**Statistics**

# Complete lecture :

http://statisticslectures.com/topics/statistics/

**statistics**

Different b/w math’s and statistics : Math’s - always talk about specific,e.g Tamil mark- 80 Statistics talk about approximation: Class avg score 80

Why statistics required?

Collet the data- organize the data- Present the data –analysis/interpret the data using some lib.

Using data we can take decisions.

Statistics is a branch of applied mathematics that involves the collection, description, analysis, and inference of conclusions from quantitative data.

Reference: wikipidea , <https://www.investopedia.com/terms/s/statistics.asp>

Example:

Person English – 88 and Maths -98 for looking into data I have done very well in maths.

I have collected all person marks & take avg

Will get mean of English – 80

Mean of maths -98

& Now tell who perform well,

Now the interspersion will vary. Because English avg mark is 80, But I took 88 so I performed well in English.

So data won’t tell everything we need to interpret.

Two types:

(1)Descriptive – Huge amount data – from the data we will get the conclusion/ statistics quantative.

Method Collet the data- organize the data- Present the data informative way.(EDA)

(2)Inferential - Method to which determine to understand the population from sample.

Descriptive statistics mostly focus on the **central tendency, variability(spread), and distribution** of sample data. Central tendency means the estimate of the characteristics, a typical element of a sample or population, and includes descriptive statistics such as [mean](https://www.investopedia.com/terms/m/mean.asp), [median](https://www.investopedia.com/terms/m/median.asp), and [mode](https://www.investopedia.com/terms/m/mode.asp).

[Variability](https://www.investopedia.com/terms/v/variability.asp) refers to a set of statistics that show how much difference there is among the elements of a sample or population along the characteristics measured, and includes metrics such as [range](https://www.investopedia.com/terms/r/range.asp), [variance](https://www.investopedia.com/terms/v/variance.asp), and [standard deviation](https://www.investopedia.com/terms/s/standarddeviation.asp).

The [distribution](https://www.investopedia.com/terms/p/probabilitydistribution.asp) refers to the overall "shape" of the data, which can be depicted on a chart such as a histogram or dot plot, and includes properties such as the **probability distribution function,** **skewness, and kurtosis**. Descriptive statistics can also describe differences between observed characteristics of the elements of a data set. Descriptive statistics help us understand the collective properties of the elements of a data sample and form the basis for testing hypotheses and making predictions using inferential statistics.

## Inferential Statistics

Inferential statistics are tools that statisticians use to draw conclusions about the characteristics of a population from the characteristics of a sample and to decide how certain they can be of the reliability of those conclusions. Based on the sample size and distribution of the sample data statisticians can calculate the probability that statistics, which measure the central tendency, variability, distribution, and relationships between characteristics within a data sample, provide an accurate picture of the corresponding parameters of the whole population from which the sample is drawn.

Inferential statistics are used to make generalizations about large groups, such as estimating average demand for a product by surveying a sample of consumers' buying habits, or to attempt to predict future events, such as projecting the future return of a security or asset class based on returns in a sample period.

Example: we are collecting marks from different college. But impossible to collect all data.

Based on section of data ( sample) we will take decision . so we are conclude population from characteristic from sample.

Another example: Decide the below poverty line – city to city the income will vary COB may be < 1000 consider BPL or BAN may be >1500. Entire people belong to BAN is population, So it’s not possible collect all data.so we need to take decision based on fractional ( sample) data.

[Regression](https://www.investopedia.com/terms/r/regression.asp) analysis is a common method of statistical inference that attempts to determine the strength and character of the relationship (or [correlation](https://www.investopedia.com/terms/c/correlationcoefficient.asp)) between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables). The output of a regression model can be analyzed for [statistical significance](https://www.investopedia.com/terms/s/statistical-significance.asp), which refers to the claim that a result from findings generated by testing or experimentation is not likely to have occurred randomly or by chance but are instead likely to be attributable to a specific cause elucidated by the data. Having statistical significance is important for academic disciplines or practitioners that rely heavily on analyzing data and research.

### What is the difference between descriptive and inferential statistics?

Descriptive statistics are used to describe or summarize the characteristics of a sample or data set, such as a variable's mean, standard deviation, or frequency. Inferential statistics, in contrast, employes any number of techniques to relate variables in a data set to one another, for example using correlation or regression analysis. These can then be used to estimate forecasts or infer causality.

1.Descriptive – Describe the data

Example: Age feature – from this we can conclude few characteristics like mean,median,mode,etc..

Population: Exit poll, 100K Emp – We need to buy T-shirt, The entire 100K is population, from that we will select some set is called sample.

Two term Population and sample

Population mean formula : 1/N

Sample mean formula: 1/n

Data type in statics  
==================

1 . Categorical – Group or some categorical – Every time we are dividing the data

Yes/No

Rating: 1,2,3,4,5 - Even numeric its cat

Theater ticket A,B,C,D - Even string also cat.

2. Numerical – Descriptive - whole number - Range

Continues - with floating number

Label of measurements: ( Inferential statstics)

=====================

1. Qualitative – Variable that categories or describe the quality of population/sample elements.

Nominal : Not in order example Brand of car

Ordinal : Put into order – Example review rating

1. Quantitative: Variable that describe the magnitude of the population

Interval :

Ratio:

1. **Measure of Center Tendency:**
2. Mean
3. Media
4. Mode

**Mean:**

Two type of mean : Sample mean & population mean

50,40,,30,40,10 Avg/Mean = ( 50+40+30+40+10)/5 = 34

Suppose the late element is 100. Then Mean: 50,40,,30,40,10 Avg/Mean = ( 50+40+30+40+100)/5 = 52

Huge difference due to outliers.

We will remove outlier & cal correct?

May be huge data – huge outlier or the outlier has some info. So we will go median

**Median :** reduce the impact of outlier

Sort the value & take middle value.

Suppose we have even number – we need to consider middle of two numbers

**Mode:** Most frequency value. Maximum **three or four modes.**

All are unique we will get error.

1. **Skewness**

**Why Skewness?:**

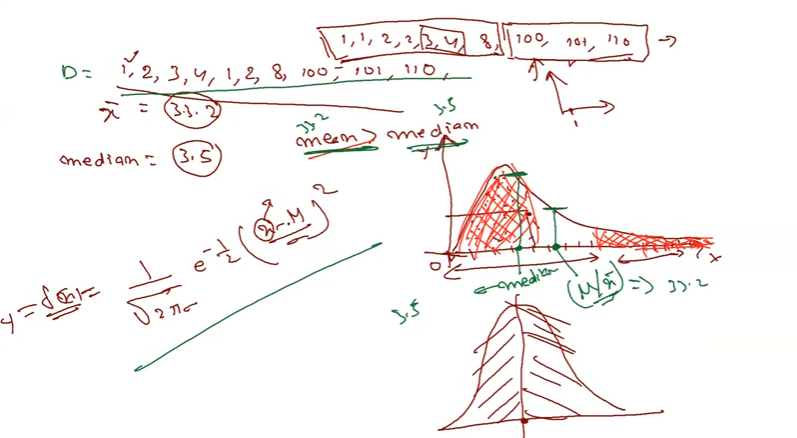
We should not take all the data Normalize & standardization –Based on the analysis

Here positive skewness – Right side is outlier

Number of frequency in the right side is low & high magnitude ( Mean high data values)

Mean is > median

X-axis – Range(distribution) – Y-axis – (Frequency of data)



The above example mean > median – The avg the data b/w 1 to 8. It’s right sknewess.

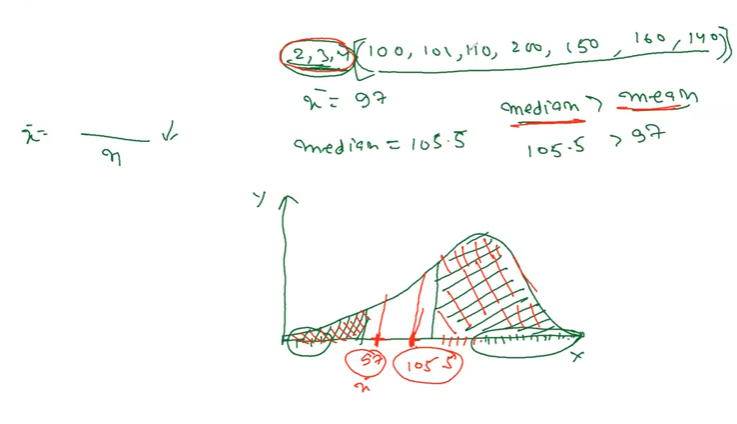
The above formula is Normal distribution formula.

