09/10/2021

Data Science

Analytics Vidhya

SATYAM SETH
INTERNSHALA TRAINING

1

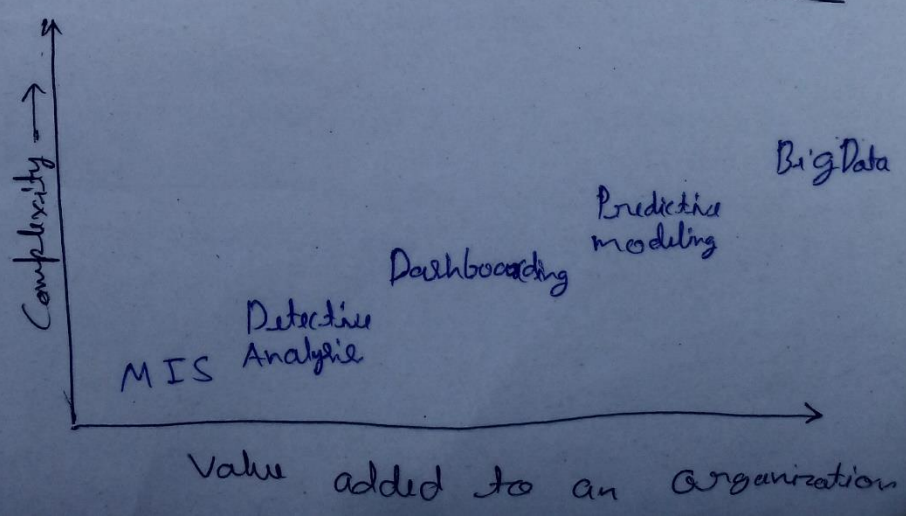
What is Analytics / data Science?

- MIS / Reporting
- Predictive Modeling
- Business Intelligence
- Forecasting
- Machine Learning
- Big Data

Data Science = Data + Science

" The field of bringing out insights from data using scientific thechnique is called Data Science."

Spectrum of Business Analytics -



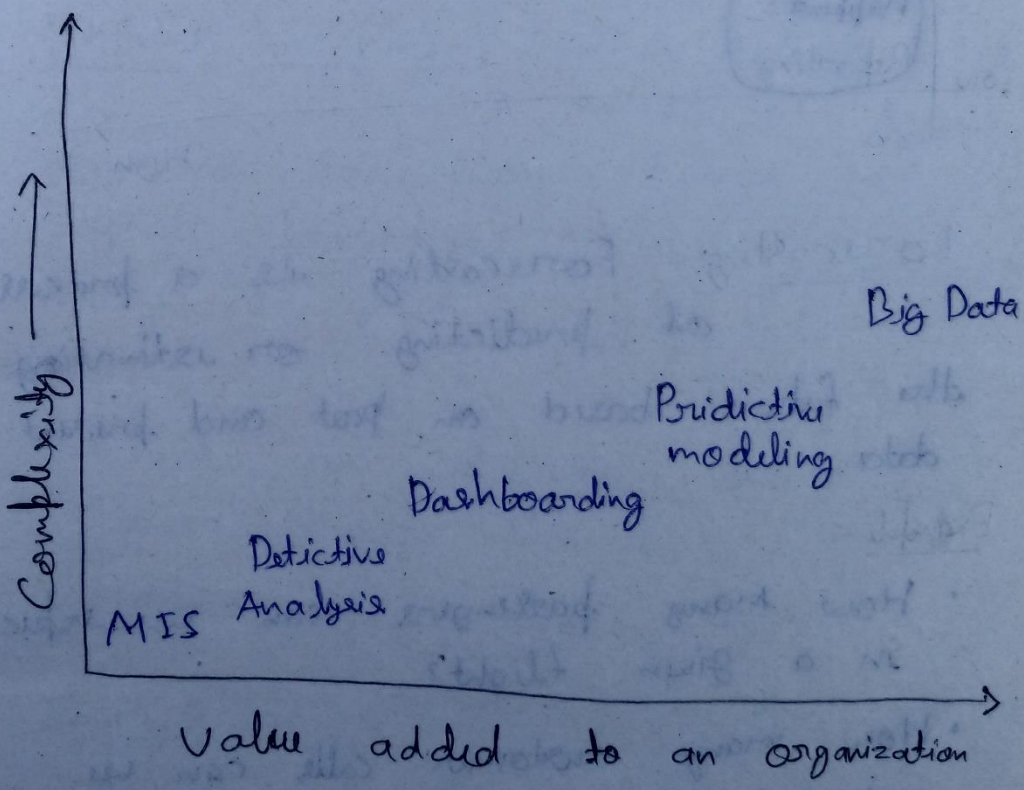
②

Q. What is Analytics / data Science?

- Data Science = Data + Science

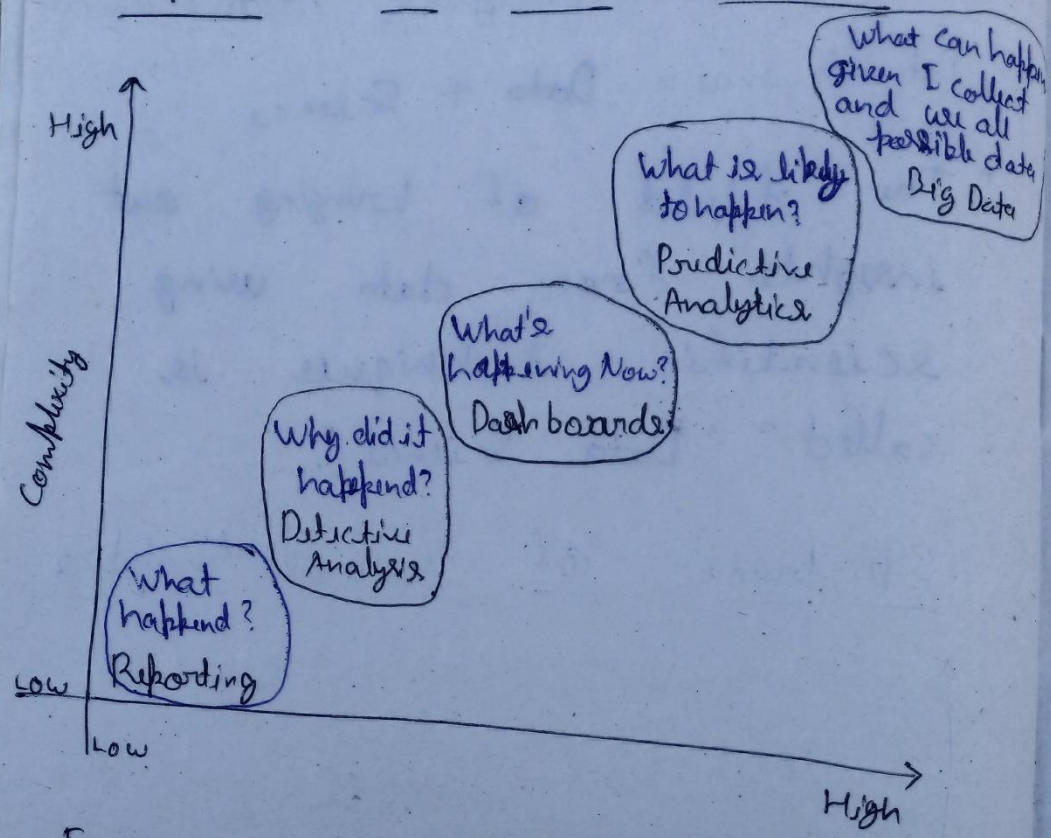
"The field of bringing out insights from data using scientific techniques is called Data Science".

Spectrum of Business Analytics -



③

Spectrum of Business Analytics -



Forecasting - Forecasting is a process of predicting or estimating the future based on past and present data.

Example -

- How many passengers can we expect in a given flight?
- How many customer calls can we expect in next hour?

Predictive Modeling - Predictive modeling, used to perform prediction more granular like "who are the customers who are likely to buy a product in next month?" and then act accordingly.

Identify right customer and taking actions accordingly.

Machine Learning - Machine Learning, is a method of teaching machine to learn things and improve predictions / behaviour based on data on their own.

Example -

- Create an algorithm which can power Google Search.
- Amazon recommendation system.

---

import Pandas -

```
import pandas as pd
df = pd.read("data.csv");
```

(5)

Pandas –

Topics to be covered

1. Introduction to Pandas
2. Why to use Pandas?
3. Reading a spreadsheet (CSV and Excel)
4. What is DataFrame and its operations
5. Indexing / Selection in DataFrame
6. Column selection, addition and deletion
7. Subsetting in DataFrame
8. Exporting DataFrame to spreadsheet (CSV and Excel)

Introduction to pandas -

Pandas is powerful python data analysis toolkit for reading, filtering, manipulating, visualizing and exporting data.

Why to use Pandas?

Wide range of functionalities such as

1. Reading different varities of data.
2. Function for filtering, searching and manipulating data.
3. Plotting data for visualization and exploration purpose.

Huge contribution and support from the community.

Pandas can help you read data from different types of files—

Format Type	Data Description	Reader
text	CSV	read_csv
text	JSON	read_json
text	HTML	read_html
text	Local clipboard	read_clipboard
binary	MS Excel	read_excel
binary	HDF5 Format	read_hdf
binary	feather Format	read_feather
binary	Msgpack	read_msgpack
binary	Stata	read_stata
binary	SAS	read_sas
binary	Python Pickle Format	read_pickle
SQL	SQL	read_sql
SQL	Google Big Query	read_gbq

7

What is DataFrame and its operations -

A DataFrame is similar to Excel workbook tabular datasheet

Excel

	A	B	C	D
1	Passen	Survived	Pclass	Name
2	1	0	1	Braund
3	2	1	3	Cumings
4	3	1	1	Allen
5	4	0	4	Saija.

DataFrame

	Passen	Survived	Pclass	Name
0	1	0	1	Braund
1	2	1	3	Cumings
2	3	1	1	Allen
3	4	0	4	Saija
4	5	1	3	Ram

What is DataFrame and its Operations.

Basic operations with DataFrame -

Operation	Command to perform
Dimension of the DataFrame "row X column"	DataFrame.shape
Access Top / Botton <n> records	DataFrame.head(n) / DatFrame.tail(n) (by defallet it gives 5 records)
Access all column name	DataFrame.columns
Access data of one column	DataFrame["column-name"]
Access data of multiple columns	DataFrame["column1", "column2", ----]]

Selecting row by their position	df.iloc[:5] (return row 0 to 4)
Selecting columns by their positions	df.iloc[:,:2] row column.

tips- for selecting specific column data -

df[df['columnname']==2]

(condition ?

[column].dtype check the datatype of the column

Descriptive ~~Stats~~ Statistics -

(9)

- **Mode -**
  - It is the number which occurs most frequently in the data series.
  - It is robust and is not generally affected much by addition of a couple of new values.

Syntax - mode_data = df ['Columnname'].mode()

Types of Variables -

1- Continuous Variables
2- Categorical Variables
   (i) Nominal Variables
   (ii) Ordinal Variables.

1- **Continuous Variables -** They have values having continuous numerical values.

Ex - Marks in Students data.

(10)

2- Categorical Variable -

(i) Nominal - Variable having categories.

Ex- Gender have only two value Male or female.

(ii) Ordinal - Variable having categories with an order.

Ex- Year of Study - Have value -
Senior > Mid Year > Junior.

Mean - often called as average

$$Mean = \frac{\Sigma x}{n}$$

where Σx is the sum of the quantities while n is the number of quantities.

Syntax - ~~mode data~~
meandata = df['columnname'].mean()

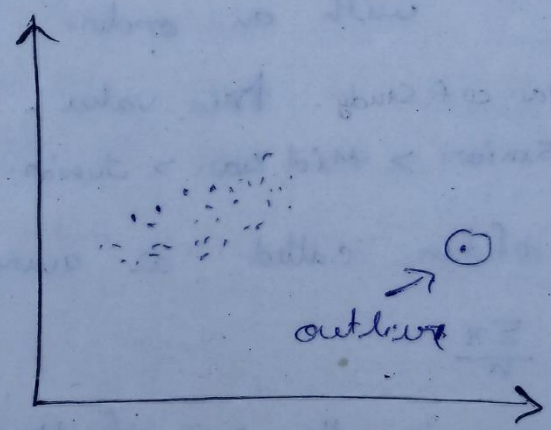
Covered -

- Why we cannot always use mode to summarize our continuous data?
- What is mean?
- How to calculate mean of the data?
- Mean is not robust

(11)

- What do we really mean by robust?
- What is outlier?
- Reasons for having outliers in data.

Outliers- Any values will fall really outside the range of the data is termed as a outlier.



Mean and Mode how affected by outliers-

without Outlier	with Outlier
4, 4, 5, 5, 5, 5, 6, 6, 6, 7, 7	4, 4, 5, 5, 5, 5, 6, 6, 6, 7, 7, 30
Mean = 5.45	Mean = 30.00
Mode = 5.00	Mode = 5.00

(12)

How to summarize data when we have outliers?

Median - It is nothing but the absolute central value of the data.

Ex- 1, 3, 3, 6, 7, 8, 9 , Median = 6

Reasons for Outliers in the data -

Types - Outliers during data collection. For example adding an extra zero by mistake.

Measurement error - Outliers in data due to measurement operator being faulty.

Intentional error - These are errors which are induced by people intentionally. For example teens might claim they've had a smaller amount of alcohol than they actually have.

Legit outliers - These are values which are not actually errors but are in the data due to legitimate reasons. For example a CEO's salary might actually be very high as compared to other employees.

