STATISTICS

Visualization

ONE WAY TABLE

Lets clear somethings first if we talk about data - A set of Information can be called as a data. Now this data term holds many power like for Machine learning, Deep learning, Data mining etc, we use data.

ONE WAY TABLE

Data Table -

A tabular representation of data is known as a Data Table.

ex-

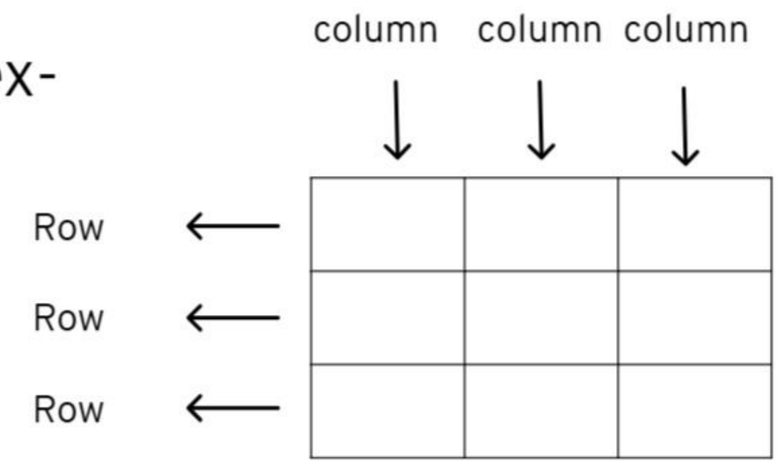
ONE WAY TABLE

Data Table -

A tabular representation of data is known as a Data Table.

We also have Rows and Columns in this Data.

ex-



ONE WAY TABLE

A basic table look like this
it has 4 rows and 2 columns.

With that the Name is known as
a individual and Age is known as
a Variable.

Name	Age
Akarsh	22
Sarthak	23
Harsh	24
Vedant	25

But it's not necessary to have only one variable...

ONE WAY TABLE

Like here you can see we have 3 variables Age, salary, and gender. and one individual name.

Name	Age	salary	gender
Akarsh	22	1.5	M
Sarthak	23	1.6	M
Harsh	24	1.7	M
Vedant	25	1.8	M
Muskan	22	2.0	F

ONE WAY TABLE

Name	Age	salary	gender
Akarsh	22	1.5	M
Sarthak	23	1.6	M
Harsh	24	1.7	M
Vedant	25	1.8	M
Muskan	22	2.0	F

Analysing this data we can see Age is represented in numeric values and gender is represented in M and F only.

ONE WAY TABLE

Name	Age	salary	gender
Akarsh	22	1.5	M
Sarthak	23	1.6	M
Harsh	24	1.7	M
Vedant	25	1.8	M
Muskan	22	2.0	F

So continuous numeric data like Age, Salary, Height etc are known as **Quantitative variables**.

Where as categorical data like gender, Genre etc are known as **Qualitative variables.**

ONE WAY TABLE

But the Question is what makes this Table a One way table?

Name	Age	salary	gender
Akarsh	22	1.5	M
Sarthak	23	1.6	M
Harsh	24	1.7	M
Vedant	25	1.8	M
Muskan	22	2.0	F

The answer is simple - We just have to focus on counter Question

Like - What is the age ?

What is the salary?

What is the Gender ?

Counter question is Simple - "of which person".

ONE WAY TABLE

Also focus on the orientation if we have more columns than rows its better to flip the orientation.

1	2	3	4	5
2				
3				

1	2	3
2		
3		
4		
5		

BAR GRAPH

Visualization

BAR GRAPH

As you can see we have a data table with Individual person and variable fruits.

PERSONS	FRUITS
A	Apple
B	Banana
C	Apple
D	Apple
E	Banana
F	Banana
G	Apple
H	Kiwi
I	Kiwi
J	Apple
K	Mango
L	Mango
M	Apple
N	Banana

BAR GRAPH

PERSONS	FRUITS
A	Apple
B	Banana
C	Apple
D	Apple
E	Banana
F	Banana
G	Apple
H	Kiwi
I	Kiwi
J	Apple
K	Mango
L	Mango
M	Apple
N	Banana

This Data can even be big like we can have more Rows.

Now I have a Question to ask.

Which is most liked fruit ?

Which is the least liked fruit ?

Its tough to answer these questions right we have to count the fruits.

BAR GRAPH

PERSONS	FRUITS
A	Apple
B	Banana
C	Apple
D	Apple
E	Banana
F	Banana
G	Apple
H	Kiwi
I	Kiwi
J	Apple
K	Mango
L	Mango
M	Apple
N	Banana

But the answers to these questions become easier if we visualise the data using bar graph.

But for making a bar graph we need **Frequency Table** for Fruits.

BAR GRAPH

PERSONS	FRUITS
A	Apple
B	Banana
C	Apple
D	Apple
E	Banana
F	Banana
G	Apple
H	Kiwi
I	Kiwi
J	Apple
K	Mango
L	Mango
M	Apple

FRUITS	COUNT
Apple	6
Banana	4
Kiwi	2
Mango	2

So here you got the frequency table of fruits.

BAR GRAPH

FRUITS	COUNT
Apple	6
Banana	4
Kiwi	2
Mango	2

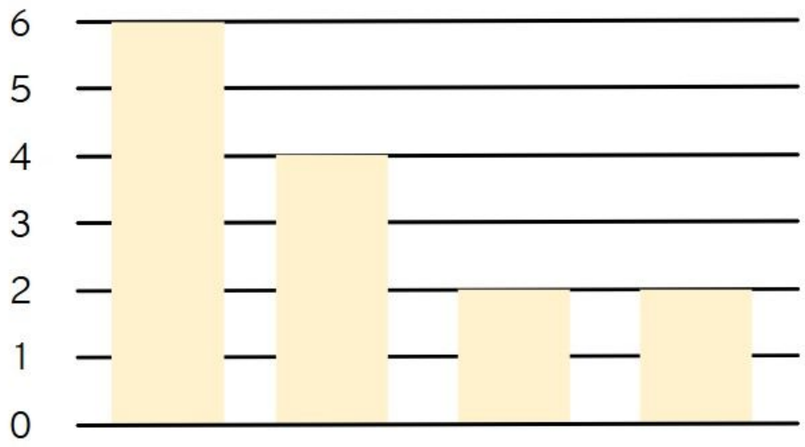
Now to make the just make some horizontal rows and every row has a value according to the count of Fruits.

BAR GRAPH

FRUITS	COUNT
Apple	6
Banana	4
Kiwi	2
Mango	2

Ok after adding the values of count now we have to use bars representing the fruits.

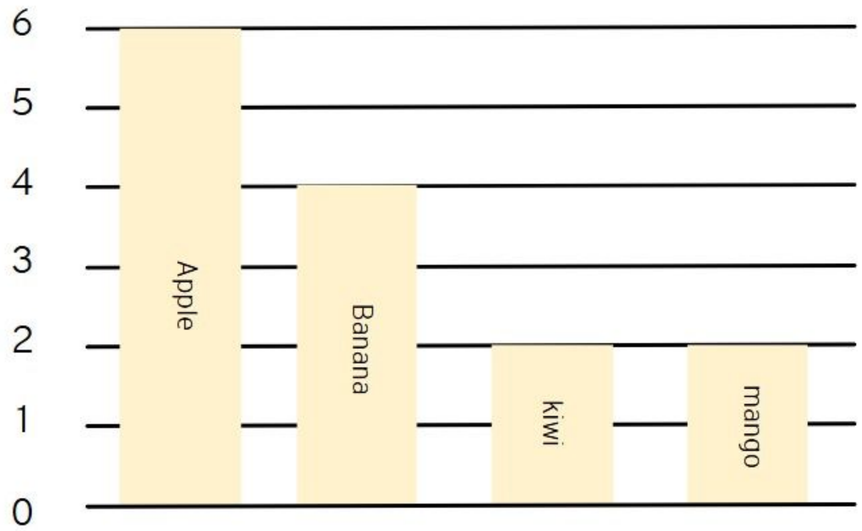
BAR GRAPH



FRUITS	COUNT
Apple	6
Banana	4
Kiwi	2
Mango	2

There you go we got the bar graph.

BAR GRAPH



PERSONS	FRUITS
A	Apple
B	Banana
C	Apple
D	Apple
E	Banana
F	Banana
G	Apple
H	Kiwi
I	Kiwi
J	Apple
K	Mango
L	Mango
M	Apple
N	Banana

Now I again ask the questions.

Which is most liked fruit ?

Which is the least liked fruit ?

BAR GRAPH

So taking a direct glance at the data table it was tough to answer the questions. But using the bar graph it became easy.

PIE CHART

Now Bar graph was good but the next thing we use for visualisation is Pie chart.

As bar graph used the frequency counts of fruits. pie chat also uses the frequency count of a particular variable.

PIE CHART

FRUITS	COUNT
Apple	6
Banana	4
Kiwi	2
Mango	2

But here we have to use the percentage occurrence of every fruits. so lets calculate.

PIE CHART

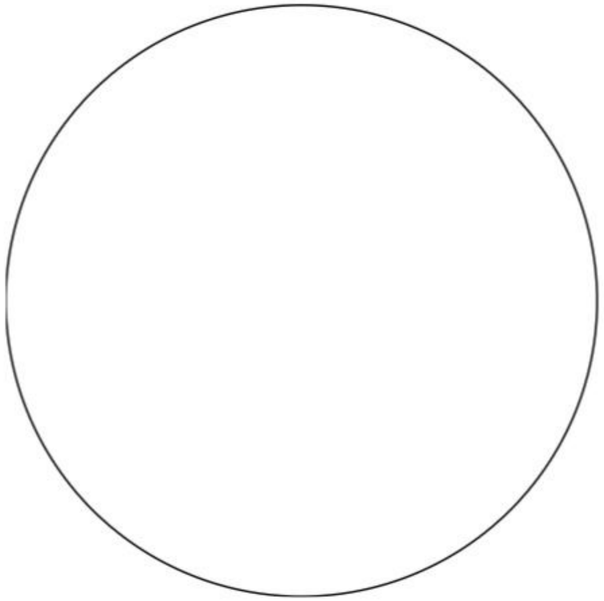
FRUITS	COUNT		
Apple	6	→	42.9%
Banana	4	→	28.6%
Kiwi	2	→	14.3%
Mango	2	→	14.3%

After calculation we got these percentages.

PIE CHART

Now to actually make the pie chart you have to use a pie shaped circle.

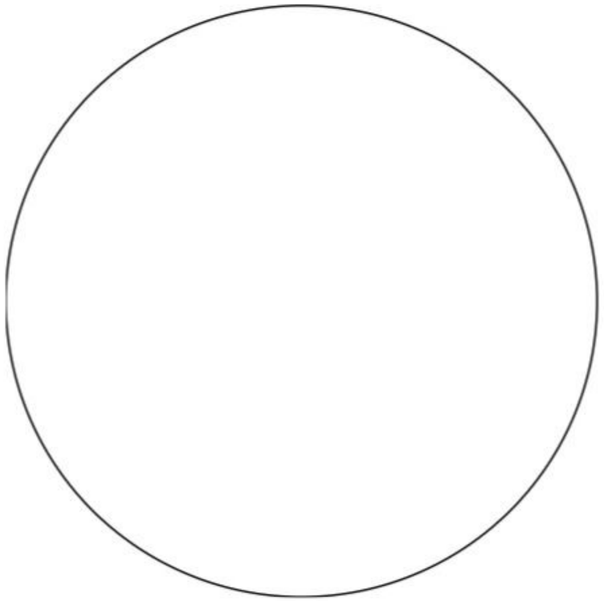
FRUITS	COUNT	
Apple	6	→ 42.9%
Banana	4	→ 28.6%
Kiwi	2	→ 14.3%
Mango	2	→ 14.3%



PIE CHART

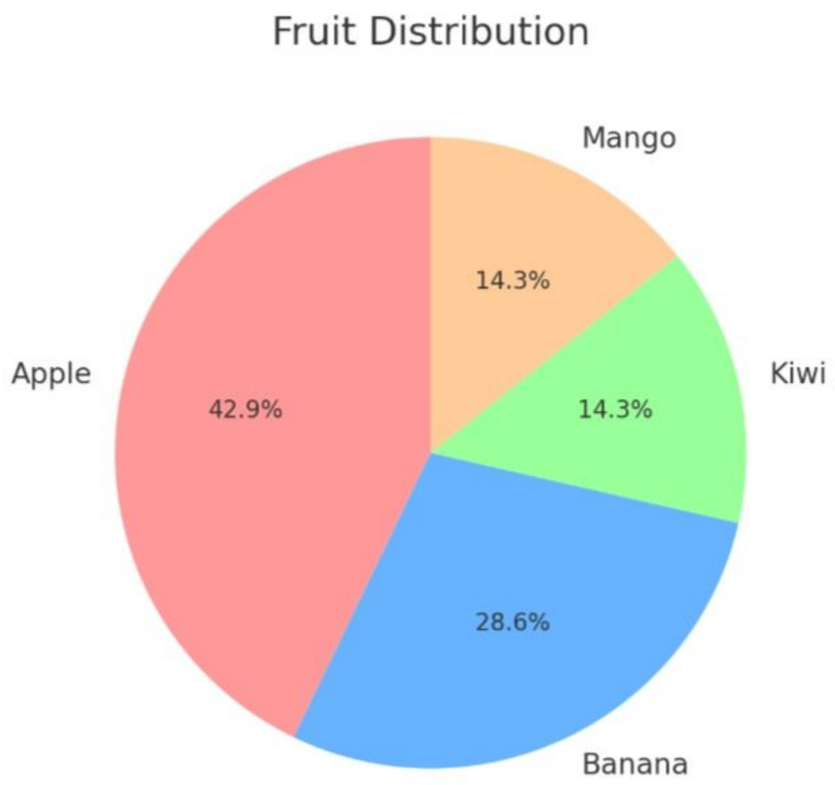
This Circle is representing 100% and you have to slice out the circle like you slice a pie.

FRUITS	COUNT	
Apple	6	→ 42.9%
Banana	4	→ 28.6%
Kiwi	2	→ 14.3%
Mango	2	→ 14.3%



**I am not good with drawing.

PIE CHART



Here is your pie chart it will look like this.

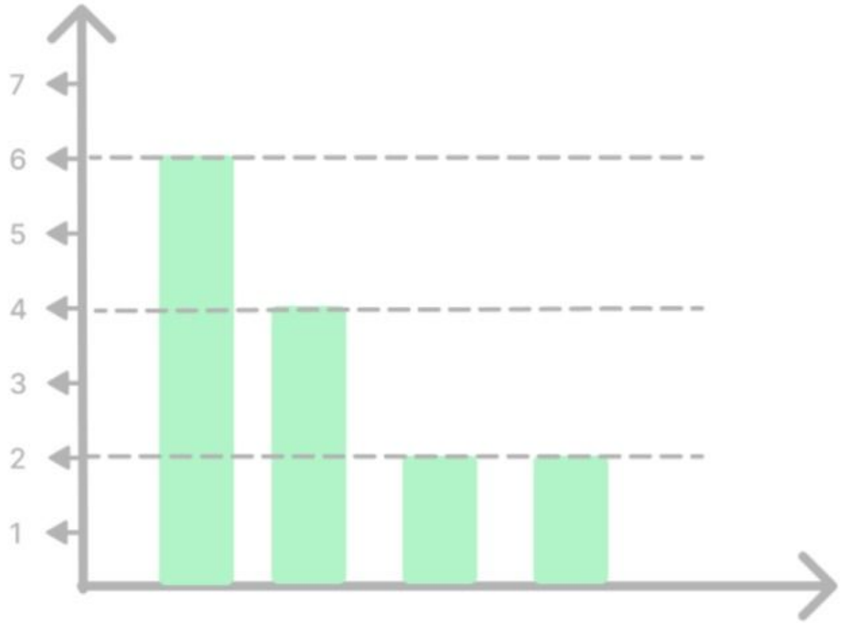
LINE GRAPH

Line Graph are interesting and they have a separate purpose. But first lets understand how to make a line Graph.

For making a Line Graph you must remember how we made a Bar graph.