In statistics, there is a relationship between the mean, median and mode that is empirically based. ... Mean – Mode = 3(Mean – Median)

**Left sknewness:** ( Negative) – Some time will not impact much – so we can take the data as it is not an issue.

Median > Mean

Short Cut: Little Indian ->Left median>Mean

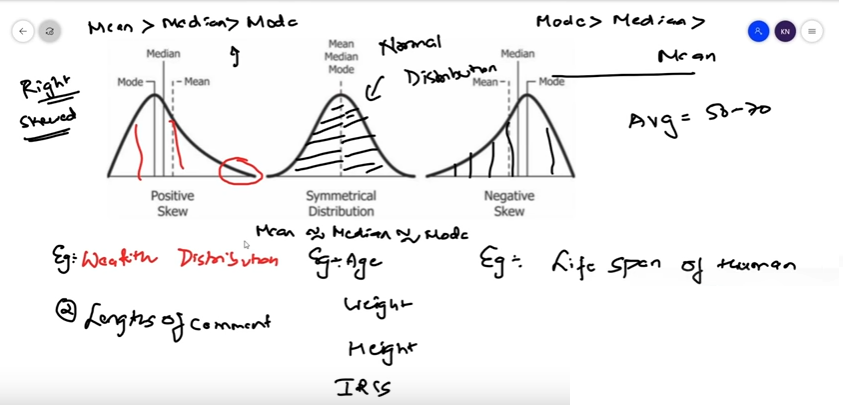
Left skewness will not impact much we can directly take .

Skewness will be find based on comparison of mean & median

Symmetrical- Left & right curve are same ( mirror)

Question:

**Example of right & left skewness and relationship?**



Right: Mean>Median>Mode

Symmetrical : Mean =(approximate)=Meidan(approximate)=Mode

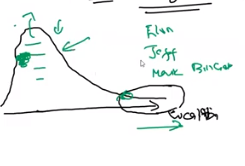
Left: Mean<median<mode or Mode>median>mean

Measure of dispersion:

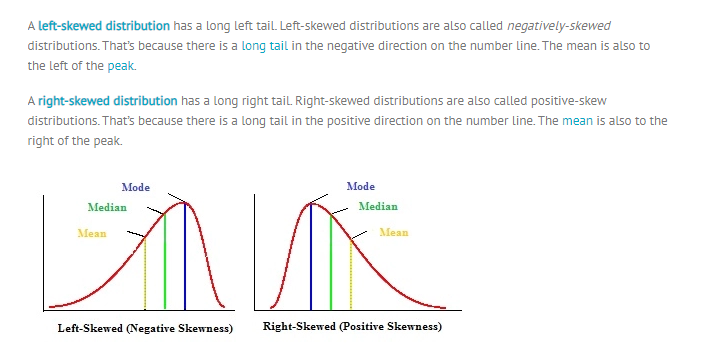
Each data we will get the symmetric curve, why different curve ? because the vraince.

Ultimate Goal is convert any skewness data to Normalize

Right skewness example: **-**Measure the wealth of the population . Whether the data is Normal distribute? No.

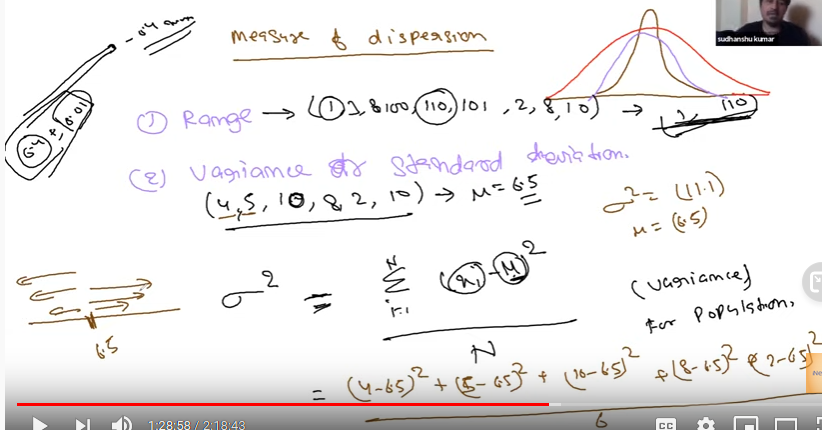


Left skewness example : Retirement Age



1. **Dispersion: Variance(spread), Std**

Variance : How far my point from mean



Every bell curve is symmetric,

But Each bell curve is different size. How do we find the measure of dispersion ?

We can use (1) range (2) Variance or std

If the dispersion in not even the ML algo will not learn properly.

If we send raw data to ML algo – my data is not evenly distributed – ML algo will not learn relationship.

What is the range of data ? example : 1 to 110

Variance will help to understand how my data vary with respect to my mean?

If the variance is high we can say my data is too far from mean.

Example:

e.g:

I am manufacture of stra – The avg radius +/ 0.01 – This is desperation.(range)

When will you complete your task = +/ 3 days

**Difference between Variance & standard deviation?**

**Variance is how the data is spread from the mean**

**Standard deviation is exact distance or absolute distance from mean.**

Variance is used to measure the spread, Standard deviation will help to identify whether element will be located from mean

Above figure Formula for population variance :

Given data set variance = 11.1

Mean is 6.5

The data will spread across the mean. Either left and right.

Will tell How much data variance in my data or how far my data from mean?

The variance is increase the dispersion curve spread will increase. In this case standard deviation is not significant.

If the variance is lot we need to take some treatment in both ML & DL. We need to normalize , then std deviation is significance.

Variance is summation of distance or dispersion from mean

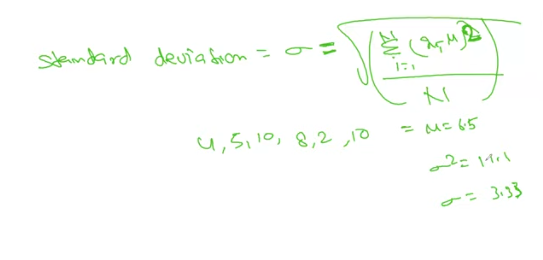
If the variance is very high : Example: Travel from point A to B

Speed is : 10,120,20,150,10,180 ,20,100

Range is 10 to 180

What will be the conclusion? Huge variance – So the driver is not good.

Standard deviation:

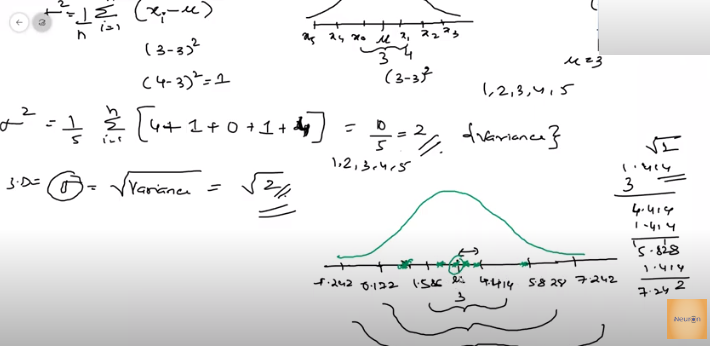


Every data is fluctuated around the mean.

6.5+/3.33 Then my data is fluctuated between 3.17 to 9.83 . it true? If I look the data I can find some data set 2 and 10 also. So it’s not true. – This is called one standard deviation.( Around 64%)

6.5+/ (2 \* 3.33) – Then we will get 2 standard deviation data ( Around 95%)

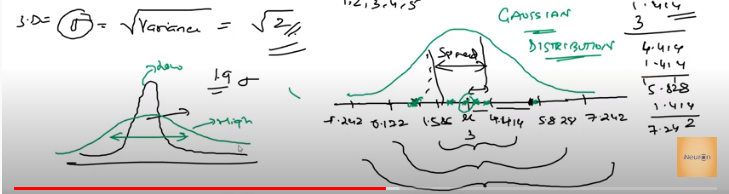
6.5+/ (3 \* 3.33) – Then we will get 3 standard deviation data (Around 99)



Mean + std = right value

Mean –std = left values

Most of the data b/w with two standard deviation - Gaussian distribution



Normally distributed data : The higher stand deviation higher spread.

Above green one have higher standard deviation, so higher spread.

All the data will come under bell curve? No.

**III.Random variables:**

Roll the dice = {1,2,3,4,5,6} - discrete ( whole number), Number of children, Number of eating plate

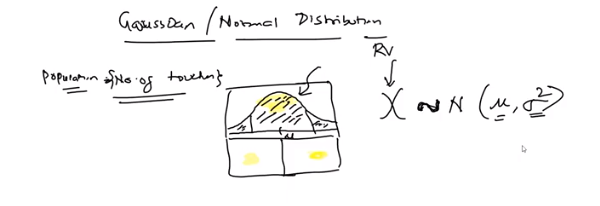
1000 members measure height = Continues. Weight , KM which I run.

Age is – both Continues( if we mention with Year,month, date) & discrete( whole number)

1. Gaussian/Normal distribution:

Laptop touching in touch pad.

