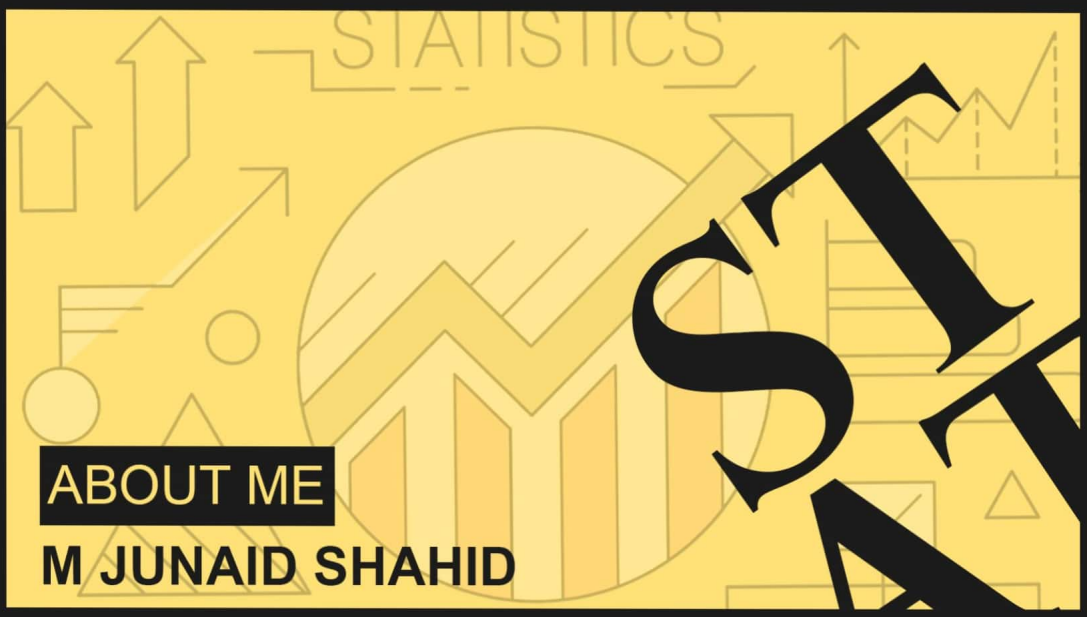
DATA SCIENCE

STAT

lecture 1

https://www.linkedin.com/in/iam-junaid/



I am a student of MATHEMATICS and studyi\g in the field of data science and data analysis. I am learning online at CAMPUSX DSMP (data science mentorship program). I have done PYTHON with numpy, pandas, matplotlib, seaborn, SQL(structured query language), MATH(linear algebra, vectors, calculus), STATISTICS(Inferential statistics, descriptive statistics, and theorems) and I am still learning to get and polish my new skills to secure my dream DATA SCIENTIST job.

LEC-1

For Examiner's Use

Descriptive Statistics

What is Statistics :-

:- Statistics is a branch of mathematics that involves collecting, analysing, interpreting, and presenting data. it provides tools and methods to understand and make sense of large amounts of data and to drawn conclusions and make decision based on the data.

In practice, statistics is used in a wide range of field, such as buisness, economics, social science, medical and engineering. it is used to conduct research studies, analyze market trend, evaluate the effectiveness of treatements and interventions, and make forecasts and predictions.

[Turn over

Example:-

Buiseness (identifying Customer behaviour)
Medical (identify efficiency of new medicine (clinical Trials)
Government & politics (Conducting Surveys)
Enviornmental Science (climate research).

Types of Statistics

- Descriptive Statistics
- Inferential Statistics.

Descriptive Statistics :-

Descriptive statistics deals with the collecting, organization, analysis interpretation and presentation of data it focus on Sumarizing & describing the main feature of the set of data. without making inference or prediction about the large population.

Inferential Statistics :-

deals with making conclusion and predidions about a population based on Sample . it involves the use of probability theoreoy to estimate the likehood of certain event ocuring . hypothesis testing to determine if a certain claim about a population is supported by the data and regression analysis to examine the relationship between Variables.

Population Vs Sample :-

=>Population :-

population refers to the entire group of invidual or object that we are intrested in Studing it is Complete Set of observation that we want to make enference about . For example The population might

be all the students in a particular school or all the cars in particular city.

Sample :-

A Sample, on the other hand, is a Subset of the population. it is a Smaller group of individuals or objects that we select from the population to study. Samples ase used to estimate characteristics of the population Such as the mean or the proportion with a certain attributes. for example we might randomly select 100 students.

Example :-

(1) All students (population) vs {who visit school for lectures} (Sample)

Things to be carefull while creating Samples:

1. Sample Size
2. Random
3. Representative.

Parameter Vs Statistics

Parameter:-

A parameter is a characteristics of a population.

