Akash Kumar

Statistics

Data:

Data are fact and figure that are collected analyzed and summerized for presentation and interpretation.

Statistics is ast of learning from data

- Descriptive statistics:

It is part of statistics that is concerned with discription and summerization of data In this we explore data for purpose of analysis.

- Inference statistics:

It is concerned with drawing conclusion from data.

☐ Population and Sample:

- Population:

Population is total collection of all elements that we are intrested in is called population.

- Sample:

The subgroup of the population that will be studied in detail is called sample.

☐ Structure and unstructured data:

- Structured data:

These are data which are organised

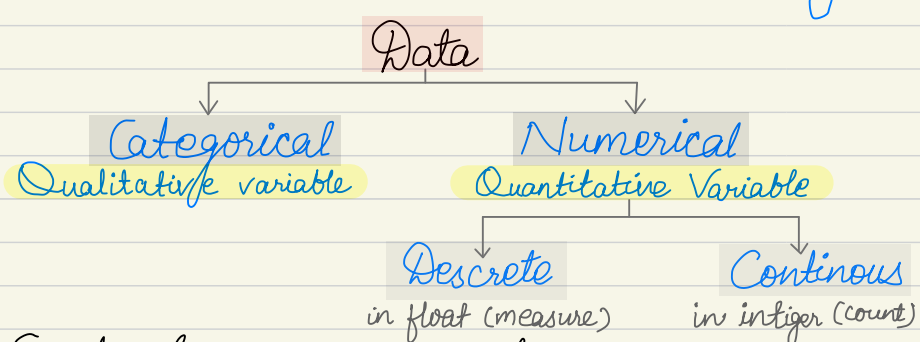
in predefined fashion.

- Unstructured data :
  These are data which are not organised in pre-defined fashion or lack data model.

☐ Variable and cases :

- Cases : It is a unit for which data is collected.
- Variable : It is characteristic that varies across all units.

Eg : case (each student)
Variable (name, marks, board, gender...)



☐ Scale of measurement :-

→ Nominal, } Categorical Data
→ Ordinal, }
→ Interval, and } Numerical Data
→ Ratio

- Nominal scale:
  When data is consist of labels or name as characteristic of observation is known as nominal scale.

  Eg: Name, Board, Gender, Blood grp, etc.
  Weather → comfertable, ok or uncomfertable

→ Some nominal variable can be numerically coaded. (like M/F = 0/1)
→ It doesn't have any order.

- Ordinal scale:
  When data exhibits property of nominal data but here rank are meaningful, this scale of measurement is considered as ordinal scale.

  Eg: Excellent, good or poor.
  Cold, warm or hot.

→ In this diff. b/w Excellent to good may not equal to diff. b/w good to poor.

- Interval scale of measurement:
  It has all the property of ordinal data and in this interval b/w values is expressed in term of a fixed unit of measure, than scale of measure is called interval scale.

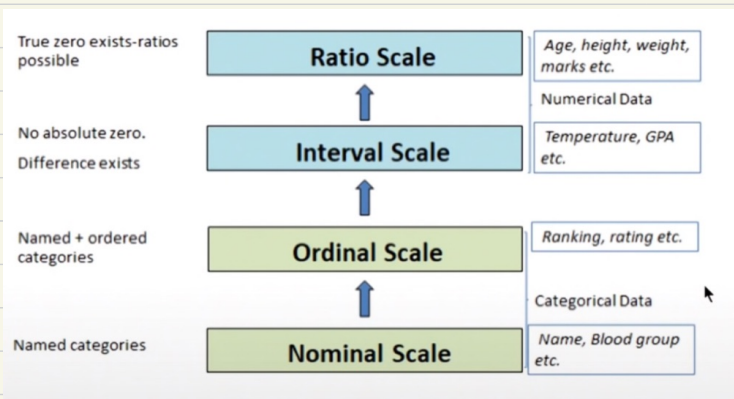
→ Interval data are always numeric. can find diff. b/w 2 value.

Eg → Temperature 40°C, 20°C, etc.

Here Ratio has no meaning and you, can't tell 40°C is twice hot as 20°C.

- Ratio scale:
  It has all property of interval scale and here ratio of 2 scale are meaningful, then scale of measurement is called ratio scale.

Eg → Height, marks, Run, weight, etc.



- Categorical data:
  - Frequency distribution:
    Frequency distribution of qualitative data is listing of distinct value and their frequency.

