Inferential Stats

* Basics of Probability

Many a time you may require large amount of data for analysis which may need too much time and resources to acquire.

→ In such situations we are forced to work with a "smaller sample of data", instead of having the entire data to work with

→ Situations like these arise all the time at big companies like Amazon, for eg Amazon QC dept wants to know what proportion of the product in it's warehouse are defective, so instead of going through all their products, the Amazon QC team can just check a small sample of 1000 products and then find for this sample the defect rate (i.e. the proportion of defective rate)

⇒ then based on the sample's defect rate, the team can 'infer' what the defect rate is for all the products in the warehouse.

⇒ "the process of "inferring" insights from sample data is called Inferential Statistics"

So in easy words "Inferential statistics" is a way of making guesses or conclusions about a larger group (called as Population) by looking at the data from a smaller group (called as Sample)

Eg Instead of Asking 1000000s of people in the city what their favourite fruit is, we could ask 100 people (a sample) and use that info to predict what the whole city (population) might prefer.

Probability

↳ How likely something is to happen.

Many events can't be predicted with total certainity, the best we can say is how likely they are to happen, using the idea of probablity.

⇒ Tossing a Coin : When a coin is tossed, there are two possible outcomes
Heads (H) or Tails (T)

Also:
→ the probablity of the coin landing H is $1/2$
→ the probablity of the coin landing T is $1/2$

⇒ Throwing a Dice : When a single dice is thrown there are 6 possible outcomes 1, 2, 3, 4, 5, 6
the prob of any 1 of them is $1/6$

Probablity in general : $\frac{\text{Nos of ways it can happen}}{\text{total Nos of outcomes}}$

Eg chance of rolling a 4 with dice

(↙ Nos of ways it can happen : 1 (there is only 1 face with a 4 on it)

total nos of outcomes : 6 (∵ there are 6 faces altogether)

$\Rightarrow$ Probablity $= 1/6$

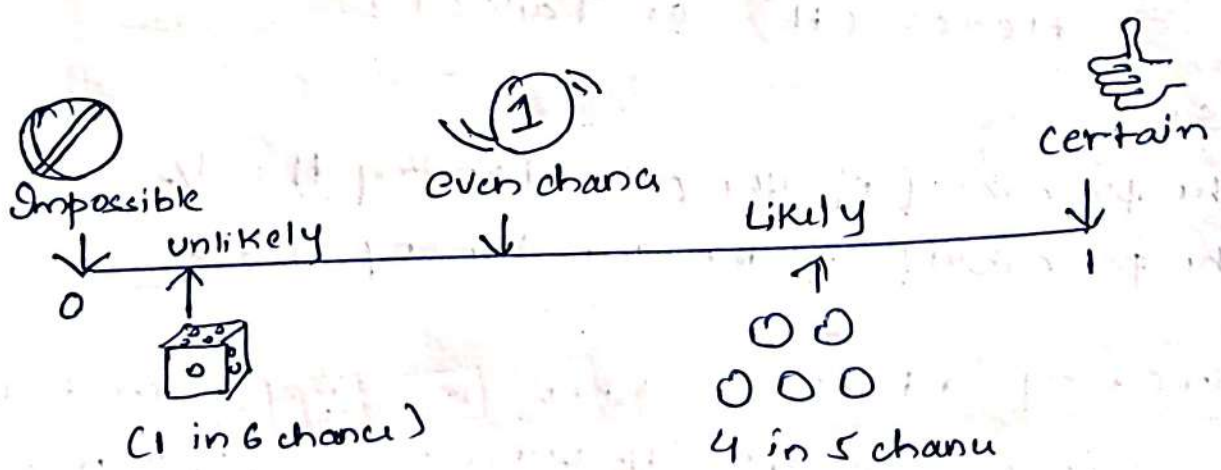
Eg 2 5 Marbles in a bag → 4 blue and 1 red.

(P(of blue marble gets picked)) $\Rightarrow 4/5$

∵ 4 → there are 4 blue marbles

5 → there are 5 marbles in total

Probablity Line : we can show probablity on a Probablity line



Probablity is always b/w 0 and 1

Probablity is just a guide

It does not tell us exactly what will happen It is just a guide.

eg : toss a coin 100 times, how many heads will come up

↳ Probablity says heads have 1/2 chances so we can expect 50 heads, but in reality we can get 48 heads or 55 heads etc.

⊛ Some words have special meaning in Probably

⇒ Experiment : A repeatable procedure with a set of possible results

eg : throwing dice

we can throw the dice again and again so it is a repeatable.

→ the set of possible outcomes for any single throw is {1, 2, 3, 4, 5, 6}

⇒ Outcome : A possible result

eg : 6 is one of the outcomes of a throw of a dice.

⇒ Trial : A single performance of an experiment

eg : conducted a coin toss experiment & after 4 trials the results recieved

outcome	Trial 1	Trial 2	Trial 3	Trial 4
H	✓	✓		✓
T			✓	

So the 3 trials had the outcome Head and One trial had the outcome Tail.

⇒ Sample Space : All the possible outcomes of an experiment.

Eg choosing a card from a deck
↳ 52 cards in deck
So sample space is all 52 possible cards

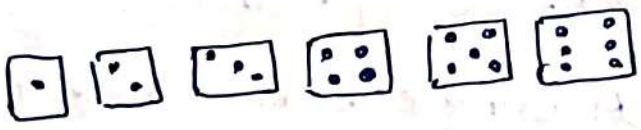
the sample space is made up of sample Points

⇒ Sample point : Just one of the possible outcom

Eg : Deck of cards
→ 5 of the clubs is a sample point
→ the King of hearts is a sample Point

Note "King" is not a sample point, there are 4 Kings so 4 different sample points

Eg : throwing dice
there are 6 diff sample points in the sample space



⇒ Event : One or more outcomes of an experiment

Eg of Events

An event can be just one outcome
→ getting a tail when tossing a coin
→ Rolling a "5"

An event can include more than one outcome
→ choosing a king from a deck of cards (any of 4 kings)
→ Rolling an "even no" (2, 4 or 6)

⊛ Additional Rule of Probablity

the addition rule of probablity is a fundamental concept used to find the probablity of the union of two events (i.e. the probablity that at least one of the events occur). the formula depends on whether the events are mutually exclusive or not.

① When events are mutually exclusive

If two events A and B cannot occur at the same time (no overlap), the probablity of A or B occuring is the sum of their individual probablities.

$$P(A \cup B) = P(A) + P(B)$$

eg Rolling a dice, let A be rolling a 2 and B be rolling a 5, ∴ we can't roll both a 2 and a 5 in one roll, A and B are mutually exclusive.

$P(A) = 1/6$, $P(B) = 1/6 \Rightarrow P(A \cup B) = P(A) + P(B)$

$\Rightarrow P(A \cup B) = 1/6 + 1/6 = 2/6 = 1/3$

(2) When events are not mutually exclusive

→ If two events A and B can occur simultaneously (overlap exists), the formula accounts for the overlap by substracting the probablity of both events happening together.

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

eg Drawing a card from a deck, A be drawing a heart, B be drawing a face card

Some cards (eg the king of hearts) satisfy both A and B so they aren't mutually exclusive.

$P(A) = 13/52$ $P(B) = 12/52$ $P(A \cap B) = 3/52$

$\Rightarrow P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$= (13 + 12 - 3)/52 = 22/52 = 11/26$

Keypoints to remember here

→ $P(A \cup B) = P(A) + P(B)$ if events are mutually exclusive

→ $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ for non mutually exclusive events

⇒ $P(A \cap B)$ is critical in the second case to avoid double counting.

Examples

① Mutually Exclusive Events

toss a coin

↳ A getting Heads ⇒ $P(A) = 0.5$

↳ B getting tails ⇒ $P(B) = 0.5$

What is the prob of getting heads or tails $P(A \cup B)$

⇒ $0.5 + 0.5 = 1$

② Non Mutually exclusive

In a class of 30 students

18 students like math ⇒ $P(A) = 18/30$

12 students like science ⇒ $P(B) = 12/30$

5 like both $P(A \cap B) = 5/30$

⇒ $P(A \cup B) = 18/30 + 12/30 - 5/30 = 25/30 = 5/6$

⊛ Multiplication Rule of Probablity

Used to find the probablity of two events occuring together (i.e. the intersection of two events)

this formula depends on whether the events are dependent or independent

① Independent Events

two event A and B are independent, if the occurance of one does not affect the probablity of the other.

→ for independent events, the multiplication rule is $P(A \cap B) = P(A) \cdot P(B)$

eg tossing 2 coins ⇒ A getting Heads on first coin $P(A) = 0.5$

and B ⇒ getting heads on second coin $P(B) = 0.5$

∴ the outcome of two coins are independent

$$P(A \cap B) = P(A) \cdot P(B) = (0.5)^2 = 0.25$$

② Dependent Events

: 2 events are said to be dependent if the occurrence of one affects the probablity of the other

for Dependent events the multiplication rule is

$$P(A \cap B) = P(A) \cdot P(B/A)$$

Here $P(B/A)$ is the conditional prob of B given that A has occured.

eg Drawing cards from a deck without replacement

A ⇒ drawing a heart on first draw. $P(A) = 13/52$

B ⇒ drawing another heart on 2nd draw

After drawing 1 heart, there are only 12 hearts left out of 51 cards

$$P(B/A) = 12/51$$

$$\Rightarrow P(A \cap B) = P(A) \cdot P(B/A) = \frac{13}{52} \times \frac{12}{51} = \frac{1}{17}$$

Examples

① for Independent events

you roll a dice and flip a coin

A : Rolling a 6 ⇒ $P(A) = 1/6$

B : getting Heads ⇒ $P(B) = 1/2$

$\Rightarrow P(A \cap B) = 1/12$

② Deck of 52 cards (without replacement)

A = Draws a king : $4/52$

B: Draws another king : $3/51$

$$: \frac{4}{52} \times \frac{3}{51}$$

$$: \frac{3}{13 \times 51}$$

OR → Union

AND → Intersection

Permutations And Combinations

Permutations and combinations are fundamental concepts in probability and counting.

they help us calculate the nos of ways to arrange or select objects depending on whether order matters.

① Permutations (Order Matters)

Permutations count the arrangements of objects where the order is important.

formula : $P(n, r) = \frac{n!}{(n-r)!}$

where

$n \rightarrow$ total nos of objects

$r \rightarrow$ Nos of objects to arrange

$! \Rightarrow$ Factorial (eg $4! = 4 \times 3 \times 2 \times 1$)

eg How many ways can you arrange 3 letters from A, B, C, D?

Here $n = 4$, $r = 3$

$$P(4, 3) = \frac{4!}{(4-3)!} \quad \frac{4 \times 3 \times 2 \times 1}{1} = 24$$

So there are 24 arrangements ABC, ACB, BAC, BCA, ...

② Combinations (order doesn't matter)

Combinations count selections of objects where the order does not matter.

formula for combinations

$$C(n, r) = \frac{n!}{(n-r)!\, r!}$$

$n \Rightarrow$ total nos of objects

$r \Rightarrow$ Nos of object to choose

eg =

How many ways can you choose 3 letters from A, B, C, D where order doesn't matter

$n = 4, \; r = 3$

$$C(4,3) = \frac{4!}{1!\, 3!} = \frac{4 \times 3!}{3!} = 4$$

So total of 4 combinations

ABC, ABD, ACD, BCD

Definations

* factorial : there are $n!$ ways of arranging n distinct objects into an ordered sequence, permutations where $n = r$.

* Combination : the number of ways to choose a sample of r elements from a set of n distinct objects where order does not matter and replacements are not allowed.

⊛ Permutation : the nos of ways to choose a sample of r elements from a set of n distinct objects where order does matter and replacements are not allowed. when $n = r$ this reduces to $n!$. a simple factorial of n.

⊛ Combination Replacement : the nos of ways to choose a sample of r elements from a set of n distinct objects where order does not matter and replacements are allowed.

⊛ Permutation Replacement : the nos of ways to choose a sample of r elements from a set of n distinct objects where order does matter and replacements are allowed.

* n → Population r → subset of n or sample

Combinations formula :

$$C(n, r) = \frac{n!}{(n-r)!\, r!}$$

the formula shows us the number of ways a sample of "r" elements can be obtained from a larger set of "n"

Combination Problem!

* Choose 2 Prizes from a set of 6 Prizes

Q You won first place in a contest and are allowed to choose 2 prizes from a table that has 6 Prizes (1 to 6), How many diff combination of 2 prizes can you possibly choose.

→ $r = 2, n = 6$

$$6C_2 = \frac{6!}{4!\,2!} = \frac{6 \times 5}{2} = 15 \text{ ways}$$

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

* Choose 3 students from a class of 25 & she wants to find out how many unique teams of 3 can be created from her class of 25

$$25C_3 = \frac{25!}{23! \times 2!} = \frac{25 \times 24^{12}}{12} = 2300$$

* Choose 4 menu items from a menu of 18 items (Here order is not important just the combination matters)

$$18C_4 = \frac{18!}{14! \times 4!} = \frac{18 \times 17 \times 16^{2} \times 15^{5}}{4 \times 3 \times 2} = 18 \times 17 \times 10 = 3060$$

Handshake Problem

In a group of n people, how many diff handshakes are possible?

