BASIC CONCEPT OF STATISTICS

❖ Population

A group of all the units under the study is called a **population**

For example, to study the standard of living workers of factory, a group of all the workers of factory become a population for the study. The total number of units contain in the population is called the size of the population. If the total number of units of the population is finite say N call it finite population.

❖ Sample

A set or group of units selected from the population on the basis of some defined criteria is called a **sample**. The number of units in the sample is term as the sample size. For example, if we selected 150 workers from above population out workers by any statistical method then the group of selected workers is called a sample and the sample size n is equal to 150.

❖ Quantitative Data (Numerical variable)

If the variable characteristic is numerical then it is known as **the numeric variable** and the collection of the observations on the numerical variables it is called as the Quantatived data.

Ex.: - Income of workers, their age, weight, Hight.

❖ Qualitative Data (categorical variable)

If the variable characteristic is non numerical then it is called **Qualitative variable.** We shall call such qualitative characteristic as an attribute.

Ex.: - Gender of workers, their education level, feedback level.

❖ Discrete Data

if a variable can assume definite or countable values within the specific range, then it is called **discrete variable.**

Example: - Number of children per family, Number of accidents per day, number of printed errors per page in book etc example of discrete variables

❖ Continuous Data

If the variable which take all possible values within a given range and its value not countable than it is called **continuous data**

Example: - Height of the students, price of the product temperature of any place etc are the example of continuous data

❖ DESCRIPTIVE STATISTICS

Descriptive Statistics refers to those methods which are used for the collection, presentation as well as analysis of data. These methods relate measurement of central tendencies, measurement of dispersion, measurement of correlation etc. For Example: Descriptive statistics is used when you estimate average height of the secondary students in your school. Descriptive statistics is also used when you find the marks in science and mathematics of the students in all classes are intimately related to each other.

MEASURES OF CENTRAL TENDENCY

❖Measures of Central Tendency

The value of the variable is concentrated around a certain central value this characteristic of data is called as **central tendency**.

The central value around which values of the variables are concentrated is called as **measured of central tendency**. **OR** The value lies within the range of data it is known as a **measure of central tendency**.

❖Ideal or Good Measures of Central Tendency

This ideal measure is given by prof yules.

1) It should be well defined and rigid.
2) It should be based on all the observation of the data.
3) It should be stable measure. If means that values od averages found for different samples of same size from the same population should be almost same.
4) It should not be unduly affected by a few very large or very small observation.
5) It should be affected as little as possible by fluctuations of sampling.

❖ Name the Measures of Central Tendency

There are three the measures of central tendency.

i) Mean
ii) Median
iii) Mode

i) Mean (Arithmetic mean) (Average)

Arithmetic mean of a set of observations is their sum divided by the number of observations.

- **Merits: -**

a) It is rigidly defined.
b) It is easy to understand and easy to calculate.
c) It is based upon all the observations.
d) It is most suitable to algebraic treatment. The mean of the composite series in terms of the means and sizes of the component series.
e) Of all the averages, arithmetic mean is affected least by fluctuations of sampling. This property is sometimes described by saying that arithmetic mean is, a stable average

- **Demerits: -**

a) It cannot be determined by inspection nor it can be located graphically.
b) Arithmetic mean cannot be used if we are dealing with qualitative characteristic which cannot be measured quantitively; such as, intelligence, honesty, beauty, etc. In such cases median (discussed later) is the only average to be used.
c) Arithmetic mean cannot be obtained if a single observation is missing or lost or is illegible unless we drop it out and compute the arithmetic mean of the remaining values.
d) Arithmetic mean is affected very much by extreme values.
e) In extremely asymmetrical (skewed) distribution, usually arithmetic mean is not a suitable measure of location.
f) It cannot be calculated for frequency distribution with open end classes.

FROMULA

No.	Ungrouped data	Grouped data (Discrete & Continuous Data)
1.	$\bar{x} = \frac{\sum x}{n}$	$\bar{x} = \frac{\sum fx}{\sum f}$
2.	$\bar{x} = A + \frac{\sum d}{n}$ where, d = X – A A = Assume no. in X	$\bar{x} = A + \frac{\sum fd}{\sum f}$ where, d = X – A A = Assume no. in X **OR** $\bar{x} = A + \frac{\sum fd}{\sum f} \times c$

EXAMPLE

UNGROUPED DATA

1. The marks obtained by ten students, in an examination are given below. Find mean of the marks. 18, 16, 20, 8, 22, 19, 17, 10, 6, 12

X	18	16	20	8	22	19	17	10	6	12	**148**

$$\bar{x} = \frac{\sum x}{n}$$
$$\bar{x} = \frac{148}{10}$$
$$\bar{x} = 14.8$$

X	15	20	23	24	29	42	55	68	79	58	Total
d = X-A, A = 29	-14	-9	-6	-5	0	13	26	39	50	29	123

2. Find the mean of the following observation.
   15, 20, 23, 24,29, 42, 55, 68, 73, 58.

$$\bar{x} = A + \frac{\sum d}{n}$$
$$\bar{x} = 29 + \frac{123}{10}$$
$$\bar{x} = 29 + 12.3$$
$$\bar{x} = 41.3$$

3. Find the mean of the following observation.
   47, 53, 52, 59, 72, 83, 92, 94, 98, 99 **[Ans : $\bar{x}$ =74.9]**
4. Find the mean of the following observation.
   126, 110, 91, 115, 112, 80, 101, 93, 97, 113 **[Ans : $\bar{x}$ =103.8]**

GROUPED DATA (Discrete Data)

1. The following is the distribution of the number of children in 60 families find the average number of children per family.

Number of children	0	1	2	3	4	5	6
Number of Families	3	8	12	21	10	4	2

Number of children x	Number of families f	fx
0	3	0
1	8	8
2	12	24
3	21	63
4	10	40
5	4	20
6	2	12
Total	**60**	**167**

$$\bar{x} = \frac{\sum fx}{\sum f}$$
$$\bar{x} = \frac{167}{60}$$
$$\bar{x} = 2.78$$

2. Find mean.

Observation	50	51	52	53	54	55	56	57	58
Frequency	1	4	8	13	9	7	4	3	1

observation x	Frequency f	d = X -A A = 54	fd
50	1	-4	-4
51	4	-3	-12
52	8	-2	-16
53	13	-1	-13
54	9	0	0
55	7	1	7
56	4	2	8
57	3	3	9
58	1	4	4
Total	**50**		**-17**

$$\bar{x} = A + \frac{\sum fd}{\sum f}$$
$$\bar{x} = 54 + \frac{-17}{50}$$
$$\bar{x} = 54 + (-0.34)$$
$$\bar{x} = 54 - 0.34$$
$$\bar{x} = 53.66$$

3. Find mean.

Observation	69	89	109	129	149	169	189
Frequency	5	8	9	14	3	1	40

Price x	No. of shops f	$d = \frac{X - A}{c}$, A = 129, c= 20	fd
69	5	-3	-15
89	8	-2	-16
109	9	-1	-9
129	14	0	0
149	3	1	3
169	1	2	2
189	40	3	120
Total	**75**		**100**

$$\bar{x} = A + \frac{\sum fd}{\sum f} \times c$$

$$\bar{x} = 129 + \frac{100}{75} \times 20$$

$$\bar{x} = 129 + 1.33 \times 20$$

$$\bar{x} = 129 + 26.6$$

$$\bar{x} = 155.6$$

4. The price of a item change from shop to shop. The following data are available find the mean price

Price	206	212	218	220	224	230
No. of shops	5	8	9	14	3	1

[Ans : $\bar{x}$ =216.75]

5. The following is the distribution is obtained 60 students od standard 11. find the mean for the distribution.

Age	15	16	17	18	19	20	21
Number of Students	3	10	15	20	8	3	1

[Ans : $\bar{x}$ =17.55]

6. Find mean

Observation	20	19	18	17	16	15	14	13
Frequency	3	10	21	25	19	15	5	2

[Ans : $\bar{x}$ =16.78]

GROUPED DATA (Continuous Data)

1. Calculate the mean of the following frequency distribution.

Class-interval	0-8	8-16	16-24	24-32	32-40	40-48
Frequency	8	7	16	24	15	7

Class-interval	Frequency f	x	d = X - A, A = 28	fd
0-8	8	4	-1.2	-9.6
8-16	7	12	-0.8	-5.6
16-24	16	20	-0.4	-6.4
24-32	24	28	0	0
32-40	15	36	0.4	6
40-48	7	44	0.8	5.6
Total	**77**			**-10**

$$\bar{x} = A + \frac{\sum fd}{\sum f}$$

$$\bar{x} = 28 + \frac{(-10)}{77}$$

$$\bar{x} = 28 + (-0.13)$$

$$\bar{x} = 27.87$$

2. The frequency distribution of the marks in statistics of 100 students is given below. Obtain the average marks from it.

Marsk	5-14	15-24	25-34	35-44	45-54	55-64	65-74
No. of students	8	11	15	25	20	13	8

Marks	No of students f	x	$d = \frac{X - A}{c}$, A = 39.5, c= 10	fd
5-14	8	9.5	-3	-24
15-24	11	19.5	-2	-22
25-34	15	29.5	-1	-15
35-44	25	39.5	0	0
45-54	20	49.5	1	20
55-64	13	59.5	2	26

65-74	8	69.5	3	24
Total	**100**			**9**

$$\bar{x} = A + \frac{\sum fd}{\sum f} \times c$$

$$\bar{x} = 39.5 + \frac{9}{100} \times 10$$

$$\bar{x} = 39.5 + 0.09 \times 10$$

$$\bar{x} = 39.5 + 0.9$$

$$\bar{x} = 40.4$$

3. Obtain the mean of the following frequency distribution.

Class	0-10	10-20	20-30	30-40	40-50	50-60	60-70
Frequency	1	7	15	27	20	8	2

[Ans : $\bar{x}$ =36.25]

4. Calculate the mean of the following frequency distribution

Class	0-2	2-5	5-10	10-20	20-50	50-100	100-200	200-500	500-1000
Frequency	2	6	8	20	45	22	18	6	3

[Ans : $\bar{x}$ =81.98]

5. The data about weight of mangoes from a tree are as follows. Find the mean of weight mangoes.

Weight of mangoes	410-420	420-430	430-440	440-450	450-460	460-470
No. of mangoes	17	13	15	20	30	33

[Ans : $\bar{x}$ =441.72]

6. Find the mean of the following frequency distribution

Class	2-6	6-10	10-14	14-18	18-22	22-26
Frequency	6	10	20	15	7	2

[Ans : $\bar{x}$ =12.87]

MISSING FREQUENCY

1. If the mean of the following frequency distribution is 122.7 then find the missing frequency.

Class	60-79	80-99	100-119	120-139	140-159	160-179	180-199
Frequency	7	4	?	18	5	5	2

Class	Frequency f	x	$d = \frac{X - A}{c}$, A = 109.5, c= 20	fd
60-79	7	69.5	-2	-14
80-99	4	89.5	-1	-4
100-119	F1	109.5	0	0
120-139	18	129.5	1	18
140-159	5	149.5	2	10
160-179	5	169.5	3	15
180-199	2	189.5	4	8
Total	**41 + F1**			**33**

Mean value $\bar{x}$ **= 122.7 is given**

$$\bar{x} = A + \frac{\sum fd}{\sum f} \times c$$
$$122.7 = 109.5 + \frac{33}{41 + F1} \times 20$$
$$122.7 - 109.5 = \frac{660}{41 + F1}$$
$$13.2 = \frac{660}{41 + F1}$$
$$13.2(41 + F1) = 660$$
$$541.2 + 13.2F1 = 660$$
$$13.2F1 = 660 - 541.2$$
$$13.2F1 = 118.8$$
$$F1 = \frac{118.8}{13.2}$$
$$F1 = 9$$

So the frequency of class 110 – 119 is 9.

2. If the mean of the frequency distribution is 27.75 then find the missing frequency.

Class	0-10	10-20	20-30	30-40	40-50
Frequency	4	6	?	17	4

ii) Geometric mean

Geometric mean of a set of n observations is the n^{th} root of their product. It is denoted by G

Uses: -

a) Suppose we study a variable that change over time. If you want to find average rate change the variable, the application of the arithmetic mean is not appropriate that time we use the geometric mean.
b) To find the rate of population growth and the rate of interest.
c) It is appropriate for averaging the ratio of the change, for average of the proportion, etc.
d) It is most suitable average for index number.

