Introduction to Statistics in Data Science

we have two statistics < Descriptive stats
Inferential stats

Statistics is a field that deals with collection, organization, Analysis, interpretation & presentation of the Data

⇓ Decision Making

Descriptive stats

⇓

- measure of Central Tendency
- measure of Dispersion

Summarizing the data

Histograms, PDF, CDF....

probability, Permutations, mean, median, mode, variance, Standard deviation.

Different types of distributions

Inferential stats

Z test
t test
Anova test
CHI Square test
Hypothesis test
Confidence intervals...

eg: I have an Age feature, it is for online shopping

Age = {24, 27, 14, 13, 28, 29, 31, 32}.

I want to make a decision, i.e promotional offers.

For which age we should give promotional offers.

By using this data, we calculate some imp things, by the help of statistics (mean, median, distribution of age)

What is "STATISTICS"

Statistics is the science of collecting, Organizing and Analyzing data. { we are doing this for better Decision making }

Definition of DATA

Data => facts (or) pieces of information that can be measured.

eg: The IQ of a class

{98, 97, 60, 55, 75, 65}

Ages of students of a class

{30, 25, 24, 23, 21, 28}

Types of Statistics → ~~differential~~ descriptive, Inferential

① DESCRIPTIVE STATS

It consists of organizing and Summarizing of data

a) measure of central tendency (mean, median, mode) b) measure of dispersion (var, σ)

② INFERENTIAL STATS

(using some experiments) → z test, t test ...

It is a technique where, we used the data that we have measured to form Conclusions (or) Inferences

(we use) Sample data → perform experiments → Conclusions → then conclusion for other data {population data}

eg: let us say I have a classroom of Maths students (26)

want to find marks of the 1st sem

marks ⇒ 84, 66, 78, 72, 75, 65, 80, 81, 92, 95, 96, 97 ....

Q ⇒ what is the average ~~age~~ Marks of the students in the class (eg of descriptive stats)

Q ⇒ Are the ~~ages~~ marks of the students of this classroom similar to the ~~age~~ marks of the ~~college~~ melting classroom (eg of inferential stats)

Population and Sample

(N) (pop.) island → 100k people

Sample (10k) (n)

Collect all the weights of all people

eg Elections

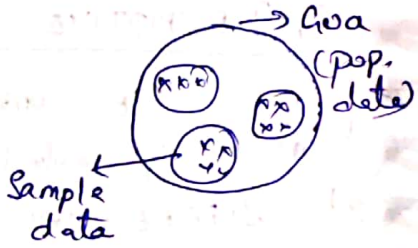
Which state are probably having the election recently

let (Goa, UP)

let election has finished and we really need to find out the exit poll

exit poll ⇒ they cannot go and ask each & every person that whom they have voted. The reports what they do is, they take up sample of population from different different regions and they ask that whom do you vote, based on that maximum number of people whom they vote, they actually create their exit poll.

In this example, my population data is the entire population of Goa and we take some pepple in some regions, those are sample data.



population is denoted by 'N'
Sample is " " 'n'

Note Q ⇒ Why you have selected samples randomly (or) is there any better ways to do sampling also coz just we need to do the sample randomly.

SAMPLING TECHNIQUES

① Simple Random Sampling ⇒

In this, we pick randomly some samples in the population there is no any format, just we take randomly.

in exit poll, we can use random sampling,

Suppose, If we want to test for the medicines, then we cannot use simple random sampling.

Dg ⇒ When performing SRS, every member of the population has an Equal chance of being selected for your sample(n)

② Stratified Sampling ⇒

dg ⇒ It is a technique where the population (N) is split into non-overlapping groups.

eg:- Gender → male, female

Let us say I want to do survey, for this I require some people (M & F), based on that gender, my samples will be devided

eg. I want to do survey on 0 to 10 years of kids

Based on Age, I will take samples

(0-10) (10-20) (20-40) ...

Q ⇒ Can we do Stratified Sampling based on profession ?

No. Dotnet, pyltton, C, ...

we can apply this stratified sampling on doctors, engineers .....

③ **Systematic Sampling** ⇒

In this, from population (N), we pick up every nth individual and will take as a sample.

Eg:- I am outside the mall & I want to do a survey regarding Covid

I will do, every 7th or 8th person I see, for this person do the survey. This is called as Systematic Sampling.

There is no reason for selecting 8th or 9th person, you just said that take every 7th person & do survey.

④ **Convenience Sampling** ⇒

Lets consider that I am doing a survey only those people who are a domain expert, is that particular survey, will be participating in this survey.

Let us say, I am doing the survey related to data science, I will say that any person who is probably interested in data science, if you consider only those people then it is basically becomes a Convienent Sampling.

→ RBI do survey with respect to household, for this household Surveys, what kind of Sampling they may use?

In household Surveys, RBI make sure that they have to fill the Survey from women, where probably they are trying to find out what is the cost expenditure running a house etc..

Here we can Do Stratified Sampling / Convienence Sampling

Note Sampling techniques are completely dependent on the use case that we are following.

Eg. ① A Drug need to be tested, what kind of Sampling we use.?!

Sol here I can bring up multiple use cases, 1st of all to whom this drug need to be tested, If I get that specific information

I will basically do the age groupings and then I apply .... Let us consider, the drug is for everyone, then I may consider up some samples, but I will put atleast put a condition that, it should be >15 years. So it is based on use case you do, based on use case you will select the sampling technique.

Variables

A Variable is a property that can take on any value.

eg hight, weight

{188,178,168} {78,99,100,60}

Here we have 2 kinds of Variables

1. Quantitative Variables
2. Qualitative / Categorical Variables

1. Quantitative Variables

It can measured numerically

{ Add, subtract, multiply, divide }

eg Age, hight, weight

Quantitative

- Discrete Variables: whole number
  - eg: No of Bank Accounts
  - Total children in a family (no. of)
- Continous Variables: any values
  - eg: hight (172.5, 162, 163.5 cm)
  - weight (100kg, 99.5, 99.25)
  - amt of rainfall (1.1 inches, 1.25 inches etc ...)

2. Qualitative / Categorical Variables

Based on some characteristics we can derive some categorical variable. In this we have some categories.

eg Height < M, F

IQ < 0-10 => less IQ; 10-85 => medium IQ; 50-100 => Good IQ

Blood group, T-shirt size ....

eg:

1) what kind of Variable Gender is ? Categorical Variable

2) what kind of " Marital status ? Categorical Variable

3) what kind of " River length ? Continous Quantitative Variable

4) what kind of " Population of State is ? Discrete " "

5) what " " Song length ? Continous " "

6) " " Blood pressure ? Continous " "

pin code ? Discrete

Variable Measurment scales

1. Nominal
2. Ordinal
3. Interval
4. Ratio

1. Nominal

These are specifically categorical data

eg. Gender, colour, type of flower

2. Ordinal data

order of the data matters but value does not

eg:

student marks	Rank	
100	1	ordinal data
96	2	
57	4	there we focus more on the order
85	3	
44	5	

3. Interval data

order matters, value also matters.,

natural zero is not present

eg Temperature

10-80 F 80-90F 90-100F

0F ⇒ not meaningful

distance

10-20 20-30 30-40

4. Ratio Data

Frequency Distribution

Suppose I have a sample dataset, which consists of 3 types of flowers.

Rose, lily, Sunflower, Rose, Lilly, Sunflower, Rose, Lilly, Lilly.

Based on the flower type, how much is the frequency.

Flower	Frequency
Rose	3
Lilly	4
Sunflower	2

freq distribution

cummulative freq

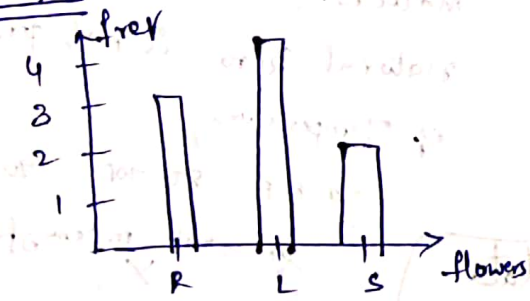
3

3+4 = 7

7+2 = 9 ← Total no of flowers

From this freq distribution, we can derive bar charts, pie charts etc

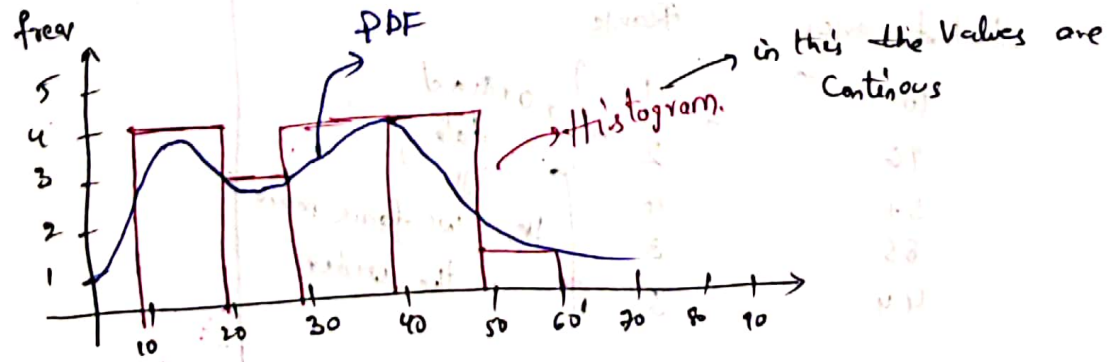
Bar graph (Data should be discrete)



Histogram (Data should be continuous)

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

In histogram, we make some kind of bins, by default the bin size = 10



bin 10-20 ⇒ we have 4 values

PDF ⇒ Smoothing of histogram (it can be done using Kernel density estimator)

probability distribution function

BAR Vs Histogram

used for discrete → ← used for continous

Day 2 **Basics to intermediate of statistics**

Topics discussed ⇒
- measure of central tendancy
- measure of dispersion
- Gaussian Distribution
- Z score
- Standard normal distribution

" It refers to the measure used to determine the centre of the distribution of data "

Central Tendency
- mean
- median
- mode

Arthemetic mean for population and sample.