Statistics:-

A Statistics is a characteristics of a Sample.

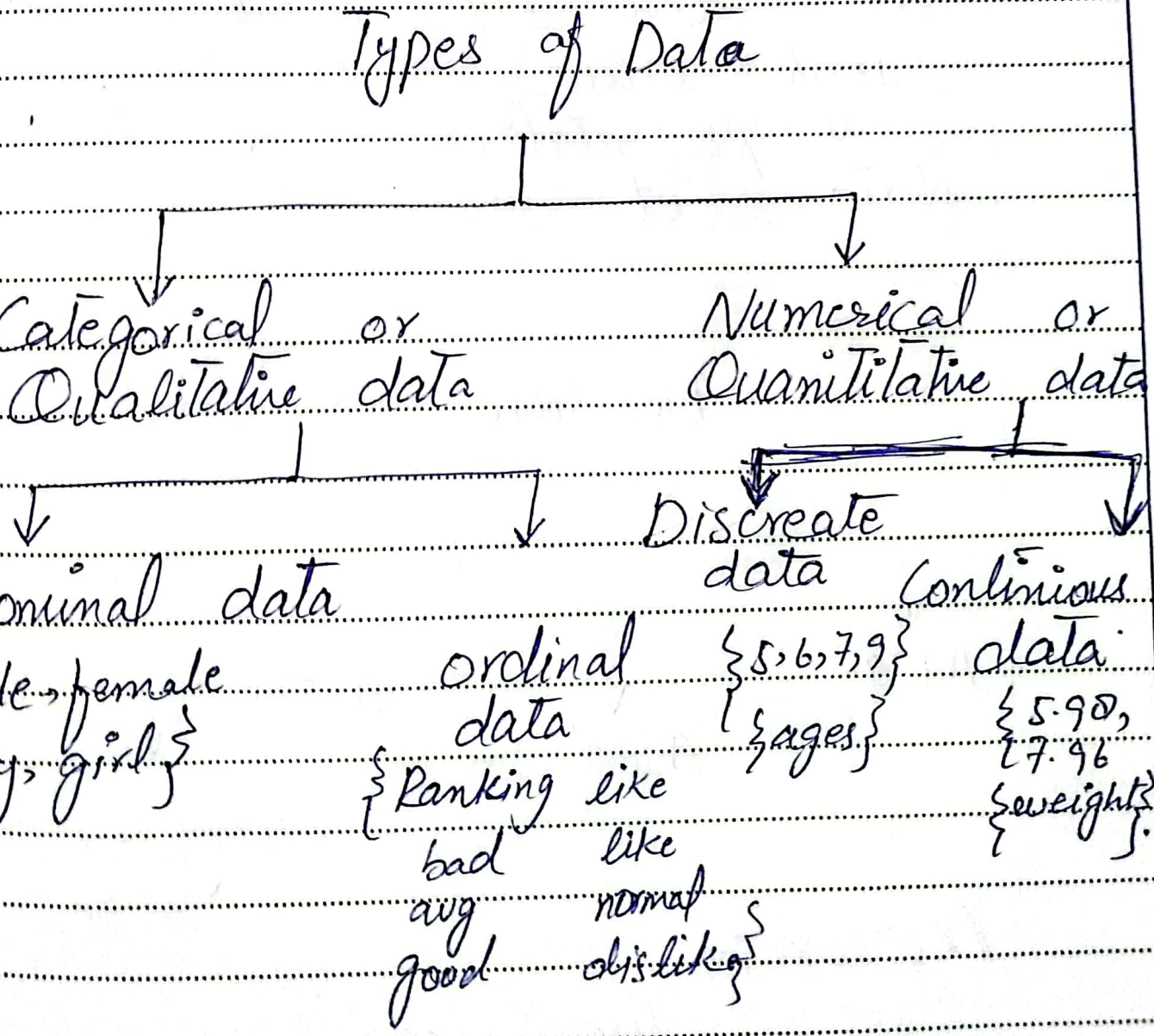
Some of Topic that Come under inferential Statistics

=> Hypothesis Testing:-

=> Confidence Interval:-

⇒ Analysis of variance (ANOVA)
⇒ Regression Analysis
⇒ Chi-Square Test
⇒ Sample techniques
⇒ Bayesian statistics.

{Types of Data}



3 (a) Mu'awiya was opposed to 'Ali's caliphate. Give reasons for his opposition and write a account of the Battle of Siffin which resulted from this opposition. [1

(b) In your opinion what was the most serious consequence of the outcome of this battle? Giv reasons for your answer. [4

Measure of Central Tendency

A measure of Central Tendency is a statical measure that represents a Typical Central value of dataset. it provide a Summary of a data by identifying a Single value that is most representative of the data set as whole.

