (H_1) = Coin is not fair

Reality Check :- Null Hypothesis is True or Null Hypothesis is False.

Decision :- Null Hypothesis is True or Null Hypothesis is False.

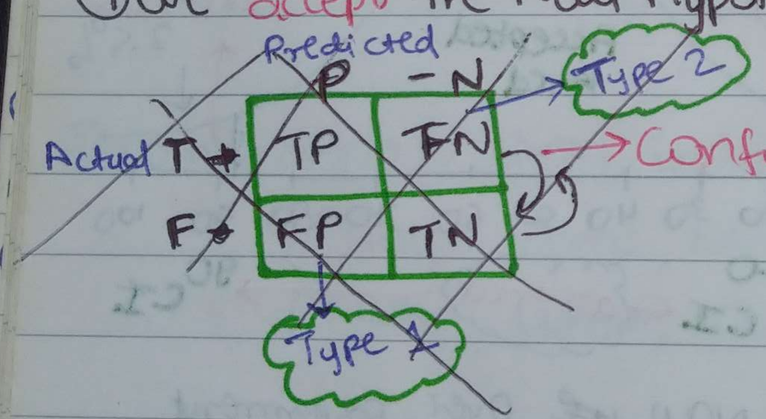
Outcome :-

① we reject the Null Hypothesis, when in reality it is false → Yes

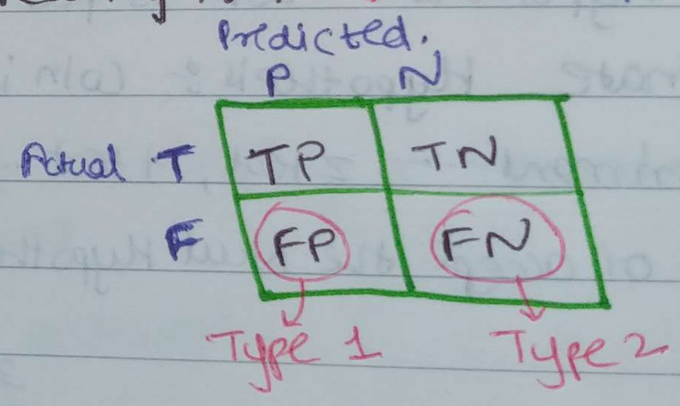
② we reject the Null Hypothesis, when in reality it is true → Type 1 Error

③ we accept the Null Hypothesis, when in reality it is false → Type 2 Error

④ we accept the Null Hypothesis, when in reality it is true → Yes



→ Confusion Matrix

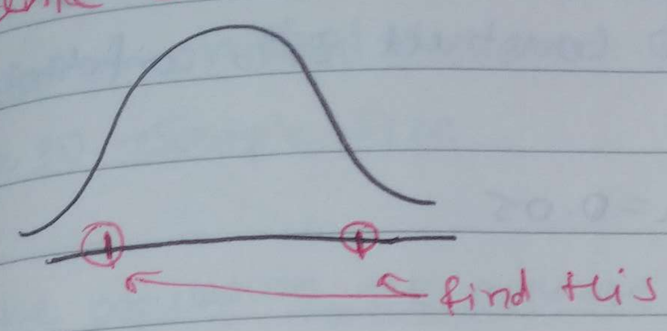


1 Tail and 2 tail Test

Q.] 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? $\alpha = 0.05$

→ Two tail test because percentage can increase or decrese of placement rate.

Confidence Interval



Point Estimate :- The value of any statistic that estimates the value of a parameter.

Sample Mean → Population Mean

$\bar{x} \longrightarrow \mu$

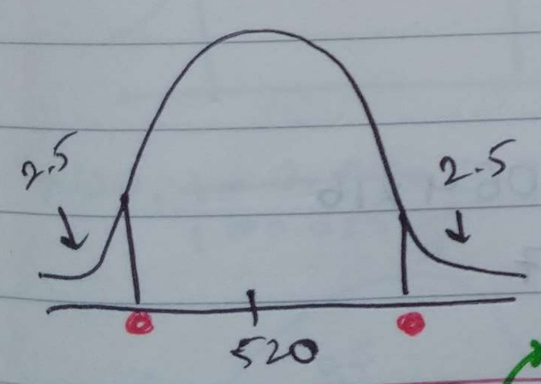
$\bar{x} = 2.9$ $\quad$ $\mu = 3$

$$\boxed{\text{Confidence Interval} = \text{Point Estimate} \pm \text{Margin Error}}$$

Q) On the quant test of CAT exam the population standard deviation is known to be 100. A sample of 25 test takers has a mean of 520 score. Construct a 95% CI about the mean?