Mean (Average)

Population (N)

Random variable

let $X = \{1,1,2,2,3,3,4,5,5,6\}$ → pop' dataset → Sample (n) → Same data

$$\mu = \frac{\sum_{i=1}^{N} x(i)}{N}$$

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

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

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

then we get same value

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

Central Tendancy

Median

will take same data set

$\{1,1,2,2,3,3,4,5,5,6,100\}$ ← outlier

$$\text{mean} = \frac{32+100}{11}$$

$$= \frac{132}{11} = 12$$

it is completely outside the values in dataset.

Outlier has major impact

when '100' is not added then mean = 3.2, if 100 is added in the same distribution then it became '12' so there is huge difference in mean because of this 100., so this no. is called as

OUTLIER

So in this case we use median.

median =>

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

Step1 Sort the no's

in this eg, already it is sorted

Step2 Take the central element

If total count = odd no.

then we take the central element.

If total count = even no.

⇓

we take middle 2 nos

& we take avg of the 2 nos.

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

median = 3 (as it is odd no)

if 2 outliers are added

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

Avg = $\frac{3+4}{2}$ = 7/2 = 3.5

median = 3.5 (it is even no)

Note:- when we use Median for one Outlier, i.e. 0, then the median = 3

when we have 2 outliers, then Median = 3.5

So there is less difference

** Median works well with the Outlier

Mode

Suppose my Dataset - {1,2,2,3,4,5,6,6,6,7,8,100,200}

To calculate mode, we find out the most frequent element in 'D'

& for our eg =>

Mode = 6

Disad

Suppose I have many outliers

Dataset

	Type of flower	PL	PW
	Rose		
	Lily		
	Sunflower		
	⋮		

In this Dataset, Suppose we have missing data (10%)

Here we use mode, because the missing value is replaced by most frequent occuring element

Mode is specially used for Categorical Value

eg: Age of students

25
26
27
-
-
-
32
34
36

for this what should I apply? (mean/mode/median)

So I prefer mean in this example because, I know students age will basically range from one value to another value.

So here Domain Knowledge will also used.

Measure of Dispersion

Variance Standard Deviation

Dispersion ⇒ Spread (how well spread the data is)

let
$D_1 = \{1,1,2,2,4\} \Rightarrow \mu_1 = 2$

$D_2 = \{2,2,2,2,2\} \Rightarrow \mu_2 = 2$

for both distributions, we get same mean, so how these 2 distributions are different?

In order to find how two distributions are different, at that pt, we use Variance

$D_1 = \{2,2,4,4\}$

$\mu_1 = 3$

the 2 & 4 are closer to mean ⇒ Disp is less

$D_2 = \{1,1,5,5\}$

$\mu_2 = 3$

here 5 & 1 are far away from mean ⇒ Dispersion is more

** Variance ⇒ it is used to differentiate the data

Find the diff b/w from mean

Pop. Variance

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

N → pop. size

Sample Variance

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

why (n-1)? → Sample size.

eg:

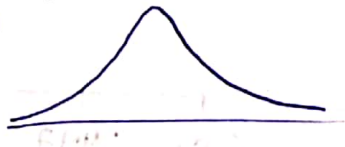
x	μ	x-μ	$(x-\mu)^2$
1	2.83	-1.83	3.34
2	2.83	-0.83	0.6889
2	2.83	0.83	0.6889
3	2.83	+0.83	0.03
4		1.17	1.37
5	2.83	2.17	4.71
			10.84

$\mu = 2.83$

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

Let say if I have a dataset, which looks like

$D_1 \Rightarrow$



$D_2 \Rightarrow$

comparing these 2 data, where the variance is more?

In 2nd Distribution, variance is more because the spread is more.

Variance ↑ when spread ↑↑ ⇓ elements present in the central region is more

Standard deviation ⇒ It tells how far away from mean.

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

$\sigma = \sqrt{\text{Variance}}$

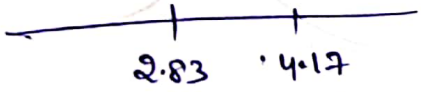
$= \sqrt{1.81}$

$\boxed{\sigma = 1.345}$

→ In this eg, mean = 2.83, from this mean, the data is distributed because mean is basically specifying your measure of central tendency it basically says that, where the center is there, for that particular distribution.

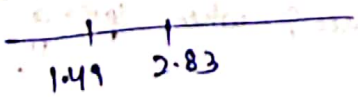
→ form the mean, if we go one standard deviation to the right, the next element that may probably found / fall b/w 2.83 + 1.345 = 4.17

one S.D

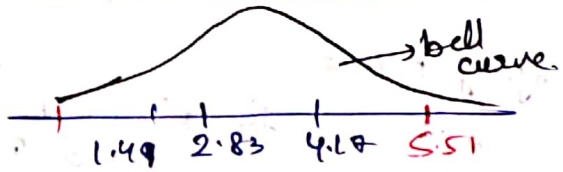


whatever elements are basically present b/w 2.83 & 4.17 will be falling with in the 1st standard deviation

If we consider something towards the left, we will substract 2.83 - 1.34 = 1.49



Any element falls b/w 1.493 to 2.83 will be falling in this region ie one S.D to the left



S.D is a very small number graph looks like above.

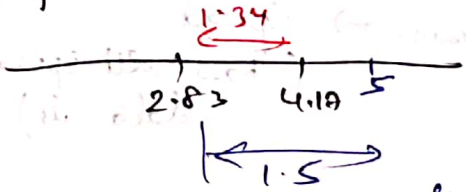
Based on SD and Variance we will be able to decide a imp things

Variance ⇒ spread how data is

S.D ⇒ B/w one S.D to the right and the left what may be the range of data that may be falling.

Standard deviation is square root of Variance ie from the mean how far a element can be

eg: if I consider 5



if ya'll try to calculate, it's may fall somewhere here

So 5 is represented as 1.5 σ from the mean.

Percentiles & Quantiles

This is the 1st step to find Outliers

Percentage ⇒ 1, 2, 3, 4, 5

% of the no's that are odd?

$\% = \frac{\text{\# of numbers that are odd}}{\text{total no's}}$

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

Percentile ⇒

(GATE, CAT, GMAT, SAT)

Definition:- It is a value below which a certain % of observation lie?

Eg Dataset

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

Q? I want to find out, what is the percentile ranking of 10?

Sol percentile Ranking of 10

let x = 10

$\text{percentile ranking of } x = \frac{\text{\# of Values below } x}{n} \times 100$

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

$= 80 \text{ percentile}$

ie 80 percentage of the entire distribution is less than 10

Real meaning

Q) what is the percentile ranking of 11?

percentile ranking of 11 $= \frac{17}{20} \times 100$

$= 85\%$

b) what value exists at percentile ranking of 25%

Formulae ⇒

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

$\text{Value} = \frac{25}{100}(20+1)$

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

Value = 5.25

it is the index position, not the value.

in my dataset

2, 2, 3, 4, 5, 5, 5, 6, 7, 8, 8, 8, 8, 9, 9,
1st 2nd 3rd 4th 5th 6th
10, 11, 11, 12

5.25 index is not there so we take average of 5th & 6th index's

$\frac{5+5}{2} = \frac{10}{2} = 5$ //

Q) what value exists at percentile ranking of 75%.

$\text{Value} = \frac{75}{100}(21)$

$= 15.75 \text{ (ordinal)}$

15th → 8
16th → 9
$\frac{8+9}{2} = \frac{17}{2}$

=

Five Number Summary

1. minimum value
2. Ist Quartile (Q_1)
3. median
4. Third Quartile (Q_3)
5. Maximum value

Removing the Outlier

$D = \{1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 27\}$

what is the outlier in this dataset

27

In order to remove outlier, we need to find lower fence & higher fence

nos < lowerfence & nos > higher fence → treated as an outlier

Lower fence = $Q_1 - 1.5(IQR)$

Upperfence = $Q_3 + 1.5(IQR)$

IQR = Inter Quartile Range

$= Q_3 - Q_1$

$Q_3 = 75\%$

$Q_1 = 25\%$

$Q_1 \to 25\%$

Value = $\frac{25}{100}(19+1)$

$= \frac{25}{100}(20)$

Value = 5 (index) $\Rightarrow Q_1 = 3$

$Q_3 \to 75\%$

value = $\frac{75}{100} \times (20)$

= 15 (index)

$Q_3 = 7$

$IQR = Q_3 - Q_1$

$= 7 - 3$

$IQR = 4$

Lower fence = $Q_1 - 1.5(IQR)$

$= 3 - 1.5(4)$

$= -3$

Upper fence = $Q_3 + 1.5(IQR)$

$= 7 + 1.5(4)$

$= 13$

[Lower fence $\longleftrightarrow$ Higher fence]

[$-3 \longleftrightarrow 13$]

So anything < -3 and anything > 13 is considered as the dataset as outlier

So '27' becomes as outlier and we should remove it.