→ let's find the total handshakes possible, that is to say if each person shook hands once with every other person in group, what is the total nos of handshakes that occur?

⇒ each person in the group will make a total of n-1 handshakes, o.o other are n people, that would be n times (n-1) total handshakes.

In other words total nos of people multiplied by nos of handshakes that each can make will be the total handshakes.

A grp of 3 would make a total of 3(3-1) = 6.

each Person registers 2 handshakes with the other 2 people in grp. o.o total handshakes ⇒ n(n-1)

G this includes each handshake twice

1 with 2, 2 with 1, 1 with 3,
3 with 1, 2 with 3, 3 with 2

o.o Original Ques wants to find diff handshakes

⇒ $6/2 = 3$

Handshake Prob as a combination prob

We can solve Handshake Problem as a combinatn Problem as $C(n,2)$

$n \rightarrow$ nos of people in grp
$r \rightarrow 2$ (the nos of people involved in a diff handshake)

the order of the items chosen in the subset does not matter so for a group of 3 it will count 1 with 2, 1 with 3, 2 with 3

$$nC_2 = \frac{n!}{(n-2)!\,2!} = \frac{n(n-1)}{2}$$

(#) Sandwich Combination Problem

It's a classic math problem and asks something like how many sandwich combinations are possible? Calculate the possible sandwich combinations if we can choose 1 item from each of the 4 categories

- → 1 bread from 8 options
- → 1 meat from 5 options
- → 1 cheese from 5 options
- → 1 topping from 3 options

As the multiplication of number possible options in each category → $8 \times 5 \times 5 \times 3 = 600$ combinatn

In terms of combination each the nos of possible options for each category is equal to the nos of possible combinations for each category ∵ we are only making 1 selection. for eg C(8,1) = 8, C(5,1) = 5 and C(3,1) = 3

$$\Rightarrow {}^{8}C_1 \times {}^{5}C_1 \times {}^{5}C_1 \times {}^{3}C_1$$

$$8 \times 5 \times 5 \times 3 = 40 \times 15 = 600$$

Random Variables

Random Variable (X) converts outcomes of experiments to something measurable.

for eg : As a data analyst at a banks, we are trying to find out which of the applicants are more likely to default on their loan.

cust_no	Income	Loan_amt	Dependants	Is_default
1	10lakh	75lakh	3	Yes
2	15lakh	50lakh	2	No
3	20lakh	40lakh	1	No.

Now instead of processing Yes/No, It's easier to define a random variable X indicating whether the customer is predicted default or not.

the values will be assigned according to below rule

X = 1 ⇒ customer defaults

X = 0 ⇒ customer doesn't default

So data will be like

cust_no	Income	Loan_amt	Dependents	X (random var)
1	10 lakh	75	3	1
2	15 lakh	50	2	0
3	20 lakh	40	1	0

Now the table is more quantified, ie converted to nos now the data is entirely on quantative terms

Probablity Distribution

A Probablity distribution is a statistical func that describes all the possible values and likelihoods that likelyhood that a random variable can take within a given range. the range is bounded by the variable's minimum and maximum possible values but the specific shape of distribution depends on the type of variable and the context.

Key Concepts of Probablity Distribution

① Random Variable : A variable that can take on different values each with an associated Probablity.

→ Disvrete random variable : taken on a finite or countable set of values
eg (rolling a dice, numbers of heads in coin tossed).

→ Continious Random Variable : Can take any value within a range eg height, weight, temperature

② Probablity Distribution Function : (PDF)

2.1

↳ for desvrete variables, It is the probablity mass function (PMF) for continious variables It gives the probablity density.'

2.2 Cumulative Distribution Function (CDF)

the cumulative probablity that a random variable Is less than or equal to a particular value.

3. Types of Probablity Distributions.

⇒ Discrete Distributions

→ Binomial Distributon : Models the nos of successes in a fixed number of trials
eg flipping a coin

→ Poisson Distribution : Models the nos of occurance of an event in a fixed interval or time or space.

⇒ Continious Distributions

→ Normal distribution : the classic bell shaped curve, many natural phenomena follow this distributon

→ Uniform distributon : Every outcome in a range is eavually likely.

Note : the probablities of all outcomes sums up to 1.
for continious distributons, the area under the curve (total probablity) is

Eqs For Discrete Probablity Distributon!.
↳ tossing a fair dice
Random variable x outcomes (1,2,3,4,5,6

↳ For continious Probablity distributon
Heights of students on class, Modeled by Normal dist with Mean and SD

Expected Value

Defination: Random variable X can take values $x_1, x_2, x_3 \dots x_n$ with Probability such that

Expected value (EV) $= x_1 * P(X = x_1) + x_2 * P(X = x_2) + x_3 * P(X = x_3) + \dots + x_n * P(X = x_n)$

Note: Here $P(X = x_i)$ denotes the probability that X is equal to x_i

eg $EV = 0 * P(X=0) + 1 * P(X=1) + 2 * P(X=2) \dots$

the expected value should be interpreted as the "avg value" you get after the experiment has been conducted as infinite number of times

Q A casino game contains numbers 0 to 36 written in irrequlars sequence, the player can bet on any no from 0 to 36, for eg Kaiti bets £100 on nos 5, Now a ball would be dropped on a wheel which is then given a spin, if the ball lands on the Pocket marked 5, Kaiti would win (£100) x 36 = £3600 resulting in net winnings of £3600 - £100 = £3500, However if the ball lands on any other pocket → Loss of £100

let's see what is the expected value for Kriti net winnings if she plays this game and bet £100 on Number 5.

$$E(x) = \sum_{i} P(i)\, x_i$$

Here

$P_i \Rightarrow$ Probability of outcome i

$x_i \Rightarrow$ Net winning for outcome i

Step1 : Identify outcomes and probabilities

① there are 37 pockets on the wheel (nos 0 to 36)

② the probability of the ball landing on any specific no

$P(\text{Landing on 5}) = 1/37$, $P(\text{Not Landing on 5}) = \frac{36}{37}$

③ the net winning for each case

if ball lands on 5 $\Rightarrow$ 3500 $\Rightarrow$ (3600 - 100)

if ball lands on another no $\Rightarrow$ $-100 \rightarrow (0 - 100)$

Step2 : Apply formula

$EV(x) = P(\text{Landing on 5}) * X(\text{Landing on 5}) + P(\text{Landing on another no}) * X(\text{Landing on another no})$

$E(x) = \frac{1}{37}(3500) + \left(\frac{36}{37}\right)(-100)$

$\Rightarrow 94.59 - 97.30 = -2.71$

So the expected value of Kriti net winning is -2.71

Probablity Without Experiment

↳ for this we will use Binomial Distribution

the probablity of an event without conducting any experiment is determined through theoritical Probablity, which is based on reasoning and calculations rather than actual outcomes.

formula for theoritical Probablity :

$$P(E) = \frac{\text{Number of favourable outcomes}}{\text{total nos of Possible outcomes}}$$

Key points :

favourable outcomes : the outcomes that meet the condition of the event you are studying.

total outcomes : All the outcomes that can possibly occur in the given situation.

Eqs ① Rolling a fair die

→ total possible outcomes = 6

→ favourable outcome for rolling a 4 ⇒ 1

⇒ Probablity = 1/6

② Drawing a card from a standard deck

→ total Possible outcomes = 52

→ favourable outcome for drawing heart = 13

→ Prob = 13/52 = 1/4

Q A bag contains 5 red balls and 3 blue balls. If a ball is selected at random, what is the probability of

① Picking a red ball? ② Picking a blue ball?

Ans ① Identify the total nos of balls

total balls = $5+3 = 8$

② Calculating Probabilities

$P(Red) = 5/8$ | $P(Blue) = 3/8$

Q If we randomly pick two balls from the bag without replacement what is the probability that

① Both Balls are red?

② One is red & another is blue?

step ① use combination to calculate probability

total ways to select 2 balls from the bag:

$$^8C_2 = \frac{8!}{6!\,2!} = \frac{8\times 7}{2} = 28$$

Prob of selecting 2 red balls

5 ways to choose 2 reds out of 5 red

$$^5C_2 = \frac{5!}{3!\,2!} = \frac{5\times 4}{2} = 10$$

$$\Rightarrow \text{Prob}(2 \text{ red}) = \frac{10}{28} = \frac{5}{14}$$

Prob of 1 red ball and 1 blue ball

ways to choose 1 red ball from 5 red and 1 blue ball from 3

$$^5C_1 \times {}^3C_1 \quad \frac{5!}{4!\times 1!} \times \frac{3!}{2!\times 1!} = 15$$

$$\Rightarrow P(1 \text{ red } 1 \text{ blue}) = \frac{15}{28}$$

Q Bag Contains → 5 red balls
↳ 3 blue balls
} Balls are drawn with replacement

If we draw 4 balls what's the probability of getting exactly 3 red balls and 1 blue ball?

(Step1) Understanding Key points

→ Since the balls are drawn with replacement, the probability of drawing a red or blue ball remains constant for each draw.

$P(\text{Red}) = 5/8$ $\quad P(\text{Blue}) = 3/8$

→ Order doesn't matter, and sequence with 3 red ball and 1 blue ball count

⇒ for this problem we will use "Binomial formula for exact probability"

the Binomial formula :

It is used to calculate the probability of achieving a specific number of success (K) in a fixed no of independent trials (n), when the probability of success (P(success)) remains constant for each trial.

→ formula :

$$P(k) = {}^{n}C_{K} * P^{K} \cdot (1-P)^{n-K}$$

where :

n : total nos of trials (eg 4 draws in probl

K : Desired number of success (eg 3 red ba on problem)

P : Probability of success on a single trial

eg P(Red) = 5/8

1-P : Probability of failure on a single trial

eg (P(B) = 3/8)

${}^{n}C_{K}$: Binomial coefficient, which calculates the nos of ways to arrange K succes in n trials

Understanding Components

① Binomial Coefficient (nC_k) represents the nos of ways to choose K successes from n trials and it's denoted by

$$nC_k = \frac{n!}{(n-k)!\,k!}$$

It's important because in binomial distribution the order of success doesn't matter for eg
→ If you are drawing 4 balls and want 3 red balls, the arrangement RRRB, RBRR etc are all valid and counted in $4C_3$

② Probability of Success : p^k

this is the prob of getting K success in n trials for eg
→ Prob of drawing red ball is $P = 5/8$
→ If we want 3 red balls → this term becomes $(5/8)^3$ representing the likelyhood of getting a red ball 3 times.

③ Prob of failure $(1-P)^{n-k}$

this account for trials that don't result in success
→ Prob of drawing blue ball = $3/8$
→ if there is 1 blue ball $(n-k=1) \Rightarrow (3/8)^1$

Applying on our ques $(n=4,\ k=3 (red),\ P(R) = 5/8$
$1-P = 3/8)$

$$P(K=3) = {}^4C_3 \left(\frac{5}{8}\right)^3 \left(\frac{3}{8}\right)^1 \Rightarrow \frac{4!}{1!\,3!} \left(\frac{5}{8}\right)^3 \left(\frac{3}{8}\right)^1$$

$$\Rightarrow 4 \times 375/4098 = 375/1024 = 0.366$$

$P \Rightarrow$ Prob of choosing red ball

$1-P \Rightarrow$ Prob of choosing blue ball

⇒ Probability Distribution for General Probability P =

Now lets extend this case to the more genrex one

X	P(X=x)
0	$(1-P)^4$ → OR 4 Blue
1	$4P(1-P)^3$ → 1R 3B
2	$6P^2(1-P)^2$ → 2R 2B
3	$4P^3(1-P)$ → 3R 1B
4	P^4 → 4R 0B

$$P(X=r) = {}^nC_r (P)^r (1-P)^{n-r}$$

X	P(X=x)
0	$nC_0 (P)^0 (1-P)^n$
1	$nC_1 (P)^1 (1-P)^{n-1}$
2	$nC_2 (P)^2 (1-P)^{n-2}$
⋮	⋮
n	$nC_n (P)^n (1-P)^0$

Conditions for Binomial Probability Distribution

① total nos of trials is fixed at n

② Each trial is binary ie has only two possible outcomes (success or failure)

③ Prob of success is same in all trials

④ $P(X=r)$ = P(Getting r successes in n trials) = $nC_r (P)^r (1-P)^{n-r}$

Cumulative Probablity

Cumulative probablity is the probablity that a random variable takes a value less than or equal to a given value, In other words it is the sum of probabilities for all possible outcomes upto and including a specific outcome.

for binomial distribution, cumulative probability answers Ques like

→ what is the probablity of getting atmost 3 red balls?

↳ this would mean summing the Prob of getting 0, 1, 2, and 3 red balls.

(*) Cumulative Probablity in a Binomial Distribution

↳ for this we sum the probablities for each k (number of success) from 0 upto a target value k.

$$P(x \leq K) = \sum_{i=0}^{K} \binom{n}{i} p^i (1-p)^{n-i}$$

where $n \Rightarrow$ total trials

$P \Rightarrow$ Probability of succes

$K \Rightarrow$ upper limit for succes

$\binom{n}{i} \Rightarrow$ Binomial Coefficient for i successes in n trials.