(13)

Steps to Calculate median-

- So let's say we have a series of numbers [2, 4, 5, 7, 7, 4, 2, 7, 8, 9]
- Arrange the series in ascending order - So we have 2, 2, 4, 4, 5, 7, 7, 7, 8, 9
- Count the number of elements in dataset - So in this case we have 10
- If the number of elements in the data set is odd, the middle most element is the median i.e. if we have n elements, the median would be $\frac{n+1}{2}$ th element in the sorted series.
- If the number of elements in the data is even. then the median would be the average of two central elements. So the median would be the average of $\frac{n}{2}$ th and $\frac{n}{2}+1$ th elements

Ex- 2, 2, 4, 4, 5, 7, 7, 7, 8, 9, 10

Median

Median of left sub part = 25% of data is called first q

Median of right subpart = 75% of data is called third

(14)

The median lie between 1st & 3rd quartile is called 2nd q

Difference between Quantile & Quartile.

Quantile are 1/4th of quartile.

Finding Quantile in Pandas.-

```
Q1 = data['columnname'].quantile(0.25)
Q2 = data['columnname'].quantile(0.5)
Q3 = data['columnname'].quantile(0.75)
Q4 = data['columnname'].quantile(1)
```

Spread of data- Spread of the data describes how similar or varied are the set of observations.

Range-

- The difference between the smallest and largest value.
- The bigger the range the more spread out is the data.

Ex- 2, 2, 4, 4, 5, 7, 7, 7, 8, 9 , range = 9 - 2 = 7

2, 2, 4, 4, 5, 7, 7, 7, 8, 9, 1000 , range = 1000 - 2 = 998

finding range

Syntax-

```
min_data = data['Overall Marks'].max()
min_data = data['Overall Marks'].min()
range_data = max_data - min_data
```

Finding Inter Quartile range -

```
Q1 = data['Overall Marks'].quantile(0.25)
Q3 = data['Overall Marks'].quantile(0.75)
IQR = Q3 - Q1
```

Range And IQR (Difference) -

Range -	IQR -
• Susceptible to outliers	• Robust to outliers
• Range = Max - Min	• IQR = Q3 - Q1
MIN MAX 0, 12, 20, 31, 74, 61, 89, 99, 100, 9700	Q1 Q3 0, 12, 20, 31, 74 | 61, 89, 99, 100, 9700
• Range = 9700 - 0 = 9700	• IQR = 99 - 20 = 79
• Takes Entire data set into consideration.	• Takes 50% of data into consideration

Variance -

ID	Marks	Teacher feedback	Gender	Distance from mean (Marks - Mean)
X001	80	good	Male	12
X002	40	bad	Female	-28
X003	75	good	Male	7
X004	90	excellent	Female	22
X005	40	bad	Female	28
X006	69	good	Female	1
X007	72	good	Male	4
X008	34	bad	Male	-34
X009	99	excellent	Male	31
X010	79	good	Female	11

Mean = 67.8 Median = 73.5 Mode = 40 (16)

Possible methods to calculate spread -

- Take sum of all distance -

  Sum of all distance = 0.

- Take sum of squared sum of distance

  144 + 784 + 49 + 484 + 784 + 1 + 16 + 1156 + 961 + 121 = 4500

  So, in this case average of squared distance is 450

Variance - Variance is the average squared deviations from the mean.

$$\sigma^2 = \frac{\Sigma (x-\bar{x})^2}{n}$$

Steps to calculate variance -

1- Calculate Mean -

mean = data['Overall Marks'].mean()

2- Calculate the distance from mean for each element -

difference = data['Overall Marks'] - mean

(17)

3- Calculate Squared distance -

squared_difference = difference **2

4- Take average of Squared distance

variance = squared_difference.mean()

Note- We are also calculate the variance by this method

var_data = data ['Overall Marks'].var(ddof=0)

Here Standard Deviation = $\sqrt{\text{variance}}$

Syntax- Standard_diviation = variance (1/2)

or std = data['Overall Marks'].std(ddof=0)

Note- When we divide the sum of squared difference with n-1 and take a square root, it is know as sample standard diviation.

In Statistics till now -

- The measure of central tendency -
  - Mean
  - Median
  - Mode
- Outlier
- Measure of spread -
  - Range
  - IQR
  - Variance
  - Standard Deviation.

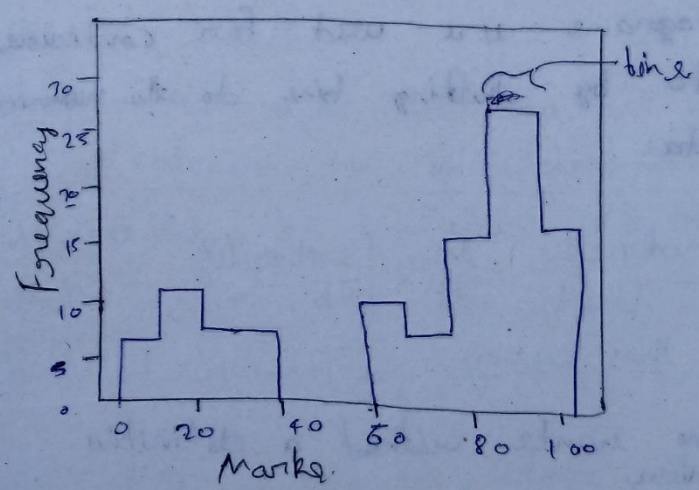
(18)

Frequency Table-

Red		Red 3
Red	→	Green 1
Blue		Blue 1
Green		
Red		

Syntax. freq_date = data['Subject'].value_counts()

Histogram- (It is use for continuous data.)



Generating histogram -

```
import matplotlib.pyplot as plt
%matplotlib inline
```
(— for draw the graph in side JupyterNotebook)

```
plt.hist (x='Column_name', data= df)
plt.show()
```

```
plt.hist (x = 'Column_name', data = df, bins = n)
plt.show()
```
No. of bins.

(19)

Properties of Histogram-

- The bins should be of equal size
- The bins should not overlap.
- The height of the bar denotes the frequency.
- We get an idea of the distribution from the histogram.
- Histograms are used for continuous data by putting bins to the numerical data.

Introduction to Probability-

In this section-

- Define events without a definitive answer.
- Define likelihood assigned to the events
- Make decisions according to the likelihood.

~~Terminology~~

Terminology - (20)

- Experiment - An experiment could be something like - whether it rains in Delhi on a daily basis or not.
- Outcome - Outcome is the result of a single trial. If it rains today, the outcome of today's trail would be "it rained"
- Event - An event is one or more outcomes of an experiment. For the experiment of whether it rains in Delhi every day the event could be "it rained" or "it didn't rain".
- Probability This is simply the likelihood of an event. So if there's a 60% chance of it raining today, the probability of raining would be 0.6

Yes! Definitely -

- High chances of rain - carry your umbrella
- 85% Chance of defaulting on a loan. Should I approve? No.

How to convert events to numbers (21)

favorable event / total event = probability

Random Variable -

- Each value of the random variable may or may not be equally likely. If I start fighting a wrestling champion our chance of winning are not equally likely.
- Sometime the random variable only take fixed or discrete value like here, the sum of value on two dice can either be 2 or 3 and not 2.5.

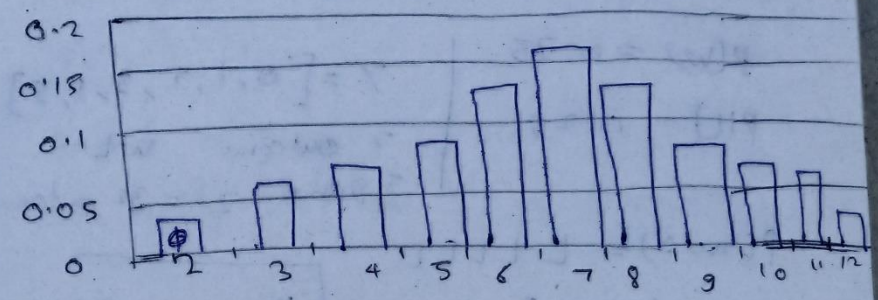
Graphical representation probabilities -

Sum on the dice	Frequency of occuring	Probability
2	1	1/36
3	2	2/36
4	3	3/36
5	4	4/36
6	5	5/36
7	6	6/36
8	5	5/36
9	4	4/36
10	3	3/36
11	2	2/36
12	1	1/36

(22)

Representing probabilities Graphically -

Probability distribution



Probability Mass Function.

Events with just two outcomes -

- Toss of a coin - head and tail
- The outcome of a fight between two people - win and lose.

Bernoulli Trials -

- An experiment which has exactly two outcomes.
- Probability distribution of the number of success in n Bernoulli trials is known as a Binomial distribution.

Ex- Suppose your favorite football team is playing against a weaker team. Chances for your team to win are 75% while for the other team are 25%. Your team wins the series if

they win more game out of 5. (22)
Figure out the probability of your team wining the series.

$P(W) = 0.75$

$P(L) = 0.25$

$X = [0, 1, 2, 3, 4, 5]$

2 outcome WL

5 games - $2^5 = 32$ outcomes

$P(x=0) = LLLLL$

$= (0.25)^5$

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

$P(x=1) =$ WLLLL
LWLLL
LLWLL
LLLWL
LLLLW

$= 5 \times (0.75 \times (0.25)^5) = 0.0146$

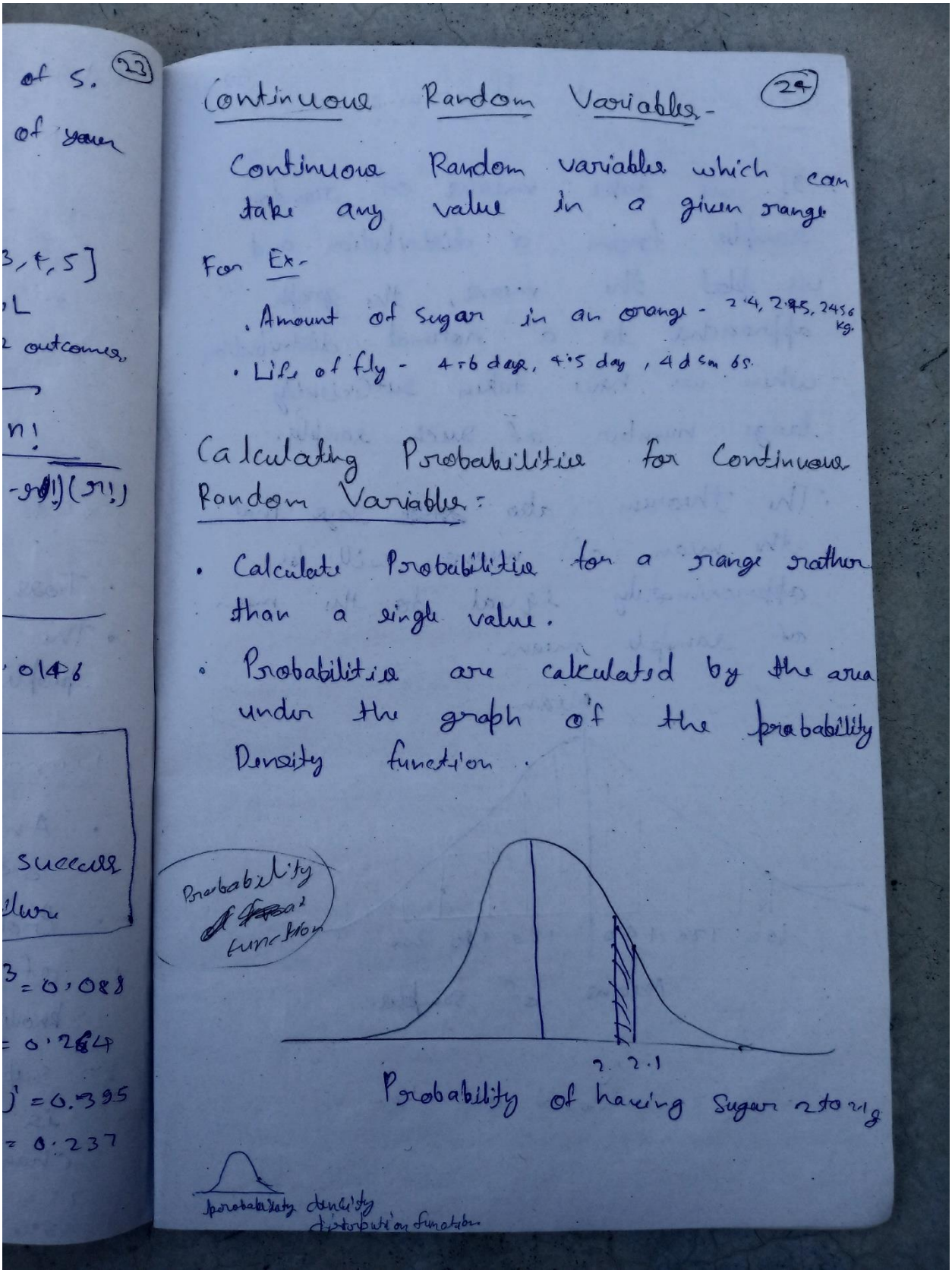
$$\boxed{\begin{array}{l} P(X=k) = {}^nC_k \times p^k \times q^{n-k} \\ \text{where } p \text{ is Probability of success} \\ q \text{ is Probability of failure} \end{array}}$$

$P(X=2) = 5C_2 \times (0.75)^2 \times (0.25)^3 = 0.088$

$P(X=3) = 5C_3 \times (0.75)^3 \times (0.25)^2 = 0.264$

$P(X=4) = 5C_4 \times (0.75)^4 \times (0.25)^1 = 0.395$

$P(X=5) = 5C_5 \times (0.75)^5 \times (0.25)^0 = 0.237$



(23)

of S.

of year

3, 4, 5]

L

outcomes

n!

(n!)