LINE GRAPH

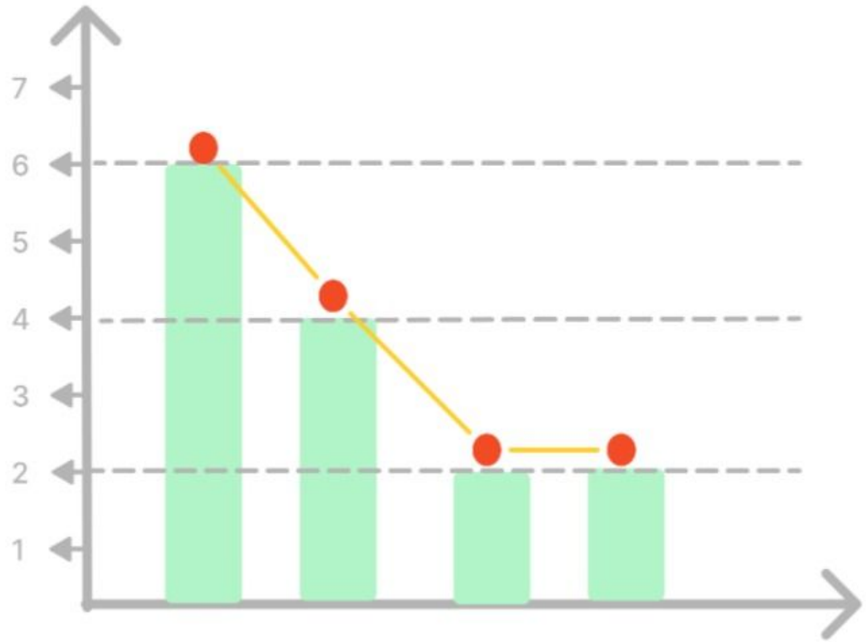
So this was the graph that we made for the fruits, in line graph you just have to plot a dot on top of the bars.



LINE GRAPH

Then after connecting the dots it appears like a line, pretty good huh!

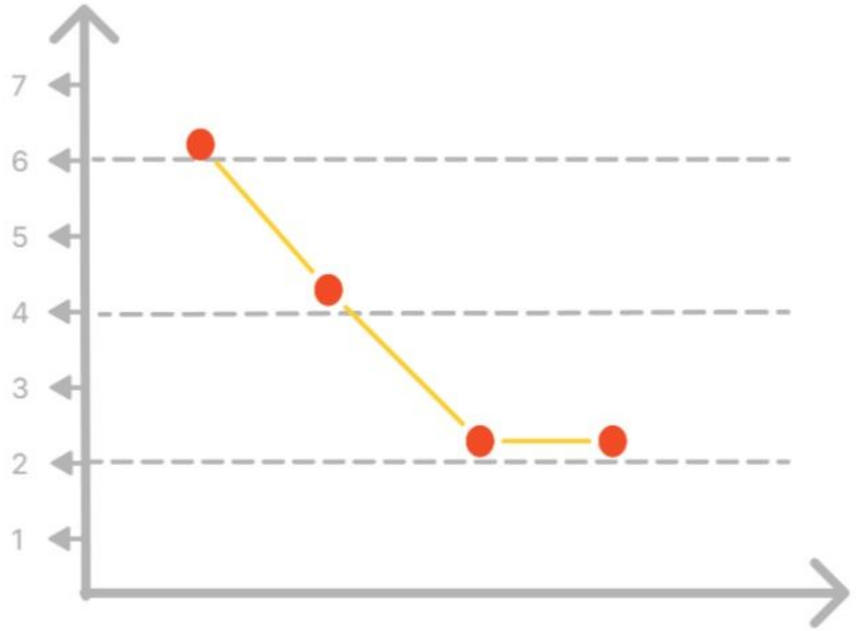
Now remove the bars as well 😎



LINE GRAPH

There we go we got our line graph 🧐

But the question arises is when to use bar graph and when to use line graph.



LINE GRAPH

To explain this answer we will again see a data.

as you can see we have monthly expense.

Month	Expense
Jan	10k
feb	12k
mar	18k
apr	12k
may	5k
jun	10k
jul	20k
aug	25k
sep	30k
oct	15k
nov	10k

LINE GRAPH

You cannot create any Frequency table for this still you can create a bar graph for this data with 12 bars for monthly expenses.

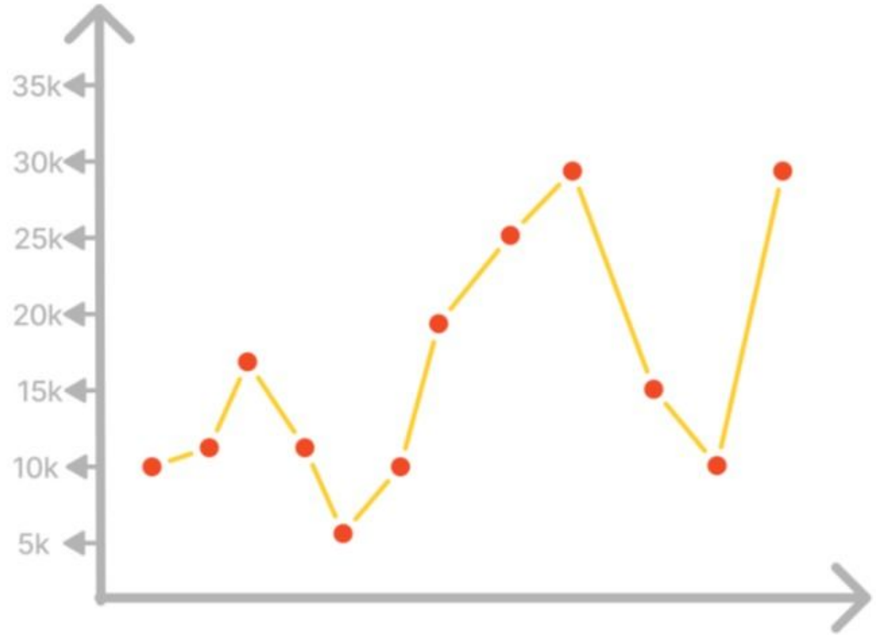
But the Bar Graph will not show us the flow of our data, like what trends my expense is following.

Month	Expense
Jan	10k
feb	12k
mar	18k
apr	12k
may	5k
jun	10k
jul	20k
aug	25k
sep	30k
oct	15k
nov	10k

LINE GRAPH

Month	Expense
Jan	10k
feb	12k
mar	18k
apr	12k
may	5k
jun	10k
jul	20k
aug	25k
sep	30k
oct	15k
nov	10k

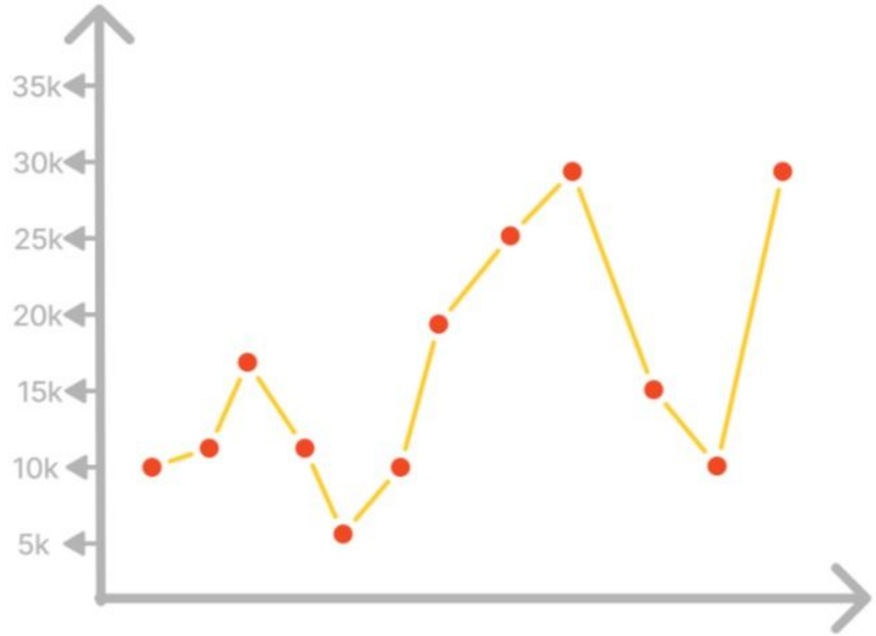
So best here is a line graph as it will show the trends of our continuous data.



LINE GRAPH

Month	Expense
Jan	10k
feb	12k
mar	18k
apr	12k
may	5k
jun	10k
jul	20k
aug	25k
sep	30k
oct	15k
nov	10k

So best here is a line graph as it will show the trends of our continuous data.



O-GIVES

Month	Expense
Jan	10k
feb	12k
mar	18k
apr	12k
may	5k
jun	10k
jul	20k
aug	25k
sep	30k
oct	15k
nov	10k

→

Month	Expense
Jan	10k
feb	24k
mar	42k
apr	54k
may	59k
jun	69k
jul	89k
aug	114k
sep	144k
oct	159k
nov	169k

To make an O-Gives you have to take the cumulative frequencies.

TWO WAY TABLE

Visualization

TWO WAY TABLE

For understanding the two way table lets take an example of a data imagine you have a data with some months and revenue of those particular months.

TWO WAY TABLE

ok don't imagine it here is the data. 😇

Month	revenue
Jan	50k
mar	80k
may	70k
jul	90k
sep	50k
nov	30k

TWO WAY TABLE

Here in this table Month is individual (independent) where as revenue is a dependant variable. as revenue depends on month.

so currently it's a one way table and we can create graphs for this data as well.

Month	revenue
Jan	50k
mar	80k
may	70k
jul	90k
sep	50k
nov	30k

TWO WAY TABLE

Now someone told the information that this data was of 4 years 2020 - 2023 and every months total is given in revenue.

So this current table is a great summary table for us like july must be a good month every year.

but still we don't know.

cause we don't have the data of every year .

Month	revenue
Jan	50k
mar	80k
may	70k
jul	90k
sep	50k
nov	30k

TWO WAY TABLE

Month	revenue
Jan	50k
mar	80k
may	70k
jul	90k
sep	50k
nov	30k

So now we broke down the data for every year and it looks like this

and this data is known as a two way table.

Month	2020	2021	2022	2023	Total
Jan	5k	10k	15k	20k	50k
mar	10k	15k	25k	30k	80k
may	10k	35k	10k	15k	70k
jul	15k	20k	25k	30k	90k
sep	5k	5k	20k	20k	50k
nov	5k	5k	10k	10k	30k

TWO WAY TABLE

Month	2020	2021	2022	2023	Total
Jan	5k	10k	15k	20k	50k
mar	10k	15k	25k	30k	80k
may	10k	35k	10k	15k	70k
jul	15k	20k	25k	30k	90k
sep	5k	5k	20k	20k	50k
nov	5k	5k	10k	10k	30k
Total	50k	90k	105k	125k	370k

How are we calling this data as a two way table. ok again lets play the game of questions.

Q- What is the revenue?

TWO WAY TABLE

Month	2020	2021	2022	2023	Total
Jan	5k	10k	15k	20k	50k
mar	10k	15k	25k	30k	80k
may	10k	35k	10k	15k	70k
jul	15k	20k	25k	30k	90k
sep	5k	5k	20k	20k	50k
nov	5k	5k	10k	10k	30k
Total	50k	90k	105k	125k	370k

Exactly now we have to ask two counter questions which year and which month. so basically here the data is dependent on both rows and columns to answer this question so this is a two way table.

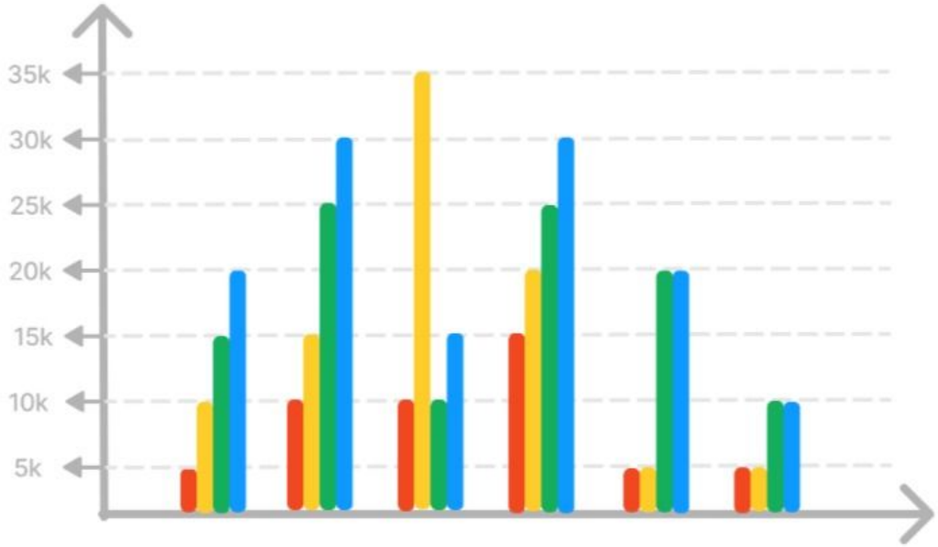
TWO WAY TABLE

Month	2020	2021	2022	2023	Total
Jan	5k	10k	15k	20k	50k
mar	10k	15k	25k	30k	80k
may	10k	35k	10k	15k	70k
jul	15k	20k	25k	30k	90k
sep	5k	5k	20k	20k	50k
nov	5k	5k	10k	10k	30k
Total	50k	90k	105k	125k	370k

Now using this data we can create a line graph for year 2022 specific or 2021.

TWO WAY TABLE

Month	2020	2021	2022	2023	Total
Jan	5k	10k	15k	20k	50k
mar	10k	15k	25k	30k	80k
may	10k	35k	10k	15k	70k
jul	15k	20k	25k	30k	90k
sep	5k	5k	20k	20k	50k
nov	5k	5k	10k	10k	30k
Total	50k	90k	105k	125k	370k



You can also create a bar graph for the data it will look like this.

TWO WAY TABLE

One more thing its a quick filler you should also know what is a **Dot plot.**

For creating a dot plot you have to use the frequency table of specific data which fulfil the need.

Data like this-

Age Group	Members
10 - 20	5
20 - 30	8
30 - 40	12
40 - 50	6
50 - 60	3

TWO WAY TABLE

Dot plot is just like a bar graph just stacking dots on top on each other other than bars but is is important cause it will build a base for understanding histogram.

Age Group	Members
10 - 20	5
20 - 30	8
30 - 40	12
40 - 50	6
50 - 60	3

RELATIVE FREQUENCY TABLE

Visualization

RELATIVE FREQUENCY TABLE

A relative frequency table can be made through a one way table and a two way table, but most of the data that we see around is two way so we will focus on that only.

So lets get the data.

RELATIVE FREQUENCY TABLE

Now as we can see we have flowers Rose and Lily and the color of these flowers are Blue and White. So this is a two way table.

	Red	White
Rose	20	50
Lily	5	30

RELATIVE FREQUENCY TABLE

	Red	White
Rose	20	50
Lily	5	30

Now if we only focus on only Red flowers we can capture many things like but one of the thing is percentages.

Lets calculate the percentage of Red rose and Red lily.

RELATIVE FREQUENCY TABLE

	Red	White
Rose	20	50
Lily	5	30

So after doing the calculation we got this table.

Column frequency	Red	White
Rose	0.80	0.71
Lily	0.20	0.29

RELATIVE FREQUENCY TABLE

	Red	White
Rose	20	50
Lily	5	30

Column frequency	Red	White
Rose	0.80	0.71
Lily	0.20	0.29

Now one important thing about the table is we have calculated the column related frequency. that's why column total is coming out to be 100% but rows are not giving any specific information.

So you can also create a row related frequency table.

JOINT DISTRIBUTION

A joint Distribution table is similar to a relative frequency table as it also contains percentages. But you know why it is called as a joint distribution because the table will compare two different distributions.

Lets see an example.

JOINT DISTRIBUTION

There you go this is a joint distribution as there is a distribution for weight loss and distribution of total hours workout and we are comparing both the distribution so JOINT DISTRIBUTION 😁

		Weight Lost			
		0-2	2-4	4-6	
Total hours workout	3-6	13%	5%	7%	25%
	6-9	5%	15%	10%	30%
	9-12	7%	13%	25%	45%
		25%	33%	42%	100%

JOINT DISTRIBUTION

Its a kind of relative frequency table made for columns and rows.

		Weight Lost			
		0-2	2-4	4-6	
Total hours workout	3-6	13%	5%	7%	25%
	6-9	5%	15%	10%	30%
	9-12	7%	13%	25%	45%
		25%	33%	42%	100%

HISTOGRAM

Visualization

HISTOGRAMS

Histograms look exactly like Bar Graph, except in histogram we are grouping the data into buckets or bins.

For understanding lets take a dummy data.

HISTOGRAMS

So this is a data with an individual candidate and a variable age.

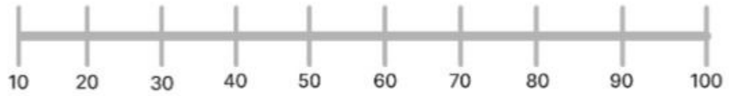
But for clearing things up this data is incomplete and there are multiple other rows.

Candidates	age
p1	21
p2	22
p3	45
p4	56
...	

HISTOGRAMS

Now to make a histogram we have to again use a X axis, and on that axis we can create the bins portrayed below.

Candidates	age
p1	21
p2	22
p3	45
p4	56
...	