- **Merits: -**

a) It is rigidly defined.
b) It is based upon all the observation.
c) It is suitable for further mathematical treatment.
d) It is not affected much by fluctuations of sampling.
e) It gives comparatively more weight to small items.
f) It is least affected by extreme values

- **Demerits: -**

a) If anyone of the observations is zero, geometric mean becomes zero.
b) If any one or more values are negative, either geometric mean will not be calculated or absurd value will be obtained.
c) Its calculation in somewhat complicated.

FROMULA

$$G = \sqrt[n]{x_1 \times x_2 \times x_3 \times \dots \dots \times x_n}$$

EXAMPLE

1. The population of an area increased by 15% ,18%, 13% and 20 % respectively for four years. Find the average rate increase in the population. **[Ans. : G = 16.47%]**
2. The value of machine depreciates at the rate of 10%, 7%, 5% and 2% in its first four years respectively. Find the average rate of depreciation using an appropriate method. **[Ans. : G = 6.05%]**
3. Find the geometric mean from the following observation.
   12,18,32,36 **[Ans. : G = 22.33]**

iii) Harmonic mean

Harmonic mean of number of observations is the reciprocal of the arithmetic mean of the reciprocals of the given values.

Harmonic mean is the inverse of the arithmetic mean of the reciprocals of the observations of set.

Uses: -

Harmonic mean is suitable when the values are pertaining to the rate of change per unit time such as speed, number of items produced per day, contracts completed per year, etc. In general, harmonic mean is suitable for time, speed, rates, prices, etc.

- **Merits: -**

a) It is based on all the observations of a set.
b) It is a good mean for a highly variable series.
c) It gives more weightage to the small values and less weightage to the large values.
d) It is better than weighted mean since in this value are automatically weighted.
e) It is suitable for further mathematical treatment.
f) It is useful only when small items have to be given a greater weightage.

- **Demerits: -**

a) Harmonic mean it is not easily understood and is difficult to compute.
b) If the any value is zero, it cannot be calculated.

FROMULA

No.	Ungrouped data	Grouped data (Discrete & Continuous Data)
1.	$H.M = \frac{n}{\sum\left(\frac{1}{x}\right)}$	$H.M = \frac{\sum f}{\sum\left(\frac{f}{x}\right)}$

EXAMPLE

UNGROUPED DATA

1. Find the harmonic mean from the following observation.
   2574, 475, 75, 5, 0.8, 0.08, 005, 0.0009 **[Ans. : H.M. = 0.006]**
2. Calculated the harmonic mean from the following observation.

3834, 382, 63, 8, 0.4, 0.03, 0.009,0.0005 **[Ans. : H.M. = 0.0037]**

GROUPED DATA (Discrete Data)

1. From the following data computer value of harmonic mean.

Marks	10	20	25	40	50
Number of students	20	30	50	15	5

[Ans. : H.M. = 20.08]

2. From the following data obtained from the survey. computer harmonic mean.

Speed of the car	130	135	140	145	150
Number of cars	20	30	50	15	5

[Ans. : H.M. =]

GROUPED DATA (Continuous Data)

1. From the following data computer value of harmonic mean.

Class	10-20	20-30	30-40	40-50	50-60
Frequency	4	6	10	7	3

[Ans. : H.M. =]

2. From the following data computer value of harmonic mean.

Class	2-6	6-10	10-14
Frequency	10	12	18

[Ans. : H.M. =]

iv) Median

Median is defined as the value of Middle east observation when the data are arranged in the either ascending or descending order. It is most preferable measure of the location for asymmetric distribution.

- **Merits: -**

a) Median is the only average to be used while dealing with qualitative data.
b) It is rigidly defined.
c) It is easily understood and easily calculated.
d) It is not at all affected by its value.
e) It can be calculated for a distribution with open index class.
f) Median can be located even if the data are uncompleted.

- **Demerits: -**

a) In case of the even number of observation median cannot be determined exactly We merely estimated it by taking the mean of the two middle terms.
b) It is not based on the all observations.
c) It is not used for further algebraic treatment.
d) As compared with mean, it is affected much by fluctuation of the sampling.

FROMULA

No.	Ungrouped data	Grouped data (**Discrete**)
1.	$M = \left(\frac{n+1}{2}\right)^{th} observation$	$M = \left(\frac{n+1}{2}\right)^{th} observation$ $n = \sum f$
		Grouped data (**Continuous)** $class\ of\ M = \left(\frac{n}{2}\right)^{th} observation$ $M = L + \frac{\left(\frac{n}{2}\right) - cf}{f} \times c$ Where, L = lower boundary point of Midian class Cf = cumulative frequency of the class prior to . median class

		F = frequency if median class C = length of median class

EXAMPLE

UNGROUPED DATA

1. The weight of 9 Pearsons are as follows. Find median weight.
   52, 48, 60, 50, 68, 51, 55, 62, 58
   Ans: - Arrangement of observation in ascending form is as follows.
   48, 50, 51, 52, 55, 58, 60, 62, 68.

$$\text{Median } M = \left(\frac{n+1}{2}\right)^{th} observation$$
$$M = \left(\frac{9+1}{2}\right)^{th} observation$$
$$M = (5)^{th} observation$$
$$M = 55\ kg$$

2. The marks obtained by 12 students, in an examination are given below. Find median of the marks. 4, 6, 5, 8, 12, 7, 10, 5, 15, 9, 10, 11
   Ans: - Arrangement of observation in ascending form is as follows.
   4, 5, 5, 6, 7, 8, 9, 10, 10, 11, 12, 15.

$$\text{Median } M = \left(\frac{n+1}{2}\right)^{th} observation$$
$$M = \left(\frac{12+1}{2}\right)^{th} observation$$
$$M = (6.5)^{th} observation$$
$$M = \frac{6^{th}\ obs. + 7^{th}\ obs.}{2}$$
$$M = \frac{8\ +9}{2}$$
$$M = 8.5\ marks$$

3. The number of items produced by factory in different weeks are 85, 92, 68, 80, 72, 63, and 55 find the median of the production. **[Ans. : M = 80]**
4. Find the median of the following observation.
   13, 17, 22, 8, 19, 14, 20, 6 **[Ans. : M = 15.5]**
   Find median. 8, 6, 7, 0, 2, 4, 6, 5, 5, 4, 8, 9, 3, 6, 7. **[Ans. : M = 6]**

5.

GROUPED DATA (Discrete Data)

1. The frequency distribution of the number of mistakes committed by 335 candidates in typing examination is given below find the median of the distribution

Number of mistakes	0	1	2	3	4	5	6	7	8
No. of candidates	35	82	107	40	31	15	12	9	4

No. of mistakes	No. of candidates	CF
0	35	35
1	82	117
2	107	224
3	40	264
4	31	295
5	15	310
6	12	322
7	9	331
8	4	335
Total	**335**	

$$M = \left(\frac{n+1}{2}\right)^{th} observation$$
$$M = \left(\frac{335+1}{2}\right)^{th} observation$$
$$M = (168)^{th} observation$$

From the column of CF it is seen that the value of observation numbers from 118 to 224 is 2. hence the value of 168 observation is also 2.

M = 2 mistakes

2. The time require for typing or report by the different typists is given in the following data. Find the median typing time using it.

Time for typing	10	11	12	13	14
No. of typists	5	7	8	15	5

Time for typing	No. of typists	CF
10	5	5
11	7	12
12	8	20
13	15	35
14	5	40
Total	**40**	

$$M = \left(\frac{n+1}{2}\right)^{th} observation$$

$$M = \left(\frac{40+1}{2}\right)^{th} observation$$

$$M = (20.5)^{th} observation$$

$$M = \frac{20^{th}\ obs. + 21^{th}\ obs.}{2}$$

$$M = \frac{12\ + 13}{2}$$

$$M = 12.5\ marks$$

3. We have the following information from the survey of hundred customers from the bank their number of visits to bank during the month

No. of visits	0	1	2	3	4	5	6
No. of customer	12	22	40	15	6	4	1

No. of visits	No. of customer	CF
0	12	12
1	22	34
2	40	74
3	15	89
4	6	95
5	4	99
6	1	100
Total	**100**	

$$M = \left(\frac{n+1}{2}\right)^{th} observation$$

$$M = \left(\frac{100+1}{2}\right)^{th} observation$$

$$M = (50.5)^{th} observation$$
$$M = 2$$

4. The following table shows the record of the absent students of the class during a month. Find the median of the number of absent days per students

No. of absent days of students	0	1	2	3	4	5
No. of students	8	12	18	9	5	1

[Ans. : M = 2 days]

5. The following table give ages 80 students selected from a college.

Age	17	18	19	20	21	22	23
No. of Students	11	14	22	15	8	6	12

[Ans. : M = 19 students]

GROUPED DATA (Continuous Data)

1. Calculate the median of the following frequency distribution.

Class-interval	10-29	30-49	50-69	70-89	90-109	110-129
Frequency	13	22	48	57	25	5

Class-interval	Frequency	CF
10-29	13	13
30-49	22	35
50-69	48	83
70-89	57	140
90-109	25	165
110-129	5	170
Total	**170**	

$$class\ of\ M = \left(\frac{n}{2}\right)^{th} observation$$
$$= \left(\frac{170}{2}\right)^{th} observation$$
$$= (85)^{th} observation$$

From the cf column 85th observation lies in the class 70-89.

Taking limit points median class = 70 – 89

= 69.5 – 89.5

$$M = L + \frac{\left(\frac{n}{2}\right) - cf}{f} \times c$$
$$M = 69.5 + \frac{85 - 83}{57} \times 20$$
$$M = 69.5 + \frac{40}{57}$$
$$M = 69.5 + 0.70$$
$$M = 70.20$$

2. The data about monthly expenditure on petrol for 75 families is given in the following table. Find the medial expenditure on petrol for these families.

Expenditure on petrol	Upto 200	Upto 400	Upto 600	Upto 800	Upto 1000	Upto 1200
No. of families	2	8	17	32	57	75

Expenditure on petrol	No. of families	CF
upto 200	2	2
200-400	6	8
400-600	9	17
600-800	15	32
800-1000	25	57
1000-1200	18	75
Total	**75**	

$$class\ of\ M = \left(\frac{n}{2}\right)^{th} observation$$
$$= \left(\frac{75}{2}\right)^{th} observation$$
$$= (37.5)^{th} observation$$

From the cf column 37.5^{th} observation lies in the class 800-1000.

Taking limit points median class = 800 - 1000

$$M = L + \frac{\left(\frac{n}{2}\right) - cf}{f} \times c$$
$$M = 800 + \frac{37.5 - 32}{25} \times 200$$
$$M = 800 + \frac{1100}{25}$$
$$M = 800 + 44$$
$$M = 844$$

3. The level of air pollution in a city on different days is as follows find the median level of pollution.

Level of pollution	250 and above	270 and above	290 and above	310 and above	320 and above	330 and above	340 and above
No. of days	150	133	108	76	41	20	7

Level of pollution	No. of days	CF
250-270	17	17
270-290	25	42
290-310	32	74
310-320	35	109
320-330	21	130
330-340	13	143
340 & above	7	150
Total	**150**	

$$class\ of\ M = \left(\frac{n}{2}\right)^{th}\ observation$$
$$= \left(\frac{150}{2}\right)^{th}\ observation$$
$$= (75)^{th} observation$$

From the cf column 75th observation lies in the class 310-320.