○ Relative frequency : The ratio of frequency to total no. of observation. It is generally used to compare 2 data set.

Eg:- AAABBCCCCD

Category	tally	Freq	RF
A	III	3	.3
B	II	2	.2
C	IIII	4	.4
D	I	1	.1
	Total	10	

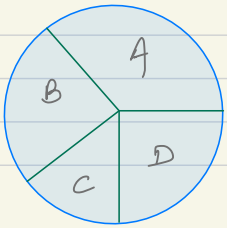
AAAAABCCCCCDDD

Category	tally	Freq	RF
A	𝍸	5	0.3
B	I	1	0.06
C	𝍸 I	6	0.4
D	III	3	0.2
	Total	15	

□ Charts of categorical data :

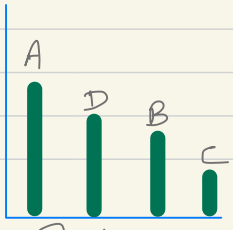
Barchart and piechart are 2 most common display of categorical variable.

- In pie chart frequency convert in 360° angle. We use pie chart to know share of each cat data.
- Bar display distinct value of qualitative data on horizontal axis with relative frequency. We use bar chart to compare each category to other.
- When barchart have shorted frequency it is called pareto chart.
- If categorical variable is ordinal than chart must be in order (Pareto chart)



Pie chart

Barchart

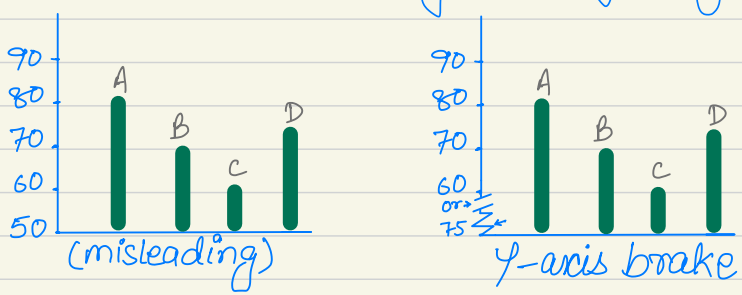


Pareto chart

○ Misleading graph: voilate area principle.

→ Area principle say that area occupy by part of graph should corresponf to amount of data represented. (Diff. bar should not be of diff. width.)

→ Trunkated graph (baseline of graph not start with zero is not good practice. You may use trunkated graph by using y-axis brake.



☐ Measure of central tendency:

It is of 2 type in catagorical data:

Median,
Mode.

○ Mode: Mode of of a categorical column is most common category.

Eg → A, B, B, A, C, A, A, C, B → Mode = A

Mode will have largest bar or largest pie.

→ If we have 2 category with highest value than it is bimodel data

→ If there is more than 2 category than it is multimodel.

○ Median : Median is middle observation of sorted value. In ordered form.

Eg → A A A B C A B A A C C C D A B C

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

median (even) ↑ ↑ median (odd) — if added

It tries to divide data in 2 half.

○ Numerical data :

→ Descrete data : It is considered as count of something.

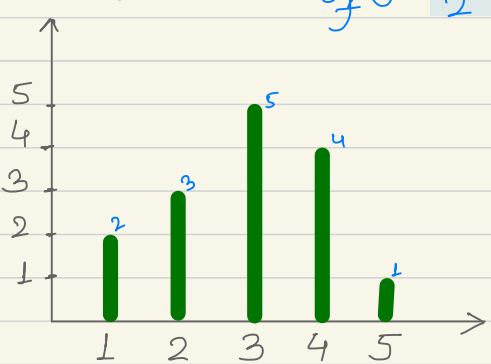
→ If descrete data is single value data then each value treated as categories.

→ By finding frequency you can know about data.

No of people in diff. house → 2, 1, 3, 4, 5, 2, 3, 3, 3, 4, 4, 1, 2, 3, 4

Converting in category :-

	1	2	3	4	5
f	2	3	5	4	1



10 — 20 ⇒ 20 − 10 = 10 (Class width)