Sol:- $\sigma = 100$, ~~$\bar{x} = 25$~~, $n = 25$, $\alpha = 95 = 0.05$, $\bar{x} = 520$

Note:- Population standard deviation is given then use z-test, $n \geq 30$



$$\therefore \boxed{CI = \bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}}$$

standard Error

$$= 520 \pm z_{\frac{0.05}{2}} \left(\frac{100}{\sqrt{25}} \right)$$

~~$= 520 \pm z_{0.025}$~~

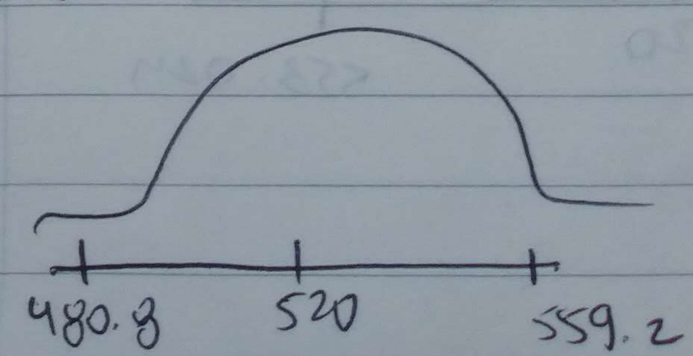
$= 520 + z_{0.025}(20)$ $\qquad$ $= 520 - z_{0.025}(20)$

Now, $1 - 0.025 = 0.975$

Now $0.975 = 1.96$

$\therefore CI = 520 + 1.96(20)$ $\qquad$ $CI = 520 - 1.96(20)$

$= 559.2$ $\qquad$ $= 480.8$



Q) On the quant test of CAT exam, a sample of 25 test takes has a mean of 520 with a standard deviation of 80. Construct 95% confidence interval about the mean?

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

Note:- Here population standard deviation is not given then we will use t-test.

CI = Point Estimate ± Margin of Error

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

→ Standard Error

Upper Bound $= \bar{x} + t_{0.05/2}\left(\frac{s}{\sqrt{n}}\right)$

Lower Bound $= \bar{x} - t_{0.05/2}\left(\frac{s}{\sqrt{n}}\right)$

Imp

$\boxed{\text{Degree of freedom} = 25 - 1 = 24}$

$\therefore$ Upper Bound $= 520 + t_{0.025}\left(\frac{80}{\sqrt{25}}\right)$

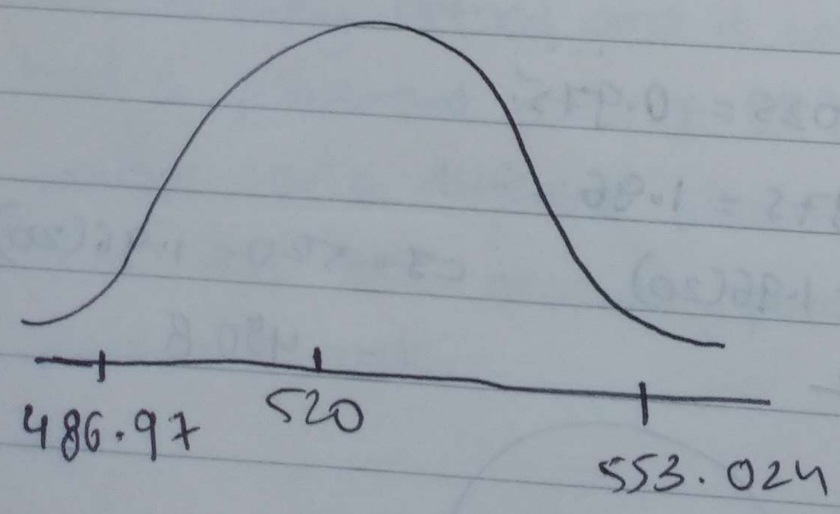
$= 520 + 2.064 \times 16$

$= 553.024$

Lower Bound $= 520 - t_{0.025}\left(\frac{80}{\sqrt{25}}\right)$

$= 520 - 2.064 \times 16$

$= 486.97$



One Sample z-test

① Population standard deviation is given

② $n \geq 30$ → Sample size

a) In the population, the average IQ is 100 with a standard deviation of 15. Reaserchers wants to test a new meditation to see if there is +ve or -ve 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 affect the intelligence? $\alpha = 0.05$, CI = 95%

