First, while indexing refers to columns, slicing refers to Rows.

⇒ slices refer to row by number rather than by Index.

In[34]: data[1:3]

Out[34]:

	area	pop	density
Florida	170	195	1.14
Illinois	149	128	0.8

Operations on Data In Pandas :-

NumPy has the ability to perform basic arithmetic operations (addition, subtraction, Multiplication) and with more sophisticated operations (trigonometric functions, exponential, logarithmic functions

Pandas inherits this functionality from numpy, and introduce use Ufuncs

"Computation on NumPy Arrays: Universal Functions"

Pandas Includes

For Binary Operations such as addition, Multiplication Pandas will automatically align indices when passing the object to the ufunc

For Unary Operations like negation and trigonometric functions pandas use preserve index and Column labels

(1) Ufuncs: Index Preservation :-

Pandas is designed to work with NumPy, NumPy ufunc will work on Pandas Series and Dataframe objects.

Let's us start by defining a simple series and Dataframe on which to demonstrate this,

```
In [1]: import pandas as pd
        import numpy as np
```

→ random no. array generator

```
In [2]: rng = np.random.RandomState(42)
        ser = pd.Series(rng.randint(0, 10, 4))
        ser
```

Create a one dimensional array of 4 values

ndim → No. of dimensions.
shape → size of each dimension.
size → total size of the array.

```
out[2]: 0    6
        1    3
        2    7
        3    4
        dtype: int64
```

Create 3×4 array of random integers in the interval [0, 10]

```
In [3]: df = pd.DataFrame(rng.randint(0, 10, (3, 4)),
                          columns = ['A', 'B', 'C', 'D'])
        df
```

randint → Random integers

```
out[3]:    A  B  C  D
        0  6  9  2  6
        1  7  4  3  7
        2  7  2  5  4
```

If we apply a NumPy ufunc on either of these objects. The results will be another pandas object with the indices preserved

```
In [4]: np.exp(ser)
out[4]: 0    403.428793
        1     20.085537
        2   1096.633158
        3     54.598150   dtype: float64
```

Also performs the complex calculation.

```
np.sin(df * np.pi/4)
```

Funcs : Index Alignment.

For binary operations on two Series or Dataframe object, Pandas will align indices in the process of performing the operation.

Index Alignment In Series :-

Suppose we are combining two different data sources, and find only the top three US states by area and top three US states by Population.

```
In [6]: area = pd.Series({'Alaska': 1723337,
                          'Texas': 695662, 'California': 423967},
                         name='area')
Population = pd.Series({'California': 38332521,
                        'Texas': 26448193, 'New York': 19651127},
                       name='population')
```

To compute the population density, we divide these data sources.

```
In [7]: population / area
Out[7]: Alaska        NaN
        California    90.413926
        New York      NaN
        Texas         38.018740
        dtype: float64
```

NaN → Not a Number.

↳ Anything does not have an entry is ~~form~~ is marked with NaN.

· Set Arithmetic.

The resulting array contains the union of indices of the two input arrays. By using a standard python set Arithmetic.

```
In [8]: area.index | population.index
Out[8]: Index(['Alaska', 'California', 'New York',
               'Texas'], dtype='object')
```

⇒ Index Matching is implemented by Python's built-in arithmetic expressions.

⇒ Any missing values are filled with NaN by default

```
In [9]: A = pd.Series([2,4,6], index=[0,1,2])
        B = pd.Series([1,3,5], index=[1,2,3])
        A+B
Out[9]: 0    NaN
        1    5.0
        2    9.0
        3    NaN
        dtype: float64
```

Fill Value → By using appropriate object Methods in place of the operators.

Calling A.add(B) —Equivalent to→ A+B.

```
In [10]: A.add(B, fill_value=0)
Out[10]: 0    2.0
         1    5.0
         2    9.0
         3    5.0
         dtype: float64
```

Index Alignment In Data Frame.

Alignment can be taken in both columns and indices when you are performing operations on Data Frames.

```
In [11]: A = pd.DataFrame(rng.randint(0,20,(2,2)),
                          columns = list('AB'))
```

A

out [11]:

	A	B
0	1	11
1	5	1

```
In [12]: B = pd.Dataframe(rng.randint(0,10,(3,3)),
                          columns = list('BAC'))
```

B

out [12]:

	B	A	C
0	4	0	9
1	5	8	0
2	9	2	6

```
In [13]: A+B
```

out [13]:

	A	B	C
0	1.0	15.0	NaN
1	13.0	6.0	NaN
2	NaN	NaN	NaN

```
In [14]: fill = A.stack().mean()
         A.add(B, fill_value = fill)
```

Out [14]:

	A	B	C
0	1.0	15.0	13.5
1	13.0	6.0	4.5
2	6.5	13.5	10.5

Fill with the mean of all values in A (which we compute by first stacking the rows of A)

Mapping between Python operators and Pandas Methods

Python operator	Pandas Methods (s)
+	add()
-	sub(), subtract()
*	mul(), multiply()
/	truediv(), div(), divide()
//	floordiv()
%	mod()
**	pow()

Ufuncs: operations Between Dataframe and Series:

When we are performing operations between Dataframe and Series, are similar to the operations between two-dimensional and one-dimensional Numpy Array.

We find the difference of a two-dimensional array and one of its rows.

```
In [15]: A = rng.randint (10, size=(3,4))
         A
Out[15]: array ([[3, 8, 2, 4],
                 [2, 6, 4, 8],
                 [6, 1, 3, 8]])

In [16]: A - A[0]
Out[16]: array ([[0, 0, 0, 0],
                 [-1, -2, 2, 4],
                 [3, -7, 1, 4]])
```

Subtraction between a two-dimensional array and one of its rows is applied row-wise.

In Pandas, row-wise operation by default.

```
In [17]: df = pd.DataFrame (A, columns = list('QRST'))
         df = df.iloc [0]
Out[17]:
```

	Q	R	S	T
0	0	0	0	0
1	-1	-2	2	4
2	3	-7	1	4

operate column-wise → specify axis keyword.

```
In [18]: df.subtract (df['R'], axis=0)
Out[18]:
```

	Q	R	S	T
0	-5	0	-6	-4
1	-4	0	-2	2
2	5	0	2	7

Unit II.

Types of Data - Types of Variables - Describing Data with Tables and Graphs - Describing Data with Averages - Describing Variability - Normal Distributions and Standard (z) scores.

Types of Data :-

Data is the collection of actual observations or scores in a survey or experiment which can be used for statistical Analysis.

Traditional data is a data

i.e) structured and stored database.

It is in table format containing Numeric or text Va

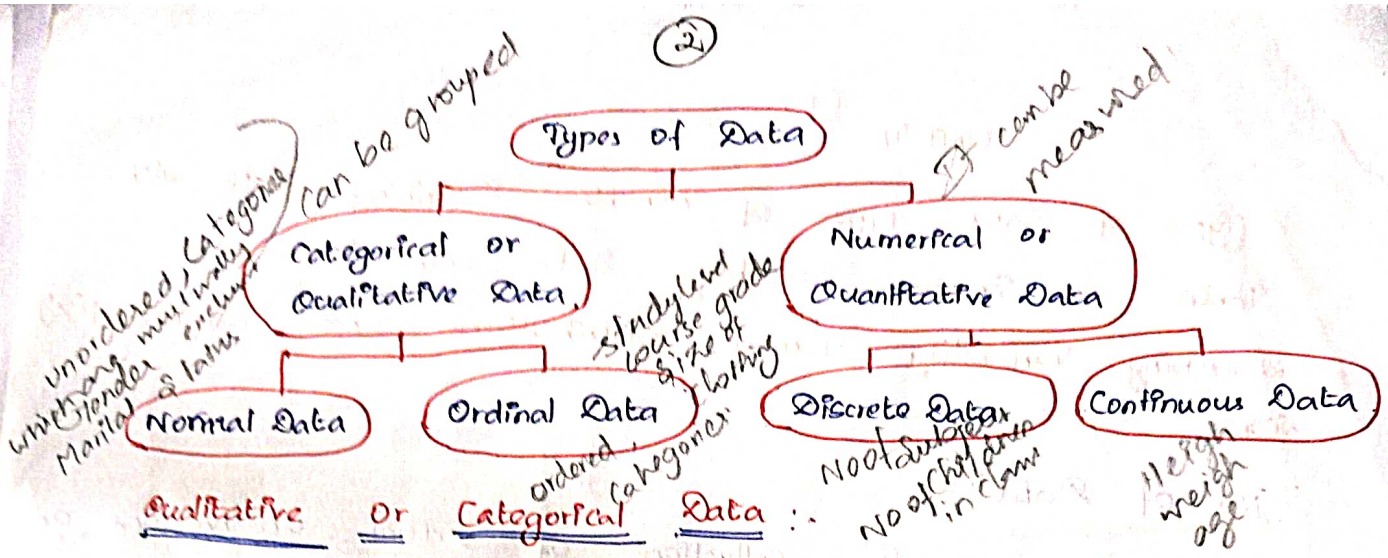
There are two types of data → Qualitative Data
→ Quantitative Data.

Further classified into Nominal, Ordinal, Discrete and Continuous

These are Collected, Analysed, Interpreted and presented.

The data is classified into mainly four Categories.

* Nominal Data
* Ordinal Data
* Discrete Data
* Continuous Data.



Qualitative or Categorical Data :-

- Qualitative Data, also known as Categorical Data,
- Describes the data that fits into the Categories.
- Qualitative data are not Numerical

Eg) Categorical variables describes features such as persons' gender, home, town etc,

some Categorical data can hold numerical value, but those values donot have a mathematical sense.

Eg) Categorical Data are birthdate, favourite sport, School postcode

Qualitative Data consist of words (Yes or No)
letters (Y or N)
Numerical Code (0 or 1)

Eg) Do you like Cartoons. Reply from few students as below.

Students	Result.
1	Y
2	Y
3	Y
4	N
5	N

eye color, Gender

Nominal Data

Nominal data is one of the type of qualitative information which helps to label the variables without providing the numerical value used to But the data can be qualitative and quantitative.

Eg) Nominal data are letters, symbols, words, gender etc

Ordinal Data:-

Ordinal data / variable is type of data that follows a natural order. is a kind qualitative data that group the variable into ordered categories.

Eg) This variable is found in surveys, finance, economics, questionaries. The ordinal data is commonly represented using a bar chart. These data are investigated and interpre

Ranked Data:-

Customer feedback, Economic status.

Ranking is the data transformation in which numerical or ordinal values are replaced by their rank when the data are sorted.

For eg) Numerical data 3.4, 5.1, 2.6, 7.3 are observed the ranks of these data items would be 2, 3, 1, & 4 respectively, Arranging in ascending order and ranking from low to high.

Ordinal data:

The categories have a natural order or rank based on some hierarchical scale, like from high to low

Rank Educational Level

* Elementary 1
* High School 2
* College 3
* Graduate 4
* Post Graduate 5

2 Quantitative or Numerical Data :-

Quantitative data represents the numerical value like how much, how often, how many.

Eg) Numerical data are height, size, weight.

Eg) Height of 5 students

Students	Height (In Inch)
1	160
2	165
3	170
4	160
5	172

2.1 Discrete Data :-

Discrete information contains only a finite number of possible values. Those values cannot be subdivided meaningfully. It can be counted in whole Numbers.

Discrete data key characteristics :

* You can count the data. It is usually units counted in whole numbers.
* The values cannot be divided into smaller pieces and add additional Meaning.
* You cannot measure the data. By nature discrete data cannot be measured.
* It has a limited number of possible values.

eg) days of the month

eg) Number of students in the class, Number of workers in a company.

Discrete Data can be represented by.

- ⇒ Bar Graph
- ⇒ Frequency table
- ⇒ Line Plot (number line).

Bar Graph is the most suitable way to represent discrete data, as finite values can be presented clearly through vertical bars.