Population : Number of touch - 80% data will be fall in to



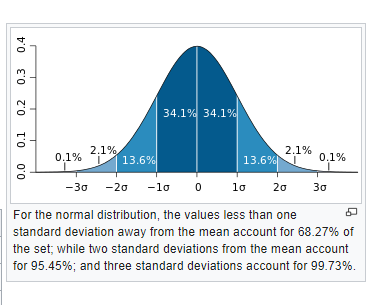
Whether all the data come under Gaussian or Normal? Depends on data

Height of the population : we need domain expertise to conclude whether data will come under Gaussian or not.

Random variable belong to Normal/Gaussian then it has mean & variance

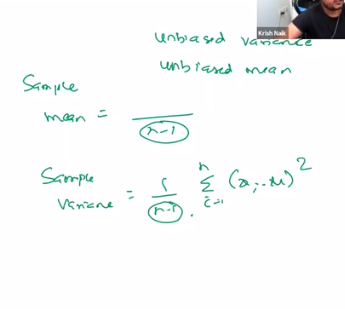
For example, the [normal distribution](https://www.statisticshowto.com/probability-and-statistics/normal-distributions/) is a [symmetric distribution](https://www.statisticshowto.com/symmetric-distribution-2/)with no skew. The tails are exactly the same.

1. **Empirical formula – How much data distributed – based on that we will say 68.27 % fall in two standard deviation**



1. **unbiased variance: (**[**https://www.khanacademy.org/math/ap-statistics/summarizing-quantitative-data-ap/more-standard-deviation/v/another-simulation-giving-evidence-that-n-1-gives-us-an-unbiased-estimate-of-variance**](https://www.khanacademy.org/math/ap-statistics/summarizing-quantitative-data-ap/more-standard-deviation/v/another-simulation-giving-evidence-that-n-1-gives-us-an-unbiased-estimate-of-variance) **)**

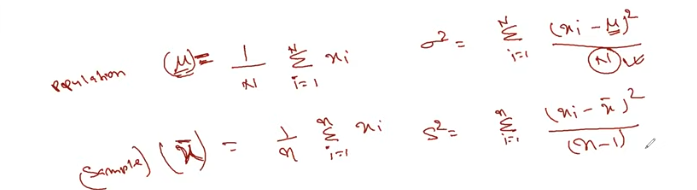
Use N-1 for formula ( when you use different different sample )



**Normalize formula: X- axis data , Y –axis is frequency of data.**

Ultimate Goal is convert any skewness data to Normalize

Summary of day1: Mean,median,mode,population variance, population std, Skewness



Why n-1? Based on experiment sample is subset of population.

So we need to take n-1

DAY2:

Standard Normal distribution or standard Normal Variate ( Z)

Normal distribution means any random X variable belongs to (Mu, sigma) ( Means any value of Mu & any value of standard deviation )

Standard normal distribution means any random variable y belongs to ( zero mean, one standard deviation) e.g(0,1)

If we have mean & standard deviation we can form the curve.

Why we need Standard Normal distribution?

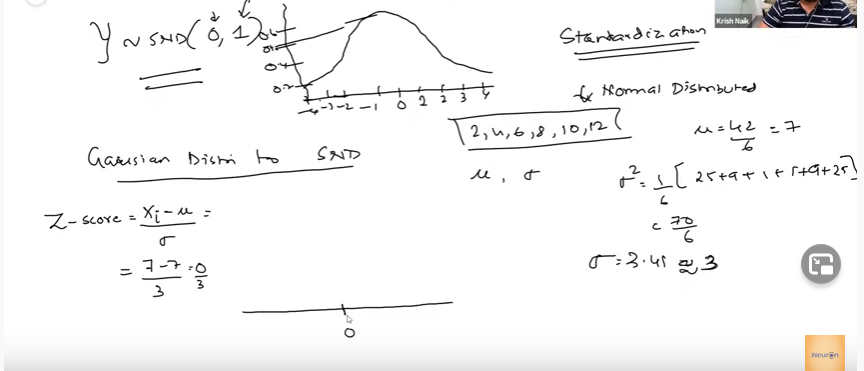
In machine learning we need common standard( scaling down the feature data) so that prediction will be more accuracy.

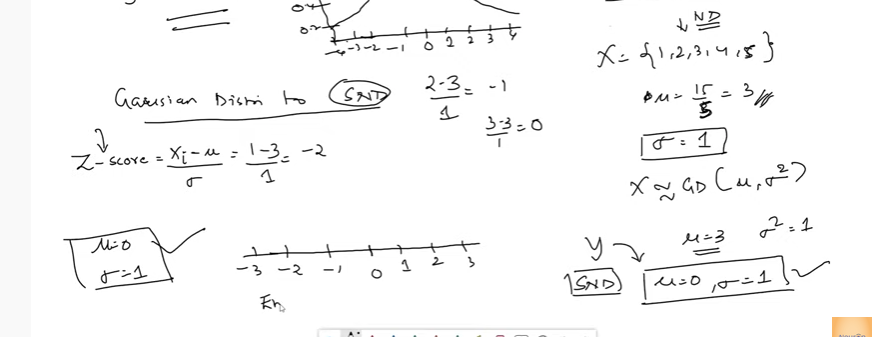
It’s also called standardization

How to convert Normal distribution to Standard normal distribution?

Use Z-score

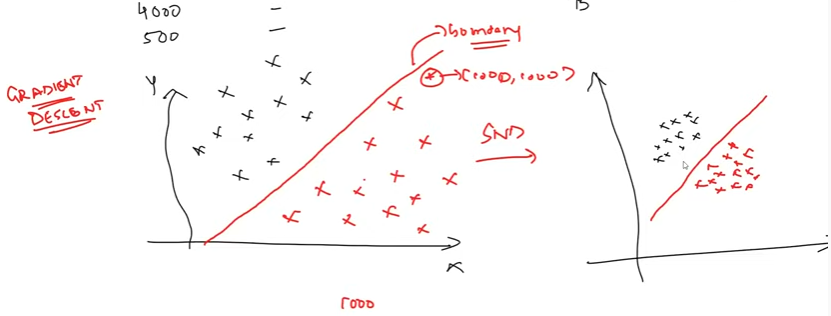
Whether the standard deviation will reduce the outlier impact? It’s depends on algorithm





Example my X, Y large value & I will convert Standard Normal distribution so the more points are close

& easy to classify.



**After conversion whether empirical formula will hold or not? Still will hold Empirical formula .**

Why we need?

When we convert Normal data into Standard Normal distribution

Data will fall under -3 to 3

So the calculation will be easy when go for complex machine learning algorithms.

Some ML algorithm we will use equalization distance.

Can we change other distribution to Standard Normal distribution?

Yes, But we need to know the original data belong to which distribution ?

**Sample distribution:**

Consider random variable – it can be any distribution

Consider I have 1000 random variable out of 1000 we will take random of 30 sample

Here m = 1000 n= 30 or more

This we can call as S1, Similarly we can take another Sample S2, like wise we can take n number of sample.

Next we need to calculate the mean for each sample, sample mean we will call as X1 bar, X2 bars..

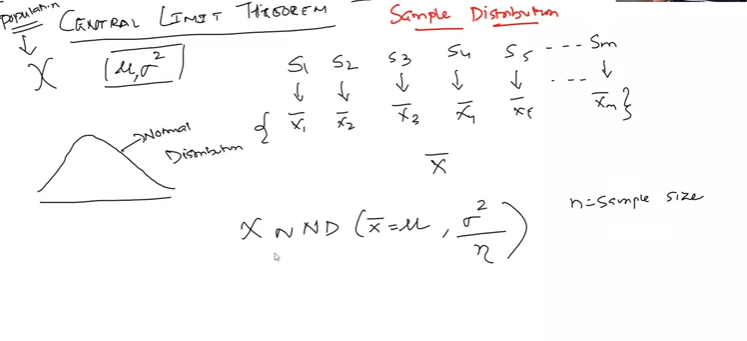
Here X bar is collection of {x1bar,x2bar,x3 bar,etc…)

If we try the plot using these means values we will get bell curve/Normal distribution/Gaussian distribution.

This curve mean (sample mean) is equal to 1000 variable mean or population mean.

The variance is equal to sigma squre/n

Here n is sample size.

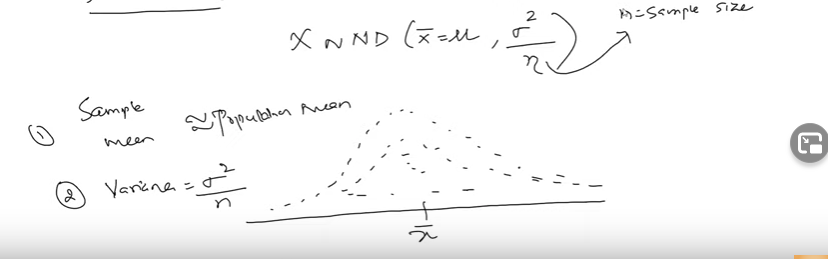


**Central Limit Theorem:**

Central limit theorem says my enter population I will take multiple sample & calculate the mean and create the distribution

(a)Then my sample means approximately equal to population mean.