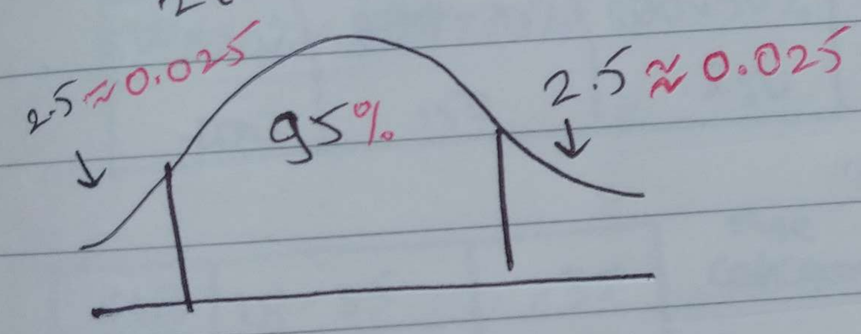
Sol) ① Define Null Hypothesis $(H_0) = \mu = 100$

② Alternate Hypothesis $(H_1) = \mu \neq 100$

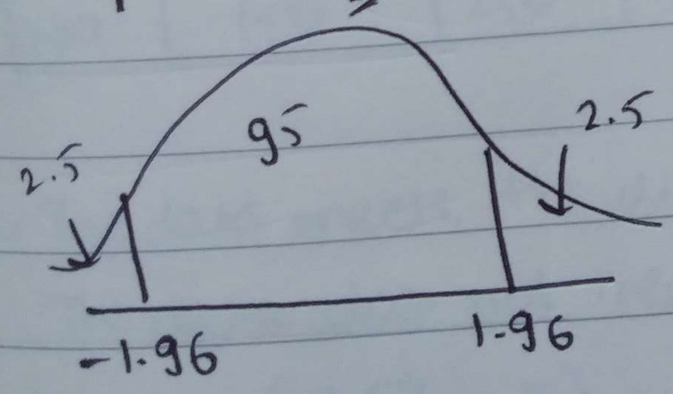
③ State Alpha $\alpha = 0.05$

④ State Decision Rule {Two Tails Test}.

2 tail Test



Now, ~~1 - 0.05~~ 0.975 ≈ 1.96
1 - 0.025



⑤ Calculate Test Statistis.

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

→ standard Error

$$z = \frac{140 - 100}{15/\sqrt{30}}$$

$$= \frac{40}{15/\sqrt{30}}$$

$$= \frac{40}{15} \times \sqrt{30}$$

$$\boxed{z = 14.60}$$

⑥ State Our Decision

$14.60 > 1.96$

∴ If z is less than -1.96 or greater than 1.96, reject the Null Hypothes

∴ Meditiation increse IQ.

One Sample t test

→ $\mu = 100$, $n = 30$, $\bar{x} = 140$, $s = 20$, Did medication after intelligence?
$\alpha = 0.05$

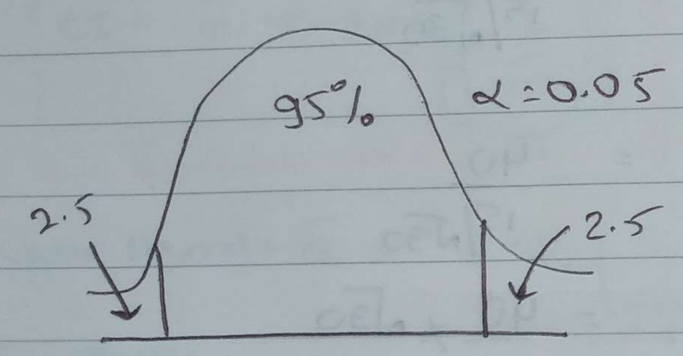
Sol:-

① $H_0 = \mu = 100$

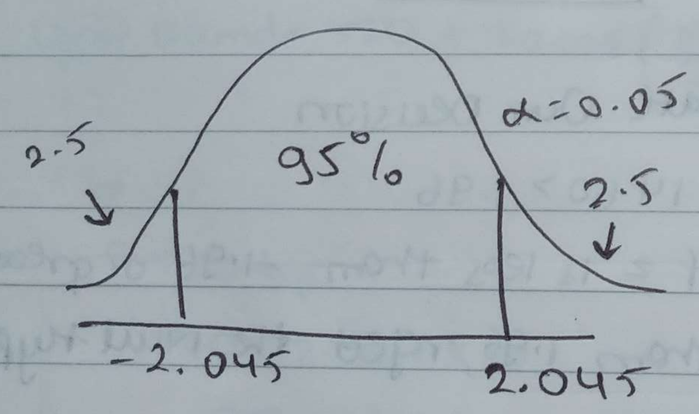
② $H_1 = \mu \neq 100$

③ Degree of freedom
$n-1 = 30-1 = 29$

④ State Decision Rule



$\because t_{0.05} = 2.045$, DOF = 29



⑤ T test

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

$$= \frac{140 - 100}{20/\sqrt{30}}$$

$$\therefore t = \frac{40}{20} \times \sqrt{30}$$

$$\cancel{= 20 \times \sqrt{30}}$$

$$= 2 \times \sqrt{30}$$

$$\boxed{t = 10.95}$$

⑥ We Reject Null Hypothesis.