0146

success

lure

3 = 0.088

= 0.264

) = 0.395

= 0.237

(24)

Continuous Random Variables -

Continuous Random variables which can take any value in a given range.

For Ex -

- Amount of sugar in an orange - 2.4, 2.45, 2.456 kg.
- Life of fly - 4.6 days, 4.5 days, 4 days 6 hrs.

Calculating Probabilities for Continuous Random Variables :

- Calculate Probabilities for a range rather than a single value.
- Probabilities are calculated by the area under the graph of the probability Density function.

Probability function

2 2.1

Probability of having sugar 2 to 2.1 g

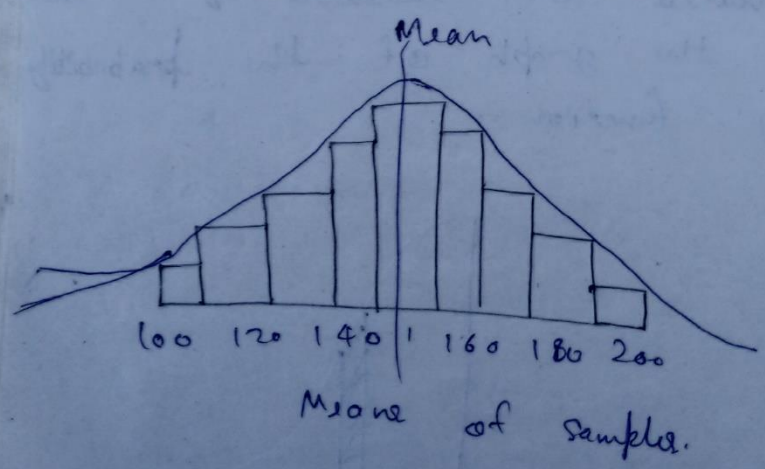
probability density distribution function

Central Limit Theorem.-

(25)

- If we take means of random samples from a distribution and we plot the means, the graph approaches to a normal distribution when we have taken sufficiently large number of such samples.

- The theorem also ~~says~~ says that the mean of means will be approximately equal to the mean of sample means.

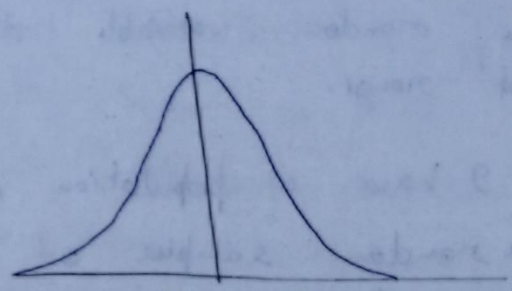


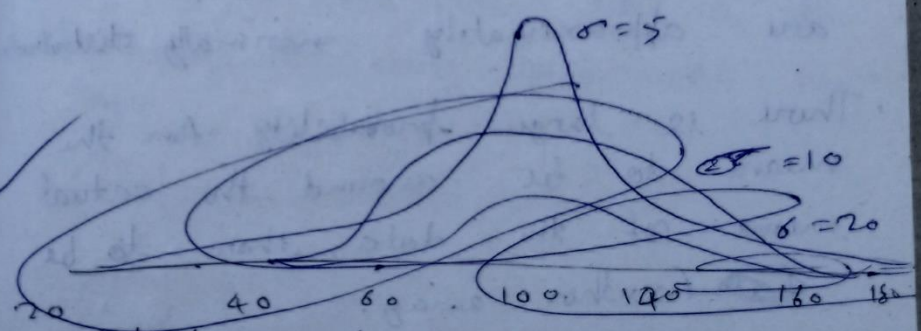
Normal distribution

Properties of the normal distribution

(26)

- The distribution is symmetric about the mean.





Standard normal distribution

- Replace Frequencies with probabilities
- Area under the curve would be equal to 1

Normal Distribution

- The equation of normal distribution is -

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\sigma^2\pi}}\left(e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right)$$

- Here μ is the mean of the data while σ is the standard deviation of the data.

Summarizing normal distribution (27)

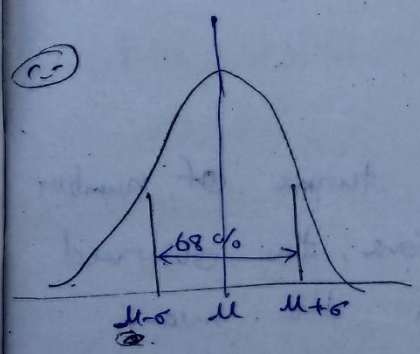
- Area under a probability density function gives the probability for the random variable to be in that range.
- If I have a population data and I take random samples of equal size from the data, the sample means are approximately normally distributed.
- There is large probability for the means to be around the actual means of the data, than to be farther away.
- Normal distributions for higher standard daviations are flatter as compared to those for lower standard deviations.

(28)

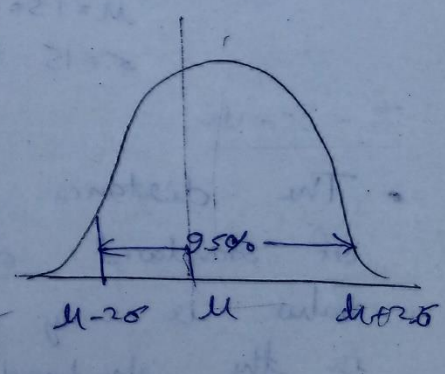
Note.

1. Total area under the standard normal curve is 1.
2. Area around 1 standard deviation around the mean is 0.68
3. Area around 2 standard deviations around the mean is 0.95
4. Area under the curve tells us the probability of data lying within that range.

(2)



(3)



Z score

$$\boxed{\mu + Z \times \sigma = \text{value}}$$

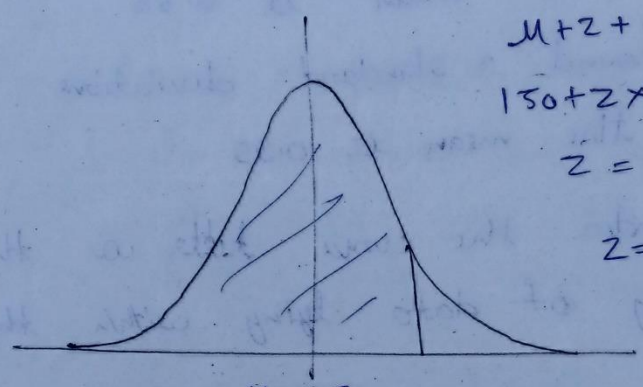
if $Z = +ve$ right side of the mean

$Z = -ve$ left side of the mean

The Problem Statement -

(29)

Let's see where 172 falls on the distribution. How many people have cholesterol less than 172.



$\mu = 150$

$\sigma = 15$

$\mu + z + \sigma = 172$

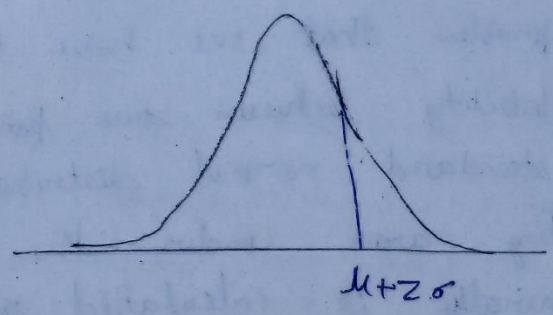
$150 + z \times 15 = 172$

$z = \frac{172 - 150}{15}$

$z = 1.47$

Z - Score -

- The distance in terms of number of standard deviations, the observed value is away from the mean, is the standard score or the Z - score.
- A positive Z score indicates that the observed value is Z standard deviations above the mean.
- Observed value $= \mu + z\sigma$ [μ is the mean and σ is the standard deviation]



Z - table -

For a particular Z score, we can look into the table to find the probability for value to fall low than that particular Z value.

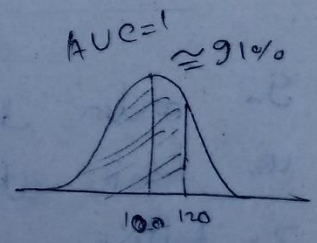
Q. The avarage IQ is 100, with a standard deviation of 15, what percentage of the population would you expect to have an IQ more than 120

$\mu + Z \times \sigma = 120$

$\mu = 100, \quad \sigma = 15$

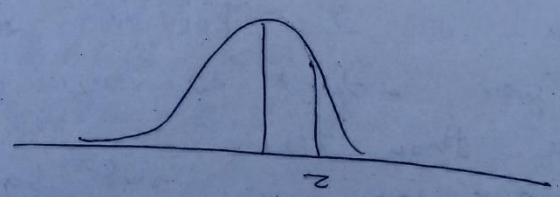
$100 + Z \times 15 = 120$

$Z = \frac{20}{15} = 1.33$



(31)

Z-Score- We often encounter a problem that we have to find probability between two points under a standard normal distribution by calculating area under the curve which generally is calculated using integral calculus. But for the sake of simplicity, the statisticians define a z-table, with the help of which the area under the standard normal distribution can be calculated. Note that using the left-t-table, which gives the area to the left of z.



In order to use the z-table, let us look at the structure of the z-table:-

along the column is the first ~~decimal~~ decimal of the z value and on the x axis is the second decimal of the z value.

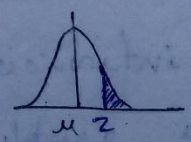
1- To use this table, we need to locate the first and the second decimal of the z value on the x axis and the y axis respectively and through their intersection, the area under the curve from the left is obtained.

2- To calculate the area to the left of z using the procedure 1 and referring to the table, let us name it k.

We know that the area of Standard Normal Distribution is 1

Therefore the area to the right of the z (shadded region) value can be calculated as- Area = 1 - k.

~~the area is~~



3- To calculate area between two points (two values of z): z1 (left) and z2 (right) respectively.

Where z1 = 0.52 and z2 = 1.47

33

In order to calculate this area, we must calculate the area to the left of both z_1 and z_2 and then subtract z_1 from z_2, what is left is the area we need.

Area = $z_2 - z_1$

~~Area = 0.92 - 0.69~~

~~Area~~

Q- For a particular z score, we can look into the table to find the probability for value to fall lee than. For z = 1.47, let's calculate the probability - 0.9292 = 92.92%.

Inferential Statistics -

- Making inferences about the population from the sample.
- Concluding whether a sample is significantly different from the population
- Hypothesis testing in general

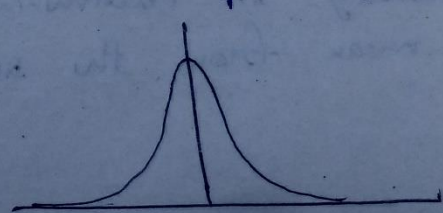
(34)

Terminology-

- **Statistic** - A single measure of some attribute of a sample
- **Population Statistic** - The statistic of the entire population in context.
- **Sample Statistic** - The statistic of a group taken from a population.
- **Standard Deviation** - It is the amount of variation in the population data. It is given by σ.

Sampling Distribution-

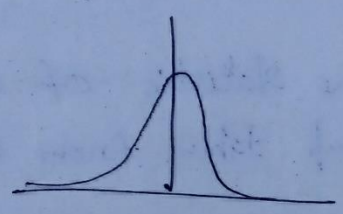
- We have the salaries of all Data Scientists in ~~sta~~ india.
- We take random sample of 200 people and calculate and plot their means.



(35)

The Central Limit Theorem-

- While plotting a sampling distribution the mean of sample means is equal to the population mean.
- The sampling distribution approaches a normal distribution.



Standard deviation $= \frac{\sigma}{\sqrt{n}}$

- Central Limit Theorem holds true irrespective of the type of distribution of the population.
- Now, we have a way to estimate the population mean by just making repeated observations of samples of a fixed size.
- Greater the sample size, greater is the accuracy in determining the population mean from the sample mean.

(35)

rubution

ie

che

done

ate the

rupiated

fised

is

the

mean.

read after CI & Margin of error goto end page (36)

Mean Estimation -

Ex Estimating the mean salary of all the data scientists in idea can be done in following ways -

- Point Estimation
- Interval Estimation

Point Estimation -

Pick a random sample from the population and calculate mean of the sample ($\bar{x}$)

And say $\bar{x} \approx \mu$

Population mean (μ) is lying between [LB, UB] with 95% confidence level.

$$L.B = \bar{x} - y \quad \text{and} \quad U.B. = \bar{x} + y$$

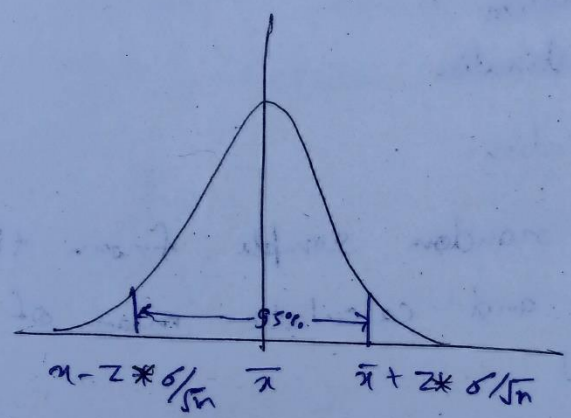
Confidence Interval and the Margin of Error -

Confidence Interval - Confidence interval in a given sample is an interval [LB, UP] with K% of confidence level.

(37)

Sample Mean $\bar{X}$ $\mu_{sample} = \mu_{population}$

$$\sigma_{sample} = \frac{\sigma_{population}}{\sqrt{n}}$$



The mathematical formula for confidence interval is-

$$C.I. = \bar{x} \pm Z_{\alpha/2}\, \sigma/\sqrt{n}$$

~~Q. Confidence~~

Q. Calculate the 95% confidence interval for a sample mean of 40 and sample standard deviation of 40 with sample size equal to 100.