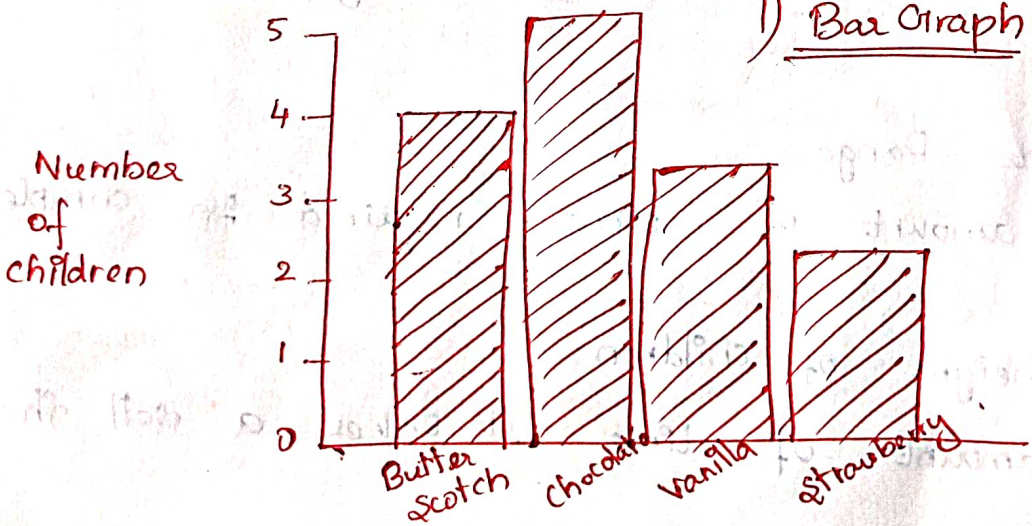
Example :-

In a survey with 14 children on their favourite Ice-Cream flavour, it was found that 4 children like butterscotch flavour, 5 children like chocolate flavour, 3 children like vanilla flavour and 2 children like strawberry flavour of ice-cream.

It is an eg, discrete data we can count the number of ~~data~~ children who like a particular flavour of ice-cream.

Graphical Representation of these data through all three modes of Representation.

1) Bar Graph.

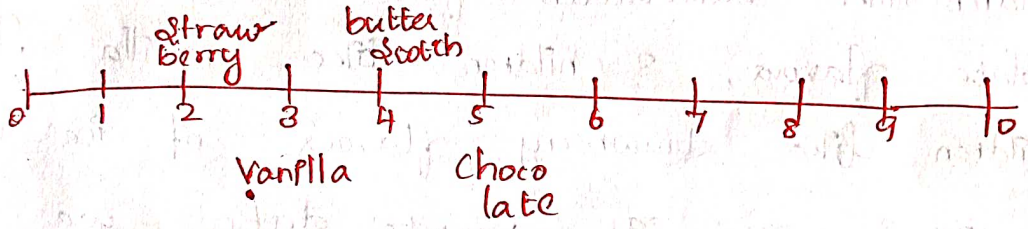


2) In frequency table, values are represented through tally marks and frequency of each variable.

Favourite Ice Cream Flavour.

Flavour	Tally Marks	Number of children
Butterscotch	IIII	4
Chocolate	IIII	5
Vanilla	III	3
Strawberry	II	2

3) On Number line, we marked the value of each variable on the number line



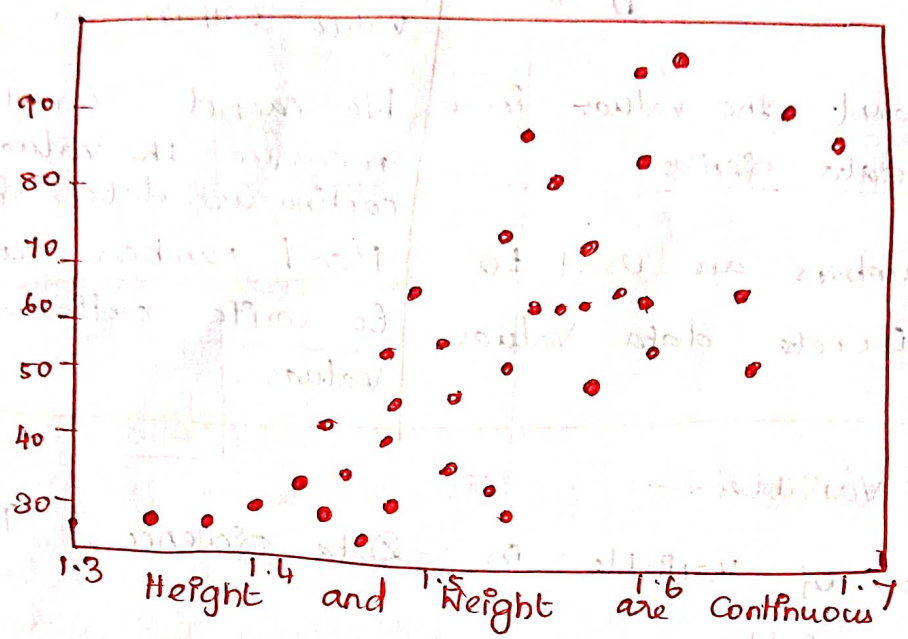
2.2 Continuous Data :- can take any value within a range

Continuous data is data that can be calculated. It has an infinite number of probable values that can be selected within a given specific range.

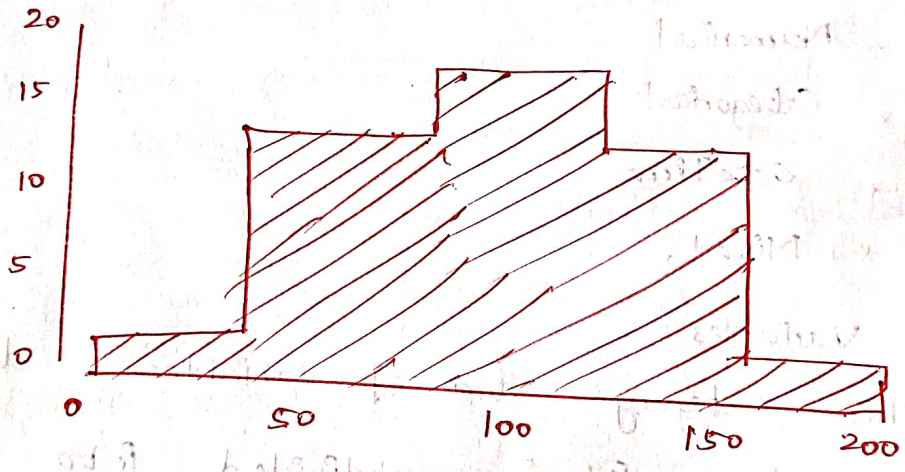
Eg) Temperature Range.

* The amount of time required to complete a project. (2-8) hours.
* The height of children
* The amount of time it takes a doll shoes.

Histograms and scatter plots to graph continuous data. These graphs are designed to handle values that fall on a continuous spectrum and have decimal places.



Height and Weight are continuous variables.



Histogram of the companies with respect to the Annual Response.

Difference Between Discrete and Continuous Data.

Discrete Data	Continuous Data.
Variable can take only specific values.	Variables can take any value within a range
We can count the values in a discrete data series	We cannot count but measure the values in a continuous data series.
Whole numbers are used to write discrete data values.	Real numbers are used to write continuous data values.

Types of Variables.

Types of variable in Data science. There are four types of variables.

* Numerical
* Categorical
* DateTime
* Mixed.

Numerical Variables :-

This category ~~dea~~ of variables deals with the Numbers only. It is subdivided into

* Discrete
* Continuous.

Categorical Variables :.

This Category deals with the Categories. It is subdivided into

⇒ Ordinal
⇒ Nominal

Date & Time Variable :-

This category of variable deals with the date & time aspects. This category can contain the type of values.

* Only having date
* Only having time.
* Having both date & time.

Eg)

* Birthdate
* Time on which the log of the system is generated.
* Date of Application
* Date of ordering a product online.

Mixed variables :-

This category of variable deals with the collection of multiple values for the multiple observations of a specific variables.

⇒ Number or label/strings in different observations

⇒ Number or label/strings in the same observations

Eg of Number or label/strings in different observations

Performance of a student (CGPA, Percentage, or grades → Includes Numbers as well as label or strings)

Eg of Numbers & Labels/strings in the same observation.

* Cabin number (A1, C3, E5)
* Vehicle Registration Number

Independent Variables :-

An independent variable is a singular characteristic that the other variables in your experiment cannot change.

Eg Age what they eat or how much they exercise are not going to change their age.

Dependent Variable :-

A dependent variable relies on and can be changed by other companies.

Eg) A grade on an exam is an example of a dependent variable, because it depends on factors such as how much sleep you got and how long you studie

Independent variable can influence dependent variable, but dependent variables cannot influence independent variables.

Eg) Time you spent studying (dependent)

2.3. Describing data with Tables and Graphs :-

There are 4 ways to represent the data,

* Tables are the simplest way to represent data.
* A pictograph uses images to represent a certain number of items.
* A line graph plots individual data points as dots and connect them with lines.
* Bar graph use bars of different heights to represent data.

Tables (Frequency Distributions)

Organization of Data :-

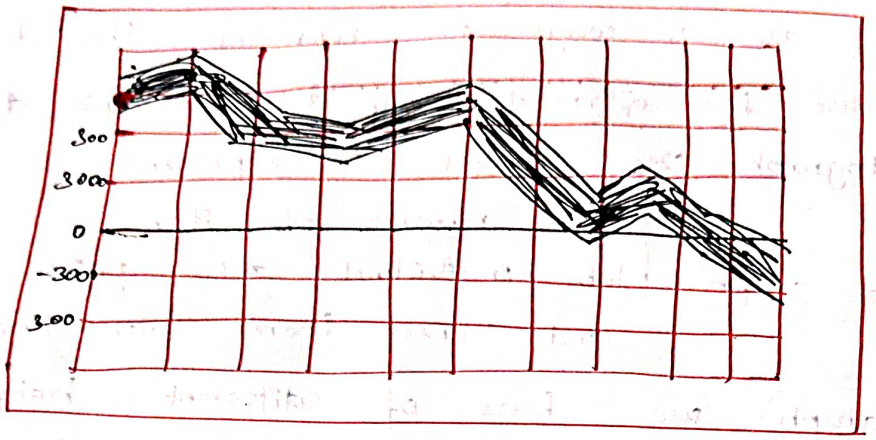
Statistics refers to the collection, organization, distribution, and interpretation of data or a set of observations.

A frequency distribution is a collection of observation produced by sorting observations into classes and showing their frequency (f) of occurrence in each class.

Types of frequency Distribution :-

* Ungrouped frequency distribution.
* Grouped frequency distribution
* Cumulative frequency distribution
* Relative frequency distribution.
* Relative Cumulative frequency distribution.

Ungrouped Frequency Distribution : -



Frequency distribution provided with raw data or random values. In this case, we may have to make either for individual observations or class intervals.

Let the test scores of all 20 students be as follows.

23, 26, 11, 18, 09, 21, 23, 30, 22, 11, 20 11, 13, 23, 11
29, 25, 26, 26

If we count all the occurences of a single data values or class interval in one go. We will have to cross check the entire list again and again for the next observation or class interval.

Hence this will take lot of time to finish. The complexity of this can be reduced by making use of tally marks.

Marks obtained in the test	Tally Marks	No of students (Frequency)
09	I	1
11	IIII	4
13	I	1
18	I	1
20	I	1
21	II	2
22	I	1
23	III	3
25	I	1
26	III	3
29	I	1
30	I	1
Total		20

Ungrouped frequency distribution, provided with raw data or random values.

Count all the occurrences of single data value. Frequency means total no. of observations Check the entire list again and again for the next observation or class interval. Hence this will take a lot of time to finish.

