STATISTICS-1

WEEK 1 - CLICK HERE	Introduction and type of data, Types of data, Descriptive and Inferential statistics, Scales of measurement
WEEK 2	Describing categorical data Frequency distribution of categorical data, Best practices for graphing categorical data, Mode and median for categorical variable
WEEK 3	Describing numerical data Frequency tables for numerical data, Measures of central tendency - Mean, median and mode, Quartiles and percentiles, Measures of dispersion - Range, variance, standard deviation and IQR, Five number summary
WEEK 4	Association between two variables - Association between two categorical variables - Using relative frequencies in contingency tables, Association between two numerical variables - Scatterplot, covariance, Pearson correlation coefficient, Point bi-serial correlation coefficient
WEEK 5	Basic principles of counting and factorial concepts - Addition rule of counting, Multiplication rule of counting, Factorials
WEEK 6	Permutations and combinations
WEEK 7	Probability Basic definitions of probability, Events, Properties of probability
WEEK 8	Conditional probability - Multiplication rule, Independence, Law of total probability, Bayes' theorem
WEEK 9	Random Variables - Random experiment, sample space and random variable, Discrete and continuous random variable, Probability mass function, Cumulative density function
WEEK 10	Expectation and Variance - Expectation of a discrete random variable, Variance and standard deviation of a discrete random variable
WEEK 11	Binomial and poisson random variables - Bernoulli trials, Independent and identically distributed random variable, Binomial random variable, Expectation and variance of abinomial random variable, Poisson distribution
WEEK 12	Introduction to continous random variables - Area under the curve, Properties of pdf, Uniform distribution, Exponential distribution

WEEK-1

10/09/22.

Statistics - The art of learning from data.
- descriptive
- inferential

Descriptive statistics: The part of statistics concerned with description and summarization of data.

Inferential statistics: The part of statistics concerned with drawing of conclusions from data.
- Draws conclusions from sample to population (leads to probability).

Population: The total collection of all the elements that we are interested in.

Sample - A subgroup of the population that will be studied in detail.

Purpose of statistical analysis:
- If the purpose of the analysis is to examine and explore information for its own intrinsic interest only, the study is descriptive.
- If the information is obtained from a sample of a population and the purpose of the study is to use that information to draw conclusions about the population, the study is inferential.
- A descriptive study may be performed either on a sample or on a population

▶ when an inference is made about the population, based on information obtained from the sample, does the study become inferential.

What is Data?

Data are the facts and figures collected, analyzed and summarized for presentation and interpretation.

▶ Statistics relies on data, info that is around us.

Data collection:

- Data available: Published data
- Data not available: need to collect, generate data.

(we assume data is available and our objective is to do a statistical analysis of available data)

Data.

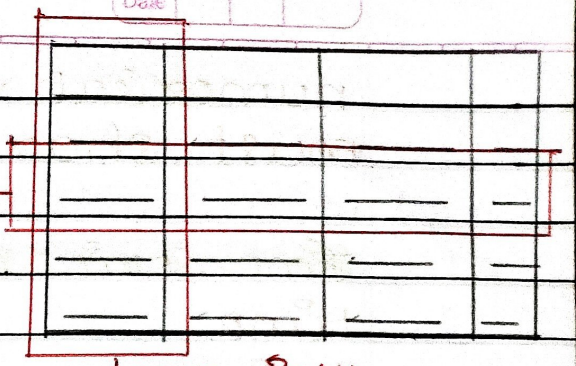
- Unstructured data
- Structured data.

when they are scattered about with no structure, the info. is of very little use, we need to organize data.

Dataset - Structured collection of dataset.

▶ it is a collection of values - could be nos, names, row nos.

Each variable must have its own column
Each observation must have its own row.



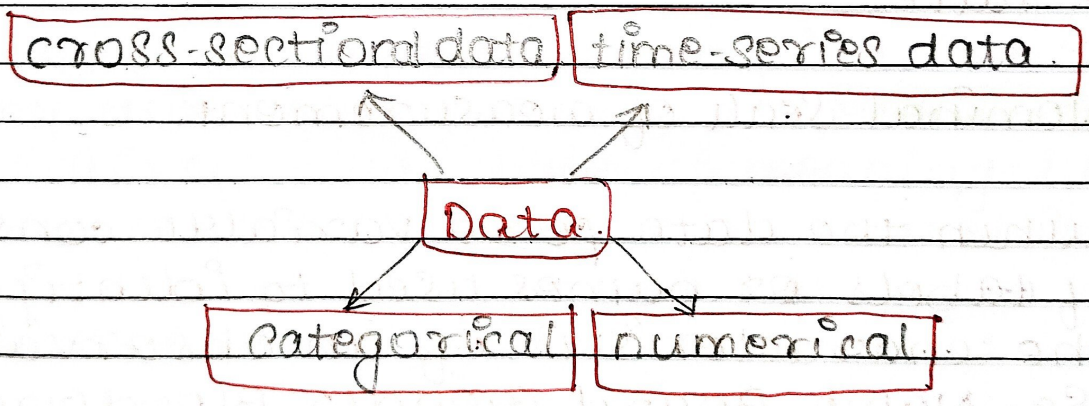
Variables and cases.
(columns) (ROWS)
(observation) case

Case: A unit from which data are collected.

Variable:
- Intuitive - A variable is that 'varies'.
- formally - A characteristic or attribute that varies across all units.

1. Rows represent cases: for each case, same attribute is recorded.
2. columns represent variables: for each variables, same type of value for each case is recorded.

Classification of data:



Categorical data: qualitative variables, identify group membership.
Numerical data: quantitative variables, describe numerical properties of cases, have measurement units.

measuring units: scale that defines the meaning of numerical data, such as weights in kgs. ▶ Those data that make up a

numerical variable in a data table must share a common unit.

Time-series data: data recorded over time

▶ Timeplot - graph of a time series showing values in chronological order

Cross-sectional data: data observed at the same time.

Scales of measurement:

Data collection requires one of the following scales of measurement:

- nominal
- ordinal
- interval
- ratio.

Nominal Scale of measurement.

when the data for a variable consist of labels or names used to identify the characteristic of an observation.

Eg: Name, Board, Gender, Blood Group.

Sometimes nominal variables might be numerically coded.

Eg: men as 1, women as 0, ...

There is no ordering.

Ordinal Scale of measurement

Data exhibits properties of nominal data and the order or rank of data is meaningful.

Eg: Rating: Good, poor, worst

- the data have properties of nominal data.
- the data can be ordered wrt the servic quality.