Example cumulative prob for Atmost 3 Red balls

Problem: Using the same scenario of 5 red balls and 3 blue balls, if we draw 4 balls with replacement, what is the probablity of getting at most 3 red balls.

Given: $n = 4$ (4 draws)

$P = 5/8$ (Prob of red)

$1 - P = 3/8$ (Prob of blue)

what we want

$$P(X <= 3) = P(X=0) + P(X=1) + P(X=2) + P(X=3)$$

We calculate the prob for each K using binomial formula.

1) $\text{Prob}(X=0) = {}^4C_0 (5/8)^0 (3/8)^{4-0}$

$$1 \times 1 \times \frac{3 \times 3 \times 3 \times 3}{8 \times 8 \times 8 \times 8} = \frac{81}{4096}$$

2) $\text{Prob}(X=1) = {}^4C_1 (5/8)^1 (3/8)^3 = \frac{540}{4096}$

3) $\text{Prob}(X=2) = \frac{1350}{4096}$

4) $\text{Prob}(X=3) = \frac{1500}{4096}$

$\therefore P(X \leq 3) = \frac{81 + 540 + 1350 + 1500}{4096} = \frac{3471}{4096} = 0.84$

✳ Understanding PDF and CDF

When dealing with random variables, Probability Density functions (PDFs) and Cumulative Distribution functions are essential tools to describe the behaviour of continious random variables.

① What is a PDF (Probability Density Func)?

⇛ Imagine you are studying the daily commute times of people, the commute time vary continiously like someone's commute could be 15.4 minutes, 20.8 minutes, or even 30.1 minutes

→ the PDF is a func that shows how likely it is for a random variable (like commute time) to be in a small range of values.

→ For eg the PDF tells us how likely it is for a commute to be in between 20 and 25 minutes

Key Idea

the area under the curve of the PDF blw two values (say 20 and 25) gives the prob of the random variable falling within that range.

Why can't we find the probablity at an exact value

For continious data the probablity of a single exact value (eg P(X=20)) is 0 ∵ there are infinitely many possible values for a continious var

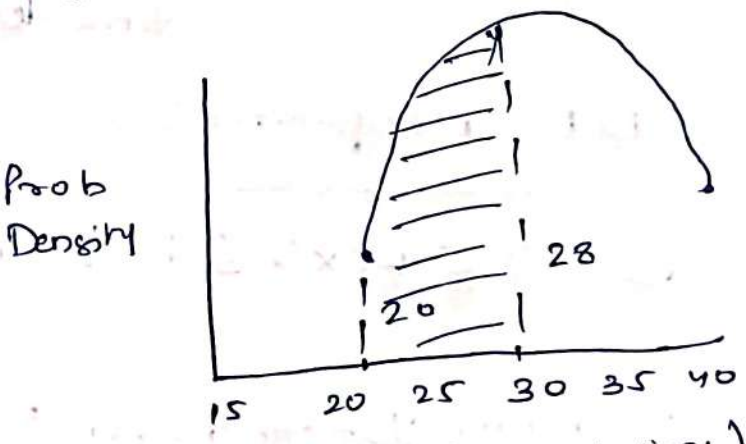
and the probablity gets spread out over a range Instead we will focus on Intervals like $P(20 \leq x \leq 25)$

PDF Properties

① The PDF is non-negative : $f(x) \geq 0$ for all x

this means the graph of PDF is always above horizontal axis.

Probability Density func is a function in which area under the curve gives us the cumulative Probability.



for eg the Area under the curve b/w 20 the smallest Possible value of x and 28 gives the cumulative Prob of $x = 28$.

Using PDF we can calculate the probability like $P(25 \leq x \leq 35)$: the prob that a commute is b/w 25 and 35 minutes

✳ What is CDF (Cumulative Distribution func)

the CDF is the cumulative probability upto a certain value of x, It tells us the prob that the random variable X is less than or equal to x.

$$F(x) = P(X \leq x)$$

How does CDF work?

The CDF is calculated by adding up all the probabilities from the smallest possible value upto x.

$$F(x) = \int_{-\infty}^{x} f(t)\,dt$$

It is essentially the area under the PDF curve from the leftmost side (usually $-\infty$) to x.

Eg $F(35) = .85$ ⇒ Means 85% of people have commute time less than or equal to 35 min.

CDF Properties

① $0 \le F(x) \le 1$: Cumulative prob grows from 0 to 1

② Non Decreasing : The CDF either stays constant or increases as x increases. It can't decrease because probability only accumulates.

③ Boundary Values

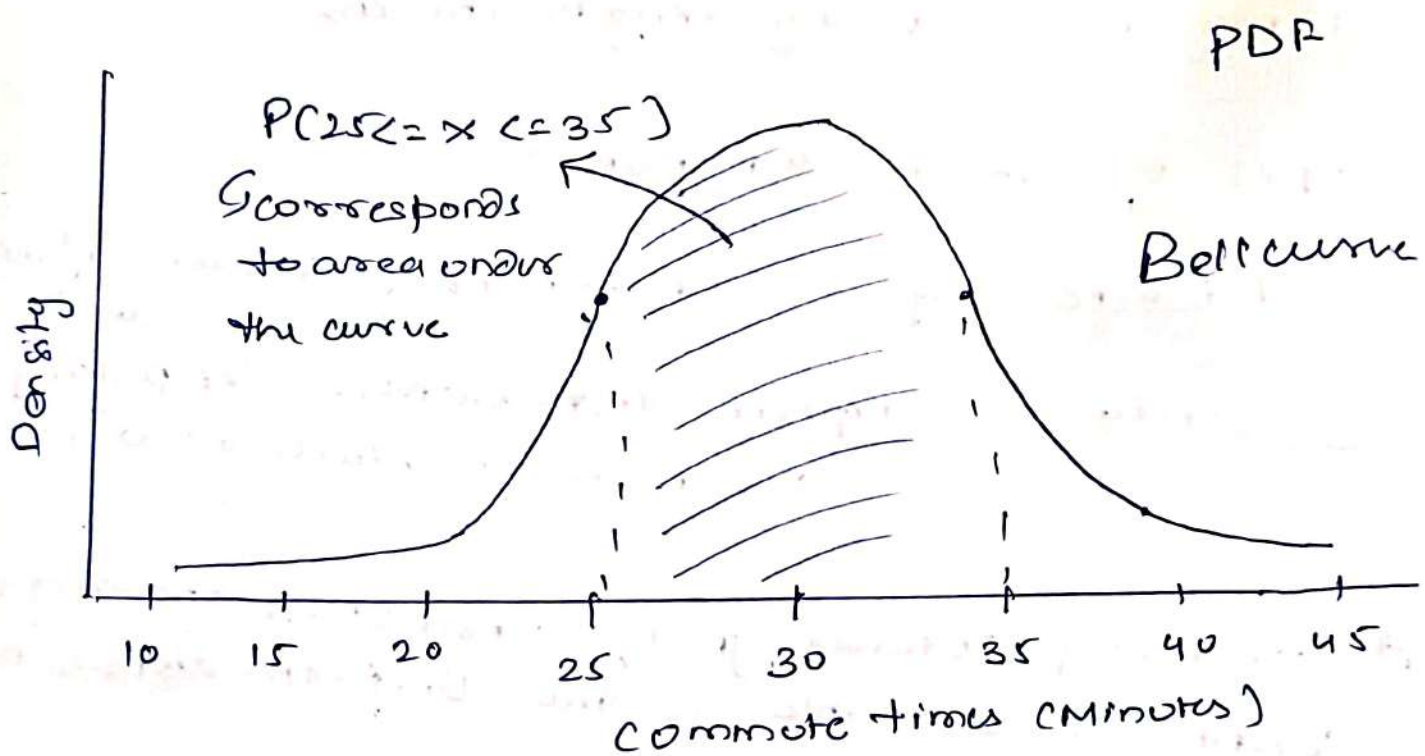
→ $F(x) \to 0$ as $x \to -\infty$: the prob of very small values is close to 0.

→ $F(x) \to 1$ as $x \to \infty$: the cumulative prob eventually reaches 1 as we include all possible values.

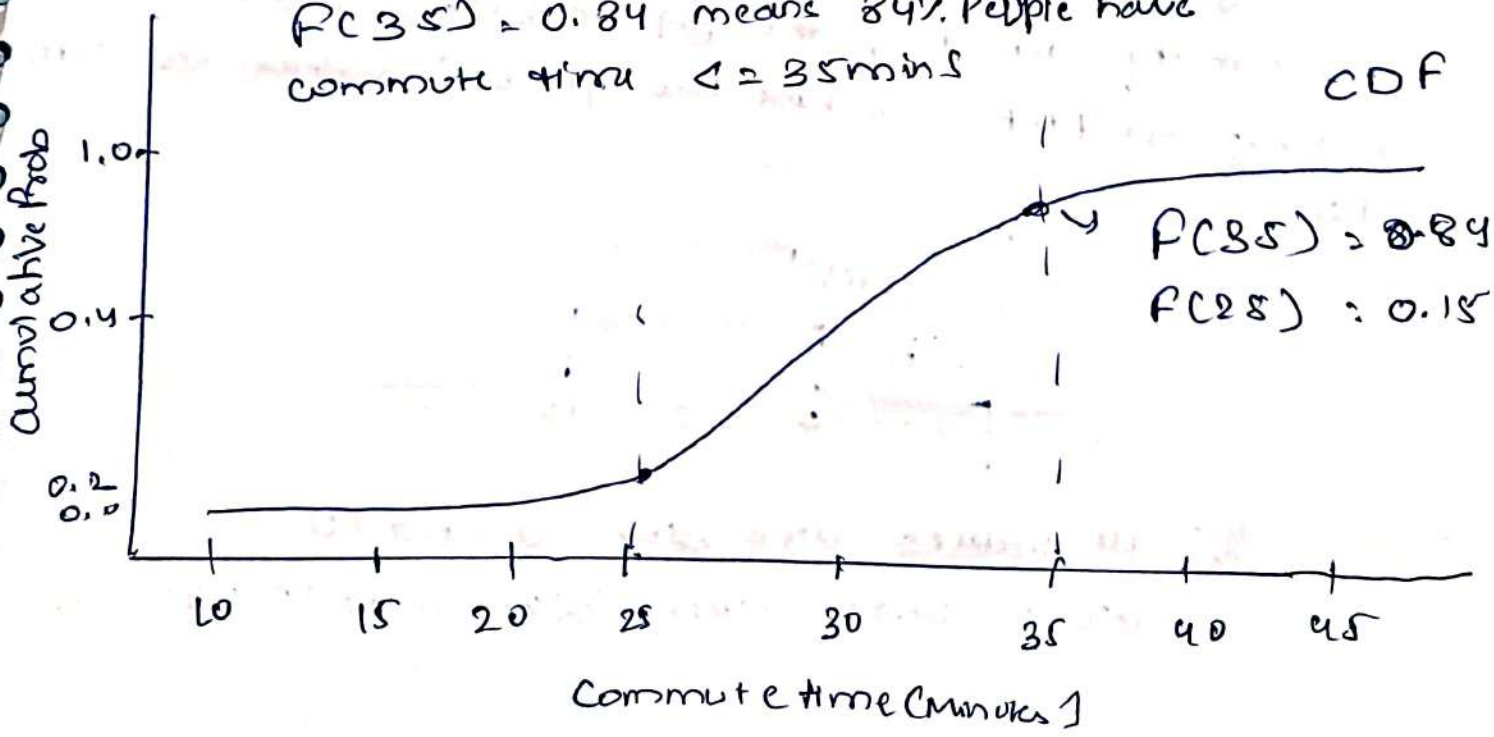
Visualizing PDF and CDF

① PDF curve showing the likelyhood of diff commute times.

② CDF curve showing cumulative probablity



F(35) = 0.84 means 84% People have commute time <= 35mins



Uniform Distribution

In Uniform distribution, every outcome has same chance of happening, you can think of it like drawing a number from a hat where all nos are equally likely to chosen.

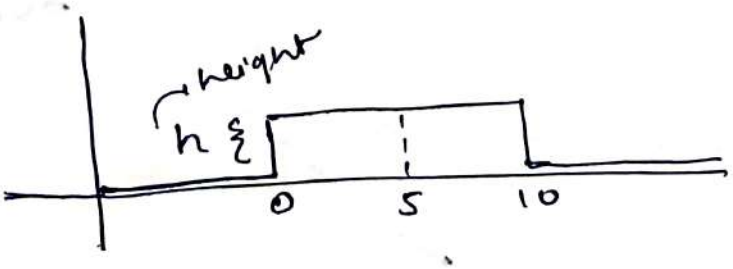
types of Uniform distribution

① Discrete Uniform distribution (like rolling a fair die)

② Continious Uniform distribution (like picking a random nos b/w 0 to 10)

(*) Commonly observed type of Distribution among Continious variable is the Uniform distribution

In Uniform PDF, all possible values have same Probablity density, the figure below shows such a uniform PDF, where the possible values are 0 to 10.