Taking limit points median class = 310-320

$$M = L + \frac{\left(\frac{n}{2}\right) - cf}{f} \times c$$

$$M = 310 + \frac{75 - 74}{35} \times 10$$
$$M = 310 + \frac{10}{35}$$
$$M = 310 + 0.2857$$
$$M = 310.29$$

4. Find median from the following frequency distribution.

class	Less than 20	20-40	40-60	60-80	80-100	More than 100
No. of students	8	11	15	25	20	13

[Ans. : M = 53.33]

5. Find median age.

Age	55-60	50-55	45-50	40-45	35-40	30-35	25-30	20-25
No. of Pearsons	17	13	15	20	30	33	28	14

[Ans. : M = 35.83 year]

6. Find median of the following distribution.

Class	118-126	127-135	136-144	145-153	154-162	163-171	172-180
Frequency	17	13	15	20	30	33	28

[Ans. : M = 146.75]

7. Use the following guitar to find a median salary of employees in a firm.

salary	5 or more	10 or more	15 or more	20 or more	25 or more	30 or more
No. of employees	150	133	108	76	41	20

[Ans. : M = 21.78]

8. The following table shows the monthly expense for entire any of group of hundred students find the median of these expense

Expense	Less than 200	Less than 400	Less than 600	Less than 700	Less than 800	Less than 900
No. of students	8	31	71	88	95	100

[Ans. : M = 495]

MISSING FREQUENCY

1. Following is an incomplete frequency distribution marks obtained by the 120 students in any one examination if the median score of the student is 75 then find the missing frequency

Marks	30-39	40-49	50-59	60-69	70-79	80-89	90-99
Students	1	3	?	20	?	33	9

Marks	Students	CF
30-39	1	1
40-49	3	4
50-59	F1	4 + F1
60-69	20	24 + F1
70-79	F2	24 + F1+ F2
80-89	33	57 + F1 + F2
90-99	9	66 + F1 + F2
Total	**120**	

Here, n = 120

66 + F1 + F2 = 120
F1 + F2 = 120 – 60
F1 + F2 = 54
F1 = 54 - F2 ……. eq. (1)

M = 75 is given
class of M = 70 - 79
= 69.5 – 79.5

$$M = L + \frac{\left(\frac{n}{2}\right) - cf}{f} \times c$$

$$75 = 69.5 + \frac{60 - (24 + f1)}{f2} \times 10$$

$$75 - 69.5 = \frac{60 - 24 - f1}{f2} \times 10$$

$$5.5 = \frac{36 - f1}{f2} \times 10$$

$$5.5f2 = (36 - f1) \times 10$$

$$5.5f2 = 360 - 10f1 \quad \ldots\ldots eq(2)$$

Now, from equation (1) f1 = 54 – f2 put in the equation (2)

$$5.5f2 = 360 - 10(54 - f2)$$
$$5.5f2 = 360 - 540 + 10f2$$
$$5.5f2 - 10f2 = 360 - 540$$
$$5.5f2 - 10f2 = 360 - 540$$
$$-4.5f2 = -180$$
$$f2 = \frac{-180}{-4.5}$$
$$f2 = 40$$

Now $f2 = 40$ value put in equation (1)

$$f1 = 54 - f2$$
$$f1 = 54 - 40$$
$$f1 = 14$$

2. If the median of the frequency distribution is 146.75 then find the missing frequency.

Length	118-126	127-135	136-144	145-153	154-162	163-171	172-180	Total
Frequency	10	20	?	40	?	25	16	40

[Ans. : f1 = 12 & f2 = 5]

3. If the median of the frequency distribution is 35 and total frequency = 170 then find the missing frequency.

salary	0-10	10-20	20-30	30-40	40-50	50-60	60-70
No. of workers	10	20	?	40	?	25	16

[Ans. : f1 = 35 & f2 = 24]

v) Mode

The value which data repeated maximum number of I term both the value occurring with maximum frequency in a given data is called a mode.

- **Merits: -**

a) Mode is readily comprehensible and easy to calculate. Like median, mode can be located in some cases merely by inspection.
b) Mode is not at all affected by extreme values.
c) Mode can be conveniently located even if the frequency distribution has class-intervals of unequal magnitude provided the modal class and the classes preceding

and succeeding it are of the same magnitude. Open end classes also do not pose any problem in the location of mode.

- **Demerits: -**

a) Mode is ill-defined. It is not always possible to find a clearly defined mode. In Some cases, we may come across distributions with two modes. Such distributions are called bi-modal. If a distribution has more than two modes, it is said to be multimodal.
b) It is not based upon all the observations.
c) It is not capable of further mathematical treatment.
d) As compared with mean, mode is affected to a greater extent by fluctuations of sampling.

FROMULA

UNGROUPED DATA

M_0 = The observation which is required for maximum number of times

GROUPED DATA(DISCREATE)

M_0 = maximum frequency of the distribution

GROUPED DATA(CONTINUOUS)

$$M_0 = L + \frac{f_m - f_1}{2f_m - f_1 - f_2} \times c$$

Where,

$$L = Lower\ limit\ of\ the\ modal\ class$$
$$f_m = Frequency\ of\ modal\ class$$
$$f_1 = Frequency\ of\ the\ class\ preceding\ the\ model\ class$$
$$f_2 = Frequency\ of\ the\ class\ succeeding\ the\ model\ class$$
$$c = class\ lengh\ of\ the\ modal\ class$$

- **The above formula we can not used when following condition appear in the data**
  a) More than one observation in a frequency distribution appears with highest frequency
  b) The continuous distribution has classes of unequal length.
  c) The frequency distribution is mixed distribution that is a part of it is discrete and the rest is continuous

d) The right or left end of the curve of the frequency distribution is too extended.

We have observed that mode is not well, defined in many cases. The noted statistician Karl Pearson established a relation between mean, median and mode by studying their vales for different data sets, He observed that for data that are not evenly distributed around average, difference between mean and mode is approximately 3 times the difference between mean and median.

That is, (Mean - Mode) =3 (Mean - Median)

The following formula is obtained to find the mode using these relations

Mode - 3 (Median) - 2(Mean)

This is written in notations as $M_0 = 3M - 2\bar{x}$

This formula to find mode is called as an empirical formula because the value obtained from observation and not from the theory. The value of mode found using this formula can be negative if the frequency distribution is not evenly distributed around the average.

EXAMPLE

UNGROUPED DATA

1. The number of accidents occur during last 15 days in a city are given below. obtain mode of the series
   2, 3, 0, 1, 4, 0, 1, 3, 1, 2, 5, 4, 1, 0, 1
   Ans: - Here observation 1 is repeated for maximum number of times
   Mode Mo = 1
2. The following figures are the number of mistakes committed by typist in different pages. find mode of the mistake.
   2, 1, 3, 3, 4, 2, 0, 3, 2, 1, 2, 0, 3, 5, 4 **[Ans. : M_0 = 1 & 2]**
3. The number of books purchased by each of the 15 persons from the bookstore are followers
   1, 0, 2, 2, 3, 4, 2, 7, 2, 2, 5, 4, 2, 1, 2. **[Ans. : M_0 = 2]**

GROUPED DATA (Discrete Data)

1. Find the mode of the following frequency distribution.

Observation	10	11	12	13	14	15	16
Frequency	15	30	38	52	40	12	10

Ans: - In the given frequency distribution maximum frequency is 52 and it is against observation 13

Mode Mo = 13

2. Find the mode of the following frequency distribution.

Observation	10	11	12	13	14	15	16	17	18
Frequency	3	10	19	28	13	11	7	2	1

[Ans. : M_0 = 13]

3. The Number of trips made by 24 taxis driver in our day are shown in the following data find the mode number of trip.

No. of trips	1	2	4	5	6	7
No. of drivers	3	7	4	7	2	1

[Ans. : M_0 = 2 & 5]

GROUPED DATA (Continuous Data)

1. The following table gives output of workers in a factory find the mode output.

Output	150–160	160-170	170-180	180-190	190-200	200-210	210-220	220-230
No. of workers	4	5	19	33	48	22	12	6

Ans: - The maximum frequency 48 is in the class 190 – 200

L = 190, f_m = 48, f_1 = 33, f_2 = 22, C = 10

$$M_0 = L + \frac{f_m - f_1}{2f_m - f_1 - f_2} \times c$$

$$M_0 = 190 + \frac{48 - 33}{2(48) - 33 - 22} \times 10$$

$$M_0 = 190 + \frac{150}{41}$$

$$M_0 = 190 + 3.66$$

$$M_0 = 193.6$$

2. Find mode from the following frequency distribution.

Class	10-14	15-19	20-24	25-29	30-34	35-39
No. of students	12	24	35	15	8	2

[Ans. : M_0 = 21.27]

3. The table below shows the data about the weight of 86 apples from a garden find the mood from the weight of apples.

Weight of apples	120–130	130-140	140-150	150-160	160-170	170-180	180-190
No. of Apples	8	13	19	23	10	8	5

[Ans. : M_0 = 152.35]

4. Find mode from the following frequency distribution.

Class	0-5	5-10	10-20	20-30	30-50	50-70	70-100
Frequency	4	7	12	20	18	12	7

Class	Frequency	x	$d = \frac{X - A}{c}$, A = 25, c= 5	fd	CF
0-5	4	2.5	-4.5	-18	4
5-10	7	7.5	-3.5	-24.5	11
10-20	12	15	-2	-24	23
20-30	20	25	0	0	43
30-50	18	40	3	54	61
50-70	12	60	7	84	73
70-100	7	85	12	84	80
Total	**80**			**155.5**	

Mean

$$\bar{x} = A + \frac{\sum fd}{\sum f} \times c$$
$$\bar{x} = 25 + \frac{155.5}{80} \times 5$$
$$\bar{x} = 25 + 1.94 \times 5$$
$$\bar{x} = 25 + 9.7$$
$$\bar{x} = 34.7$$

Median

$$class\ of\ M = \left(\frac{n}{2}\right)^{th}\ observation$$
$$= \left(\frac{80}{2}\right)^{th}\ observation$$

$$= (40)^{th} observation$$

From the cf column 40 observation lies in the class 20 - 30.

Taking limit points median class = 20 - 30

$$M = L + \frac{\left(\frac{n}{2}\right) - cf}{f} \times c$$
$$M = 20 + \frac{40 - 23}{20} \times 10$$
$$M = 20 + \frac{170}{20}$$
$$M = 20 + 8.5$$
$$M = 28.5$$

Mode

$$\boldsymbol{M_0 = 3M - 2\bar{x}}$$
$$\boldsymbol{= 3(28.7) - 2(34.7)}$$
$$\boldsymbol{= 85.5 - 69.44}$$
$$\boldsymbol{= 16.06}$$

5. The time between placing and order and it is delivery was noted for the certain wholesaler as follows find the mode of this time.

Time	20-25	25-30	30-35	35-40	40-45	45-50	50-55
No. of orders	2	5	7	5	6	7	3

[Ans. : $\bar{x}$ = 38.36, M = 38.5, M_0 = 38.78]

6. Find mode from the following frequency distribution.

Age	50-60	60-65	65-70	70-85	85-100
No. of persons	6	10	19	9	4

1 **[Ans. : $\bar{x}$ = 68.85, M = 67.11, M_0 = 63.63]**

MISSING FREQUENCY

1. The incomplete frequency distribution of monthly expenditure detail of 360 families is as follows. If their mode expenditure his rupees 1376 then find the missing frequency.

Expense	Less than 400	400-800	800-1200	1200-1600	1600-2000	2000-2400	2400-2800	2800 and more
No. Of families	14	22	?	124	?	32	15	5

[Ans. : f1 = 80 & f2 = 68]

2. If the mode of the frequency distribution is 341. then find the missing frequency.

Length	200-210	210-220	220-230	230-240	240-250	250-260	260-270	Total
Frequency	4	16	?	?	40	6	4	230

[Ans. : f1 = 60 & f2 = 100]

3. If the mode of the frequency distribution is 74 and total frequency = 163 then find the missing frequency.

Class	40-50	50-60	60-70	70-80	80-90	90-100	100-110	110-120
No. of workers	4	12	?	50	?	13	9	4

[Ans. : f1 = 38 & f2 = 32.6]

MEASURES OF DISPERSION

❖ Meaning of dispersion.

A measure which shows how far the observations of the data are scattered from the measure of average is terms as dispersion.

The degree to which numerical data trend to spread about an average value is called variation or dispersion of the data.

Dispersion or spread is the degree of the scatter or variation of the items.

❖ Characteristics for an Ideal measure of Dispersion

i) It should he rigidly defined.
ii) It should be easy to calculate and easy to understand.
iii) It should be based on all the observations.
iv) It should be amenable to further mathematical treatment.
v) It should be affected as little as possible by fluctuations of sampling.

❖ Concept of Absolute Measure and Relative measure

➢ Absolute Measure: -

A measure of dispersion which is expressed in the same unit in which the observation of the data is expressed is called an absolute measure of dispersion.

Example: - Height in cm, weight in kg, Income in rupees, etc.

In this case two series having different unit of dispersion cannot be compered. Absolute measure is not free from the unit measurement.

➢ Relative measure: -

A measure of dispersion which is free from the unit of measurement is called relative measure of dispersion. It is also known as coefficient of dispersion.

The variability of two or more sets of data having different units of measurement can compared only by relative measure of dispersion.

It is independent of units of measurement of the observation of data.

❖ Measures of dispersion

i) Range
ii) Standard deviation
iii) Variance

Range and quartile deviation are the positional measure of dispersion while mean deviation and standard deviation are known as the measures of dispersion representing the summary of devotions of observations from the measure of central tendency

i) Range

The difference between the largest and smallest values of a set of data is called its range.

Range = largest value - lowest value

Coefficient of range = largest value - lowest value / largest value + lowest value

Lesser the Coefficient of range, better it is.