Interval Scale of measurement

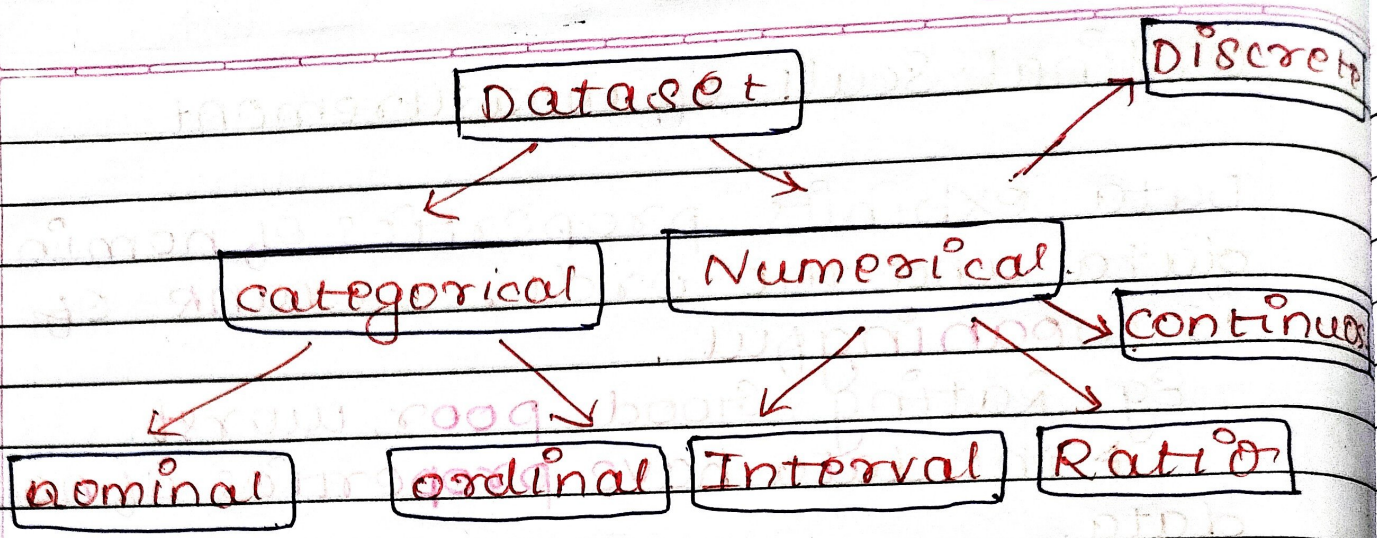
Data have all properties of ordinal data and the interval between values is expressed in terms of a fixed unit of measure.

- Always numeric, can find out diff b/w any two values.
- Ratio of values have no meaning here because the value of zero is arbitary.

Ratio Scale of measurement

Data have all the properties of interval data and the ratio of two values is meaningful

Eg: height, weight, age, marks.



Nominal

(name categories without order).

↓

ordinal

↓

(name categories with order)

↓

Interval

(num values that can be added/sub)

(no abs zero)

↓

Ratio

(num values that can be +/-/*/÷)

(Ratio comparisons possible)

(abs zero possible)

- Eg: temp - uncomfortable/comfor (Nominal)
  - good/bad/worst (ordinal)
  - 20°C/40°C/60°C (interval.)
    - ↳ Difference is possible

WEEK 2

Page No.
Date

Frequency distributions

A frequency distributions of of qualitative data is a listing of the distinct values and their frequencies.

Each row of a frequency table lists a category along with the no. of cases in this category.

1. A, A, B, C, A, D, A, B, D, C.

FREQUENCY DISTRIBUTION TABLE

	category / column	Tally mark	Frequency
Distinct values.	A	IIII	4
	B	II	2
	C	II	2
	D	I	1

Relative frequency

The ratio of the frequency to the total number of observations is called relative frequency.

2. A, A, B, C, A, D, A, B, D, C, A, B, C, D, A

Category	Tally mark	Frequency
A	IIII I	6
B	III	3
C	III	3
D	III	3
		Total = 15

Why relative frequency?

- For comparing two data sets.
- Because relative frequencies always fall b/w 0 and 1, they provide a standard for comparison.

Charts of categorical data:

- The two most common displays of a categorical variable are **bar chart** and **a pie chart**.
- Both describe a categorical variable by displaying its frequency table.

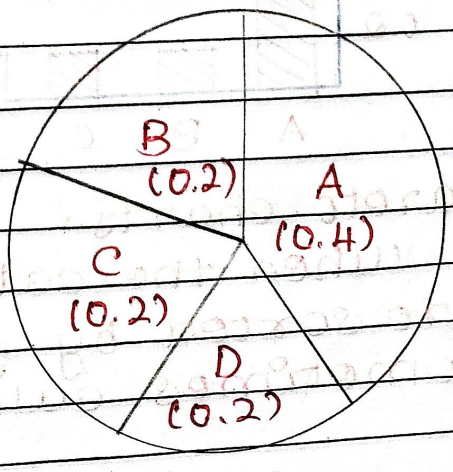
Pie chart:

It is a circle divided into pieces proportional to the relative frequencies of qualitative data. (RF × 360°) = Degrees.

3. A, A, B, C, A, D, A, B, D, C

Chart: Category	Tally mark	Freq.	RF	Degrees.
A	IIII	4	0.4	144 (0.4×360)
B	II	2	0.2	72 (0.2×360)
C	II	2	0.2	72 (0.2×360)
D	II	2	0.2.	72 (0.2×360)

Relative frequency
0.4 (6/15)
0.2 (3/15)
0.2 (3/15)
0.2. (3/15)
1

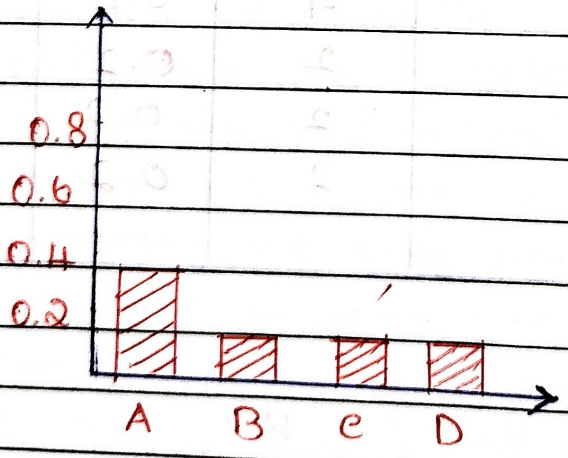


- A pie chart is used to show the proportions of a categorical variable.
- A pie chart is a good way to show that one category makes up more than half of the total.

Bar chart: chor/ver)

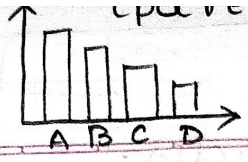
A bar chart displays the distinct values of the qualitative data on a horizontal axis and the relative frequencies (Freq / %) of ~~ve~~ those values on a vertical axis.

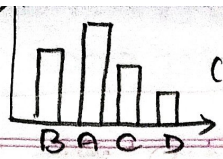
- The frequency / relative ~~r~~ of each distinct value is represented by a vertical bar whose height is equal to the ~~r~~ / RF of that value.
- The bars should be positioned so that they do not touch each other.



Pareto charts:

When the categories in a bar chart are sorted by freq. the bar chart is sometimes called a Pareto chart.