∴ all values are b/w 0 and 10
area under curve b/w 0 and 10 is 1

Clearly this area is the area of rectangle with length 10 and unknown height h

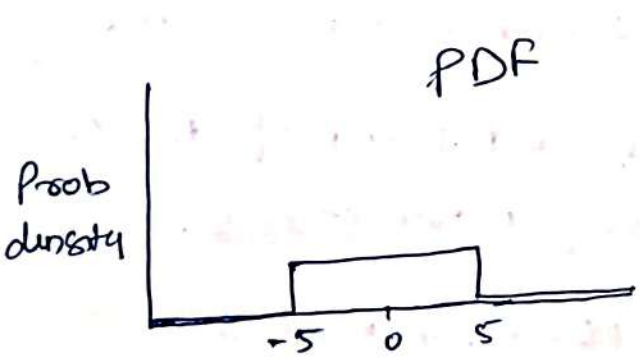
$\Rightarrow 10 \times h = 1 \quad \Rightarrow h = 1/10 = \underline{0.1}$

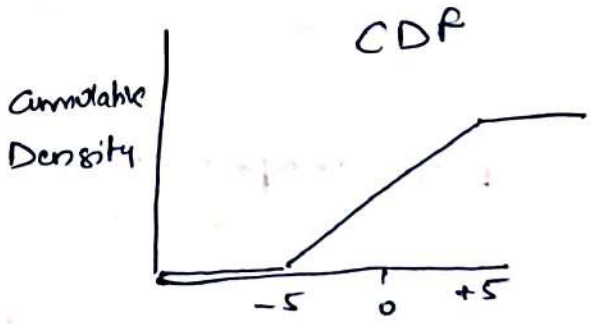
$\because l \times b = \text{Area} \quad \therefore$ PDF for all values b/w 0 to 10 is $\underline{0.1}$

Now cumulative prob for $X = 0.5$

$0.1 \times \underbrace{0.5}_{\text{length}} = 0.05$
$\quad\; h$

Note: PDFs are more commonly used in real life, the reason is that it is much easier to see patterns in PDFs as compared to CDFs





PDF shows uniformity as the probablity density value remains constant for all possible values. whose CDF doesn't show any trends that shows a variable is uniformly distributed.

(*) Normal Distribution

Normal distribution also called as bell curve or Gaussian distribution is one of the most common distribution, It represents a situation where data is symmetrically distributed around the mean, meaning most of the values cluster around the avg and fewer values are found as you move away from it.

Key features of Normal distribution

→ Symmetry : the curve is perfectly symmetric around the mean (μ)

→ Mean = Median = Mode : these 3 measures of central tendency are same

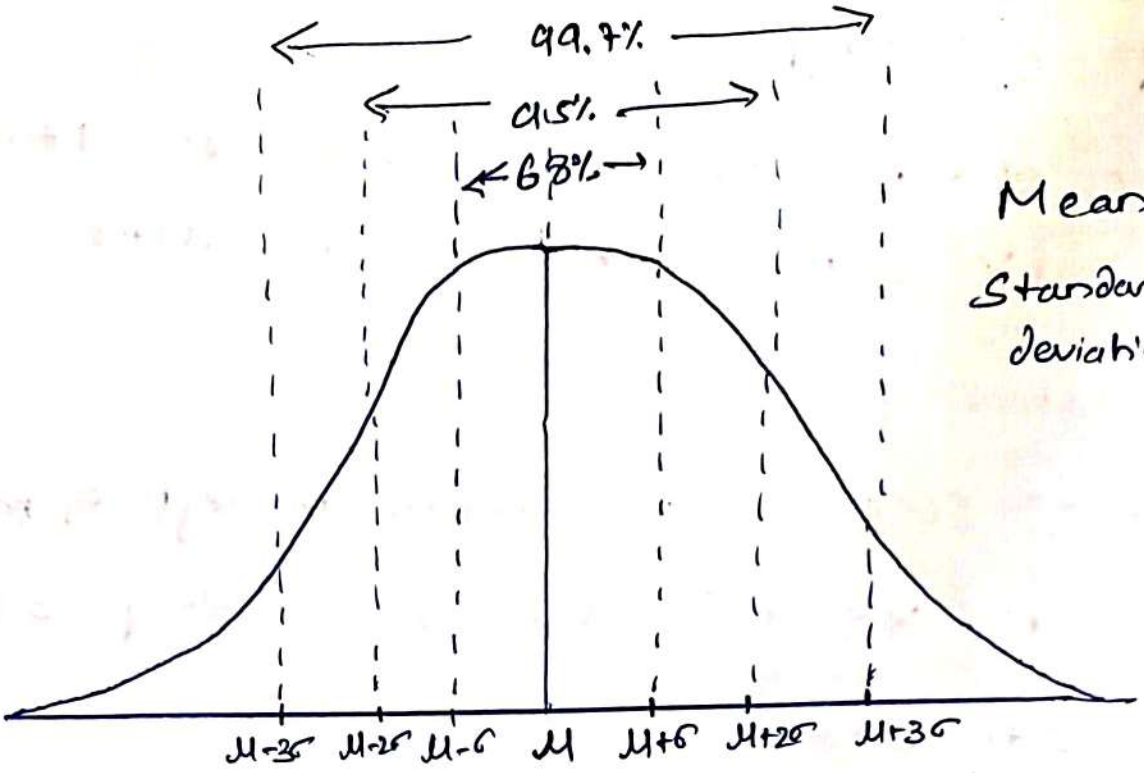
→ Bell shaped curve : Highest point of the curve occurs at the mean.

→ Empirical Rule (68-95-99.7 Rule)

① 68% of data lies within 1st standard deviation (σ) of the mean.

② 95% of data lies within 2 standard deviations

③ 99.7% of data lies within 3 standard deviations



Eg of Normal distribution

Scenario

↳ let's say the heights of adult males in a country are normally distributed with
mean (μ) of 170 cm
standard deviation σ of 10 cm

} Most Men are around 170 cm tall.
Fewer Men are shorter than 140 cm ($\mu-3\sigma$) and longer than 200 cm ($\mu+3\sigma$)

Q Let's say we need to find the cumulative prob for a random variable x which is normally distributed, we do not know what the value of X is or, for that matter, what is the value of μ and σ is we only know that $x = \mu+\sigma$
Can you find cumulative prob ie prob of var being less than $\mu+\sigma$?

Given: random var X is normally distributed
we know $X = \mu + \sigma$ we need to find $P(x \leq \mu+\sigma)$

Step by step soln

→ 68% of data lies b/w $\mu-\sigma$ to $\mu+\sigma$

⇒ 34% data lies b/w μ and $\mu+\sigma$

Cumulative prob

⇒ 50% of data lies to the left of mean

$P(x \leq \mu) = 50\%$) ∴ adding 34% from

μ to $\mu+\sigma$

⇒ 84%.

⇒ $P(x <= \mu+\sigma)$ is 84%

⊛ Standard Normal Variable (Z score)

A Z score (also called as standard score) tells us how many standard deviations a data point is from the mean of the dataset in a normal distributions, It allows us to standardize different datasets so we can compare them or calculate probabilities.

Key Concepts:

① Position Relative to mean

→ A Z score measures the position of a value x relative to the mean μ of dataset

→ A +ve z score means X is above mean.

→ A -ve z score means X is below mean

→ A z score of 0 means X is equal to mean

② Scaling by Standard deviation

The z score considers how far a value is from the mean, scaled by standard deviation (σ) this ensures that the comparison is not biased by spread of data.

③ Standard Normal Distribution

→ By converting values to z score any normal distribution can be transformed into standard normal distribution. (mean=0, standard dev = 1)

How z scores work ?

→ Let's say we have students dataset

Mean height μ = 170 cm

standard deviation (σ) = 10 cm

If student height is 180cm, z score tells us how many standard deviations 180 is above the mean

Interpreting z scores :

z = 0 (exactly at mean) | z = 1 (1 standard dev above mean)

z = -1 (1 deviation below mean) z = 2 (2 dev above mean)

Eg Comparing heights

Dataset : Heights of men → $\mu = 170$, $\sigma = 10$

A man's height is $X = 180$

$$Z \text{ score} = \frac{\text{Value} - \text{Mean}}{\text{Standard dev}} = \frac{180 - 170}{10} = 1$$

⇒ $X = 180$ is 1 standard deviation above the mean

if height were $X = 160$

$$Z = \frac{160 - 170}{10} = -1$$

We can refer to Z table for Probability using Z score

eg 1.65 → row will be 1.60

and col will be 0.05

Q regulatory authority selects a random tablet from a batch, based on Prev Knowledge you know that batch Z2 has

mean → 510mg std dev = 20mg

Prob that tablet select by authority has Paracetamol level < 550mg

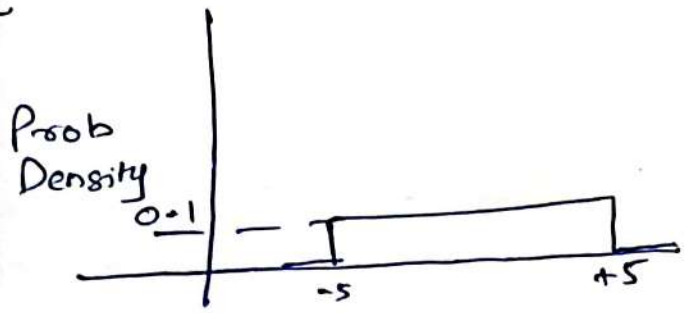
$$\frac{x_i - \mu}{\sigma} = \frac{550 - 510}{20} \quad (2)$$

$$\Rightarrow Z = 2$$

$\mu + 2\sigma$

from Z table $P(Z < 2) = 0.977$

Q



this graph represents the PDF of a uniformly distributed random variable X.

As we can see the prob density is equal for all the possible values of X (-5 to +5)

what is the prob of the random variable X lying b/w -1.5 to +2.5 i.e. $P(-1.5 < X < 2.5)$

Ans: the prob of the variable lying b/w -1.5 and 2.5 would be equal to the area under the PDF

→ Calculating the area of rect from (-1.5 to 2.5) and of height (0.1) ⇒ $(0.1)(2.5 - (-1.5))$

$= 0.1(2.5 + 1.5) = 0.1(4)$

⇒ 0.4

Q An astronomer's error in estimating the dist b/w earth and uranus follows a normal distribution with $\mu = 0$ km and standard deviation $\sigma = 1000$ km

what's the prob that the error is

① overestimating by 2330 km or more

② within ±500 km

Soln ① we need to find $P(X \geq 2330)$

$Z = 2330 - 0 / 1000 = 2.33$

Now using Z table

$P(Z \geq 2.33) = 1 - P(Z \leq 2.33)$

$\Rightarrow \quad 1 - (0.9901) = 0.0099 = \text{or } 0.99\%$

(2) $P(-500 \leq x \leq +500)$

for $x = -500 \Rightarrow Z = \frac{-500-0}{1000} = -0.5$

for $x = +500 \Rightarrow Z = +0.5$

$P(Z \leq 0.5) = 0.6915 \qquad P(Z \leq -0.5) = 0.3085$

$P(-0.5 \leq Z \leq 0.5) = P(Z < 0.5) - P(Z < -0.5)$

$= 0.6915 - 0.3085$

$= 0.3830 \Rightarrow 38\%$

(*) Central limit theorem

Sample : In statistics, a sample is a subset of individuals, items or observations taken from a larger population to study and make conclusion about the population. ∵ studying an entire population is often impractical.

Key points: A sample should represent the population well to ensure that inferences made from the samples are valid.

→ Bias in sampling can lead to inaccurate results

⊛ Notations and formulae related to population and samples

Population/Sample	Term	Notation	Formula
$(X_1, X_2, X_3 \dots, X_n)$ Population	Population Size	N	Nos of items/ele in population
	Population Mean	μ	$\sum_{i=1}^{N} \frac{X_i}{N}$
	Population Variance	σ^2	$\sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N}$
Sample. $(X_1, X_2, X_3 \dots X_n)$ sample of Population	Sample Size	n	Nos of items/ele in the sample
	sample mean	$\bar{X}$	$\sum_{i=1}^{n} \frac{X_i}{n}$
	Sample Variance	s^2	$\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$

Now suppose we conducted a survey and got that 50.5% of sample are preferring option 2 over option1, is it fair to conclude that 50.5% of the people in population will also Prefer option2 over option?

↳ Write now can't say....

Sample Distribution

A sampling distribution is the probability distribution of a statistic (like the sample mean, proportion, variance etc) calculated from multiple random samples of the samples of the same size drawn from a population.