$Z = 1.96$

$CI = [(40 - (1.96 \times 40/\sqrt{100})), (40 + 1.96 \times 40/\sqrt{100})]$

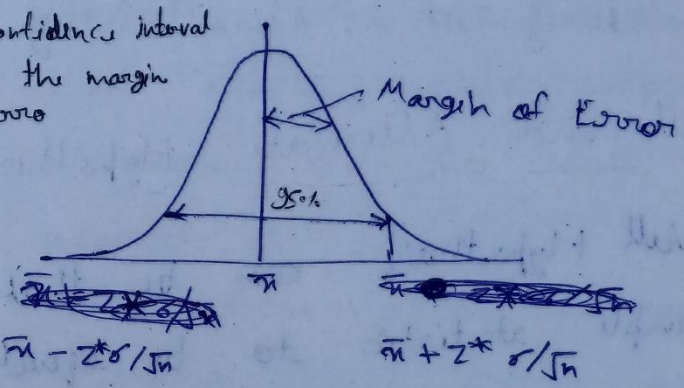
$= [32.16, 47.84]$

(38)

Margin of Errors.

The margin of error can be defined as -

- Half of confidence interval
- Here z^* is the margin of error



- Confidence Intervals can be build with different degree of confidence suitable to a user's need lik 70%, 90% etc.

Note - We assumed that Population statistic was known.

- Often not available in Real data
- Solution to this: Student's T-distribution

Go to last page

39

Hypothesis Testing – A hypothesis is nothing but a proposed explanation for a phenomenon

Always Test – Always Verify

Null and Alternate Hypothesis

- Null Hypothesis can be that the sample statistic to be equal to the population statistic or that the intervention doesn't bring any difference to the sample.
- Alternate Hypothesis basically negates the null hypothesis or says that the intervention brings a significant difference to the sample or that the sample is significantly different from the population

Hypothesis Testing.

(40)

Ex: class 8th student have a mean score of 40 marks out of 100. The principal decided that some music while teaching would improve the performance of the class. The class scored an average of 45 marks out of 100 after ~~random chance~~ taking these class. Can we be sure whether the increase in marks is a result of the music or is it just by random change?

- Check a sample statistic against a population statistic
- Check a sample against another sample to check for difference
- Check the sample with the same post intervention.

$H_0 = \mu = 40$. $H_1 = \mu <> 40$

$$z = \frac{x - \mu}{s/\sqrt{n}} = \frac{40 - 45}{26/10} = \frac{.5}{2.6} = 1.92$$

$z = 0.9726$, $1 - 0.9726 = 0.0274$

Hypothesis Testing

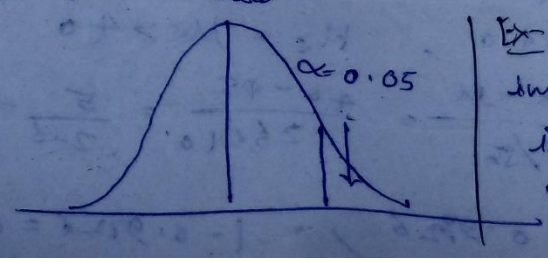
(41)

- Hypothesis Testing is done on different levels of confidence, and makes use of ~~z score~~ z-score to calculate the probability.
- We cannot accept the null hypothesis, only reject it or fail to ~~reject~~ reject it.
- As a ~~for~~ practical tip, Null hypothesis is framed on the statement we want to disprove.
  Check a sample to another sample to check for difference.
- We might errors while making conclusions using hypothesis testing.

Go to page last to 1 about

Directional Hypothesis Testing

For directional hypothesis the null hypothesis is tested in only one direction. In this case, one tail test is used.

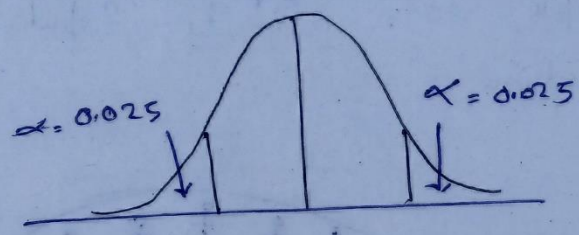


Ex- Did the performance improve on introducing music?

(41)

(42)

Non Directional Hypothesis Testing -



For non directional hypothesis, the null hypothesis is tested in both the directions. In this case, a two tail test is used.

Ex- Did the performance improve or decrease on introducing music?

Errors -

- A hypothesis is not factual. It is highly dependent on the researcher

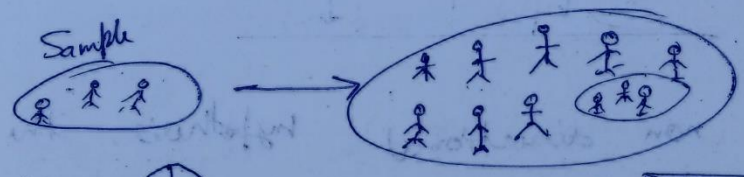
	H_0 True	H_0 False
Reject H_0	Type I Error	Correct Rejection
Fail to Reject H_0	Correct Decision	Type II Error

- Type 1 error - Rejecting the null hypothesis when it's actually true. (false positive)
- Type 2 error - Failing to reject the null hypothesis when it's actually false. (false Negative)

43

T test :-

T test is used when we do not know the population standard deviation and we sample to estimate population std. deviation.



$$S = \frac{\sqrt{\Sigma (x-\bar{x})^2}}{(n-1)}$$

Degree of freedom

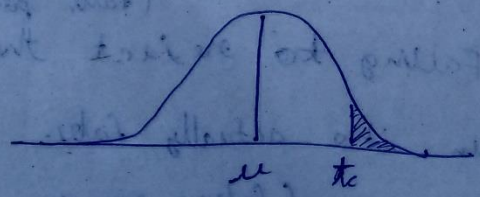
T - tests -

- Very similar to z. score.
- Use sample standard deviation to estimate population standard deviation.

T Table -

One tailed test -

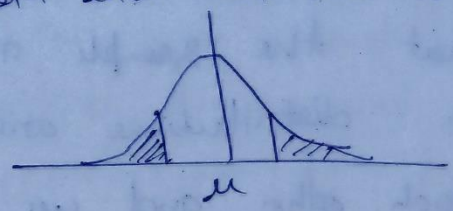
On Either end of the T-distribution



Two tailed Test -

(44)

On Both end of the T- distribution.



Degree of Freedom $= (n-1)$

sum of digit = 10

2, 3, 5, 0 → [illegible] = 3
1 6, 2, 1

T - Table -

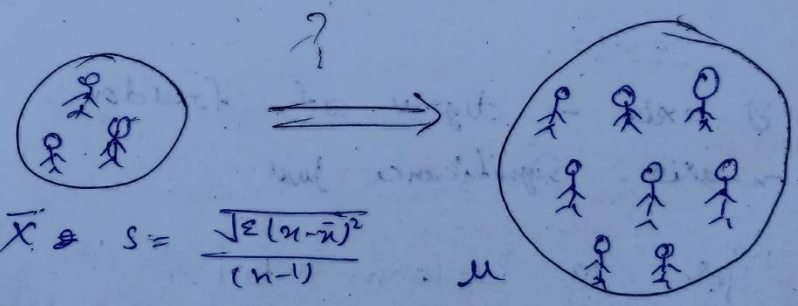
y - axis - degree of freedom
x - axis - Significance level.

Steps to Perform T-test -

1. Define null and alternate hypothesis
2. Compute T statistic
3. Get T critical from the t - table
4. Compare the computed T - statistic and the critical T - value.

Note- When t-statistic is more than t critical value - it means that the sample and population distributions are different from each other and we reject the null hypothesis.

One Sample T-test -



$\bar{x}$ & $s = \frac{\sqrt{\Sigma(x-\bar{x})^2}}{(n-1)}$

- Calculating the T Statistic

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

where- $\bar{x}$ - mean of the sample
μ - mean of the population distribution
s - Standard deviation of sample
n - number of instances.

2. Problem-

(46)

Evolution has been there for million years. the mean length of an insect in earlier time was 6.09 cms. Now we wish to find out if the length now are significantly different from earlier time. Let's say we have 29 sample with the sample mean being 6.49 cms and sample standard deviation being 0.4. we are taking significance level to be 0.05

$\mu = 6.09$ $\quad H_0 : \bar{x} = \mu$

$\bar{x} = 6.49$ $\quad H_a : \bar{x} \neq \mu$

$s = 0.4$, $d = 28$

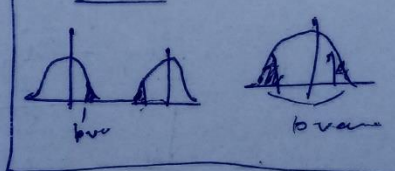
$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

$$t = \frac{6.49 - 6.09}{0.4 / \sqrt{29}} = 5.46$$

using T table- $t_c = 2.048$

$t > t_c$

P value-



Note- H_0 - Null Hypothesis

H_a - Alternate Hypothesis

Using T table Degree of freedom & significant level.

Perform one sample T test using Pandas

```
# importing libraries-
import pandas as pd
import scipy.stats as stats
from scipy.stats import ttest_1samp
#Loading data
data = pd.read_csv("one sample.csv")
data = data['Insect Length']
t_statistics, p_value = ttest_1samp(data, 6.09)
```

mean

Two sample T tests-

- $t = \frac{\text{difference}}{\text{standard error}}$

- Difference is the difference in their mean and the standard error would be the combine standard error of the two sample

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}}}$$

- Degree of freedom = $N1 + N2 - 2$

(48)

You have two choice to go to eat One is Hauz Khas and the other is Defence colony. You don't want to spend a lot of money, so you decide to figure out if there is a significant difference in the meal price of these places.

Mean Hauz khas = 178.89, Mean Def Col = 222.86

n Hauz = 18, n Def Col = 14

sd Hauz = 52.9, Sd Def Col = 43.58

use alpha 0.05

$$t = \frac{222.86 - 178.89}{\sqrt{\frac{(52.9)^2}{18} + \frac{(45.58)^2}{14}}} = 2.58$$

$df = 18 + 14 - 2 = 30$

$t_c = 2.042$

$t > t_c$

So we can reject null hypothesis

Note- Two sample t-test (49)

used ~~in two~~; we take 2 independent samples and compare the difference between them.

Infer - 2 sample t test via Pandas

```
# importing libraries
import pandas as pd
import scipy.stats as stats
from scipy.stats import ttest_ind
# Loading datasets
data = pd.read_csv('Datafor.csv').
# Calculating t-statistic and p value using 2 sample t test.
tstatistic, _ = ttest_ind(data['col1'], data['col2'])
```

(50)

Paired T test -

- We check if the same sample has changed it's behavior after an intervention, or behave significantly in two conditions.

- Null Hypothesis would be: That the difference between two sample means to 0 ie $\bar{x}_1 - \bar{x}_2 = 0$
- Alternate Hypothesis will be: That there is a significant difference between the two ie $\bar{x}_1 - \bar{x}_2 <> 0$

Formula for Paired T test -

- $$t = \frac{\bar{d}}{S_d / \sqrt{n}}$$

- where, d(bar) = mean of the case wise difference between before and after, S_d is the standard deviation of the difference and n is the sample size.

(51)

Perfor Paired T test-

```
# importing libraries
import pandas as pd
import scipy.stats as stats
from scipy.stats import ttest_rel.

# Loading data
data = pd.read_csv("Data.csv")

# calculating t and p-value using
scipy library
t_statistic, _ = stats.ttest_rel(data['col1'], data['col2'])
```

Note: In paired t-test, we compare sample before and after intervention. So, null hypothesis will be that the mean of sample differs significantly from the mean of the sample after intervention.

(51)

(52)

Chi Squared Tests.

Ex. There is a training institute called "winsit", which claims that 41 out of 100 students in it's class qualify for premier engineering institutes. Normally each year only 33% able to make it into premier engineering institute. So should you enroll yourself in "winsit"?

- Expectation : that 33 out of 100 qualify for premier institute.
- Observation : 41 out of 100 qualified.

Frequency Table.

	Successful	Unsuccessful
Observed	41	59
Expected	33	67

Null Hypothesis - No sig. diff. between observed and expected.

Alternate Hypothesis - Significant difference

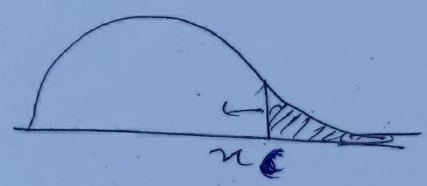
formula of Chi Square Test -

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

Here O_i is the observation frequency while E_i is the expected frequency

53

Chi Square Distribution -



If chi Square value exceed χ_c the reject null hypothesis

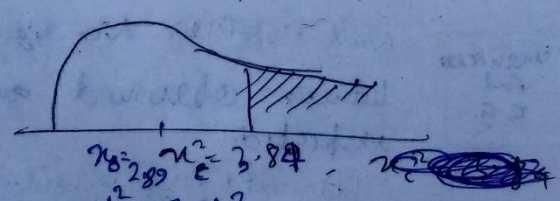
$$\chi_c^2 = \Sigma \frac{(41-33)^2}{33} + \frac{(59-67)}{67}$$

$$\chi_c^2 = 2.89$$

D.F = no. of categories − 1

D.F. = 2 − 1 = 1

(Success full & Unsuccessful)



$\chi_0^2 < \chi_0^2$ the fail to reject null hypothesis

Note - T - distribution is more prone to error than normal distribution

Chi square table - use degree of freedom & α values

Perform Chi Square using Pandas - (54)

```
# importing libraries.
import pandas as pd
import numpy as np
import scipy.stats as stats.
from scipy.stats import chisquare.

data = pd.read_csv("chi_quarucer")

# Calculating t-statistic and p-value of
  the chi square test.

t_statistic, p_value = chisquare(f_obs =
        data['col1'], f_exp = data['col2'])
```

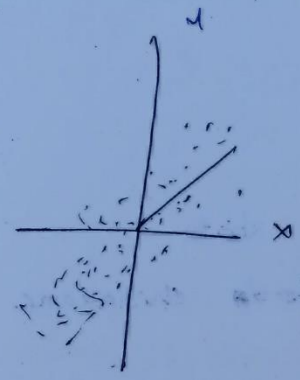
Correlation -

Example -

- More height more weight
- More hours into study more marks
- More the age of the car lower the value.