HISTOGRAMS

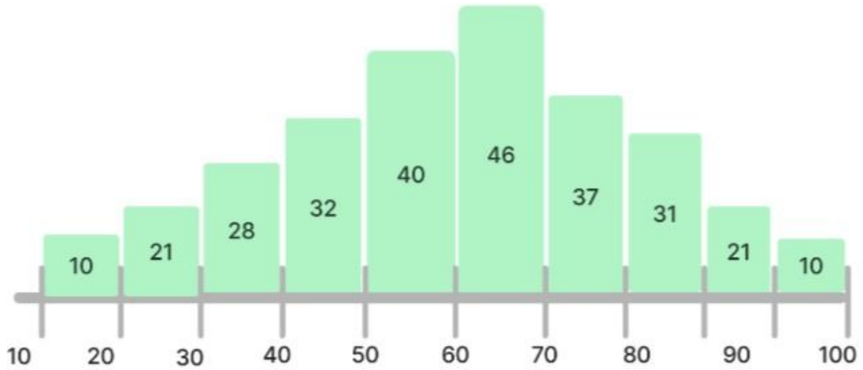
Now inside these bins we have to put the candidates according to there age. Its kind of capturing the frequency of ages like, how many times specific portion of ages occurred. After this you have to plot the bars according to the number of candidates in bins.

Candidates	age
p1	21
p2	22
p3	45
p4	56
...	

HISTOGRAMS

Now after making all the bars it will look like this.

Candidates	age
p1	21
p2	22
p3	45
p4	56
...	



HISTOGRAMS

Now work on an actual data.

There are several persons going to gym and the hours they are working are

3,4,4,5,5,5,6,6,7,8,9,10,10,10,10,11,11,12,13,14,15,15,18,18,20,20,20,21,24,24.

Now lets see the range of the data set largest - smallest = 24 - 3 = 21

After finding the range now we have to fine the number of bins, the bins can be 6 or

7 usually you have to play around with bins try different bins values.

HISTOGRAMS

Now work on an actual data.

There are several persons going to gym and the hours they are working are

3,4,4,5,5,5,6,6,7,8,9,10,10,10,10,11,11,12,13,14,15,15,18,18,20,20,20,21,24,24.

Now lets create 6 bins - 21/6 = 3.5

but its always good to round up so we round up to 4

So our bins are - 1-5, 5-9, 9-13, 13-17, 17-21, 21-25.

HISTOGRAMS

3,4,4,5,5,5,6,6,7,8,9,10,10,10,10,11,11,12,13,14,15,15,18,18,20,20,20,21,24,24.

So our bins are - 1-5, 5-9, 9-13, 13-17, 17-21, 21-25.

MEASURE OF CENTRAL TENDANCY

Analysing Data

OUTLINE

In last section we saw how we can manage showcase our data using graphs and charts. But this section will focus on how we can do things mathematically to better understand our data.

MEASURE OF CENTRAL TENDENCY

First thing we are gonna talk about is measure of central tendency, which is different ways we can come up to describe the middle part of our data.

many people think there is only one way to describe the middle but actually there are several.

MEASURE OF CENTRAL TENDENCY

Mean - The first thing is something you hear all the time and that is Mean. It is also know as Average.

lets see with an example-

MEASURE OF CENTRAL TENDENCY

Mean

Here we have the data - 1,2,3,4,5

MEASURE OF CENTRAL TENDENCY

Mean

It was easy to find the mean of continuous data but what about this data - 2,2,7,9,10

mean = Sum of all the numbers / total numbers

MEASURE OF CENTRAL TENDENCY

Mean

Fancy representation in statistics is - $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$

Look at the mean as a balancing point on a sea saw
and try for 2,4,7,7 and distance will be same from the mid point.

MEASURE OF CENTRAL TENDENCY

Median

Median is also used to find the middle point of our data, but her we are just looking for the middle number.

Lets take the example of - 4,7,9,10

MEASURE OF CENTRAL TENDENCY

Median

In even numbers of data points we have to find the average of the middle 2 numbers.

But for odd data points we can directly find the mid point.

MEASURE OF CENTRAL TENDENCY

Mode

Mode is simple whichever number occurred most in our data is the mode.

For example - 2,2,6,6,8,8,9,0,10,10,10,10

MEASURE OF CENTRAL TENDENCY

Mode

now lets see this data - 1,2,3,4,5,6

we can say this data has no mode cause all the values are distinct.

MEASURE OF CENTRAL TENDENCY

Mode

What about this data 1,2,1,1,2,2,7,8,9

here we can say this is a base of bimodal as we have 2 modes 1 and 2.

Or we can say whenever we have modes more than one it is a case of bimodal.

MEASURE OF SPREAD

Analysing Data

MEASURE OF SPREAD

We talked about finding the central point of our data but we also have to talk about measure of spread, which are kind of the opposite , Its how much and how are data is spread out around its center.

Just for an example is it really tightly clustered around the center, or is it scattered far away from the center in an uneven, or unbalanced sort of way.

MEASURE OF SPREAD

We will talk about Range and IQR(Interquartile range) variance and standard deviation are also measures of spread but we will talk about that later in the course.

MEASURE OF SPREAD

Range

To find the range you have seen before we just have to subtract the highest data point with the smallest data point.

MEASURE OF SPREAD

Range

Lets take two examples and understand what range wants to tell us.

Ex1 - we have a golf game and there were different scores the highest score was 72 and lowest was 62

Ex2- There are various heights of candidates in a class. the height of tallest student is 172cm and shortest candidate is 162cm

MEASURE OF SPREAD

Range

So Range tells us the distribution of our data, how and in what range they are distributed.

so whatever the data is you just have to find the minimum and maximum point of our data and you will get the range by subtracting those points.

$$\text{Range} = \text{maxV} - \text{minV}$$

MEASURE OF SPREAD

Interquartile Range(IQR)

After understanding range now we have to focus on interquartile range for our data.

MEASURE OF SPREAD

Interquartile Range(IQR)

But for understanding IQR we first have to understand Quartiles because in IQR we are measuring the distance between one quarter and another quarter of our data.

MEASURE OF SPREAD

Interquartile Range(IQR)

For starters lets look at this data and understand quarters and IQR.

12,12,13,14,15,17,20,20,28,29,31,35,46,46,48,51,59,60

MEASURE OF SPREAD

Interquartile Range(IQR)

Now in this data we have to find the quarters of our data or we can say percentiles of our data.

12,12,13,14,15,17,20,20,28,29,31,35,46,46,48,51,59,60

MEASURE OF SPREAD

Interquartile Range(IQR)

So first see the terminology in a data there is

Q1 - 25 percentile.

Q2 - 50 percentile.

Q3 - 75 percentile.

Q4 - 100 percentile.

12,12,13,14,15,17,20,20,28,29,31,35,46,46,48,51,59,60

MEASURE OF SPREAD

Interquartile Range(IQR)

So after finding the percentile we get.

12,12,13,14,$\boxed{15}$,17,20,20,28 29,31,35,46,$\boxed{46}$,48,51,59,60

↓ Q1 ↓ Q2 ↓ Q3

MEASURE OF SPREAD

Interquartile Range(IQR)

Now to find the IQR you just have to subtract Q1 from Q3

12,12,13,14,$\boxed{15}$,17,20,20,28 29,31,35,46,$\boxed{46}$,48,51,59,60

↓ Q1 ↓ Q2 ↓ Q3

MEASURE OF SPREAD

Interquartile Range(IQR)

So if the IQR is low it means the data is close to the median

if the IQR is high it means the data is highly spread from the center.

OUTLIERS

Analysing data

OUTLIERS

Before starting you must know what is outliers.

Just see this example - 1,2,4,7,9,10,101

OUTLIERS

Before starting you must know what is outliers.

Just see this example - 1,2,4,7,9,10,101

ok now lets find the central tendency of this data.

OUTLIERS

Before starting you must know what is outliers.

Just see this example - 1,2,4,7,9,10,101

mean - 1 + 2 + 4 + 7 + 9 + 10 + 101 = 134

134/7 = 19.14

median - midpoint of our data - 7

OUTLIERS

Just see this example - 1,2,4,7,9,10,101

mean - 19.14

median - 7

As you can see both mean and median are used to find the central tendency of our data and still both have a huge difference.

OUTLIERS

Just see this example - 1,2,4,7,9,10,101

mean - 19.14

median - 7

This is happening because of a large value(101) present in our data and mean depends on all the values median depends on the positions of our data.

OUTLIERS

Just see this example - 1,2,4,7,9,10,101

mean - 19.14

median - 7

So 101 is considered to be a outliers.

Now lets see some points how outliers are damaging our data.

OUTLIERS

- Outliers can distort measures like mean and standard deviation, making them less representative of the data.
- In machine learning, especially regression models, outliers can disproportionately influence the model, leading to poor generalization.
- outliers impact our model performance our statistical testing and many other things so we have to get rid from them.

OUTLIERS

So we have to get rid of the outliers but how, we use upper fence and lower fence in our data and by using that we can get rid of that thing.

OUTLIERS

here we will use our IQR, Q1 and Q3 to find the upper and lower fence.

Lets understand with the example.

1, 1, 2, 2, 3, 4, 5, 5, 6, 6, 7, 8, 9, 9, 10, 10, 11, 13, 13, 14, 15, 101

OUTLIERS

here we will use our IQR, Q1 and Q3 to find the upper and lower fence.

1, 1, 2, 2, 3, 4, 5, 5, 6, 6, 7, 8, 9, 9, 10, 10, 11, 13, 13, 14, 15, 101

↓ Q1 ↓ Q2 ↓ Q3

IQR = Q3 - Q1

IQR = 11 - 4 = 7

Q1 = 4

Q3 = 11

OUTLIERS

1, 1, 2, 2, 3, 4, 5, 5, 6, 6, 7, 8, 9, 9, 10, 10, 11, 13, 13, 14, 15, 101

IQR = 7

Q1 = 4

Q3 = 11

Now to find the upper fence and lower fence you have to follow this formula.

Upper Fence = Q3 + 1.5(IQR)

Lower Fence = Q1 - 1.5(IQR)

lets find out...

OUTLIERS

1, 1, 2, 2, 3, 4, 5, 5, 6, 6, 7, 8, 9, 9, 10, 10, 11, 13, 13, 14, 15, 101

IQR = 7

Q1 = 4

Q3 = 11

Upper Fence = 11 + 1.5(7) = 21.5

Lower Fence = 4 - 1.5(7) = -6.5

OUTLIERS

1, 1, 2, 2, 3, 4, 5, 5, 6, 6, 7, 8, 9, 9, 10, 10, 11, 13, 13, 14, 15, ~~101~~

so 21.5 is upper fence and -6.5 is lower fence.

That means values above 21.5 and values below -6.5 are outliers. So in our data 101 is the only element that is out of our fence so final step is to remove the data.

5 NUMBER SUMMARY

Analysing data

5 NUMBER SUMMARY

The **five-number summary** is a statistical measure used to describe the distribution of a dataset. It provides a quick overview of the data's spread and center, and it includes the following five key values:

- Minimum
- First Quartile(Q1)
- Median (Q2)
- Third Quartile(Q3)
- Maximum

5 NUMBER SUMMARY

- Minimum
- First Quartile(Q1)
- Median (Q2)
- Third Quartile(Q3)
- Maximum

Now lets see the working with the data

2, 4, 7, 10, 15, 20, 25,205

5 NUMBER SUMMARY

- Minimum
- First Quartile(Q1)
- Median (Q2)
- Third Quartile(Q3)
- Maximum

Now lets see the working with the data

2, 4, 7, 10, 15, 20, 25,205

If you see this data you have outliers as well and for making the 5 number summary you have to remove the outliers first.

5 NUMBER SUMMARY

- Minimum - 2
- First Quartile(Q1) - 4
- Median (Q2) - 10
- Third Quartile(Q3) - 20
- Maximum - 25

Now lets see the working with the data

2, 4, 7, 10, 15, 20, 25, ~~205~~

5 NUMBER SUMMARY

- Minimum - 2
- First Quartile(Q1) - 4
- Median (Q2) - 10
- Third Quartile(Q3) - 20
- Maximum - 25

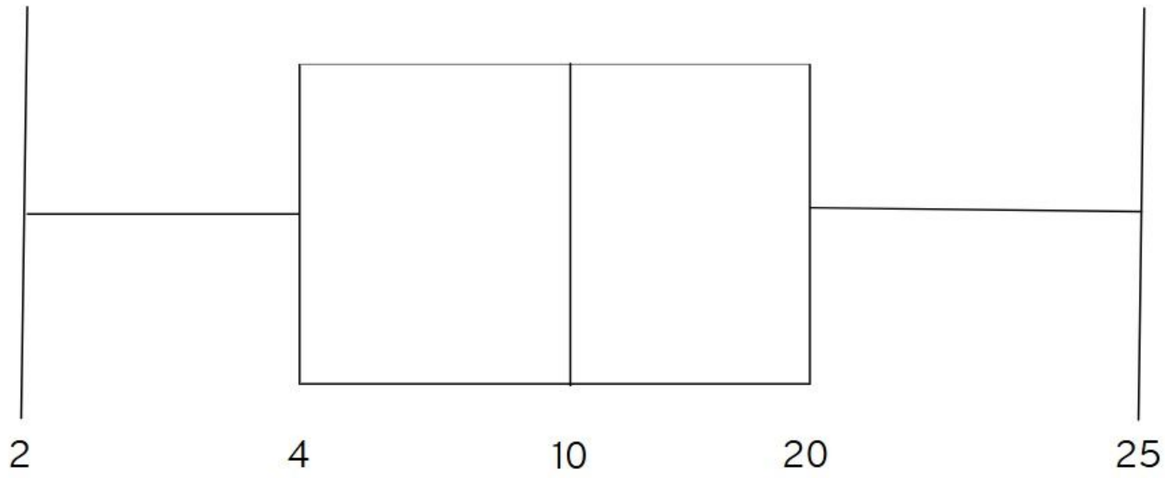
Now this 5 number summary is used to create a box plot as well lets see the box plot.

5 NUMBER SUMMARY

2, 4, 7, 10, 15, 20, 25,205

- Minimum - 2
- First Quartile(Q1) - 4
- Median (Q2) - 10
- Third Quartile(Q3) - 20
- Maximum - 25

Box Plot

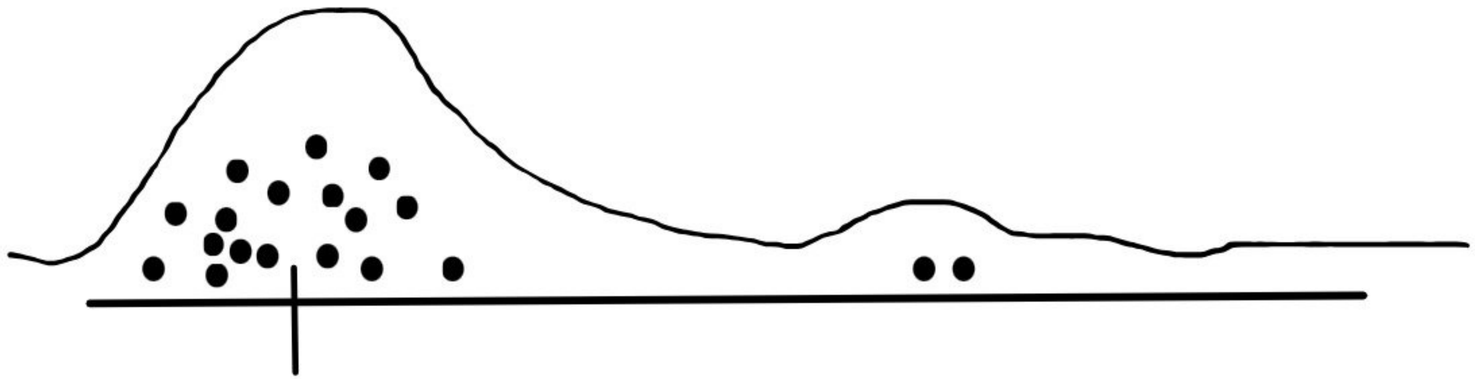


INTRODUCTION

Data Distribution

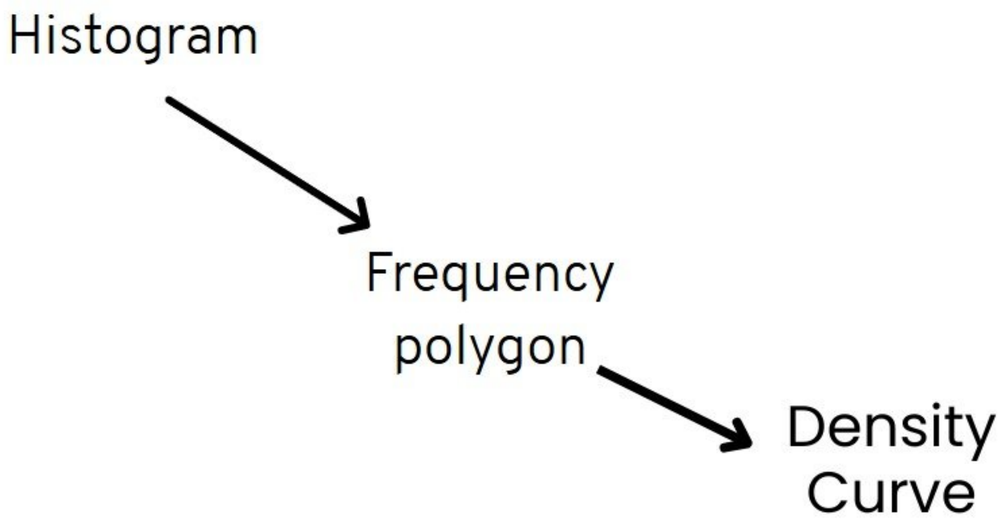
INTRODUCTION

In data distribution we will see how data is spread around the mean for example. In the below distribution data is tightly clustered around the mean and because of 2 outliers the the mean is behind.