Day - 6.

Chi Square Test :- claims about population proportions. It is a non parametric test that is performed on categorical (nominal or ordinal) data.

Eg:- 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 $n = 500$ individual were sampled. Below are the results.

<18	18 - 35	> 35
121	288	91

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

Sol:-

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

Population 2000 Expected

<18	18-35	>35
121	288	91

$n = 500$ Observed

500×20%	~~288~~ 500×30%	500×50%
100	150	250

→ Expected.

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

Age category ($n = 3$)
Observation
Expected

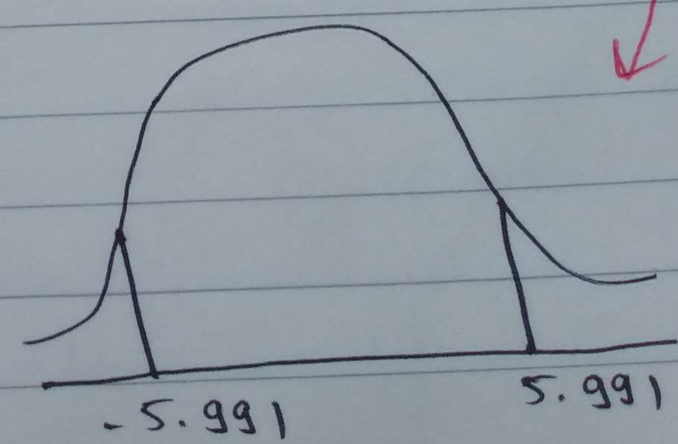
Using Chi-Square Table df = 2 and 0.05 CI is equal to 5.991

① H_0 = The data meets the distribution 2000 censis.
H_1 = The data does not meets the distribution 2000 censis.

② $\alpha = 0.05$ (95% CI)

③ Degree of freedom $= n-1 = 3-1 = 2$
df

④ Decision Boundary



P.T.O →

If χ^2 is greater than 5.991 reject (H_0) Null Hypothesis.

(5) calculate test statistics.

$$\boxed{\chi^2 = \sum \frac{(fo - fe)^2}{fe}}$$

fo → Observed
fe → Expected.

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

$$= \cancel{234}$$

$$\boxed{\chi^2 = 232.484}$$

$\chi^2 = 232.484 > 5.99$ {Reject Null Hypothesis}

Covariance.

weight(x)	Height(y)
50	160
60	170
70	180
75	181

$\boxed{X\uparrow Y\uparrow}$ $\boxed{X\downarrow Y\downarrow}$ +ve correlation

No. of hours Study	Play
2	6
3	4
4	3

$\boxed{X\uparrow Y\downarrow}$ $\boxed{X\downarrow Y\uparrow}$ -ve correlation

$$\boxed{Cov(x,y) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1}} = +ve \text{ or } -ve \text{ or } 0 \rightarrow \text{No relation}$$

Disadvantage of Covariance

How much negative correlative or how much positive correlated.