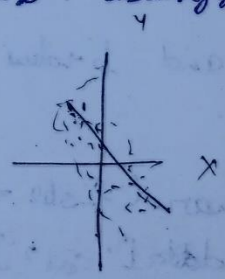
55

Scatter Plots

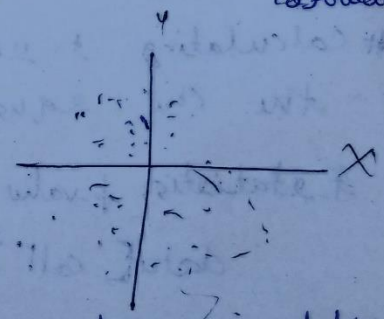


x & y strongly correlated

x & y weakly correlated



High correlation



Less correlation

Correlation.

- It is used to determine the relationship between two variables.
- It is denoted by r
- The value range from -1 to 1. Here 0 means no correlation.

(55)

$$R = \frac{Cov(X,Y)}{S_X, S_Y}$$

(56)

- The covariance shows how much of ther variable vary with each other while the standard deviation shows how much ther variable vary apart from each other.

- $Cov(X,Y) = \frac{\Sigma (X-\mu)(Y-\nu)}{n-1}$

Calculating Covariance and Correlation –

Q. Calculate covariance for the following data set : x : 2.1, 2.5, 3.6, 4.0 (mean=3.1), y : 8, 10, 12, 14, (mean=11).

Steps.

- Calculate the standard deviation for x and y.

$$\sigma_x = 0.896, \quad \sigma_y = 2.582.$$

- Divide covariance by multiplication of standard deviation of x and y

$$Cov = (2.1 - 3.1)(8 - 11) + (2.5 - 3.1)(10 - 11) + (3.6 - 3.1)(12 - 11) + (4.0 - 3.1)(14 - 11)$$

$$= 2.267.$$

$$R = \frac{2.267}{(0.896 \times 2.582)} = 0.979$$

Properties of the correlation coefficient

- If the data falls perfectly in the positive direction then the value of $r = 1$ and if the data falls in the negative direction, $r = -1$
- R lies within -1 and +1
- Data that is not correlated at all has r value of 0
- R-square is known as the coefficient of determination and can be interpreted in the terms of %

Note:- Correlation between variable tells us that how correlated they are and it also tells the direction of correlation. i.e positive or negative correlation.

find Correlation using Python (58)

```
import pandas as pd
import numpy as np

data = pd.read_csv('file.csv')

# calculate correlation between Item_weight and
  item_MRP.

data[['Item_weight', 'Item_MRP']].corr()

# calculate correlation between Item_MRP
  and Item_Outlet_Sales.

data[['Item_MRP', 'Item_Outlet_Sales']].corr()

# Import plotting libraries.
import matplotlib.pyplot as plt
% matplotlib inline
import seaborn as sns

data.corr()

# plotting correlation between different features
cor = data.corr()
plt.figure(figsize=(10,10))
sns.heatmap(cor)
```

Module 4

Predictive Modeling -

(59)

- Making Use of past Data and other attributes.
- Predict the future using this data.

Type of Predictive Models -

1. Supervised Learning
2. Un-Supervised Learning

1- Supervised Learning -

(a) Regression - Continuous target value.

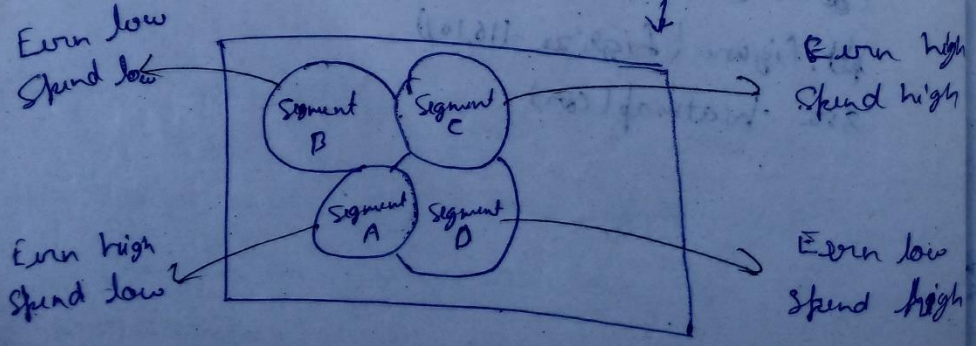
Ex - House Price Prediction

(b) Classification Learning - Discrete target value

Ex Cancer Prediction - 0 or 1

2- Un-Supervised Learning - Clustering.

Ex- Document Clustering, Market Segmentation



(60)

Stages of Predictive Model

We can broadly divide the model building life cycle in six stages.

1- Problem Definition.
2- Hypothesis Generation
3- Data Extraction / Collection
4- Data Exploration and Transformation
5- Predictive Modeling
6- Model Deployement / Implementation.

Ex:

1- Problem Definition-

Bad Problem Statement - Want to improve the profitability of credit card customers.

- Want to increase the APR of credit cards.
- Want to deploy different APR for different segment of customers.
- Want to identify the customer segments having lower default rate.
- Want to have different APR and other benefits for different customers segments (On expected default rate)

to maximize profit.

(61)

Good Problem Statement :- Want to predict the default rate of customer.

2- Hypothesis generation

"List down all possible variables which might influence problem objective."

Ex: What are factors, those can impact the customer will default or not?

- Income - Higher Income means higher chance of financial stability that may lead to lower default rate.
- Job Type - If job is more stable that can also lead to lower default rate. Exple: Front End Sales job is less stable than data science job.
- Credit history - Your Previous good repayment behaviour suggest that you know how to use credit product and that leads to lower defalt rate.

(62)

. Education: Higher education is directly proportionate to awareness about the use of credit products and that may impact the default rate also

Common Questions -

6. Should Hypothesis Generation be done before or after looking at the data?

Hypothesis generation to be done before looking at the Data.

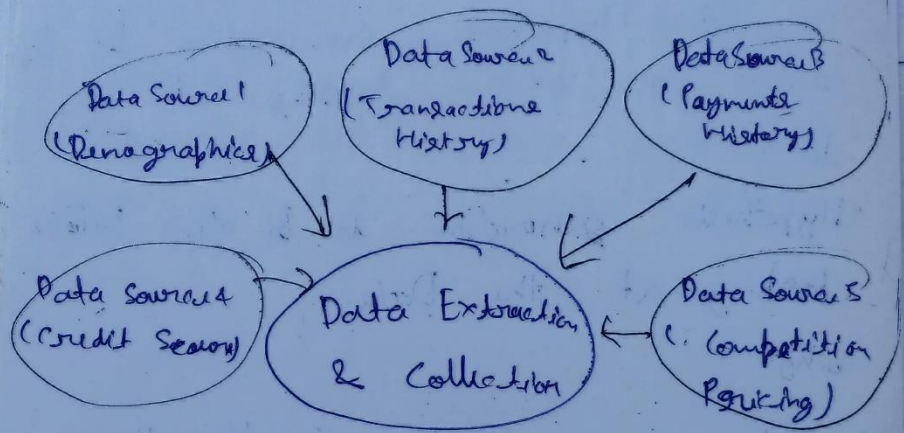
why?

1- Let's you think of all the factors which might affect the problem without being biased.

2- Stops time wastage in analyzing all available data.

3- Data Extraction / Collection-

Extract / collect data from different source and combine those for exploration and model building



4- Data Exploration-

Good Analyst vs Bad Analyst.

Good Analyst - Knows his / her data well.

Bad Analyst - Relies on tools and libraries

Steps for Data Exploration-

Reading the Data → Univariate Analysis → Missing Value Treatment → Variable Transformation

Variable Identification

Bi-variate Analysis

Outlier Treatment

(i) Reading the Data - (64)

Two common types of file formats

1- CSV format - .csv

2- Excel format - .xlsx/.xls

Note Simply use Pandas

(ii) Variable Identification.

Q. What is Variable Identification?

1- Independent and Dependent variables

2- Continuous and Categorical variable.

Q. Why do we perform variable identification?

1- Technique like Supervised Learning require identification of Dependent Variable.

2- Different data processing technique for Categorical and Continuous data

Variable Transformation

(65)

Difference between Dependent and Independent Variables.

1- Dependent Variable - The variable we are trying to predict.
Ex- Survived variable in titanic problem

2- Independent Variable - The variables which help in predicting the Dependent Variable.
Ex- Sex, Fare etc. in titanic problem

Note- Can only be identified from the problem Statement.

Difference between Categorical and Continuous Variables.

1- Categorical Variable - Discrete in nature. Ex- Survived, Sex

2- Continuous Variable - Can have infinite number of possible values.
Ex- Fare, Age

(65)

Note - In Pandas - (66)

- Categorical Variables - Stored as object.
- Continuous Variables - Stored as int or float.

(iii) Univariate Analysis -

Q. What is Univariate Analysis?

1- Explore one variable at a time
2- Summarize the variable
3. Make sense out of that summary to discover insights, anomalies etc.

Ex age of human ~~[illegible]~~ under '60 not 400

(note last page to above 3)

for Continuous Variable -

- Central Tendency and dispersion - mean, median, standard deviation
- Distribution of variable - Symmetric / Right Skewed / Left Skewed.
- Presence of Missing Values.
- Presence of Outliers

(67)

2 methods -

1- Tabular methods- Mean, Median, Standard Deviation and missing values.

function describe()

2- Graphical Methods- Distribution of variable, Presence of Outliers.

function 1- Histrogram- Distribution of a variable.

→ pd.DataFrame.hist()

2- Boxplot - Detection of Outliers

→ pd.DataFrame.boxplot()

Ex- df['Age'].plot.hist()

df['Age'].plot.

~~Stage of Data Exploration~~

for Catagorical Variable-

Q. What do we look for in Univariate Analysis-

- Count - absolute frequency of each category in a catagorical variable.
- Count % - proportion of different categories in a categorical variable expressed as %

for this 2 methods are available

(68)

1. Tabular methods - Frequency tables
2. Graphical Methods - Bar plots.

1. Tabular method - frequency table.

Ex. male - 577 } Count
female - 314

male . 0.647587 } Count %
female - 0.352413

→ df['Sex'].value_counts()

% → df['Sex'].value_counts()/len(df['Sex'])

2. Graphical Methods - Bar Plot.

Count Count %.

→ df['Sex'].value_counts().plot.bar()

Count % -

df['column'].value_counts()/len(df['column']).