Eqs and Visualizations

Imagine we have a population of 10 students with the following test scores

Population : 50, 55, 60, 65, 70, 75, 80, 85, 90, 95

↳ Population Mean (μ) ⇒ 72.5

Step1 : take random samples

Suppose you take random samples of size 2 ($n = 2$) repeatedly from this population and Calculate the sample mean ($\bar{x}$) for each sample

→ Here are some possible samples and their means

Sample1 {50, 55} → Mean ⇒ (50+55)/2 = 52.5

Sample2 {65, 70} → Mean = 67.5

Sample3 {85, 95} → Mean = 90

Sample4 {60, 80} → Mean = 70

Similarly many samples can be withdrawn

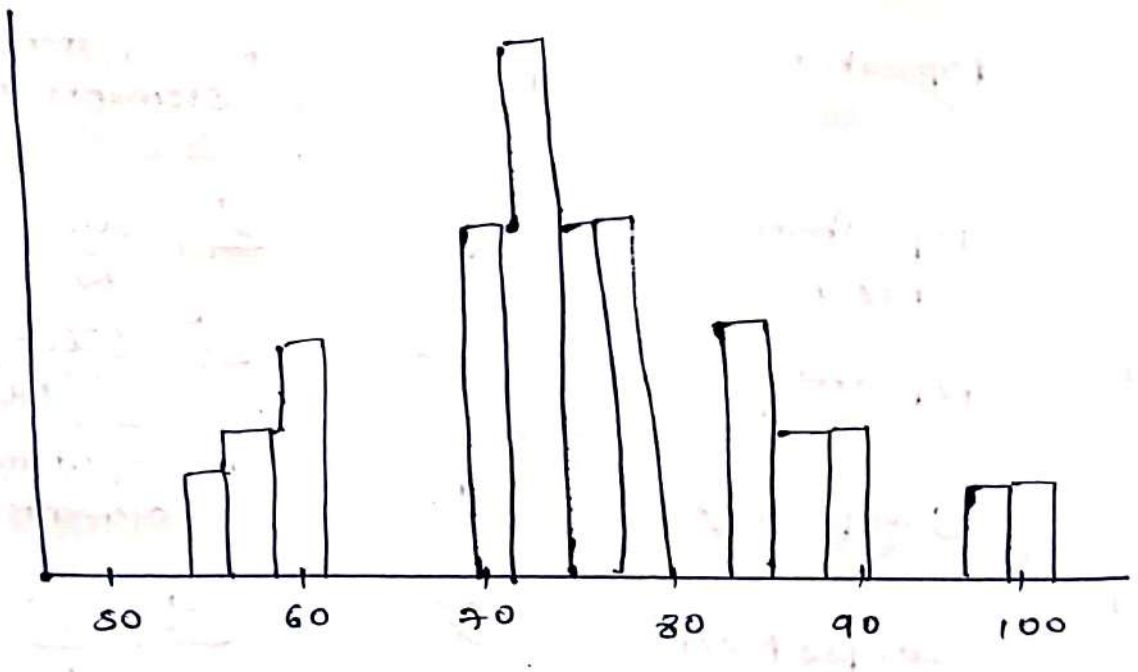
Step2: Create a sample distribution

If we list the means from all possible samples of size 2 we get the sampling distribution of the sample mean.

the sampling dist will look like a histogram showing freq of each sample mean.

Visualization

Sampling Distribution of Sample Mean
(Sample Size = 2)



Observations:

→ the sample means are centered around the Population mean ($\mu = 72.5$)

→ the distribution is symmetric because population itself is symmetric

→ As the sample size increases the spread (variability) of sampling distribution decreases making it more concentrated around mean.

* The standard deviation of the distribution of the sampling means is given by

$$\text{Sampling distribution's Standard deviation} = \frac{\sigma}{\sqrt{n}}$$

(σ → Population SD; n → sample size)

$=$

the higher the value of the standard deviation the fatter the curve will be, hence the values will be farther apart.

* Notations and formula related to sampling Distribution

Population/ sample	Term	Notation	formula
Population $(X_1, X_2, X_3 \ldots X_n)$	Population Size	N	Nos of items/ elements in population
	Population Mean	μ	$\sum_{i=1}^{N} \frac{x_i}{N}$
	Population variance	σ^2	$\sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N}$
Sample $(x_1, x_2 \ldots x_n)$ (Sample of Population)	Sample Size	n	Nos of items/ elements in sample
	Sample Mean	$\bar{X}$	$\sum_{i=1}^{n} \frac{x_i}{n}$
	Sample Variance	s^2	$\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$
Sampling distribution of sample mean $(\bar{X}_1, \bar{X}_2, \ldots \bar{X}_K)$ K sample Means	Sampling distribution Mean	$\mu_{\bar{x}}$	$\mu_{\bar{x}} = \mu$
	Sampling dist Standard dev	S.E (Standard error)	$\sigma/\sqrt{n}$

Central Limit Theorem (CLT)

It's a fundamental concept in statistics that explains how the sampling distribution of the sample mean approaches a normal distribution as the sample size increases, regardless of the original population distribution.

Keypoints of CLT

① Sample Mean approximates population Mean: the mean of sampling distribution of $\bar{x}$ is equivalent to the population mean (μ)

② Standard error ($\sigma_{\bar{x}}$): the standard deviation of the sampling distribution (called as standard error) is given by $\sigma_{\bar{x}} = \sigma/\sqrt{n}$

Where
$\sigma \rightarrow$ Standard Deviation of Population
$n \rightarrow$ sample size

③ Normality of large sample sizes: for $n > 30$ the sampling distribution of sample mean becomes approximately normal, even if the population distribution is not normal

⇒ CLT helps us make predictions and perform statistical tests, even if the original data is not normally distributed

Every day examples

If we measure the avg height of 30 students in different classrooms (multiple samples)

→ their avg heights will
center around population mean (true avg height of all students)

→ form a normal distribution even if individual heights don't follow normal distribution

So for an office of 30000 employees, we wanted to find the avg commute time
So instead of asking all, we can ask only 100
So for sample size of 100 → Mean ($\bar{x}$) = 36.6
Standard dev. (s) = 10

However it would not be fair to conclude the same for total population because the flaws of sampling process must have lead to some error
Hence sample mean's value has to be reported with some margin of error.

36.6 ± marginal error

Practice Ques

→ Exit Poll for MCD, we asked 100 randomly selected voters to name the party they voted for

Data Collected is

Contesting Party	No. of Voters
BJP	58
INC	42

from this sample we have to estimate the % of voters that might have voted for BJP

So we defined X as the proportion of people that voted for BJP

X	freq
1	58
0	42

} frequency Distribution of X

Mean of X which is equal to (0+0+0 ..+42 times and (1+1+1 ... 58 times)

$$\text{Mean} = \frac{(0 \times 42) + (1 \times 58)}{100} = 0.58$$

Mean in terms of X (voted for BJP)

Standard Deviation (σx) $\frac{(x_i - \mu)^2}{N-1}$

$$\Rightarrow \frac{((0 - 0.58)^2 \times 42) + ((1 - 0.58)^2 \times 58)}{100 - 1}$$

$$\Rightarrow 0.2461$$

Q We define Y (Proportion of people voted for Congress

	Y	votes
Freq Distribution	0	58
	1	42

Mean = 0.42

$$\sigma^2 : \frac{(0 - 0.42)^2 \times 58 + (1 - 0.42)^2 \times 42}{100 - 1}$$

$$\Rightarrow \frac{(-0.42)^2 \times 58 + (0.58)^2 \times 42}{99}$$

$$\Rightarrow \frac{0.1764 \times 58 + 0.3364 \times 42}{99}$$

$$\Rightarrow \frac{10.231 + 14.128}{99} = 0.2460$$

$$\Rightarrow SD = \sqrt{\sigma^2} : \sqrt{0.2460} = 0.4959$$

$\Rightarrow 49.59\% = 49.6\%$

Q3 Voter Sample

let's say we have a sample distribution of Y the proportion of people that voted for INC

Mean of sample distribution is $\mu_{\bar{x}} = 0.50$

SE $(\sigma/\sqrt{n})$ is 0.048

to find $\sigma \Rightarrow$ for population

$$0.048 = \frac{\sigma}{\sqrt{n}} \qquad n = 100$$

$$\sigma = 0.048 \times \sqrt{100} = 0.48 \Rightarrow 48\%$$

(*) Estimating Mean using CLT

Earlier we tried to estimate the mean commute time of 30,000 employees of an office by taking a small sample of 100 employees and finding their mean commute time,

the sample mean was $\bar{x} = 36.6$ minutes &
standard deviation was $s = 10$ minutes.

Recall that we also said about the population mean ie daily commute time of all 30,000 employees

$\mu = 36.6$ (sample mean) $\pm$ some margin of error

We can find the margin of error using CLT.

⇒ To find the margin of error using the Central Limit theorem (CLT) let's go step by step

Given Data:

① Sample mean ($\bar{x}$) = 36.6 minutes

② Sample standard deviation (s) = 10 minutes

③ Sample size (n) = 100

④ the population mean (μ) is approximately

$\bar{X} \pm$ margin of error

Step 1: Standard error of Mean

By Central limit theorem

$$SE = \frac{S}{\sqrt{n}} = \frac{10}{\sqrt{100}} = 1$$

∴ Sample size 100 > 30 so the sampling distribution is a normal distribution.

But the Mean is unknown.

let's see $P(\mu - 2 < 36.6 < \mu + 2) = ?$

⇒ 0.954 (by normal distribution rule 68, 95, 99.7)

$P(36.6 - 2 < \mu < 36.6 + 2) = P(\mu - 2 < 36.6 < \mu + 2)$

= 95%

$P(36.6 - 2 < \mu < 36.6 + 2) = 95.4\%$

Prob associated with the claim is called confidence level (Here its 95.4%.)

Maximum error made in sample mean is called Marginal error. (Here it is 2mins)

→ Final interval of values is called confidence interval (34.6, 38.6)

Step by step approach using Z score

Given data

① sample mean ($\bar{x}$) = 36.6 minutes

② sample standard deviation (s) = 10 minutes

③ sample size (n) = 100

→ SE (standard error) = $s/\sqrt{n} = 10/\sqrt{100} = 1$

⇒ SE = 1

Margin of error is calculated as

Margin of error = Z . SE

o.s sample data is normally distributed

Based on CLT

$P(\mu - 2 < \mu < \mu + 2) = 0.95$

Z for 95% value is 1.96

⇒ Margin of error = $1.96 \times 1 = 1.96$

$\vec{\mu} = \bar{x} \pm$ margin of error

$\Rightarrow 36.6 \pm 1.96 \Rightarrow (34.64 - 38.56)$

* Now we can say that for y% confidence level the confidence interval is given by

$$\left(\bar{x} - \underbrace{z \times \frac{s}{\sqrt{n}}}_{\text{z * (SE)}} , \bar{x} + z \times \frac{s}{\sqrt{n}} \right)$$

↳ $\sigma/\sqrt{n}$ or $s/\sqrt{n}$ for sample

Q Maggie Noodles example

↳ Max Permissible lead in products 2.5 PPM

↳ Sample size = 100

Sample Mean ($\bar{x}$)
↳ for lead content = 2.3 PPM

Standard dev (s) = 0.3 PPM

Now to find the confidence interval

↳ Confidence level ⇒ ∵ it's a sensitive test
let's take it as 99%. z (99%) = 2.57

$\Rightarrow$ Confidence Interval ⇒ $2.3 \pm \frac{(2.576)(0.3)}{\sqrt{100}}$

$\Rightarrow 2.3 \pm \frac{2.576(0.3)}{10}$

$(2.223, 2.377)$ ✓

Q Amt of Paracetamol specified by drug regulatory authority → 500mg

→ allower error ⇒ 10%.

Below 450mg → Quality issue & abave 550mg would be serious regulatory issue

→ Given sample of 100 tablets
→ Mean Paracetamol ⇒ 530mg
→ Standard dw : 100mg
→ CI : 95%.

CI : $\mu \pm Z \cdot SE$

Z score for 0.95 (95%.)

Note : 95% of confidence level implies that 95% of the data lies within the confidence interval

This leaves 2.5% in each tail split equally

2.5% on the lower side and 2.5% on the upper side.

to determine the upper bound z score, we need the cumulative prob uptill that point

(ie the central tendency (95%)
+
the lower tail (2.5%)

Hence the cumulative probability becomes

$95\% + 2.5\% = 0.975$

↳ Hence Z score for $0.975 = 1.96$

so margin of error for 95% Confidence level

$Z * SE \Rightarrow (1.96) \times \frac{\sigma}{\sqrt{n}} = 1.96 \times \frac{100}{\sqrt{100}}$

$\frac{1.96 \times 100}{10} = 19.6$

⇒ Confidence Interval for 95% confidence level

$530 \pm 19.6 \Rightarrow (510.4, 549.6)$

⊛ Confidence Interval for 90% confidence level

90% Confidence level Mean 5% left on each tail

⇒ Z for 90% + 5% ⇒ for 0.95 is 1.64

⇒ $Z \times SE = (1.64)\left(\frac{100}{\sqrt{100}}\right) = \frac{1.64 \times 100}{10}$

16.4

$530 \pm 16.4 \quad (513.6, 546.4)$

✳ Hypothesis Testing

Let's understand the basic difference b/w Inferential statistics and hypothesis testing.

Inferential Statistics

Purpose : to make generalizations about a population based on data collected from a sample.

Key concept : The population mean (or another parameter) is unknown, so we estimate it using :

→ Sample statistics : such as sample mean ($\bar{x}$)

→ Confidence Intervals : these provide a range within which the population parameter is likely to fall, along with confidence level (eg 95% confidence interval)

Use case : When you don't have prior knowledge about the population and need to draw conclusions from it

Hypothesis Testing

⇒ Purpose : To test an assumption or claim (hypothesis) about a population parameter using sample data. Hypothesis testing is used to confirm your conclusions (or hypothesis) about the population parameter (which you know from EDA or your intution).
through hypothesis testing, you can determine whether there is enough evidence to conclude if the hypothesis about the population parameter is true or not.

to understand about Hypothesis, let's take an eg

" A criminal trial → A person has been charged with crime, Jury has to decide whether the dependent is innocent or guilty.