(Bar)

Page No.
Date

- Pareto charts are popular in quality control to identify problems in a business process.
- If the **categorical variable is ordinal**, then the bar chart must preserve the ordering.

Know your purpose

Have a purpose for every table or graph you create. Choose the table/graph to serve the purpose

Pie charts are best to use when you are trying to compare parts of whole.

Bar graphs are used to compare things b/w different groups.

Label your data

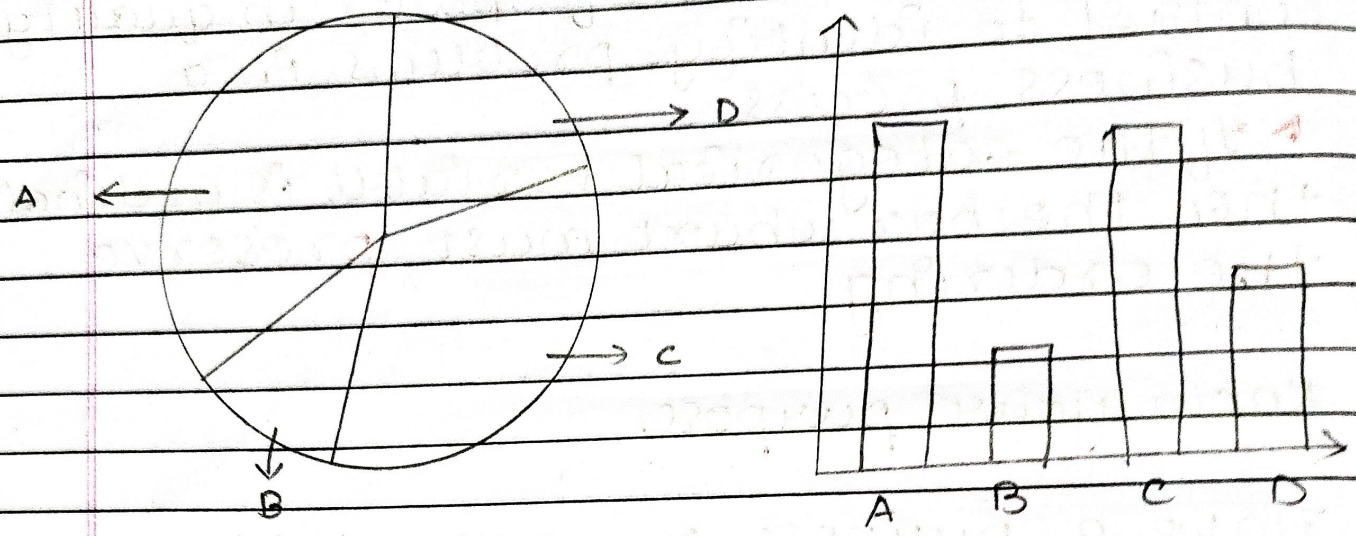
Label your chart to show the categories & indicate whether some have been combined or omitted.

Name the bars in the bar chart and slices in the pie chart.

If you have omitted some of the cases make sure the label of plot defines the collection that is summarized.

Pareto - ascending / descending
↳ depends on situation

Distribution of grades.



many categories

A bar chart or pie chart with too many categories might conceal the more important categories. In some case, grouping other categories together might be done.

Area Principle

- Displays of data must obey a fundamental rule called the area principle.
- The area principle says that the area occupied by a part of the graph should correspond to the amount of data it represents.
- Violations of the area principle are a common way to mislead with statistics.

Misleading graphs.

1. **Violating area principle.**
   - Decorated graphics: charts decorated to attract attention often violate area principle.
   - (No baseline and the chart shows bottles on top of labelled boxes of various sizes and shapes.)

2. **Truncated graphs.**
   - Another common violation is when the baseline of a bar chart is not at zero.

3. **Round-off errors.**
   - Important to check for round-off errors.
   - When table entries are % or proportions, the total may sum to a value slightly different from 100% or 1. This might result in a pie chart where the total does not add up.

4. **Indicating a y-axis break.**

Know the purpose

↓

Label the data

↓

multiple category

↓

Area principle

↓

misleading graphs

↓

violating area principle

truncated graphs

indicating a y-axis break

round-off errors

Summarizing categorical data

Graphical summaries of categorical data: bar chart and pie chart

Numbers that are used to describe data sets are called descriptive measures.

Descriptive measures that indicate where the center or most typical value of a data set lies are called measures of central tendency.

→ mean, median, mode

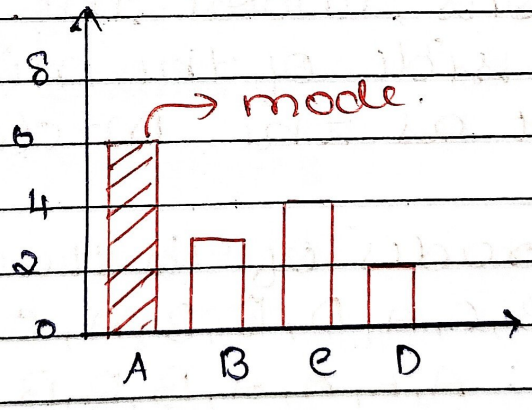
mode: The mode of the categorical variable is the most common category, the category with the highest frequency.
→ The longest bar in a bar chart
→ The widest slice in a pie chart
→ In a pareto chart, the mode is the first category

Eg: A, A, B, C, A, D, A, B, C, C, A, B, C, D, A.

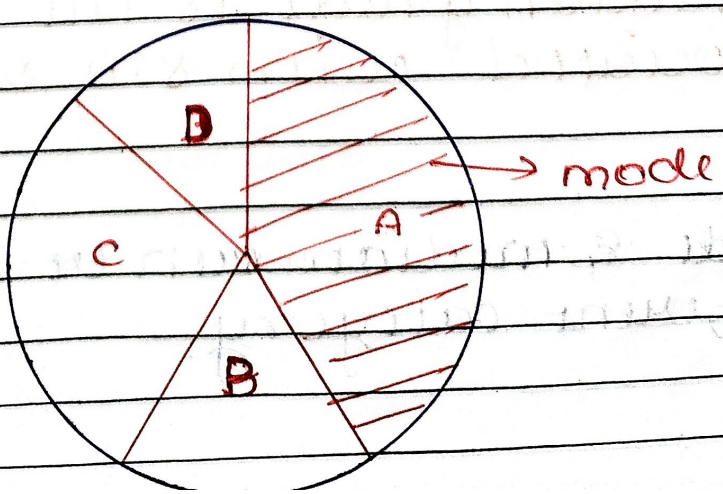
Frequency Distribution Table.

category	Tally	Frequency.
A	𝍸 I	6 → mode.
B	III	3
C	IIII	4
D	II	2.
		15.

Bar graph.



Pie chart.



Bimodal and multimodal data

If two or more categories tie for the highest frequency, the data are said to be bimodal (in case of two) or multimodal (more the two).

Median

Ordinal data offer another summary, the median, that is not available unless the data can be put into order.