Range is useful in statistical quality control for the construction of control charts which measure the variability within the samples taken from the production. If variation is not very large then range is useful in measuring variation in money rates, exchange rates, share prices etc. In our day-to-day problem like 'daily sales in a supermarket', 'temperature of a city', 'expense of petrol by two-wheeler or car', etc. are generally expressed in the form of the interval in which it lies, and range of the data can be known from it.

Merits: -

a) It is the easiest measure of dispersion.
b) It can always be found out visually.
c) It is one of the largely used measure of dispersion.

Demerits: -

a) It depends on two extreme values of a series. Thus, it gives no information about the observations lying between smallest and largest values.
b) It is highly susceptible to sampling fluctuations.
c) It is not suitable for further mathematical treatment.
d) It cannot be calculated for the frequency distribution having open-ended classes.

FORMULA

Absolute Range = $x_H - x_L$

Relative Range = $\frac{x_H - x_L}{x_H + x_L}$

EXAMPLE

UNGROUPED DATA

1. The run score by batsmen in his last 10 innings of cricket matches are given below. Find the range and relative range of run.
   48, 75, 37, 52, 93, 81, 25, 72, 18, 60.
   Ans: - $x_H = 93 \quad x_L = 18$

$$\text{Absolute Range} = x_H - x_L$$
$$= 93 - 18$$
$$= 75$$

$$\text{Relative Range} = \frac{x_H - x_L}{x_H + x_L}$$
$$= \frac{93 - 18}{93 + 18}$$
$$= \frac{75}{111}$$
$$= 0.68$$

2. The following data refers to the height of 10 students of the class. Find the range and relative range of height of students.
   162, 145, 170, 181, 167, 151, 175, 185, 169, 156.

GROUPED DATA (DISCRETE)

1. Find absolute range and relative range of the following data.

Observation	100	101	102	103	104	105	106	107
Frequency	8	21	36	50	28	20	10	5

Ans: - $x_H = 100 \quad x_L = 107$

$$\text{Absolute Range} = x_H - x_L$$
$$= 107 - 100$$
$$= 7$$

$$\text{Relative Range} = \frac{x_H - x_L}{x_H + x_L}$$

$$= \frac{107 - 100}{107 + 100}$$

$$= \frac{7}{207}$$

$$= 0.0034$$

2. A bus company has 77 busses who travel in the city. The information of the number of passengers in the bus on the particular days at a particular time is given below. Find the range and the relative range of the number of passengers.

No. of passengers	2	7	10	18	25	30	37
No. of buses	1	4	11	17	23	16	5

GROUPED DATA (CONTINUOUS)

1. Find absolute range and relative range of the following data.

Class	44-46	46-48	48-50	50-52	52-54	54-56
Frequency	7	15	35	30	8	2

Ans: - $x_H = 56 \quad x_L = 44$

$$\text{Absolute Range} = x_H - x_L$$

$$= 56 - 44$$

$$= 12$$

$$\text{Relative Range} = \frac{x_H - x_L}{x_H + x_L}$$

$$= \frac{56 - 44}{56 + 44}$$

$$= \frac{12}{100}$$

$$= 0.12$$

2. Find absolute range and relative range of the following data.

Class	10-15	15-20	20-25	25-30	30-35
Frequency	8	15	26	47	4

3. Find the range and the relative range of the marks from the following frequency distribution of the Marks score by 50 students of the school in certain examination.

Marks	50-59	60-69	70-79	80-89	90-99
No. of students	2	15	23	6	4

ii) Standard Deviation (S.D.) & Variance

We have seen that the definition of mean deviation is based on the absolute value of the deviation of observations of the data from the mean. Since the algebraic sign of the observation are ignored, mean deviation is less used in advanced study of statistics. This limitation of Mean deviation is overcome by an important measure of dispersion known as the Standard Deviation. Instead of talking the absolute value of deviation each observation from the mean the square of the deviation is taken. If the sum of the square of this deviation is divided by the total number of observations, we get an important measure of dispersion known as the variance. This positive square root of the variance is called as the standard deviation.

Well known statistician Karl Pearson defined the "standard deviation as the standard deviation is the positive square root of the mean of the square of the deviation's measures from the mean."

After mean, standard deviation is another very useful measure which gives information about values of the observation of a population.

Standard deviation is also known as mean error, mean square error or root mean square deviation from mean. As a matter of fact, S.D. is nothing but the quadratic mean of the deviation from the data.

Merits: -

a) All the observation are used in its computation.
b) Standard deviation is the most effective and widely used measure of dispersion among all the measures of dispersion.
c) Standard deviation is a suitable measure for algebraic calculation.
d) Standard deviation carries great importance in sampling methods.
e) It is least sensitive to sampling fluctuations.
f) With the help of S.D., it is possible to ascertain the area under the normal curve.
g) It has great utility in testing of hypotheses which other measures of dispersion hardly do.

Demerits: -

a) The extreme observations get undue importance in the value this measure.
b) It cannot be obtained if the frequency distributions have open-ended classes.

FROMULA

No.	Ungrouped data	Grouped data (Discrete & Continuous Data)
	Standard Deviation Formula	
1.	$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n}}$	$s = \sqrt{\frac{\sum f(x-\bar{x})^2}{n}}$
2.	$s = \sqrt{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2}$	$s = \sqrt{\frac{\sum fx^2}{\sum f} - \left(\frac{\sum fx}{\sum f}\right)^2}$
3.	$s = \sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2}$	$s = \sqrt{\frac{\sum fd^2}{\sum f} - \left(\frac{\sum fd}{\sum f}\right)^2}$ $\boldsymbol{whene\ d = x - A}$ $s = \sqrt{\frac{\sum fd^2}{\sum f} - \left(\frac{\sum fd}{\sum f}\right)^2} \times C$ $\boldsymbol{whene\ d = \frac{x-A}{c}}$
	Variance Formula	
1.	$s^2 = \frac{\sum(x-\bar{x})^2}{n}$	$s^2 = \frac{\sum f(x-\bar{x})^2}{n}$
2.	$s^2 = \frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2$	$s^2 = \frac{\sum fx^2}{\sum f} - \left(\frac{\sum fx}{\sum f}\right)^2$
3.	$s^2 = \frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2$	$s^2 = \frac{\sum fd^2}{\sum f} - \left(\frac{\sum fd}{\sum f}\right)^2$ $\boldsymbol{whene\ d = x - A}$ $s^2 = \frac{\sum fd^2}{\sum f} - \left(\frac{\sum fd}{\sum f}\right)^2 \times C$ $\boldsymbol{whene\ d = \frac{x-A}{c}}$

EXAMPLE

UNGROUPED DATA

1. The run score by a batsman in his last seven matches are given below 52, 58, 40, 60, 54, 38, 48. find the variance and the runs of batsmen and also find the standard deviation

X	$x - \bar{x}$	$(x - \bar{x})^2$
52	2	4
58	8	64
40	-10	100
60	10	100
54	4	16
38	-12	144
48	-2	4
350		**432**

Mean $\bar{x} = \frac{\sum x}{n}$

$$= \frac{350}{7}$$
$$= 50\ runs$$

Standard deviation $s = \sqrt{\frac{\sum (x-\bar{x})^2}{n}}$

$$= \sqrt{\frac{432}{7}}$$
$$= \sqrt{61.41}$$
$$= 7.86\ runs$$

Variance $s^2 = (7.86)^2$

$$= 61.78\ runs$$

2. The time taken to slow puzzle by the 5 students are 5, 8, 3, 6, 10. Compute standard deviation of the time taken to solve the puzzles and also find the variance.

X	X^2
5	25
8	64
3	9
6	36

10	100
32	**234**

Mean $\bar{x} = \frac{\sum x}{n}$

$= \frac{32}{5}$

$= 6.4\ minutes$

Standard deviation $s = \sqrt{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2}$

$= \sqrt{\frac{234}{5} - \left(\frac{32}{5}\right)}$

$= \sqrt{46.8 - 40.96}$

$= \sqrt{5.84}$

$= 2.42\ minutes$

Variance $s^2 = (2.42)^2$

$= 5.86\ minutes$

3. The following are closing prices of 5 shares 132, 147 120, 152 and 125. Find standard deviation and variance.

X	$d = x - A$ A = 120	d^2
132	12	144
147	27	729
120	0	0
152	32	1024
125	5	25
676	**76**	**1922**

Standard deviation $s = \sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2}$

$= \sqrt{\frac{1922}{5} - \left(\frac{76}{5}\right)}$

$= \sqrt{384.4 - 231.04}$

$= \sqrt{153.36}$

$= 12.38$

Variance $s^2 = (12.38)^2$

$= 153.26$

4. The monthly pocket expense of 10 students are rupees 225, 190, 220, 245, 270, 180, 250, 300, 175 and 265 find the standard deviation of the expense and also find the variance of the expense. **[Ans: s = 35.45 & s^2 = 1556]**
5. Find the S.D. and variance of the following data.
   60, 45, 25, 40, 70, 32. **[Ans: s = 15.51 & s^2 = 240.55]**
6. A time of the 10 competitors in 400 meters running rays are respectively 70, 75, 90, 95, 72, 80, 88, 85, 75 and 80 seconds. Find the variance and standard deviation of the data. **[Ans: s = 7.86 & s^2 = 61.78]**

GROUPED DATA (DISCRETE DATA)

1. The distribution of the number of absent days of 15 students of a class in the month of January is given below find a standard deviation and variance of their numbers of absent students.

No. of absent days	0	1	2	3	4
No. of students	1	3	7	3	1

No. of absent days x	No. of students f	fx	$x - \bar{x}$	$(x - \bar{x})^2$	$f(x - \bar{x})^2$
0	1	0	-2	4	4
1	3	3	-1	1	3
2	7	14	0	0	0
3	3	9	1	1	3
4	1	4	2	4	4
Total	**15**	**30**	**0**	**10**	**14**

Mean $\bar{x} = \frac{\sum fx}{\sum f}$

$= \frac{30}{15}$

$= 2\ days$

Standard deviation $s = \sqrt{\frac{\sum f(x-\bar{x})^2}{\sum f}}$

$= \sqrt{\frac{14}{15}}$

$$= \sqrt{0.9333}$$
$$= 0.97\ days$$

Variance $s^2 = (0.97)^2$
$$= 0.94\ days$$

2. The following frequency distribution represent the amount of deposit and the number of depositors in the bank find the standard deviation and variance of the deposits.

Deposits	5	10	15	20	25	30	35
No of depositors	2	7	11	15	10	4	1

Observation x	Frequency f	fx	Fx^2
5	2	10	50
10	7	70	700
15	11	165	2475
20	15	300	6000
25	10	250	6250
30	4	120	3600
35	1	35	1225
Total	**50**	**950**	**20300**

Standard deviation $s = \sqrt{\frac{\sum fx^2}{\sum f} - \left(\frac{\sum fx}{\sum f}\right)^2}$
$$= \sqrt{\frac{20300}{50} - \left(\frac{950}{50}\right)^2}$$
$$= \sqrt{406 - 361}$$
$$= \sqrt{45}$$
$$= 6.71$$

Variance $s^2 = (6.71)^2$
$$= 45.02$$

Using other formula

Observation x	Frequency f	$d = x - A$ A = 20	fd	fd^2
5	2	-15	-30	450
10	7	-10	-70	700
15	11	-5	-55	275
20	15	0	0	0
25	10	5	50	250
30	4	10	40	400
35	1	15	15	225
Total	50		-50	2300

Standard deviation $s = \sqrt{\frac{\sum fd^2}{\sum f} - \left(\frac{\sum fd}{\sum f}\right)^2}$

$$= \sqrt{\frac{2300}{50} - \left(\frac{-50}{50}\right)^2}$$

$$= \sqrt{46 - 1}$$

$$= \sqrt{45}$$

$$= 6.71$$

Variance $s^2 = (6.71)^2$

$$= 45.02$$

3. Find the standard deviation.

Observation	62	67	72	77	82	87	92
Frequency	5	7	10	8	5	3	2

[Ans: s = 8.135]

4. The information of the number of mobile phones sold in 35 days in the mobile shop is given below find the standard deviation variance of number of mobile phones sold.

No. of mobile sold	5	6	7	8	9	10
No. of days	2	5	8	12	7	1

[Ans: s = 1.20 & s^2 = 1.45]

5. Find S.D. and variance.

Observation	70	60	50	40	30	20	10
Frequency	7	18	40	45	63	68	75

[Ans: s = 17.39 & s^2 = 302.41]

GROUPED DATA (CONTINUOUS DATA)

1. Find standard deviation from the following frequency distribution of marks obtained by 200 students of a school in an examination.