Measure of C.T

- Mean
- Median
- Mode
- Trimmed Mean
- weighted mean

Mean :-

The mean is a Sum of all values in

the dataset devided by the number of values.

Population Mean

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

$N \longrightarrow$ no of items in population

Sample Mean

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

$n \longrightarrow$ no. of items in Sample.

Marks
5
4
3
2
1
7

$$\bar{x} = \frac{5+4+3+2+1+7}{6}$$

$$\boxed{\bar{x} = 3.66667}$$

Note:-

But if there an a outlier in our data set so the the mean

will be not acurate

like

Marks
1
2
3
5
7
9
10
69

with outlier

$$\bar{x} = \frac{1+3+2+5+7+9+10+69}{8}$$

$$\boxed{\bar{x} = 13.25}$$

69 → outlier

Marks
1
2
3
5
7
9
10

without outlier.

$$\bar{x} = \frac{1+2+3+5+7+9+10}{7}$$

$$\boxed{\bar{x} = 5.28}$$

So that's why we used median when an outlier in our data set.

Median:-

The median is

(a) Islamic teachings revolve around six main Articles of Faith. Write about the following two:

- belief in God, and
- belief in angels. [10]

(b) Why is the belief in angels important for Muslims? [4]

is a ~~medle~~ middle value in our data set. when the data is arranged in order.

Marks
1
5
7
6
4
3
2

without Outlier

Arrange them

1, 2, 3, (4), 5, 6, 7

middle value

Median = 4

Marks
1
2
3
4
5
6
7
69
70

Arrang

1, 2, 3, 4, (5), 6, 7, 69, 70

Median = 5

This is actually fine from Mean & more accurate.

Mode :-

The mode is a value that appears most frequently in the dataset

Marks
1
5
9
7
5
6
5
4
5
9
7

Mode = 5

↓

because 5 contain more times in dataset.

Weighted Mean :-

The weighted mean is a sum of the product of each value and its weight, devided by the sum of the weights. it is used to calculate a mean when the value in the

data set have different importance or frequency.

let we predict on model.

L.R → accuracy = 0.2 → on this 10 data
R.F → accuracy = 0.3 → on this 15 data
Xgboo → accuracy = 0.5 on this 12.

product weight value

W_m = Sum of product of value with its weight / Sum of weights

$$= \frac{0.2 \times 10 + 0.3 \times 15 + 0.5 \times 12}{0.2 + 0.3 + 0.5}$$

W_m = Ans

Trimmed Mean:-

A trim mean is calculated by removing a certain percentage of the smallest and largest value from the dataset and then taking the mean

5 (a) 'Prophets played a central part in conveying God's message to humanity.' Write an account Muslim belief in prophets. [1

(b) Why do you think God gave miracles to his chosen prophets?

of the remaining values.
the percentage of the value removed is called trimmed percentage.

=> withoutliers & not trimmed
20k, 22k, 23k, 25k, 28k, 30k, 32k, 35k, 50k, 80k

$\bar{X} = 36k$

=> By trimmed 20% from largest & smallest side

25k, 28k, 30k, 32k, 35k

$\bar{X} = 30k$

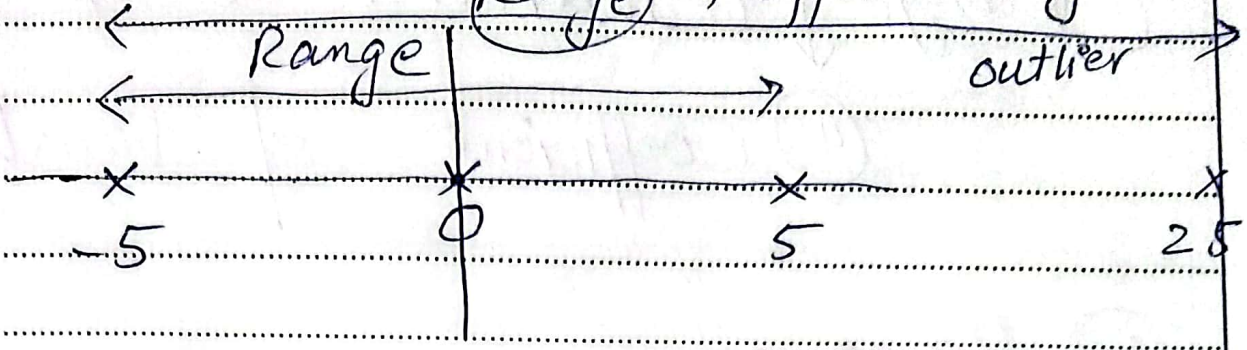
Measure of Dispersion:-

A measure of dispersion is a statical measure that describe the spread or variability of dataset. it provide information about how much data set is distributed around the

