Annexure No :

Descriptive Statastic

- when you are into account the entire population in your data analysis is called descriptive statastic

Inferential Statastic

- when you take a samples for your process and then you seeg this analysis applies to the entire population, It is called Inferential statastic.

outlier:- outlier is a datapoint that is very different than rest of the data point.

IQR - Interquartile Range

Interquartile Range

Ronak meet | Jenish Hali om Appit Alarmest

4000 5000 | 6000 7000 8000 | 9000 10 million

25th Percentile

75th Percentile

Range b/w 25% percentile and 75% is Interquartile Range

How is Percentile used in data science

Semester

Income Dataset

outlier Removal

General Data Analysis

↳ example is CISCE board exam result on percentile

What is mean?

Ronak	4K
meet	5K
Keval	6K
Kunal	7K
John bezos	10 million

⇒ If, I want to open a ~~not~~ premium car showroom in a specific area of a city, so as a Data scientist we will do analysis

⇒ So, our first step is that area which have higher Income people are leaving we choose o that area so, we will do average/mean of specific area.

⇒ $\frac{\text{total Income}}{\text{total person.}}$

What is median.

⇒ In previous example one outlier is there because of which our mean value is higher so, mean is not good choice.

So, Instead of this we can use median.

So, firstly we arrange in ascending order and we find middle element

⇒ What is mode?

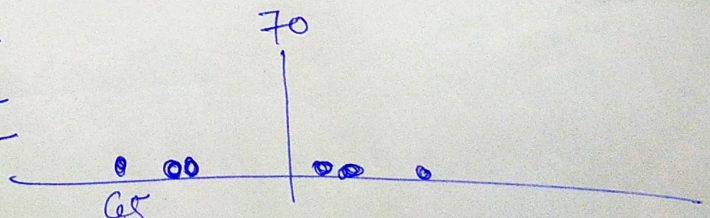
- mode here is mexican.
- mode means most freaquently occuring value.

Q. What is mean absolute Deviation and Standard deviation?

History test

Name	score	(score − Avg) for data = 70
Monan	75	5
Andrea	72	2
Sofia	68	2
Joe	65	5
Virat	67	3
Abdul	73	2

Mean = 3.16

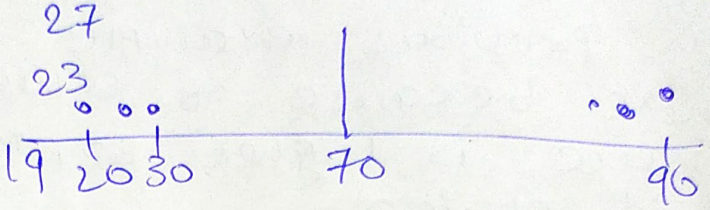


Points are nearby Average

Math Test (score-Avg)

Name	Score	MAD
Mohan	93	23
Andrea	96	26
Sofia	43	27
Joe	47	23
Visat	51	19
Abdul	50	20
	MAD	23

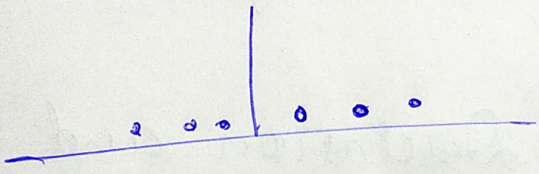
Average = 70



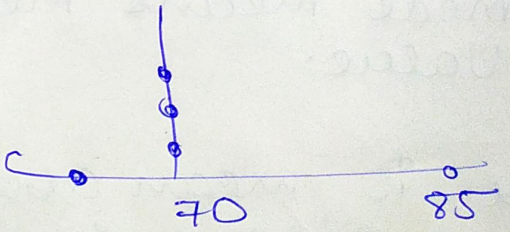
data point are spreaded out then Average

By seeing this we can tell that Math Test dataset is widespread than ~~history~~ test, so, this is called ~~Abs~~ mean Absolute Deviation.