The median of an ordinal variable is the category of the middle observation of the sorted values.

▶ If there are even number of observations, choose the category on either side of the middle of the sorted list as the median.

Eg: Grades of 15 students.

A, B, B, C, A, D, B, B, A, C, B, B, C, D, A

ordered: A, A, A, A, B, B, B, B, B, B, C, C, C, D, D

median grade is the category associated with 8th obs which is 'B'.

mode & median can be same or different category.

WEEK-3

DESCRIBING NUMERICAL DATA

(Both Discrete & con)

Frequency tables:
- Organizing numerical data.

Graphical Summaries:
- Histograms
- Stem-and-leaf diagram.

Numerical Summaries.
- measures of central tendency
- measures of dispersion
- Percentiles.

Organizing discrete data (single value)

If the data set contains only a relatively small number of distinct or different values, it is convenient to represent it in a frequency table.

Each class represents a distinct value (single value) along with its frequency of occurrence.

Eg:
Suppose the dataset reports the no of people in a household. The following data is the response of 15 individuals
2, 1, 3, 4, 5, 2, 3, 3, 3, 4, 4, 1, 2, 3, 4

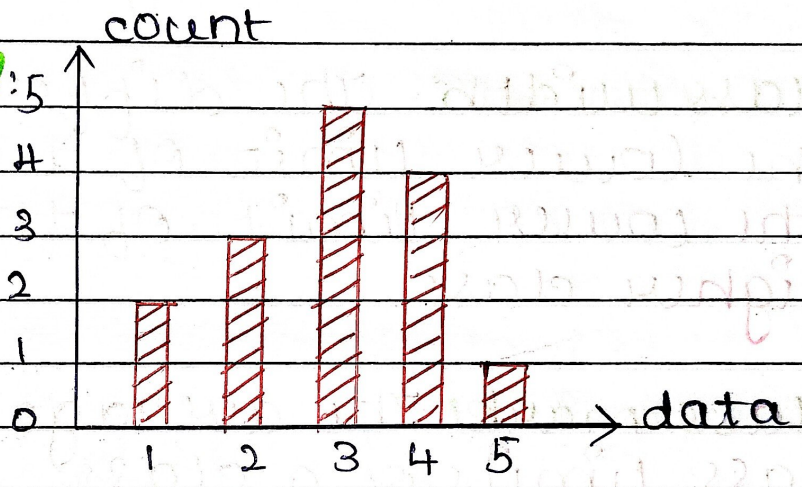
92.5 138 126 125.

each one of distinct value as a category.

Frequency table:

Value	Tally	Frequency	RF
1	II	2	0.133 (2/15)
2	III	3	0.200 (3/15)
3	IIII	5	0.333 (5/15)
4	IIII	4	0.266 (4/15)
5	I	1	0.066 (1/15)

Bar Graph:



Organizing continuous data

1. Organize the data into a no. of classes to make the data understandable
   - No. of classes: The appropriate no is a subjective choice, the rule of thumb is to have between 5 and 20 classes.
   - Each observation should belong to some class and no observation should belong to more than one class.
   - It is common, although not essential, to choose class intervals of equal length

Page No.
Date

Some new terms

1. **Lower class limit**: The smallest values that could go in a class.

2. **Upper class limit**: The largest value that could go in a class.

3. **Class width**: The difference b/w the lower limit of a class & the lower limit of the next-higher class.

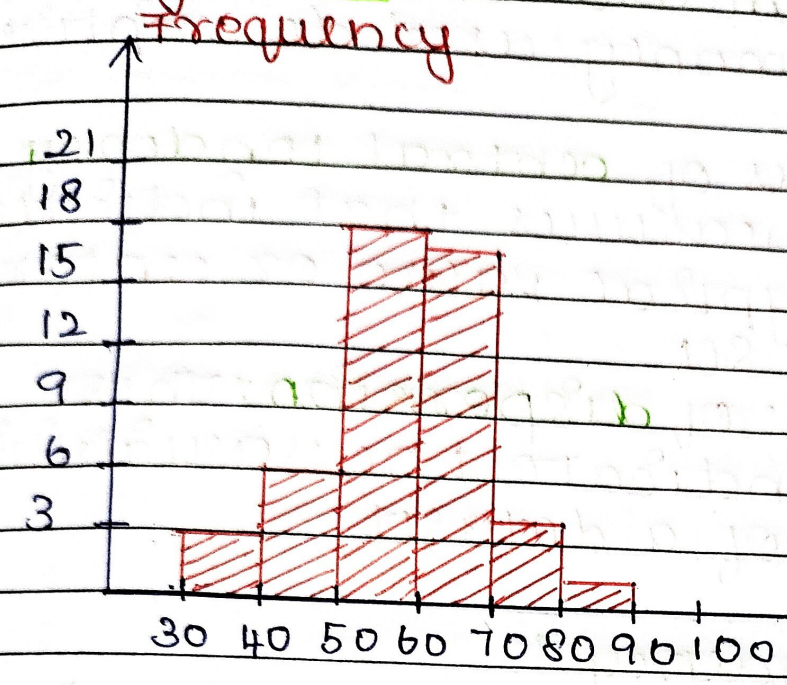
4. **Class mark**: The average of the two class limits of a class.

5. Any class interval contains its **left end** but not its right-end.

Frequency table:

Graphical summary:

histogram.



30-40	3
40-50	6
50-60	18
60-70	17
70-80	4
80-90	2

Stem-and-leaf diagram

In a stem-and-leaf diagram (stemplot), each observation is separated into two parts, namely, a stem - of all but the rightmost digit - and a ~~leaf~~ leaf the rightmost digit.

Eg: 15

Stem ↓ ↳ Leaf.

The two values 75, 78 is

Stem ~~&~~ leaf.

7 | 5, 8

Eg 1: The following are the ages, to the nearest year, of 11 patients admitted in a certain hospital: 15, 22, 29, 36, 31, 23, 45, 10, 25, 28, 48.

Stem	Leaf
1	0, 5
2	2, 3, 5, 8, 9
3	1, 6

Numerical summary
Descriptive measures

▶ most commonly used descriptive measures:

1. measures of **central tendency**: These are measures that indicate the most typical value or center of a data set.
2. measures of **dispersion**: These measures indicate the variability or spread of a dataset.

Central tendency

1. **mean** (most common measure) (average)

The mean of a data set is the sum of the observations divided by the no. of observations.

for discrete observations:

for ungrouped data:

Sample mean: $\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$

Population mean: $\mu = \frac{x_1 + x_2 + \dots x_N}{N}$

Mean for grouped data: discrete single value data

▶ The following data is the response from 15 individuals.
2, 1, 3, 4, 5, 2, 3, 3, 3, 4, 4, 1, 2, 3, 4

▶ $\bar{x} = \frac{f_1x_1 + f_2x_2 + \dots + f_nx_n}{n}$

Value(x_i)	Tally mark	Frequency(f_i)	f_ix_i
1	||	2	2
2	|||	3	6
3	𝍸	5	15
4	𝍸|	4	16
5	|	1	5
Total		15	44

Mean $= \frac{44}{15} = 2.93$

MORE VIDEOS

mean is affected by the outliers in data.