The complexity can be reduced by making use of tally marks.

Grouped Frequency Distribution :-

Grouped data are presented by frequency table.

A grouped frequency distribution shows the scores by grouping the observations into intervals and then lists these intervals in the frequency distribution table. The intervals in grouped frequency distribution are called class intervals or limits.

Groups of data in the form of class intervals to tally the frequency for the data that belongs to that particular class interval. The lowest number in a class interval is called lower limit. The highest number is upper limit.

Marks obtained in the test (Class Interval)	No. of students (frequency)
0-5	3
5-10	11
10-15	38
15-20	34
20-25	9
25-30	5
Total	100

Cumulative Frequency Distributions :-

The cumulative frequency is the total of frequencies, in which the frequency of first class interval is added to the frequency of the second class interval and then the sum is added to the frequency

steps to Construct Less than Cumulative Frequency Curve.

1. Mark the upper limits on the horizontal axis or x-axis
2. Mark the cumulative frequency on the vertical axis or y-axis.
3. Plot the points (x, y) in the coordinate plane where x represents the upper limit value and y represents the cumulative frequency.
4. Finally, join the points and draw the smooth curve.
5. The curve so obtained gives a cumulative frequency distribution graph of less than type.
6. To draw the frequency distribution, of less than type.

Table Cumulative Frequency Distribution Table of more than type

Level of Essay	Age Group class interval	Age Group	Number of Participants (Frequency)	Cumulative Frequency.
Level 1	10-15	Less than 15	20	20
Level 2	15-20	Less than 20	32	52
Level 3	20-25	Less than 25	18	70
Level 3	25-30	Less than 30	30	100

of the second class interval and then the sum is added to the frequency of the third class interval and so on.

Types of Cumulative Frequency Distributions :-

* Less than Cumulative Frequency
* Greater than Cumulative Frequency

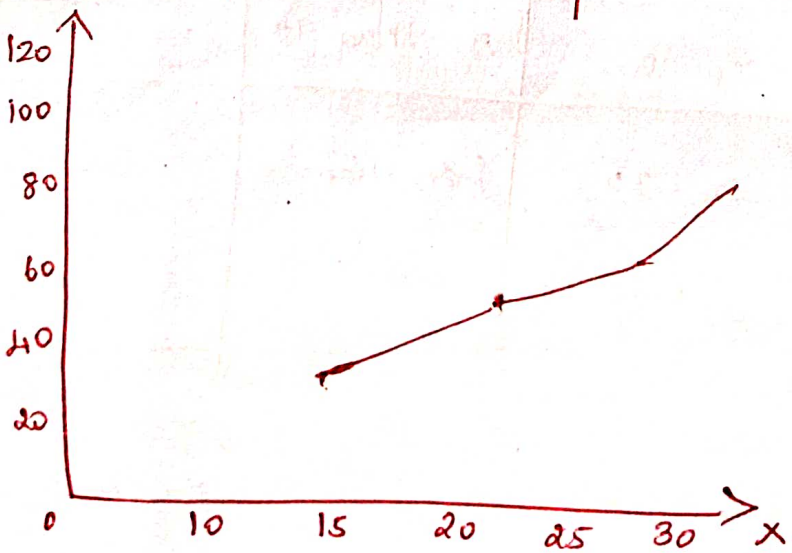
Less than Cumulative Frequency :-

The cumulative begins from the Lowest the highest size

Greater Cumulative Frequency :-

The cumulative total Frequencies starting from the highest class to the Lowest class.

Level of Essay	Age Group class interval	Age Group	Number of Participants (Frequency)	Cumulative Frequency
Level 1	10-15	Less than 15	20	20
Level 2	15-20	Less than 20	32	52
Level 3	20-25	Less than 25	18	70
Level 4	25-30	Less than 30	30	100



Relative frequency Distribution : (17)

A relative frequency distribution shows the proportion of the total number of observations associated with each value or class of values and is related to a probability distribution.

Height	Frequency	Rel. Frequency.
57 or less	1	0.025
57.1 to 58.6	1	0.025
58.6 to 60.1	3	0.075
60.1 to 61.7	6	0.15
61.7 to 63.3	8	0.2
63.3 to 64.8	11	0.275
64.8 to 66.4	3	0.075
66.4 to 68.0	7	0.175
Total	40	1

Cumulative Relative frequency.

It can be found by dividing the cumulative frequency of each interval by the total number of observations.

$$\text{Cumulative Relative frequency distribution} = \frac{\text{Cumulative frequency of each interval}}{\text{Total no. of observations}}$$

Eg. following dataset.

{1, 1, 1, 1, 1, 3, 3, 3, 3, 3, 5, 5, 5, 5, 5, 5, 11, 11, 11, 11, 11, 11, 11}

First are → find frequency table then we find the cumulative frequency. Cumulative Relative frequency.

Count	Frequency	Cumulative Frequency	Cumulative Relative Frequency
1	5	5	$5/25 = 0.2$ 20%.
3	5	10	$10/25 = 0.4$ 40%.
5	6	16	$16/25 = 0.64$ 64%.
7	1	17	$17/25 = 0.6$ 68%.
11	8	25	$25/25 = 1.0$ 100%.

Qualitative data from frequency table are presented by Histogram.

Histogram :- Linegraph, Scatter diagram.

- Histogram :-

It is used to visualize the distribution of data among different intervals as a series of vertical bars

Eg) Construct a Histogram for the following frequency distribution table describe the frequency of weight of 25 student in a class.

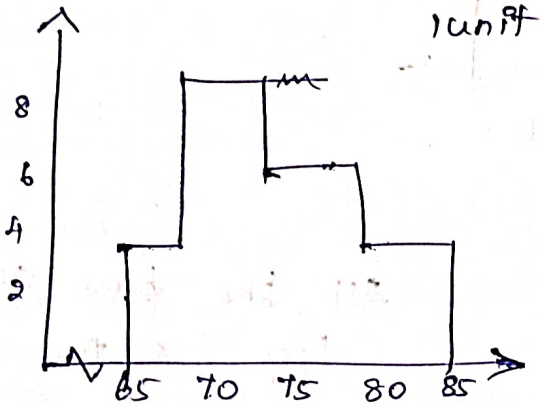
Weight	Frequency (Numbers of students.
65-70	4
70-75	10
75-80	8
80-85	4

Horizontal axis - start from 65 not from 0

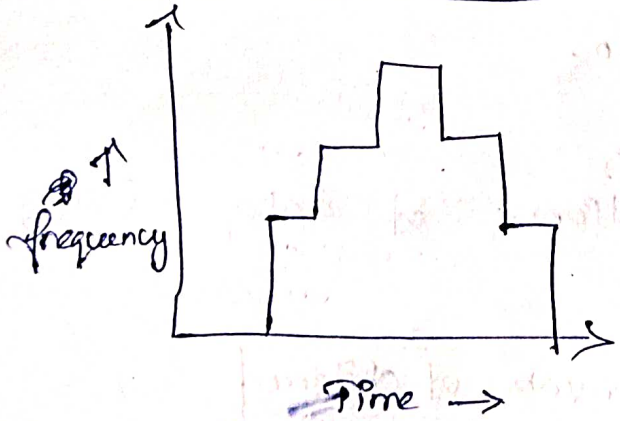
1 unit = 15

Vertical axis - frequency varying from 4 to 1

1 unit = 2



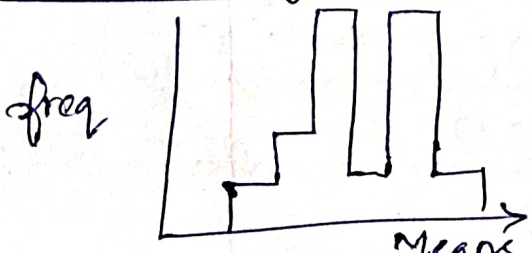
1) Bell shaped Histogram :-



→ It has a single peak

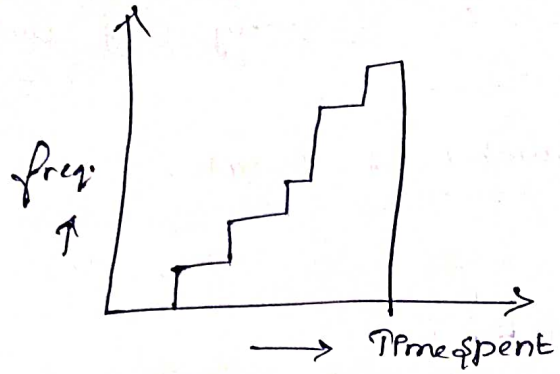
→ Has only one peak at the time.

2. Bimodal Histogram :-



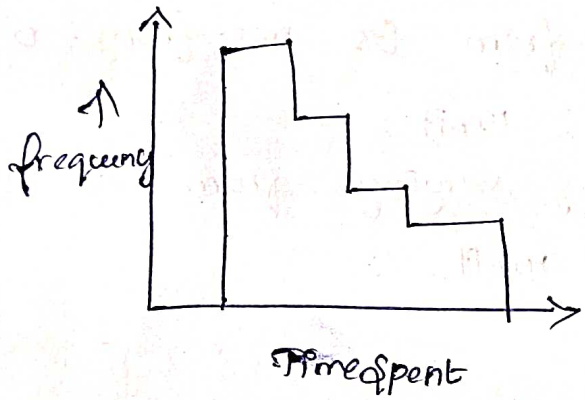
→ It has two peaks.

Skewed Right Histogram : (20)

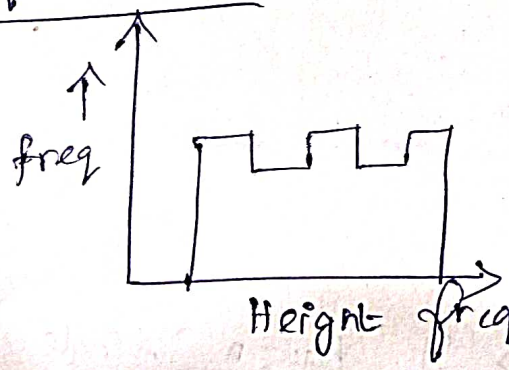


Bars of the Histogram are skewed to Right

Skewed Left Histogram :-



Uniform Histogram :-



All the bars are more or less of the same height.

Pictograph :-

Pictorial representation of data using images, icons, or symbols.

~~Atum~~ flavour	Number of children
chocolate	○○◁
butterscotch	○○○○△
Vanilla	○○○.
Strabery	○○

Key : ○ Represents 4 children.

$2 \times 4 \times \frac{1}{4} \times 4 = 9$

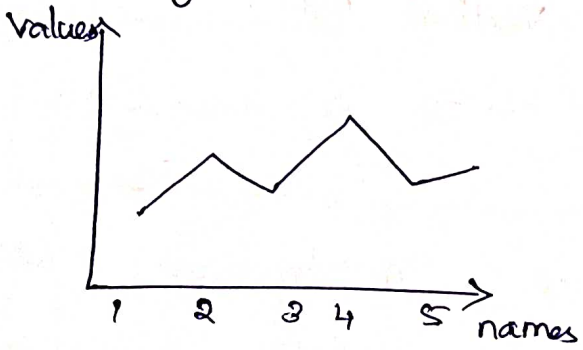
$8 + 1 = 9$

Line Graph :-

Draw the horizontal and vertical axis

x-axis → names

y-axis → values.

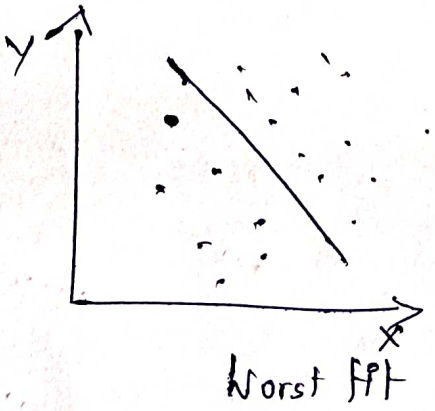


Plot the point on each plots

Match	1	2	3	4	5
Runs	10	30	50	20	40

Scatter graph :-

A scatter or xy plot has points that show the relationship b/w two sets of data.