Remaining data is

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

Minimum value = 1

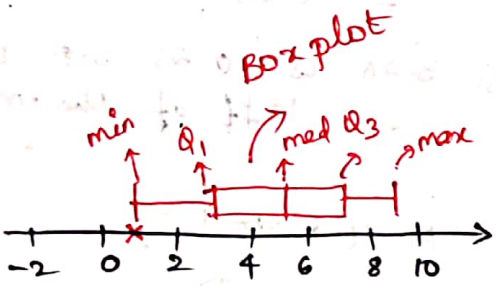
1st Quartile (Q_1) = 3

median = 5

3rd Quartile (Q_3) = 7

Max value = 9

By this specific data we can draw a **BOXPLOT**



This technique of removing an outlier we basically say with respect to lower fence & higher fence & we use IQR.

Day 3

Distributions

let we have a dataset

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

Here we have a dataset, So how do we basically see this dataset in a Visualized way.

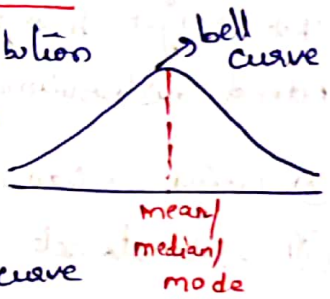
If I want to start analysis I really need to see lot of Visualized diagrams & that is where, when I Consider this entire distribution there are multiple ways to Visualize this data through various graphs.

As we have ages dataset, I want to plot this data and the best & easy way that can probably think about is your HISTOGRAM

So we will go to distributions

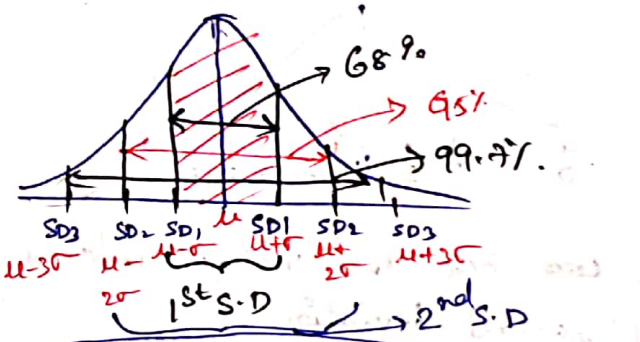
① Gaussian Distribution / Normal distribution

If I have a distribution & if this distributions is like a bell curve, main thing is this bell curve is symmetrical on b.s



it is

The right part data present = left part data.



Emperical formulae (68-95-99.7% rule)

(68) → B/w the 1st standard deviation region, around 68% of the distribution is present

Let D = 100 pts

So out of 100 pts, 68 pts present in the 1st S.D region (so it is in form of bell curve)

(95) → with in 2nd S.D region, around 95% of the entire data lie in this region

99.7% → within 3rd S.D region, around 99.75% of the entire distribution will fall in this region.

Eg: ① Height

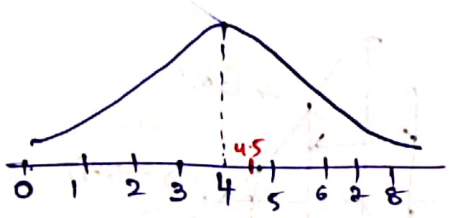
It is normally distributed

So Domain expert (data) says that height belongs to gaussian/ normal distribution

② weight → Gaussian dist

③ Iris data set

Eg: Suppose I have $\mu = 4$, $\sigma = 1$ then can I find distribution



Q→ where does 4.5 fall in terms of standard deviation.

Sol: it is 0.5 SD right

Q)→ 4.25 ?

the S.D = 1, in case of 4.25 it is very much difficult to do do the calculation

So we can use Zscore, it used to find out how much SD away from the mean

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

$$Z\text{score} = \frac{4.5 - 4}{1}$$

$$= 0.75$$

∴ 4.25 is 0.75 S.D to the right

As it is +ve value ⇒ So it is right to mean

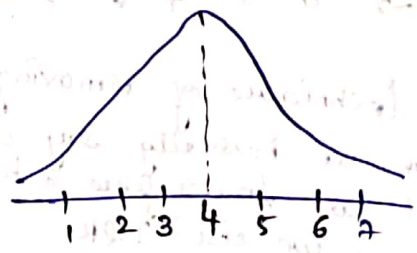
Q) 3.75 ?

$$Z\text{score} = \frac{3.75 - 4}{1}$$

$$= -0.25$$

So 3.75 is .0.25 S.D to the left of the mean

→ $\mu = 4$ & SD = 1



lets apply Zscore to every values

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

initally my distribution is

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

after applying Zscore my distribution is

{−3, −2, −1, 0, 1, 2, 3}

$Z(1) = \frac{1-4}{1} = -3$ $Z(5) = 1$

$Z(2) = \frac{2-4}{1} = -2$ $Z(6) = 2$

$Z(3) = \frac{3-4}{1} = -1$ $Z(7) = 3$

$Z(4) = 0$.

{1, 2, 3, 4, 5, 6, 7} → normal/gaussian distribution

↓

ZScore

↓

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

↓

This distribution is called as

STANDARD NORMAL DISTRIBUTION

↓

$\mu = 0$ & $\sigma = 1$

$\therefore Y \in SND(\mu = 0, \sigma = 1)$

Standard normal distribution

Q2:- Why do we do this?

we do this in ML algorithms.

Related eg

Suppose Dataset is

Age (years)	Salary (Rs)	weight (Kg)
24	40K	70
25	80K	80
26	60K	55
27	70K	45

One main target is that we should bring up in a form where $\mu = 0$ & $\sigma = 1$ at that pt of time I can apply Standard normal distribution i.e I can take up this entire data and apply ZScore & convert this into SND.

This process is called as

STANDARDIZATION

↓

$\mu = 0$ & $\sigma = 1$

whenever we talk about Standardization, internally there is a ZScore formulae applying on it.

NORMALIZATION

→ Let us say I want to change my all values in Dataset b/w 0 to 1 then we use Normalization

This normalization can be done by MinMax Scaler. In this you have to provide 0 to 1 & automatically this kind of normalization happens.

Normalization gives you a process where you can basically define the Lower bound and upper bound and you can convert your data b/w them.

where do we use normalization?

In CNN, whenever you are doing image Training / Object detection, every image has a pixels, each and every pixel ranges b/w 0-255

Before we start training this can be applied with min-max scaler and it gets converted b/w 0 to 1 where the min value '0' is assigned to 0 & the max value 255 is assigned to '1'.


4x4 image

So here we are doing Normalization.

another type of normalization process $\Rightarrow$ $\frac{0 \text{ to } 255}{255}$

↓

0 to 1

Practical eg :- (1)

Cricket match
{ India vs South Africa }

ODI series (2021)

The series average 2021 = 250

The S.D of the score = 10
(Standard deviation)

Rishabh score = 240
(team) avg

(2020)

Series average 2020 = 260

S.D = 12

Rishabh score = 245
(Team) avg

Question =) Compare to both the series, in which year Rishabh Pant final score was better ?

Ans

2021

$$\text{Zscore} = \frac{x_i - \mu}{SD}$$

$$= \frac{240 - 250}{10} = -1$$

2020

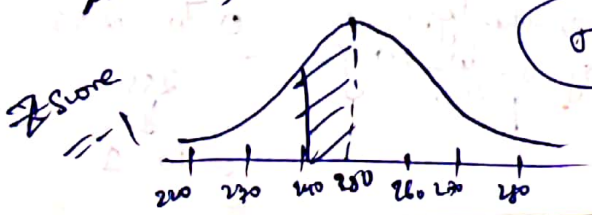
$$\text{Zscore} = \frac{x_i - \mu}{\sigma}$$

$$= \frac{245 - 260}{12} = \frac{-15}{12}$$

$$= -1.25$$

In 2021,

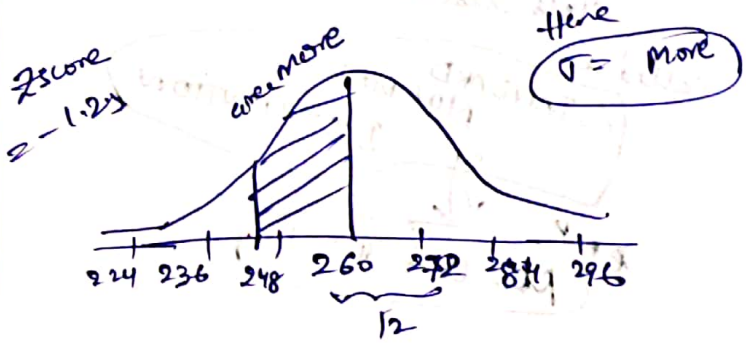
$\mu = 250$, $x_i = 240$, $\sigma = 10$



240 is falling into -1 standard deviation

In 2020

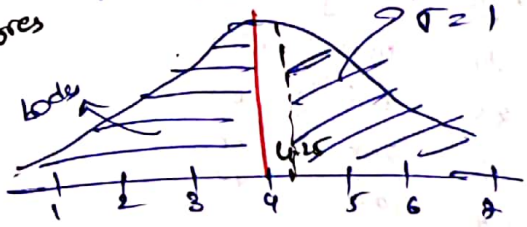
$\mu = 260$, $x_i = 245$, $\sigma = 12$



∴ India performed well in 2021 because Zscore is more for 2021

Practical eg (2)

X = it has the below kind (r.v) of distribution



Q) what % of scores that fall above 4.25 ?

meaning is, what is the % of scores that are greater than 4.25

Ans Here we use Zscore

$$\text{Zscore} = \frac{x_i - \mu}{\sigma}$$

$$= \frac{4.25 - 4}{1}$$

Zscore = 0.25

∴ 4.25 falls 0.25 Standard deviation from the mean.