(b)Variance is equal to sigma square/ number of sample



Various statistics say the sample size should be equal to 30 or more then only we will get the sample mean approximate population mean.

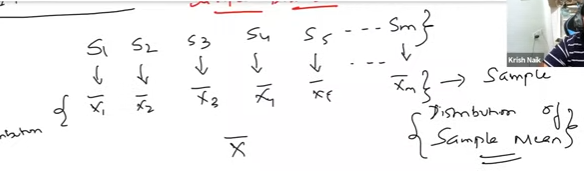
**Sample distribution of sample mean.**

X1, X2 bar are sample distribution

If we calculate mean of x bar collection { X1 bar,X2 bar,x3 bar ,etc..} it’s called sample distribution of sample mean.

If we take mean of the sample distribution we can call as sample distribution of sample mean.

sample distribution of sample mean will follow normal distribution if plot using collection values



**Importance of CLT:**

**===============**

Even though we have sample distribution we can find the approximately population mean.

Using this we can assume lot of statistics

Why distribution are used?

To data analysis and get the information from data.

**Pdf: Probability density function**

**It kind the math equation help to find the probability of occurrence in the experiment.**

Example: Roll the dice

{1,2,3,4,5,6} probability of getting 3 is 1/6

If we roll two dice what is the probability of 3 getting both dice?

If we roll multiple dice what is the probability of total 10 getting?

Example: Particular region how much fire or ambulance will deploy. Based on probability.

We can solve using pdf.

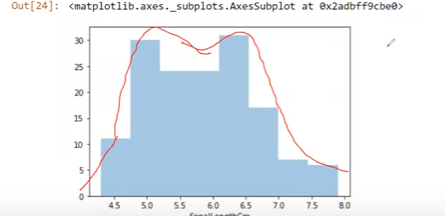
**Different Type of PDF:**

1. Normal distribution
2. Standard Normal distribution
3. Bernoulli distribution.
4. Binomial distribution
5. Uniform distribution
6. Standard T -distribution
7. Passion distribution



# Green on has more standard deviation

**Pdf is smooth version on top of histogram.**



**1.Binomial distribution:**

Two output from the certain experiment. (Each event is independent)

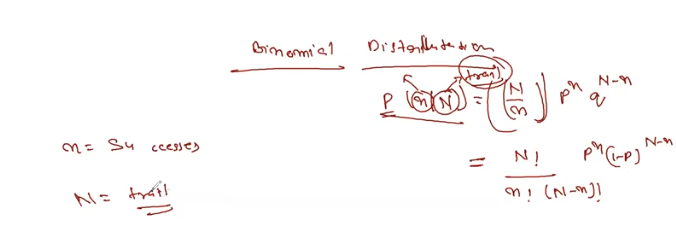
1.Whether I will pass or Not

Formula:

N= No of Trail , n is success from Trial.

N! Formula is expansion formula of 1st one.

N/n is co-efficient



Practical Example: HOSPTIAL 100 patient taking treatment out of 75 die.

If I select randomly 6 person (sample) from that probability of 4 person should be recover.

Probability of recovery is ( %) = 100-75 => 25% or 0.25

Probability of Failure is ( %) = 75 => 75% or 0.75

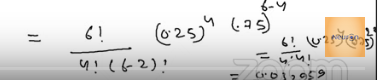
Why Binominal – we will get output either success or failure.

Here also person will recover or not.

Based on question we can see 6 person so trial will be N= 6.

n (success rate) = 4

Sub suite all the value into formula:



N = 6! n =4! P = 0.25 q is 0.75

= 0.03

The conclusion is 0.03 % will be recovery of select 6 person from 100 people.

Another Problem:

1. If I roll the dice 3 time what is the probability of not getting 5?
2. What is the probability of at least one time 5
3. What is the probability of all time getting 5.

Trail = N = 3

No of success n = 0 ( Not getting 5)

Probability getting 5 is = 1/6

Probability of not getting 5 is => 1-1/6 => 5/6

Total probability is always one – so we are subtract the value

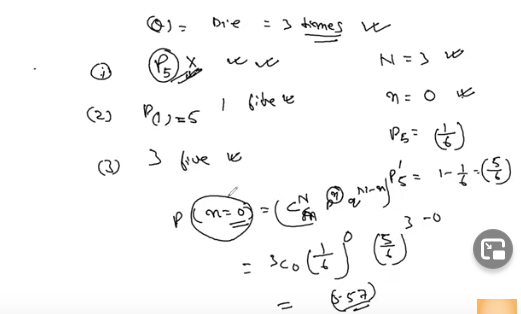
Here I am consider P is success means ( not getting 5)

P(n=0) = N/ni \*pn\*q(N-n)

= 3/0\*(1/6)\* (5/6)3-0

=0.57

So 57% we will not get 5 out of 3 trials.

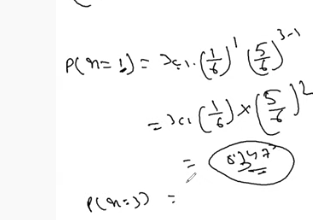


Question2 :

Probability of getting 5 at least once 1/6

Probability of not getting 5 is => 1- 1/6 =>5/6

No of success n = 1 ( Getting 5 )

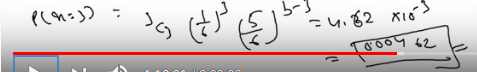


Question 3:

Probability of getting 5 all time 3/6

Probability of not getting 5 all time 3/6

No of success n = 1 ( Getting 5 all time)



Conclusion :

If we are expecting more time the probability is less 0.0004 ( 0.4%)

If we are expecting one time the probability is more 0.34

Another example:

(1)A person going to hit the arrow into target: A per son claim out of 5 times he hit target 4times

what is the probability of hit more than 2 times in the Target

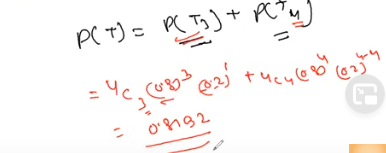
Now we need to consider the probability of hitting Target.

P(T) = P(T3) + P(T4)---------🡪 here 2 & more so need find probability of hit 3 times and also 4 times.so we need to consider both for final probability.

This also success or failure case . So use binomial distribution.

Success rate of hitting the Target = 4/5 = 0.8

Failure rate of hitting Target is = 1/5 = 0.2



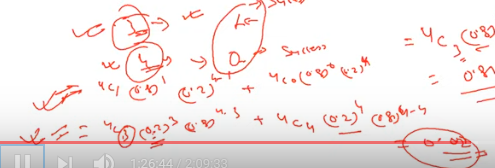
So 87% chance he was able to hit more than 2 times in the Target.

(2) what is the probability of the person missing 3 times given chance of 4?

The possible option 3 missing 1 hit – Probability success

Another possible option all 4 missing – Probability success.

Here we can use either success 4C1 & 4C0 or Failure 4C4 & 4C3



**2.Poisson Distribution**

Using Poisson distribution we can find the probability b/w the some time period or event of occurrence or occurrence of some kind of event in certain time.

Example : In 1 hour 5 accident happen what will be the probability of 10 accident will happen another 1 hr or 1 day?

In Car wash 10 cars I am getting 1 hr. What will be probability of getting 100 cars in one day?

Both case we are using Time period. When ever we talk about time interval use Poisson distribution.

Formula: P(*x*; μ) = (e-μ) (μx) / x!

Another example: Sales person sales 3 insurance policy /week. We can find what is the probability of selling 1 policy, Not selling policy ?

(1)Find the probability of at least selling something?

1-P(X0)

Here μ = 3

P(0;3) = (2.71828-3) (20) / 0!

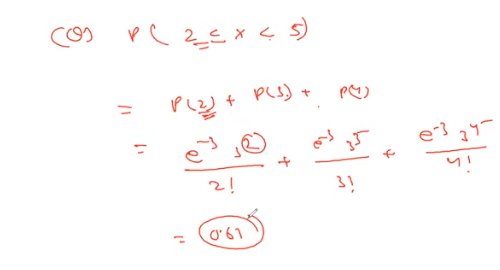
probability of not selling something = 0.0497

probability of selling something = 1-0.0497=>0.95621

(2) What is the probability of selling more than 2 and lesser than 5.

Here we need to consider P(2)+P(3)+P(4)

Here μ = 3



(3)The average number of homes sold by the Acme Realty company is 2 homes per day. What is the probability that exactly 3 homes will be sold tomorrow?

*Solution:* This is a Poisson experiment in which we know the following:

* μ = 2; since 2 homes are sold per day, on average.
* x = 3; since we want to find the likelihood that 3 homes will be sold tomorrow.
* e = 2.71828; since *e* is a constant equal to approximately 2.71828.

We plug these values into the Poisson formula as follows:

P(*x*; μ) = (e-μ) (μx) / x!

P(3; 2) = (2.71828-2) (23) / 3!