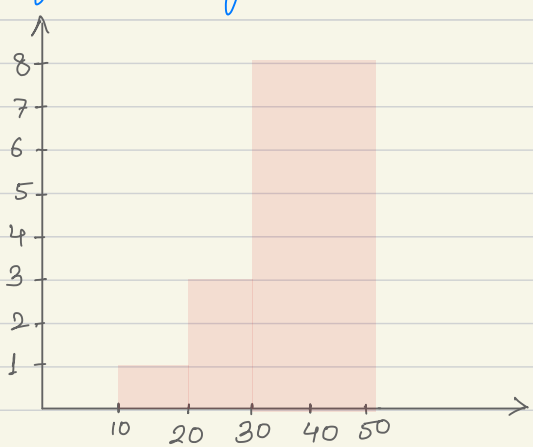
lower class, Upper class ⇒ $\frac{20+10}{2}$ = 15 (Class marks)

→ Continous data: It is considered as measure of something.
→ Here organise data in number of classes to make data understandable.
→ Each observation should belong to exactly 1 class.

- Marks is measured not counted so continous data.
Marks of 20 student = 33, 39, 30, 49, 40, 30
30, 40, 30, 11, 27, 34, 45, 48, 41, 43, 21, 47, 36, 33,

Class interval	f	Rf
10 - 20	1	.0
20 - 30	3	.05
30 - 40	8	.4
40 - 50	8	.4
Total	20	1

- Continous data is shown graphical summary through histogram.



○ Stem-and-leaf diagram:
It separate once & 10th position of dictionary. In smallest to largest.

	Stem	leaf
Eg → 75 =	7	5
75, 78 =	7	5, 8

Eg 2 → 15, 22, 29, 36, 31, 23, 45, 10, 25, 28, 48

Steam	leaf
1	0 5
2	2 3 5 8 9
3	1 6
4	5 8

○ Descreptive measure:
→ Measures of central tendency: It indicate most typical value or centre data set.

→ Measure of dispersion: These measures indicate variablity / spread in data.

□ Measures of central tendency:
It capture centre or typicalness of datasef.
→ Mean
→ Median
→ Mode

○ Mean :- It is most commonly used measure. It is average means sum of all observation divided by no of observation.

$n \rightarrow$ Sample Size $N \rightarrow$ Population Size

→ mean reffers as average.

For descrete observation:-

▷ Sample mean $= \bar{x} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n}$

▷ Population mean $= \mu = \frac{x_1 + x_2 + x_3 + \dots + x_n}{N}$

Example →

	mean
2, 12, 5, 7, 6, 7, 3	$\bar{X} = \frac{42}{7} = 7$
2, 105, 5, 7, 6, 7, 3	$\bar{X} = \frac{135}{7} = 19.285$
2, 105, 5, 7, 6, 7	$\bar{X} = \frac{128}{6} = 21.33$

○

x_i	1	2	3	4	5	Total
f_i	2	3	5	4	1	15
$x_i f_i$	2	6	15	16	5	44

$$\text{mean} = \frac{\sum_{i=n}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} = \bar{X}$$

$$\text{mean} = \frac{44}{15} = 2.93$$

○ Mean for continous data:

Class interval	f	Rf	mp	fimi
10 - 20	1	.0	15	15
20 - 30	3	.05	25	75
30 - 40	8	.4	35	280
40 - 50	8	.4	45	360
Total	20	1		$\Sigma fimi$ 730

$$Avg = \frac{730}{20} = 36.5$$

→ 36.5 is approximation not mean bcz we are not looking to data but only seeing midpoint.

→ If you add constant to every point in dataset than your new $\bar{x}$ = old $\bar{x}$ + constant.

→ If you multiply constant to every point in dataset than your new $\bar{x}$ = old $\bar{x}$ * constant.

→ Highly affected by outlier.

○ Median :-

It is of a data set is middle value of ordered list. It is another frequently used measure of central tendency. It divide dataset in top 50% & botom 50%.

In ordered list: (n = no' of observation)

— If no' of observation is odd than the data in $(n+1)/2$ th value is median.

— If no' of observation is even than the median is avg. of $(\frac{n}{2}$ and $\frac{n}{2}+1)$.

Example → 2, 9, 4, 6, 7, 8

In odd.

arrange data → 2, 4, 6, 7, 9

$n=5$ median = $\frac{5+1}{2} = 3^{rd}$

Median = 3rd element = 6

In even,

2, 4, 6, 7, 8, 9

$n=6$

median = $\frac{6}{2} = 3$, $\frac{6}{2}+1 = 4$

median = $\frac{6+7}{2} = 6.5$

- Very less affected by outlier.
- If constant is added to each point of data set, length doesn't change so new median will be old median + C.
- If constant is multiplied to each data-point. Than due to same length. median will be old median × C.