But If we have



MAD = 3.33



MAD = 3.33

so In this scenario we can't predict.

= SD = √Variance

Parul University

Faculty of Engineering & Technology
Subject Name:
Subject Code:
B.Tech. ___Year____ Semester____

Annexure No :

What is Standarddeviation

$$\sigma = \sqrt{\frac{\Sigma (xi - u)^2}{N}}$$

N = size of Population

Standard deviation is a square root of Variance

Avg = 70

Name	score	Abs score (score − Avg)	(score − Avg)^2
Mohan	75	5	25
Andrea	72	2	4
Sofia	68	2	4
Joe	65	5	25
Virat	67	3	9
Abdul	73	3	9

Avg = 12.66

√Avg = 3.55

Standard deviation.

Mean test Name	score	Abs(score − Avg)	(score − Avg)^2
Mohan	83	13	169
Andrea	70	0	0
Sofia	70	0	0
Joe	63	7	49
Virat	70	0	0
Abdul	70	0	0

Avg = 36.33

√Avg = 6.02

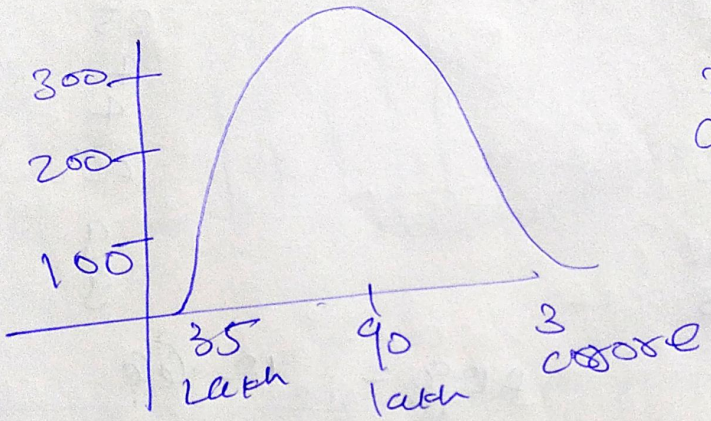
Page No :

Enrollment No :

So, using standard deviation we can say that math test data are wide spread.

What is Normal Distribution

⇒ You will have most of your data sample around average value and you have some data sample that is far away from average.

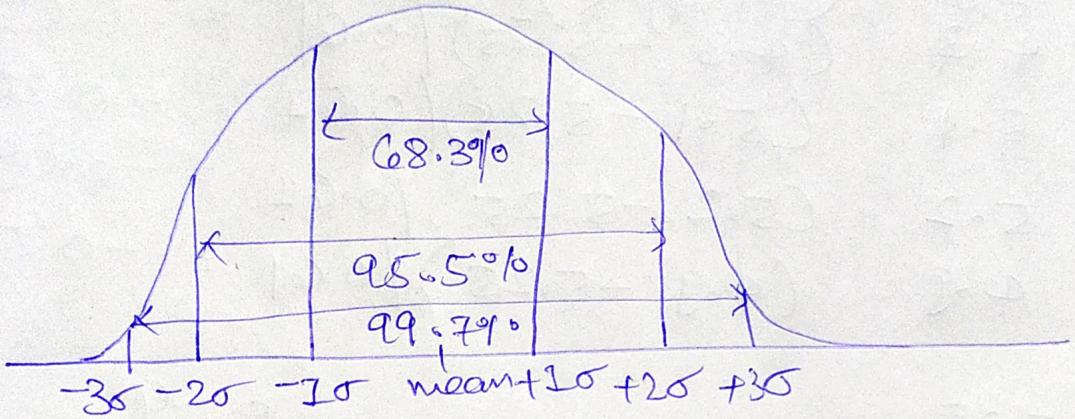


How can I use this in data analysis.

- outlier Removal.

⇒ what formula do we use to remove outliers.

How standarddeviation is used to Remove Standard deviation.



⇒ any datapoint that is below -3σ and above 3σ is treated as a outlier.