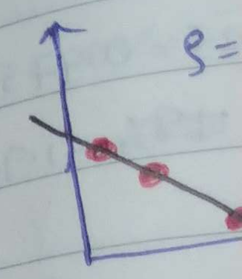
Pearson Co

The more t

The more

$\rho(x,y)$

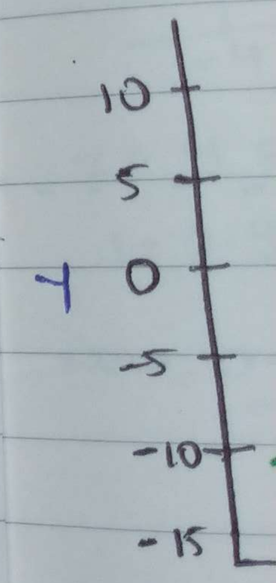
$\rho =$



$X\uparrow$

$X\downarrow$

Spearmann



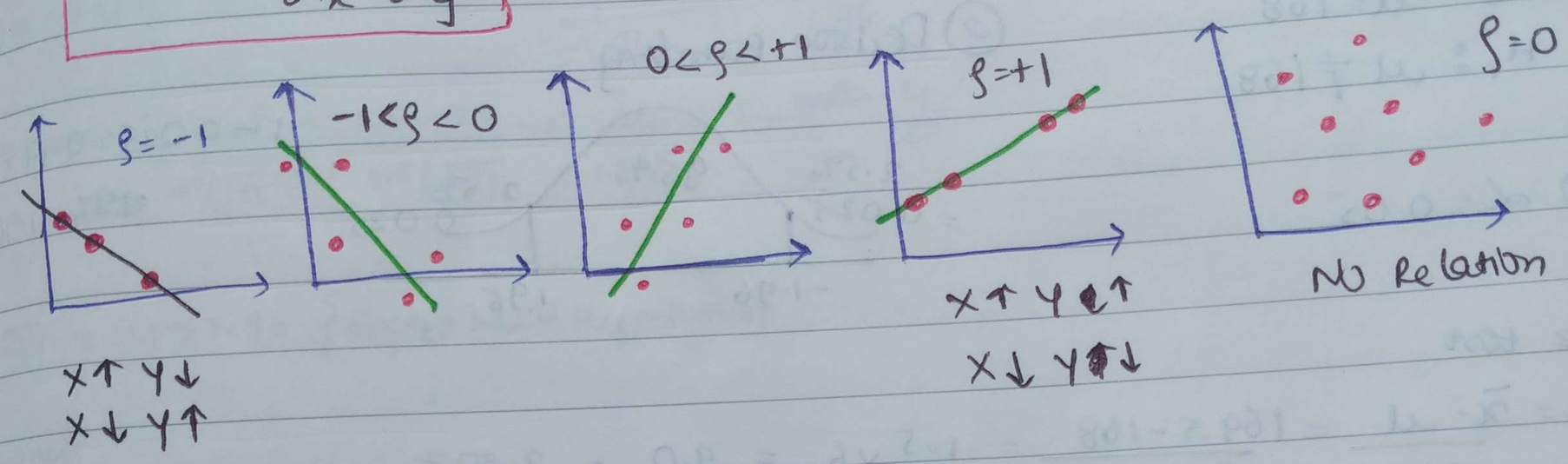
Eg:-

Pearson Correlation Coefficient {-1 to 1}

The more towards +1 more positively correlation.
The more towards -1 more negatively correlation.

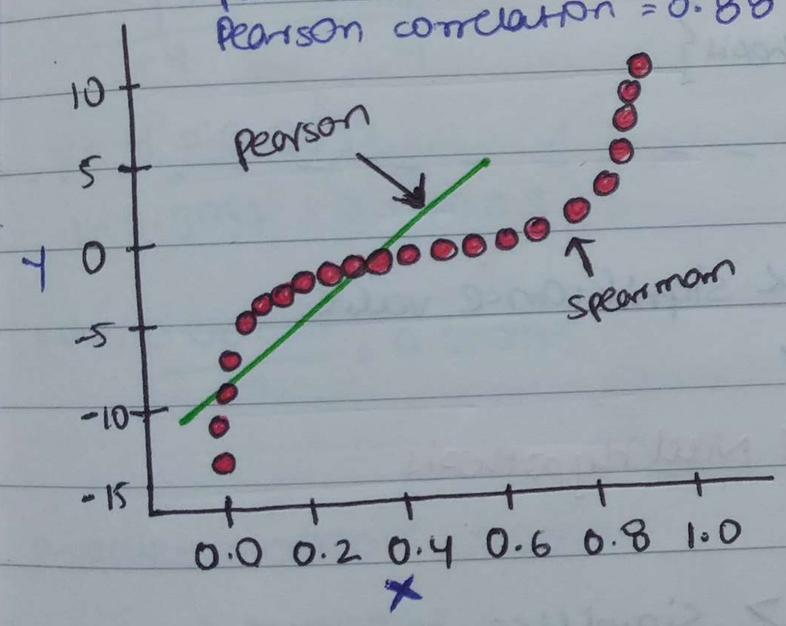
$$\rho_{(x,y)} = \frac{Cov(X,Y)}{\sigma x \ \sigma y}$$

Note:- It capture only linear property.



Spearman rank correlation Coefficient

Spearman correlation = 1
Pearson correlation = 0.88



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

Note:- It capture both linear and non linear property.