Worst fit

Best fit

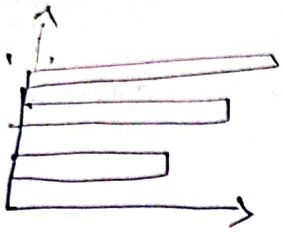
The line is drawn which is nearest to almost all the points.

Bar Graph :-

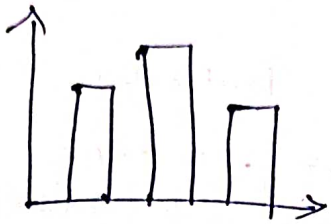
It is the pictorial representation of data, in the form of vertical and Horizontal bar.

Types of Bar Graph : (22)

1) Vertical Bar Chart
2) Horizontal Bar chart



Horizontal Bar



Vertical Bar.

2.4. Describing Data With Averages :-

Different measures can be computed to describe how data are distributed. These are classified into 3 Categories.

* Measures of Location
* Measures of Variation
* Measures of Shape

Location measure is understood as the "center" of the data eg) Average price of a new product, middle price of a new product, or the most frequent price of a new product.

Most widely used Measures of the data, "Center" are the ⇒ Mean (average)
⇒ Median
⇒ Mode

Average :-

Average represents the whole value and it lies between the minimum and maximum value of the data. Hence it is the representative figure of the entire data.

Definition :-

Clark and Sekkade defines as "average is an attempt to find one single figure to describe whole of figures.

Objectives of an Average :-

1) The representation value explains the character of the cohole data. The one value may represent ever thousand and millions of valuos.

2) Policy decision may be taken with the help of one representative figure.

3) It helps to compare with other data.

characteristics of an Average :-

1) Everyone can easily Understand the single Representative value.

2) It is very simple to calculate.

3) It should be deffned rigidly.

4) It should Represent the entire data

Types of Averege :-

1) Arithmetic Mean

(i) Simple Arithmetic Mean

(ii) Weighted Arithmetic Mean

2) Median

3) Mode

4) Geometric Mean

5) Harmonic Mean.

Arithmetic Mean :-

The mean of data indicates an average of the given collection of data.

It is equal to the sum of all the values in the group data divided by the total number of values.

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots x_n}{x_n}$$

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

Calculating the mean when the frequency of the observations is given

$$\bar{x} = \frac{f_1 x_1 + f_2 x_2 + f_3 x_3 + \cdots f_n x_n}{f_1 + f_2 + f_3 + \cdots f_n}$$

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i}$$

In a class 30 students, marks obtained by students in mathematics out of 50 is tabulated below. Calculate the mean of the data.

marks obtained	Number of student	Class mark	$f_i x_i$
10-20	5	15	75
20-30	5	25	125
30-40	8	35	280
40-50	12	45	540
Total	$\sum f_i = 30$		$\sum f_i x_i = 1020$

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{1020}{30} = 34.$$

The mean of the given data is 34.

Median :-

→ Median is the middle most value in a given dataset after it is arranged in ascending order

If the number of items in the data set is odd = $\left[\frac{n+1}{2}\right]^{th}$ term.

" even = $\left[\left[\frac{n}{2}\right]^{th} \text{term} + \left[\frac{n}{2}+1\right]^{th} \text{term}\right] / 2$

n → total no. of observations.

Find the median of the following.

data 48, 20, 50, 69, 73

(3)

Arranged in ascending order

No. of observations = 5

⇒ Median of odd data

$= \left[\frac{n+1}{2}\right]$

$= 5+1/2$

$= 6/2$

$= 3$

Median is 3 observation.

Median = 50.

Find the median of the above grouped data.

To find Median, we need cumulative frequency.

Class Intervals	No. of girls (f)	Cumulative frequency
120 - 130	2	2
130 - 140	8	2+8 = 10
140 - 150	12	10+12 = 22
150 - 160	20	2+22 = 42
160 - 170	8	42+8 = 5

n = sum of cf $n = 50$ $n/2 = 25$

Median class = 150 - 160

$l = 150$, $c = 22$, $f = 20$, $h = 10$

$$\text{Median} = l + \left[\left(\frac{n}{2} - c\right)/f\right] * h$$

$$= 150 + \left[\left(\frac{50}{2} - 22\right)/20\right] \times 10$$

$$= 150 + 1.5$$

Median = 151.5

Find.

Weekly Expenditure	0 - 1000	1000 - 2000	2000 - 3000	3000 - 4000	4000 - 5000	Total
No. of families	34	12	43	60	51	200

Mode :-

It is the value which has the maximum frequency. It is possible to have more than one value which has the same maximum frequency.

Finding the Mode of Ungrouped data.

Wickets taken by a bowler in 10 cricket Matches are, 2, 6, 4, 5, 0, 2, 1, 3, 2, 3

To find the Mode of Ungrouped data.

Numbers of Wickets	0	1	2	3	4	5	6
Number of Matches	1	1	3	2	1	1	1

Maximum no. of Matches = 3

Mode of the given data = 2.

$$\text{Mode} = f + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right) \times h.$$

} It is ungrouped data.

Where f_1 is the frequency of the modal class

f_0 is the frequency of the class preceeding the modal class

f_2 is the frequency of the class succeeding the modal class.

h is the size of the class Interval

f is the Lower Limit of the modal class

Mode of grouped data

Eg)

Size of family	1-3	3-5	5-7	7-9	9-11
No. of families	7 (f_0)	8 (f_1)	2 (f_2)	2	1

Maximum class frequency = 8

Class Interval = 3-5

Lower Limit of the modal class = $l = 3$

class size (or) size of the interval is $h = 2$

frequency of the modal class $f_1 = 8$

frequency of the class proceeding to modal class $f_0 = 7$

frequency of the class succeeding to modal class $f_2 = 2$

We know that the formula to find the mode of the grouped data.

$$\text{Mode} = l + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right) \times h$$

$$\text{Mode} = 3 + \left(\frac{8-7}{2(8)-7-2}\right) \times 2$$

$$= 3 + \left(\frac{1}{16-7-2}\right) \times 2$$

$$= 3 + \left(\frac{1}{7} \times 2\right)$$

$$= 3 + 2/7$$

$$= \frac{21+2}{7}$$

$$= 23/7$$

2.5 Describing variability :-

* Variability (also called spread or dispersion) refers to how spread out a set of data is.

* Variability gives you a way to describe how much data sets vary and allows you to use statistics to compare your data to other sets of data. The four main ways to describe variability

⇒ Range
⇒ Interquartile Range
⇒ Variance
⇒ Standard deviation

Several Measures of variability, including the range, the interquartile range, the variance, the most important, the standard deviation.

Range :-

Range is the simplest method of studying dispersion. It is defined as the difference between the value of the smallest item and value of the Largest item included in the distribution.

$$\text{Range} = L - S$$

$L \rightarrow$ Largest Item

$S \rightarrow$ Smallest Item.

The relative measures of Corresponding to the range is called the coefficient of range.

Coefficient of range = $\frac{L-S}{L+S}$

Eg)1 The following are the prices of shares of AB company Ltd. From monday to saturday, Monday, Tuesday, Wednesday, Thursday, Friday, Saturday.

Price : 200, 210, 208, 160, 220, 250

Calculate Range and its coefficient

Range = L − S

Range = 250 − 160

= 90

Coefficient of Range = $\frac{L-S}{L+S}$

$= \frac{250-160}{250+160}$

$= \frac{90}{410}$

$= 0.22.$

Eg)2 Calculate Coefficient of range from the following data

Marks: 10-20, 20-30, 30-40, 40-50, 50-60

Number of stu: 8, 10, 12, 8, 4.

L = 60

S = L − S

Range = L − S

= 60 − 10

= 50

Coefficient of Range = $\frac{L-S}{L+S}$

$= \frac{50}{70}$

$= 0.7142.$

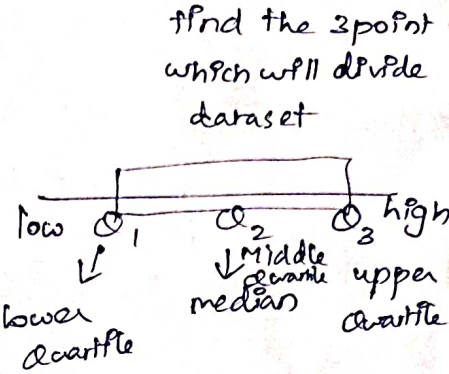
Interquartile Range (IQR)

→ It is the range for the middle 50 percent of the scores.

→ They are values that divide the dataset into 4 equal parts.

$$IQR = Q_3 - Q_1$$

Inter Quartile Range



Interquartile range gives you the spread of the middle of your distribution.

The Interquartile range is the third quartile (Q_3) minus the first quartile (Q_1). It will provide the range of the middle half of a dataset

* Lowest value
* Q_1 : 25 th percentile
* Q_2 : The median
* Q_3 : 75th percentile
* Highest value (Q_4)

Quartile deviation $QD = \frac{Q_3 - Q_1}{2}$.

Coefficient $QD = \frac{Q_3 - Q_1}{Q_3 + Q_1}$

Eg) Find out the value of QD and the coefficient from the following data.

Roll No : 1, 2, 3, 4, 5, 6, 7

Marks : 20, 28, 40, 12, 30, 15, 50

Calculating Quartile Deviation

Marks arranged in ascending order 12, (15), 20, 28, 30, (40), 50

$n = 7$

$Q_1 = \frac{1}{4}(N+1)^{th}$ term

$= \frac{1}{4}(7+1)$

$= 8/4$

$Q_1 = 2^{th}$ term is (15)

odd = n+1 term

even = n th term.

$Q_3 = \frac{3}{4}(N+1)^{th}$

$= \frac{3 \times 8}{4}$

$Q_3 = 6^{th}$ term is (40)

$QD = \frac{40 - 15}{2}$

$= \frac{25}{2}$

$= 12.5$

Coefficient of $QD = \frac{Q_3 - Q_1}{Q_3 + Q_1}$

$= \frac{40 - 15}{40 + 15}$

$= \frac{25}{55}$

$= 0.45$

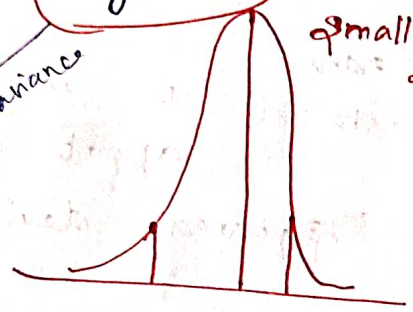
Standard deviation :- (34)

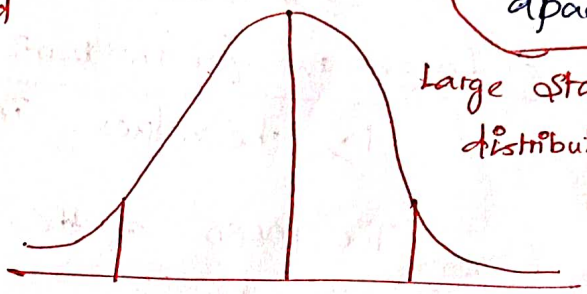
The std deviation tells you how tightly your data is clustered around the mean (the average)

A small SD indicates that your data is tightly clustered

A Large SD tells you that your data is more spread apart.

Std deviation = $\sqrt{\text{variance}}$





standard Deviation Formula for population.