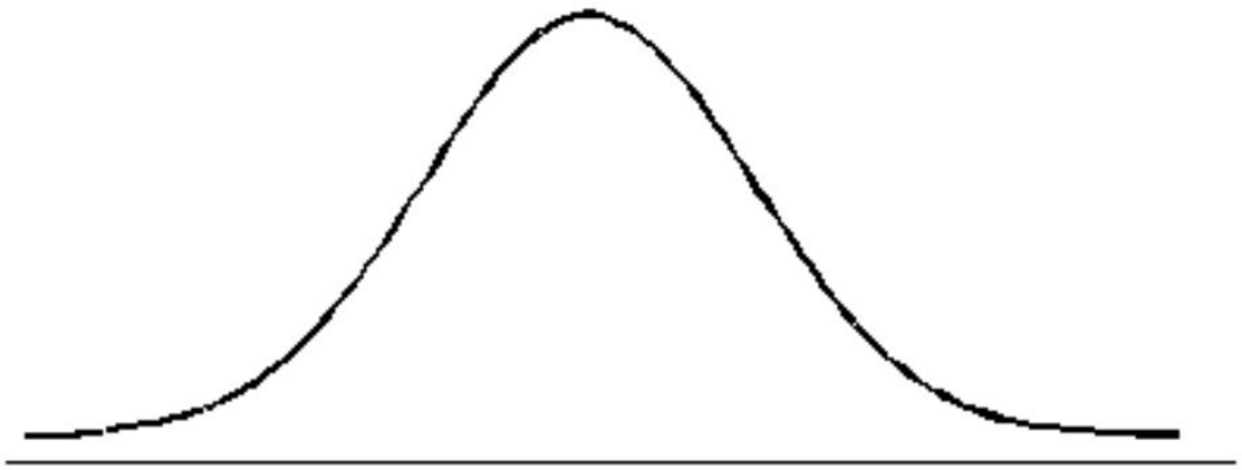
INTRODUCTION

In this section we will see how data distribution starts with a Histogram then it becomes a frequency polygon and eventually turns into a density curve that represents the distribution.



INTRODUCTION

We are also going to see a important - normal distribution which always forms a bell shaped curve.



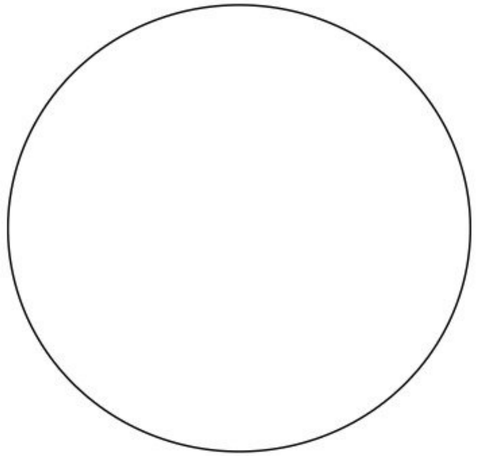
MEAN, VARIANCE AND STANDARD DEVIATION

Data Distribution

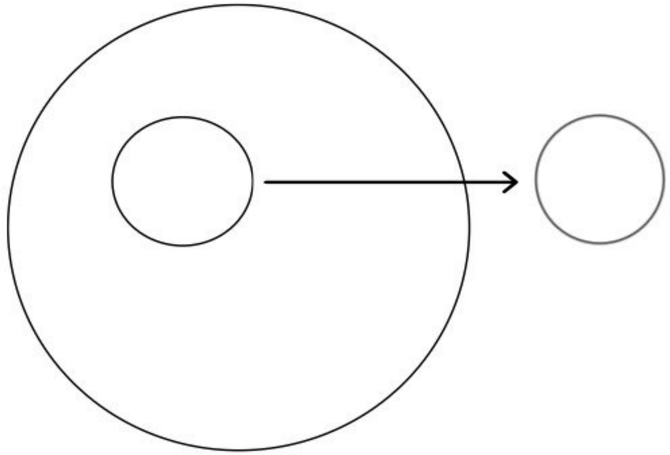
MEAN, VARIANCE AND STANDARD DEVIATION

Before understanding this thing first we have to understand sample and population in statistics.

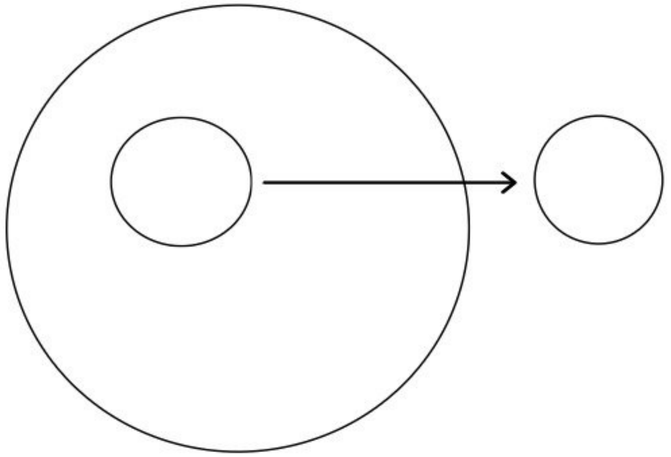
MEAN, VARIANCE AND STANDARD DEVIATION



Lets say this circle is representing the population.



If we extract some part of the population it is called as sample.



Now this sample is some what representing the population.

For example, when testing a vaccine, it is impractical to test it on the entire global population of 8 billion people. Instead, a sample of 500 individuals from diverse regions around the world can be selected for testing. Based on the results from this sample, conclusions can be drawn about the vaccine's effectiveness for the entire population.

MEAN, VARIANCE AND STANDARD DEVIATION

When discussing mean, variance, and standard deviation, it is important to clarify whether we are referring to a population or a sample, as the formulas used for each differ.

MEAN, VARIANCE AND STANDARD DEVIATION

Now lets understand variance- **Variance** is a statistical measure that quantifies the spread or dispersion of a set of data points around their mean (average).

MEAN, VARIANCE AND STANDARD DEVIATION

To find the variance we are going to first find the mean of our data.

$85,90,95,100,105$

mean = (85 + 90 + 95 + 100 + 105) / 5 = 95

MEAN, VARIANCE AND STANDARD DEVIATION

Now after finding the mean we have to compute the difference from each point and mean and squaring them.

$85, 90, 95, 100, 105$

mean = 95

$$(85-95)^2 + (90-95)^2 + (95-95)^2 + (100-95)^2 + (105+95)^2 = 250$$

MEAN, VARIANCE AND STANDARD DEVIATION

Now after finding the mean we have to compute the difference from each point and mean and squaring them.

85,90,95,100,105

mean = 95

$$(85-95)^2 + (90-95)^2 + (95-95)^2 + (100-95)^2 + (105-95)^2 = 250$$

Now just divide the computed difference from the numbers of points in your data

250/5 = 50

Now after finding the mean we have to compute the difference from each point and mean and squaring them.

$85,90,95,100,105$

250/5 = 50

so 50 is the variance of this data.

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

MEAN, VARIANCE AND STANDARD DEVIATION

Now to find the standard deviation you just have to take the square root of the variance.

$85, 90, 95, 100, 105$

variance = 50

standard deviation = $\sqrt{50}$ = 7.071

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

DENSITY CURVE

Data Distribution

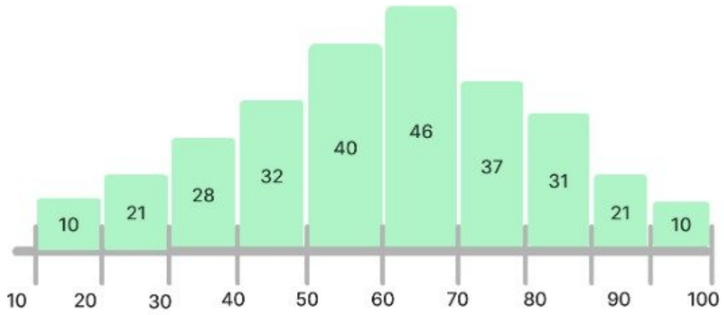
DENSITY CURVE

To understand a density curve should first know about histogram. which we have discussed previously.

Now lets create a histogram and then understand.

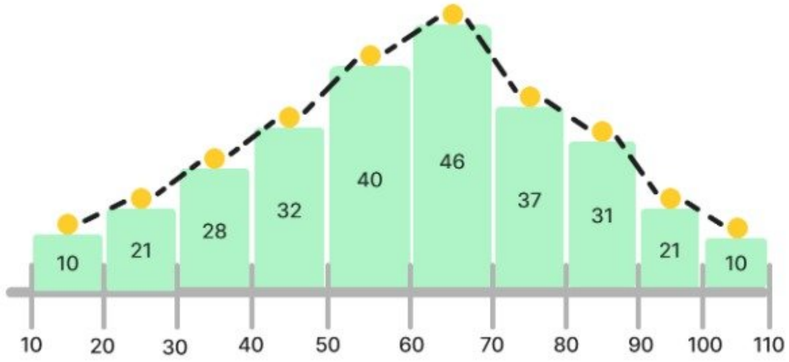
DENSITY CURVE

So this is a basic histogram representing total data distribution of age present in a specific ranges.



DENSITY CURVE

Now we just have to plot points over the bars and connect them kind of making a line graph.



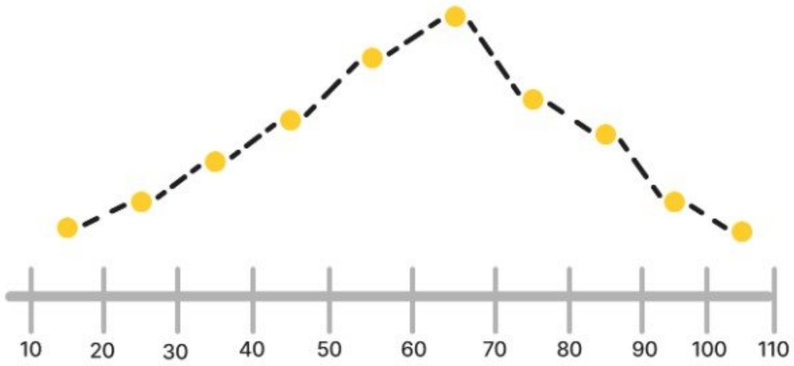
DENSITY CURVE

After this now we just have to remove the bars and you will get a frequency polygon.

the fewer bins we use the more

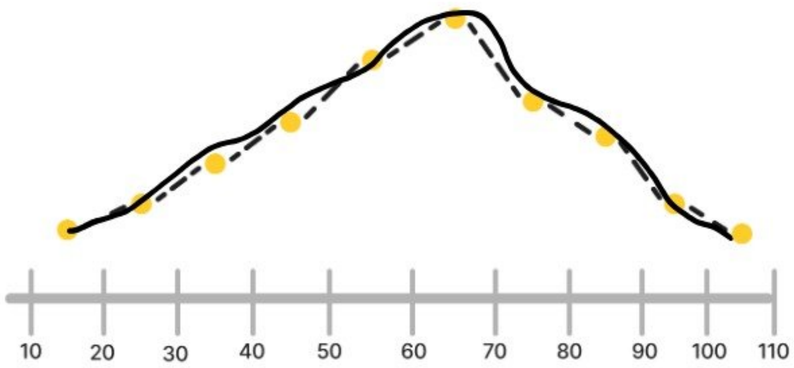
inaccurate our polygon will

be.



DENSITY CURVE

now imagine you are using more and more bins the dots will be very closer and eventually it will become a density curve.

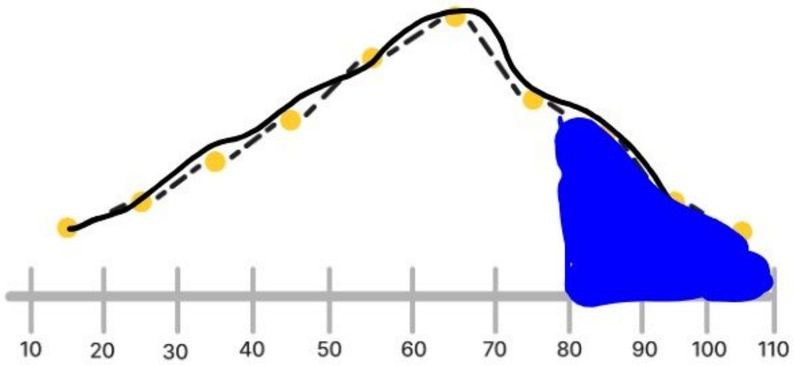


DENSITY CURVE

Now this density curve tells us many things like our total area inside the density curve is 100% and we want to know how much percentage of our data is above 80 age we can

easily get it in our density

curve.



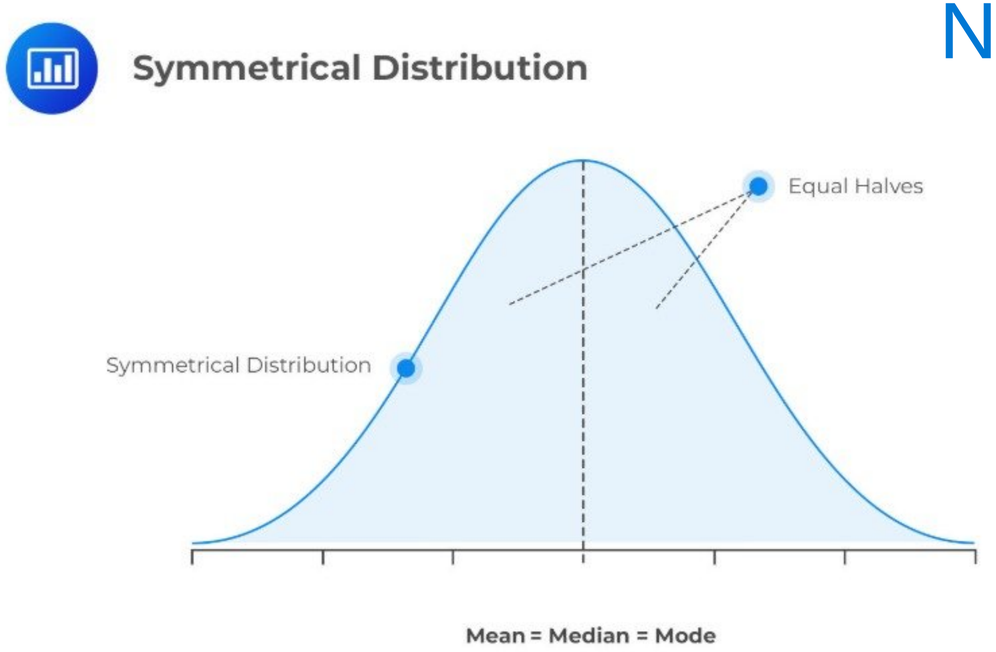
DENSITY CURVE

This density curve is telling us the distribution.

Gaussion Distribution

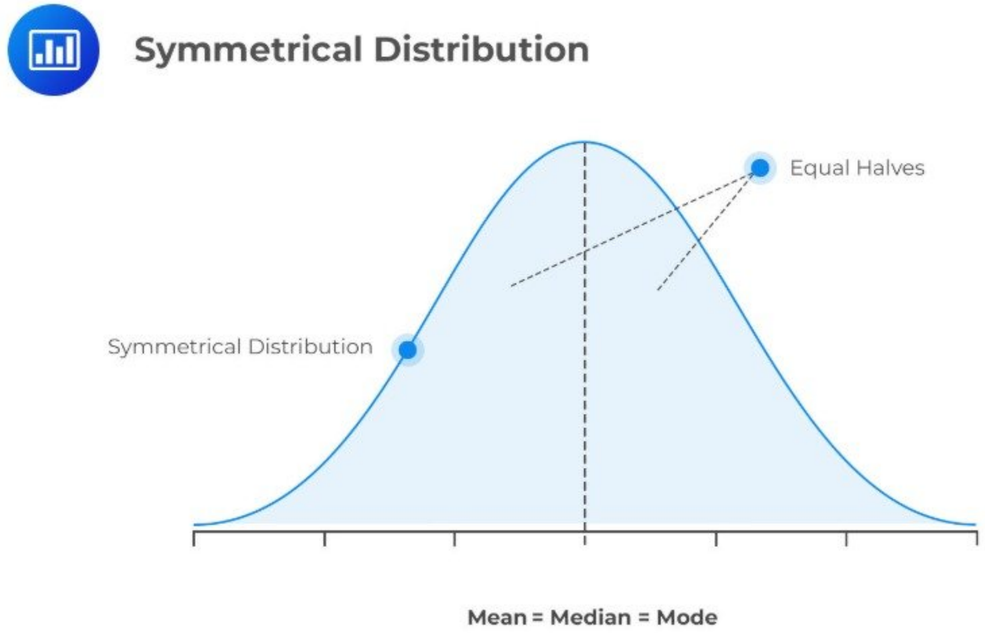
and one of the distribution is symmetrical distribution it looks like this.