Zscore = 0.25

This is a symmetrical bell curve. i'e the entire area can consider it as one.

Since I am interested in above 4.25 region, I say as tail region, this region has tail the remaining region is said as Body.

Based on Zscore, I need to find the area of the body curve.

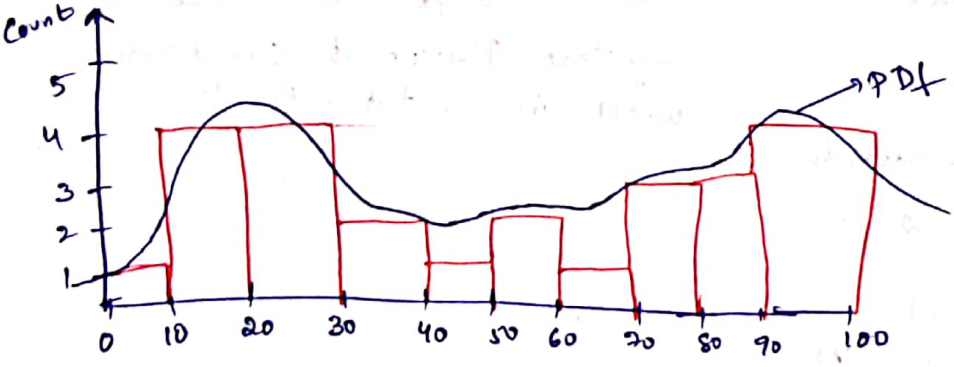
Histograms

Dataset =

10, 12, 13, 14, 20, 22, 24, 25, 26, 35, 38, 42, 55, 56, 68, 82, 84, 86, 92, 93, 94, 100

These nos are Quantitative no's.

bin = 10



If we Smoothen this histogram, it gets converted into PDF

Smoothen version of Histogram = PDF

Eg: This histogram will also useful for Qualitative data

Standard Normal distribution

(continue of Gaussian dist)

$Y \approx SND(\mu=0, \sigma=1)$

$X \approx N(\mu, \sigma)$

$\Downarrow$

$Y \approx SND(\mu=0, \sigma=1)$

To convert Normal distribution to SND, we use Z-score.

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

Eg: $D = 1, 2, 3, 4, 5$

$\mu = 3$

$\sigma = 1.1$ (let $\sigma = 1$)

$$Z = \frac{x_i - \mu}{\sigma} = \frac{1-3}{1} = -2$$

$$Z_2 = \frac{2-3}{1} = -1$$

$$Z_3 = \frac{3-3}{1} = 0$$

$$Z_4 = \frac{4-3}{1} = 1$$

$$Z_5 = -1$$

$D = \{-2, -1, 0, 1, 2\}$

$\mu = 0$

$\sigma = 1$

Q) Why we are doing this? (converting any dist. into SND)

eg Suppose Dataset

Age (years)	Wt (kg)	distance (km)
24	100	100
25	120	200
26	75	500
27	80	600

1) In this dataset, all the units to measure the features are different.

In our ML model, we should try to bring all the features in the same scale.

In this kind of dataset, inorder to bring into same scale, we apply z score to each & every feature & convert into SND.

Application of Z score

Z score ⇒ these are standardized values that can be used to compare scores in different distributions.

eg:- Cricket (India vs Australia)

2020 2021

2020 (3 tests)
Test avg = 181
SD = 12
Rishabh pant = 187

2021
Test avg = 182
SD = 5
Rishabh pant = 185 in 1st test

Compared to the rest of the entire test series in which year was Rishabh pants score in 1st game better?

Sol $Z = \frac{x-\mu}{\sigma}$

2020

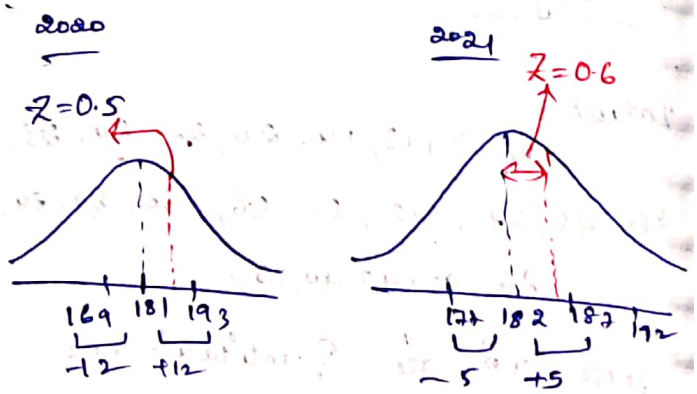
$x = 187$, $\mu = 181$

$Z = \frac{187-181}{12}$

$Z = 0.5$

2021

$Z = \frac{185-182}{5} = \frac{3}{5} = 0.6$



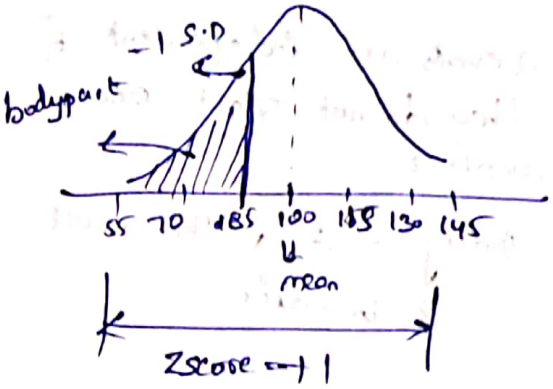
Since $Z = 0.6$ for 2021, so we can say that Rishabh pant scored very well in 2021 test series. (0.6 > 0.5)

eg of Zscore

a) In india the avg IQ = 100, with a standard deviation of 15. what % of the popuatn would you expect to have an IQ lower than 85?

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

$= \frac{85 - 100}{15} = \frac{-15}{15} = -1$



from Z table, the area of the body curve is 0.84134 ⇒ for +1

$1 - 0.84134 = 1.5866.$

for -1 ⇒ 1.58661%

Basics of probability

Probability ⇒ it is a measure of the likelihood of an Event/experiment

eg: flipping of a coin.
what is the prob. of getting head

Sol Sample space = {H, T}

$$\text{probability} = \frac{\text{No. of ways an event can occur}}{\text{no of possible outcomes}}$$

$= \frac{1}{2} = 0.5$ //.

Addition Rule

mutual exclusive events ⇒ 2 events are mutual exclusive if they cannot occur at the same time.

eg: when flipping a coin, the 2 events Head & tail are mutual exclusive.

non mutual exclusive events ⇒ 2 events are non mutual exclusive if they can occur at the same time.

eg: Deck of Cards ⇒ when picking a card randomly, the 2 events Queen & Hearts.

Multiplication Rule

Independent Event

2 events are independent if they do not effect one another

Eg: Rolling a '6' & then roll '3' in a dice

Dependent Event

2 events are dependent, if they effect one another

Eg: Consider we have Bag of marbles which has 4 red & 2 black marbles.

$P_r(\text{Red}) = \frac{4}{6} = \frac{2}{3}$

$P_r(B) = \frac{2}{5}$

multiplcatn rule for indep. events

Q: what is the prob. of rolling 6 & then 3, with a normal 6 sided die?

Sol $P(A \cap B) = P(A) * P(B)$

$P(6 \cap 3) = P(6) * P(3)$

$= 1/6 * 1/6 = 2/6 = 1/3$

Q) multiplcatn rule for dependent events.

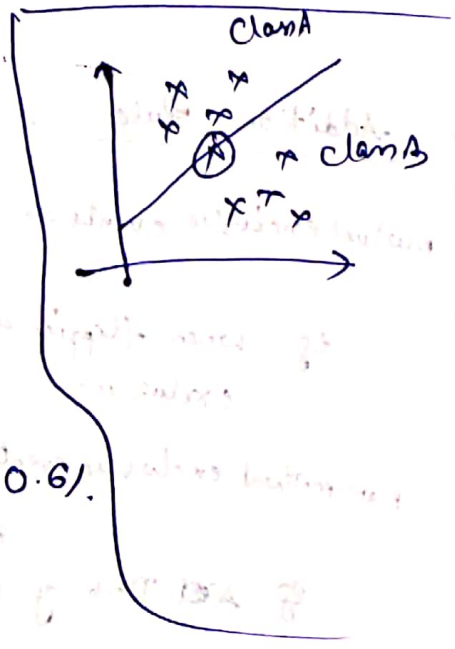
what is the prob. of drawing a Queen & then drawing a king from a deck of cards?

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

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

$P(Q) = \frac{4}{52}$ $P(K/Q) = \frac{4}{51}$

$P(Q \text{ and } K) = \frac{4}{52} \times \frac{4}{51} = 0.06 = 0.6/.$

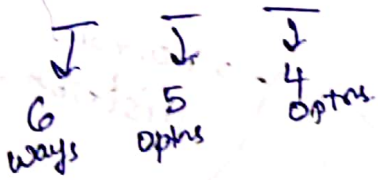


Permutations and Combinations

Permutations =>

chocolate factory (6 chocolates)

eg: In a school trip, they are 50 students, we go to a chocolate factory, I have given a task that, note down the 1st three choclats that you probably see?



=> 6 × 5 × 4 = 120 => ie, they have 120 options (or) permutations to write the chocolate name.