Eg:-

Height (X)	weight (y)	R(x)	R(y)
170	75	2	2
160	62	3	3
150	60	4	4
145	55	5	5
180	85	1	1

Ignore — will used in spearman Rank

Day 7

P-value and Significance value

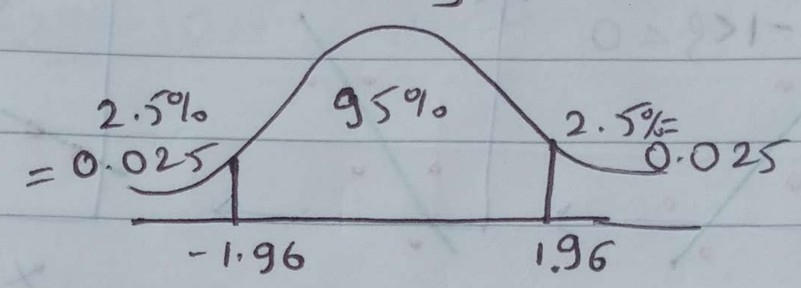
→ The average weight of all residents in Banglore city is 168 pounds with a standard deviation 3.9. we take a sample of 36 individuals and the mean is 169.5 pounds. C.I. = 95%

→ $\mu = 168$, $\sigma = 3.9$, $n = 36$, $\bar{x} = 169.5$, {C.I = 95 ~~$\alpha = 1 - 0.05 = 0.025$~~ no}

① $H_0 : \mu = 168$
$H_1 : \mu \neq 168$

② $\alpha = 0.05$

③ Decision Boundary



$1 - 0.025 = 0.975$

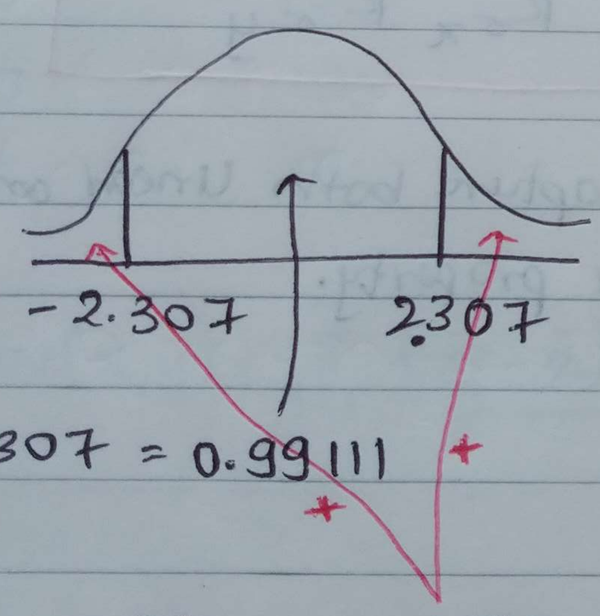
$= 1.96$

④ z test

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{169.5 - 168}{3.9/\sqrt{36}} = \frac{1.5}{3.9} \times 6 = \frac{9.0}{3.9} = \underline{\underline{2.307}}$$

⑤ $z = 2.307 > 1.96$ {Reject Null Hypothesis}

P value.



$z_{table} = 2.307 = 0.99111$

Now, $1 - 0.99111 = 0.00889$

Now, $\frac{0.00889}{2} = 0.0044$

Both side area

Now, pvalue $= 0.0044 + 0.0044$
$= 0.0088$

p value < 0.05

$0.0088 < 0.05$ {Reject Null Hypto}

Note

p-value < significance value.
↓
Reject Null Hypothesis

p-value > significance value
↓
Fail to Reject the Null Hypothesis.

→ Average age of a college is 24 years with a standard deviation 1.5. Sample of 36 students mean is 25 years. with $\alpha = 0.05$, C.I = 95%, do the age vary?

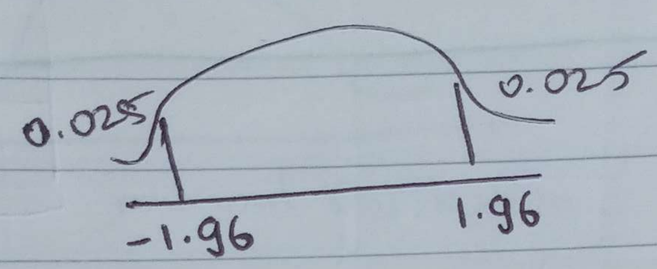
Sol:- $\mu = 24$, $\sigma = 1.5$, $n = 36$, $\bar{x} = 25$, $\alpha = 0.05$

① $H_0 : \mu = 24$

$H_1 : \mu \neq 24$

② $\alpha = 0.05$

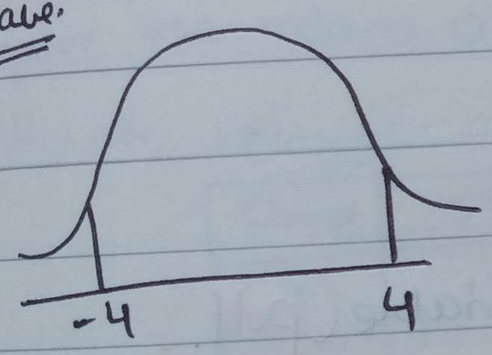
③ Decision Rule.