○ Mode :

It is most frequently occuring value of dataset.

- If no value occur more than 1 than there is no mode.

Eg: 1, 2, 3, 7, 7, 3, 2, 1 — mode = 7
2, 9, 3, 4, 2, 7 — no mode.

- If constant is added to each point of data set, length doesn't change so new mode will be old mode + C.
- If constant is multiplied to each data-point. Than due to same length. mode will be old mode × C.
- Don't affect by outlier.

Let's compare 2 data:

D1 → 3, 3, 3, 3, 3 D2 = 1, 2, 3, 4, 5

→ finding measure of central tendency :

	D1	D2
Mean	3	3
Median	3	3
Mode	3	—

However mean & median are same for dataset but dataset are not same.

- **Measure of Dispersion:** To describe the above difference quantatively, we use discriptive measure that indicate amount of variation, spread in data. These is called measure of dispersion/variance/Spread.

  Measure of dispersion are:
  - → Range,
  - → Variance,
  - → Standard daviation,
  - → Interquartile range.

- **Range:-** It is defined as difference b/w largest and lowest value of a dataset.

  Range = Largest − lowest

  Let's compare 2 data:

  D1 → 3,3,3,3,3 — Range = 3-3 = 0

  D2 = 1,2,3,4,5 — Range = 5-1 = 4

  D3 = 1,2,3,4,15 — Range = 15-1 = 14

  As we can se it do well in D1 & D2 but in D3 not properly

  So, Range is extremly sencitive to outlier.

- **Variance:-** In contrast to range variance takes into account all observations.
  - One way to measure variablity of dataset is consider daviation of data value from centre value. It is affected by outlier.

Eg:-

	1	2	3	4	5
	-2	-1	0	1	2

centre point = 3 = $\bar{X}$

difference from centre

Population variance :

$$\sigma^2 = \frac{(x_1-\mu)^2+(x_2-\mu)^2+\ldots+(x_N-\mu)^2}{N}$$

Sample variance :

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

Eg :-

D1.

	Data	Deviation from mean ($x_i-\bar{x}$)			Squared Deviation ($x_i-\bar{x}$)^2
1	68	68 - 59	=	9	81
2	79	79 - 59	=	20	400
3	38	38 - 59	=	-21	441
4	68	68 - 59	=	9	81
5	35	35 - 59	=	-24	576
6	70	70 - 59	=	11	121
7	61	61 - 59	=	2	4
8	47	47 - 59	=	-12	144
9	58	58 - 59	=	-1	1
10	66	66 - 59	=	-7	49
Total	590			0	1898

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

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

- If a constant is added to all data point than the variance doesn't change.
- If a constant is multiplied to all data point than new variance = old variance $\times$ (constant)^2

○ Standard deviation: It is square root of variance.

- Square root of sample variance is sample standard deviation.
- Square root of population variance is population standard deviation.

Population standard deviation:

$$\sigma = \sqrt{\frac{(x_1-\mu)^2+(x_2-\mu)^2+\ldots+(x_N-\mu)^2}{N}}$$

Sample standard deviation:

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

D1. Population std. deviation = $\sqrt{189.8}$

Sample std. deviation = $\sqrt{210.88}$

Unit of standard deviation:

u^2 → Variance is recorded as unit^2

u → Standard deviation by root convert it back

- If a constant is added to all data point than the standard deviation doesn't change.
- If a constant is multiplied to all data point than new ST deviation = old ST deviation × constant.
- It is affected by outlier.

○ Percentile:- Percentile indicates the percent of distribution of data.

- To find percentile:-
  - Arrange data in ascending order.
  - Divide data into half if n is even we get 'integer' if it is odd we don't get integer. Than determine smallest integer greater than np. (np = middle of that part)
  - It is value in that position of 100p percentile.
  - But if np is integer in case of even. than take avg of np and np+1.

Example:- Arranged data:
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
0.5	5	(61+66)/2	= 63.5
0.75	7.5		= 68
1	10		= 79

→ Quartile:
- Sample 25th percentile is called 1st quartile.
- The sample 50th percentile is called median / Second quartile.
- The sample 75th percentile is called 3rd quartile.