Formula	Explanation
$\sigma = \sqrt{\frac{\Sigma(X-\mu)^2}{N}}$	$\sigma \rightarrow$ population std deviation Σ – Sum of X – Each value μ = population mean $N \rightarrow$ Number of values in the population

Standard Deviation Formula for samples.

Formula	Explanation
$s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$	$s \rightarrow$ Sample standard deviation $\Sigma \rightarrow$ sum of $x \rightarrow$ each value $\bar{x} \rightarrow$ Sample mean $n \rightarrow$ number of values in the samples

Variance :

It is the square root of standard deviation. Variance of a data set is how spread out your data

A small no for the variance Means.

↳ dataset is tightly clustered together,

A Larger no. for the variance Means

↳ The values in the dataset is more spread apart.

The mean ~~of the~~ of all squared deviations squares

$$\sigma^2 = SS/N = \frac{\text{Sum of Square}}{N}$$

$$\text{Std deviation} = \sqrt{\text{variance}}$$

Variance formula for (population).

↳ Entire or Complete set

Formula	Explanation
$\sigma^2 = \frac{\Sigma(X-\mu)^2}{N}$	σ^2 = population variance Σ = sum of each value μ = population mean = No. of values in the population

(Sample) ↳ Subset of population.

Variance Formula for (samples):

Formula	Explanation
$s^2 = \frac{\Sigma(X-\bar{x})^2}{n-1}$	s^2 = sample variance Σ = sum of each value. $\bar{x}$ = sample mean n = number of values in the sample

Data values are arranged in ascending order.

7, 11, 11, 15, 20, 20, 28

x	$\bar{x}$	$x - \bar{x}$	$(x - \bar{x})^2$
7	16	-9	81
11	16	-5	25
11	16	-5	25
15	16	-1	1
20	16	4	16
20	16	4	16
28	16	12	144

$\Sigma (x - \bar{x})^2 = 308$

Mean

$$\bar{x} = \frac{\Sigma x}{n} = \frac{7+11+11+15+20+20+28}{7}$$

$$\bar{x} = \frac{112}{7}$$

$$\boxed{\bar{x} = 16}$$

Variance formula for samples.

$$s^2 = \frac{\Sigma (x - \bar{x})^2}{n-1} = \frac{308}{7-1} = \frac{308}{6} = 51.333$$

$$\boxed{s^2 = 51.333}$$

Standard Deviation :-

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

$$= \sqrt{\frac{308}{6}}$$

$$= \sqrt{51.333}$$

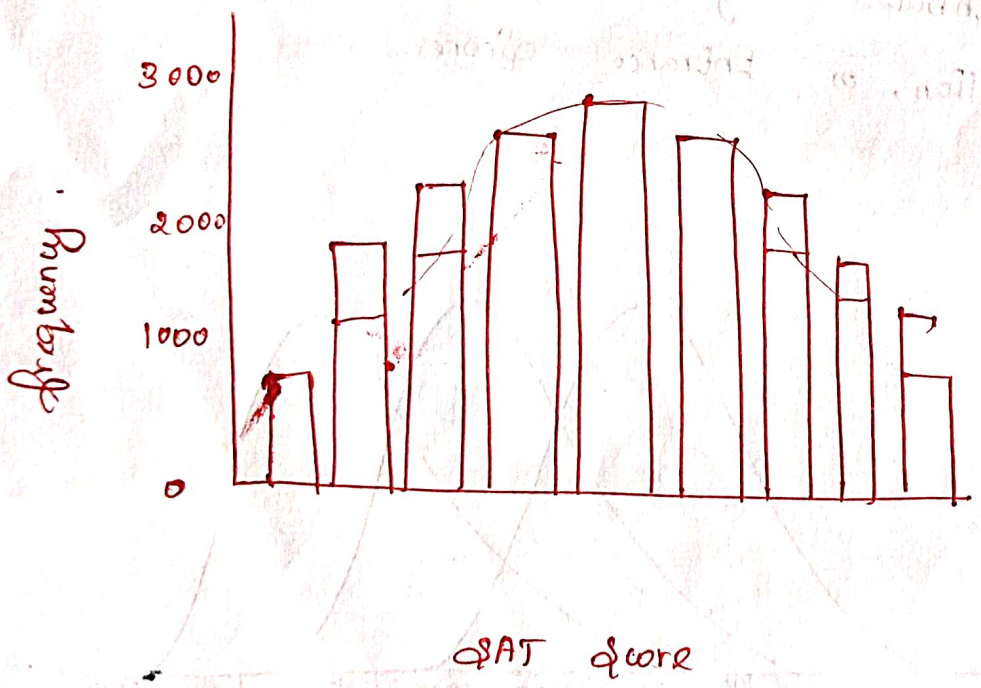
$$\boxed{s = 7.165}$$

2.6. Normal Distributions and standard (z) scores:-

Normal distribution, also known as Gaussian distribution, is a probability distribution that is symmetric about the mean, showing the data near the mean are more frequent in occurence than data far from the mean.

In graphical form, the normal distribution appears as a "bell Curve"

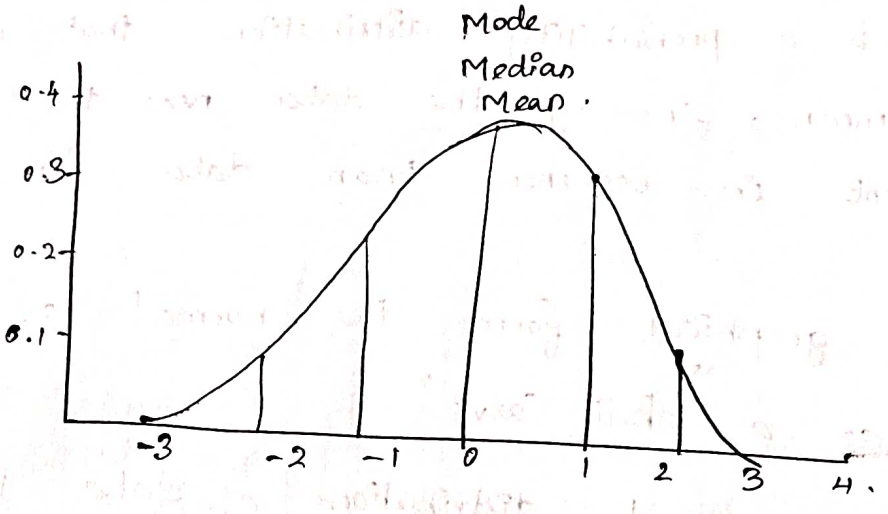
In a normal distributions, data is symmetrically distributed with no skew. When plotted on a graph the data follows a bell shape, with most values clustering around a Central region.



Properties of Normal Distribution:-

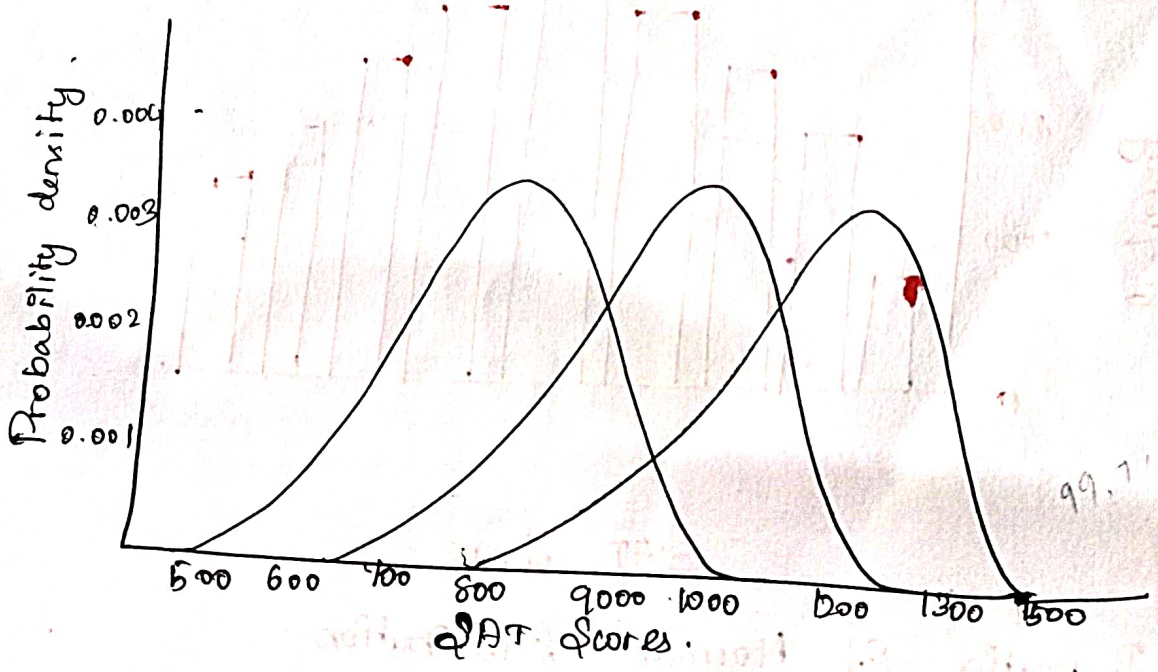
* The mean, median, mode are exactly the same
* The distribution is symmetric about the mean. half the values fall below the mean and half above the mean.

* The distribution can be described by two values. the mean and the standard deviation.



Normal distributions Important :-

All kinds of variables in natural and Social Sciences are normally or approximately ~~in~~ normally distributed. Height, birth weight, Reading ability, Job satisfaction, or Entrance Scores



Normal distributions with different means.

1. Outlier :-

An unusual high or low piece of data could be outlier.

1. Sort your data :-

To identify Outlier, to sort your data, which allows you to see any unusual data points within your information.

For eg) 3, 6, 7, 10 and 54 arranged in ascending order you can see, 54 is not larger than the rest of the data point

2. Graph your data

To find outliers, graph your data in scatter plots or histograms

↳ Useful for visualizing outliers, because one dot is far away from the other dots.

Histograms :-

→ It displays data in groups called "bins"
→ Most of your data points, are on the ~~left~~ Right side of the graph, and one bin of data is on the left side of the graph.

Left bin → outlier.

3. Calculate the Z-score :-

(40)

⇒ To calculate z-score, subtract the mean from the raw measurement and divide it by the standard deviation.

$$Z = \frac{X - \mu}{\sigma}$$

$X \rightarrow$ Raw measurement

$\mu \Rightarrow$ mean

$\sigma \rightarrow$ Standard deviation.

Describing Relationships.

Correlation - scatter plots - Correlation coefficient for quantitative data - computational formula for correlation coefficient - regression - regression line - least squares regression line - standard error of estimate - interpretation of r^2 - multiple regression equations - regression toward the mean.

Correlation :-

Definition :-

According to Smith 'correlation is the relationship that exists between two variables, when one ~~value~~ variable's value changes, the other variable changes as well (decreases or increases)

Mathematically, we can say a function has a purpose to predict a value, by converting input(x) to output (f(x)).

Types of correlation :-

* Positive and Negative correlation
* Simple Correlation and Multiple Correlation
* Partial and Total correlation.
* Linear and Non-linear Correlation.

Positive and Negative correlation :-

In positive correlation, increasing or decreasing the values of one variable necessarily leads to increasing

(or decreasing) the values of another variables.

In this case of negative correlation increasing or (decreasing) the values of one variable certainly leads to decreasing (or increasing) the values of another variable

Positive Correlation :-

Two variables change in the same direction.

Negative correlation :-

Two variables change in opposite direction.