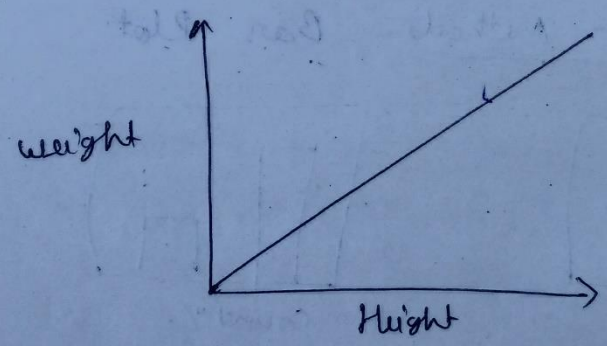
(iv) Bivariate Analysis

69

Q. What is Bivariate Analysis?

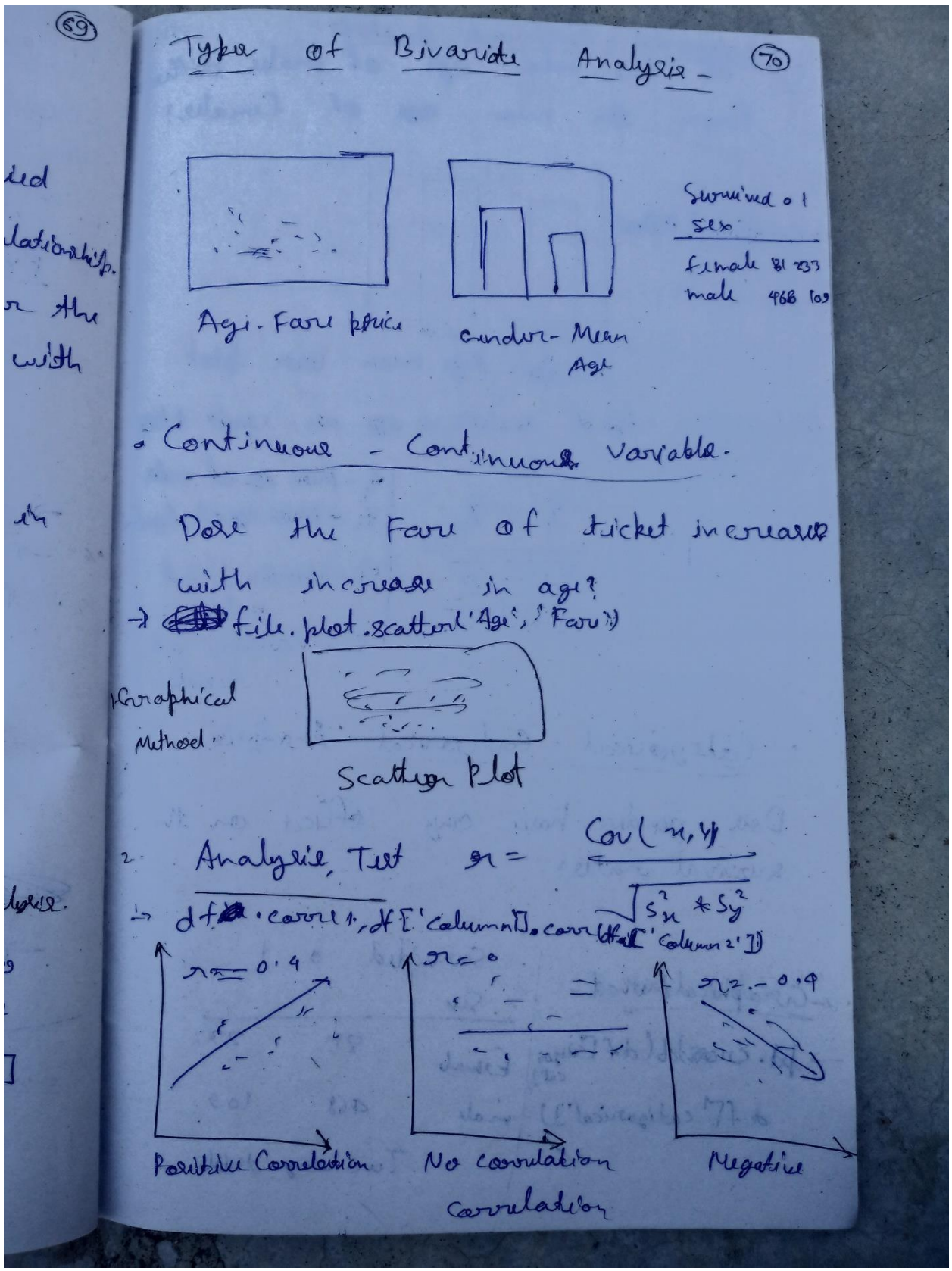
- When two variables are studied together for their empirical relationship.
- When you want to see whether the two variables are associated with each other.

For Ex:- Increase in height is associated with increase in weight and vice versa.



Q. Why do we need Bivariate Analysis.

- It helps in prediction: [When two variables are associated one may be used to infer the other.]
- It helps in detecting anomalies.



Type of Bivariate Analysis - (70)

Agi - Fare price

Gender - Mean Age

Survived o 1
Sex
female 81 233
male 468 109

- Continuous - Continuous Variable.

Does the Fare of ticket increase with increase in age?

→ file.plot.scatter('Age', 'Fare')

Graphical Method.

Scatter Plot

2. Analysis Test $r = \frac{Cov(x,y)}{\sqrt{S_x^2 * S_y^2}}$

→ df.corr(), df['Column1'].corr(df['Column2'])

$r = 0.4$ — Positive Correlation

$r = 0$ — No correlation

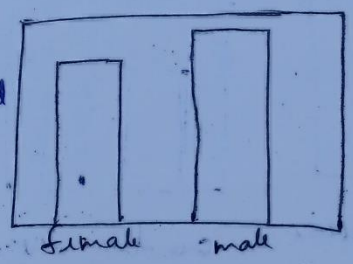
$r = -0.4$ — Negative correlation

(71)

• Categorical - Continuous Analysis -

Ex- Is the mean age of male differ from the mean age of female?

* Graphical Method



Gender - Age mean bar plot

* For this type of analysis we use - 2 sample t test.

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

$\bar{y}_1$ - Mean age of male,
$\bar{y}_2$ - Mean age of female
s_1^2 - Variance of male
s_2^2 - Variance of female

• Categorical - Categorical Analysis.

Does gender have any effect on the survival rates?

1- Graphical Method -

→ pd.crosstab(df['Categorical'], df['categorical'])

Survided	0	1
Sex		
Female	81	233
male	468	109

Two way table

(72)

Analysis Test.

$$\chi^2 = \frac{\sum (O - E)^2}{E}$$

chisquare test

χ^2 = the test statistic Σ = the sum of

O = observed frequence. E Expected frequence

→ fill.corr() for Continuous - Continuous

→ fill['Age'].corr(fill['Fare'])

See the Jupy

Categorical - Continuous Bivariate Analysis

→ file.groupby('sex')['Age'].mean().plot.bar()

into Jupyter notebook

Note- for Categorical - Continuous - Analysis first we group categorical & continuous variable-

→ df.groupby('Categorical')['continuous'].mean().plot.bar()

(V) Missing Value Treatment -

Reasons for Missing Value -

- Non-Response - Example when you collect data on people's income and many choose not to answer
- Error in Data Collection
- Error in Reading Data.

Types of Missing Value -

1 - Missing Completely at Random (MCAR)
2 - Missing at Random (MAR)
3 - Missing not at Random (MNAR)

1 - Missing Completely at Random (MCAR) -

Missing Value have no relation to -

- The variable in which missing variable exist.
- Other Variable in the data set.

Complete data

Age	IQ
10	77
15	114
25	98
35	72
40	119

Missing होने का किसी भी value का कोई effect नहीं है

Incomplete data

Age	IQ
10	
15	114
25	
35	72
40	

(74)

2- Missing at Random (MAR) -

Missing values have no relation to -

. The variable in which missing value exist.

Missing values have relation to -

. The variable other than the variable in which missing values exists.

Complete Data

Age	IQ
10	77
15	114
25	98
35	72
40	119
45	76
55	98

missing data for Age is greater than 35.

Incomplete Data

Age	IQ
10	
15	
25	
35	72
40	119
45	76
55	98

3- Missing not at Random (MNAR) -

Missing values have relation to -

. The variable in which missing value exist.

Complete data

Age	IQ
10	77
15	114
25	98
35	72
40	119
45	76
55	98
60	114
65	76

Incomplete data

Age	MNAR
10	
15	114
25	98
35	
40	119
45	
55	98
60	114
65	

(75)

· Identify Missing Value -

describe()

Continuous Variable

Isnull()

All Variable.

(if missing give true otherwise false)

· Different methode to deal with missing value -

1- Imputation

2- Deletion

→ df.dropna() { drop all the row which have any missing value }

→ df.dropna(how="all") { drop all row which have all value are missing }

1- Imputation -

- Continuous variable
  - Mean
  - Median
  - Regression model.

(Go to last page above 4)

- Categorical Variable.
  - Mode
  - Classification model.

2- Deletion -

· Row wise deletion

· Column wise deletion

-drop all the column which have any missing value -

→ df.dropna(axis=1)

-drop all columns which have all value are missing

→ df.dropna(axis=1, how=all)

(vi) Outlier Treatment - (76)

Reasons for Outliers -

1. Data Entry Errors - Entering salary of a professional as $1000000 instead of $ 100000

2. Measurment Errors - Measuring some particular data in km instead of m.

3. Processing Errors.

4. Change in the underlying Population.

Type of Outliers -

- Univariate Outliers. (only one variable)
- Bivariate Outliers. (Two variable)

Identifying Outliers -

1. Graphical Methods -
   - Boxplot, (for Univariate Outliers)
   - Scatter plot (for Bivariate Outliers)

77

2. Formula Methods

$< Q1 - 1.5 * IQR$ or $> Q3 + 1.5 * IQR$

where - $IQR = Q3 - Q1$

$Q3$ = Value at 3rd quartile

$Q1$ = Value at 2nd quartile.

↳ to last page above 5

Treating Outliers

1- Deleting Observations

2- Transforming and Binning Values.

3- Imputing Outliers like Missing Values

4- Treat them as separately.

(VII) Variable transformation

Variable transformation is the process by which -

1- We replace a variable with some function of that variable.

Ex- replacing a variable x with its logarithm

(78)

2- We change distribution or relationship of a variable with others.

Q Why do we use Variable Transformation?

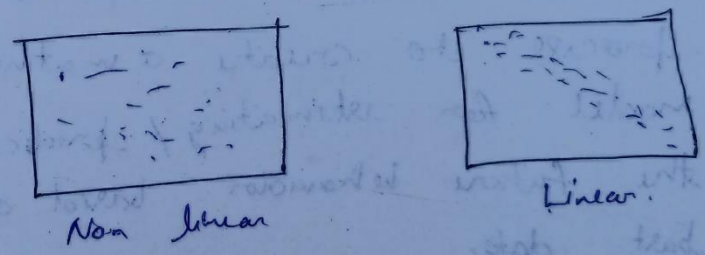
Variable Transformation is used to -

1. Change the scale of a variable.

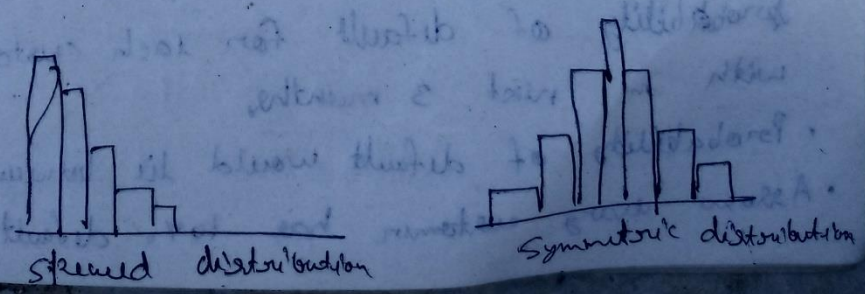
Ex- If 10 variable are measured in km and 1 in miles.

~~Variable~~

2- Transforming non-linear relationship into linear relationship.



3- Create symmetric distributions from skewed distributions



Methods - of variable Transformation - (79)

1- Logarithm - Taking log of the variable reduces right skewedness of the variable. -> np.log(df['col'']).plot.hist()

2- Square root - Used for right skewed variable with positive value only.
-> np.sqrt(df['Age']].plot.hist()

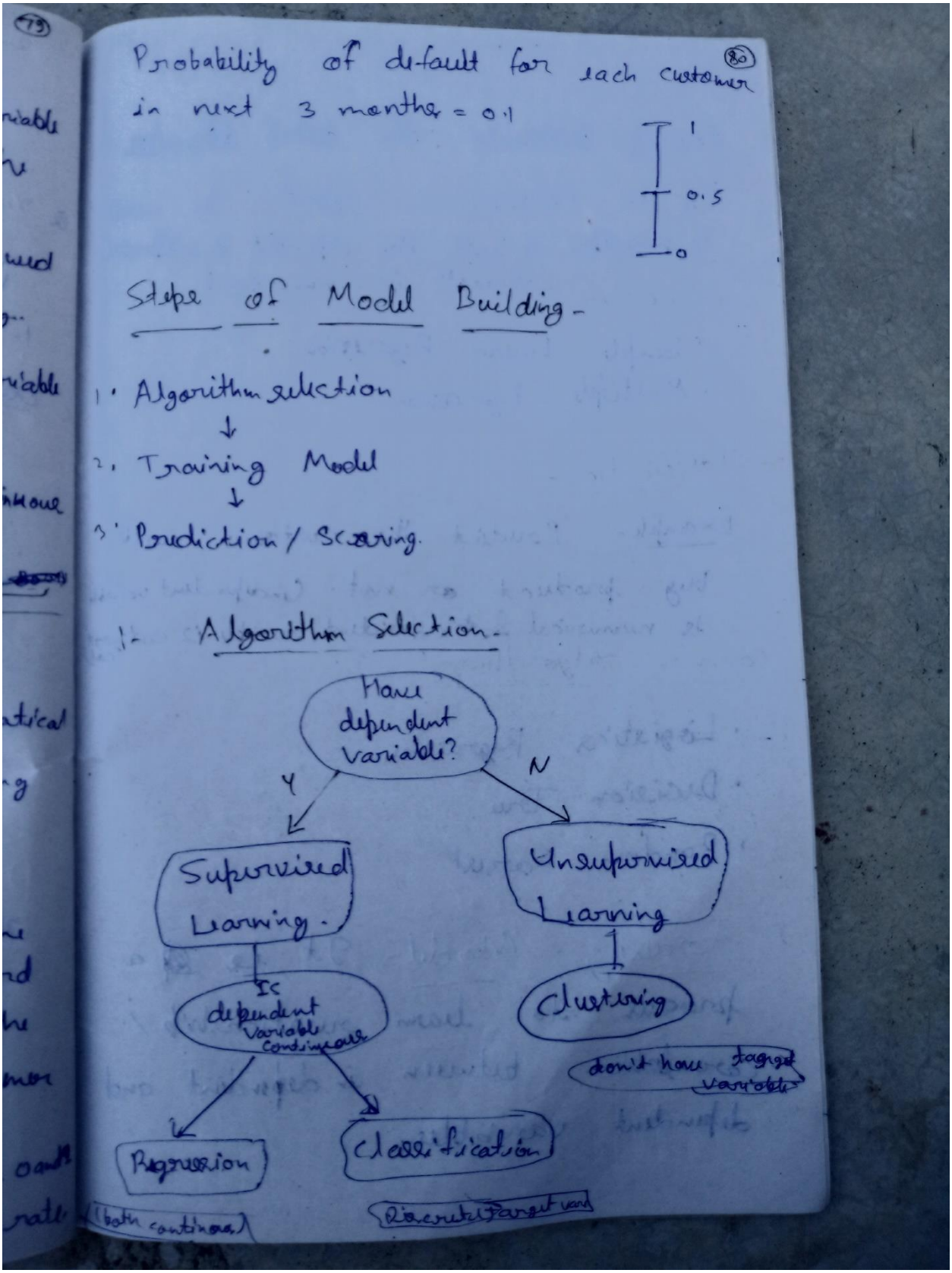
3- Cube root - Used for skewed variable with positive or negative value.
-> np.power(df['Age'], 1/3).plot.hist()

4- Binning - Used for converting continuous variable to categorical variables.
Ex - in Age - 10-35 (A), 40-60 (B), 60-80 (C),
Go to last page above 6

5- Model Building - It is a process to create a mathematical model for estimating / predicting the future behavior based on past data.