The five number summary:
- Minimum
- 1st quartile (lower)
- 2nd quartile (median)
- 3rd quartile (Upper)
- Maximum

→ Interquartile range:

The interquartile range (IQR) is difference b/w first and third quartile.

$$IQR = Q_3 - Q_1$$

Eg →

1st quartile, $Q1 = 49.75$
3rd quartile, $Q3 = 68$
IQR $(Q_3 - Q_1) = 18.25$

It is also measure of dispersion.

○ Contingency table: It is also called two-way frequency table, is a tabular mechanism with atleast two row & two column used in statistics to present categorical in term of frequency count.

Eg:

	Gender	No	Yes	Total
Owns phone	Male	10	34	44
	Female	14	42	56
	Total	24	76	100

nominal data

	Income (coded)	Yes	No	Total
By income	1	2	18	20
	2	27	39	66
	3	9	5	14
	Total	38	62	100

Ordinal data

→ Row relative frequencies: It is dividing each row by its row total.

Eg:

Owns phone

Gender	No	Yes	Total
Male	10/44	34/44	44
Female	14/56	42/56	56
Total	24/100	76/100	100

nominal data

⇒

Gender	No	Yes	Total
Male	22.7%	77.2%	44
Female	25.0%	75.0%	56
Total	24.0%	76.0%	100

→ Column relative frequencies: It is dividing each column by its row total.

Owns phone

Gender	No	Yes	Total
Male	10/24	34/76	44/100
Female	14/24	42/76	56/100
Total	24	76	100

nominal data

Gender	No	Yes	Total
Male	41.6%	44.7%	44.0%
Female	58.3%	55.2%	56.0%
Total	24	76	100

→ Association b/w two variable:
It is finding whether information about one variable provide information about another variable.

→ If the row or column relative frequency are same for all rows then two variable are not associated with each other.

Row related frequency

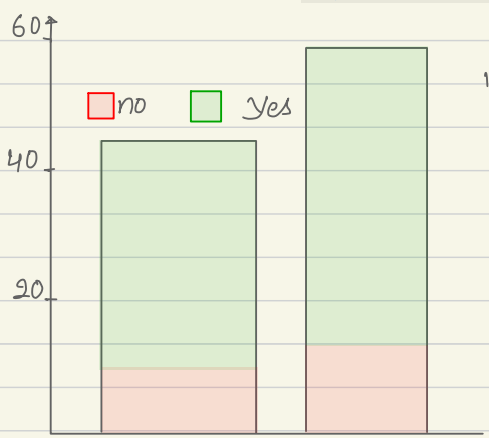
Gender	No	Yes	Total
Male	22.7%	77.2%	44
Female	25.0%	75.0%	56
Total	24.0%	76.0%	100

Both total% are for male % & female %

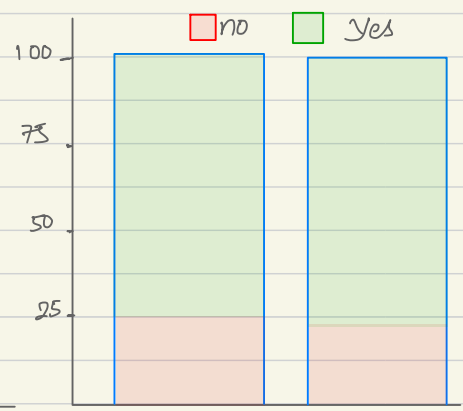
Column related frequency

Gender	No	Yes	Total
Male	41.6%	44.7%	44.0%
Female	58.3%	55.2%	56.0%
Total	24	76	100

Both male & female % are similar to total.



(Standard barchart)



(100% stacked barchart)

→ Stacked barchart :- It summarise data in form of barchart with proportion in respect to category.

→ If the row or column relative frequency are diffrent for some rows then two variable are associated with each other.

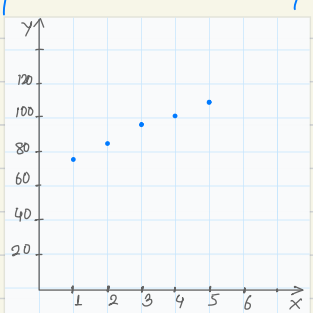
Here, in both row & column freq. are diff. from total so, Income & phone are connected.

Income level	Yes	No	total
High	10.0%	90.0%	20
Medium	40.9%	59.1%	66
Low	64.1%	35.9%	14
Total	38.0%	62.0%	100

Income level	Yes	No	total
High	5.2 %	29.0 %	20.0%
Medium	71.0 %	62.9%	66.0%
Low	23.6 %	8.1 %	14.0%
Total	38	62	100

○ Scatter plot: It is a graph that displays pair of values as points on 2-D plane.