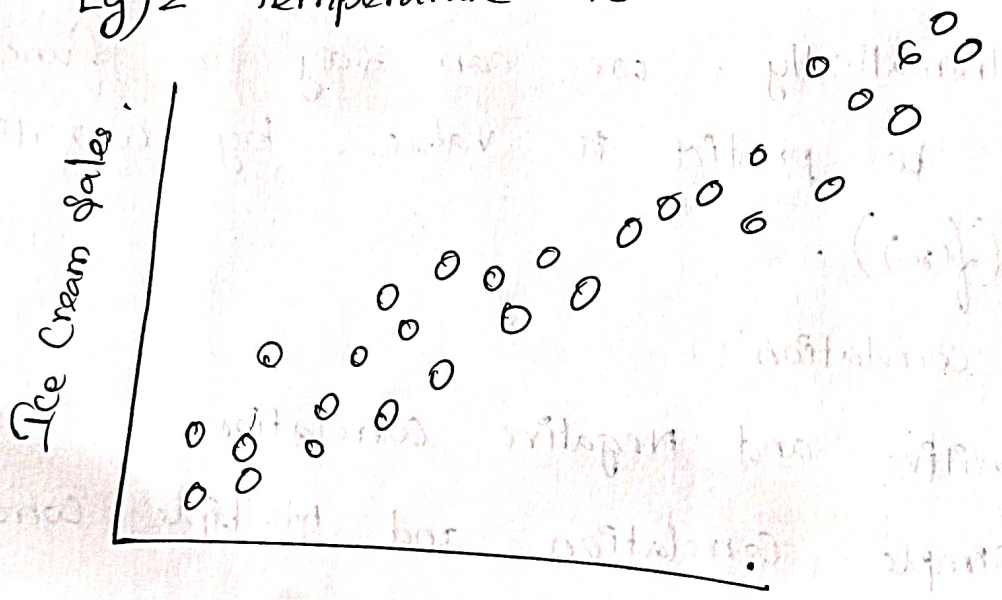
No Correlation :-

There is no association or relavent relationship between the two variables.

Positive Correlation : Eg

Eg)1 Height Vs Weight

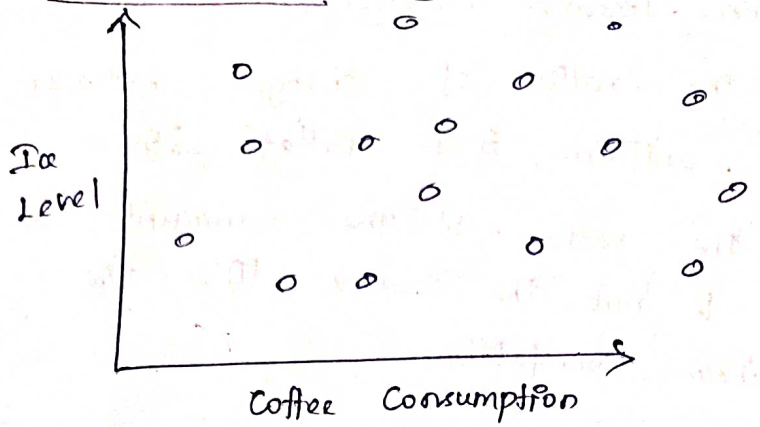
Eg)2 Temperature Vs Ice Cream Sales.



Negative Correlation : Eg)

Eg) Time Spent Running Vs Body Fat.

No correlation (3)



Amount of Coffee and IQ Level has a Correlation of zero.

Simple and Multiple Correlation :-

Simple

When we study the relationship between only two variables, it is called Simple Correlation

Eg) Price & Demand.

Multiple :-

When we study the relationship between more than two variables, it is called Multiple Correlation

Eg) Price, Demand & Supply of a product.

Partial and Total Correlation :-

When we study the relationship between only two variables, by eliminating other variable is called Partial correlation.

Eg) Studying Price & Demand Eliminating Supply of a product

In Total correlation

All the facts are taken into account

Linear and Non-Linear correlation :-

Linear

When the ratio of changes between two variables is uniform, it is called Linear.

Non-Linear

If the amount of change in one variable is not the same like the other, it is called Non-Linear correlation.

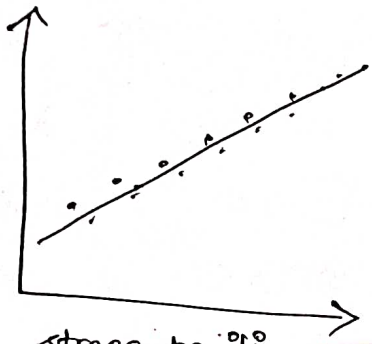
Scatter Plots :-

Scatter diagram is a diagram that shows the values of two variables X and Y. The way to relate to each other.

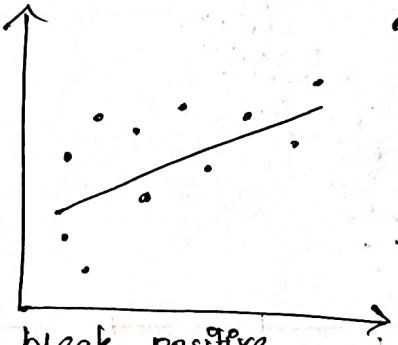
The values of

Scatter Plots :-

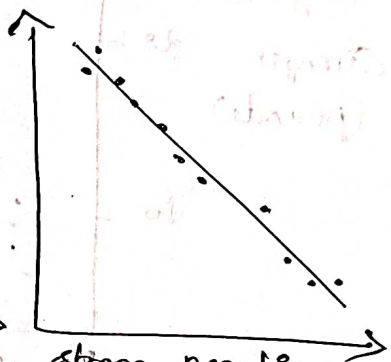
A scatter diagram is a diagram that shows the values of two variables X and Y, along with the way in which these two variables relate to each other. The values of variable X are given along the horizontal axis, with the values of variable Y given on the vertical axis.



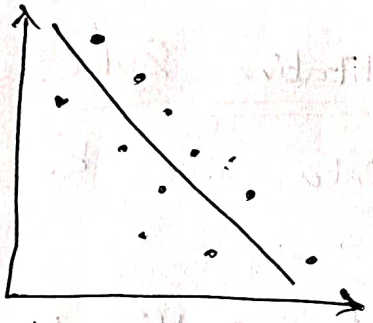
strong positive Correlation.



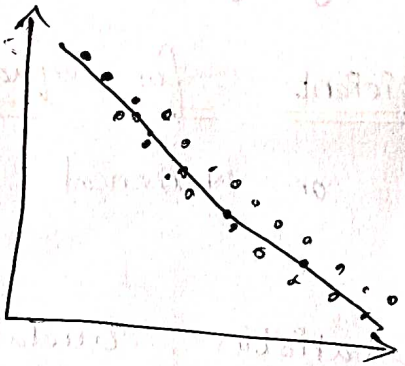
Weak positive Correlation.

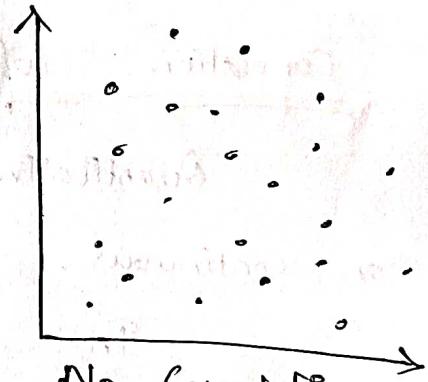


strong negative Correlation



Weak negative correlation.



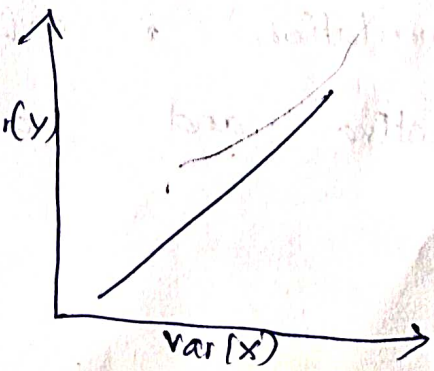


No Correlation

Linear Relationship :-

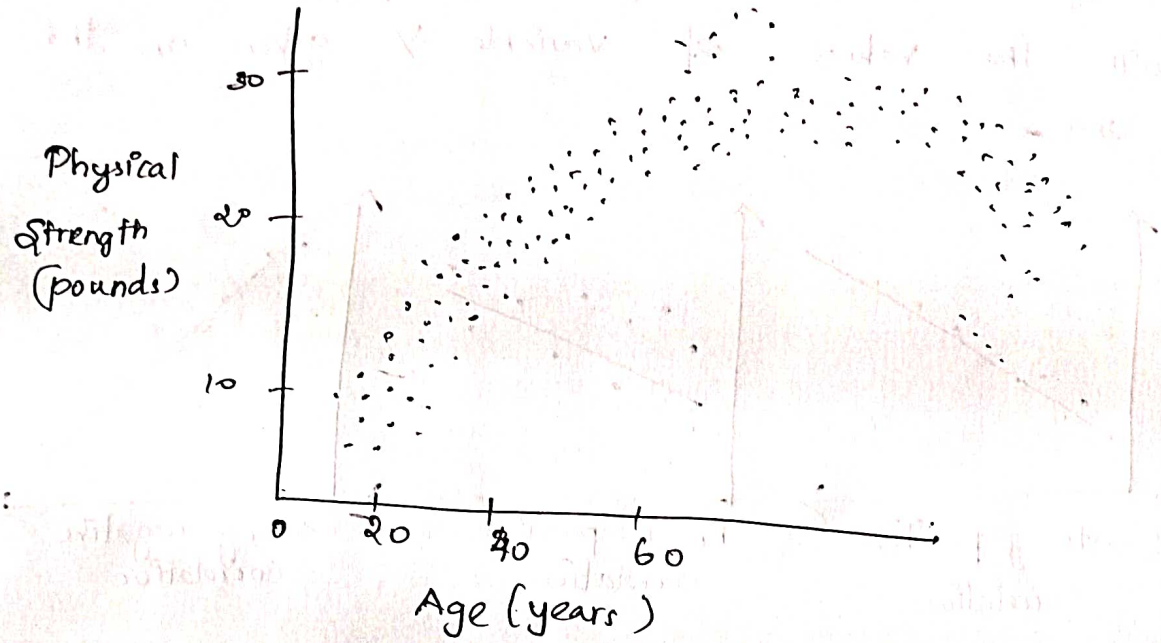
If the dot clusters approximates a straight line and, it reflects a linear relationship between X and Y

X	:	Y
1	:	10
2	:	20
3	:	30



Curvilinear Relationship:-

If the dots cluster approximates a bent or curved line and therefore reflects a curvilinear relationship



Correlation Coefficient for quantitative Data:

Quantitative or Numerical Data can be Discrete or Continuous.

In statistics, correlation coefficients are a quantitative assessment that measures both the direction and strength of the's tendency to vary together.

Correlation coefficient can be calculated using "Pearson correlation r coefficient" formula. Both variables are quantitative and normally distributed with no outliers.

Pearson's correlation coefficient is represented by the Greek Letter rho ρ (ρ) → population parameter

r → Sample statistic.

Correlation coefficient is a single numba that measures both the strength and direction of the linear relationship between two continuous variables. Values can Range from −1 to +1.

strength:-

Greater the Absolute Value of the pearson Correlation Coefficient ↓ Stronger Relationship.

Direction:-

The sign of the pearson correlation coefficient represents the direction of the relationship

Positive coefficients:-

⇒ It indicate that when the value ~~the~~ of one variable increases, the value of the other variable also tends to increase.

⇒ Positive relationships produce an upward slope on a scatterplot

Negative coefficients:-

⇒ It represent cases when the value of one variable increases, the value of the other variable tends to decrease.

Negative relationships produce a downward slope.

Pearson Correlation Coefficient Formula :-

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

When n = Quantity of Information

Σx = Total of the first Variable Value.

Σy = Total of the second Variable Value.

Σxy = Sum of the product of first & second Value.

Σx^2 = Sum of the squares of the First Value.

Σy^2 = Sum of the squares of the second value.

Correlation coefficient Measures the relationship between two variables.

1 → there is a perfect linear relationship b/w variable

0 → There is a No linear relationship b/w variables.