Central tendency (mean, median, mode) of dataset

Range → effected by outlier

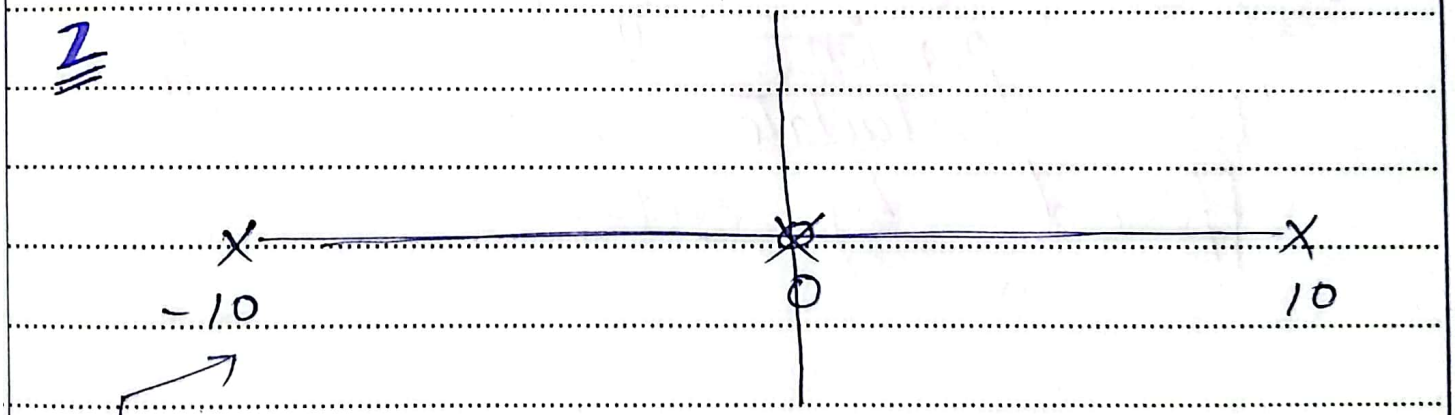
1



$\bar{X} = \frac{-5+0+5}{3} = \frac{0}{3} \Rightarrow 0$ (Mean)

Range = $5 - (-5) = 10$

2



$\bar{X} = \frac{-10+0+10}{3} = 0$ (Mean)

Range = $+10 - (+10) = 20$

Both Same in two example So how we identify we identify it by how the data is spread by measure of dispersion.

Measure of dispersion

(1) Range | (2) Variance | (3) S.D

(4) Coefficient of Variation

(1) Range :-

The Range is the difference between the maximum and minimum values in the dataset. it is simple measure of dispersion that is easy to calculate but can be effected by outliers

=> Range = Max-Value — Minimum

(2) Variance :-

The variance is the average of the squared difference between each data

point and the mean. it measures the average distance of each data point from the mean and is usefull in comparing the dispersion of dataset with different means

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{N} \quad \text{population}$$

$$S^2 = \frac{\sum_{i=1}^{n} (x_1 - \bar{x})}{n-1} \quad \text{Sample}$$

Data	$x_i - \bar{X}$	$(x_i - \bar{x})^2$
3	3-3	0
2	2-3	1
1	1-3	4
5	5-3	4
4	4-3	1

$\bar{X} = \frac{3+2+1+5+4}{5} = 3$

$$S^2 = \frac{0+1+4+4+1}{5}$$

$S^2 / \sigma^2 =$ Ans

Note

So we used Square on each

Value for if we get the answer in negative after to sort out we used Square one person said we can used absolute in / on other hand => Yes, we can but one markdown is the absolute formulla not provide enference about data we will use the name of Question is

=> Mean absolute deviation :-

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

:. STANDARD DAVIATION :-

S.D is a square root of variance it is wieldy used measure of dispersion that is usefull in describing the shape of distribution.

its also used to change unit of data Same as data => mean if I have.

writing?

if i have a dataset have unit kgs then by using variance the unit is $(kgs)^2$ & again taking S.D it convert the $\sqrt{(kgs)^2}$ it in same unit of data

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

$$S = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n-1}} \quad \text{Sample.}$$

=> Coefficient of Variation:-

Coefficient of variation (CV) The ratio of the standard deviation to the mean expressed as a percentage. it is used to compare the variability of datasets

with different mean and is commonly used in field such as biology, Chemistry & engeenering.

Formulla

$$CV = \frac{\text{Standard daviation}}{\text{Mean}} \times 100$$

it tells us also the distribution between two Comparing Quantities how much data is spread.