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70
No. of students	5	12	30	45	50	37	21

Marks	No, of students f	X	$d = \frac{x - A}{c}$ A = 35, c =10	fd	fd^2
0-10	5	5	-3	-15	45
10-20	12	15	-2	-24	48
20-30	30	25	-1	-30	30
30-40	45	35	0	0	0
40-50	50	45	1	50	50
50-60	37	55	2	74	148
60-70	21	65	3	63	189
Total	**200**			**118**	**510**

Standard deviation $s = \sqrt{\frac{\sum fd^2}{\sum f} - \left(\frac{\sum fd}{\sum f}\right)^2} \times C$

$$= \sqrt{\frac{510}{200} - \left(\frac{118}{200}\right)^2} \times 10$$

$$= \sqrt{2.55 - 0.3481} \times 10$$

$$= \sqrt{2.2019} \times 10$$

$$= 14.84 \text{ marks}$$

2. Find the standard deviation of the daily wages from following information of wages workers of a factory.

Daily wages	130-150	150-170	170-190	190-120	210-230
Frequency	7	18	40	45	63

3. Find the standard deviation following frequency distribution.

Class	100-200	200-300	300-400	400-500	500-600	600-700	700-800
Frequency	13	25	38	45	36	30	13

[Ans: s = 162.43]

4. Find the standard deviation of the age of the persons from the following frequency of 125 persons living in the society.

Age	0-10	10-20	20-0	30-40	40-50	50-60	60-70	70-80
No. of persons	15	15	23	22	25	10	5	10

[Ans: s =19.76]

Unit 2) Data Representation and Sampling Technique

❖Graphical representation of data

As more and more data are available to us today, there are several varieties of charts and graphs than before. In reality, the amount of data that we produce, acquire, copy, and use now will be nearly doubled by 2025. Data visualisation is therefore crucial and serves as a powerful tool for organisations. One can benefit from graphs and charts in the following ways:

- Encouraging the group to act proactively.
- Showcasing progress toward the goal to the stakeholders.
- Displaying core values of a company or an organization to the audience.

Moreover, data visualisation can bring heterogeneous teams together around new objectives and foster the trust among the team members. Let us discuss about various. Here we want to discuss three graphs.

1. Histogram
2. Box plots
3. Scatter plots

1. Histogram

A histogram visualises the distribution of data across distinct groups with continuous classes. It is represented with set of rectangular bars with widths equal to the class intervals and areas proportional to frequencies in the respective classes. A histogram may hence be defined as a graphic of a frequency distribution that is grouped and has continuous classes. It provides an estimate of the distribution of values, their extremes, and the presence of any gaps or out-of-the-ordinary numbers. They are useful in providing a basic understanding of the probability distribution.

Constructing a Histogram:

To construct a histogram, the data is grouped into specific class intervals, or “bins” and plotted along the x-axis. These represent the range of the data. Then, the rectangles are constructed with their bases along the intervals for each class. The height of these rectangles is measured along the y-axis representing the frequency for each class interval. It's important to remember that in these representations, every rectangle is next to another because the base spans the spaces between class boundaries.

Use Cases:

When it is necessary to illustrate or compare the distribution of specific numerical data across several ranges of intervals, histograms can be employed. They can aid in visualising the key meanings and patterns associated with a lot of data. They may

help a business or organization in decision-making process. Some of the use cases of histograms include-

- Distribution of salaries in an organisation
- Distribution of height in one batch of students of a class, student performance on an exam,
- Customers by company size, or the frequency of a product problem.

Best Practices

- **Analyse various data groups:** The best data groupings can be found by creating a variety of histograms.
- **Break down compartments using colour:** The same chart can display a second set of categories by colouring the bars that represent each category.

Example

The following daily wages and workers create histogram chart using data.

Daily wages	36-38	38-40	40-42	42-44	44-46	46-48	48-50	50-52	52-54	54-56
workers	2	4	10	15	19	19	15	10	4	2

2. Box plots

When displaying data distributions using the five essential summary statistics of minimum, first quartile, median, third quartile, and maximum, box-and-whisker plots, also known as boxplots, are widely employed. It is a visual depiction of data that aids in determining how widely distributed or how much the data values change. These boxplots make it simple to compare the distributions since it makes the centre, spread, and overall range understandable. They are utilised for data analysis wherein the graphical representations are used to determine the following:

1. Shape of Distribution
2. Central Value
3. Variability of Data

Constructing a Boxplot:

The two components of the graphic are described by their names: the box, which shows the median value of data along with the first and third quartiles (25 percentile and 75 percentile), and the whiskers, which shows the remaining data. The 3rd quartile's difference from the first quartile of data is called the interquartile range. The highest and minimum points in the data can also be displayed using the whiskers. The points beyond 1.5 ´ interquartile range can be identified as suspected outliers.

Use Cases:

A boxplot is frequently used to demonstrate whether a distribution is skewed and whether the data set contains any potential outliers, or odd observations. Boxplots are also very useful for comparing or involving big data sets. Examples of box plots include plotting the:

- Gas efficiency of vehicles
- Time spent reading across readers

Best Practices

- **Cover the points within the box:** This aids the viewer in concentrating on the outliers.
- **Box plot comparisons between categorical dimensions:** Box plots are excellent for quickly comparing dataset distributions.

3. Scatter plots

Scatter plot is the most commonly used chart when observing the relationship between two quantitative variables. It works particularly well for quickly identifying possible correlations between different data points. The relationship between multiple variables can be efficiently studied using scatter plots, which show whether one variable is a good predictor of another or whether they normally fluctuate independently. Multiple distinct data points are shown on a single graph in a scatter plot. Following that, the chart can be enhanced with analytics like trend lines or cluster analysis. It is especially useful for quickly identifying potential correlations between data points.

Constructing a Scatter Plot:

Scatter plots are mathematical diagrams or plots that rely on Cartesian coordinates. In this type of graph, the categories being compared are represented by the circles on the graph (shown by the colour of the circles) and the numerical volume of the data (indicated by the circle size). One colour on the graph allows you to represent two values for two variables related to a data set, but two colours can also be used to include a third variable.

Use Cases:

Scatter charts are great in scenarios where you want to display both distribution and the relationship between two variables.

- Display the relationship between time-on-platform (How Much Time Do People Spend on social media) and churn (the number of people who stopped being customers during a set period of time).

• Display the relationship between salary and years spent at company.

Best Practices

- **Analyse clusters to find segments:** Based on your chosen variables, cluster analysis divides up the data points into discrete parts.
- **Employ highlight actions:** You can rapidly identify which points in your scatter plots share characteristics by adding a highlight action, all the while keeping an eye on the rest of the dataset.
- **mark customization:** individual markings Add a simple visual hint to your graph that makes it easy to distinguish between various point groups.

❖ Probability theory

Concept of probability

There are a number of events in day-to-day life about which one is not sure whether it will occur or not. But one is always curious to know what chance is there for happening or event to occur. For instance, one may be interested to estimate whether it will rain today or not, one would like to evaluate his chance of winning for head is a definite number of tosses, what is the chance that there are four aces in one hand in a game of card among four players, etc.

The numerical evaluation of chance factor of an event is known as a probability.

The probability is to numerically express the possibility of this uncertain event.

Define the random experiment

The exponent which can be independently repeated under the identical condition and all its possible outcomes are known but which of the outcomes will appear cannot be predicting with certainty before conducting the experiment is called a random experiment.

An experiment whose outcomes are known before the experiment is done but certain outcome cannot be predicted before the experiment is happens is known as a random experiment.

Define the random variable

Random variable is real valued functions consider each and every value of sample space. A variable whose values can be obtained from the results of a random experiment is called a random variable. A random variable is a function associated with a sample space of a random experiment, and it takes different values with different probabilities. A random variable can either be discrete or continuous.

Define sample space

The set of all possible outcome of a random experiment is called the sample space.

- **Finite Sample Space: -**If the total number of a possible outcomes in the sample space is finite then it is called a finite sample space.
- **Infinite Sample Space: -**If the total number of possible outcomes in the sample space is infinite then it is called an infinite sample space.

Define the Events

A subset of the sample space of random experiment is called Event.

The collection of possible outcomes which are favourable to a happening out of total outcomes is called an Event.

- **Impossible Event: -**An event which is certain not to occur is called an impossible event. It denoted by ϕ or { }.
- **Certain Event: -** A special subset of the sample space of random experiment is called certain event.
- **Complementary Event:** - The complement of any event A is the event [not A], i.e., the event that A does not occur. It is denoted by A'.
- **Interception of Event: -** Suppose A and B are two events of finite sample space where event A and Occur simultaneously is called the intersection event. It is denoted by $A \cap B$.
- **Union of Events: -** Suppose event A and B are any two events of the finite sample space where the event A occurs or the B occurs or both the event A and B occurs is called the union event. It is denoted by $A \cup B$.
- **Mutually Exclusive Event: -** Suppose A and B are any two events of finite sample space where event A and B do not occur together means $A \cap B = \emptyset$ or in other word event B does not occur when event A occur and event A does not

occur when event B occurs then the event A and B are called mutually exclusive event.

- **Exhaustive Event: -** If the group of the favourable outcomes of event of random experiment in the sample space, then the events are called the executive event. Suppose A and B are any two events of sample space where the event A and B are called exhaustive events if the union $A \cup B$ of the two event A and B is a sample space. That is $A \cup B = U$.
- **Mutually Exclusive and Executive Events: -** Suppose A and B are two event of a finite sample space these two events A and B are called the mutually exclusive and exhaustive events. That is $A \cap B = \emptyset$ and $A \cup B = U$.
- **Elementary Event: -** The event formed by all the subset of single element of the sample space of a random experiment are called elementary event. The elementary events are mutually exclusive and executive event.

Property of probability of an element A

a) Probability is a pure number. i.e., it has no unit.
b) It is a relative measure.
c) $0 \leq P(A) \leq 1$, i.e., probability can never be negative and cannot exceed unit(one).
d) If A1, A2, …., Ak are k mutually exclusive and exhaustive events in the sample than $\sum_{i=1}^{k} P(A_i) = 1 \rightarrow P(U) = 1$.
e) P(ϕ) = 0, i.e., Probability of an impossible event is 0.

Mathematical definition of probability:

Suppose there are the total an outcome in finite sample space of random experiment which are mutually exclusive in exhaustive and equiprobable. if m outcome among there is favourable for an event A then the probability of the event A is $\frac{m}{n}$. The probability of event A denoted by P(A).

P(A) = probability of event A

$$= \frac{\text{Favourable outcomes of event A}}{\text{Toatl number of mutually exclusive, exhaustive and eui}-\text{probable outcomes of sample specae}}$$

$$= \frac{m}{n}$$

Other formula

$$P(A') = 1 - P(A)$$

$$P(A' \cup B') = P(A \cap B)' = 1 - P(A \cup B)$$

$$P(A' \cap B') = P(A \cup B)' = 1 - P(A \cup B)$$
$$P(A - B) = P(A \cap B') = P(A) - P(A \cap B)$$
$$P(B - A) = P(A' \cap B) = P(B) - P(A \cap B)$$
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Example

1. **If 3 coins are tossed at a time find the probability of following event.**
   - **A. Getting at least one head**
   - **B. Getting at most one tail**
   - **C. Getting more than one tail**
   - **D. Getting head and tail alternately**
   - **E. Getting more number of tail than head**

Ans: - The sample space of random experiment of tossing coin 3 time

Total outcomes (n) = $(2)^3$

= 8

U = {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}

A. Getting at least one head

Favourable outcomes (m) =7

A = {HHH, HHT, HTH, HTT, THH, THT, TTH}

$$P(A) = \frac{m}{n}$$
$$= \frac{7}{8}$$

B. Getting at most one tail

Favourable outcomes (m) =4

B = {HHH, HHT, HTH, THH}

$$P(B) = \frac{m}{n}$$
$$= \frac{4}{8}$$
$$= \frac{1}{2}$$

C. Getting more than one tail

Favourable outcomes (m) =4

C = {HTT, THT, TTH, TTT}

$$P(C) = \frac{m}{n}$$
$$= \frac{4}{8}$$

$$= \frac{1}{2}$$

D. Getting head and tail alternately

Favourable outcomes (m) =2

D = {THT, HTH}

$$P(D) = \frac{m}{n}$$

$$= \frac{1}{2}$$

E. Getting more number of tail than head

Favourable outcomes (m) =4

E = {HTT, THT, TTH, TTT}

$$P(E) = \frac{m}{n}$$

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

$$= \frac{1}{2}$$

2. **Two balance dice are thrown simultaneously. Find the probability of the following events:**
   - **A. The sum of number on the dice is 8**
   - **B. The sum of number on the dice is more than 9**
   - **C. The sum of numbers on the dice multiple of 3**
   - **D. The product of number on the dice is 12.**