P(3; 2) = (0.13534) (8) / 6

P(3; 2) = 0.180

Thus, the probability of selling 3 homes tomorrow is 0.180 .

1. Cumulative Poisson Probability

A **cumulative Poisson probability** refers to the probability that the Poisson random variable is greater than some specified lower limit and less than some specified upper limit.

**Cumulative Poisson Example**  
  
Suppose the average number of lions seen on a 1-day safari is 5. What is the probability that tourists will see fewer than four lions on the next 1-day safari?

*Solution:* This is a Poisson experiment in which we know the following:

* μ = 5; since 5 lions are seen per safari, on average.
* x = 0, 1, 2, or 3; since we want to find the likelihood that tourists will see fewer than 4 lions; that is, we want the probability that they will see 0, 1, 2, or 3 lions.
* e = 2.71828; since *e* is a constant equal to approximately 2.71828.

To solve this problem, we need to find the probability that tourists will see 0, 1, 2, or 3 lions. Thus, we need to calculate the sum of four probabilities: P(0; 5) + P(1; 5) + P(2; 5) + P(3; 5). To compute this sum, we use the Poisson formula:

P(x < 3, 5) = P(0; 5) + P(1; 5) + P(2; 5) + P(3; 5)

P(x < 3, 5) = [ (e-5)(50) / 0! ] + [ (e-5)(51) / 1! ] + [ (e-5)(52) / 2! ] + [ (e-5)(53) / 3! ]

P(x < 3, 5) = [ (0.006738)(1) / 1 ] + [ (0.006738)(5) / 1 ] + [ (0.006738)(25) / 2 ] + [ (0.006738)(125) / 6 ]

P(x < 3, 5) = [ 0.0067 ] + [ 0.03369 ] + [ 0.084224 ] + [ 0.140375 ]

P(x < 3, 5) = 0.2650

Thus, the probability of seeing at no more than 3 lions is 0.2650.

1. A person go through 0 -100 page and he find average 1.5 mistake .

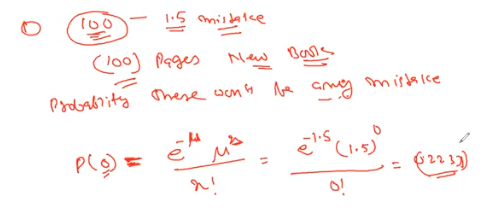
The same person took another new 100 page what is the probability of wont able to find any mistake?

Again here 0 to 100 Interval:

Zero error I am looking?

Here μ is 1.5

Here x is 0.



(6) The same person took 400 page what is the probability of wont (Zero) able to find any mistake?

Interval of 100 page is 1.5

Interval of 400 page is 4\* 1.5 = 6

Here μ is 6

Here x is 0

DAY3:

**3.Normal Distribution:**

Percentile:

How to calculate the percentile?

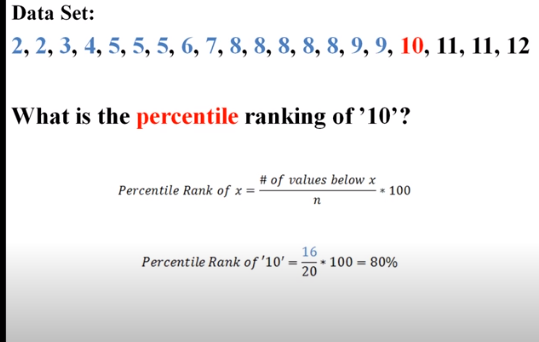
Height = {168,170,150,160,182,140,175,180,170,190}

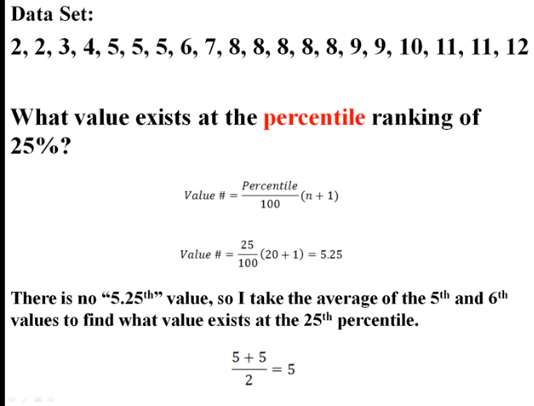
Order the data = {140,150,160,168,170,170,175,180,182,190}

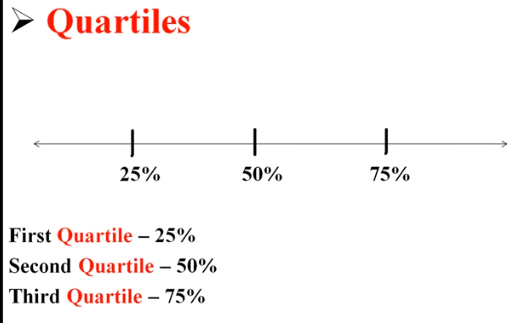
5th number is 50th Percentile means 50% less than 170.

190 is 90% percentile

Another example:





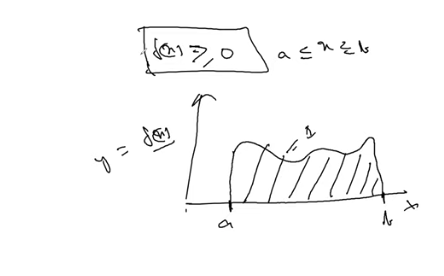


PDF – Probability ,CDF - Cumulative

PDF:

What is mean by PDF?

Assume y =f(x) is >0 and x event b/w a & b. area= 1



Based on the area I can represent the probability of occurrence for that particular data.

Uni varite , - Single plane, histogram , histogram with KDE, PDF,CDF graph

bi-variate - Scatter

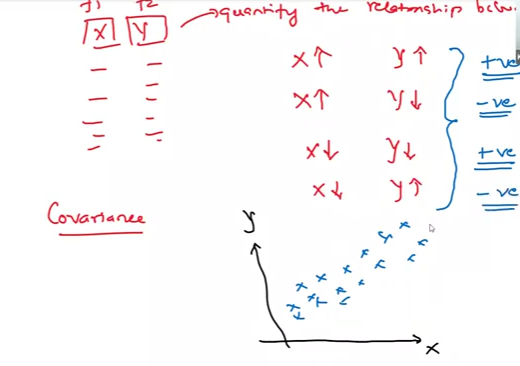
multi varite –pair plot

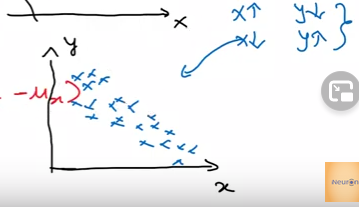
Refer : Statistics EDA Note book

Variance of x is Var(x) is covariance(X,X)

Covariance:

Quantify the relationship b/w feature.

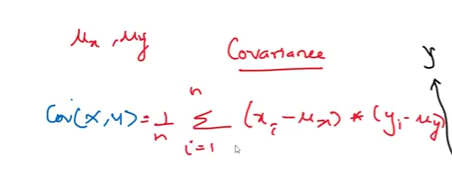


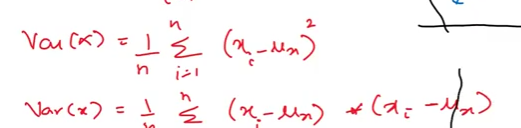


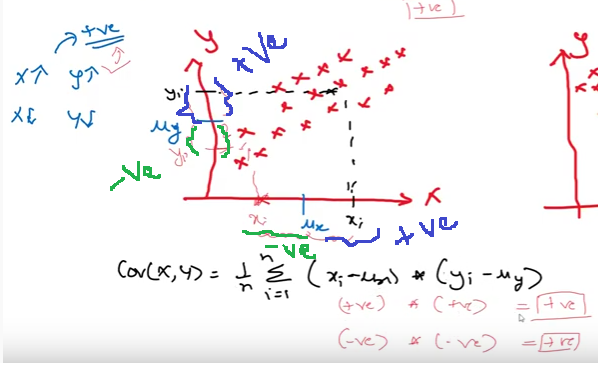
The above curve we can say linear relationship , x axis increase the y axis also increase.

And also we can say positive co-variance.

Formula ( Variance with respect to feature) ( Here mu is mean of feature)



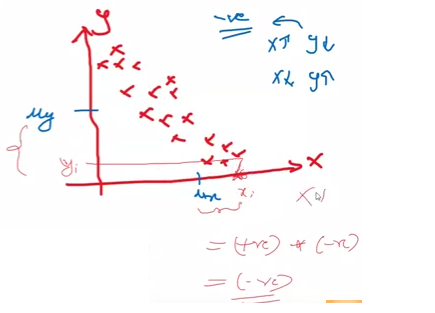


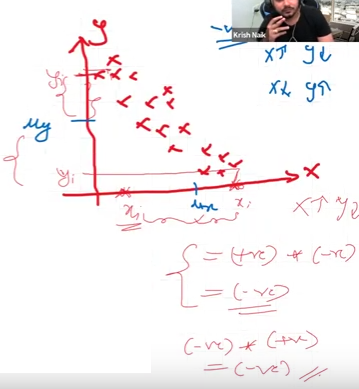


From the above graph if the new point is xi above the mean ( x increase Y increase) - both points are positive from mean – so positive covariance.