-1 → There is a perfect negative linear relationship

Find the Karl Pearsons correlation coefficient b/w variables.

x	x^2	y	y^2	xy
6	36	9	81	54
2	4	11	121	22
10	100	5	25	50
4	16	8	64	32
8	64	7	49	56
30	220	40	340	214

x : 6, 2, 10, 4, 8

y : 9, 11, 5, 8, 7

$$r = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{N\Sigma x^2 - (\Sigma x)^2}\sqrt{N\Sigma y^2 - (\Sigma y)^2}}$$

$$= \frac{5 \times 214 - 30 \times 40}{\sqrt{5 \times 220 - (30)^2}\sqrt{5 \times 340 - (40)^2}}$$

$$= \frac{1070 - 1200}{\sqrt{1100 - 900}\sqrt{1700 - 1600}}$$

$$= \frac{-130}{\sqrt{200}\sqrt{100}}$$

$$= \frac{-130}{14.14(10)}$$

$$= \frac{-130}{141.4}$$

$$= -0.9194$$

2.4. Computational Formula For Correlation Coefficient :-

The four types correlation coefficients are given by

* Pearson Correlation Coefficient
* linear correlation Coefficient
* Sample Correlation Coefficient
* Population Correlation Coefficient.

Correlation show the relation between two variable Correlation coefficient shows the measure of correlation. To compare two datasets, we use the correlation formulas.

Pearson Correlation Coefficient Formula :-

The most common formula is the pearson Correlation Coefficient used for linear dependency between the dataset. The value of the coefficient lies between -1 to +1. When the coefficient comes down to Zero. the the data is considered as not related.

If we get the value of +1 → then the data are positively correlated

-1 → Negatively Correlated.

$$r = \frac{n\Sigma(xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x^2)][n\{\Sigma y^2\} - (\Sigma y)^2]}}$$

Where n = Quantity of Information

Σx ⇒ Total of the First Variable Value.

Σy ⇒ Total of the Second Variable Value.

Σxy = Sum of the product of first & second value.

Σx^2 = Sum of the squares of the First value.

Σy^2 = Sum of the squares of the second value.

Linear correlation coefficient Formula :-

The formula for the Linear correlation coefficient is given by ;

$$r_{xy} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

Sample correlation coefficient formula :-

The formula is given by ,

$$r_{xy} = s_{xy} / s_x s_y$$

Where s_x and s_y are the sample standard deviations, and

s_{xy} is the sample covariance.

Population correlation coefficient Formula :-

The population correlation coefficient uses σ_x and σ_y as the population standard deviations and σ_{xy} as the population covariance

$$r_{xy} = \sigma_{xy} / \sigma_x \sigma_y .$$

Rank correlation ⇒ Charles Edward Spearman developed the method of Rank correlation coefficient.

$$R = 1 - \frac{6\Sigma D^2}{N^3 - N}$$

⇒ This method can be applied where the population is not known.

Eg 3

Two Judges in the beauty context rank the 12 entries as follows.

X : 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12

y : 12, 9, 6, 10, 3, 5, 4, 7, 8, 2, 11, 1

Compute the rank correlation between the two

Soln

Calculation of Rank Correlation Coefficient.

Rank X	Rank y	$D = R(x) - R(y)$	D^2
1	12	-11	121
2	9	-7	49
3	6	-3	9
4	10	-6	36
5	3	2	4
6	5	-1	1
7	4	3	9
8	7	1	1
9	8	1	1
10	2	8	64
11	11	0	0
12	1	11	121
$N = 12$			$\Sigma D^2 = 416$

$$R = 1 - \frac{6\Sigma D^2}{N^3 - N}$$

$$= 1 - \frac{6 \times 416}{12^3 - 12}$$

$$= 1 - \frac{2496}{1728 - 12}$$

$$= 1 - \frac{2496}{1716}$$

$$= 1 - 1.4545$$

$$\boxed{R = -0.4545}$$

Regression :-

Regression is defined as a statistical method that help us to analyze and understand the relationship between two or more variables of interest.

In regression, we normally have one dependent variable and more or more independent variables.

Regression Analysis :-

Regression Analysis is a branch of statistical theory that is widely used in almost all the scientific disciplines.

	Correlation	Regression
1.	Relationship b/w two or more variable.	It is a mathematical measure showing the average relationship between variables
2.	X and Y are random variables	X is a Random Variables Y is a fixed variables.
3.	It gives limited information verifying the relationship b/w the variables	It is used for the prediction of one value, in relationship to other given value.
4.	The range of relationship between the variables.	It studies Regression value is absolute figure.
5.	It studies linear relationship between the variables.	It studies linear and not linear relationship b/w variables.
6.	If the coefficient of correlation is positive, then the two variable are positively correlated and vice versa.	The regression are then explains that the decrease in one variable is associated with the increase in another variable

Dependent Variables :-

This is the variable that we are trying to understand or forecast.

Independent Variable :-

These are factors that influence the analysis or target variable and provide us with information regarding the relationship of the variable with the target variables.

Applications of Regression Analysis :-

The regression analysis method of forecasting generally involves five basic applications.

* Predictive Analysis :
* Operation Efficieny
* Supporting decision
* Correcting errors.
* New Insights.

Regression Line :-

→ The regression line is a straight line rather than a curved line because of the linear relationship between the two variables.

⇒ The regression line is often referred to as the least squares regression line.

⇒ If we take two variables x & y we have 2 regression lines

Regression Equation :-

The algebric equation of the two regression lines are called regression equation

Regression of x on y

$$x = a + by$$

The values of a and b can be calculated by solving the two normal equations,

$$\Sigma x = Na + b\Sigma y.$$

$$\Sigma xy = a\Sigma y + b\Sigma y^2$$

Regression of y on x

$$y = a + bx$$

The values of a and b can be calculated by solving the two normal equations

$$\Sigma y = Na + b\Sigma x$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2$$

Find Regression line (17)

x : 3, 5, 6, 8, 9, 11

y : 2, 3, 4, 6, 5, 10

x	$x = x - \bar{x}$	x^2	y	$y = y - \bar{y}$	y^2	xy
3	−4	16	2	−3	9	12
5	−2	4	3	−2	4	4
6	−1	1	4	−1	1	1
8	1	1	6	1	1	1
9	2	4	5	0	0	0
11	4	16	10	5	25	20
$\Sigma x = 42$	$\Sigma x = 0$	$\Sigma x^2 = 42$	$\Sigma y = 30$	$\Sigma y = 0$	$\Sigma y^2 = 40$	$\Sigma xy = 38$

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

$$= 42/6 = 7$$

$$\boxed{\bar{x} = 7}$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{30}{6} = 5$$

$$\boxed{\bar{y} = 5}$$

Regression Equation of x on y

$$(x - \bar{x}) = r\frac{\sigma x}{\sigma y}(y - \bar{y})$$

$$r * \frac{\sigma_x}{\sigma_y} = \frac{\Sigma xy}{\Sigma y^2}$$

$$\frac{\Sigma xy}{\Sigma y^2} = 0.95$$

$$(x-7) = 0.95(y-5) \Rightarrow$$

$$x-7 = 0.95y - 4.75$$

$$x = 0.95y - 4.75 + 7$$

$$x = 0.95y + 2.25$$

Regression equation of y on x

$$y-\bar{y} = r\frac{\sigma_y}{\sigma_x}(x-\bar{x})$$

$$\frac{r\sigma_y}{\sigma_x} = \frac{\Sigma xy}{\Sigma x^2} \qquad \frac{\Sigma xy}{\Sigma x^2} = 0.904.$$

$$y-5 = 0.904(x-7)$$

$$y = 0.904x - 6.328 + 5$$

$$y = 0.904x - 1.238$$

3.7 Least Squares Regression Equation :-

An equation pinpoints the exact least squares regression line for any scatter plot.

Where ⇒ Y represents the predicted value.

⇒ X represents the known value

⇒ b & a represents the Calculated from the original correlation analysis.

Using linear regression, we can find the line the best "fits" our data. This line is known as the least squares regression line and it can be used to help us understand the relationship between weight and height.

The formula for the line of best fit is written as,

$$Y = b + ax$$

Y → predicted value of the response variable,

b_0 → intercept

a → Regression Coefficient

x → value of the predictor variable.

Notice how our data points are scatted closely around this line. That's because this least squares regression line is the best fitting line for our data.

Fitting Least squares Regression lines.

Eg) Predictor Value : 140, 155, 179, 192, 200, 212.

Response Value : 60, 62, 67, 70, 71, 72, 75

Find a Regression line using the Calculator also answer.

a) For a person who weighs 170 pounds, how tall would we expect them to be?

b) For a person who weighs 150 pounds, how tall would we expect them to be?

Linear Regression Equation

$$\hat{y} = 32.7830 + (0.2001) * x$$

The calculator automatically finds the least squares regression line:

$$y = 32.7830 + 0.2001x$$

How to Use the ~~Regression~~ Least Squares Regression Line

Using this least squares regression line, we can answer the questions like

a) To answer this, we can simply substitute $x = 170$ into our regression line for x and solve for y.

$$y = 32.7830 + 0.2001(170) = 66.8 \text{ inches.}$$

b) To answer this, We can substitute for $x = 150$ into our regression line for x and solve for y

Standard Error of Estimate :-

The standard Error of Estimate is the measure of variation of an observation made around the computed regression line. Simply it is used to check the accuracy of predictions made with the regression line.

The regression equations helps us to predict the values of Y for values of X or the value of X for values of Y.

The deviation of each dot from the regression line is symbolised by $Y - Y_c$. Thus the square root of the mean of the squared deviation.

$$\sigma_y = \sqrt{\frac{\Sigma (Y - Y_c)^2}{N}}$$

Similarly

$$\sigma_x = \sqrt{\frac{\Sigma (X - X_c)^2}{N}}$$

In fact, $(X - X_c)$ and $(Y - Y_c)$ represent the unexplained variation in X and Y series.

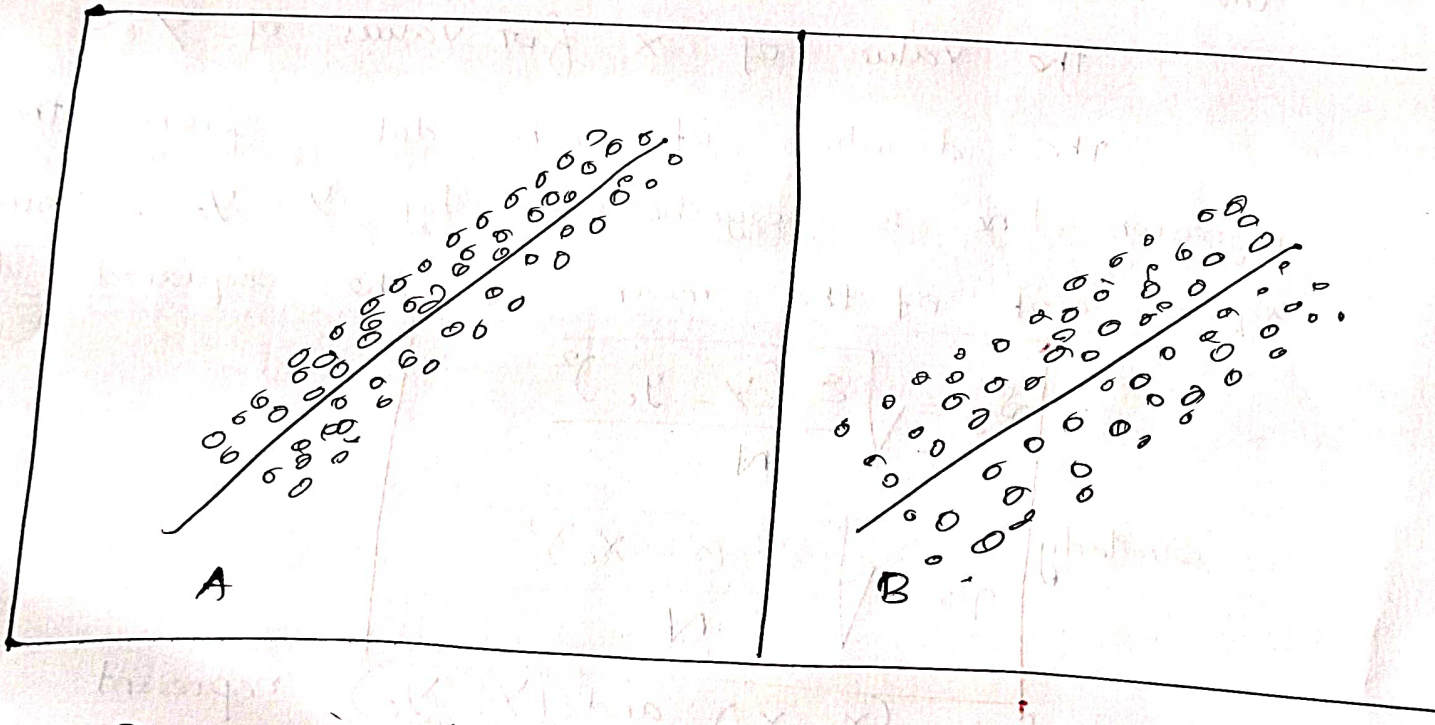
$$\sigma_y = \sqrt{\frac{\text{unexplained Variation in } Y}{N}}$$

$$\sigma_x = \sqrt{\frac{\text{unexplained Variation in } X}{N}}$$

The computation of standard error or estimate by the above formula is quite tedious. More convenient formula.