Ans: - The sample space of random experiment of two balance dice are thrown simultaneously

Total outcomes (n) = $(6)^3$

= 36

U = {(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6),
(2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6),
(3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6),
(4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6),
(5, 1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6),
(6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6)}

A. The sum of number on the dice is 8

Favourable outcomes (m) = 5

A = {(2, 6), (3, 5), (4, 4), (5, 3), (6, 2)}

$$P(A) = \frac{m}{n}$$
$$= \frac{5}{36}$$

B. The sum of number on the dice is more than 9

Favourable outcomes (m) = 6

B = {(4, 6), (5, 5), (5, 6), (6, 4), (6, 5), (6, 6)}

$$P(B) = \frac{m}{n}$$
$$= \frac{6}{36}$$
$$= \frac{1}{6}$$

C. The sum of number on the dice is multiple of 3

Favourable outcomes (m) = 12

C = {(1, 2), (2, 1), (1, 5), (2, 4), (3, 3), (4, 2), (5, 1), (3, 6), (4, 5), (5, 4), (6, 3), (6, 6)}

$$P(C) = \frac{m}{n}$$
$$= \frac{12}{36}$$
$$= \frac{1}{3}$$

D. The product of the number o the dice is 12

Favourable outcomes (m) = 4

D = {(3, 4), (4, 3), (6, 2), (2, 6)}

$$P(D) = \frac{m}{n}$$
$$= \frac{4}{36}$$
$$= \frac{1}{9}$$

3. **The marks of the ten students in the subject of statistics are given below 50, 64, 65, 35, 55, 52, 63, and 30 if one is selected at randomly then what is the probability that his marks would be greater than the average marks.**

Ans: Total no. of Students (n) = 10

Probability that one is selected at randomly and his marks would be greater than the average marks.

$$\bar{x} = \frac{\sum x}{n} = \frac{495}{10} = 49.5$$

Favourable outcomes (m) = 6

A = {50, 60, 65, 55, 52, 63}

$$P(A) = \frac{m}{n}$$
$$= \frac{6}{10}$$
$$= 0.6$$

4. **Find the probability of getting vowels in the first, third and sixth place when all the letters of the word ORANGE are arranged in the all-possible ways.**

Ans: - The total number of ways of arranging these these six letters.

Total outcomes (n) = 6!

= 720

A = the probability of getting vowels in the first, third and sixth place when all the letters of the word ORANGE.

O			A			E

Favourable outcomes (m) = $3! \times 3!$

= 36

$$P(A) = \frac{m}{n}$$
$$= \frac{36}{720}$$
$$= \frac{1}{20}$$

5. **Find the probability of 5 Mondays in the months of February of a leap year**

Ans: - In a leap year total no. of days in the months of February = 29

- 7 days × 4 weeks in every month = 28

Extra day 1

In a week everyday come only once. Hence in 4 weeks every day comes 4 times the sample space for extra day is expressed as follows

U = {Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday}

Total outcomes (n) = 7

A = The probability of 5 Mondays in the months of February of a leap year

Favourable outcomes (m) = 1

$$P(A) = \frac{m}{n}$$
$$= \frac{1}{7}$$

6. **Find the probability of having 53 Thursday in day a leap year**

Ans: - Total no. of days in a leap year = 366

- 7 days × 52 weeks in a year = 364

Extra day = 2

The additional 2 days can be as follows which given the sample space for this experiment.

U = {Sunday-Monday, Monday-Tuesday, Tuesday-Wednesday, Wednesday-Thursday, Thursday-Friday, Friday-Saturday, Saturday-Sunday}

Total outcomes (n) = 7

A = The probability of 5 Mondays in the months of February of a leap year

Favourable outcomes (m) = 2
{Tuesday-Wednesday, Wednesday-Thursday}

$$P(A) = \frac{m}{n}$$
$$= \frac{2}{7}$$

7. **One card is randomly selected from the packet of 52 cards find a probability that the selected card is**
   i) Diamond or king card
   ii) neither a diamond nor a king card.

Ans: one card is randomly selected from the packet of 52 cards

Total outcomes (n) = 52

i) Selected card is diamond or king card

$$P(A \cup B) = \text{Selected card is diamond or king card}$$

Event A: Selected card is diamond

Favourable outcomes (m) = 13

$$P(A) = \frac{m}{n}$$
$$= \frac{13}{52}$$

Event B: Selected card is king

Favourable outcomes (m) = 4

$$P(B) = \frac{m}{n}$$
$$= \frac{4}{52}$$

Event A∩B: Selected card is diamond and king

Favourable outcomes (m) = 1

$$P(A \cap B) = \frac{m}{n}$$
$$= \frac{1}{52}$$

$$P(A \cup B) = \text{Selected card is diamond or king card}$$
$$= P(A) + P(B) - P(A \cap B)$$
$$= \frac{13}{52} + \frac{4}{52} - \frac{1}{52}$$
$$= \frac{4}{13}$$

ii) Selected card neither a diamond nor a king card

A' = Event that the selected card is not diamond

B' = Event that the selected card is not king

A'∩B = Event that the selected card neither a diamond nor a king card

$$P(A' \cap B') = P(A \cup B)'$$
$$= 1 - P(A \cup B)$$
$$= 1 - \frac{4}{13}$$
$$= \frac{9}{13}$$

8. **The probability that a Pearson from a group reads English newspaper is 0.55, the probability that he reads Hindi newspaper is 0.69 and the probability that he reads both the newspaper English and Hindi is 0.27. find the probability that a person**

selected at random from this group

i) read at least one of the English and Hindi newspaper.

ii) does not read any of newspaper English and Hindi.

Ans: Event A = Probability that a Pearson from a group reads English newspaper

Event B = Probability that a Pearson from a group reads Hindi newspaper

Event A ∩ B = The probability that he reads both the newspaper English and Hindi

P(A) = 0.55, P(B) = 0.69, P(A ∩ B) = 0.27

i) Read at least one of the English and Hindi newspaper

P(A ∪ B) = The probability that a person selected at random from this group read at least one of the English and Hindi newspaper

$$\begin{aligned} \mathrm{P(A \cup B)} &= \mathrm{P(A) + P(B) - P(A \cap B)} \\ &= 0.55 + 0.69 - 0.27 \\ &= 0.97 \end{aligned}$$

ii) Does not read any of newspaper English and Hindi

Event A’ = Probability that a Pearson from a group not reads English newspaper

Event B’ = Probability that a Pearson from a group not reads Hindi newspaper

Event A’ ∩ B’ = The probability that he not reads both the newspaper English and Hindi

$$\begin{aligned} \mathrm{P(A' \cap B')} &= \mathrm{P(A \cup B)'} \\ &= 1 - \mathrm{P(A \cup B)} \\ &= 1 - 0.97 \\ &= 0.03 \end{aligned}$$

9. **For the two event A and B is the sample space of a random experiment P(A') = 0.3, P(B)= 0.6 and P($\mathbf{A} \cup \mathbf{B}$) = 0.83. Find P($\mathbf{A} \cap \mathbf{B}'$) and P(A'$\cap \mathbf{B}$).**

Ans.: Here, P(A') = 0.3 P(A) = 1-P(A') = 1- 0.3 = 0.7

P(B) = 0.6 and P($A \cup \mathrm{B}$) = 0.83

First, we will find P(A ∩ B').

P(A ∪ B) = P(A) + P(B) - P(A ∩ B)

0.83 = 0.7 + 0.6 - P(A ∩ B)

P(A ∩ B) = 0.7 + 0.6 – 0.83

P(A ∩ B) = 0.47

Now,

P(A ∩ B') = P(A) - P(A ∩ B)

= 0.7 – 0.47

= 0.23

Required probability = 0.23

P(A'∩ B) = P(B) - P(A ∩ B)

= 0.6 – 0.47

= 0.13

Required probability = 0.13

10. **For two events A and B in the sample space of a random experiment P(A) = 2P(B) = 4P(A ∩ B) = 0.6. Find the probability of the following events: (1) $A' \cap B'$ (2) $A' \cup B'$ (3) A – B (4) B – A**

It is given that P(A) = 2P(B) = 4P(A ∩ B) = 0.6. Hence,

P(A) = 0.6 2P(B) = 0.6 4P(A ∩ B) = 0.6

∴ P(B) = 0.3 ∴ P(A ∩ B) = 0.15

(1) Probability of events A' ∩ B' = P(A' ∩ B')

= P(A ∪ B)'

= 1 – P(A ∪ B)

= 1 - [P(A) + P(B) - P(A ∩ B)]

= 1 - [0.6 +0.3 - 0.15]

= 1- 0.75

= 0.25

Required probability = 0.25

(2) Probability of events A' ∪ B' = P(A' ∪ B')

= P(A∩ B)'

= 1 - P(A ∩ B)

= 1 - 0.15

= 0.85

Required probability = 0.85

(3) Probability of events A – B = P(A-B)

= P(A) – P(A ∩ B)

= 0.6 – 0.15

= 0.45

Required probability = 0.45

(4) Probability of B – A = P(B-A)

= P(B) - P(A ∩ B)

= 0.3-0.15

= 0.15

Required probability = 0.15

11. **Two events A and B in the sample space of a random experiment mutually exclusive. If 3P(A) = 4P(B) = 1 then find P($A \cup B$).**

Ans: Since 3P(A) = 4P(B) = 1

3P(A) = 1 4P(B) = 1

$\therefore$ P(A) = $\frac{1}{3}$ $\therefore$ P(B) = $\frac{1}{4}$

As the events A and B are mutually exclusive ($A \cap B = \emptyset$),

P($A \cup B$) = P (A) + P(B)

$= \frac{1}{3} + \frac{1}{4}$

$= \frac{7}{12}$

Required probability = $\frac{7}{12}$

12. **For three mutually exclusive and exhaustive events A, B and C in the sample space of a random experiment 2P(A) = 3P(B) = 4P(C). Find P(A $\cup$ B) and P(B $\cup$ C).**

Ans: Taking 2P(A) = 3P(B) = 4P(C) = x,

2P(A) = x 3P(B) = x 4P(C) = x

$\therefore$P(A) = $\frac{x}{2}$ $\therefore$P(B) = $\frac{x}{3}$ $\therefore$P(C) = $\frac{x}{4}$

Since A , B and C are mutually exclusive and exhaustive events,

$$P(A \cup B \cup C) = P(A) + P(B) + P(C)$$

$\therefore \frac{x}{2} + \frac{x}{3} + \frac{x}{4} = 1$

$\therefore \frac{6x+4x+3x}{12} = 1$

$\therefore 13x = 12$

$\therefore x = \frac{12}{13}$

Thus,

$P(A) = \frac{x}{2} = \frac{\frac{12}{13}}{2} = \frac{6}{13}$

$P(B) = \frac{x}{3} = \frac{\frac{12}{13}}{3} = \frac{4}{13}$

$P(C) = \frac{x}{4} = \frac{\frac{12}{13}}{4} = \frac{3}{13}$

Now, the probability of required events,

$$P(A \cup B) = P(A) + P(B)$$

$$= \frac{6}{13} + \frac{4}{13}$$

$$= \frac{10}{13}$$

Required probability = $\frac{10}{13}$

$$P(B \cup C) = P(B) + P(C)$$

$$= \frac{4}{13} + \frac{3}{13}$$

$$= \frac{7}{13}$$

Required probability = $\frac{7}{13}$

QUESTIONS

1. If 3 coins are tossed at a time find the probability of following event.
   A. Getting at least one tails
   B. Not getting a single head
   C. Getting at most two head
   D. Getting more than one head
   E. Getting more number of head than tail
   F. Getting at most three tails

[Ans: 1/8, 1/8, 7/8, 1/2, 1/2, 1]

2. If two balanced coins are tossed, then find the probability of
   A. Getting at most one tail
   B. Getting one head and one tail

[Ans: 3/4, 1/2]

3. Two balanced dice marked with number 1 to 6 are thrown simultaneously. Find the probability that
    A. The sum of number on the dice is 5
    B. The sum of number on the dice is more than 10
    C. The sum of numbers on the dice multiple of 4
    D. Both the dice show same numbers
    E. Sum of numbers on the dice is 1
    F. Sum of the number on the dice is 12 or more
    G. Sum of the number on the dice is 12 or less

[Ans: 1/8, 1/12, 1/4, 1/6, 0/36, 0/36, 1]

4. One family randomly selected from the families having two children find the probability that
    A. One child is boy and one child is girl
    B. At least one child is boy among the two children of the selected family

[Ans: 1/2, 3/4]

5. The sample space for the random experiment of selecting numbers is U {1, 2, 3,……,150} and all the outcomes in the sample space are equiprobable. Find the probability that the number selected is:
    A. a multiple of 2
    B. not a multiple of 2
    C. a multiple of 5
    D. not a multiple of 5
    E. a multiple of both 2 and 5

[Ans: 1/2, 1/2, 1/5, 4/5, 1/10]

6. Find the probability of getting U in the first place and E in the last place when all of the word **UNIQUE** are arranged in the possible ways.