Normal Distribution



DENSITY CURVE

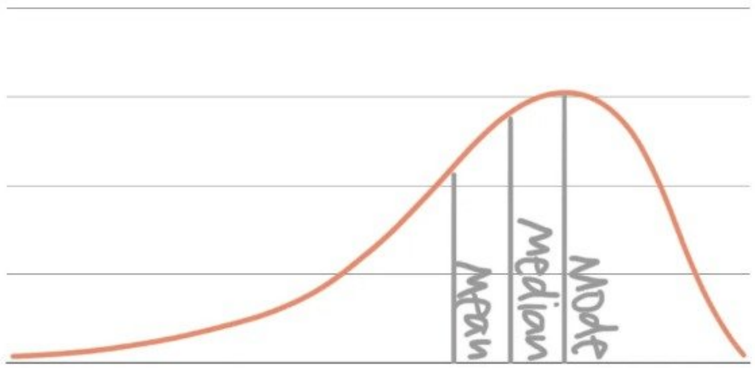
This distribution is also called as a normal distribution and it always follows a bell shaped curve and we will talk about it a lot in the following parts



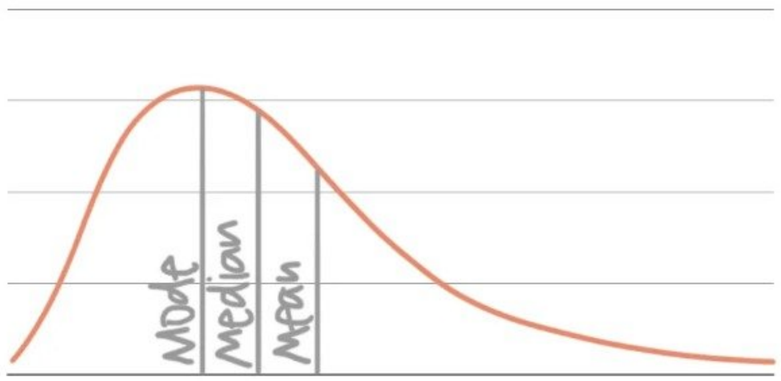
DENSITY CURVE

There is another distribution and that is skewed distribution basically there are left skewed and right skewed distributions.

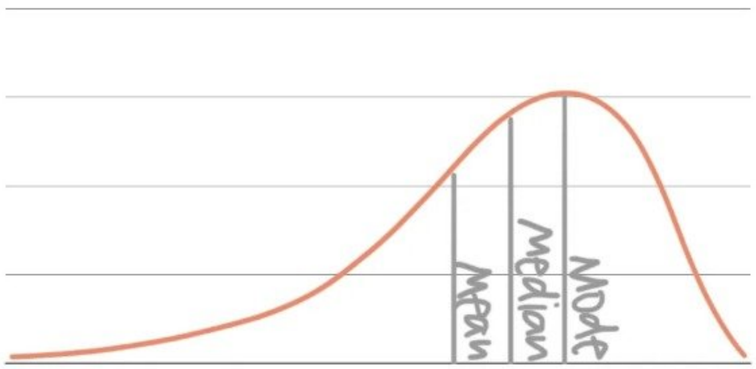
The tail is at left so left skewed distribution

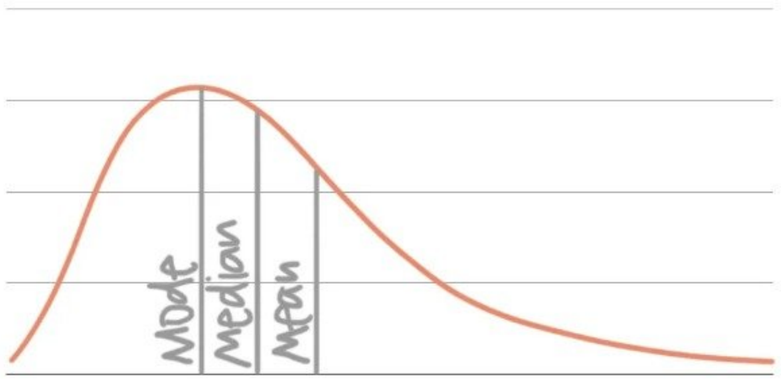


The tail is at right so right skewed distribution



DENSITY CURVE





- Now always in skewed distributions mean will be close to the tail.
- the mode will be there where most of the data exists.
- and median will be in centre as usual.

DENSITY CURVE

Usually the skewed distribution occurs when there is an outlier in the data for example. we have a data -

1, 2, 2, 3, 4, 4, 5, 5, 6, 7, 7, 8, 98, 99

DENSITY CURVE

Usually the skewed distribution occurs when there is an outlier in the data for example. we have a data -

1, 2, 2, 3, 4, 4, 5, 5, 6, 7, 7, 8, 98, 99

If we plot the points it will look like this.

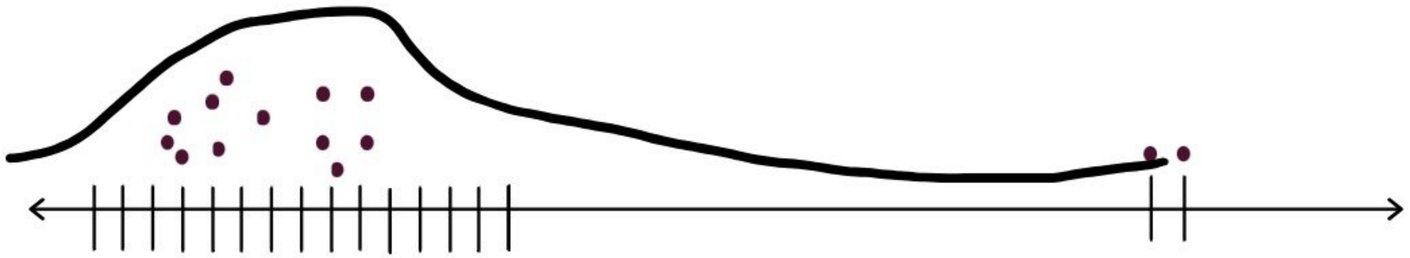
DENSITY CURVE

Normal (MCT)-> mean & (MofS)-> SD

Skewed (MCT)->median & (MofS)-> IQR

As you can see we have a right skewed distribution and if you have s skewed distribution its better to find the median than mean for central tendency.

And if we talk about measure of spread its better to check the IQR instead of Standard deviation.



DENSITY CURVE

As you can see we have a right skewed distribution and if you have s skewed distribution its better to find the median than mean for central tendency.

And if we talk about measure of spread its better to check the IQR instead of Standard deviation.

Vice versa works if you have a symmetric distribution its better to use mean and standard deviation for finding measure of spread and measure of central tendency

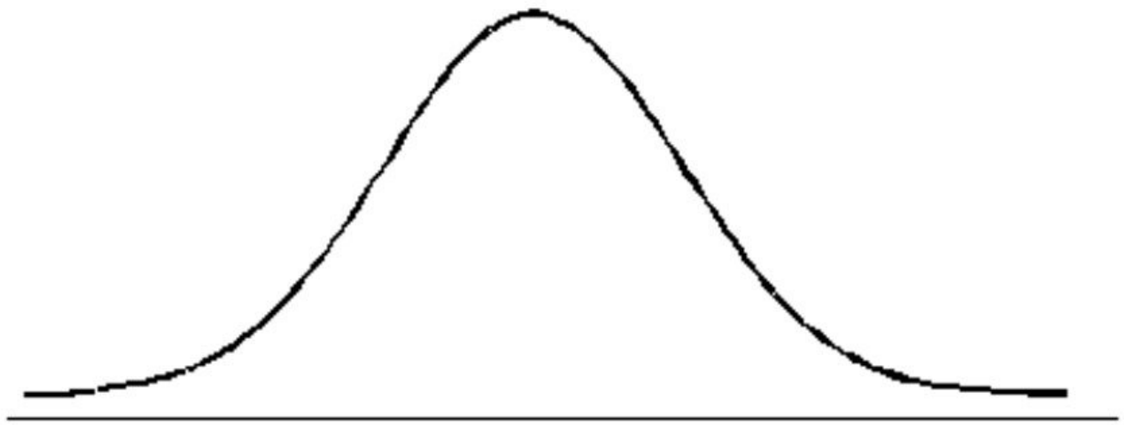
Z - SCORE

Data distribution

Z - SCORE

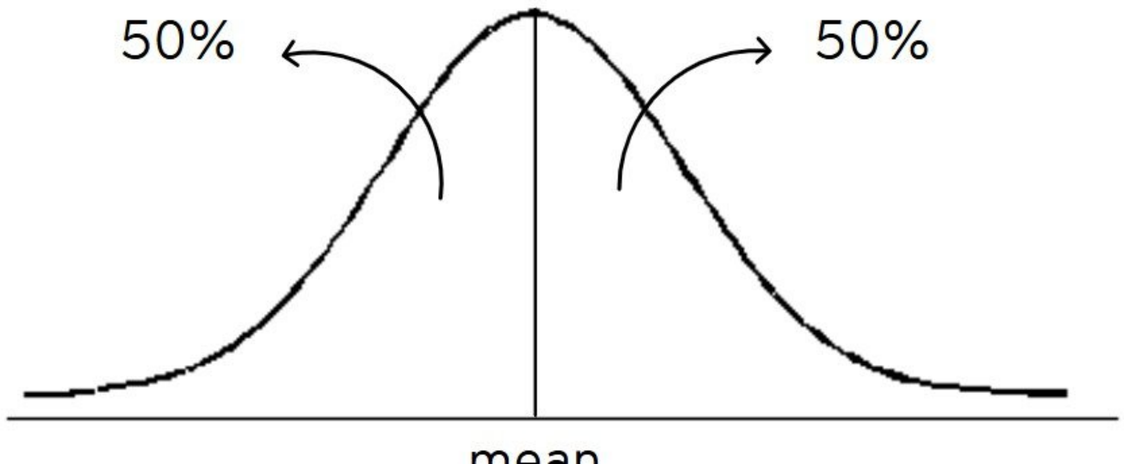
Now we have already seen the normal distribution and they are very import in data as well as probability distribution.

The area under the normal distribution is always 100% or 1.



Z - SCORE

The mean is in centre and 50% data is on left and 50% data is on right.

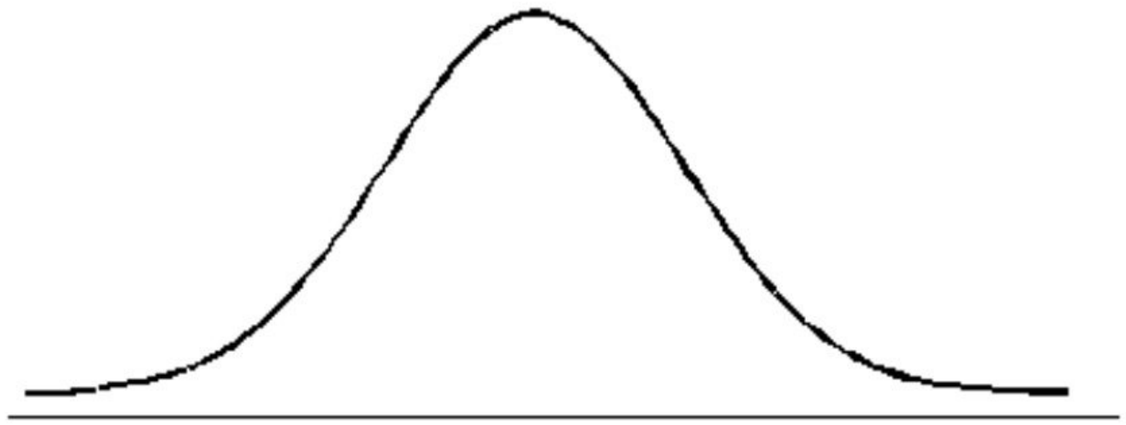


Z - SCORE

Now this Normal distribution also follows the empirical formula.

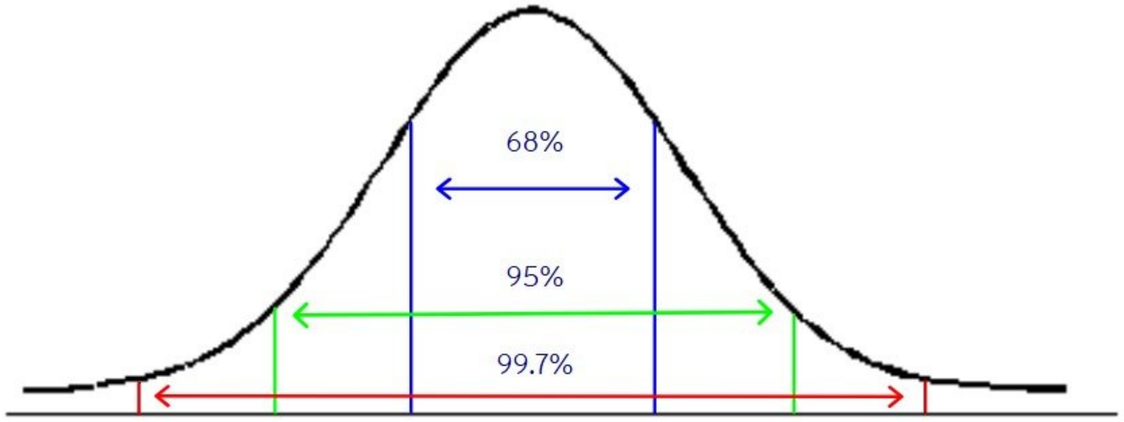
Empirical is a kind of a rule which gets followed in normal distribution.

the rule is simple - 68 - 95 - 99.7 rule



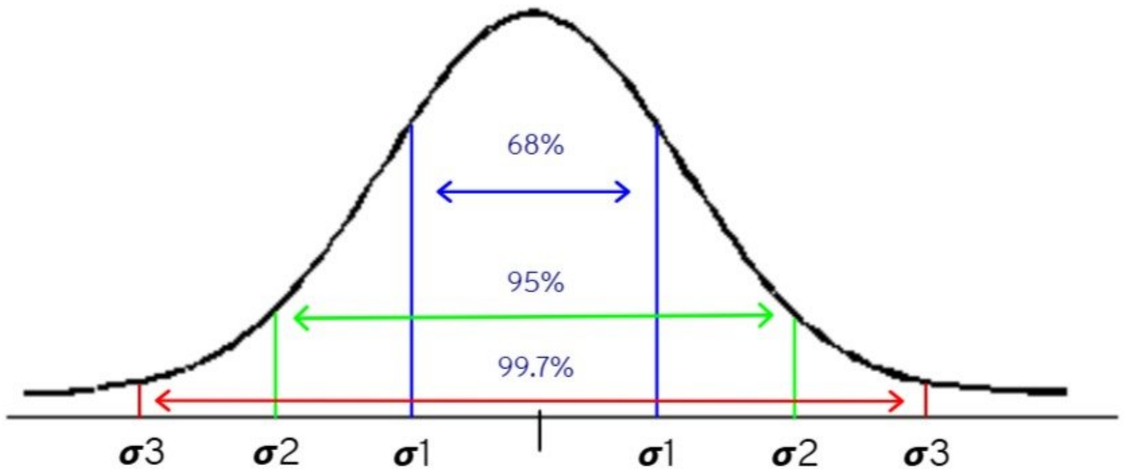
Z - SCORE

68 - 95 - 99.7 rule - you must be thinking what is this rule. we divide the distribution in 3 sections and those sections hold respective percentages.



Z - SCORE

Now the question is how these lines are made we make? These lines are divided using standard deviation points.



Z - SCORE

Now lets see with an example with a consistent data it will follow a normal distribution.