Hypothesis 1 : Defendent is innocent

Hypothesis 2 : Defendent is not innocent (guilty)

these two opposing hypothesis are called as the "null hypothesis" and the "alternate hypothesis"

eg of food product containing excess lead

⇒ Null Hypothesis

Avg lead content is within the allowed limit of 2.5PPM

⇒ Alternate Hypothesis

Avg lead content is more than 2.5PPM

In our eg of criminal trial

Null hypothesis (H_0) (Notation) ⇒ Dependent is innocent

Alternate Hypothesis (H_1) ⇒ Dependent is not innocent

If the Dependant is found Guilty ⇒ there is a "Rejection of Null Hypothesis"

Key concepts:

→ you start with null hypothesis (H_0) → a statement of no effect or no difference.

→ the alternate hypothesis (H_1) is what you aim to support

through hypothesis testing, you can check whether the sample data provides enough evidence to reject H_0 in favour of H_1

In Maggie Noodles eg

↳ Null Hypothesis (Ho) ⇒ avg lead content is less than or equal to 2.5 PPM.

If we fail to reject null hypothesis, we can conclude that Maggie Noodles doesn't contain excess lead.

Note: In hypothesis testing we never "accept" the null hypothesis we can only fail to reject it because

→ the null hypothesis (Ho) represents a default assumption such as there is no effect or the means are equal.

⇒ Failing to reject Ho doesn't ~~prove~~ prove it is true → It simply means the sample data does not provide Proper evidence against Ho

→ Lack of evidence is not equivalent to evidence of absence.

Just like the court of law, failing to reject doesn't mean innocence (acceptance) it simply means there wasn't enough evidence to prove guilty (reject Ho)

Both Null and alternate hypothesis can't be true at the same time, only one of them is true

Key Rule for formulating Hypothesis

Null Hypothesis (H_0):

→ Represents a statement of no difference, no effect or a baseline assumption.

→ Always includes the signs:

$=$, $\leq$ or $\geq$

↳ Because hypothesis testing seeks evidence to reject H_0 and these signs allows us to test for deviations.

Alternative Hypothesis (H_1):

→ Represents what we aim to support or find the evidence for.

→ Always include the signs:

$\neq$, $>$ or $<$

Because H_1 is formulated to oppose H_0.

Guidelines for the Claims

→ Claim with "atleast"

eg the mean income is atleast $50000

→ H_0 : $\mu \geq 50,000$

H_1 : $\mu < 50,000$

→ Claim with "atmost"
eg defect rate is atmost 5%.

H_0 : $P \leq 0.05$
H_1 : $P > 0.05$

→ Claim with "greater than"
eg the new drug is more effective than older.

H_0 : $\mu_{new} \leq \mu_{old}$
H_1 : $\mu_{new} > \mu_{old}$

→ Claim with equals
eg Mean life span of product is 500 hrs

$H_0 \Rightarrow \mu = 500$
$H_1 \Rightarrow \mu \neq 500$

For eg Flipkart claimed that it's total valuation in December 2016 was atleast $ 14 billion, Here the claim contains $\geq$ sign, so the null hypothesis is the original claim.

Total valuation $\geq \$14$ Billion → Null Hypothesis
Total valuation $< \$14$ Billion → Alternate Hypothesis

Making a decision

eg: Apurva's claim : Avg score in archery is equal to 70.

Over 5 games of Archery :

↳ Avg score of Apurva = 20 (Less likely to believe her claim)

↳ Avg score = 65 → More likely to believe her claim.

So we reject her claim when her avg score was 20 but fail to reject if her avg score is 65 → so what score b/w 20 and 65 is the boundary where we decide to change our opinion.

↳ this point is called as Critical point.

(#) Critical Region : It's like a danger zone for the null hypothesis (H0) in hypothesis testing, It's a range of values where, if your test statistic lands → we decide to reject the null hypothesis.

think of it in this way!

→ Imagine throwing a dart at the target

→ null hypothesis says that the dart will land on the safe zone (outside the critical region)

→ the critical region is like a danger zone on the target.

→ If your dart (test statistic) hits the red zone, it's so unexpected (based on null hypothesis) that you decide, "the null hypothesis doesn't make any sense"

Critical Point : Its the boundary value that separates the critical region from the non critical region. It is determined based on the significance level (α) and the type of test (eg one tailed or two-tailed)

How it works:

→ For a significance level of 5% ($\alpha = 0.05$)

↳ One tailed test : the critical point is the value of the test statistic that leaves 5% in one tail of distribution.

↳ two tailed test : the critical points leave 2.5% in each trial

Tailed Tests

- Determines how the rejection region is distributed

→ One tailed test : test whether the sample statistic is either significantly greater than or significantly less than the hypothesized value

→ Two tailed test : tests whether the sample statistic is significantly different (either greater or less) than the hypothesize value.

Eqs with visualizations

(e9) One-tailed test (Right tail)

Scenario: A factory claims that its machine produces bolts with an avg strength of atleast 50kg. We want to test if the avg strength is greater than 50kg.

1. Hypothesis

⇒ H_0 : $\mu \leq 50$

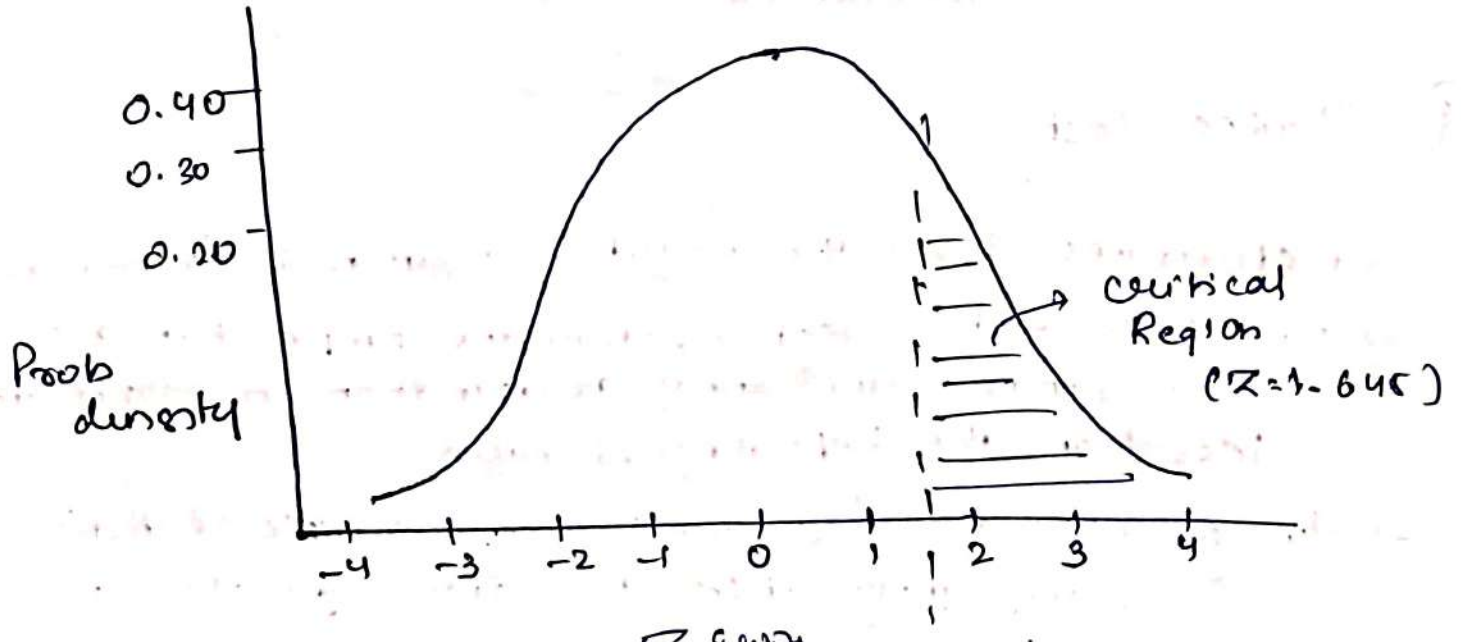
H_1 : $\mu > 50$ (Right tailed test)

2. Significance level(α) = 0.05 → 5%, the critical region lies in the right tail of distribution

3. Critical point : For $\alpha = 0.05$, the critical Z value is 1.645

↳ if the test statistic $Z > 1.645$ → reject H_0

4. Visualization



→ Here the shaded area represents the critical region for a right-tailed test ($H_1: \mu > 50$)

→ the critical point $Z = 1.645$ separates the critical region from the rest of the distribution.

If the test statistic falls in shaded region → we reject Ho.

(eg2) two tailed test

Scenario: A pharma company claims that the avg effectiveness of drug is 75%, we want to test if the actual effectiveness is diff from 75%.

1. Hypothesis:

Ho: $\mu = 75\%$.

H_1: $\mu \neq 75\%$. (two tailed test)

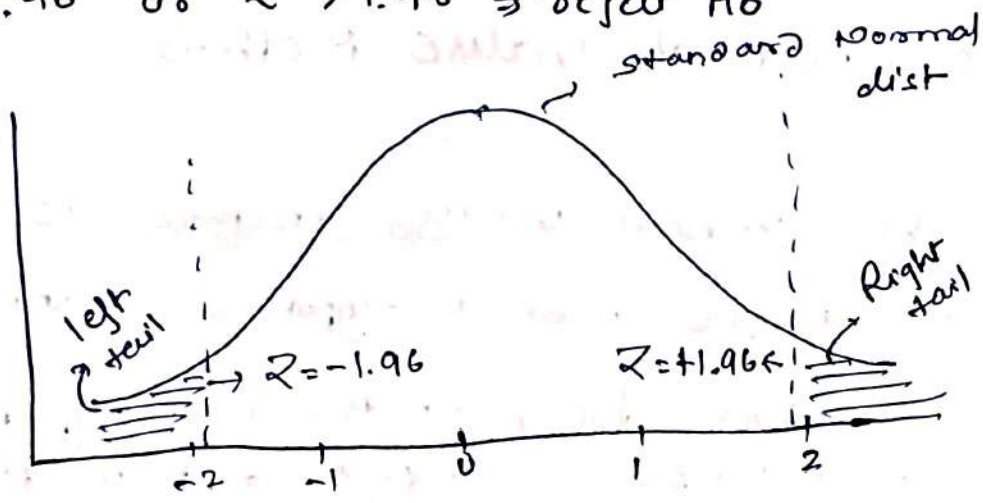
2. Significance Level (α) = 5% → Critical region is split b/w two tails of the distribution.

3. Critical points: For $\alpha = 0.05$ $Z = -1.96, 1.96$

→ If the test statistic

$Z < 1.96$ or $Z > 1.96$ → reject Ho

Visualization



(*) Quick rule to identify the type of test

(1) Two tailed test ($\neq$)

→ Use when you are testing for a difference (each direction).

→ Rejection regions are on the both tails of the distribution

eg testing if mean is not equal to 50 ⇒ $H_1 = \mu \neq 50$

(2) Lower-tailed test ($<$)

→ use when you are testing if the population parameter is less than a certain value.

→ Rejection Region is on the left side of the distribution

eg testing if mean is < than 50 ⇒ $H_1 < 50$

(3) Upper tailed test ($>$)

→ use when you are testing if the population parameter is greater than a certain value

→ Rejection Region is on the right side of distribution

$H_0 > 50$

(*) Critical Value Method

The critical value method is a statistical technique used in hypothesis testing to decide whether to reject the null hypothesis (H_0) based on a comparison of test statistic to critical values

Steps in the critical value method:

1. formulate the hypothesis
   - H_0 ⇒ Null Hypothesis ⇒ Assumes no effect or no difference (eg $\mu = \mu_0$)
   - H_1 ⇒ Alternative Hypothesis: indicates the presence of an effect or difference eg ($\mu \neq \mu_0$, $\mu > \mu_0$ or $\mu < \mu_0$)

2. Choose significance level α
   - commonly used α values 0.05 (5%) or 0.01 (1%)

3. Determine the critical value:
   - Critical values (UCV and LCV) are the boundaries of the rejection region for H_0
   - use the Z table (for normal distribution) to find these values based on α and whether the test is two tailed or one tailed.

4. Compute test statistic

   for a Z test ⇒ $$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

$\bar{x}$ → sample mean
μ_0 → hypothesized population mean
σ ⇒ Population standard deviation
n ⇒ sample size

5. Compare the test statistic with critical value

→ For a two tailed test reject H_0 if

$Z < LCVC$ (Lower critical value) or $Z > UCVC$ (upper critical value)

→ For one tailed test, reject H_0 if Z is beyond the critical value in appropriate tail.