From the above graph if the new point is xi below the mean ( x increase Y increase) - both points are negative from mean ( Negative multiply by negative- psotive) – so positive covariance.

Next scenario:





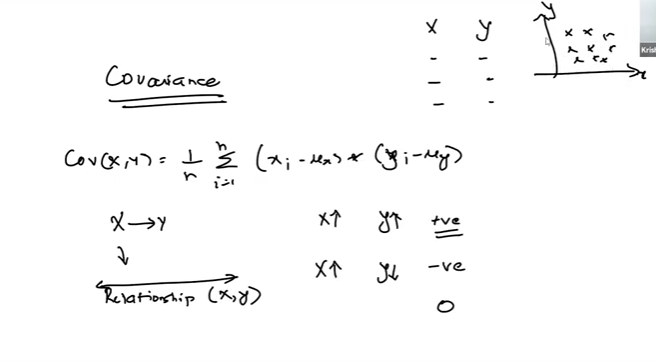
The data is scatted – Then no relationship – Zero co variance.

Day4

**(1)Covariance:**

==========

Find the relationship



Zero means there is no much relationship.

Here we will know direction whether it’s positive or negative, we can’t find magnitude

Disadvantage of covariance:

We can’t find the how much positive and how much negative.

Means can’t find strength, here it will go positive infinite to negative infinite.

**(2)Pearson correlation:**

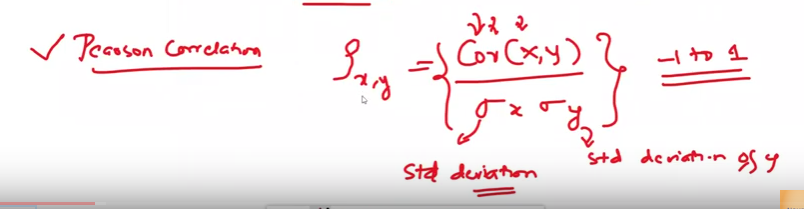
Using this we can find the type of relation & strength,

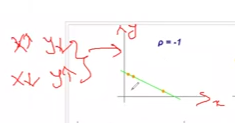
The range between {-1 to 1}

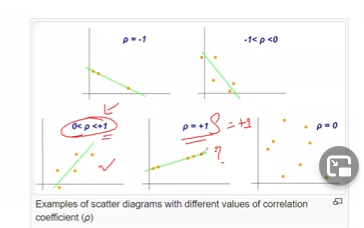
The value is near to 0.9 means highly positive co-relation.

The value is near to - 0.9 means highly negative co-relation.

Formula:







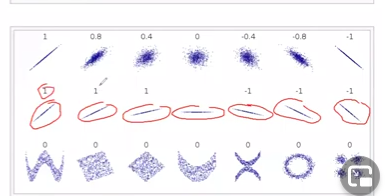
The above diagram first plot p = -1 means – best fitted line pass through all points so negatively co-related

The above diagram second plot -1<P<0 means – best fitted line pass through only few points and negatively co-related so the range will be -1 to 0

The above diagram third plot o< p + 1 means –– best fitted line pass through only few points and positively co-related so the range will be 0 to 1

The above diagram fourth plot p = +1 means – best fitted line pass through all points so positively co-related

P =0 means no co relation.



Pearson co-relation will work only with linear .The above diagram we can see even though the direction change ( second row) we can find the co-relation.

Second row third picture no – correlation because the line is horizontal. ( Constant)

**Disadvantage of Pearson co-relation:**

The above diagram in third row non linear pattern but we points are positively & negatively co-related.

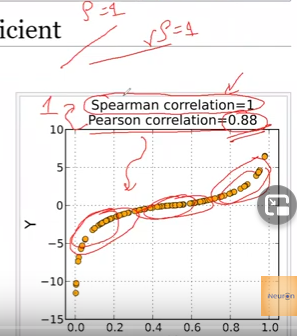
But we can’t find non linear info.

**(3)Spearman Rank co-relation:**

Correlation

Magnitude

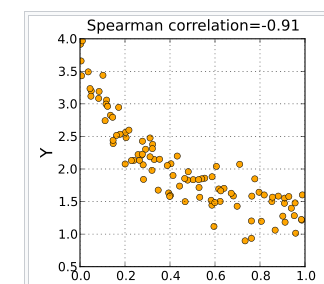
Non Linear property



The above diagram

X is increase Y also increase so the spearman correlation is =1

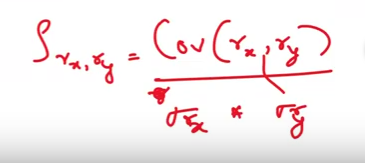
But here not all the lines are pass thorough the straight best fitted line so person co-relation is 0.88

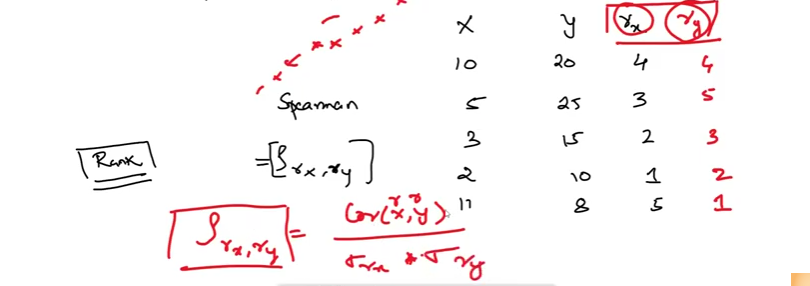


Example:

<https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php>

Formula:





**Practical:**

Statistics EDA Note book.

**Why Correlation is important?**

**Case 1:**

Assume my X feature is Age, Y feature is Weight I need to predict feature Z( Height)

I machine learning algo or feature engineering we will find the correlation b/w each in-depended feature variable. ( Example Pair plot)

Example age also 90%(0.9) correlated with Height & weight also 90%(0.9) correlated with Height

We can drop one feature to predict the height.

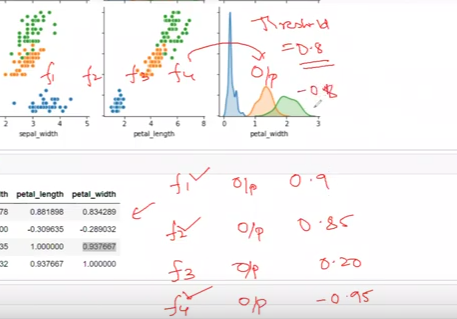
So which variable gives highest correlation we will use those feature for predication.

Assume we are adding one more feature now the age correlation is 95% and weight correlation also 95% with height. Now also we can use either one feature to predict.

Case 2:

One more scenario:

Co-relation checking with O/P feature, which one we will use predict?



We can use f1,f2,f4 –may be here we will set threshold like choose if positive correlation is 0.8 &

negative correlation is 0.8 .

others co-relation value are dropped.

Here negative co-relation also important.

Threshold – will set by domain expert.

Correlation work with Numeric feature

Whenever the depend variable is classification based we will not use co-relation.

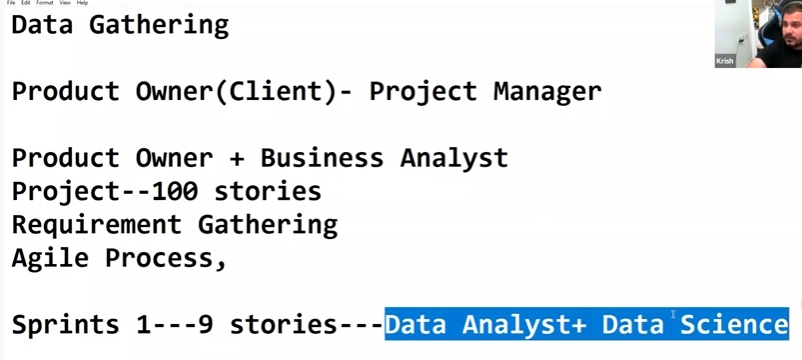
**Feature Engineering:**

Practical:

Feature Engineering – Folder Jupyter Notebook.

Data science life cycle:

1. Data Gathering



The data may be internal or external – ( may be cost involved)

1. The data may be different format.

The data may be missing values,

There is many scenarios to missing the data.

As Data eng we need to clean the data using EDA & feature engineering.

1. Missing data:

#### What are the different types of Missing Data?[¶](http://localhost:8888/notebooks/MachineLearning/iNeuron_ML/Week10/statistics_week1/Feature-Engineering-Live-sessions-master/Feature%20Engineering-%20MeanMedianDay%201.ipynb#What-are-the-different-types-of-Missing-Data?)