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

Now here we will find the

Mean - 5

Standard Deviation - 2.73

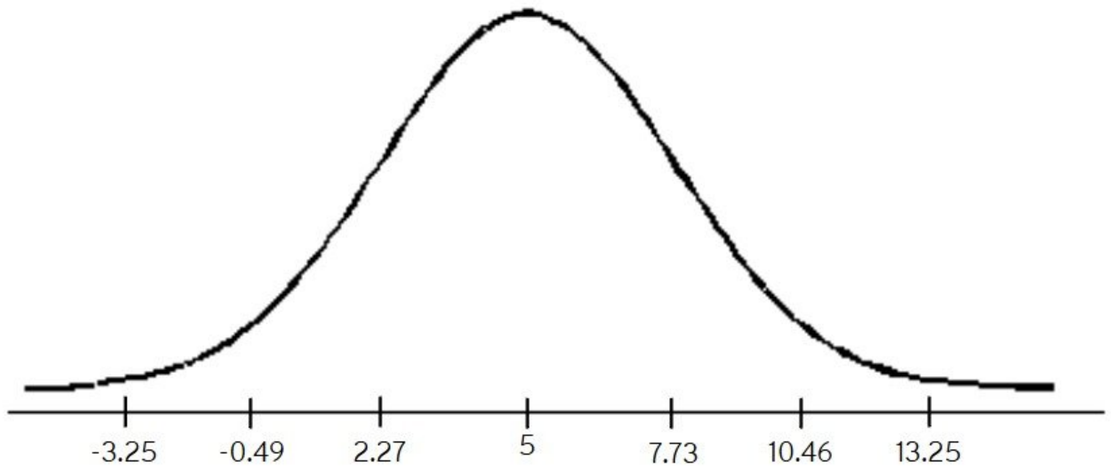
Z - SCORE

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

Mean - 5

Standard Deviation - 2.73

Now plot these standard deviation points in the graph and it will follow a bell shaped curve.



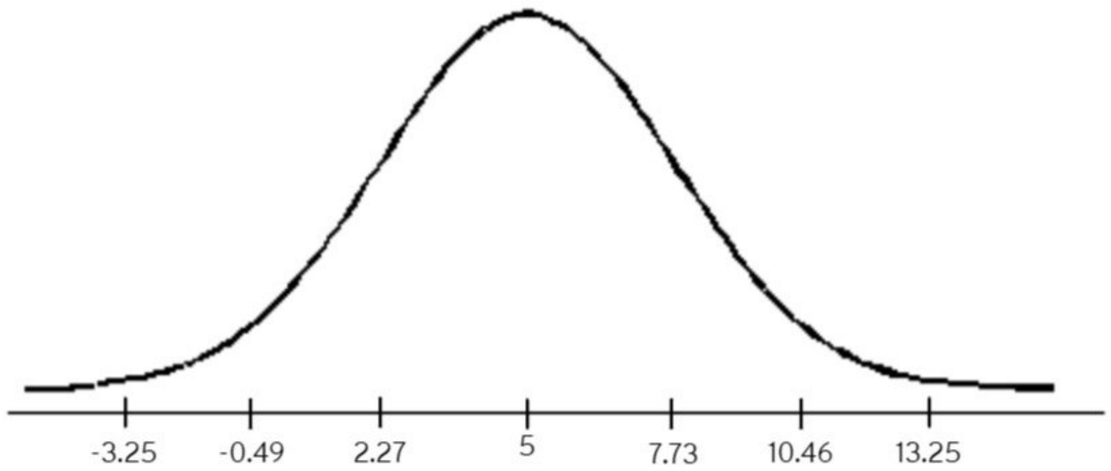
Z - SCORE

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

Mean - 5

Standard Deviation - 2.73

Now if you see clearly all the data or we can say 99.7% of our data will be inside 3 standard deviated points

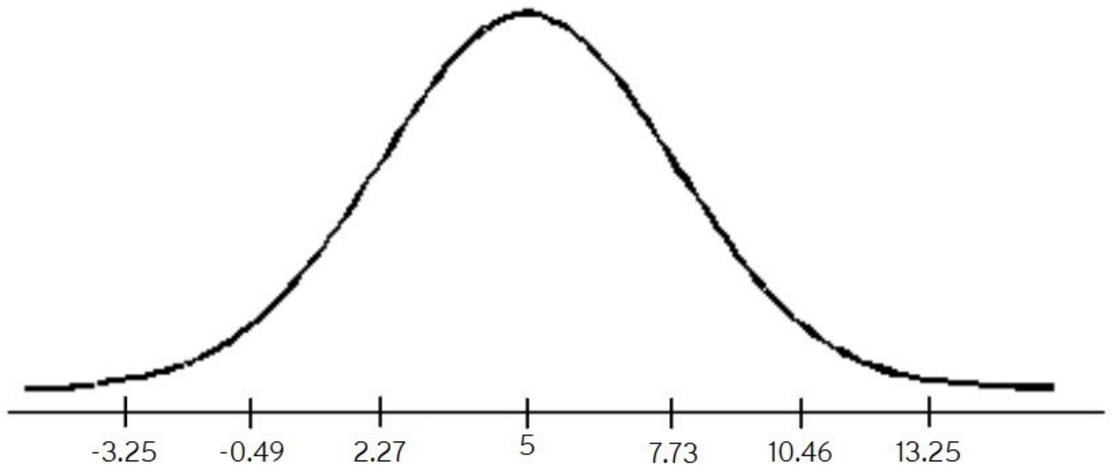


Z - SCORE

Ok we got what is empirical formula but what about Z score .

Two Questions - What is the use of Z score.

- What is the formula.



Z - SCORE

What is the use of Z score.

The **Z-score** is a powerful statistical tool used to understand how far a data point is from the mean of a dataset, measured in terms of standard deviations.

Z - SCORE

What is the formula?

Z-Score = (data point - mean)/ standard deviation

Z - SCORE

So lets say you have a data with mean on 8 and standard deviation is 1

and you have been asked how many standard deviation points away is 9.5 from the mean.

Z- score = (9.5 - 8)/1 = 1.5

That means 9.5 is 1.5 standard deviation points away from the mean.

if the value is in negative that means left side of the mean and if its positive that means right side of the mean.

Z - SCORE

So lets say you have a data with mean on 8 and standard deviation is 1

and you have been asked how many standard deviation points away is 9.5 from the mean.

Z- score = (9.5 - 8)/1 = 1.5

What other thing is Z-score telling us :

Z score can calculate the total percentage of data below the specified point.

To do this we will use Z- Table.

Z - SCORE

Search for Z-table on google and find out the value for 1.50.

we hot .93319

Multiply it with 100 to convert the value in percentage out of 100

so we got that below the data point 10 there are 93.32 percentage data.

COVARIANCE AND CORRELATION

Data distribution

COVARIANCE

Covariance tell how 2 variables change together. Covariance is used majorly in machine learning but it is a part of statistics.

Now if you don't remember what are variables.

Reminder when we have a table and in that table we want to see the connection between 2 columns we use covariance.

	Age	**Salary**
	22	26k
	23	34k
	24	40k
	25	45k
	26	50k

COVARIANCE

Now we have to check the relation between two variables Age and salary we will use covariance.

There are two types of relation.

- Positive Covariance
- Negative Covariance

	Age	**Salary**
	22	26k
	23	34k
	24	40k
	25	45k
	26	50k

COVARIANCE

Positive covariance - When one variable increases the other variable tends to increase like we can see in this example when Age is increasing the salary is increasing.

Negative covariance - When one variable increases the other variable decreases thats the negative covariance. eg time is increasing the ranks are decreasing.

	Age	Salary
	22	26k
	23	34k
	24	40k
	25	45k

	time	rank
	6	14
	7	10
	8	8
	9	4

COVARIANCE

To calculate the covariance we have to use the formula.

$$\mathrm{Cov}(X,Y) = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

After applying this formula we get the covariance of 14.75.

	Age	Salary
	22	26k
	23	34k
	24	40k
	25	45k
	26	50k

COVARIANCE

To calculate the covariance we have to use the formula.

$$\mathrm{Cov}(X,Y) = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

After applying this formula we get the covariance of 14.75.

	Age	Salary
	22	26k
	23	34k
	24	40k
	25	45k
	26	50k

COVARIANCE

So what covariance tells us basically it just tells us the direction the relation is positive or negative but to actually see the strength of relation we have to use the coorelation.

CORRELATION

So what covariance tells us basically it just tells us the direction the relation is positive or negative but to actually see the strength of relation we have to use the correlation.

CORRELATION

To find the correlation we have to find the covariance first and then divide it by standard deviation of both the variables.

$$r = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

CORRELATION

So with this data our covariance was- 14.75

Now we just have to divide it by the standard deviation of both variables.

Age σ = 1.58

Salary σ = 9.38

correlation = 14.75/14.82 = 0.99

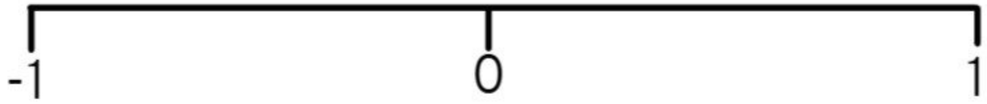
	Age	**Salary**
	22	26k
	23	34k
	24	40k
	25	45k
	26	50k

CORRELATION

So our correlation of these 2 variables are coming out to be 0.99 but what is it telling.

It is telling the strength how closely they are related to each other.

correlation will always come between -1 and 1.



	Age	Salary
	22	26k
	23	34k
	24	40k
	25	45k
	26	50k

PROBABILITY

probability and permutation & combination

PROBABILITY

Think about the probability of an event occurring as the likelihood that the event will occurre.

```
probability
=
LikeliHood
```

PROBABILITY

We will start by understanding how to find the likelihood of getting a head or a tail in a coin flip, or how to calculate the likelihood of a card in a deck of card or dice problems.

PROBABILITY

We will start by understanding how to find the likelihood of getting a head or a tail in a coin flip, or how to calculate the likelihood of a card in a deck of card or dice problems.

PROBABILITY

Statistics and probability go hand in hand, cause statistics is all about data and analysing data.

Probability is going to allow us to determine how reliable our statistical result actually are. So we really like to study them together.

PROBABILITY

Now lets start with simple probability -

A probability will always comes between 0 and 1. and its tells the likely hood of an event occurring. we can also tell it in percentage form or fraction form its same thing.

PROBABILITY

How to write and calculate the probability -

Lets take an example of a coin flip and the question is what is the probability of getting a head.

To solve this we have to use this

P(H) = True Outcome / Total Outcome

P(H) = 1/2

PROBABILITY

So here are some basic question.

What is the probability of getting a heart in deck of cards.

What is the probability of getting a 4 when we roll a dice.

A box has 3 green and 2 blue chocolates what is the probability of getting 2 green chocolates in 2 tries.

PROBABILITY

Now lets say you flipped the coin 3 times and every time you got a tail. Now what is the probability of getting a tail.

you will say it is - p(T) = 3/3 and that is 100%

But if we see it generally its not right.

this probability is because of an experiment we did and thats why this is known as experimental probability.

if you flip the coin infinite times at the end you will eventually get the probability of 1/2

UNION VS INTERSECTION

probability and permutation & combination

ADDITION RULE

First we have to see the addition rule to see this we will take one simple example.

You have flipped a coin 2 times what is the probablity of getting at least 1 heads.

ADDITION RULE

You have flipped a coin 2 times what is the probability of getting at least 1 heads.

First lets see how many outcomes we can get.

HH

HT

TH

TT

ADDITION RULE

You have flipped a coin 2 times what is the probability of getting at least 1 heads.

First lets see how many outcomes we can get.

HH, HT, TH, TT

So if we see First outcome is as head is 2 times and Second outcome as head is also 2 Times.

So the probability is - p(h1 or h2) = 2/4 + 2/4 = 1 or 100%

ADDITION RULE

You have flipped a coin 2 times what is the probability of getting at least 1 heads.

First lets see how many outcomes we can get.

HH, HT, TH, TT

So the probability is - p(h1 or h2) = 2/4 + 2/4 = 1 or 100%

But this is not possible cause we can see one of the outcome is TT so it's impossible to have 100%.

To understand this we will see a ven diagram.

ADDITION RULE

You have flipped a coin 2 times what is the probability of getting at least 1 heads.

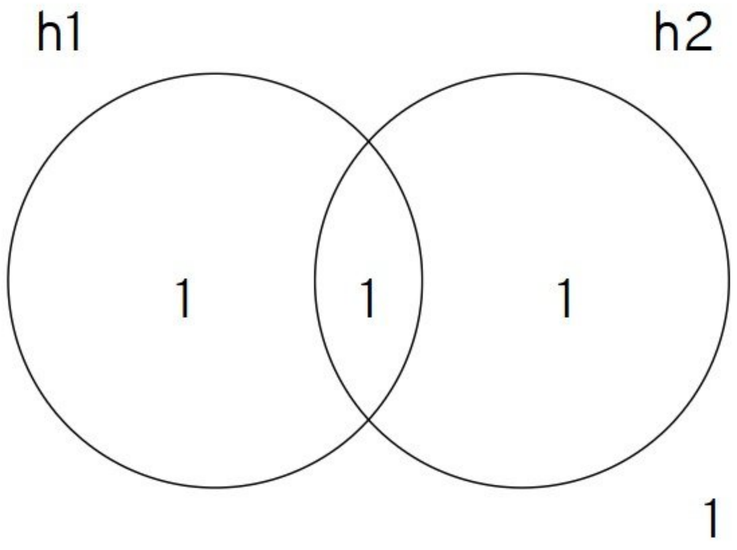
First lets see how many outcomes we can get.

HH, HT, TH, TT

So the probability is - p(h1 or h2) = 2/4 + 2/4 = 1 or 100%

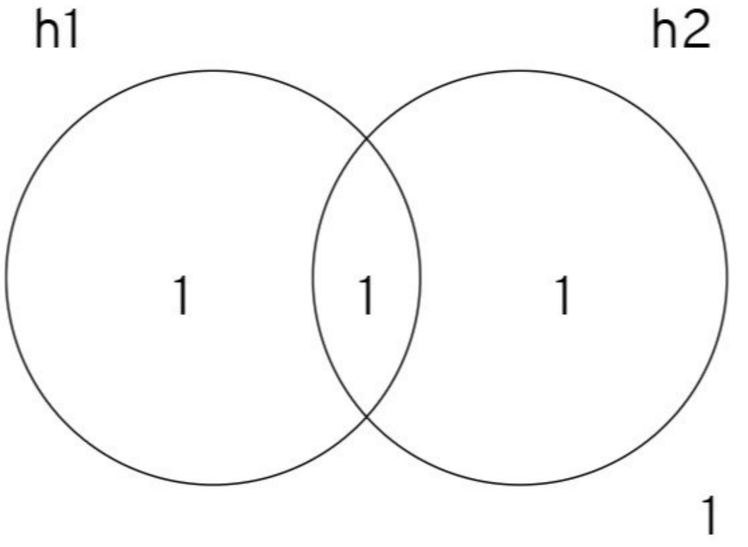
But this is not possible cause we can see one of the outcome is TT so it's impossible to have 100%.

To understand this we will see a ven diagram.



ADDITION RULE

Usually this thing happens when we have at least and we have to add the probability and this is why it is known as addition rule.



ADDITION RULE

So probability of A or B happening is

p(A or B) = p(A) + p(B) - p(A and B)

But if we have mutually exclusive event like (you roll a dice and getting a 3 or 4)

so probability of p(3) and p(4) = 0

therefore for mutually exclusive event the formula is

p(A or B) = p(A) + p(B)

ADDITION RULE

Addition Rule

p(A or B) = p(A) + p(B) - p(A and B)

we can also write this as

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

$\cup$ ➡️ this sign represent the union or say OR

$\cap$ ➡️ This sigh represent the intersection or say AND

More or less the same thing.

ADDITION RULE

So if we have data like this and you have to find the probability of male or red.

Color	male	Female	total
Red	22	16	38
white	13	8	21
Blue	25	16	41
Total	60	40	100

ADDITION RULE

So if we have data like this and you have to find the probability of male or red.

Color	male	Female	total
Red	22	16	38
white	13	8	21
Blue	25	16	41
Total	60	40	100

So we have to find the p(male U red) = p(male) + p(red) - p(male ∩ red)

p(male U red) = 60/100 + 38/100 - 22/100

p(male U red) = 19/25 or 0.79

INDEPENDENT AND DEPENDANT EVENT

probability and permutation and combination

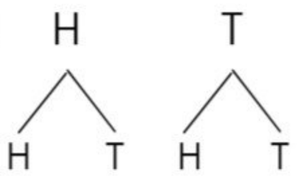
INDEPENDENT EVENTS

As the name tells independent events are independent to each other but the best way to understand is with an example.

INDEPENDENT EVENTS

Lets say you are flipping a coin the outcomes are H and T.

Now you flipped the coin again now the outcomes can be

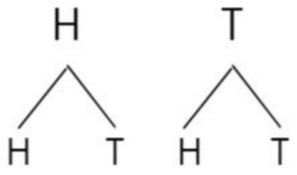


So total 4 outcomes will be there -> HH, HT, TH, TT.

INDEPENDENT EVENTS

Lets say you are flipping a coin the outcomes are H and T.