6. Make a decision

→ If the test statistic falls in the rejection region reject H_0

→ Otherwise fail to reject H_0

(eg) Company claims that avg weight of product is 500 gm, you suspect that it's incorrect and test this with a sample of 50 products, the sample Mean weight ⇒ 495 gm, standard dev ⇒ 10gms

$\alpha = 0.05$

$H_0 \Rightarrow \mu = 500$, $H_1 \Rightarrow \mu \neq 500$

$\alpha = 0.05$ (two tailed test) $\alpha = 0.025$ in each tail

Using Z table Z value for (0.025)

$LCV = -1.96$, $UCV = +1.96$

Compute test statistic $\Rightarrow \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{495 - 500}{10/\sqrt{50}}$

$\Rightarrow -3.55$

Z value (-3.55) < -1.96 It falls on rejection region ⇒ Reject Ho ⇒ Avg weight is not 500 gm.

Q what would be the value of cumulative prob of ucv of the significance level (α) for an upper tailed test is 3%?

In an upper tailed test with a significance level (α) of 3%, the cumulative prob up to the upper critical value is (UCV)

$1 - \alpha$

this is because the rejection region (critical region) is the area on the upper tail of the distribution corresponding to significance level.

Calculation:

cumulative prob: $1 - \alpha \Rightarrow 1 - 0.03 = 0.97$

⇒ the cumulative probability of upper critical value is 0.97 ⇒ 97%

Critical Value Method for Hypothesis testing

The critical value method involves comparing a test statistic to a critical value(s) to decide whether to reject or fail to reject the null hypothesis, Let's break the process into steps:

Step 1: formulate the Hypothesis

① Null Hypothesis (H_0): this is the default assumption we are testing,
for eg: H_0: the mean salary of employees is $50000

② Alternative Hypothesis (H_1): this is the opposing statement we aim to support
for eg: the mean salary of employees is not $50,000 (two tailed test)

Step 2: Determine the Significance level (α)

→ Significance level (α): this is the prob of rejecting the null hypothesis
if not specified assume $\alpha = 0.05 \Rightarrow 5\%$
for $\alpha = 0.05$ in a two tailed test
$\alpha/2 = 0.025$ in each

Step3 : Calculate the Z value (Z_c)

→ the Z value is the critical value corresponding to $\alpha/2$, it can be found using Z table

for $\alpha = 0.05$

$Z_c = 1.96$ (from the Z table for 0.025 in each tail for a standard normal distribution).

Step4 : Determine the critical values (UCV and LCV)

① Critical values

→ Upper critical value (UCV) : $\mu + Z_c \cdot \frac{\sigma}{\sqrt{n}}$

→ Lower critical value (LCV) : $\mu - Z_c \cdot \sigma/\sqrt{n}$

example:

→ Assume :

$\mu = 50,000$ (hypothesized population mean)

$\sigma = 4000$ (Population standard deviation)

$n = 25$ (sample size)

→ standard error (SE) = $\frac{4000}{\sqrt{25}}$: $\frac{4000}{5} = 800$

→ Critical values :

UCV → $\mu + Z(SE) = 50000 + 1.96(800)$

→ LCV : $\mu - Z(SE) = 5000 - 1.96(800)$

→ (48,432, 51,568)

Step 5 : Compute test statistic

→ the test statistic measures how far the sample mean (x) is from the hypothesized mean (μ)

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

Step 6 : Make a decision

① Decision Rule :

→ Reject Ho if x or (Z) lies outside the critical region defined by Lcv and Ucv

→ fail to reject Ho if x (or Z) lies inside the critical region

Eg Sample mean (x) : Assume x = 51000

test statistic (Z) : $\frac{51000 - 50000}{800}$

$\Rightarrow 1.25$

∴ 1.25 is b/w −1.96 to +1.96

↳ fail to reject Ho

Q Problem statement :

Company claims that the avg time taken to resolve the customer query is 15 min, A random sample of 36 queries had a mean resolution time of 16 min with a standard dev of 3 min, $\alpha = 0.05$

Step1 Formulating Hypothesis

H_0 (Avg resolution time is) 15min $\Rightarrow$ $\mu = 15$

H_1 (Alternate hypothesis) $\mu \neq 15$ (two tailed)

Given $\alpha = 5\%$.

$\Rightarrow$ 2nd step $\because$ it's two tailed $\Rightarrow \alpha/2 = \frac{0.05}{2} = 0.025$

3rd step $\Rightarrow$ Graph

Z critical value depends on the significance level (α) and the type of test (one tailed or two tailed)

Below is the guide for common significance level

For One tailed test, the critical value is found by $1-\alpha$

α (Significance level)	Cumulative Prob $(1-\alpha)$	Z critical value (Z_c)
0.10 (10%)	0.90	1.28
0.05 (5%)	0.95	1.645
0.01 (1%)	0.99	2.33

For two tailed test, divide the significance level by 2 ($\alpha/2$) and then find the critical value for $1-\alpha/2$

α (Significance level)	Cumulative Prob $(1-\alpha/2)$	Z critical value
0.10 (10%)	0.95	$\pm$ 1.645
0.05 (5%)	0.975	$\pm$ 1.96
0.01 (1%)	0.995	$\pm$ 2.575

Here $\alpha = 0.05$ ∴ its a two tailed test

$\Rightarrow \alpha/2 = 0.025$

for z critical or $Zc = 1 - \alpha/2 = 1 - 0.025$

$= 0.975$

$\Rightarrow Z(0.975) = \pm 1.96$

$\Rightarrow UCV = +1.96, \quad LCV = -1.96$

(Step 4) to compute test statistic

$\bar{x} = 16, \quad \mu = 15$

$\sigma = 3, \quad n = 36$

$SE = \sigma/\sqrt{n} = 3/\sqrt{36} = 1/2 = 0.5$

Z test statistic: $Z = \frac{\bar{x} - \mu}{SE} = \frac{16-15}{0.5} = \frac{1}{0.5} = 2$

(Step 5) Decision Making

Z test = 2

Z critical $\Rightarrow \pm 1.96$

∴ Z test > Z critical

∴ Reject H_0

Q A manufacturer claims that avg price life of product is 36 months. Auditor selects a sample of 49 items → Avg life for these 34.5 months Population standard dev = 4 months, $\alpha = 3\%$

$H_0 : \mu = 36$

$H_1 : \mu \neq 36$ (two tailed)

Given $\alpha = 0.03 \Rightarrow Z_c = 1 - \alpha/2 = 1 - 0.015$

0.985

$\Rightarrow$ Z score $(0.985) = \pm 2.17$

Step2 : to find UCV, LCV from Z_c

Critical value $= \mu \pm (Z_c \cdot SE)$

$$36 \pm \left(2.77 \times \frac{4}{\sqrt{49}}\right) = 36 \pm \left(2.17 \times \frac{4}{7}\right)$$

$\Rightarrow 36 \pm 1.24 \Rightarrow (34.76, 37.24)$

$\Rightarrow$ UCV $\Rightarrow$ 37.24
LCV $\Rightarrow$ 34.76

$\bar{x} = 34.5$

$\bar{x} <$ LCV $\Rightarrow$ Reject Ho

Q Govt regulatory body have specified the max permissible amount of lead in any food product is 2.5 PPM
Suppose we take 100 random samples, mean lead content in samples $\Rightarrow$ 2.6 PPM with standard dev of 0.6

$\Rightarrow SE = 0.6$

One thing to notice here is that the standard dev of sample is given as 0.6 instead of population standard dev $\rightarrow$ so we can approximate the population standard dev to sample standard dev

Given α : 0.03 $\Rightarrow$ 3%

Ho : Avg lead content $\leq$ 2.5 PPM

Hi : Avg lead content $>$ 2.5 PPM (Right tailed test)

$\alpha = 0.03 \Rightarrow Z_c$ for $1 - 0.03 = 0.97 \Rightarrow 1.89$

Q Cadbury states that avg weight of one of its chocolate Product 'Dairy Milk Silk' is 60g, $\alpha = 2\% = 0.02$

Sample size = 100, $\bar{x} = 62.69$

SE = 10.79

$$Z_c = 1 - \alpha/2 = 1 - \frac{0.02}{2}, 1 - 0.01 = 0.99$$

Z_c value of 0.99 is 2.33

$$\mu \pm Z\,SE \Rightarrow 60 \pm (2.33)\frac{10.79}{\sqrt{100}}$$

$60 \pm 2.33(1.079) \Rightarrow +62.49$

$+60.00 - (2.33 \times 1.079)$

$(57.51, 62.49)$ $\Rightarrow 57.519$

Sample mean lies out side the range → Failed test

(*) Pvalue Method

What is Pvalue?

the p value is the probability of observing the data (or something more extreme) assuming the null hypothesis (H_0) is true.

→ Null hypothesis (H_0): this is the default assumption
eg there is no difference, no effect or no relationship

→ Alternative Hypothesis (H_1) : the hypothesis you want to test eg there is a difference an effect or a relationship.

A low p-value (usually $< \alpha$, where $\alpha = 0.05$ or some significance level) suggests evidence against H_0 meaning you reject the null hypothesis

A high Pvalue suggests that the data is consistent with H_0, meaning you fail to reject the null hypothesis

Simplified steps for the Pvalue test method

① formulate the Hypothesis

eg H_0: the mean weight of apples is 150 gms

H_1: the mean weight of apples is not 150 gms

(two tailed test)

② Collect Data and Compute the test statistic

formula of Z score for test statistic ⇒ $Z = \frac{\text{Sample mean} - \text{pop mean}}{\text{standard error}}$

③ Find the Pvalue

Use the Z score and a Z table to find the cumulative prob.

④ Compare the pvalue with α

→ If $P < \alpha$, reject H_0

→ If $P \geq \alpha$, fail to reject H_0

Eg One tailed test

Scenario : A teacher claims that the avg score of her students is greater than 75

H_0 : Avg score ≤ 75

H_1 : Avg score > 75

Given Data : → sample mean = 78
→ Population mean = 75
→ Standard dev = 10
→ sample size = 30

Step1 : Calculate Z score $\Rightarrow Z = \frac{78 - 75}{10/\sqrt{30}} = 1.64$

Step2 : find the P value
from the Z table the cumulative prob for $Z = 1.64$ is 0.9495

For one tailed test $\Rightarrow$ $P = 1 - 0.9495 = 0.0505$

Step3 : Decision
if $\alpha = 0.05$
$P = 0.0505 > 0.05 \Rightarrow$ fail to reject H_0

Eg two tailed test

Scenario : A company claims that avg battery life of Product is 10 hours

H_0 : Avg battery life = 10hrs
H_1 : Avg battery life $\neq$ 10 hrs

Given Data:

Sample mean = 9.5
Population mean = 10
standard dev = 1.5
sample size = 40
$\alpha = 0.05$

Calculate z score

$$z = \frac{9.5 - 10}{1.5/\sqrt{40}} = \frac{-0.5}{0.237} = -2.11$$

finding the Pvalue based on z value

So the cumulative prob for $z = -2.11$ is 0.0174

for two tailed test $\Rightarrow P = 2 \times 0.0174 = 0.0348$

Now taking decision

$P = 0.0348 < 0.05 \Rightarrow \because P < \alpha \rightarrow$ rejecting Ho

(#) So In Short

After formulating the null and alternate hypothesis the steps to follow In order to make a decision using the pvalue method:

→ (1) Calculate the value of z score for the sample mean point on the distribution

→ (2) Calculate the Pvalue from the cumulative probability for the given z score using z table

→ (3) Make a decision on the basis of pvalue (multiply it by 2 for two tailed test) with respect to the given value of α (significance val)

Q A manufacturer claims that the avg life of its product is 36 months, we select a sample of 49 units of the product and calculates the avg life to be 34.5 months, the population standard dev is 4 months
to test the manufacturer claim at 3% significance level using p value test

⇒ $H_0 \Rightarrow \mu = 36$ months , $H_1 \Rightarrow \mu \neq 36$ months

Step1: to calculate value of z score for sample mean point on the distribution

Calculating Z-score for sample mean ($\bar{x}$) = 34.5

$$Z = \frac{\text{Sample mean} - \text{pop mean}}{4/\sqrt{49}}$$

$$\Rightarrow Z = \frac{34.5 - 36}{4/7} = \frac{-1.5 \times 7}{40} = -\frac{21}{40}$$

$$\Rightarrow -21/8 = -2.625$$

Step2: Calculing p value for $Z = -2.625$

Z value for $-2.625 \rightarrow 0.0044$

∴ its two tailed $\rightarrow 2 \times 0.0044 = 0.0088$

$P = 1 - 0.0088$ ⇒ We don't have to use this here

Note : the calculation process for pvalues can depend on which side of distribution we are dealing with, and weather its one tailed or two tailed, let's see why we didn't use 1−p here.

for a z score of $\boxed{-2.62}$ the cumulative prob from the z table (P = 0.0044) already represents the area to the left of z score on the standard normal curve

→ Since the z score is −ve, the area we need is already represented directly as 0.0044 from z table

→ For a two tailed test, we simply double the prob to account for both tails

When to use 1−p

we use 1−p when the observed z-score lies on the right hand side (Positive side of mean), the cumulative Prob from z table will give the area to the left and we subtract it from 1 to find the area to right.

Hence

① If z score is −ve (−2.62)

↳ use cumulative prob directly from z table (P = 0.0044) & for 2 tailed ⇒ 2 × 0.0044

② If z score +ve : find cumulative prob from z table P(0.9956) ⇒ For 1 tailed ⇒ 1 − P = 1 − 0.9956 = 0.0044

For two tailed ⇒ 2(1 − 0.9956) = 0.0088

Hypothesis testing for Paracetamol content

Scenario: A pharma company produces tablet with 500mg of paracetamol as the ideal content.