##### *Missing Completely at Random, MCAR:*

A variable is missing completely at random (MCAR) if the probability of being missing is the same for all the observations. When data is MCAR, there is absolutely no relationship between the data missing and any other values, observed or missing, within the dataset. In other words, those missing data points are a random subset of the data. There is nothing systematic going on that makes some data more likely to be missing than other.

Examples: The women to write the phone number – ( Generally will happen so relation ship is there – so this not MCAR)

But in Titanic data set Age & cabin info missing – Here there is no relationship. – This type of missing is MCAR

##### *Missing Data Not At Random(MNAR): Systematic missing Values*

There is absolutely some relationship between the data missing and any other values, observed or missing, within the dataset.

E.g More non survived more person cabin info missing – so relationship is there

1. *Missing at random.(MAR)*

e.g #Men---hide their salary

#Women---hide their age

**Missing value imputation:**

# 1. Mean/ Median/Mode replacement

# 2. Random Sample Imputation

# 3. Capturing NAN values with a new feature

# 4. End of Distribution imputation

# 5. Arbitrary imputation

# 6. Frequent categories imputation

DAY5:

--------------

Missing at completely random (MACR) – we will use random imputation – will give good prediction as per research.

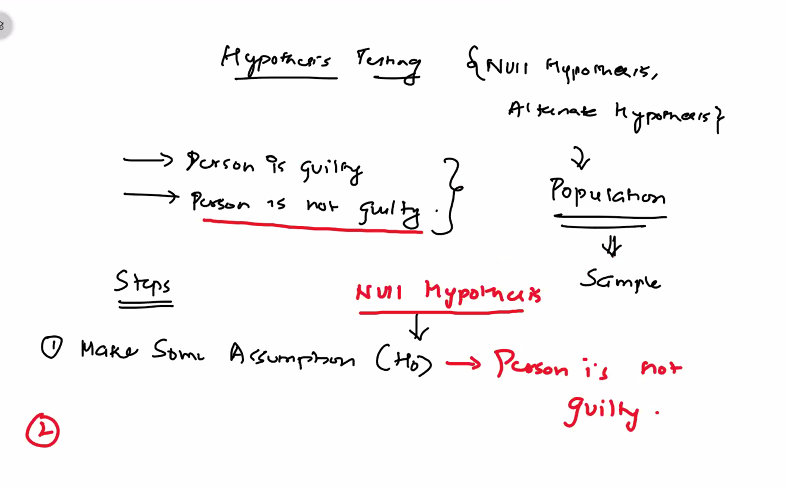
50% missing. Please drop the feature.

Hypothesis:

Hypothesis will help analysis data

Based on the sample we will derive some conclusion.

For that we will assumption something.



.

We will use data & test – if Test Pass we will say person is not guilty.

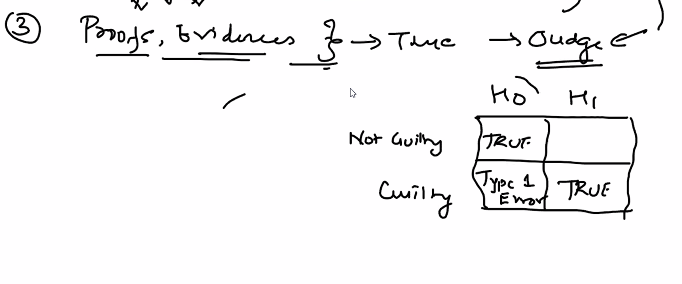
If test fail we will consider alternative hypothesis

**We can do assumption initially as person is guilty also. ( not an issue**)

What kind of data I will take?

We need more evidence to prove.

Confusion matrix



The person not done any crime, because of evidence he put as guilty.

Then it’s Type 1 Error.

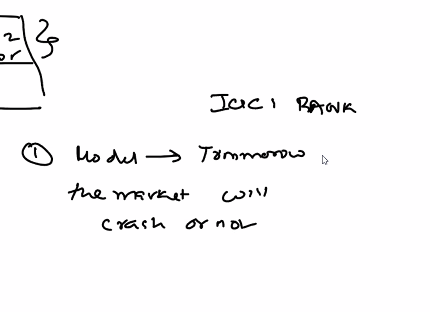
The person is done crime, But does not have any evidence so it’s shows not guilty.

This is called Type 2 error.

We need to reduce Type1 & type2 error.

Another example:

Cancer example:



Example:

**Tomorrow market will crash or not**

Ho = Market will crash

Alternate: market will not crash

2 x 2 matrix

Diagonal will be True.

HO H1

|  |  |
| --- | --- |
| True | Type 2 error. |
| Type 1 error | True |

Model predict not crash – but actually crashed - People will lose money. This is danger. Type 1 is more important.

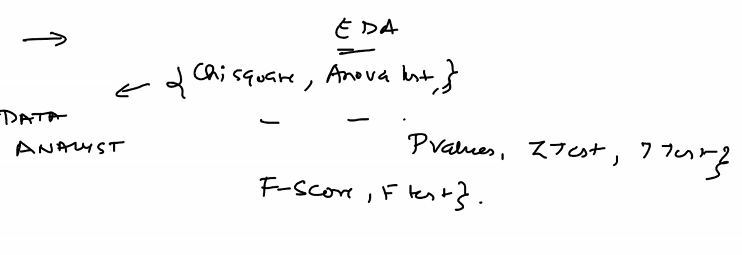
Model predict is crash – But not crashed - model predict the market will crash so people will fear & get out the money from bank.

From People perceptive Type 1 is danger, From bank perceptive Type 2 is danger.

Hope my understanding is correct?

Normally Assumption statement will be positive.

Tomorrow topics:



DAY7;

1. Hypothesis Testing jupyter notebook

ANNOVA:

* analysis of variance
* Compare mean of two or more groups - we will use Anova

Two factors we need to consider:

1. Factor
2. Level

Two types:

**One way AOVA** : It has one factor with atleast Two level ( one factor & two level ) – Levels are independent

Example: Dossage – is factor

0 mg 50 mg 100 mg person taking level

2 5 7

8 10 11

**Repeated measure of ANOVA :** one factor, atleast Two level & also levels are dependent

Dependent means if I change something will increase.

Example

Day - factor

Day night -level

Sun moon

Another example:

Feature of body - Factor

Age Height Weight - Level

50 5.5 78 Here the feature are dependent. Age increase the Height & weight increase.

30 5.2 55

**Third Example:**

Iris – factor

Petal width, sepal width - Level

Hypothesis notebook.

One Very good example:

<http://statisticslectures.com/topics/onewayanova/>

**Factorial ANOVA:** Two or more factor with 2 or more level and either dependent or independent or both

DAY8:

Confidence Interval:

When should performance – Population - Std deviation given- Z- test

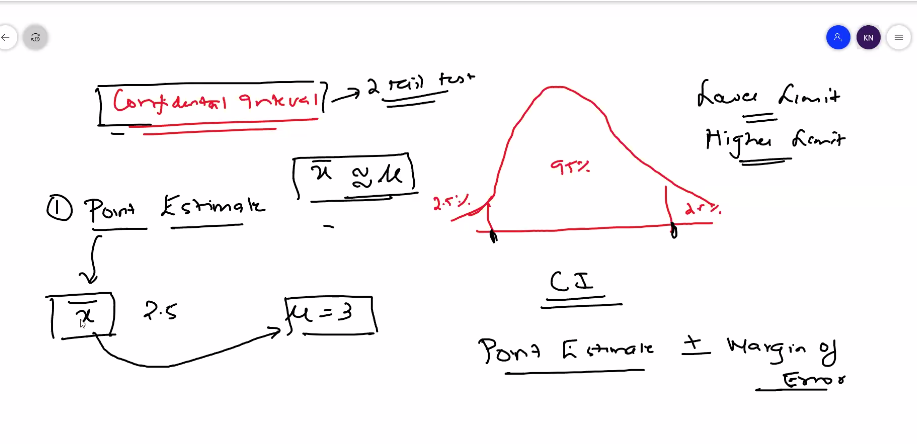
Population – Std deviation not given

T- Test sample size less than 30

Z- Test sample size greater than 30.

Confidence interval – 2 Tail Test:

Point estimate = approximately the sample mean is equal to population mean.



**Problem Statement:**

Entire class population standard deviation is sigma is 100

Sample size of n= 32 we will take & the mean is 520

Construct the 95% confidence interval about the mean?

I need to tell lower limit,higher limit using given info.

CI (population) = Point of estimate +margin of error.