Graphs For univariate Analysis

- Categorical
- Numerical

Categorical :- Frequency distribution

A frequency distribution table is a table that Summarizes the numbes of times (for frequency)

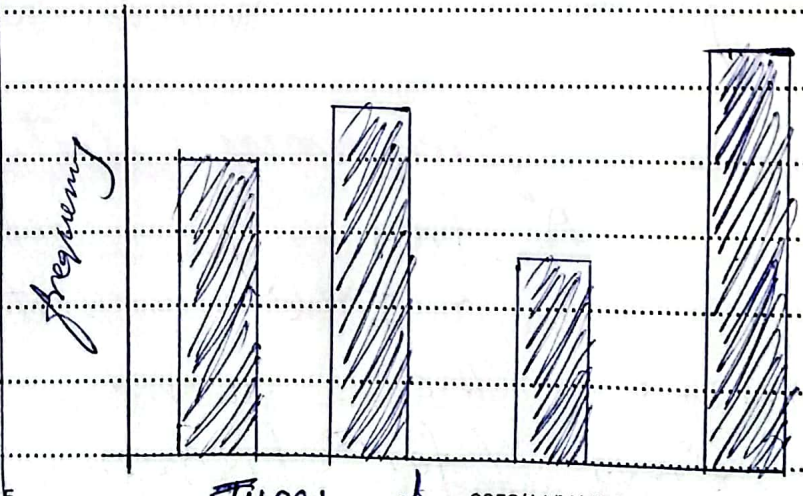
That each value occur in data set

E.G

let Survey of 200 peoples and ask about faviourite type of vacation

Type of Vacation	frequency
Beach	60
City	40
Adventure	30
Nature	20
Cruise	15
other	35

In this we created bar graph.



(a) Write about the changes in the relationship between the Prophet and the Quraysh in the years between his marriage to Khadija and the death of Abu Talib. [10]

(b) Why is it significant that the Quraysh were still willing to keep their belongings with the Prophet after he started to preach Islam? [4]

Relative frequency:- is the proportion or percentage of a category in a data set or sample. it is calculated by deviding the frequency of a category by the total number of observation in the dataset

Types of vacation	frequency	R.F = freq / 200 × 100 =
Beach	60	0.3
City	40	0.2
Adventure	30	0.15
Nature	35	0.175
Cruise	20	0.1
other	15	0.0075

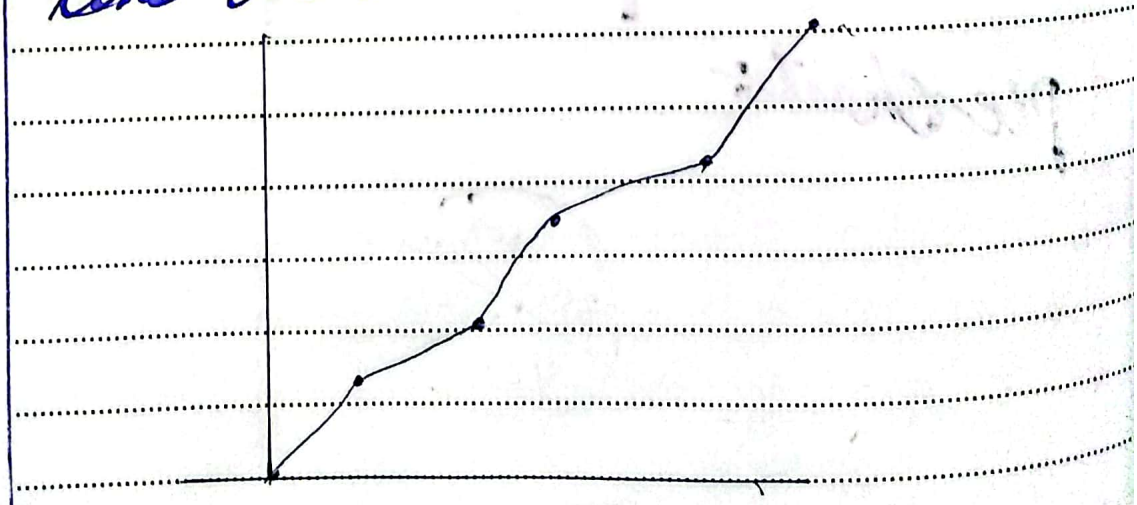
Pie Chart:-

Commulative frequency:-

is the ruming total of frequencies of a variable or category in a dataset or Sample. it is Calculated by adding up the frequencies of the Current Category and all previous Categories in the dataset or Sample.