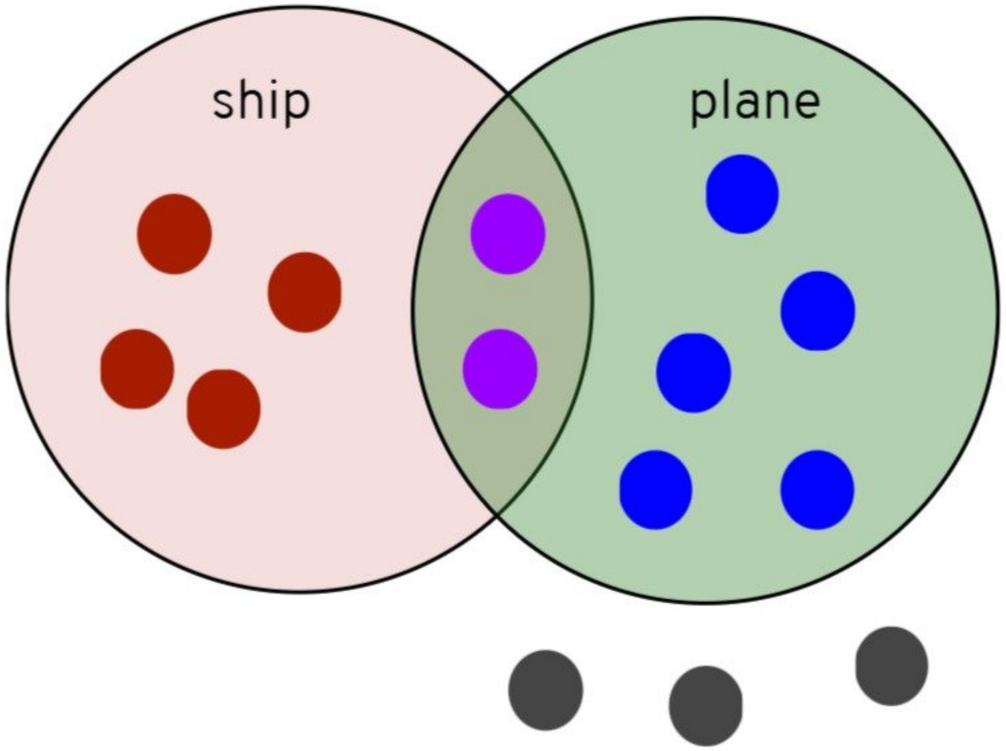
Now you flipped the coin again now the outcomes can be



That means when we have independent events we don't have any effect on the outcomes of any events. so probability of getting tails on both flips we will be following multiplication rule .

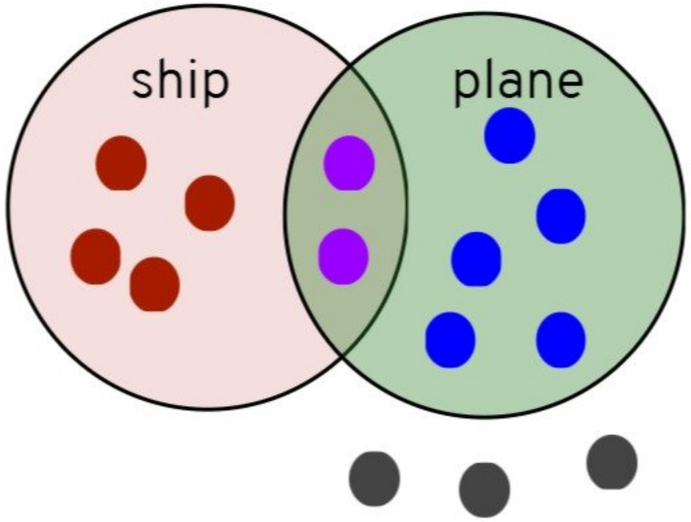
so $p(A \cap B) = p(A) * p(B)$

CONDITIONAL PROBABILITY



CONDITIONAL PROBABILITY

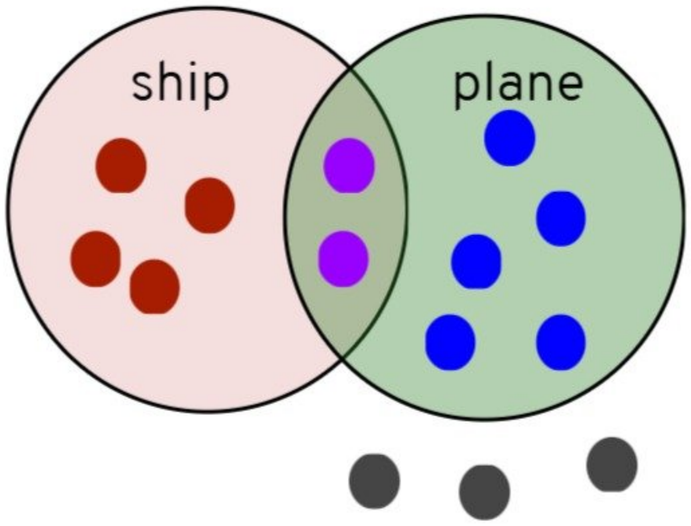
We can create a contingency table for this.



	loves ship	**no ship**
love plane	2	5
no plane	4	3

CONDITIONAL PROBABILITY

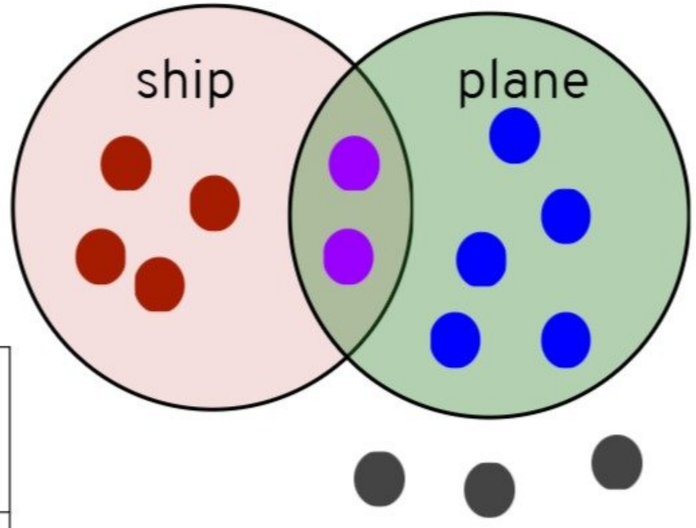
We can create a contingency table for this.



	loves ship	**no ship**
love plane	2 p - 2/14	5 p-5/14
no plane	4 p- 4/14	3 p- 3/14

CONDITIONAL PROBABILITY

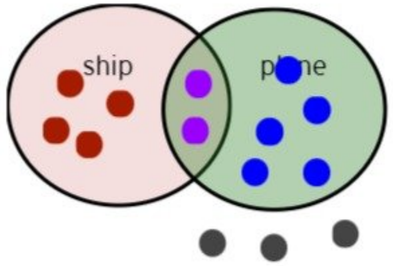
We can create a contingency table for this.



	loves ship	**no ship**	**row total**
love plane	2 p - 2/14	5 p-5/14	2 + 5 p - 7/14
no plane	4 p- 4/14	3 p- 3/14	4 + 7 7/14
column total	2 + 4 p- 6/14	5+3 p - 8/14	

CONDITIONAL PROBABILITY

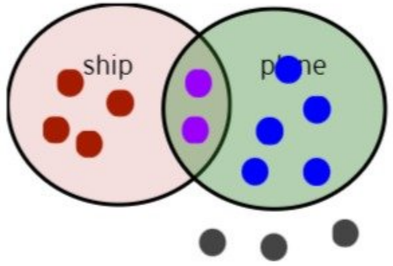
	loves ship	no ship	row total
love plane	2 p - 2/14	5 p-5/14	2 + 5 p - 7/14
no plane	4 p- 4/14	3 p- 3/14	4 + 7 7/14
column total	2 + 4 p- 6/14	5+3 p - 8/14	



Now a person came who loves plane what is the probability he also loves ship.

CONDITIONAL PROBABILITY

	loves ship	no ship	row total
love plane	2 p - 2/14	5 p-5/14	2 + 5 p - 7/14
no plane	4 p- 4/14	3 p- 3/14	4 + 7 7/14
column total	2 + 4 p- 6/14	5+3 p - 8/14	

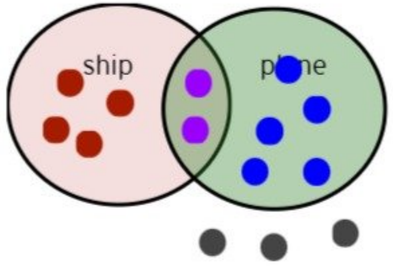


Now a person came who loves plane what is the probability he also loves ship

p(person who loves P and S | person loves plane)

CONDITIONAL PROBABILITY

	loves ship	no ship	row total
love plane	2 p - 2/14	5 p-5/14	2 + 5 p - 7/14
no plane	4 p- 4/14	3 p- 3/14	4 + 7 7/14
column total	2 + 4 p- 6/14	5+3 p - 8/14	

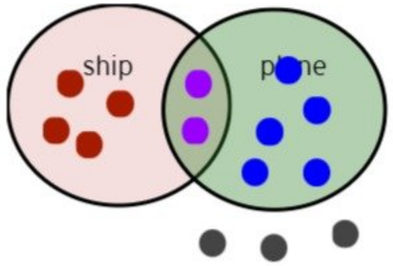


p(person who loves P and S | person loves plane) =

p(person who loves P and S) / p(person loves plane)

CONDITIONAL PROBABILITY

	loves ship	no ship	row total
love plane	2 p - 2/14	5 p-5/14	2 + 5 p - 7/14
no plane	4 p- 4/14	3 p- 3/14	4 + 7 7/14
column total	2 + 4 p- 6/14	5+3 p - 8/14	



$p(A|B) = p(A)/p(B)$

p(person who loves P and S | person loves plane) =

p(person who loves P and S) / p(person loves plane)

(2/14) / (7/14)

DEPENDANT EVENT

Dependant events means if the outcome of first event affects the outcome of second event.

there are 3 white balls and 2 red balls what is the probability of getting 2 white balls.

$p(A \cap B) = p(A) * p(B|A)$

QUESTIONS

- A die is rolled. What is the probability of getting an even number or a number less than 3?

- A card is drawn from a deck. What is the probability of getting a king or a red card?

QUESTIONS

- In a class of 40 students, 18 like Maths, 24 like Science, and 10 like both. What is the probability that a student chosen at random likes either Maths or Science?

- A bag contains **4 red** and **3 blue** marbles. Two marbles are drawn **without replacement**.
  What is the probability that both marbles are red?

BAYES THEOREM

Agar hume kisi result ke milne ka reason pata ho, to Bayes' Theorem hume yeh batata hai ki us result se hume wo reason mila ya nahi.

Ek bande ko corona hai, aur mujhe pata hai ki wo kisi corona patient ke contact mein aaya tha. Toh Bayes' Theorem meri intuition ko support karti hai.

BAYES THEOREM

Let's break it down:

- **A = Banda corona positive hai**
- **B = Banda kisi corona wale ke contact mein aaya tha**

Aapko mila hai **B (evidence)**, aur aap soch rahe ho:

> *"B milne ke baad, A hone ka chance kitna hai?"*

Yehi to hai: BAYES theorem

$$P(\text{Corona}|\text{Contact}) = \frac{P(\text{Contact}|\text{Corona}) \cdot P(\text{Corona})}{P(\text{Contact})}$$

BAYES THEOREM

$$P(\text{Corona}|\text{Contact}) = \frac{P(\text{Contact}|\text{Corona}) \cdot P(\text{Corona})}{P(\text{Contact})}$$

=

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

BAYES THEOREM

Lets say we have 2 dice one of the dice is Fair and another dice is unfair it is weighted as it is getting 6 50% of the time .

1. Fair dice
2. Unfair dice - 50% times 6

BAYES THEOREM

1. Fair dice
2. Unfair dice - 50% times 6

Scenario - we choose one of the die without knowing which die I picked.

we roll that dice and got a 6.

Question - What is the probability that we picked the biased die given that we rolled a 6.

BAYES THEOREM

A - choosing the biased die

B - Rolling a 6

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

so

p(A|B) = p(Biased|6)

p(B|A) = p(6|Biased)

p(A) = p(Biased)

p(B) = p(6)

BAYES THEOREM

A - choosing the biased die

B - Rolling a 6

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

so

p(A|B) = p(Biased|6)

p(B|A) = p(6|Biased) -> 1/2

p(A) = p(Biased) -> 1/2

p(B) = p(6) -> lets calculate

BAYES THEOREM

A - choosing the biased die

B - Rolling a 6

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

so

p(A|B) = p(Biased|6)

p(B|A) = p(6|Biased) -> 1/2

p(A) = p(Biased) -> 1/2

p(B) = p(6) -> p(biased -> 6) + p(fair -> 6)

(1/2 * 1/2) + (1/2 * 1/6)

1/4 + 1/12

1/3

BAYES THEOREM

p(A|B) = p(Biased|6)

p(B|A) = p(6|Biased) -> 1/2

p(A) = p(Biased) -> 1/2

p(B) = p(6) -> p(biased -> 6) + p(fair -> 6)

(1/2 * 1/2) + (1/2 * 1/6)

1/4 + 1/12

1/3

Now we just have to plug in the values to find p(A|B) = (1/2 *1/2)/1/3

the final result is 3/4.

BAYES THEOREM

Now we just have to plug in the values to find $p(A|B) = (1/2 * 1/2)/1/3$ the final result is 3/4.

So the probability of selecting a biased die when we already rolled a 6 is 3/4.

BAYES THEOREM

> ***Bayes' Theorem batata hai ki mere intuition ko nayi evidence milne ke baad kitna support milta hai."***

Ya aur casually:

> ***"Bayes' Theorem = Mere gut feeling + Maths se confirmation!"*** 😄

DESCRIPTIVE VS INFERENTIAL

	Descriptive	**Inferential**
Purpose	Summarize & describe data	Make predictions or draw conclusions from data
Works with	**Actual data** you collected	Goes **beyond the data** to generalize to a population
Examples	Mean, Median, Mode, Range, Standard Deviation, Graphs	Hypothesis Testing, Confidence Intervals, p-values, t-tests
Focus	"What does the data say?"	"What can we infer about the population?"

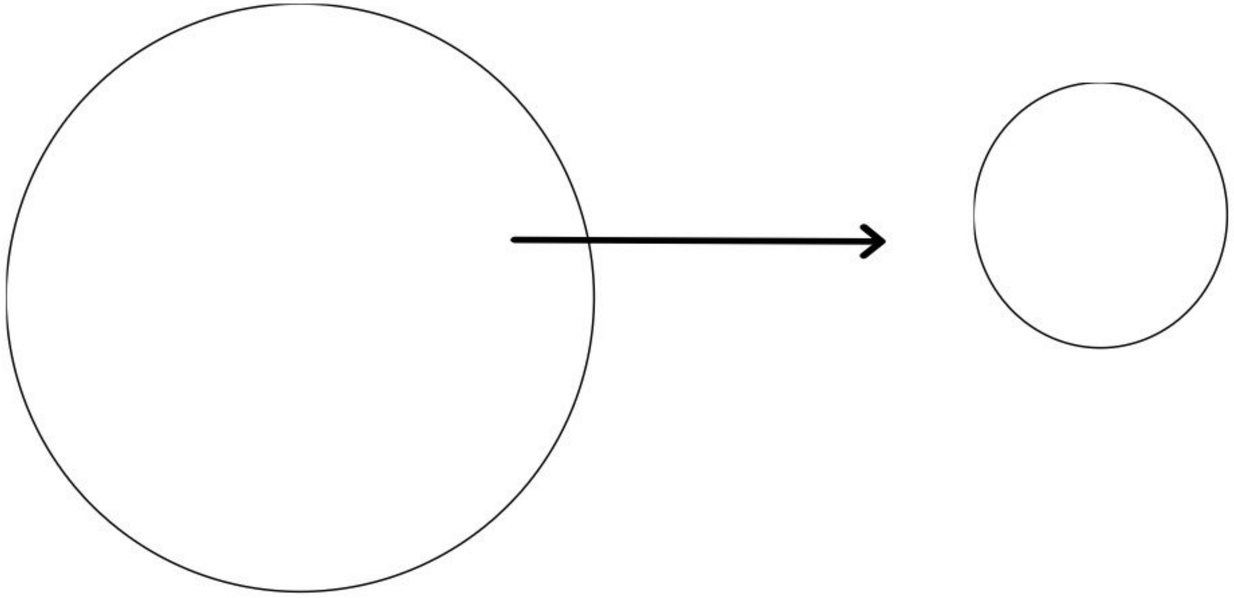
DESCRIPTIVE VS INFERENTIAL

Let's say you survey **100 students** on their exam scores.

- **Descriptive Stats**:
  "The average score is 72.3, and the standard deviation is 8.5."
  → You're just **describing what you observed**.
- **Inferential Stats**:
  "We are 95% confident that the average score of **all students in the school** is between 70 and 75."
  → You're **inferring** something about a **larger population** based on your sample.
- We have already covered descriptive stats and now it's time to start with inferential stats.

INFERENTIAL STATS

It's the branch of statistics that helps us **make predictions or decisions about a population** based on a **sample** of data.