Age	Height
1	75
2	85
3	94
4	101
5	108



You can describe association b/w variables in scatter plot by answering 4 question:

→ Direction → Does pattern Up, Down or both.

↗ ↘ ⩘⩘

→ Curvature → Is it linear or curve.

→ Variation → tightly clustered → space b/w points.

→ Outliers → Outside pattern (exceptions).

- Measure of association:
  Strength of association b/w 2 variable can be measured with: – Covariance
  – Correlation

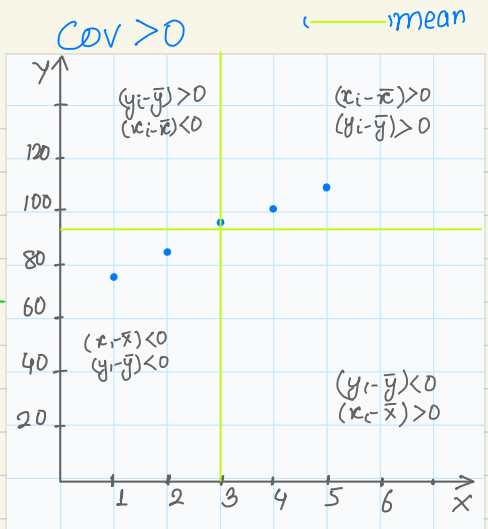
- Covariance :- It quantify strength of a linear relation b/w 2 numeric variables. (No units)

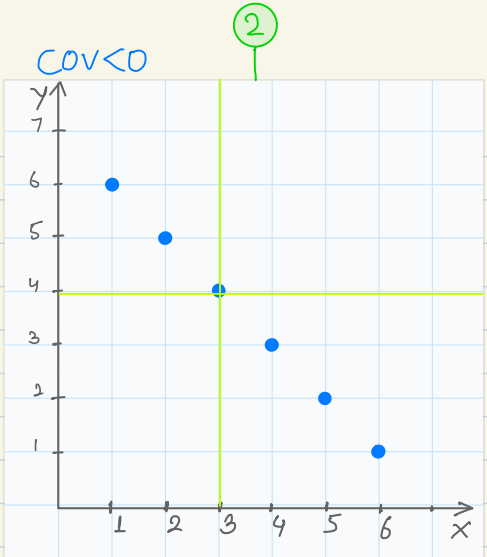
Age	Height	Deviation x $(x_i-\bar{x})$	Deviation y $(y_i-\bar{y})$	$(x_i-\bar{x})(y_i-\bar{y})$
1	75	-2	-17.6	35.2
2	85	-1	-7.6	7.6
3	94	0	1.4	0
4	101	1	8.4	8.4
5	108	2	15.4	30.8
$\bar{x}=3$	$\bar{y}=92.6$			$\Sigma=82$

(1)

Age	Height	Deviation x	Deviation y	$(x_i-\bar{x})(y_i-\bar{y})$
1	6	-2	2	-4
2	5	-1	1	-1
3	4	0	0	0
4	3	1	-1	-1
5	2	2	-2	-4
$\bar{x}=3$	$\bar{y}=4$			$\Sigma=-10$

(2)





→ When large X is associated with large Y and vice-versa the deviation sign will be same

→ When large X is associated with small Y and vice-versa the deviation sign will different..

The covariance b/w variable x & y is given by :-

- Population covariance :

$$Cov(x,y) = \frac{\sum_{i=1}^{n}(x-\bar{x})(y-\bar{y})}{N}$$

- Sample covariance :

$$Cov(x,y) = \frac{\sum_{i=1}^{n}(x-\bar{x})(y-\bar{y})}{n-1}$$

① Population covariance $= \frac{82}{5} = 16.4$ ② Pop. cov. $= \frac{-10}{5} = -2$

Sample covariance $= \frac{82}{4} = 20.5$ Sam. cov. $= \frac{-10}{4} = -2.5$

- **Correlation:** It is more easily interpreted measure of linear association b/w two numerical variable. (No units)
  - → It is derived from covariance.
  - → To find correlation b/w two variable X & Y divide covariance b/w X & Y by product of standard deviation of X & Y.