④ $z\text{-score} = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \dfrac{25 - 24}{1.5/\sqrt{36}} = \dfrac{1}{1.5/\sqrt{36}} = \underline{\underline{4}}$

⑤ $z = 4 > 1.96$ {Reject Null Hypothesis}

P-value.



$z = 4 = 0.9997$

$1 - 0.9997 = 0.00003$

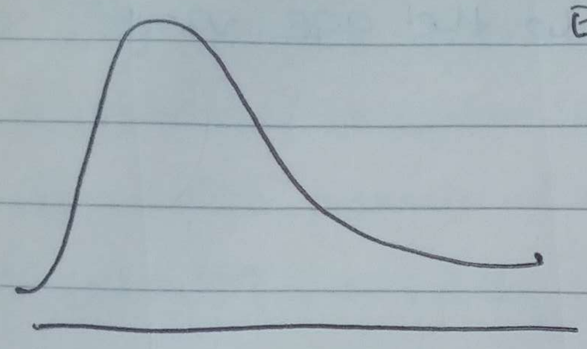
Now, $\dfrac{0.00003}{2} = 0.000015$

P-value $= 0.000015 + 0.000015$

$= 0.00003$

$0.00003 < 0.05$ {Reject Null Hypothesis}

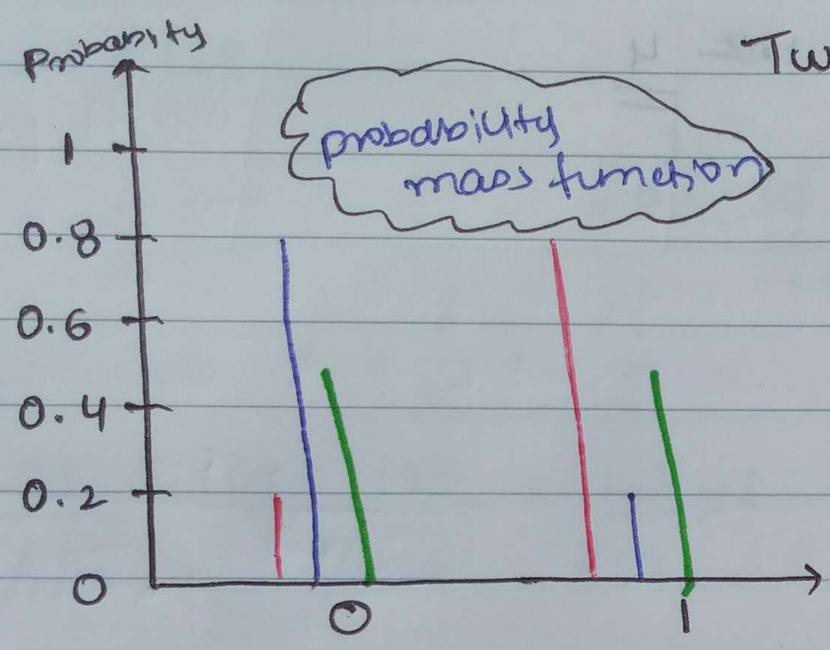
Log Normal Distribution



Eg:- ① Wealth Distribution ② People writing comments.

$y \approx$ Log Normal Distribution $\Longrightarrow \log(y) \rightarrow$ Normal Distribution

Bernoulli Distribution



Two Outcomes $\{0 \text{ or } 1\}$

$p = 0.5$

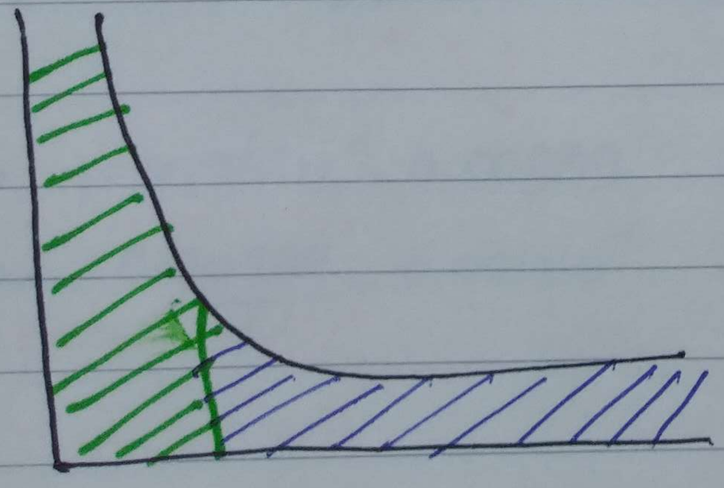
$q = 1 - p = 1 - 0.5 = 0.5$

Note:- probability density function is for continuous variable (pdf)
probability mass function is for categorical variable (pmf)

Binomial Distribution :- Combination of Bernoulli Distribution.