[Ans: 1/30]

7. The runs for by 10 players are follows 25, 32, 40, 20, 36, 45, 50, 30, 25, 37 one player is selected at randomly the find the probability that
    A. Whose run more than the average runs
    B. Whose runs are less than the average runs

[Ans: 0.5, 0.5]

8. The run of 11 cricketers in a day match argue 90, 50, 30, 100, 5, 40, 0, 25, 70, 10, 20 if one cricketer is selected at randomly the what is the probability that he's run could be less than the average runs.

[Ans: 6/11]

9. Find the Probability having 53 Fridays in a year which is not a leap year. **[Ans: 1/7]**
10. Find the probability of having 5 Tuesdays in the month of august of any year. **[Ans: 3/7]**

11. Two cards are drowned from a pack of 52 cards what is the probability that
    A. both cards are king
    B. both cards are king or both queens
    C. both cards are king and queens.

[Ans: 1/221, 2/221, 8/663]

12. Two cards are drowned from a pack of 52 cards what is the probability that
    A. Both cards are same suit
    B. Both cards are same colour

[Ans: 4/17, 25/51]

13. One card is randomly selected from the packet of 52 cards find a probability that the selected card is
    i) Heart or jack card
    ii) neither a Heart nor a jack card.

[Ans: 4/13, 9/13]

13. Two aircrafts dropped bomb to destroy a bridge the probability that a bomb dropped from the first aircraft hits the target is 0.9 and the probability that a bomb from the second aircraft hits the target is 0.7. the probability of one one bomb drops from both the aircraft hitting the target is 0.63 the bridge is destroyed even if one bomb drops on it. Find the probability that the bridge is destroyed.

[Ans: 0.97]

14. For three mutually exclusive and exhaustive events A, B and C in the sample space of a random experiment 4P(A) = 5P(B) = 3P(C). Find P(A ∪ C) and P(B ∪ C).

[Ans: 35/47, 32/47]

15. For two events A and B in the sample space of a random experiment P(A') = 2P(B') = 3P(A ∩ B) = 0.6. Find the probability of the following events:
    i. A – B
    ii. B – A

[Ans: 0.2, 0.5]

16. **F**or two events A and B in the sample space of a random experiment P(A) = 0.6 P(B) = 0.5 and P(A ∩ B) = 0.15. Find the probability of the following events: (1)
    i. P(A')
    ii. P(B – A)
    iii. $P(A \cap B')$
    iv. $P(A' \cap B')$
    v. $P(A' \cup B')$

[Ans: 0.4, 0.35, 0.45, 0.05, 0.85]

❖ Probability distribution

A probability distribution is the mathematical function that gives the probabilities of occurrence of different possible outcomes for an experiment.

Type of Probability Distribution

There are two types of Probability Distribution

1. Discrete probability distribution
2. Continuous probability distribution

1. Discrete probability distribution

If a random variable can take finite values or countable infinite values it is known as a discrete variable. The number of heads obtained in tossing of two coins is a discrete variable. Similarly, the number of children in a family, number of accidents are examples of discrete variables. If the number of defective screws manufactured by a company is denoted by x then the values of x can be 0, 1, 2, 3, 4, 5, 6..... etc. The values of discrete variable in this case are countable infinite values.

Here we study Binomial Distribution which discrete probability distribution.

Binomial Distribution

In random experiment which done in single trial and outcome are of the trial either success or failure then such trial is known as Bernoulli trials.

Considering n independent Bernoulli trials each trial has two outcome is considered as success with the the probability p and failure with probability q that is p – q = 1, among this trial here the probability of x success follows binomial distribution. Since x denoted the number of successes in n independent trials

Therefore x is random variable having value X = 0, 1, 2, …. , n let probability of x success and n - x failure having the order of probability of success and failure respectively then it is denoted with $p^x q^{n-x}$

The possible ways of x success among these trails are C_x^n

Therefore, the x[th] in n independent trial

$$P(x) = C_x^n . p^x . q^{n-x} \quad ; x = 0, 1, 2, \ldots . , n$$

- **Examples of Bernoulli's Trails are**
  - Tossing a coin (head or tail)
  - Throwing a die (even or odd number)
  - Student's performance an exam (Pass or fail)

➢ Conditions for binomial distribution

The probability of success is denoted by ‘p’ and that of failure by ‘q’, such that p + q = 1. The distribution can be obtained under the following experimental conditions:

- The number of trials ‘n’ is finite.
- The trials are independent of each other.
- Success probability ‘p’ is constant for each trial.
- Each trial has only one of the two possible results either success or failure.

The best examples of binomial distribution are, tossing of coins or throwing of dice or drawing cards from a pack of cards.

➢ Characteristics of Binomial Distribution

1) This is a distribution of a discrete variable.
2) n, p are the parameters of Binomial distribution.
3) The mean of Binomial distribution is np which shows the average number of successes.
4) Variance of binomial distribution is npq
5) Standard deviation of binomial distribution is $\sqrt{npq}$
6) When p and q are equal i.e., p = q = 1/2 Binomial distribution is a symmetrical distribution.
7) When p< 1/2, its skewness is positive and when p> 1/2 its skewness is negative.
8) The variance of Binomial distribution is always less than its mean.
9) When number of trials n is very large, and p and q are not very small, Binomial distribution tends to Normal distribution.
10) When number of trials n is very large and p or q is very small, Binomial distribution tends to Poisson distribution.

➢ Example

1. **The probability of winning a one-day cricket match of India against Australia is 1/3. India and Australia are going to play three one day cricket matches. Find the probability that**
   - **A. India will lose all the three one day matches**
   - **B. India will win at least one one-day matches.**

Ans: Probability of winning a one-day cricket match of India against Australia $= p = \frac{1}{3}$

India and Australia are going to play three one day cricket matches = n = 3

here $q = 1 - p = 1 - \frac{1}{3} = \frac{2}{3}$

A. India will lose all the three one day matches

x = 0

$$P(x) = C_x^n . p^x . q^{n-x}$$

$$P(0) = C_0^3 . \left(\frac{1}{3}\right)^0 . \left(\frac{2}{3}\right)^{3-0}$$

$$= 1 \times 1 \times \frac{8}{27}$$

$$= \frac{8}{27}$$

B. India will win at least one one-day matches

x = 1, 2, 3

P(x ≥ 3) = P(1) + P(2) + P(3)

$$= C_1^3 . \left(\frac{1}{3}\right)^1 . \left(\frac{2}{3}\right)^{3-1} + C_2^3 . \left(\frac{1}{3}\right)^2 . \left(\frac{2}{3}\right)^{3-2} + C_3^3 . \left(\frac{1}{3}\right)^3 . \left(\frac{2}{3}\right)^{3-3}$$

$$= \left(3 \times \frac{1}{3} \times \frac{4}{9}\right) + \left(3 \times \frac{1}{9} \times \frac{2}{9}\right) + \left(1 \times \frac{1}{27} \times 1\right)$$

$$= \frac{12}{27} + \frac{6}{27} + \frac{1}{27}$$

$$= \frac{19}{27}$$

OR short cut method

P(x ≥ 3) = 1 + P(0)

$$= 1 - \left(C_0^3 . \left(\frac{1}{3}\right)^0 . \left(\frac{2}{3}\right)^{3-0}\right)$$

$$= 1 - \left(1 \times 1 \times \frac{8}{27}\right)$$

$$= 1 - \left(\frac{8}{27}\right)$$
$$= \frac{19}{27}$$

2. **A chance of winning the match of India against Pakistan is 5:2 if a series of three matches is to be played than what is the probability of the India to be win at most one match.**

Ans: A chance of winning the match of India against Pakistan is 5:2

$p = \frac{5}{7}, \qquad q = 1 - p = 1 - \frac{5}{7} = \frac{2}{7}, \qquad n = 3$

Probability of the India to be win at most one match

x = 0, 1

P(x ≤ 1) = P(0) + P(1)

$$= C_0^3 . \left(\frac{5}{7}\right)^0 . \left(\frac{2}{7}\right)^{3-0} + C_1^3 . \left(\frac{5}{7}\right)^1 . \left(\frac{2}{7}\right)^{3-1}$$
$$= \left(1 \times 1 \times \frac{8}{343}\right) + \left(3 \times \frac{5}{7} \times \frac{4}{49}\right)$$
$$= \frac{8}{343} + \frac{60}{343}$$
$$= \frac{68}{343}$$

3. **30% of attacking aircraft are expected to be shoot down before reaching the targets what is the probability that one out of five attacking aircraft will reach the target.**

Ans: Attacking aircraft are expected to be shoot down before reaching the targets = 30%

q = 0.3, p = 1 – q = 1 – 0.3 = 0.7, n = 5

Probability that one out of five attacking aircraft will reach the target

x = 1

$$P(x) = C_x^n . p^x . q^{n-x}$$
$$P(1) = C_1^5 . (0.7)^1 . (0.3)^{5-1}$$
$$= 5 \times (0.7) \times (0.0081)$$
$$= 0.02835$$

4. **The probability that a bomb dropped from the plane will hit target is 2/5. Two bombs are enough to destroy a bridge. If 4 bombs are dropped on the bridge, find the probability that**

A. **The bridge will be destroyed**
B. **The bridge will be partially destroyed**
C. **The bridge will be saved.**

Ans: The probability that a bomb dropped from the plane will hit target = 2/5

$$p = \frac{2}{5}, \qquad q = 1 - p = 1 - \frac{2}{5} = \frac{3}{5}, \qquad n = 4$$

A. The bridge will be destroyed
x = 2, 3, 4

P(2 ≤ x ≤ 4) = P(2) + P(3) + P(4)

$$= C_2^4.\left(\frac{2}{5}\right)^2.\left(\frac{3}{5}\right)^{4-2} + C_3^4.\left(\frac{2}{5}\right)^3.\left(\frac{3}{5}\right)^{4-3} + C_4^4.\left(\frac{2}{5}\right)^4.\left(\frac{3}{5}\right)^{4-4}$$

$$= \left(6 \times \frac{4}{25} \times \frac{9}{25}\right) + \left(4 \times \frac{8}{125} \times \frac{3}{5}\right) + \left(1 \times \frac{16}{625} \times 1\right)$$

$$= \frac{216}{625} + \frac{96}{625} + \frac{16}{625}$$

$$= \frac{328}{625}$$

B. The bridge will be partially destroyed
x = 1

$$P(x) = C_x^n.p^x.q^{n-x}$$

$$P(1) = C_1^4.\left(\frac{2}{5}\right)^1.\left(\frac{3}{5}\right)^{4-1}$$

$$= 4 \times \frac{2}{5} \times \frac{27}{125}$$

$$= \frac{216}{625}$$

C. The bridge will be saved
x = 0

$$P(x) = C_x^n.p^x.q^{n-x}$$

$$P(0) = C_0^4.\left(\frac{2}{5}\right)^0.\left(\frac{3}{5}\right)^{4-0}$$

$$= 1 \times 1 \times \frac{81}{625}$$

$$= \frac{81}{625}$$

5. **There are 50 fishes in a pool, out of which 10 are golden, 5 fishes are taken at random from the pool. What is the probability that 2 fishes are golden out of 5?**

Ans: there are 50 fished in a pool N = 50
out of which 10 are golden X = 10
5 fishes are taken random from the pool n = 5
Probability that 2 fishes are golden x = 2

$p = \frac{2}{5}, \qquad q = 1 - p = 1 - \frac{2}{5} = \frac{3}{5}, \qquad n = 5$

$$P(x) = C_x^n . p^x . q^{n-x}$$
$$P(2) = C_2^5 . \left(\frac{2}{5}\right)^2 . \left(\frac{3}{5}\right)^{5-2}$$
$$= 10 \times \frac{4}{25} \times \frac{9}{25}$$
$$= \frac{360}{625}$$
$$= \frac{72}{25}$$