Ex - A retail bank wants to know the default behaviour of its credit card customers. They want to predict the probability of default for each customer with in next 3 months.

- Probability of default would lie between 0 and
- Assume every customer has 10% default rate



Probability of default for each customer in next 3 months = 0.1

1
0.5
0

Steps of Model Building -

1. Algorithm selection
↓
2. Training Model
↓
3. Prediction / Scoring.

1. Algorithm Selection -

Have dependent variable?

Y → Supervised Learning.

Is dependent variable continuous

Regression (both continuous)

Classification (Discrete target var)

N → Unsupervised Learning

Clustering

don't have target variable

1. Supervised Learning =

(81)

- Regression -

Example - Estimate the total sales by an insurance agent in next 3 months. (both the dependent & independent variable is numerical)

Algorithms -

- Simple Linear Regression.
- Multiple Regression.

- Classification -

Example - Predict the customer will buy product or not. (independent variable is numerical & dependent variable is categorical)

Comman Algorithm

- Logistic Regression
- Decision Tree
- Random Forest.

2- Traning Model - It is a process to learn relationship / correlation between independent and dependent variable.

Data Set -

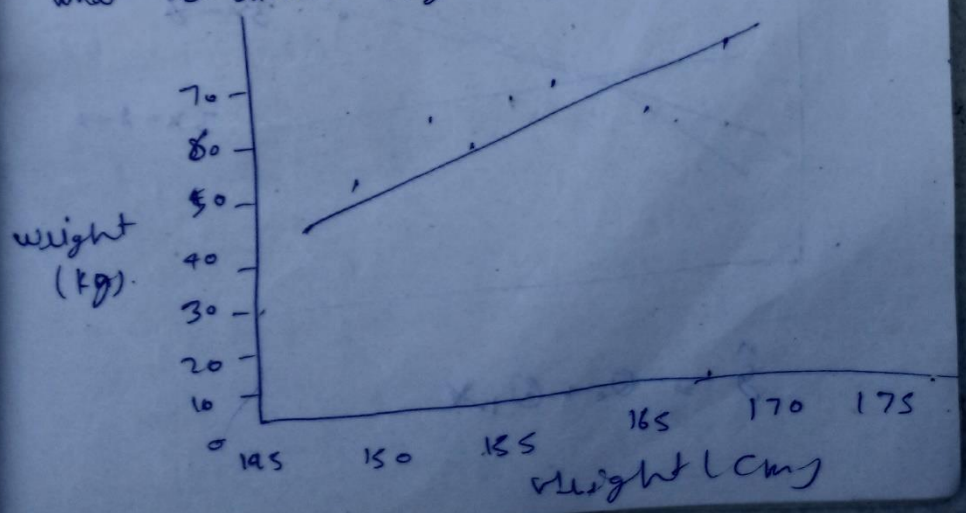
(82)

Train	Test
• Past data (known dependent variable)	• Future data (unknown dependent variable)
• Used to train model	• Used to score.

3. Prediction / Scoring - It is a process to estimate / predict dependent variable of test data set by applying model rule.

Linear Regression -

What is linear regression



83

Intercept Slope

$$\hat{y} = \theta_0 + \theta x$$

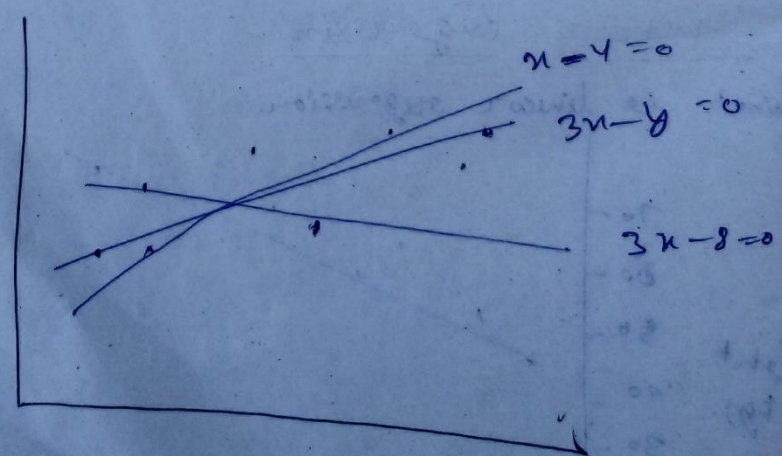
Height (cm) (X)	Weight (kg) (Y)
150	75
155	60
175	65
160	70
153	59
164	62
170	85

n - number of slope

x - independent feature

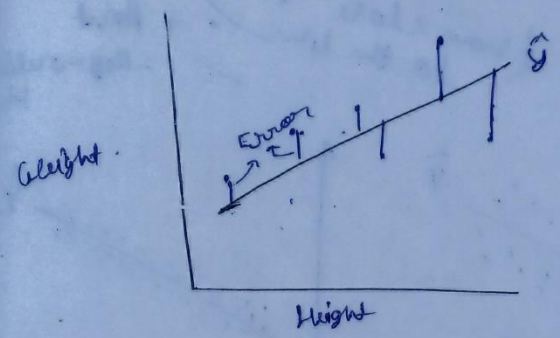
y - Target variable

• Cost function



$$\hat{y} = \theta_0 + \theta_1 x$$

(84)



$$\hat{y} = \theta_0 + \theta_1 x$$

Squared Error $= (\hat{y} - y)^2$

Cost function $\frac{1}{2n}(\theta_0 + \theta_1 x - y)^2$

- <u>Gradient Decent</u> -

$$\hat{y} = \theta_0 + \theta_1 x + \theta_2 x_2 + \cdots$$

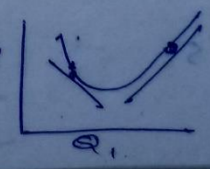
$$c = \frac{1}{2n} \Sigma (\hat{y} - y)^2$$

$$c = \frac{1}{2n} \Sigma (\theta_0 + \theta_1 x - y)^2$$

$$\frac{\partial c}{\partial \theta_0} = \frac{1}{n} \Sigma (\theta_0 + \theta_1 x - y)$$

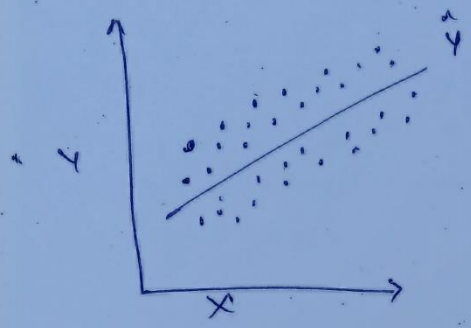
$$\frac{\partial c}{\partial \theta_1} = \frac{1}{n} \Sigma (\theta_0 + \theta_1 x - y) x$$

$$\theta_1 = \theta_1 - \frac{\partial c}{\partial \theta_1}$$

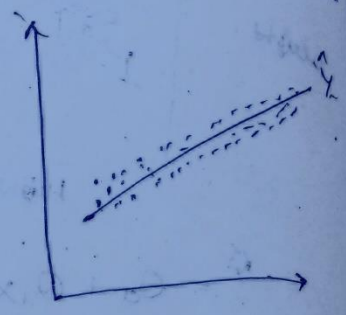


85

· R-squared (how close points to the line) • → Points
— → Regression line



R_1



R_2

$\boxed{R_2 > R_1}$

$$r^2 = \frac{\sum(\underbrace{\hat{Y}}_{\text{predicted value}} - \underbrace{\bar{Y}}_{\text{Actual mean}})^2}{\sum(Y-\bar{Y})^2} = \frac{SSR}{SST}$$

$$r^2 = 1 - \frac{SSE}{SST}$$

$SSE = \sum(\hat{Y}-Y)$

Ex → $\bar{Y} = 2$ (mean)

Y	$\hat{Y}$	Error $(\hat{Y}-Y)$	$(\hat{Y}-Y)^2$	Total variance $(Y-\bar{Y})$	$(Y-\bar{Y})^2$
0	1	−1	1	−2	4
1	2	−1	1	−1	1
2	3	−1	1	0	0
3	4	−1	1	1	1
4	5	−1	1	2	4
			5		10

(86)

$$r^2 = \frac{\Sigma(\hat{Y} - \bar{Y})^2}{(Y - \bar{Y})^2} = 1 - \frac{5}{10} = 0.5$$

Note: Value of r^2 lies between 0 & 1 (ie $0 \le R^2 \le 1$

RMSE (Root Mean Square Error).

Y	$\hat{Y}$	$Y - \hat{Y}$	$(Y - \hat{Y})^2$
1	2	-1	1
2	7	-5	25
3	5	-2	4
4	4	0	0

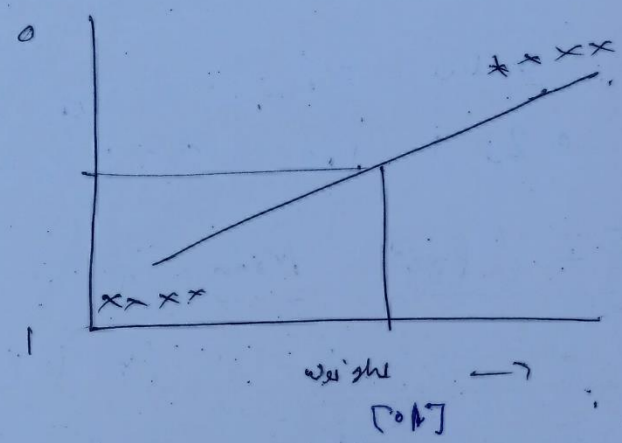
$$MSE = \frac{(Y - \hat{Y})^2}{n}, \quad RMSE = \sqrt{\frac{(Y - \hat{Y})^2}{n}}$$

$$RMSE = \sqrt{\frac{30}{4}}$$

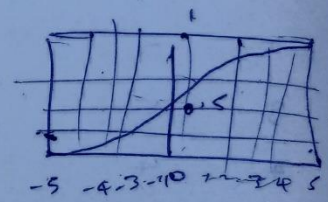
Ex- Low RMSE is better

87

Logistic Regression



$P = \text{Sigmoid} = \frac{1}{1+e^{-\hat{y}}} \Rightarrow$



$y = 0$ or 1

$P(y=1|x) =$

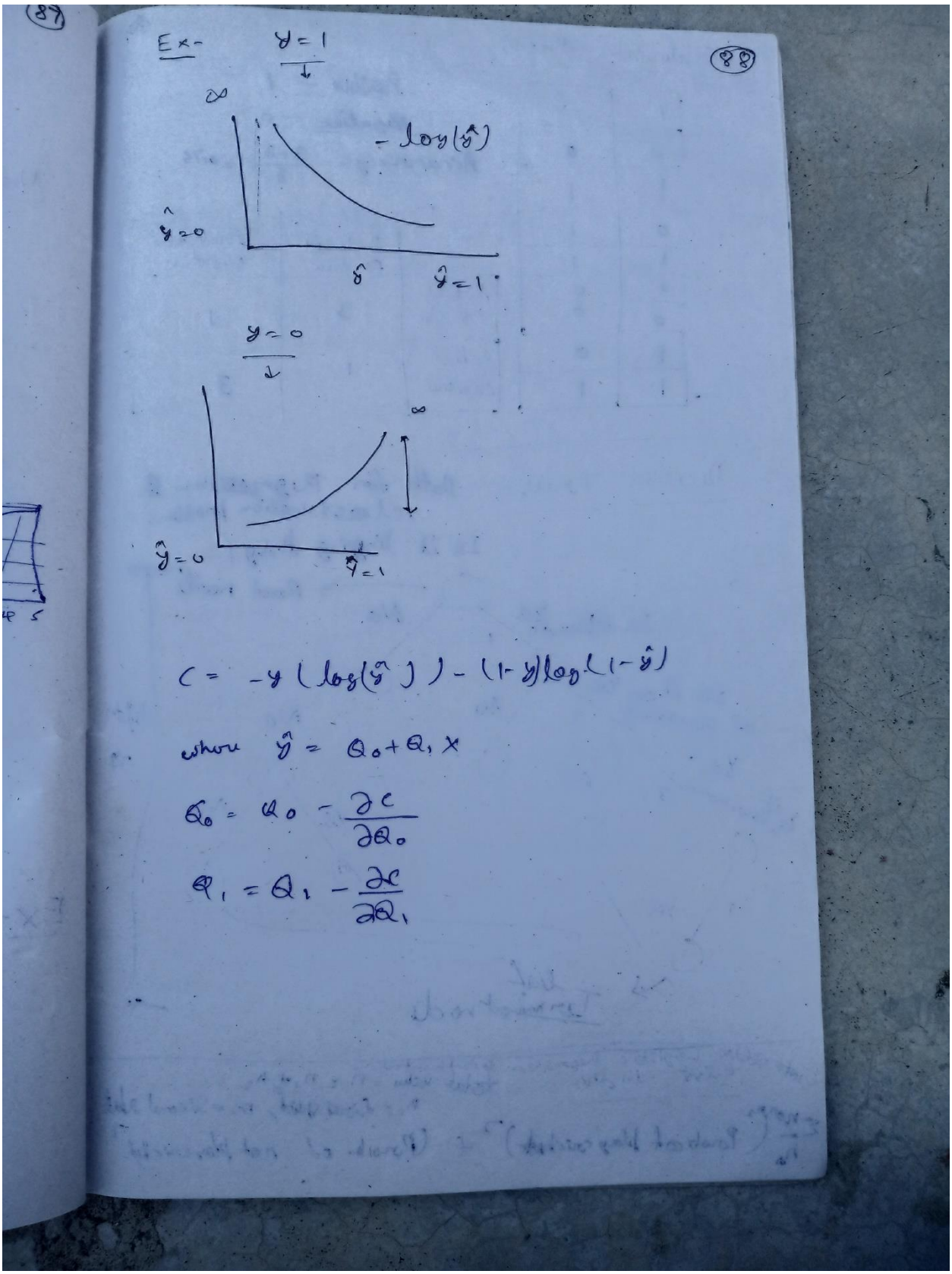
Cost function -

$$\frac{1}{1+e^{-\hat{y}}}$$

$$C = \frac{1}{2n}\Sigma(\hat{y} - y)^2$$

$$\begin{matrix} -\log(\hat{y}) \\ -\log(1-\hat{y}) \end{matrix} = \begin{bmatrix} y=1 \\ y=0 \end{bmatrix}$$