Types of Vacation	frequency	Commulative frequency
Beach	60	60
City	40	100
Adventure	30	130
Nature	35	165
Cruise	20	185
other	15	200

line chart:-



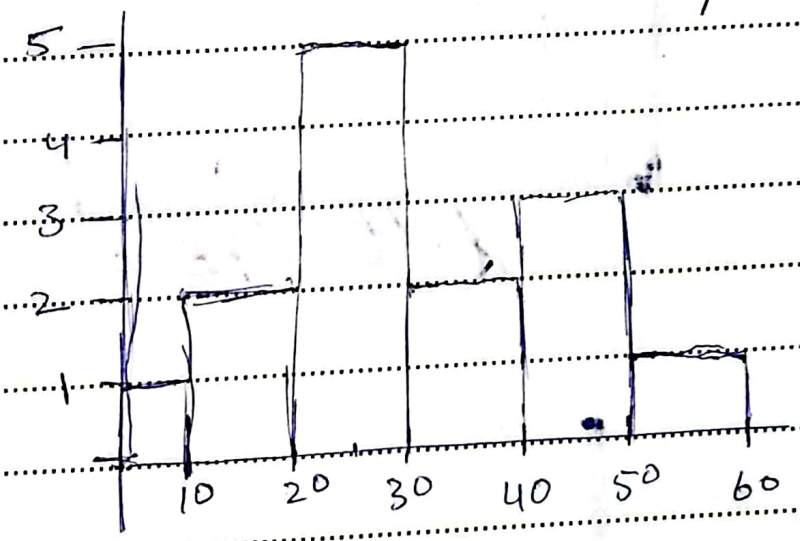
(2)

Numerical:-

In Univarient numerical column we Can draw frequency distribution Graph as well as Histogram

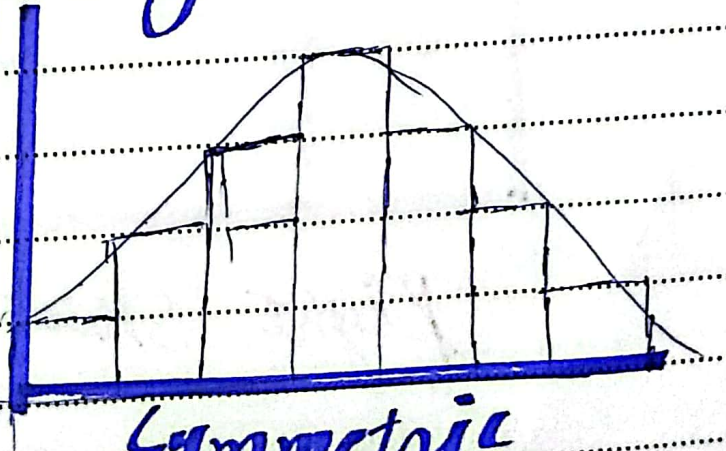
Histogram:-

it is a representation of data as a histplot In this graph we distribute our data in bins/buckets.