HYPOTHESIS TESTING

Hypothesis Testing is a **statistical method** used to decide whether there's enough evidence in a sample to support a claim about a population.

Lets see this with an example and understand how hypothesis testing works.

HYPOTHESIS TESTING

I have a coin and I want to test weather the coin is fair of not by flipping in 100 times.

$p(h) = 0.5$ and $p(t) = 0.5$

HYPOTHESIS TESTING

I have a coin and I want to test weather the coin is fair of not by flipping in 100 times.

So if I focus on heads and I flip the coin 100 times I should get head 50 times and if this happens the coin is fair.

HYPOTHESIS TESTING

I have a coin and I want to test weather the coin is fair of not by flipping in 100 times.

Now to test the coin is fair or not I will do the hypothesis testing and for hypothesis testing we have to define 2 hypothesis null and alternate.

HYPOTHESIS TESTING

I have a coin and I want to test weather the coin is fair of not by flipping in 100 times.

1) Null hypothesis - Coin is Fair.

2) Alternate Hypothesis - Coin is unfair.

Null hypothesis is always see the positive side cause there is a quote 'innocent until proven guilty'.

and alternate hypothesis is always opposite to null hypothesis.

HYPOTHESIS TESTING

I have a coin and I want to test weather the coin is fair of not by flipping in 100 times.

1) Null hypothesis - Coin is Fair.

2) Alternate Hypothesis - Coin is unfair.

3) now we will do the experiment

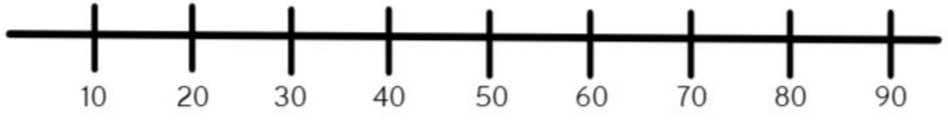
4) after Experimenting we will Accept or reject the null hypothesis.

HYPOTHESIS TESTING

I have a coin and I want to test weather the coin is fair of not by flipping in 100 times.

Experiment

Lets say before experiment we got the information that our mean is 50 and standard deviation is 10. and we have a normal distribution.

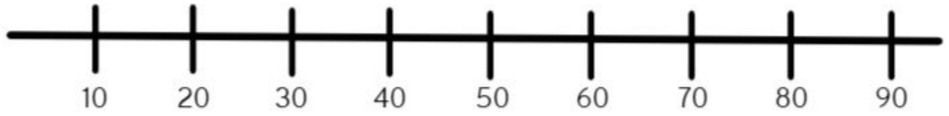


HYPOTHESIS TESTING

I have a coin and I want to test weather the coin is fair of not by flipping in 100 times.

Experiment

Now I performed the experiment and I got head 30 times so can we say our coin is fair.

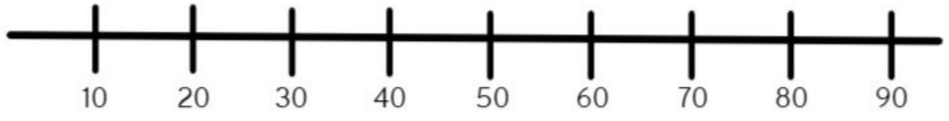


HYPOTHESIS TESTING

I have a coin and I want to test weather the coin is fair of not by flipping in 100 times.

Experiment

So to see the wether the coin is fair or not our experiment head count must be close to mean but how much close for this we will use a important concept of significance value.



HYPOTHESIS TESTING

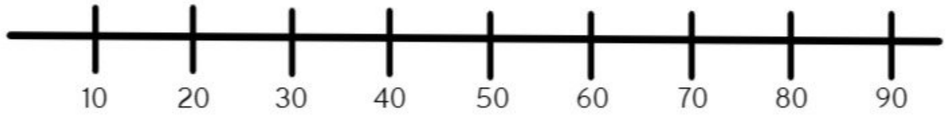
I have a coin and I want to test weather the coin is fair of not by flipping it 100 times.

Significance value is set by the domain experts and lets say the significance value is 0.05 and it is donated by α and using this significance value we calculate the confidence interval and that will be 1 - 0.05 = .95

HYPOTHESIS TESTING

I have a coin and I want to test weather the coin is fair of not by flipping it 100 times.

Significance value is set by the domain experts and lets say the significance value is 0.05 and it is donated by α and using this significance value we calculate the confidence interval and that will be 1 - 0.05 = .95



TYPE 1 AND TYPE 2 ERROR

We saw when we were doing the hypothesis testing we are having null hypothesis and alternate hypothesis and at the end we have to check wether we accept the null hypothesis or reject the null hypothesis.

TYPE 1 AND TYPE 2 ERROR

now we have reality -

where either the null hypothesis is true or it is false.

And then we have decision -

where we will decide the null hypothesis is true or false.

TYPE 1 AND TYPE 2 ERROR

So based on this information we will have 4 outcomes :

1) We reject the null hypothesis in reality it was false.
2) We reject the null hypothesis in reality it was true.
3) We accept the null hypothesis in reality it was True.
4) We accept the null hypothesis in reality it was false.

TYPE 1 AND TYPE 2 ERROR

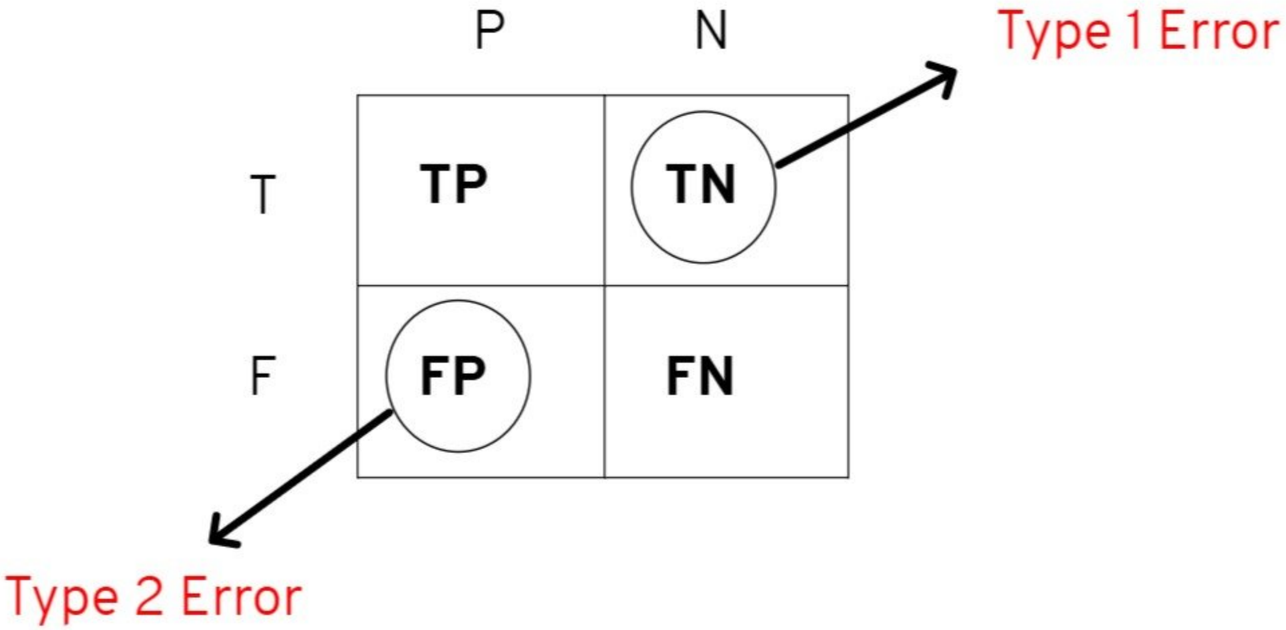
So based on this information we will have 4 outcomes :

1) We reject the null hypothesis in reality it was false.
2) We reject the null hypothesis in reality it was true. Type 1 error
3) We accept the null hypothesis in reality it was True.
4) We accept the null hypothesis in reality it was false. Type 2 error

TYPE 1 AND TYPE 2 ERROR

So based on this information we will have 4 outcomes :

Based on this in future we will also learn about confusion matrix used in classification problem.



ONE TAILED AND TWO TAILED TEST

For this first I want to give you an example we are not solving anything right now I am just providing an example to understand one tailed and two tailed test.

ONE TAILED AND TWO TAILED TEST

Question -

You created an app that helps students prepare for exams. You claim that students who use your app score **more than 70 marks on average** in a particular subject.

Now, you want to test this claim using statistics.

You collect a sample of 30 students who used the app and calculate their average score.

ONE TAILED AND TWO TAILED TEST

Question -

You created an app that helps students prepare for exams. You claim that students who use your app score **more than 70 marks on average** in a particular subject.

Now, you want to test this claim using statistics.

You collect a sample of 30 students who used the app and calculate their average score.

Here in this question you are only interested in finding the score are more than 70 marks on average. so here we will apply one tailed test.

ONE TAILED AND TWO TAILED TEST

Question -

You created an app that helps students prepare for exams. You claim that students who use your app score **more than 70 marks on average** in a particular subject.

Now, you want to test this claim using statistics.

You collect a sample of 30 students who used the app and calculate their average score.

Here in this question you are only interested in finding the score are more than 70 marks on average. so here we will apply one tailed test.

HYPOTHESIS TEST

Z-test

Z-TEST

We will use Z-test when the sample size is greater than 30.

And you must also know population variance or standard deviation.

Z-TEST

WHAT DOES IT CHECK?

> *"Is the sample mean **significantly different** from the population mean?"*

Z-TEST

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

- $\bar{x}$ = Sample mean
- μ = Population mean
- σ = Population standard deviation
- n = Sample size

Z-TEST

- Sample mean = 74
- Population mean = 70
- Population standard deviation = 10
- Sample size = 36

Z-TEST

- Sample mean = 74
- Population mean = 70
- Population standard deviation = 10
- Sample size = 36

This will give us the result of 2.4

2.4 → This means the sample mean is **2.4 standard deviations** away from the population mean.

Now, look up this Z value in a **Z-table** to get the **p-value**:

Z-TEST

If p-value	Meaning	What to Do
$p \leq \alpha$	Result is **statistically significant**	❌ Reject the null hypothesis (H_0)
$p > \alpha$	Not enough evidence	✅ Fail to reject H_0

So here we got the p value of 0.0164 is it greater or smaller than significance.

it is smaller so we reject the null hypothesis.

Z-TEST

The average marks of Sheryians students are 80 with the population standard deviation of 2.5 but a teacher believes the mean to be different so he took the marks of 36 individuals and found the average to be 82.5

Z-TEST

A health researcher claims that the average weight of adult males in a city is 70 kg.
A sample of 49 men is taken, and their average weight is found to be 72 kg.
Assume population standard deviation (σ) is 8 kg.

👉 At 5% significance level, **test the claim** using a one-sample Z-test.

Z-TEST

A company claims that its new software bug fix takes **no more than 30 minutes** on average.
A sample of 36 fixes shows an average time of 32 minutes.
Population standard deviation is known to be 6 minutes.

👉 Test the company's claim at the 0.01 significance level.

Z-TEST

It is believed that the average score in a statistics exam is **75**.
A professor thinks students this year have performed **worse**.
A random sample of 40 students shows an average score of 72,
with a population standard deviation of 10.

👉 Test the professor's belief at the 0.05 significance level.

T-TEST

When we have unknown population standard deviation we will use T-Test

Sample size is can be small than 30

T-TEST

formula is simple

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

T-TEST

But in t test we will calculate the degree of freedom and the formula for degree of freedom is n-1.

Now lets solve a simple example to get how T-test works.

T-TEST

A bakery claims that their cupcakes weigh **at least 150 grams** on average.
A customer suspects this isn't true.
They take a random sample of 10 cupcakes and find:

- Sample mean = **145 grams**
- Sample standard deviation = **8 grams**

👉 At a 5% significance level, test the customer's claim using a one-sample **T-Test**.

T-TEST

A professor believes students study an average of **20 hours/week**.
To test this, a sample of 15 students was taken.
They found:

- Mean = **18.5 hours**
- Sample standard deviation = **3.2 hours**

👉 Test the professor's belief at a 0.05 significance level.

T-TEST

A nutritionist wants to verify if a new diet plan helps people lose **more than 5 kg** in a month. From a sample of 15 people, the **average weight loss** was **5.6 kg**, with a **standard deviation of 1.2 kg**.

Test the hypothesis at a 5% significance level. Can we conclude the diet helps reduce more than 5 kg?

TWO SAMPLES

You use **two-sample tests** (whether **T-Test** or **Z-Test**) when:

> *You have **two different groups (samples)** and you want to **compare their means** to see if there's a **statistically significant difference** between them.*

TWO SAMPLES

EXAMPLE 1: COMPARING TEST SCORES

- Group 1: Students from School A
- Group 2: Students from School B
  👉 Want to see if School A performs better than School B on average.

TWO SAMPLES

EXAMPLE 1: COMPARING TEST SCORES

- Group 1: Students from School A
- Group 2: Students from School B
  👉 Want to see if School A performs better than School B on average.
- So here we have to use Two sample test

TWO SAMPLES

Formula for
two sample Z-Test

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Formula for
two sample T-Test

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

df = min(n1-1,n2-1)

TWO SAMPLES

Important Note:
When choosing between a **Z-test** and a **T-test**, your main focus should be on the **type of standard deviation provided**.

👉 If the **sample standard deviation** is given (and **population standard deviation is unknown**), you should use the **T-test** — even if the sample size is greater than 30.

TWO SAMPLES

A teacher wants to know if there's a difference in math scores between boys and girls.

- Boys: n = 30, mean = 70, std dev = 8
- Girls: n = 28, mean = 74, std dev = 6

TWO SAMPLES

A hospital compares recovery times for two treatments:

- Treatment A: n = 20, mean = 8 days, std dev = 2
- Treatment B: n = 20, mean = 10 days, std dev = 3

CHI SQUARE TEST

The **Chi-Square Test** is used to check if there is a **significant association** between two categorical variables. It helps to answer questions like:

CHI SQUARE TEST

The **Chi-Square Test** is used to check if there is a **significant association** between two categorical variables. It helps to answer questions like:

- Is there a relationship between **gender** and **voting preference**?
- Do two different **marketing strategies** result in different **customer preferences**?
- Is there a **difference** between observed and expected frequencies in a survey or experiment?

CHI SQUARE TEST

The formula for the **Chi-Square statistic** is :

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

- O_i = **Observed frequency** (actual data)
- E_i = **Expected frequency** (what we expect under the null hypothesis)
- $\sum$ = Sum over all categories

CHI SQUARE TEST

Steps are similar

State the Hypotheses:

- **Null Hypothesis (H_0):** There is no association (or no difference) between the variables.
- **Alternative Hypothesis (H_1):** There is an association (or a difference).
- **Calculate Expected Frequencies (E):** For each cell in your data, the expected frequency is calculated using:

$$E_i = \frac{(Row\,Total \times Column\,Total)}{Grand\,Total}$$

CHI SQUARE TEST

Steps are similar

Calculate the Chi-Square statistic: Apply the formula.

Determine the Degrees of Freedom (df): For a **Test of Independence**:

$$df=(r-1)(c-1)$$

- Where **r** = number of rows and **c** = number of columns.

CHI SQUARE TEST

Let's say we want to check if there's a relationship between **Gender** and **Choice of Drink**. We survey 100 people, and the data is:

	Coffee	Tea	Juice	Total
Male	30	10	5	45
Female	15	20	20	55
Total	45	30	25	100

CHI SQUARE TEST

after solving the expected frequency are

	Coffee	Tea	Juice	Total
Male	30	10	5	45
Female	15	20	20	55
Total	45	30	25	100

Gender	Coffee	Tea	Juice	Row Total
Male	20.25	13.5	11.25	45
Female	24.75	16.5	13.75	55
Col Total	**45**	**30**	**25**	**100**

CHI SQUARE TEST

Now apply chi square test:

Gender	Drink	Observed (O)	Expected (E)	$\frac{(O-E)^2}{E}$
Male	Coffee	30	20.25	$\frac{(30-20.25)^2}{20.25}$ = **4.69**
Male	Tea	10	13.5	$\frac{(10-13.5)^2}{13.5}$ = **0.91**
Male	Juice	5	11.25	$\frac{(5-11.25)^2}{11.25}$ = **3.47**
Female	Coffee	15	24.75	$\frac{(15-24.75)^2}{24.75}$ = **3.84**
Female	Tea	20	16.5	$\frac{(20-16.5)^2}{16.5}$ = **0.74**
Female	Juice	20	13.75	$\frac{(20-13.75)^2}{13.75}$ = **2.84**

CHI SQUARE TEST

so after adding all of them we got the value of 16.49

now calculate the degree of freedom

$$df=(r-1)(c-1)=(2-1)(3-1)=1\times 2=2$$

Using this we are getting the critical value as 5.991 and $16.46 > 5.991$ we reject the null hypothesis.