Mean for grouped data: continuous data

- $\bar{x} = \frac{f_1 m_1 + f_2 m_2 + \ldots + f_n m_n}{n}$

Class interval	Tally mark	Frequency(f_i)	Mid point(m_i)	$f_i m_i$
30-40	III	3	35	105
40-50	IIII I	6	45	270
50-60	IIII IIII IIII III	18	55	990
60-70	IIII IIII IIII II	17	65	1105
70-80	IIII	4	75	300
80-90	II	2	85	170
Total		50		2940

- Average $= \frac{2940}{50} = 58.8$.
- 58.8 is an approximate and not exact value of the mean

ln + constant.

$\frac{50}{10}^{5}$.

3. median

The median of a data set is the middle value in its ordered list. (less affected by outliers)

Steps to obtain median

Arrange the data in increasing order. Let n be the total number of observations in the dataset.

1. If the number of observations is odd, then the median is the observation exactly in the middle of the ordered list, i.e. $\frac{n+1}{2}$ observation
2. If the number of obsevations is even, then the median is the mean of the two middle observations in the ordered list, i.e. mean of $\frac{n}{2}$ and $\frac{n}{2}+1$ observation

Eg: 2, 105, 5, 7, 6, 3

① 2, 3, 5, 6, 7, 105.

② n = 6., even.

③ avg ($\frac{n}{2}$ & $\frac{n+1}{2}$) obs

3rd, 4th.

$\frac{5+6}{2} = 5.5$.

mean is affected by the outliers in data.

manipulating data:

Adding a constant.

$\bar{y} = \bar{x} + c$

new mean = old mean + constant.

multiplying a constant:

$\bar{y} = \bar{x}\, c$

$\frac{50}{10}^{5}$.

manipulating data set:

Adding a constant:

$y_i = x_i + c$ where c is a constant.

new median = old median + c

because here no. of observation does not change

multiplying a constant:

$y_i = x_i \cdot c$. where c is a constant.

new median = old median × c.

3. **mode:**

▶ If no value occurs more than once, then the dataset has no mode.

▶ Else, the value that occurs with greatest frequency is a mode of the data set.

(least affected by outliers).

manipulating data set:

Adding a constant.

$y_i = x_i + c$, where c is constant

new mode = old mode + c.

multiplying a constant.

$y_i = x_i \cdot c$

new mode = old mode × c.

Eg. 1. 2, 12, 5, 7, 6, 7, 3.

⇒ 7 is mode.

2. 2, 105, 5, 7, 6, 3

⇒ no mode

outliers - very odd, extreme values.

Page
Date

Dispersion:

1 Range:

- The Range of a data set is the difference between its largest and smallest value.

Range = max - min.

- Sensitive to outliers.
- Takes into account of max & min value.

2 variance:

- One way of measuring the variability of a data set is to consider the deviations of the data values from a central value.

Deviations: $(x_1 - \bar{x})(x_2 - \bar{x})(x_3 - \bar{x}) \ldots (x_n - \bar{x})$.

- Adding deviations = 0.
- Population variance: $\sigma^2 = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 \ldots + (x_N - \mu)^2}{N}$
- Sample variance: $s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 \ldots + (x_n - \bar{x})^2}{n - 1}$

↳ beyond scope.

- Takes into account of all values.

Adding a constant:

$y_i = x_i + c$. where c is a constant then,

new variance = old variance.

Multiplying a constant:

$y_i = x_i * c$,

new variance = $c^2 \times$ old variance.

	Data	Deviation from mean $(x_i - \bar{x})$	Squared deviations $(x_i - \bar{x})^2$
1	68	9	81
2	79	20	400
3	38	-21	441
4	68	9	81
5	35	-24	576
6	70	11	121
7	61	2	4
8	47	-12	144
9	58	-1	1
10	66	-7	49
Total	**590**	**0**	**1898**

RE VIDEOS

opulation variance $= \frac{1898}{10} = 189.8$

2. Sample variance $= \frac{1898}{9} = 210.88$

14.81 / 32.19 · Example

3. Standard deviation.

The quantity $$s = \sqrt{\frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_n-\bar{x})^2}{n-1}}$$

which is the square root of sample variance is the sample standard deviation.

Units of standard deviation: The sample standard deviation is measured in the same units as the original data.

Adding a constant:

$y_i = x_i + c$, where c is a const.

new variance = old variance.

multiplying a constant:

$y_i = x_i c$, where c is const.

new variance = c^2 x old variance.

Standard deviation and variance is affected by outliers.

4. Percentiles.

The sample 100p percentile is the data value have the property that at least 100(1-p) percent of the data values are greater than or equal to it.

If two data values satisfy this contd, then the sample 100p percentile is the arithmetic avg of these values.

median is the 50th percentile

Eg: Let n = 10

Arranging data in ascending order

35, 38, 47, 58, 61, 66, 68, 68, 70, 79

p	np	
0.1	1	(35+38)/2 = 36.5
0.25	2.5	47
0.50	5	(61+66)/2 = 63.5
0.75	7.5	68
1.00	10	79.

→ Quartiles:

The sample 25th percentile is called the first quartile. The sample 50th percentile is the 2nd quartile. The sample 75th percentile is called the 3rd quartile.

The five no summary:

- minimum
- 1st quartile / lower quartile
- 2nd quartile / median
- 3rd quartile / upper quartile
- maximum

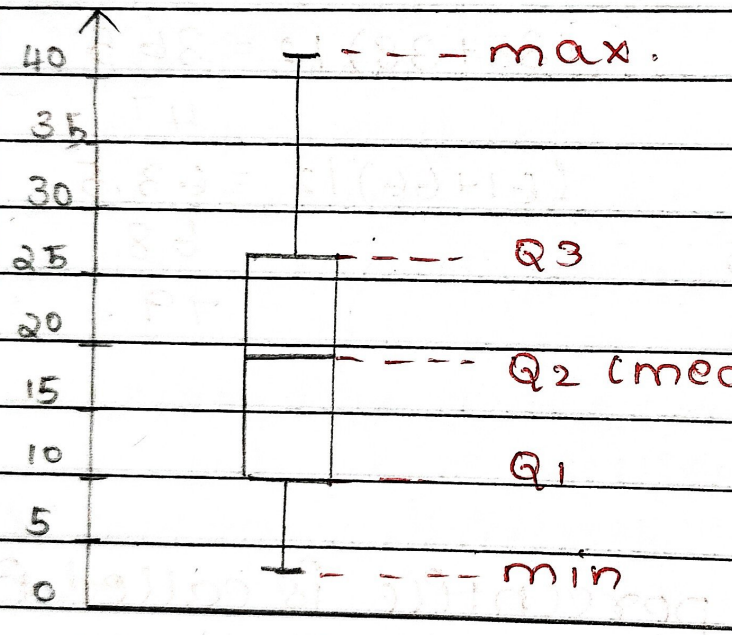
The interquartile Range (IQR):

The IQR is the diff b/w first and third quartile

$$IQR = Q3 - Q1$$

The Box plot (Candle stick plot).