$$C = -y(\log(\hat{y})) - (1-y)\log(1-\hat{y})$$



$$C = -y(\log(\hat{y})) - (1-y)\log(1-\hat{y})$$

where $\hat{y} = Q_0 + Q_1 x$

$$Q_0 = Q_0 - \frac{\partial C}{\partial Q_0}$$

$$Q_1 = Q_1 - \frac{\partial C}{\partial Q_1}$$

Evaluation Matrix-

Y	$\hat{Y}$
0	0
1	1
0	1
1	1
0	0
0	0
1	0
1	1

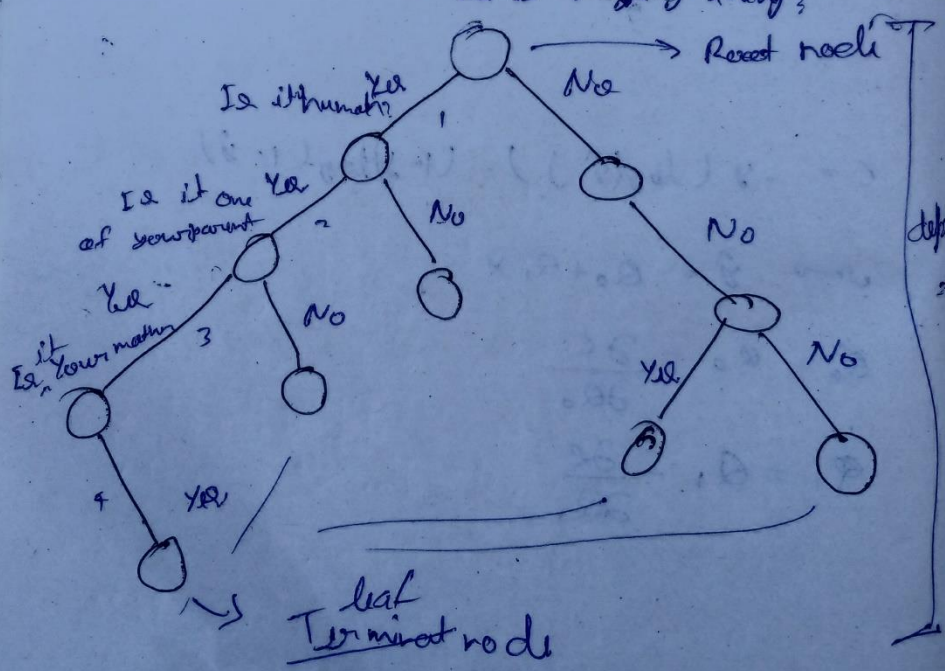
Positive – 1

Negative → 0

Accuracy – $\frac{3+3}{8} = 0.75$

	Predicted Positive	Predicted Negative
Actual Positive	3	1
Actual Negative	1	3

Decision Tree:- Both for Regression & classification problem



Gini index- total value $= n = n_1 + n_2$

n_1 = first split, n_2 = second split

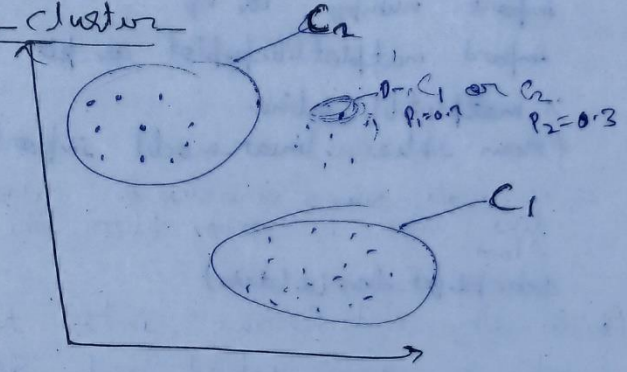
$\sum \frac{n_1 \text{ or } n_2}{n} (\text{Prob of play cricket})^2 + (\text{Prob. of not play cricket})^2$

(90)

Note- large gini index is indicate split is better.

2. Un-Supervise Learning-

Cluster



0 - C_1 or C_2
$P_1 = 0.7$ $P_2 = 0.3$

Types
1- Soft C_1 or C_2 & 2- Hard (C_1 & C_2)

Algorithm for Hard clustring-

K-means-

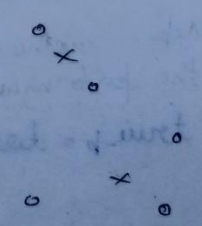
K = 2

(2, 6)
(3, 4)
(8, 3)
(4, 1)
(7, 1)

Centroid = $\frac{x_1 + x_2}{2}, \frac{y_1 + y_2}{2}$

clusterid = $\sqrt{(4-6)^2 + (1-2)^2}$
$= \sqrt{5} = 2.24$

clusterid = $\sqrt{(4-4)^2 + (1-1/3)^2}$
$= 8/3 = 2.67$



Perform Linear Regression in Pandas

1- # importing the require libraries

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
% matplotlib inline.
from sklearn.linear_model import LinearRegression
```

We create dummy variables for catogorical variables like value in string, sklearn take input only float.

```
data=pd.get_dummies(data)
```

2- Divide the data frame into two part first is train dataset and second one is test dataset

```
train = data[0:7999]
test = data[8000:]
```

3- test data set does not have target variable and train dataset have both dependent (target) & independent variable but these two are saperatly.

```
X_train.drop('Item_Outlet_Sales', axis=1)
y_train = train['Item_Outlet_Sales']
X_test = test.drop('Item_outlet', axis=1)
```

4- Make another copy of test dataset for checking the performance of the model.

```
true_p = test['Item_Outlet_Sales']
```

(92)

fill the missing values-

```
x_train.fillna(0, inplace=True)
x_test.fillna(0, inplace=True)
```

create object of LinearRegression & fit the data set (train our model)

```
lreg = LinearRegression()
lreg.fit(x_train, y_train)
```

predict on the test -

```
lreg.predict(x_test)
```

check the performance of model using rsquare. (It take 2 argument, first is independent variable of test dataset & second is their true value)

We check rsquare value of test dataset & also train data set

```
lreg.score(x_test, true_p)
lreg.score(x_train, y_train)
```

RMSE (root mean square error)

```
rmse_test = np.sqrt(np.mean(np.power((np.array(true_p) - np.array(pred)), 2)))
rmse_train = np.sqrt(np.mean(np.power((np.array(y_train) - np.array(lreg.predict(x_train))), 2)))
```

Note- first we change all categorical dependent variable to the dummy variable & fill all the missing value then perform regression.

Perform Logistic Regression - in Pandas

```
#import required libraries
import pandas as pd
import matplotlib.pyplot as plt
% matplotlib inline
import numpy as np

data = pd.read_csv("titanic.csv")
data.head()
data['survived'].value_counts()

# Make dummies -
data = pd.get_dummies(data)
# filling missing value -
data.fillna(0, inplace = True)
# classify train & test data set.
train = data[0: 699]
test = data[700: 890]
# Seperate the independent & dependent variable
x_train = train.drop('Survived', axis=1)
y_train = train['Survived']
x_test = test.drop('survived', axis=1)
true_p = test['Survived']
# import Logistic Regression
from sklearn.linear_model import LogisticRegression
logreg = LogisticRegression()
```

(94)

train ~~dataset~~ Model

logreg.fit (x_train, y_train)

prediction -

pred = logreg.predict(x_test)

check performance of model.

logreg.score (x_test, true_p)

logreg.score (x_train, y_train)

Decision tree -

Gini index -

Ex - $n = 30$

ID	class	Gender	Play

Split according to gender

F $n_1 = 10$

ID	Class	Gender	Play

$n_2 = 20$ M

ID	Class	Gender	Play

$p = 2/10 = 0.2$, $q = 0.8$

$p = 13/20$ $q = 7/20$

$0.2 * 0.2 + 0.8 * 0.8$

$0.65 * 0.65 + 0.35 * 0.35$

$$\frac{n_1}{n}\left[(0.2)^2 + (0.8)^2\right] + \frac{n_2}{n}\left[(0.65)^2 + (0.35)^2\right]$$

$$\text{Gini index} = \frac{10}{30}\left[(0.2)^2 + (0.8)^2\right] + \frac{20}{30}\left[(0.65)^2 + (0.35)^2\right]$$

Gini index = 0.59

Note - large gini index is indicate split is better

Perform the Decision Tree coding in Pandas

```
# import libraries-
import pandas as pd
import numpy as np
# reading csv file - (which is already cleaned)
data = pd.read_csv
# separating independent & dependent variables
x = data.drop['Survived']
y = data['Survived']
# split the dependent & independent variable
into two parts-
from sklearn.model_selection import train_test
-split

train_x, test_x, train_y, test_y = train_test_
split(x, y, random_state=101, stratify=y)
```

Note - random_state use to get exact same split & use stratify = y to make sure no. of 1's & 0's similar in both train & test dataset.

```
# check the % of 0's & 1's in both
the train & test data set is
similar or not.
trainy.value_counts()/len(train_y)
test_y.value_counts()/len(test_y)
# import Decision tree Classifier
from sklearn.tree import DecisionTreeClassifier
```

(96)

```
# train model
clf = DecisionTreeClassifier()
clf.fit(train_x, train_y)

# check the performance of model
clf.score(train_x, train_y)
clf.score(test_x, test_y)

# prediction
clf.predict(train_x)
clf.predict(test_x)
```

Perform the coding K-means in Pandas

```
# import library
import pandas as pd
import numpy as np
from matplotlib import as plt
%matplotlib inline
from sklearn.cluster import KMeans

# read dataframe
data = pd.read_csv("student-evaluation.csv")
data.shape
data.head()
# check there are not any missing value in dataset
pd.isnull(data).sum()
```

```
# check all the value in numerical form
data.describe()
# decide the number of clusters
kmeans = KMeans(n_clusters=2)
# train model
kmeans.fit(data)

# prediction
pred = kmeans.predict(data)
pd.Series(pred).value_counts()   {check no. of 0's & 1's}
# check clustering is good or not
kmeans.inertia_   {low inertia is better}

kmeans.score(data)   {it give -ive value of inertia}

# find the optimal value of cluster
SSE = [ ]

for cluster in range(1,20):
    kmeans = KMeans(n_jobs=-1, n_clusters=cluster
    kmeans.fit(data)
    SSE.append(kmeans.inertia)
frame = pd.DataFrame({'Cluster': range(1,20), 'SSE':SS
frame.head()

plt.figure(figsize=(12,6))
plt.plot(frame['Cluster'], frame['SSE'] marker
                                        = "o")
```

98

```
# Scale the data.
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
data_scaled = scaler.fit_transform(data)
# check optimal value of cluster.
SSE_scaled=[]

for cluster in range(1,20):
    kmeans = KMeans(n_jobs = -1, n_clusters =cluster)
    kmeans.fit(data_scaled)
    SSE_scaled.append(kmeans.inertia_)
frame_scaled = pd.DataFrame({'Cluster': range
    (1,20),'SSE' :SSE_scaled})
plt.plot(frame_scaled['Cluster'], frame_scaled
                ['SSE'], marker="O")

plt.xlabel("Clusters")

plt.ylabel("SSE")
# prediction
kmeans = KMeans(n_jobs = -1 , n_clusters = 4)
kmeans.fit(data_scaled)
pred = kmeans.predict(data_scaled)
frame['cluster'] = pred
frame.loc[frame['cluster']==3:]
```

95

4) Imputation-

- fill all the missing values in a dataframe with 0.

→ fill.fillna(0) (for only view.) temporary.

→ fill.fillna(0, inplace=True)

- fill the missing value of any column

→ df['column'].fillna(value)

5) Outliers-

- Removing outliers -

→ df = df[df['columnname'] < value]

- Replacing outliers-

→ df.loc[df['column'] > value, 'column'] = np.mean(col)

6) Variable transformation in Pandas-

first see the variable distribution using histogram

- then perform transformation.

Binning-

→ bins = [start, middle, end]

group = ['group1', 'group2']

df['New column'] = pd.cut(df['column'], bins, labels=group)

Paired ~~T~~ test.

(1) (100)

- We check if the same sample has changed it's behavior after an intervention, or behave significantly different in two conditions.

- Null Hypothesis - would be - That the difference two sample means is 0 ie $\bar{X}_1 - \bar{X}_2 = 0$
- Alternate Hypothesis - will be That there is a significant difference between the two ie. $\bar{X}_1 - \bar{X}_2 <> 0$

Formula for Paired T test -

- $$t = \frac{\bar{d}}{S_d / \sqrt{n}}$$

- where d (bar) = mean of the case wise difference between before and after. S_d is the standard deviation of the difference and n is the sample size