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

n = total no. of objects
r = no of objects you are picking

$$ {}^{6}P_3 = \frac{6!}{(6-3)!} = \frac{6!}{3!} = 120 $$

Combinations =>

By write down the 1st three chocolets that you see, but they should not repeat

$$ {}^{n}C_r = \frac{n!}{(n-r)!\,r!} = \frac{6!}{3!\,3!} $$

$$ {}^{6}C_3 = 20 \text{ unique combinations} $$

∴ combinations < permutation

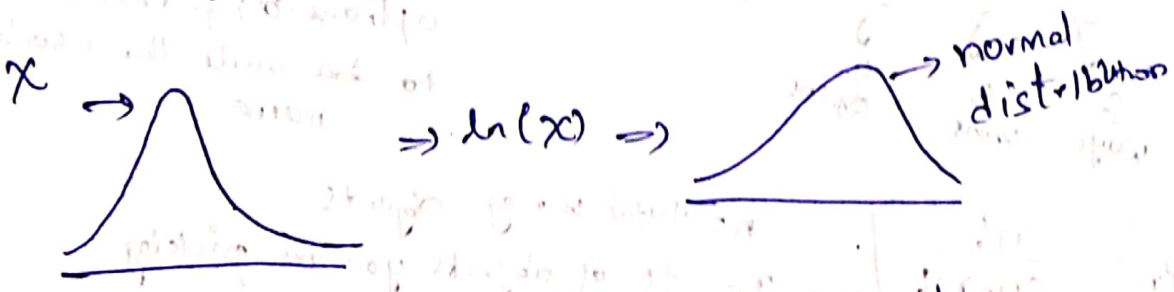
Log Normal Distribution

WKT $X \approx N(\mu, \sigma)$

$X \approx$ log normal distribution

⇓ then

$y = \ln(x)$ ⇒ it has a normal distribution



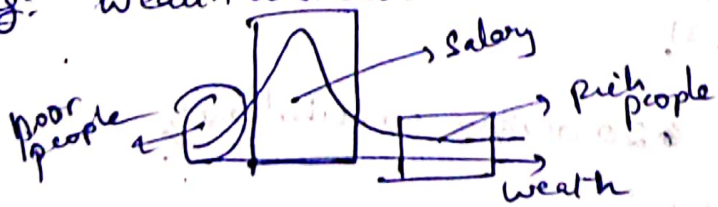
then $X \approx$ Log Normal distribution

if the R.V X is a log-normally distributed then $y = \ln(x)$ is a normal distribution.

$y = \ln(x)$

$x = \exp(y)$

eg: ① wealth distribution

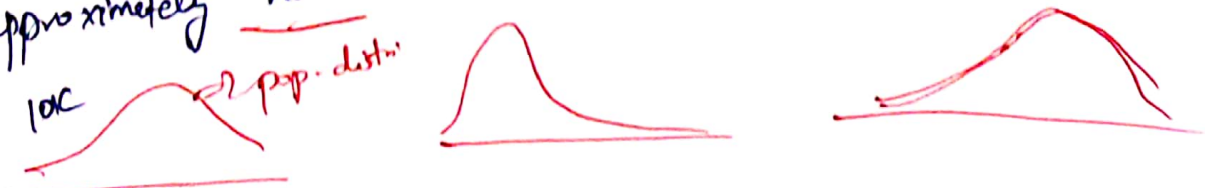


② length of the comment in any chanel.

③ Salaries of people in companies.

Central Limit Theorem

The CLT states that regardless of the shape of population distribution, the distribution of sample means will be approximately normal.



$n > 30$

⇓

Samples (50-60)

$\bar{x}_1, \bar{x}_2, \bar{x}_3, \bar{x}_4, \bar{x}_5 \ldots \bar{x}_{200}$ (normally distributed)

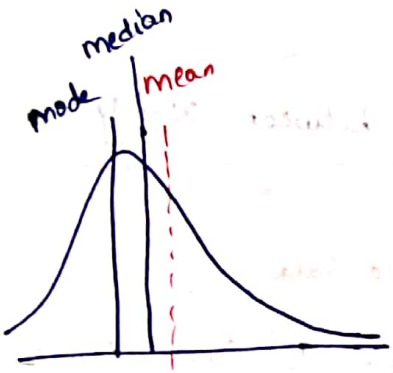
Here $n \geq 30$

2 imp points in CLT,

① The distribution of sample means will become more normal as its sample size increases.

② Good rule thumb:- Sample distribution will be appr. normal if their sample size is $n \geq 30$

Left skewed & right skewed distribution

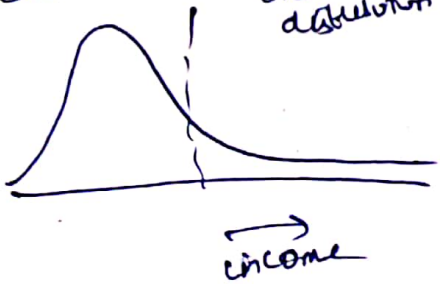


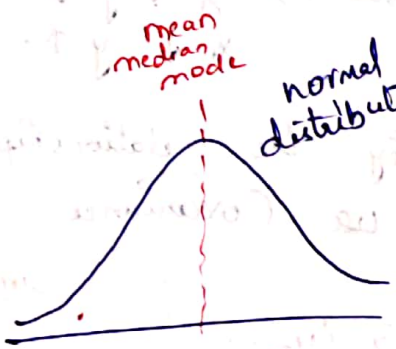
positive (Right skew skew)

eg: wealth distribution
length of comments

Mean > median > mode

In this most of the score occur at the lower end of distribution & few scores at higher end

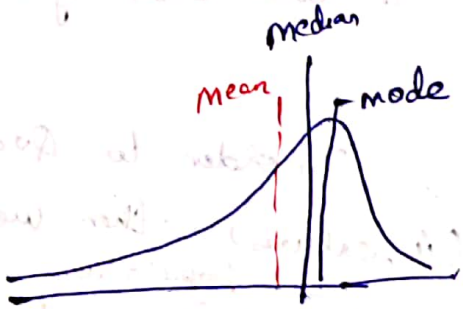




Symmetrical Distribution

eg: Age
weight
height
IRIS data set

mean = median = mode

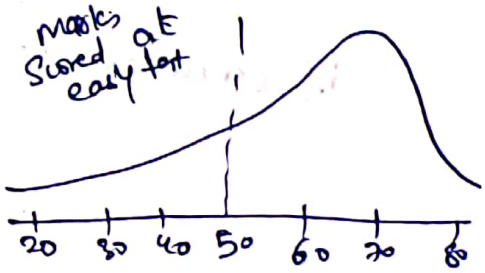


Negative skew

eg: Life span of human being

mode > median > mean

In this most of the score occur at the higher end of distribution & few score at lower end.



Co-Variance → It helps to quantify the relationship b/w features in a dataset R.V.

Covariance is used to find the direction of the Relationship

eg: Suppose I have a dataset

(X) Age (years)	weight (kg) (Y)
20	75
18	63
15	45
14	40
25	78

As the $X \uparrow \Rightarrow Y \uparrow\uparrow$ from dataset

$X \uparrow Y \uparrow$

$X \downarrow Y \downarrow$

In some cases of dataset $\Rightarrow$ $X \uparrow Y \downarrow$

$X \downarrow Y \uparrow$.

In order to Quantify the relationship between X & Y (ie features) then we use Covariance

Population

$$Cov(X,Y) = \frac{\sum_{i=1}^{N} (x-\bar{x})(y-\bar{y})}{N}$$

sample data

$$Cov(X,Y) = \frac{\sum_{i=1}^{n} (x-\bar{x})(y-\bar{y})}{n-1}$$

↓ Because of Bessels correction

eg:-

Economic Growth (%) (X)	Nifty 50 index (Y)
2.1	8
2.5	12
4.0	14
3.6	10

(yrs) Age	wt (kg)
20	75
18	63
15	45
14	40
25	78

Covariance calculation

if covariance is +ve $\begin{cases} x \uparrow y \uparrow \\ x \downarrow y \downarrow \end{cases}$

-ve $\begin{cases} x \uparrow y \downarrow \\ x \downarrow y \uparrow \end{cases}$

x	y	$\bar{x}$	$\bar{y}$	$x-\bar{x}$	$y-\bar{y}$
2.1	8	3.1	11	−1	−3
2.5	12	3.1	11	−0.6	1
4.0	14	3.1	11	0.9	3
3.6	10	3.1	11	0.5	−1

$x\uparrow \Rightarrow y\uparrow \Rightarrow$ cov = +ve

$x\uparrow\ y\downarrow \Rightarrow$ cov = −ve

$$Cov(x,y) = \frac{(-1)(-3) + (-0.6)(1) + (0.9)(3) + (0.5)(-1)}{4-1}$$

$$= \frac{3 - 0.6 + 2.7 - 0.5}{3} = \frac{4.6}{3}$$

$Cov(x,y) = 1.533$ → +ve value

you are satisfying 2 conditions

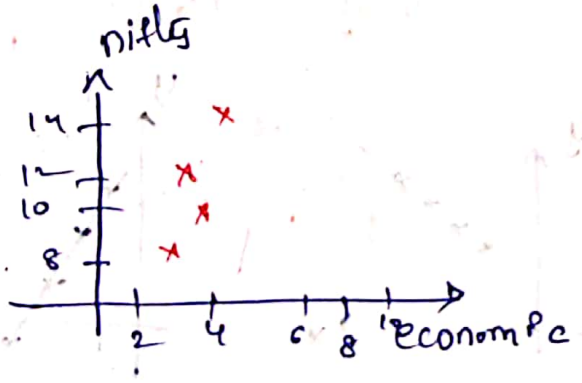
① When economic Growth ↑ ⇒ Nifty 50 index growth ↑

② " " " ↓ ⇒ " " " ↓

covariance can be positive / negative value, there is no restriction at all. but we should restrict, In order to prevent this we use pearson & Spearman Rank correlation.