Eg. min = 3, $Q_1 = 10$, $Q_2 = 19$, $Q_3 = 27$, max = 42.



Outliers.

$Q3 + 1.5IQR < outliers < Q1 - 1.5IQR.$

WEEK-4

Page No.
Date

Association between two variables -

Review (Two way contingency) ↳ categorical.
(Scatter plots, correlation) ↳ numerical.
↳ cat - num.

Association b/w two **categorical variables**.

To understand the association b/w 2 categorical variables, Learn how to construct two-way contingency table, Learn concept of relative row/column frequencies & how to use them to determine whether there is an association b/w categorical data.

Example 1: **Gender vs smartphone**.
a group of 100 college going children were surveyed about whether they owned a smartphone or not.

▶ The category variables in this example are:

Gender: Male, Female (2 categories) - Nominal variable.

Own a smartphone: Yes, No (2 categories) - Nominal variable.

Summary Statistics.

1. There are 44 female and 56 male stud
2. 76 students owned a smartphone, 24 did not own.
3. 34 female students owned a smartphone
4. 42 male students owned a smartphone

Contingency table (two-way table)

	Own a smartphone		
Gender	No	Yes	Row total
Female	10	34	44
Male	14	42	56
Column tot	24	76	100

To first construct the contingency table, consider one variable as row and another variable as column.

Example 2: **Income Vs smartphone:**
a group of 100 randomly picked individuals were surveyed about whether they owned a smartphone or not.

▶ The category variables in this example are:

Income: Low, medium, High (3 categories) - Ordinal variable.

Own a smartphone: yes, No (2 categories) - Nominal variable.

Summary statistics:

1. There are 20 High, 66 medium, 14 low
2. 62 owned a smartphone, 38 did not own
3. 18 high income owned smartphone, 39 medium income owned smartphone, 5 low income owned smartphone

Contingency table (two way table)

	Own a smartphone		
Income	No	Yes	Row tot
High	2	18	20
medium	27	39	66
Low	9	5	14
Column tot	38	62	100

Note:

1. Organize bivariate categorical data into a two-way table - contingency table.
2. If data is ordinal, maintain order of the variable in the table.

Relative frequencies:

↳ Row Relative

↳ Column Relative

Row Relative: Divide each cell frequency in a row by its row total.

	Own a smartphone		
Income	No	Yes	Row tot
High	2/20 (0.1)	18/20 (0.9)	20
medium	27/66 (0.40)	39/66 (0.59)	66
Low	9/14 (0.64)	5/14 (0.35)	14
Column tot	38 (0.38)	62 (0.62)	100

Answers:
what proportion of total participants own a smart phone?
what proportion of high income participants own a smart phone?

Column Relative %: Divide each cell freq. in a column by its column total.

	own a smartphone		
Income	No	Yes	Row tot
High	2/38	18/62	20/100
medium	27/38	39/62	66/100
Low	9/38	5/62	14/100
column tot	38	62	100

Answers:
what proportion of total participants have ~~no~~ high income?
what proportion of smart phone owners have high income.

Association:
what do we mean by stating two variables are associated?
Knowing information about one variable provides information about the other variable.

▶ If the row relative / column relative frequency are the **same** for all rows / columns then we say that the two variables are **not associated**.

If the row / column relative r are different then the two variables are associated with each other.

Stacked bar chart: Represents the count for a particular category. In addition, each bar is further ~~do~~ broken down into smaller segments, with each ~~assignment~~ segment representing the % of that particular category within the segment.

Also known as segmented bar chart.

Association blw two **numerical variables**:

To understand the association blw 2 numerical variables, learn how to construct scatter plots & interpret association in scatter plots, summarize association with a line, correlation matrix

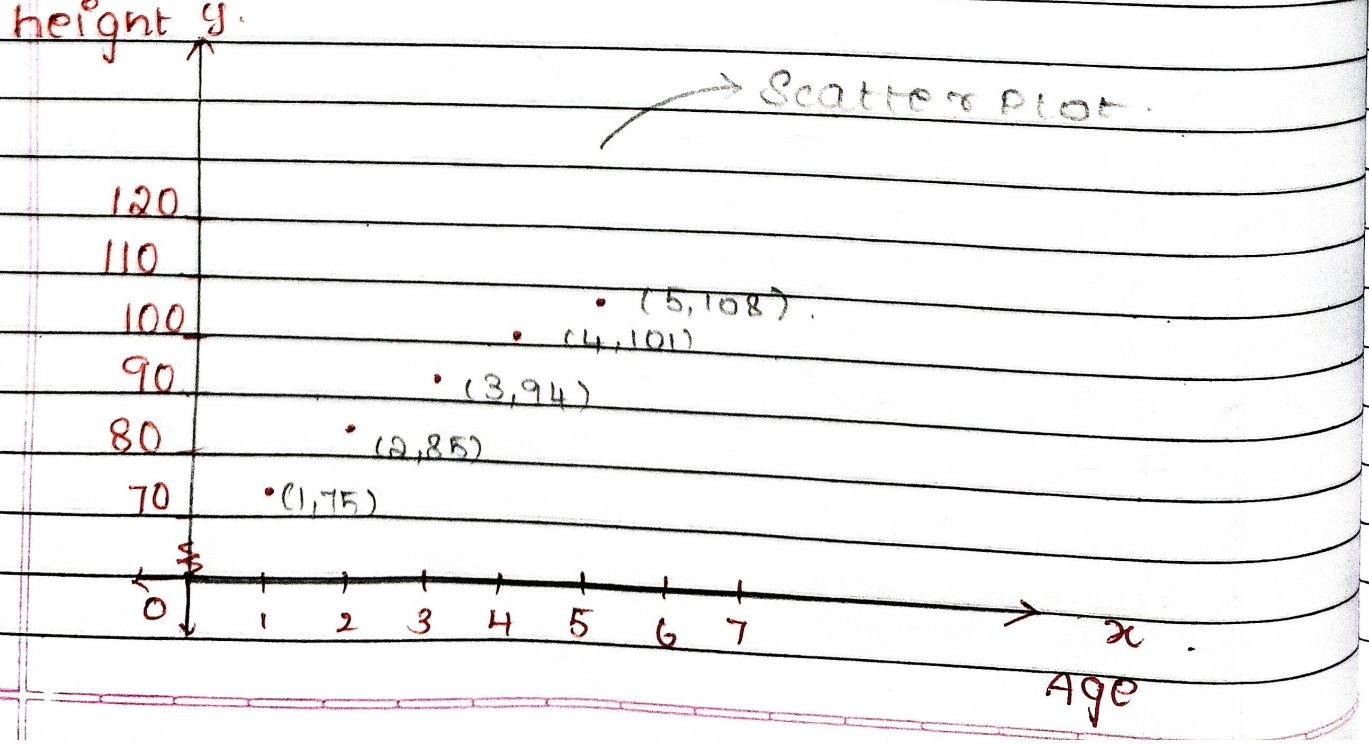
Scatter Plot: a graph that displays pairs of values as points on a 2-D plane.