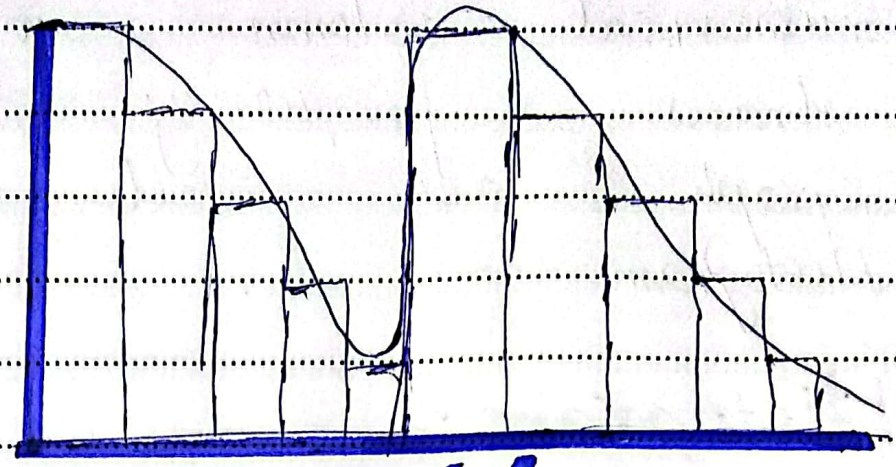


age	frequency
25	5
43	3
34	2
56	1

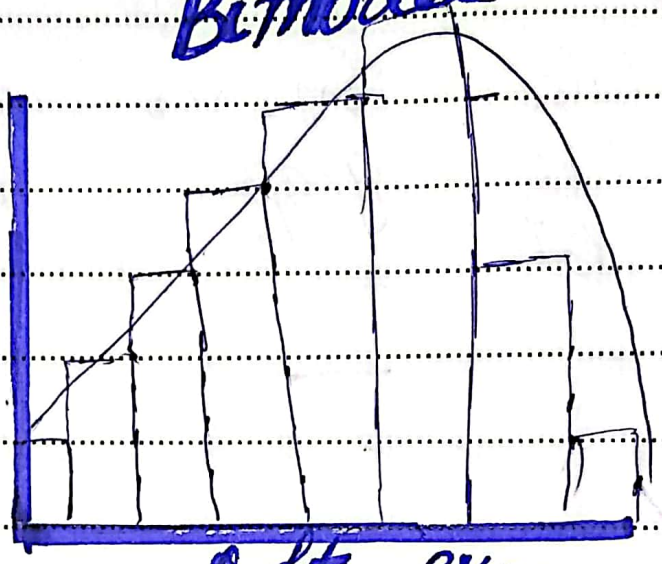
Types of histogram:-



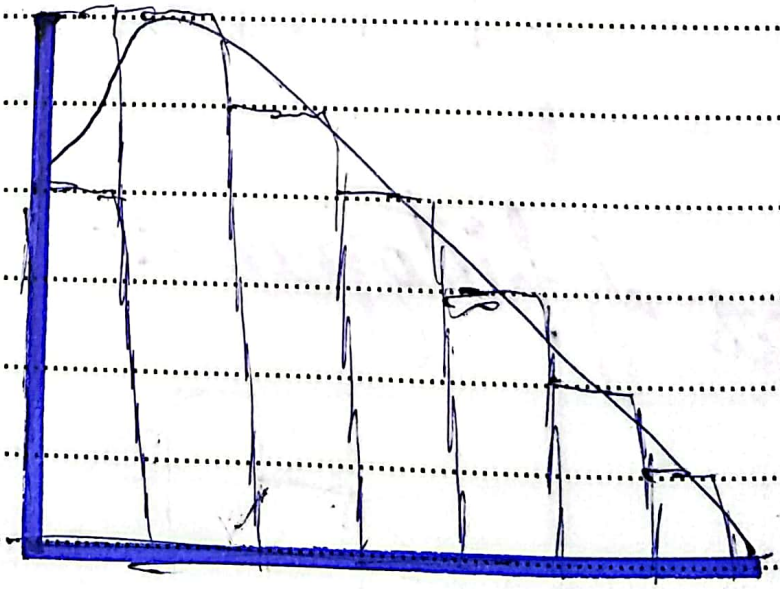
Symmetric



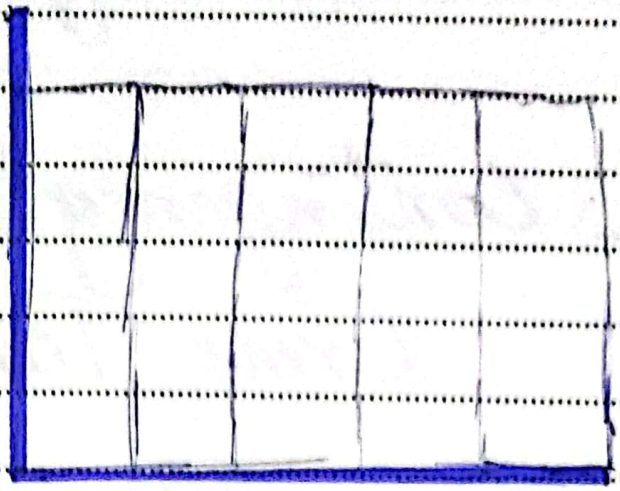
Bimodel



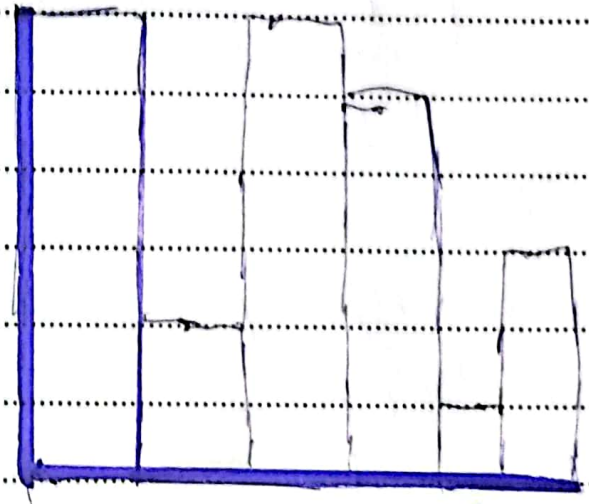
Left Skew



Right Skew



Uniform



No-pattern

≡ Graph of Bivarient Analysis

Categorical	Categorical	Numerical
Categorical	Numerical	Numerical

Categorical - Categorical

Consts Contingency Table / Cross Table

A Contingency Table also known as cross-Tabulation is a type of table used in statistics to summarize the relationship between two categorical variables.