Pearson correlation coefficient (it is used for feature selection)

Covariance is positive / negative but how strong it is +ve (or) how strong it is −ve, it is difficult to find (or) to restrict the value. In order to overcome we use pearson correlation coefficient.



$$\rho_{(x,y)} = \frac{cov(x,y)}{\sigma_x \sigma_y}$$

$$\sigma_x = \sum_{i=1}^{n} \sqrt{\frac{(x-\bar{x})^2}{n-1}} \qquad \sigma_y = \sum_{i=1}^{n} \sqrt{\frac{(y-\bar{y})^2}{n-1}}$$

$$= \frac{(-1)^2 + (-0.6)^2 + (1.1)^2 + (0.5)^2}{4-1} \qquad = \frac{(-3)^2 + (1)^2 + (3)^2 + (-1)^2}{3}$$

$$\sigma_y = 2.58$$

$$= 0.8981$$

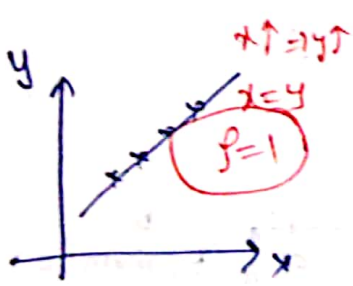
$$\rho_{(x,y)} = \frac{1.533}{(0.898)(2.58)} = 0.66 \Rightarrow +ve \text{ correlation.}$$

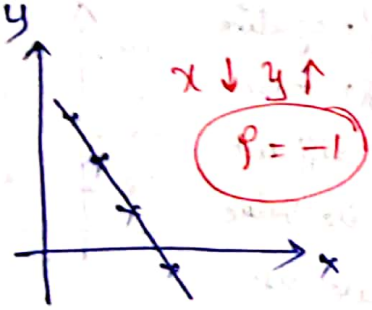
∴ 66.6% is +vely correlated.

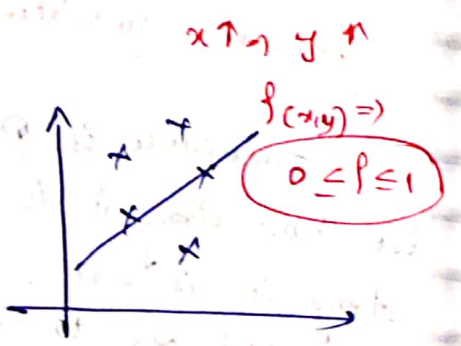
Based on variance of x & y, pearson correlation coefficient will tells you about both Strength and Direction of Relationship

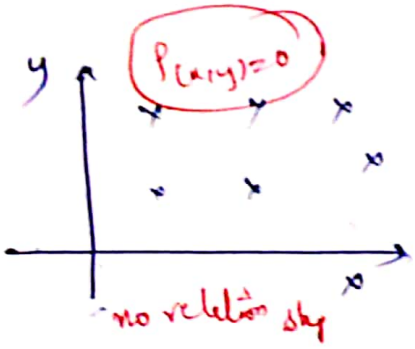
(who how much it is +ve/-ve)

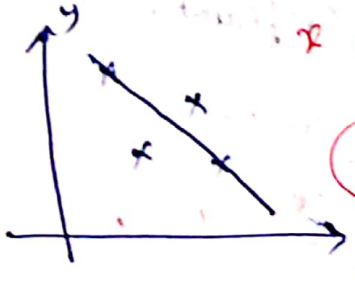
always $\rho_{(x,y)} \Rightarrow \boxed{-1 \le \rho \le 1}$









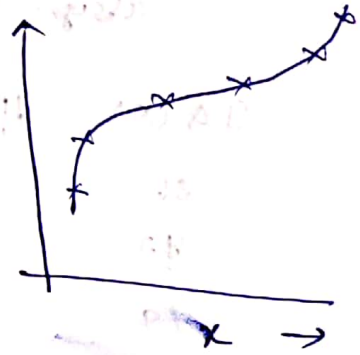


When we have 2 features x & x', suppose if $1 \uparrow \Rightarrow x' \uparrow$ then we say that $\rho = 1 \Rightarrow$ 2 features are same $\rightarrow$ so in this case we can drop one feature and apply ML models.

Spearman's rank correlation coefficient

$$r_s = \frac{\text{Cov}(rg_x, rg_y)}{\sigma_{rg_x} \cdot \sigma_{rg_y}}$$

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2-1)}$$



$\text{Cov}(rg_x, rg_y)$ = Covariance of rank variables

$\sigma_{rg_x}, \sigma_{rg_y}$ = Standard deviations of rank variables.

eg The tabel below shows to calculate the correlation between IQ of a person with the no. of hours spent infront of TV per week

IQ (X)	Hours of TV / week (y)
106	7
86	2
86	2
101	50
99	28
103	29
97	20
113	12
112	6
110	17

Sol) if we want to apply Spearman rank correlation, steps are

Step 1 ⇒ Sort the data by the 1st colo (x), Create a new Coloumn x_i & assign the ranked value 1, 2 ...

2 ⇒ Sort the data by 2nd col (y), Create a col. y_i & assign ranked values

IQ (x)	Hours of TV (y)	Rank x_i	Rank y_i	d_i	d_i^2
86	0	1	1	0	0
97	20	2	6	-4	16
99	28	3	8	-5	25
100	27	4	7	-3	9
101	50	5	10	-5	25
103	29	6	9	-3	9
106	7	7	3	4	16
110	17	8	5	3	9
112	6	9	2	7	49
113	12	10	4	6	36

Step 3 ⇒ Create a 5th col. of to find the difference b/w 2 ranks (x_i & y_i) & square the d_i value.

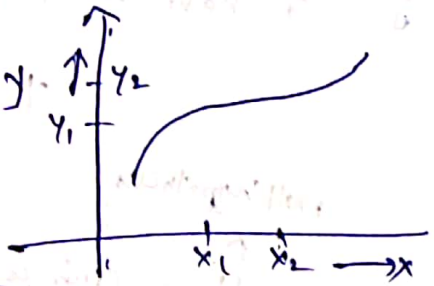
$$\rho = 1 - \frac{6\Sigma d_i^2}{n(n^2-1)}$$

$$= 1 - \frac{6 \times 194}{10(10^2-1)}$$

$\rho = -0.17575$ ⇒ very near to zero ⇒ hours spent on TV is very less

Correlation is very very less.

Eventhough the data is non-linear, we can find the correlation / we can clearly do the correlation b/w 2 features by using Spearmans rank correlation.



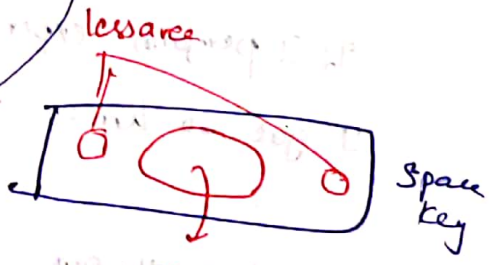
eg:

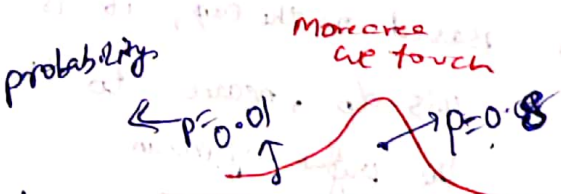
Economic growth(%)	ngty sonder(%) (y)
2.1	8
8.5	12
4.0	14
3.6	10

Cov(x, y)

x	y	$\bar{x}$
2.1	8	3.1
2.5	12	3.1
4.0	14	3.1
3.6	10	3.1

*** P Value





Where there is maximum no of touchings, there we have a bulg part, where there is less no of touchys, there it is tail part. Here we have 2 tailed test.

Let we touch at the dotted point, there let p=0.01

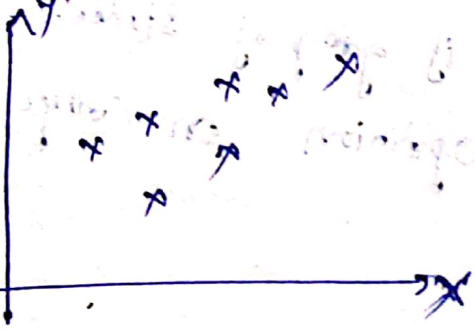
it indicates, if we repeat this experiment 100 times, the no of times we will be touching this area is 1 time

→ if p = 0.8 ⇒ it touches 80 times

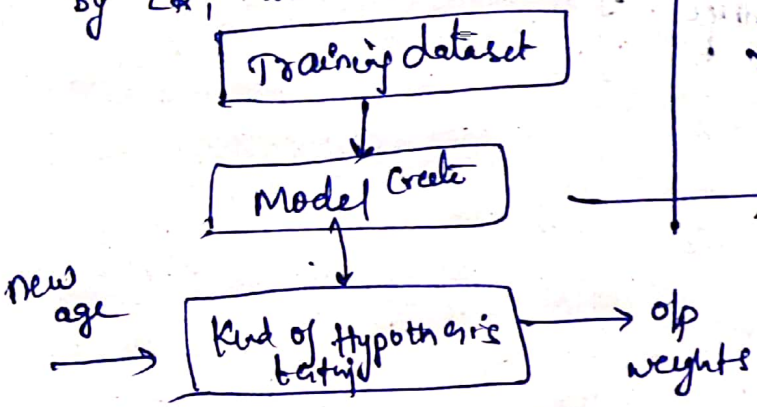
pvalue ⇒ it is the probability for the null hypothesis to be true.