Example 1: Age vs Height.

Age(yrs)	height(cm)
1	75
2	85
3	94
4	101
5	108.

As ordered pairs:
(1,75) (2,85) (3,94), (4,101), (5,108).

x-axis - independent
y-axis - dependent.

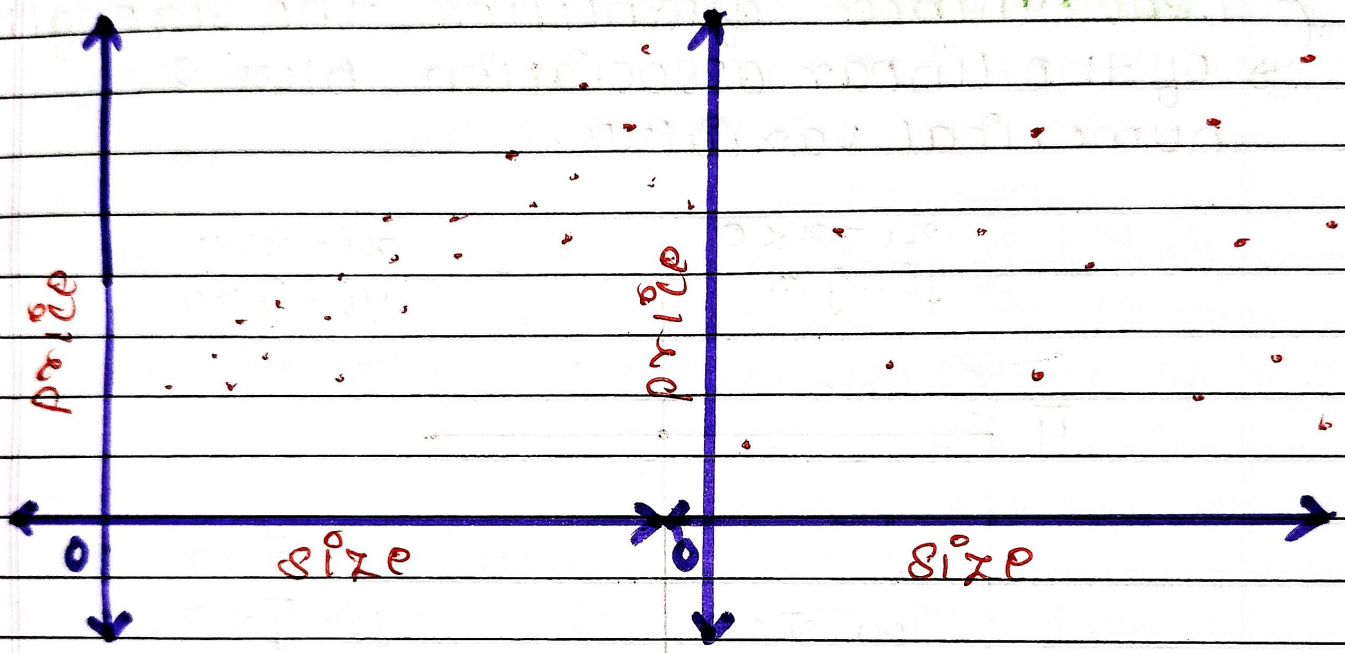


Visual test for association:

In other words, if i know about the x-value, can i use it to say something about the y-value (some pattern).

if pattern - associated, no - not associated.

Example:



(response variable)
(explanatory variable).

Describing association with scatterplots:

1. Direction - Does the pattern trend up or down or both?

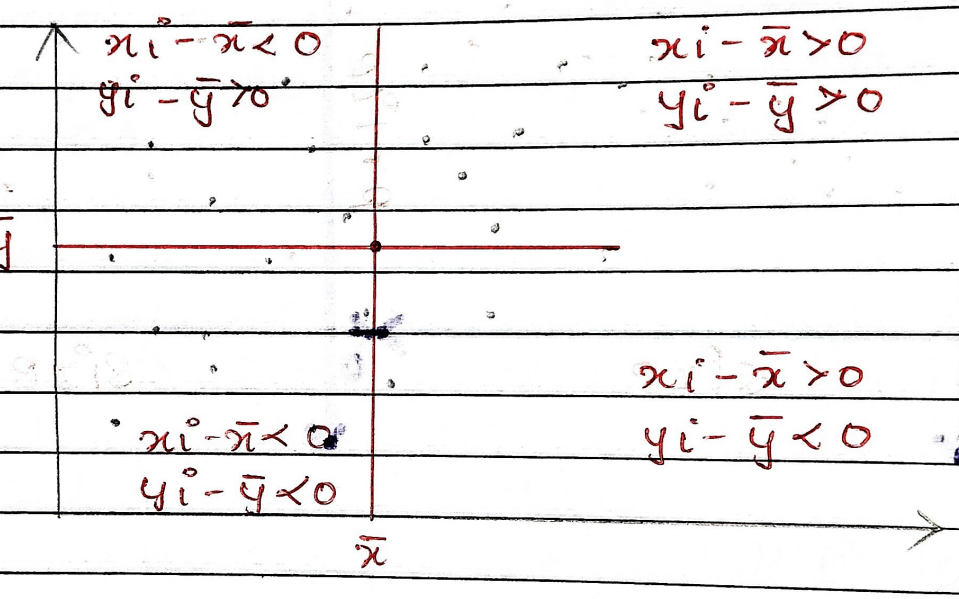
2. Curvature - linear or curve

3. variation - are the points tightly clustered along pattern / spreaded

4. Outliers: something unexpected.

measures of association:
How do we measure the strength of association b/w 2 variable?
1. Covariance } linear association
2. Correlation }

1. **Covariance** - quantifies the strength of the linear association b/w 2 numerical variables.



Example 1:

Age (yrs)	height (cm)	Devia. of x	Devia. of y
x	y	$x_i - \bar{x}$	$y_i - \bar{y}$
1	75	-2 } <0	-17.6 } <0
2	85	-1 }	-7.6 }
3	94	0	1.4 }
4	101	1 } >0	8.4 } >0
5	108	2 }	15.4 }

Key observation:

- large / small values of x – large/small value of y → the signs of deviations $(x_i - \bar{x})$ & $(y_i - \bar{y})$ will also be same.
- large/small values of x – small/large value of y → signs of deviations will also be different.

Population covariance: $cov(x,y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{N}$

Sample covariance: $cov(x,y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$

units of covariance:

- The size of the covariance, however, is difficult to interpret because the covariance has units.
- The units of the covariance are those of the x-variable times of those of the y-variable.