$$\sigma_x = \sqrt{\frac{\Sigma x_2^2 - a\Sigma x - b\Sigma xy}{N}}$$

$$\sigma_y = \sqrt{\frac{\Sigma y^2 - a\Sigma y - b\Sigma xy}{N}}$$



Regression differing in accuracy of prediction

You can see that Group A, the points are closer to the line than they are in Graph B. The prediction in Graph A are more accurate than in Graph B.

The standard error of the estimate is a measure of the accuracy of predictions.

The regression line is the line that minimizes the sum of squared deviations of predictions (also called the sum of squares of error)

$$\sigma_{est} = \sqrt{\frac{\Sigma (Y - Y')^2}{N}}$$

σ_{est} is the standard error of the estimate

y is an actual score

y' is a predicted score.

N is the number pair of scores.

$(y - y')$ — is the error of prediction

$$\sigma_{est} = \sqrt{\frac{\Sigma (y - y')^2}{N}}$$ for popupulation.

Similar formula are used when the standard error of the estimate is computed from a sample rather than a population.

$$s_{est} = \sqrt{\frac{\Sigma (y - y')^2}{N - 2.}}$$

Interpretation of r^2

R-squared

R-squared (R^2 or the coefficient of determination) is the statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by the independent variable. In other words, r-squared shows how well the data fit the regression model.

Regression Statistics.

Multiple R	0.93944053 6.
R Square	0.88254852
Adjusted R Square	0.87981709
Standard Error	207.2651402

R-squared can take any value between 0 and 1

In addition, it does not indicate the correctness of the regression model. Therefore the User should always draw conclusions about the model by analyzing r-squared together with the other variable in statical model.

How to Calculate R-Squared :-

The formula for Calculating R-square is

$$R - \text{squared} = \frac{SS_{regression}}{SS_{total}}$$

$SS_{regression}$ is the sum of squares due to regression (Explained sum of squares)

SS_{total} is the total sum of squares.

Multiple Regression :-

⇒ It is the statistical technique that can be used to analyze the relationship between a single dependent variable and several Independe variable.

⇒ Multiple Regression analysis is used whenever we wish to model the relationship between one response variable and more than one regressor variable

Multiple Linear Regression :-

Formula to fit Multiple Linear Regression

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + \dots + b_k * x_*$$

The variable in this equation are,

y is the predicted or expected value of the dependent variable

x_1, x_2, x_p are three independent or predictor variable.

b_0 is the value of y when all the independent variables

$b_1, b_2 \& b_k$ are the estimated regression coefficients. Each regression coefficients represents the change in y relative to a one-unit change in the respective independent variable.

Public health:-

If we want predict the future spread of this illness based upon current known infections Multiple independent variables can affect number of future infections including population size, population density, air temperature, asymptomatic carriers.

Difference between simple Linear and Multiple Linear Regression

simple Linear regression has only one x and one y variable. Multiple Linear regression has one y and two or more x variable.

Advantages of Multiple regression:-

→ More accurate

→ Precise Understanding of the association of each individual factor with the outcome

y	X_1	X_2
140	60	22
155	62	25
159	67	24
179	70	20

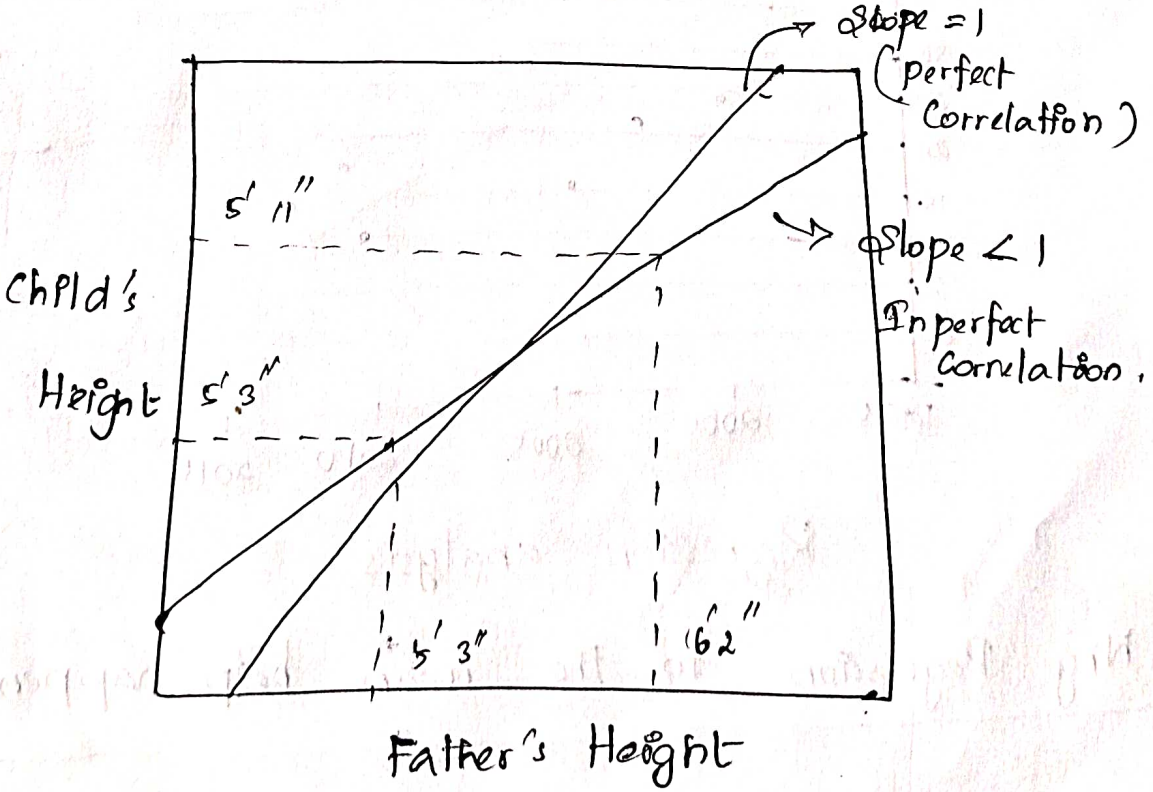
Regression to the Mean :-

Definition :-

Regression to the Mean (RTM) is a statistical phenomenon that can make natural variation in repeated data look like real change. It happens when unusually large or small measurements tends to be followed by measurements that are closer to the mean.

Mathematically, the strength of this "regression" effect is dependent on whether or not all of the random variable are drawn from the same distribution.

If data has perfect correlation, it will never regression to the mean. With an r of zero there is 100 percent regression to the mean.



The fallacy :-

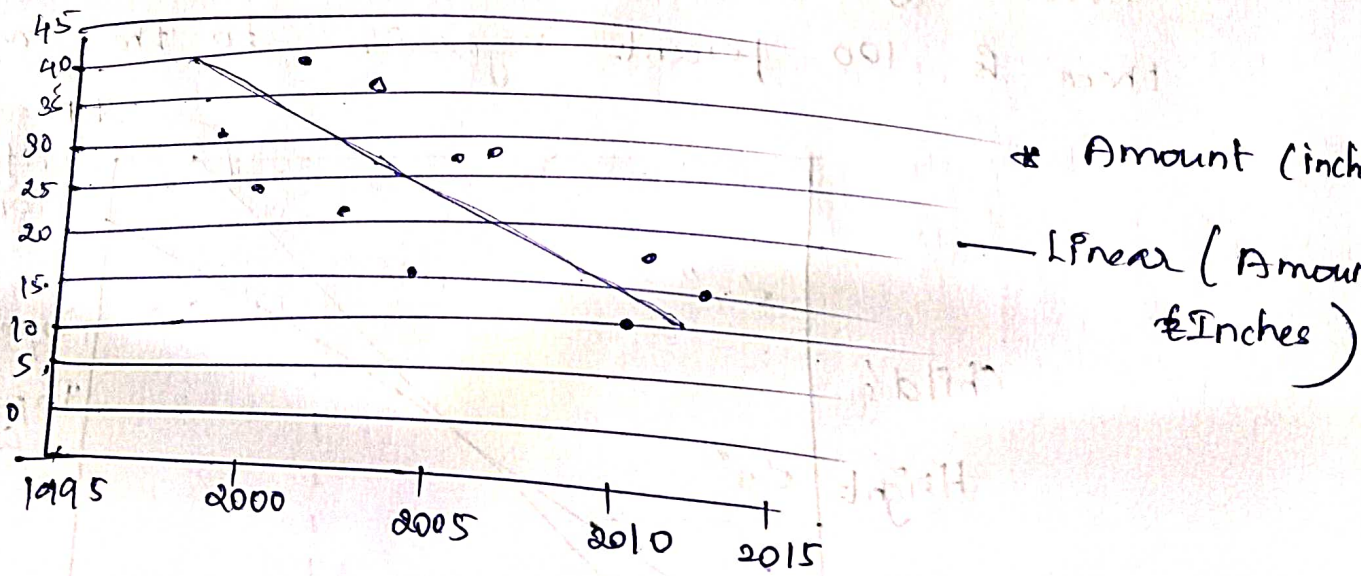
The general statistical rule is that whenever the correlation between two variable is imperfect there will be regression to the mean.

Importance of Regression to the mean.

It is important to minimize instances of bad judgement and address the weak spot in our reasoning.

Regression to the mean fallacy :-

The regression (or) regressive fallacy is an informal fallacy. It assumes that something has returned to normal because of corrective actions this fails to account for natural fluctuations.



Regression analysis.

Why Regression to the mean happen

Regression to the mean usually happens because of sampling error. A good sampling technique is to randomly sample from the population.

Formula for the percent of regression to the mean.

Percent of Regression to the mean = $100(1-r)$

Correlation coefficient and regression to Mean.

$r \rightarrow$ correlation coefficient.

If $r = 1 \longrightarrow$ perfect correlation

then $1-1 = 0$ and the regression to the mean is zero

If your data has perfect correlation, it will regress to the mean. With an r of zero there is 100 percent regression to the mean.

In other words, data with an r of zero will always regress to the mean.

Terminologies related to regression analysis.

1) outliers
2) Multicollinearity
3) Heteroscedasticity
4) Underfitting and overfitting.