What is Z score

How many standard deviation away a datapoint is from mean.

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

μ = mean

σ = Standard deviation

Name	Height (ft)	Z-Score
Ronak	6.2	$(6.2 - 5.25)/0.61 = 1.53$
meet	5.7	$(5.7 - 5.25)/0.61 = 0.72$
Keval	4.6	$(4.6 - 5.25)/0.61 = -1.06$
Arpit	5.4	$(5.4 + 5.25)/0.61 = 1.04$
Abhay	5.9	$(5.9) - 5.25)/0.61$
Hariom	4.3	$(4.3 - 5.25)/0.61 = -1.55$
Jenish	5.1	$(5.1 - 5.25)/0.61 = -0.25$
Minit	5.2	$(5.2 - 5.25)/0.61 = -0.09$
mahesh	4.9	$(4.9 - 5.25)/0.61 = -0.58$

0.23

Average = 5.25

Standard deviation = 0.61

$$z = \frac{x - \mu}{\sigma}$$

Now let's apply z-score

What is logarithm?

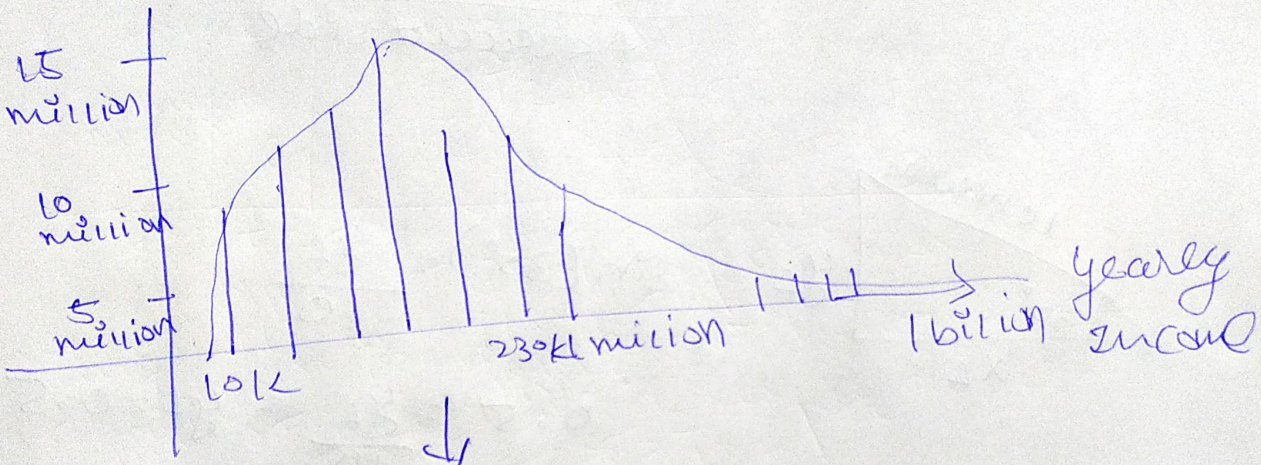
I have invested 5 $
After some point of time it gives me 125 $ so, how many years it take

$$\log_5 125 \rightarrow 3$$

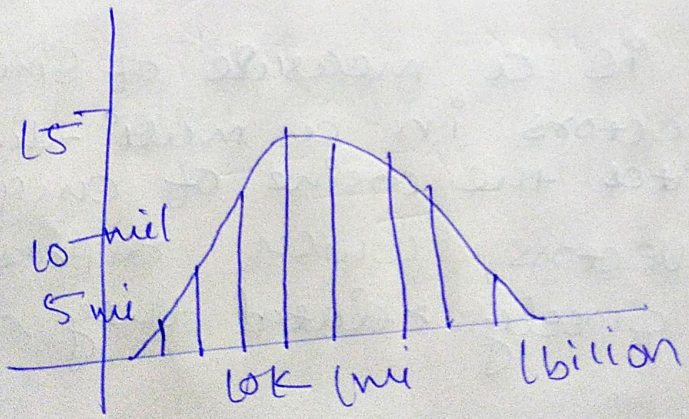
logarithm is an inverse of an exponent.