2. Correlation: $r = \frac{cov(x,y)}{S_x \cdot S_y}$

→ a more easily interpreted measure of linear association b/w two numerical variable

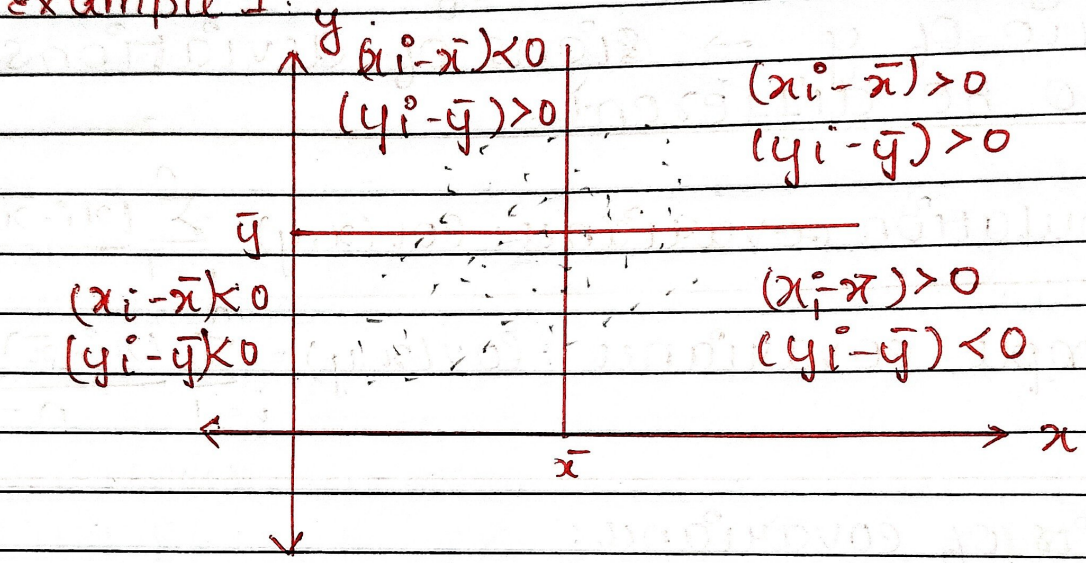
→ It is derived from covariance.

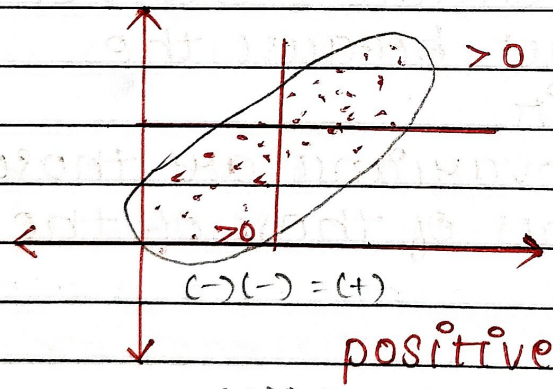
$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

units of correlation: (unitless)

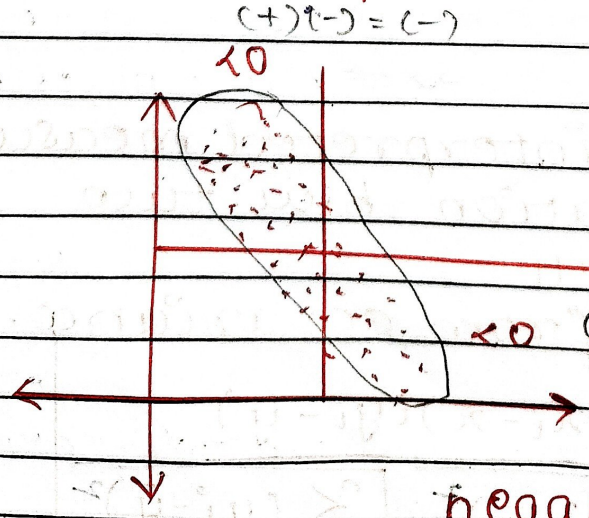
The units of standard deviations cancel out the units of covariance. correlation measure always lies between -1 and +1 ($-1 \le r \le +1$)

Example 1.

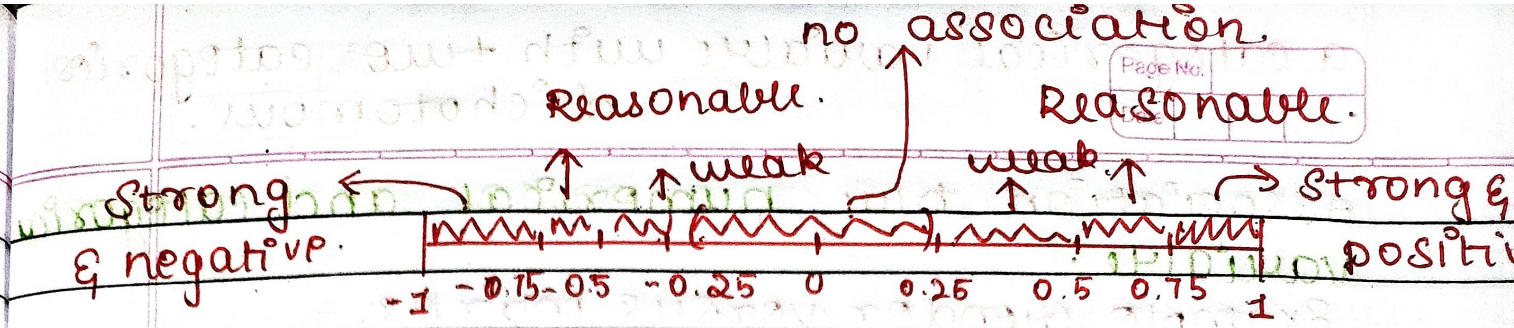




positive covariance.



negative covariance.



Summarize the linear association b/w two variables using the equation of a line.

R^2 - Goodness of fit measure.

$0 \leq R^2 \leq 1$

not good fit. | good fit

→ Square of correlation coefficient.

$y = mx + c$

→ explanatory

Response

Slope → $m > 0$ → +ve

→ $m < 0$ → -ve.

a categorical variable with ~~two~~ categories dichotomous.

Association b/w numerical and categorical variable:

Example: Gender versus marks.

Point - Bi-serial correlation coefficient.

▶ Let X be a numerical variable and Y be a categorical variable with two categories.

1. Group the data into two sets based on the value of the dichotomous variable Y. That is, assume that the value of Y is either 0 or 1.

2. ~~Calcu~~late the mean values of two groups. Let $\bar{Y}_0$ and $\bar{Y}_1$ be the mean values of groups with Y=0 and Y=1.

3. Let p_0 and p_1 be the proportion of observations in a group with Y=0 & Y=1 & S_x be the standard deviation of random variable X.

$$\therefore r_{pb} = \left(\frac{\bar{Y}_0 - \bar{Y}_1}{S_x}\right)\sqrt{p_0 p_1}$$