Linear Regression

Suppose I have 2 features (Age, wt) based on these features we have data pts.

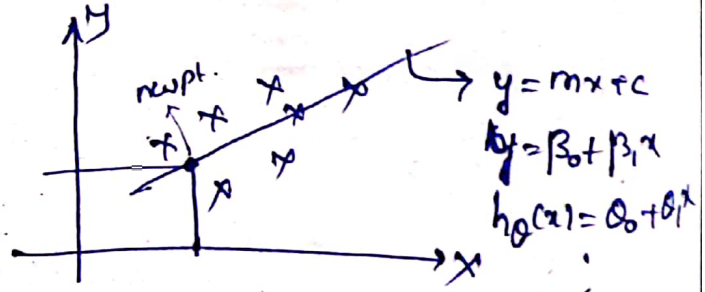


By LR, we



By LR, we will try to find out a best fit line, which will help us to do prediction

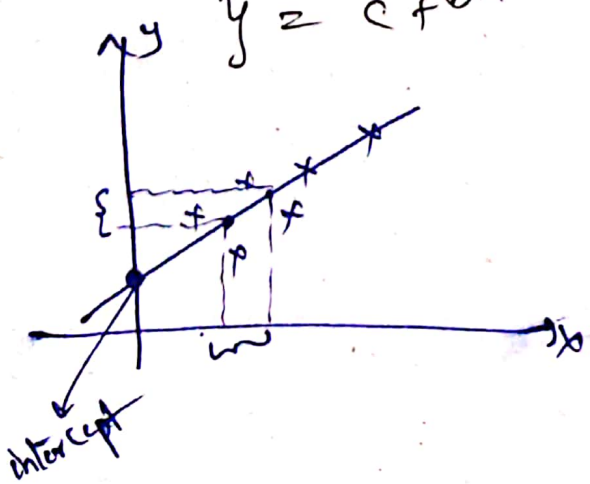
y = linear function (x)



To order to create a straight line, we use an equation

$$h_\theta(x) = \theta_0 + \theta_1 x$$

$y = c + mx$



θ_0 = intercept

when $x = 0 \Rightarrow h_\theta(x) = \theta_0$

↓

at what pt, you are meeting the 'y' axis

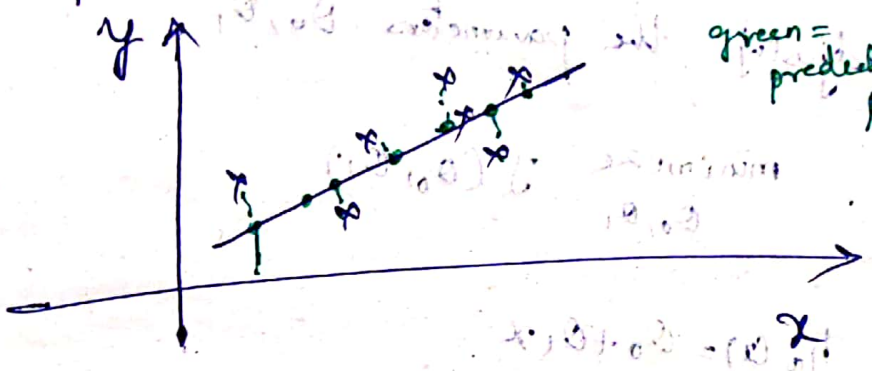
θ_1 = Slope / coefficient

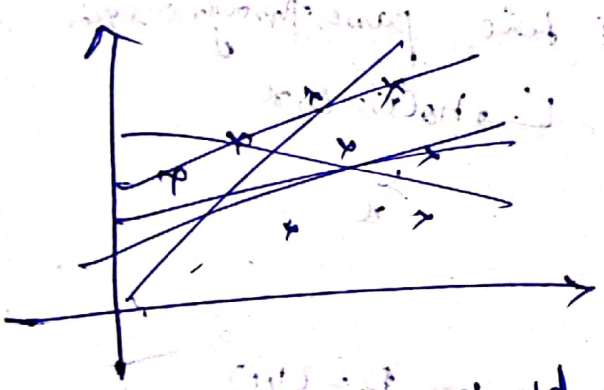
↓

in a 1-unit moment on x axis, what is the 1-unit moment in y-axis

x_p = data points.

Main aim ⇒ To create a best fit line such a way that the distance b/w the data pts & have & predicted pts should be every very less.





we should do many iterations and how do you know which line is best fit line ?

for this ⇒ we should start at one point & then go towards finding the best fit line.

Following this, we use a cost function.

Cost function

*** formulae for finding distance

$$C.f = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x) - y)^2$$

$$J(\theta_0, \theta_1) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2$$

↓ entire eqn called as squared error function

y = original pt

$h_\theta(x)$ = predicted pt.

$()^2$ ⇒ remove -ve value

m = no. of data pts.

Σ ⇒ add all the distance.

$\frac{1}{m}$ ⇒ avg of all

$\frac{1}{2m}$ ⇒ for derivation purpose

What we need to solve

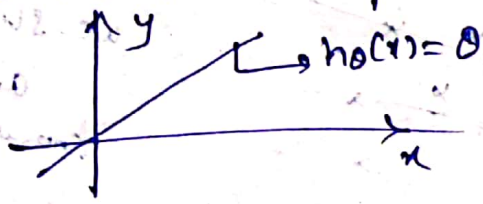
I need to minimize $\underset{\theta_0,\theta_1}{\text{minimize}} \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$

by adjusting the parameters θ_0, θ_1

$$\underset{\theta_0,\theta_1}{\text{minimize}}\ J(\theta_0,\theta_1)$$

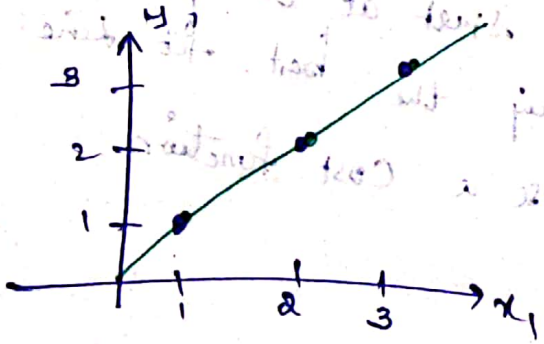
Let $h_\theta(x) = \theta_0 + \theta_1 x$

if $\theta_0 = 0 \Rightarrow$ the best fit line passes through origin



eg

(1) $h_\theta(x) = \theta_1 x$, $\theta_0 = 0, \theta_1 = 1$



$D_1 = (1,1)$
$D_2 = (2,2)$
$D_3 = (3,3)$

if $\theta_1 = 1 \Rightarrow$ then Straight line passes through all the pts.

When $\theta_1 = 1 \Rightarrow$ as predicted pts & Original datapts are same then the str. line passes through all the pts

Cost function $J(\theta_1) = \frac{1}{2m}\sum_{i=1}^{3}\left[\left(h_\theta(x^i) - (y^i)\right)^2\right]$

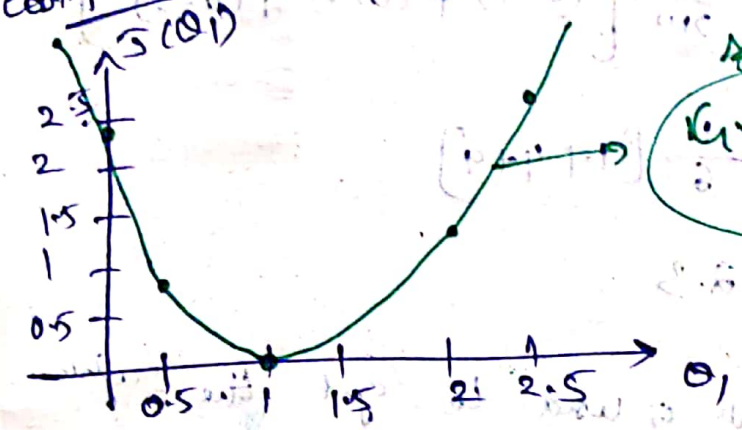
$= \frac{1}{2m}\left[(1-1)^2 + (2-2)^2 + (3-3)^2\right]$

$J(\theta_1) = 0$ → when $\theta_1 = 1$

$\theta_1 = 0 \Rightarrow J(\theta_1) = 0$

$J(1) = 0$

cost function graph



Gradient Descent

①

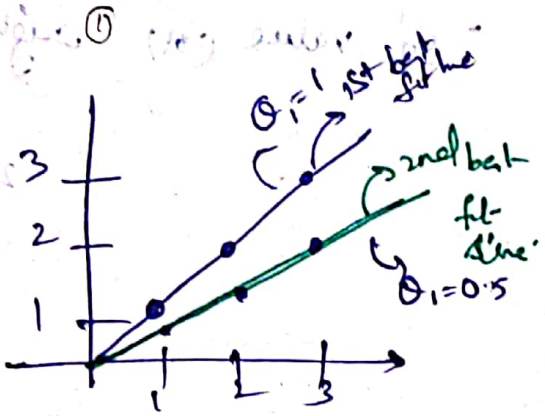
② $\theta_1 = 0.5$ $h_\theta(x) = \theta_1(x)$