Population mean not given:

**Different type of distribution:**

1. **Bernoulli distribution**

Fair coin ( Toss example)

Probability(H) = 0.5

Probability(T) = 0.5

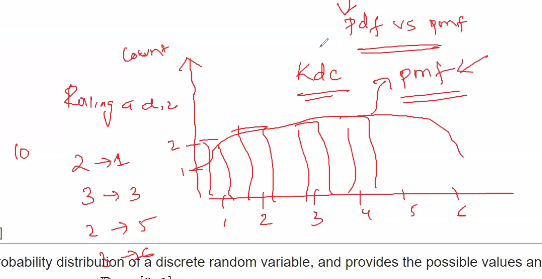
Suppose Probability is 1 Then Q = 1-P =>

**PDF & PMF:**

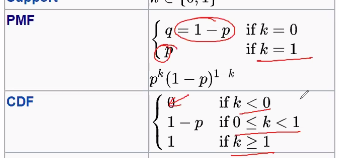
Whenever continues random variable if we plot – we will get pdf

If it’s fixed(whole number) – Discrete – if we plot – we will get pmf

Rolling die :



We will get gap b/w bin.



Outcome will be = 2

**2.Binomial distribution:**

Tossing a coin is which distribution : Bernoulli distribution:

If we do the coin toss N time then = Binomial distribution

How many time head comes = 5 like that

**3.Log Normal distribution**

When ever right skewed then it’s log Normal distribution

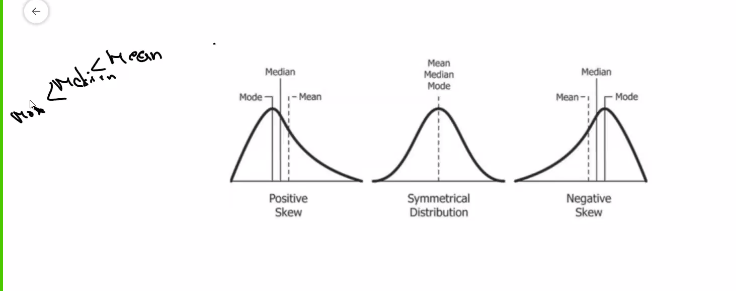
Mean>median>mode

e.g: Comments in review comment

wealth of human

Left skewed :

Death of human



Can we convert log normal to Normal distribution?

Yes, we can do.

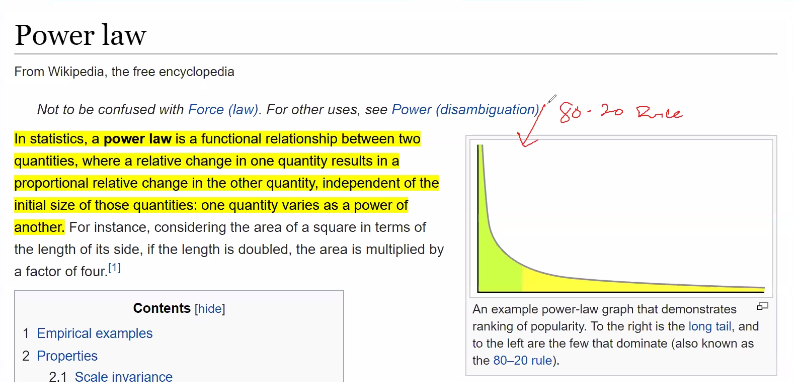
Random variable belong to log normal:

Only if log(x) belong to Gaussian distribution.

Q-Q Plot

**4.Power Law distribution(80-20 rules)-(Pareto distribution)**

The height will decided by alpha



E,g Population size of cities

Bangalore – Horizontal growth – People will move outskirt – already some people also there.

Mumbai – Vertical growth.

80% Cancer is because of toboco product.

Can we convert Power law to Normal distribution?

Yes, we can do using box cox Transform.

Feature Engineering – Normalization and standardization note book

Chi-square Test – Watch Krish You tube

**EDA - Book**

**Feature Engineering Notebook**

**Scaling**

[https://www.youtube.com/playlist?list=PLZoTAELRMXVMcRQwR5\_J8k9S7cffVFq\_U](https://us02st1.zoom.us/web_client/ehjzr5/html/externalLinkPage.html?ref=https://www.youtube.com/playlist?list=PLZoTAELRMXVMcRQwR5_J8k9S7cffVFq_U)

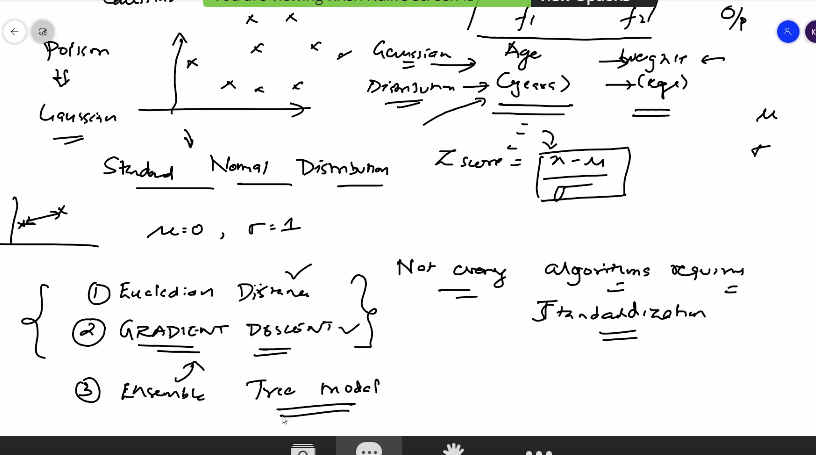
**DAY9:**

**Advanced EDA ( Automated way)**

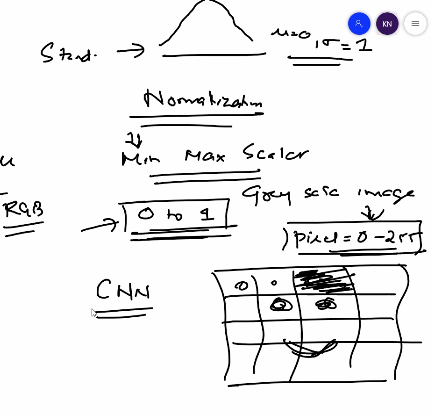
**Standardization and Normalization:**

**Standardization: ( Convert distribution)- Most of ML**

**If my distribution is Poisson or log Distribution – Convert first Normal/Gaussian distribution -then convert standardization**

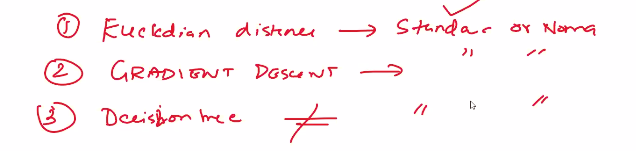


Normalization : (Min-Max) – CNN it’s important.



How to covert between 0 to 1?

X – mean(x)/(max(x)-min(x))



Decision Tree /Random Forest/Ensample will created based on feature so standardization and Normalization not required.

Feature Engineering- Normalization And Standardization- Day 5 – Jupyter notebook.

Difference b/w fit and Fit\_Transform:

Apply the algo & change the data( Standardization)

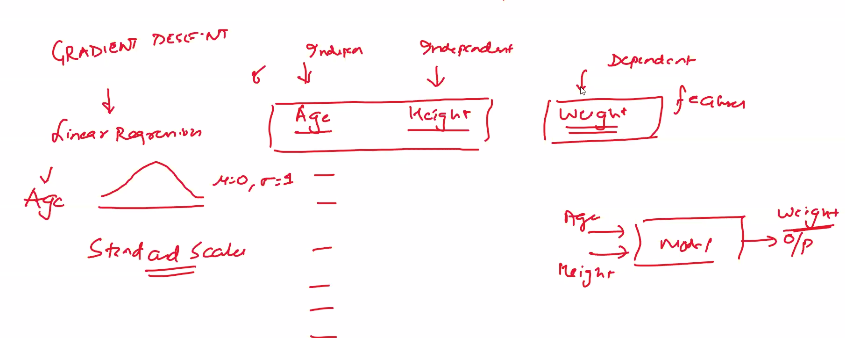
Feature Transformation:

Example: Each feature has different distribution

So first we need to convert to Gaussion/Normal distribution

Next apply standardization.

Example:



First check the independent feature has missing value

Next check whether Gaussian /Normal distribution

Linear regression – Gradient involved

Convert the data into Train \_ test Split

Now apply standard scalar in to Train data –

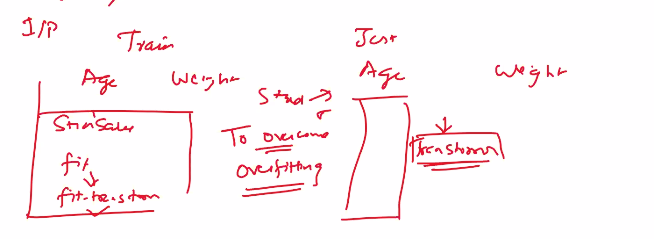
For that first Fit & Transform or do Fit – Transform Will convert data

Next pass the standard scalar data into Linear regression algo

There you have to fit the data.

For Testing data – we will only Transform ( we should not fit)

Next pass the transformed data to algo & predict.



Predict - with Test.