→ Correlation = $\frac{\text{Covariance}}{\text{Standard Deviation}}$

$$\rightarrow \text{Correlation} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{\text{Cov}(x,y)}{S_x S_y}$$

Age	Height	$(x_i-\bar{x})^2$ Sq. Dev. of X	$(y_i-\bar{y})^2$ Sq. Dev. of Y	$(x_i-\bar{x})(y_i-\bar{y})$
1	75	$-2^2 = 4$	$-17.6^2 = 309.76$	35.2
2	85	$-1^2 = 1$	$-7.6^2 = 57.76$	7.6
3	94	$0^2 = 0$	$1.4^2 = 1.96$	0
4	101	$1^2 = 1$	$8.4^2 = 70.56$	8.4
5	108	$2^2 = 4$	$15.4^2 = 237.16$	30.8
$\bar{X}=3$	$\bar{Y}=92.6$	$\Sigma 10$	$\Sigma = 677.2$	$\Sigma = 82$

$SD_x = 1.58$, $SD_y = 13.01$, covarriance $= 20.5$

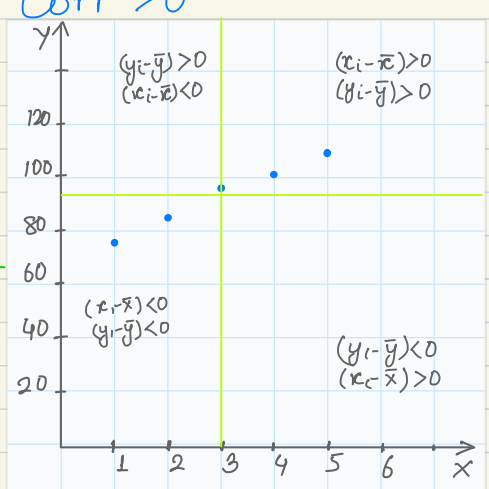
$$\text{Correlation} = \frac{\text{Cov}(x,y)}{S_x S_y} = \frac{82}{\sqrt{10 \times 677.2}} \text{ or } \frac{20.5}{1.58 \times 13.01} = 0.9964$$

Age	Height	$(x_i-\bar{x})^2$ Sq. Dev. of X	$(y_i-\bar{y})^2$ Sq. Dev. of Y	$(x_i-\bar{x})(y_i-\bar{y})$
1	6	$-2^2 = 4$	$2^2 = 4$	-4
2	5	$-1^2 = 1$	$1^2 = 1$	-1
3	4	$0^2 = 0$	$0^2 = 0$	0
4	3	$1^2 = 1$	$-1^2 = 1$	-1
5	2	$2^2 = 4$	$-2^2 = 4$	-4
$\bar{X}=3$	$\bar{y}=4$	$\Sigma = 10$	Σ 10	$\Sigma = -10$

$SD_x = 1.58$, $SD_y = 1.58$, covarriance=

$$\text{Correlation} = \frac{\text{Cov}(x,y)}{S_x S_y} = \frac{-10}{\sqrt{10} \times \sqrt{10}} \text{ or } \frac{-2.5}{1.58 \times 1.58} = -1$$

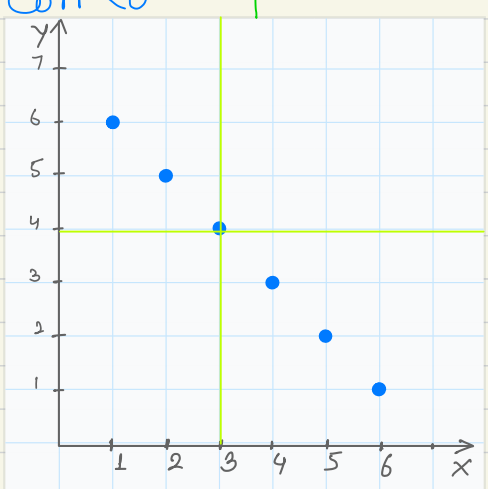
Corr >0



1

Corr <0

2



These linear relation can be summerised through line.

o Point Bi-serial Correlation coefficient :

→ Here we group our data (one numerical and one categorical column and encode over categorical column. eg→ male=0, female=1)
→ Compute mean value of numerical column in respect to encoded cat. column. eg→ mean of marks of male/Female student.
→ P_0 and P_1 is proportion of group. Eg:- For 20 student male= 12/20, female= 8/20

So, Correlation Coffecient :

$$\gamma_{pb} = \left(\frac{\bar{y}_0 - \bar{y}_1}{S_x}\right)\sqrt{P_0 P_1}$$