→ Quality issues if Paracetamol < 500mg
→ Regulatory issues if Paracetamol > 500mg

Objective to test whether the manufacturing process is running successfully using a sample of 900 tablets

Given data → Sample size (n) = 900
→ Sample mean ($\bar{x}$) = 510mg
→ Sample standard dev: 110mg (s)
→ α = 5%.

Steps: $H_0 \Rightarrow \mu = 500mg$
$H_1 \Rightarrow \mu \neq 500mg$

$$Z = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{510 - 500}{110/\sqrt{900}} = \frac{10}{3.67} = 2.72$$

for $Z = 2.72 \Rightarrow$ Prob = 0.9967

∴ two tailed $\Rightarrow P = 2 \times (1 - 0.9967)$

$P = 0.0066$

∵ $P < \alpha$

↳ Reject H_0

(*) Types of Errors

In hypothesis testing, two types of errors can occur based on whether we reject or fail to reject the null hypothesis.

These are :

(1) Type I Error (false Positive) :

→ Definition : Rejecting the Null Hypothesis Ho when it is actually true.

→ Cause : Happens due to random sampling variation or setting a very low significance level (α)

→ Probability : the significance level α represents the likelihood of making a type I error.

for eg $\alpha = 0.05$ means there is a 5% chance of rejecting Ho incorrectly.

Eg : we conclude that a new drug works better than the existing ones, but in reality it doesn't

(2) Type 2 error (false negative)

→ Definition : Failing to reject the null hypothesis (Ho) when it is actually false

→ Cause : Happens due to small sample size, large variability in data or setting a high significance level (α)

→ Probability : Represented by β (Power of test = $1 - \beta$)

Eg of type 2 error : we conclude that a new drug is no better than existing one, but in reality it is better

Comparison of Errors :

	Type I error	Type 2 error
Null Hypothesis (H_0)	Rejected Incorrectly	Not rejected when it should be rejected
Outcome	false Positive	false -ve
Risk Control	Controlled by setting significance level (α)	Reduced by increasing sample size or test power

⇒ type I error represented by α occurs when you reject a true null hypothesis.

⇒ type II error represented by β occurs when you fail to reject null hypothesis when its actually false

Q P-value method

Suppose we conduct a hypothesis test and observe that the value of sample mean and sample standard dw when $n = 25$ do not lead to the rejection of the null hypothesis, we calculate the P value as 0.0667, what would happen to Pvalue

If you observe the same sample mean and standard deviation for a larger sample size say > 50

Ans When the sample size (n) increases, while keeping the sample mean and standard deviation constant, the P value will decrease

because:

Impact of Larger sample size on p value

① Formula for test statistic (z)

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

(S) → sample standard deviation

(n) → sample size

as $n \uparrow \Rightarrow S/\sqrt{n} \downarrow \Rightarrow Z \uparrow$ (Moving further away from 0)

So the P-value Relationship

↳ the p value is the probability of observing a value as extreme as the test statistic, assuming the null hypothesis is true.

A larger Z score corresponds to a smaller area under the tail of curve, hence a smaller p value.

Q In a scenario, the Null Hypothesis (H0) is that the person does not have the disease, and the alternate hypothesis H_1 → Person have a disease, if the treatment doesn't have a serious side effect It is better to increase the probability of making a type1 error to avoid missing a critical diagnosis.

T Distribution

The T test and t-distribution test are the statistical tools used to make conclusions about the population parameters, particularly when the sample size is small or the population variance is unknown.

① What is a t test?

A t-test is a hypothesis test that helps to determine whether there is a significant diff b/w the means of two groups or between a sample mean and a known value, it uses the t distribution to calculate the test statistic when:

→ the sample size is small ($n < 30$)
→ the population standard deviation (σ) is unknown

Types of t tests:

① One-sample t test: Compares the sample mean to a known population mean.

② Independent two sample t test: Compares the means of two independent groups

③ Paired t test: Compares means from the same group at diff times
eg (before and after)

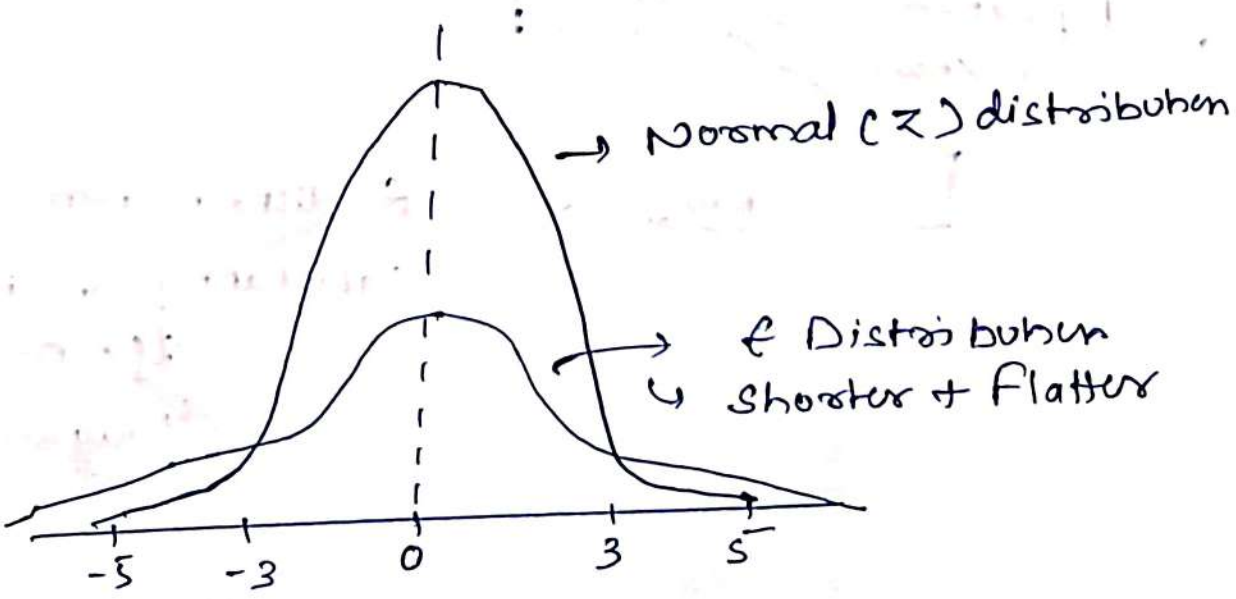
2. What is the t-Distribution?

the t-distribution is a probability distribution similar to the normal distribution (bell shaped and symmetric) but;

→ It has heavier tails, meaning it accounts for more variability in small samples.

→ the shape depends on the degree of freedom (df = n-1) as the sample size (n) increases, the t distribution approaches the standard normal distribution

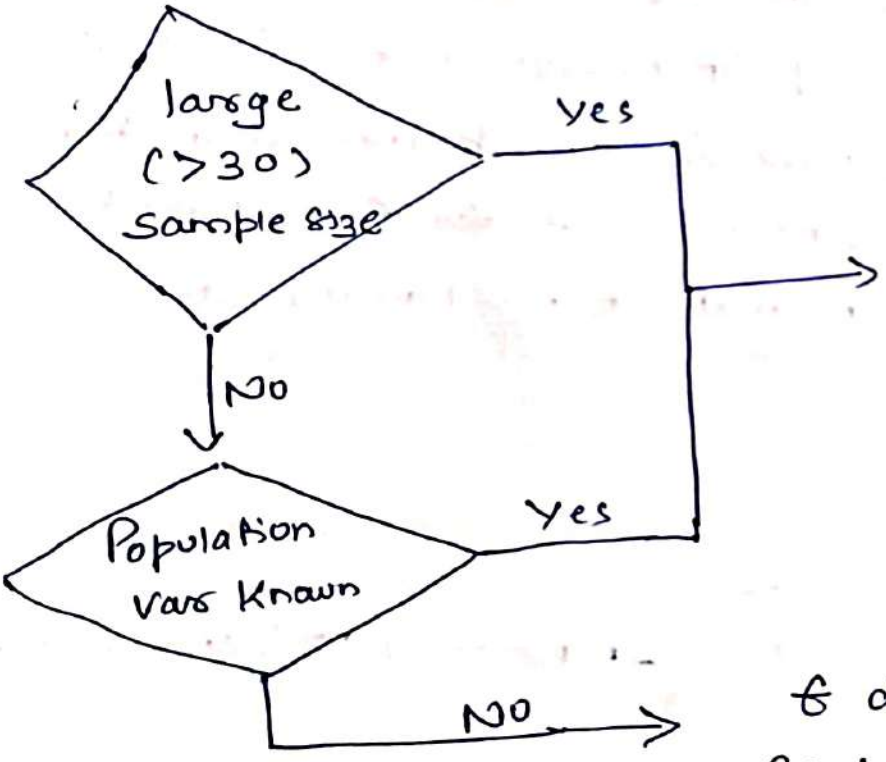
A T distribution is similar to the normal distribution in many cases, for eg it is symmetrical about it's central tendency, however it is shorter than the normal distribution and has a flatter tail, which would eventually mean that it has larger standard deviation.



At a sample size beyound 30, the t distribution approximates to the normal distribution.

the most important use of t distribution is that you can approximate the value of the standard deviation of the population (σ) from the sample standard deviation (s), however as the sample size increases more than 30, the t value tends to be equal to the Z value.

Decision Making in flowchart



Normal Distribution (Calculating Z statistic)

t distribution (Calculating t statistic)

$df = n-1$

↳ degree of freedom

Let's Look at how the method of making a decision changes, if we are using sample's standard deviation instead of populations.

Step1 was to calculate the value of Z_c from the given value of α (significance level), take it as 5% if not specified.

So for finding Z_c we would use t-table instead of Z table the T table contains the value of Z_c for a given degree of freedom and value of α, Here Z_c can also be called as T statistic

Q you are given the standard deviation of a sample size 25 for a two-tailed hypothesis test of a significance level of 5%.

Given info :

→ Sample size(n) = 25

→ Significance level(α) = 5% = 0.05

→ test type : two tailed hypothesis test

steps: find degree of freedom (df)

$$df = n-1 = 25-1 = 24$$

step2: Significance level per tail $\alpha = 0.05$

$$\Rightarrow \alpha = 0.025$$

Step3: find critical from t table (row corresponding to $df = 24$ & col corresponding to $\alpha = 0.025$)

$$\therefore \text{critical} = 2.064$$

Ans = 2.064

& because for sample size = 25, degree of freedom will be 25 - 1 = 24, so look at the value on the t table corresponding to df = 24 and $\alpha = 0.05$ for two tailed test → 2.064.

Q You are given the standard deviation of a sample of size 32 for 2tailed test of $\alpha = 5\%$.

$n = 32$

$df = 32 - 1 = 31$

$\alpha = 0.05$ (∵ two tailed → $\alpha = 0.025$ & for each tail)

$t_{critical} = +1.960$ from t table.

(*) Two Sample Mean test

Note : to enable Data Analysis tools in excel

File ⇒ Options ⇒ Add ins ⇒ Manage
↓
Excel Add ins
↓
OK ← "Analysis toolpak" ← Go

In Data tab → "Data Analysis Sub tab" gets added here

→ Paired Sample Mean test

⊙ Used when data is collected from the same group at two different conditions or at two different times.

Eqs
→ Measuring blood pressure before and after taking medicine.
→ Performance scores of students before and after training.

→ Unpaired Sample mean test (Independent Sample test)

⊙ Used when data is collected from two different groups.

eqs → comparing test scores of GroupA and GroupB under different teaching methods
→ Comparing weights of two seperate groups of individuals following two diff diets.

Two Sample Mean test

Two sample mean test is a statistical method used to compare the means of two independent groups to determine if there is a statistically significant difference b/w them.

It is commonly used in hypothesis testing, when you have two sets of data and want to know if their population means are different.

Types of Two Sample Mean test

① Independent Two Sample Mean Test: Used when the two groups are independent.

② Paired Two Sample T test: Used when two grps are related or paired (eg pre test and post test) on same objects

Key Steps in Two Sample Mean test:

① formulate hypothesis:

- Null Hypothesis (H_0): Means of the two groups are equal.

$$H_0 : \mu_1 = \mu_2$$

- Alternate Hypothesis (H_1): the means of the two groups are not equal

→ $H_1 : \mu_1 \neq \mu_2$

2. Choose the test

→ Use the t-Test if the sample size are small ($n < 30$) and the population standard deviations are unknown

→ Use the z-test if the sample size are large ($n \geq 30$) and population standard dev are known.

3. Calculate the test statistic : the test statistic depends on whether the variance of the two groups are assumed to be equal.

→ equal variances :

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$s_p^2 \rightarrow$ Pooled var

→ Unequal variance :

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

4. Determine the degree of freedom (df)

→ For equal variance $\Rightarrow df = n_1 + n_2 - 2$

→ For unequal variance $\Rightarrow df = \left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2$

$$\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n-1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2-1}$$

→ Calculate the Pvalue and test statistic