$h_\theta(1) = 0.5(1) = 0.5$

$h_\theta(2) = 0.5(2) = 1$

$h_\theta(3) = 0.5(3) = 1.5$

predicted pts



$$J(\theta_1) = \frac{1}{2m}\left[(0.5-1)^2 + (1-2)^2 + (1.5-3)^2\right]$$

$$= \frac{1}{2(3)}\left[(0.5)^2 + 1 + 2.25\right]$$

$J(\theta_1) \approx 0.58$

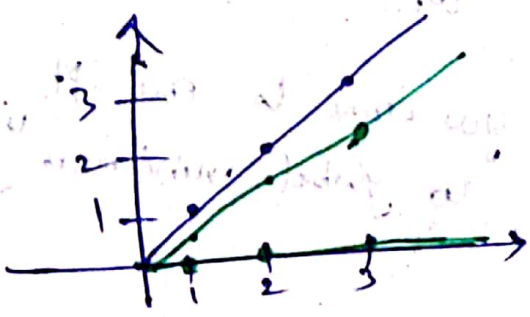
$J(0.5) = 0.58$

③ $\theta_1 = 0 \Rightarrow$ $\theta_1(x)$

$h_\theta(1) = 0$

$h_\theta(2) = 0$

$h_\theta(3) = 0$

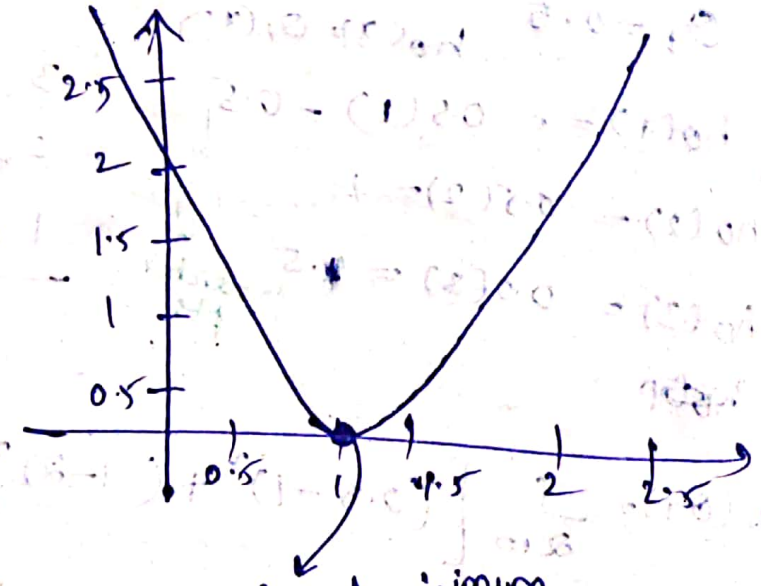


$$J(\theta_1) = J(0) = \frac{1}{2m}\left[(0-1)^2 + (0-2)^2 + (0-3)^2\right]$$

$$= \frac{1}{6}[1+4+9]$$

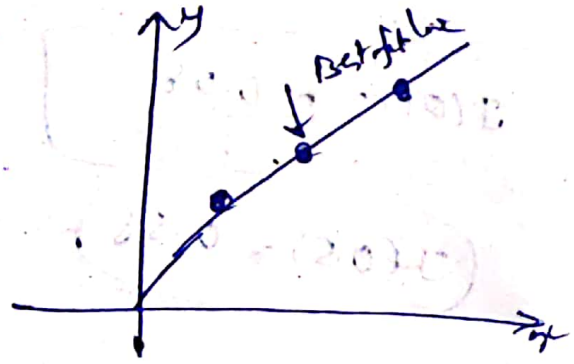
$$J(0) \approx 2.3$$

Gradient descent is used to get the right θ_1 value (or) right slope value.



we got 3 lines, which is the best fit line

When J at global minimum ⇒ the distance b/w actual pt & predicted pt is very very less.



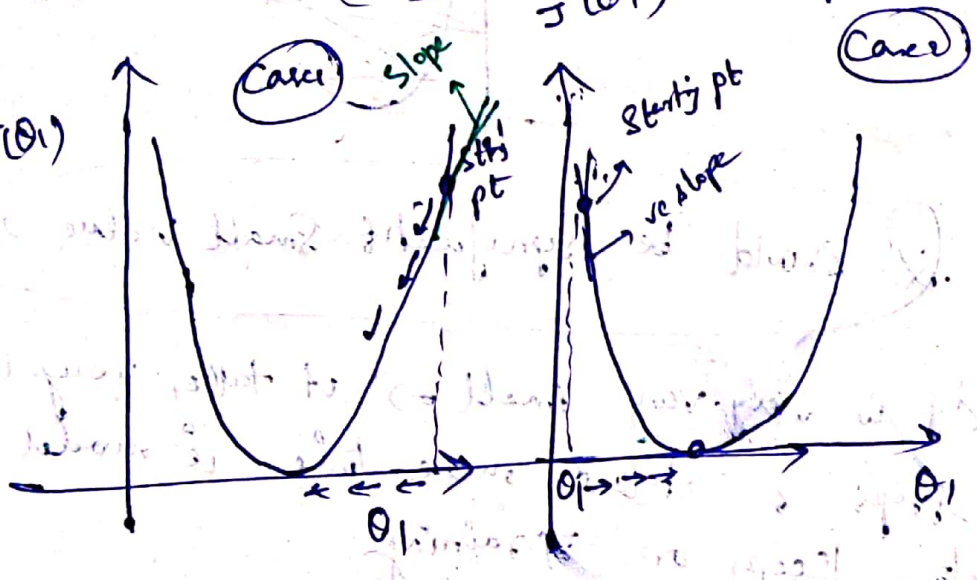
→ you come to one pt and then we reach to global minimum instead of using dy/dx value, how do you this?

for this we use Convergence Algorithm

α = learning rate

Repeat until Convergence
{
Condition updation

$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$ → derivative ⇓ slope.



after reaching to one pt, we will apply derivative ie we are trying to find out the slope.

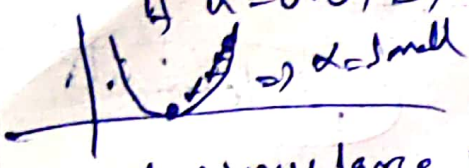
in the 1st case, ab starting pt, we calculate slope, it is +ve slope, now we should update the weights as we have +ve slope.

$\theta_1 := \theta_1 - \alpha$ (+ve value)

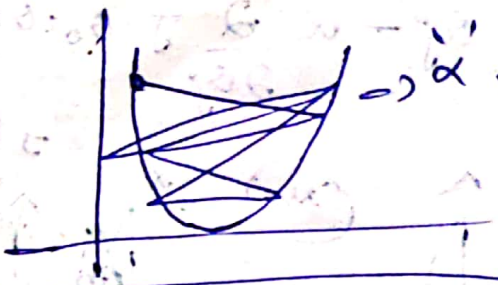
$\theta_1 := \theta_1 - \alpha$ (-ve value)
↓
-ve slope for Case

α = learning rate
⇓
by what speed we should reach to global minimum

If $\alpha = 0.01$ ⇒ low value ⇒ it is small small steps ⇒ α = small steps to reach global min



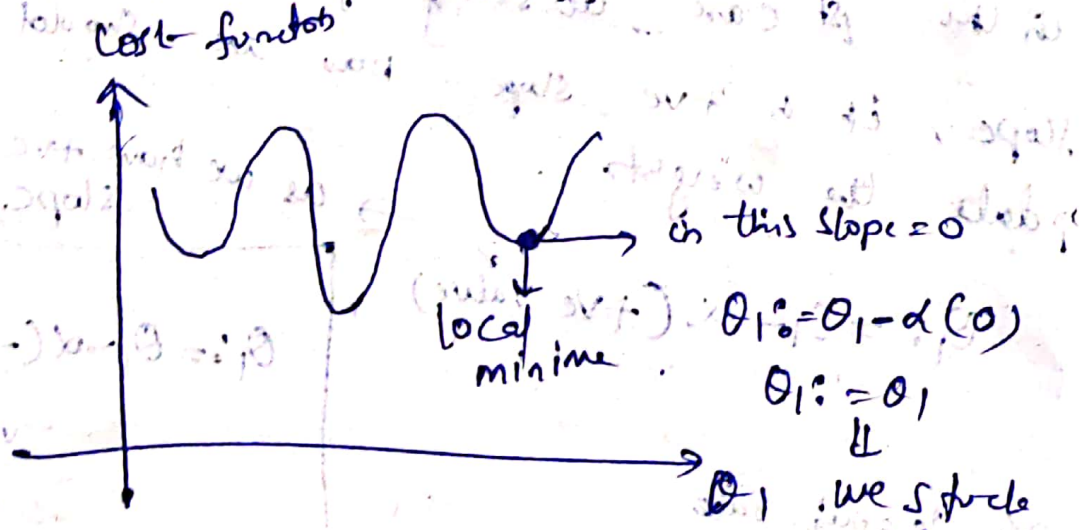
α ⇒ very large value ⇒ updation of θ_1 keeps on jumping, it will never reach c.m.



'α' should be significant small value.

α is very very small ⇒ it takes very less steps & takes more time ie model gets keeps on training.

→ If my cost function has local minima.



Local minima is not in LR, but in Deep learning.

Gradient Descent Algorithm

repeat until Convergence

{

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$$

}

$$\frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1) = \frac{\partial}{\partial \theta_j} \cdot \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$$

$\frac{1}{2m}(x^2) = \frac{x}{2m}$

if $j=0$ $\Rightarrow$

$$\frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left[h_\theta(x^{(i)}) - y^{(i)} \right]^2$$

$$= \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)$$

$j=1 \Rightarrow$ $$\frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1) = \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right) \cdot x^{(i)}$$

$$h_\theta(x) = \theta_0 + \theta_1 x$$

Repeat until Convergence

{

$$\theta_0 := \theta_0 - \alpha \frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1)$$

$$= \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)$$

$$\theta_1 := \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right) (x^{(i)})$$