Popular log is $\log_{10} 10 = 1$

If you get a normal distribution by applying a log function to a dataset then dataset is log normally distributed.

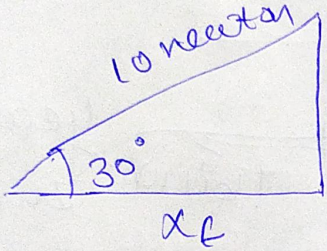


log(income)



$\sin(\theta) = \frac{\text{opposite}}{\text{hypotenuse}}$ $\cos\theta = \frac{\text{Adjacent}}{\text{Hypotenuse}}$

$\tan\theta = \frac{\text{opposite side}}{\text{Adjacent side}}$



$\sin(30^\circ) = \frac{yf}{10}$

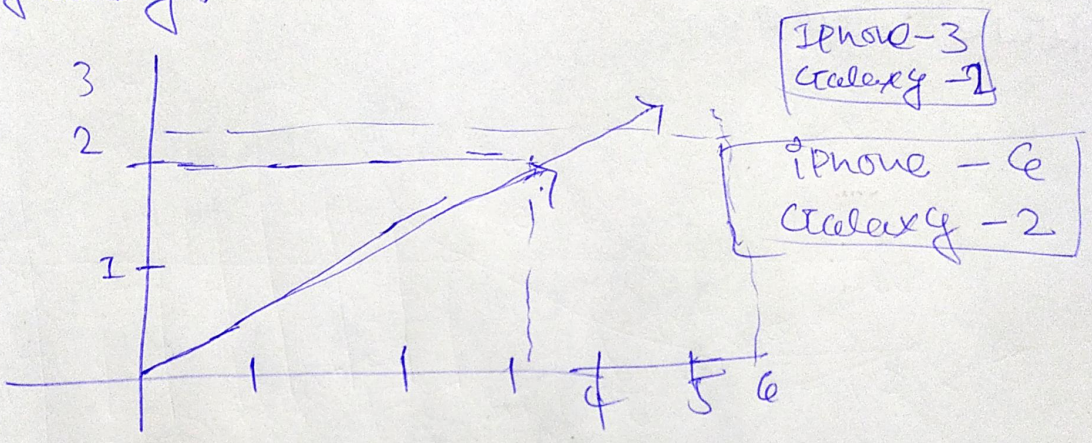
$0.5 = \frac{yf}{10} \Rightarrow yf = 0.5$

Cosine similarity

Cosine similarity is a measure of similarity between two vectors in a multi-dimensional space. It calculates the cosine of angle between the vectors, which represents how similar or closely related they are.

A example

- document similarity of IPhone and galaxy.



So here angle is 0 so, document is similar.

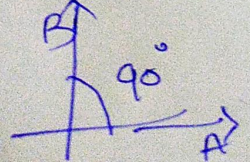
$$\text{Cosine Similarity} = \frac{A \cdot B}{\|A\| \|B\|}$$

$$\text{Cosine Similarity} = \frac{\|A\| \|B\| * \cos(\theta)}{\|A\| \|B\|}$$

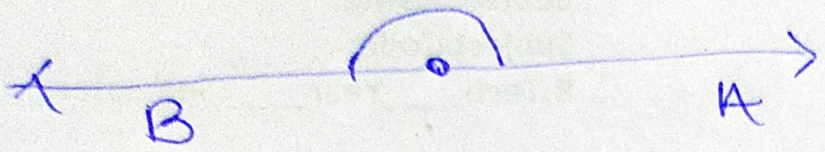
$$= \cos(\theta)$$

cos similarity = 1

cos similarity = 0



3.) cosimilarity -1



Cosine distance = 1 - cosimilarity

Cosine distance always represented in positive space.