$\boxed{B(n, p)}$ n – number of trials, p = probability.

Pareto Distribution {Power Law} {80-20}



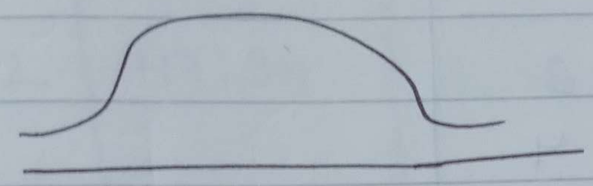
Eg:- 80% of wealth is distributed with 20% of the people.

Eg:- 80% of the company project are done by 20% of the people in a team

Central Limit Theorem

whatever distribution it may be, the sample of of the distribution will follow Normal Distribution. $\boxed{n \geq 30}$

$S_1 \rightarrow \bar{x}_1$
$S_2 \rightarrow \bar{x}_2$
$S_3 \rightarrow \bar{x}_3$
$S_4 \rightarrow \bar{x}_4$
$\vdots$
$S_n \rightarrow \bar{x}_m$

$\rightarrow$ 

Normal Distribution.

Anova {Analysis of Variance} (f-test) → Anova is a statistical method used to compare the means of 2 or more groups. ① Factors (variable) ② Levels → {IMP}

{Anxiety Reducing Medicine}

Medicine. Dosage.

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

factors :- Dosage
levels :- 0mg, 50mg, 100mg

Gender :- Factors, Male, Female :- Levels.

Types of Anova

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

Eg :- Dosage.

② Repeated Measure Anova :- One factor with atleast 2 levels, but the level are dependent.

Running {Factor}

	Day 1	Day 2	Day 3	Day 4 {levels}
1	6km	4km	2km	10km

③ Factorial Anova :- Two or more factors each of which with atleast 2 levels, levels can be either independent, dependent or both (mixed).

Eg:-

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

Factors :- Days & Gender

Levels :- D1, D2, D3, Men, Women

One way Anova (f test).

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

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

Sol:- ① H_0 :- $\mu_{0mg} = \mu_{50mg} = \mu_{100mg}$

② H_1 :- $\mu_{0mg} \neq \mu_{50mg} \neq \mu_{100mg}$

③ State the α and CI

$\alpha = 0.05$ $\alpha = 95\%$

③ Calculate Degree of freedom

$N = 21$, $n = 7$

a = level

$df_{Between} = a - 1 = 3 - 1 = 2$

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

$df_{Total} = N - 1 = 21 - 1 = 20$

④ State Decision Rule.

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

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

From F table (2, 18) = 3.5546 $\{\alpha = 0.05\}$

If f test is greater than 3.5546, Reject the Null Hypothesis

⑤ Calculate F

	Sum of SS
Between	98.
within	10.
Total	108

$SS_{between} = \Sigma$

$\Sigma(\Sigma(a_i)^2) = (9+$

$(7$

$($

$= 5$

$T^2 = [$

$\therefore SS_{Between} =$

$\therefore SS_{Betw}$

$SS_{within} = \Sigma$

$\Sigma y^2 = ($

$($

$($

$\Sigma y^2 =$

$\therefore SS_{within} =$

SS_{within}

nder
en, women

articipants
y anxiety
3 conditions.

μ_{100mg}

μ_{200mg}

2, 18)

6 $\{\alpha = 0.05\}$

546, Reject

5) Calculate F test statistics.

	Sum of Square SS	dof	Mean Square	F Test ~~value~~
Between	98.67	2	= 49.34	86.56
within	10.29	18	= 0.57	
Total	108.96	20		

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

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

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

$$\boxed{T^2} = [57+47+21]^2 = \boxed{125^2}$$

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

$$\therefore \boxed{SS_{Between} = 98.67}$$

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

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

$$\boxed{\Sigma y^2 = 853}$$

$$\therefore SS_{within} = 853 - \left[\frac{57^2+47^2+21^2}{7}\right]$$

$$\boxed{SS_{within} = 10.29}$$

$$F_{test} = \frac{MS_{Between}}{MS_{within}} = \frac{49.34}{0.57}$$

$$\therefore \boxed{F_{test} = 86.56}$$

Final Conclusion:-

$86.56 > 3.5546$, so we reject the Null Hypothesis.