A Consistengency table displays the frequency or relative frequency of the observed values of the two variables, organized into rows and colomns.

Survied	Pclass
	1
0	2
1	3

(b) What lessons can Muslims learn

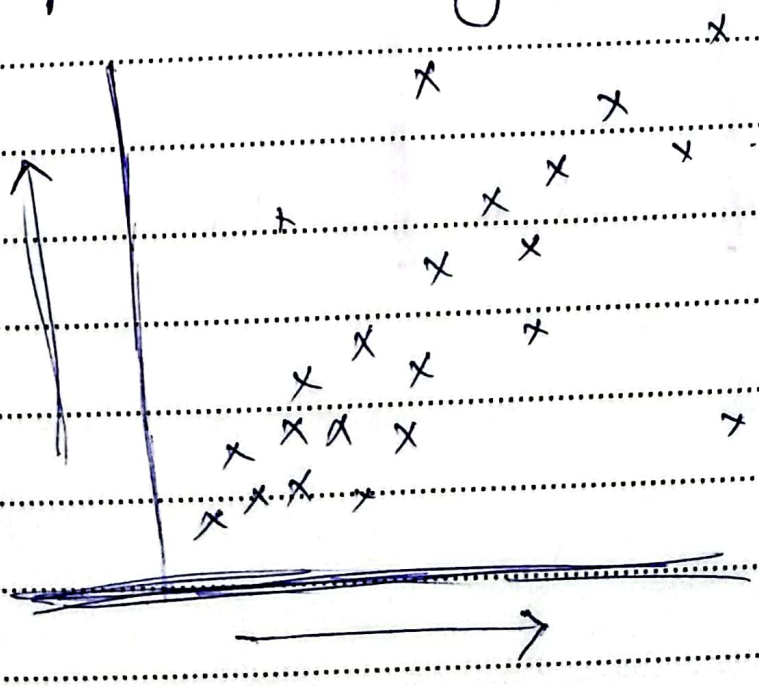
Survived	Pclass 1	2	3
1	71	31	63
0	42	118	13

Numerical - Numerical :-

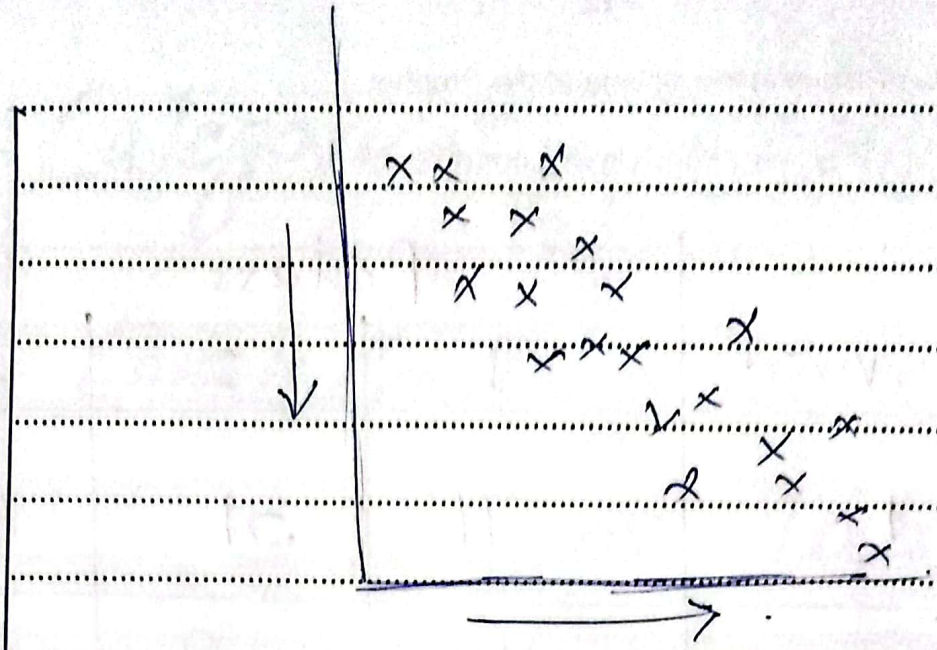
In Numerical - Numerical bivariete analysis we can draw a Scatter plot.

point where distribution of data Spread using points

Scatter plot :-



[Turn ove



Categorical - Numerical

:- In Categorical - Numerical Bivariate analysis we can make histogram Contigency table.

	0 - 10	11 - 20	21 - 30
Male	32	41	110
female	15	18	120

THANKS