6. **In a competitive test there are 6 objective questions and three alternatives are given for each question. Only one of them is correct. A candidate does not know the correct answers of any of the questions and hence in each question he ticks any one of the alternatives randomly. Find the probabilities of getting**
   A. **All correct answers**
   B. **At least 4 correct answers**

Ans: There three alternatives are given for each question only one of them is correct = 1/3

$p = \frac{1}{3}, \qquad q = 1 - p = 1 - \frac{1}{3} = \frac{2}{3}, \qquad n = 6$

A. All correct answers
x = 6

$$P(x) = C_x^n . p^x . q^{n-x}$$
$$P(2) = C_6^6 . \left(\frac{1}{3}\right)^6 . \left(\frac{2}{3}\right)^{6-6}$$
$$= 1 \times \frac{1}{729} \times 1$$
$$= \frac{1}{729}$$

B. At least 4 correct answers

x = 4, 5, 6

P(x ≥ 4) = P(4) + P(5) + P(6)

$$= C_4^6.\left(\frac{1}{3}\right)^4.\left(\frac{2}{3}\right)^{6-4} + C_5^6.\left(\frac{1}{3}\right)^5.\left(\frac{2}{3}\right)^{6-5} + C_6^6.\left(\frac{1}{3}\right)^6.\left(\frac{2}{3}\right)^{6-6}$$

$$= \left(15 \times \frac{1}{81} \times \frac{4}{9}\right) + \left(6 \times \frac{1}{243} \times \frac{2}{3}\right) + \left(1 \times \frac{1}{729} \times 1\right)$$

$$= \frac{60}{729} + \frac{12}{729} + \frac{1}{729}$$

$$= \frac{73}{729}$$

7. An accountant is to audit 24 accounts of a firm. 16 of these are high valued customers. If the accountant selects six accounts at random, what is probability of selection of:

A. At least three high valued customers?

B. At the most three high valued customers?

Ans: An accountant is to audit 24 accounts of a firm N = 24

16 of these are high valued customers X = 16

If the accountant selects six accounts at random n = 6

$$p = \frac{16}{24} = \frac{2}{3}, \qquad q = 1 - p = 1 - \frac{2}{3} = \frac{1}{3}, \qquad n = 6$$

A. At least three high valued customers

x = 3, 4, 5, 6 **OR** x = 1 – [0, 1, 2]

$$P(x \geq 3) = 1 - [P(0) + P(1) + P(2)]$$

$$= 1 - \left[C_0^6.\left(\frac{2}{3}\right)^0.\left(\frac{1}{3}\right)^{6-0} + C_1^6.\left(\frac{2}{3}\right)^1.\left(\frac{1}{3}\right)^{6-1} + C_2^6.\left(\frac{2}{3}\right)^2.\left(\frac{1}{3}\right)^{6-2}\right]$$

$$= 1 - \left[\left(1 \times 1 \times \frac{1}{729}\right) + \left(6 \times \frac{2}{3} \times \frac{1}{243}\right) + \left(15 \times \frac{4}{9} \times \frac{1}{81}\right)\right]$$

$$= 1 - \left[\frac{1}{729} + \frac{12}{729} + \frac{60}{729}\right]$$

$$= 1 - \left[\frac{73}{729}\right]$$

$$= \frac{656}{729}$$

B. At the most three high valued customers

x = 0, 1, 2, 3

$$P(x \geq 3) = P(0) + P(1) + P(2) + P(3)$$

$$= C_0^6.\left(\frac{2}{3}\right)^0.\left(\frac{1}{3}\right)^{6-0} + C_1^6.\left(\frac{2}{3}\right)^1.\left(\frac{1}{3}\right)^{6-1} + C_2^6.\left(\frac{2}{3}\right)^2.\left(\frac{1}{3}\right)^{6-2} + C_3^6.\left(\frac{2}{3}\right)^3.\left(\frac{1}{3}\right)^{6-3}$$

$$= \left(1 \times 1 \times \frac{1}{729}\right) + \left(6 \times \frac{2}{3} \times \frac{1}{243}\right) + \left(15 \times \frac{4}{9} \times \frac{1}{81}\right) + \left(20 \times \frac{8}{27} \times \frac{1}{27}\right)$$

$$= \frac{1}{729} + \frac{12}{729} + \frac{60}{729} + \frac{160}{729}$$

$$= \frac{233}{729}$$

8. Five coins are tossed simultaneously for 3200 times. Find the frequencies of different number of heads and tails and tabulate the results.

Ans: Here probability of getting head $p = \frac{1}{2}$

$\therefore q = 1 - p = \frac{1}{2}$ N = 3200, n = 5

Now for Binomial distribution

$$P(x) = C_x^n.p^x.q^{n-x}$$
$$= C_x^5.\left(\frac{1}{2}\right)^x.\left(\frac{1}{2}\right)^{n-x}$$

Putting x = 0, 1, 2, 3, 4, 5 we find the probability of different number of heads.

The results can be represented in a table as follows:

Number of heads x	Probability $P(x)$	Expected frequency $= N \times P(x)$
0	$C_0^5.\left(\frac{1}{2}\right)^0.\left(\frac{1}{2}\right)^5 = \frac{1}{32}$	100
1	$C_1^5.\left(\frac{1}{2}\right)^1.\left(\frac{1}{2}\right)^4 = \frac{5}{32}$	500
2	$C_2^5.\left(\frac{1}{2}\right)^2.\left(\frac{1}{2}\right)^3 = \frac{10}{32}$	1000
3	$C_3^5.\left(\frac{1}{2}\right)^3.\left(\frac{1}{2}\right)^2 = \frac{10}{32}$	1000
4	$C_4^5.\left(\frac{1}{2}\right)^4.\left(\frac{1}{2}\right)^1 = \frac{5}{32}$	500

5	$C_5^5.\left(\frac{1}{2}\right)^5.\left(\frac{1}{2}\right)^0 = \frac{1}{32}$	100
Total	1	3200

9. A person tosses three coins simultaneously. He gets Rs. 5 if three heads appear, Rs. 3 if two heads appear, Rs. 2 if one head appears. He has to pay a penalty of Rs. 10 if no head appears. Find his expected amount.

Ans: suppose the penalty is Rs a if no head appears.

Head	Amount x	Probability $P(x)$	$x.P(x)$
0	$-a$	$C_0^3.p^0.q^3 = \frac{1}{8}$	$-\frac{a}{8}$
1	2	$C_1^3.p^1.q^2 = \frac{3}{8}$	$\frac{6}{8}$
2	4	$C_2^3.p^2.q^1 = \frac{3}{8}$	$\frac{12}{8}$
3	8	$C_3^3.p^3.q^0 = \frac{1}{8}$	$\frac{8}{8}$

For a fair game $Ex = 0$

$$\therefore Ex = 0$$
$$\therefore -\frac{a}{8} + \frac{6}{8} + \frac{12}{8} + \frac{8}{8} = 0$$
$$\therefore -a + 26 = 0$$
$$\therefore a = 26$$

$$\therefore He\ will\ loss\ Rs.26\ if\ no\ head\ appears.$$

10. The mean and variance of binomial distribution are 15 and 6 respectively. Find the value of n and p.

Ans: $Mean = np = 15$

$$Variance = npq$$
$$= 6$$

$$q = \frac{variance}{mean} = \frac{npq}{np} = \frac{6}{15} = \frac{2}{5} = 0.6$$

$$q = \frac{6}{15} \qquad \therefore q = 1 - p = 1 - 0.6 = 0.4$$

$$np = 15$$

$$n(0.6) = 15$$

$$n = \frac{15}{0.6}$$

$$n = 25$$

11.For a binomial variate n = 10 and $P(x = 5) = 2.P(x = 4)$ find the values of p.

Ans: for binomial variate

$$P(x) = C_x^n . p^x . q^{n-x}$$

$$P(x = 5) = 2 \times P(x = 4)$$

$$\therefore \quad C_5^{10} . p^5 . q^5 = 2 \times (C_4^{10} . p^4 . q^6)$$

$$\therefore \quad 252 . p^5 . q^5 = 2 \times (210 . p^4 . q^6)$$

$$\therefore \quad \frac{p^5 q^5}{p^4 q^6} = \frac{420}{252}$$

$$\therefore \quad \frac{p}{q} = \frac{5}{3}$$

$$\therefore \quad p = \frac{5}{3}\text{q}$$

$$Now \qquad p + q = 1$$

$$\therefore \quad p + \frac{5}{3}p = 1$$

$$\therefore \quad \frac{8}{3}p = 1$$

$$\therefore \quad p = \frac{5}{8}$$

➢ Question

1. A chance of winning the match of India against Pakistan is 2:1 if a series of three matches is to be played than what is the probability of the India to be win
   A. At least one match
   B. two matches

[Ans: 26/27, 12/27]

2. A chance of winning the match of India against Pakistan is 5:2 if a series of three matches is to be played than what is the probability of the India to be win
    A. At least one match
    B. Three matches

[Ans: 0.9768, 0.3645]

3. A chance of winning the match of India against Pakistan is 4:1 if a series of three matches is to be played than what is the probability of the India to be win
    A. At least one match
    B. two matches

[Ans: 0.992, 0.384]

4. If Wining Probability of India against South Africa is 3/5 and both teams will play 5 matches. So, find the probability that India wins 2 matches at least.

[Ans: 0.9129]

5. If the probability that a man aged 60 will live to be 70 is 0.65, what is the probability that out of 10 men aged 60, at least 7 will live up to 70?

[Ans: 0.5138]

6. A brokerage survey reports that 30 per cent of individual investors have used a discount broker, i.e., one which does not charge the full commission. In a random sample of 9 individuals. What is the probability that
    A. Exactly two of the sampled individuals have used a discount broker
    B. Not more than three have used a discount broker
    C. At least three of them have used a discount broker.

[Ans: 0.2656, 0.7297, 0.5372]

7. Assuming that half the population is rice eater and assuming that 100 investigators can take a sample of 10 individuals to see whether they are rice eaters. How many investigators would you expect to report that:
    A. Three people or less were rice eater.
    B. At least 7 people were rice eater.

[Ans: 0.1719, 0.1719]

8. There are 200 farms in a taluka among the bore wells made in these two 100 farms of the taluka, salted water is found in 20 farms. Find the probability of the event of not getting salty water in three out of five randomly selected farms from the taluka.

[Ans: 0.0729]

9. Seven coins are tossed simultaneously for 512 times. Find the frequencies of different number of heads and tails and tabulate the results.

[Ans:]

No of head x	0	1	2	3	4	5	6	7

Probability $P(x)$	$\frac{1}{128}$	$\frac{7}{128}$	$\frac{21}{128}$	$\frac{35}{128}$	$\frac{35}{128}$	$\frac{21}{128}$	$\frac{7}{128}$	$\frac{1}{128}$
Expected frequency $= N \times P(x)$	4	28	84	140	140	84	28	4

10. The mean and variance of binomial distribution are 18 and 4.5 respectively. Find the value of n and p.

[Ans: n = 24, p = 3/4]

11. The mean and variance of binomial distribution are 3.9 and 2.73 respectively. Find the value of n and p.

[Ans: n = 13, p = 0.3]

12. The parameter of a binomial distribution is 10 and 2/5. Calculated mean and variance.

[Ans: np = 4, npq = 12/5]

13. If the Variance and probability of failure are 3/4 and 3/4 respectively in binomial distribution then find mean.

[Ans: np = 1]

14. Mean and S.D. of binomial variable are 20 and 2 respectively then find its parameters

[Ans: n = 25, p = 4/5, q = 1/5]

15. For a binomial variate n = 8 and $2.P(x = 4) = P(x = 3)$ find the values probability of failure also find mean and variance.

[Ans: q = 1/3, np = 16/3, npq = 16/9]

16. For a binomial variate n = 6 and $2.P(x = 4) = P(x = 2)$ find the values probability of success also find mean and variance.

[Ans: p = 1/4, np = 3/2, npq = 9/4]

17. For a binomial variate n = 5 and $4.P(x = 2) = 6.P(x = 3)$ find the values probability of success also find mean.

[Ans: p = 0.4, np = 2]

18. For a binomial variate n = 5 and $P(x = 0): P(x = 1) = 1: 36$ find the values variance.

[Ans: p =3/4, npq = 9/4]

19. A person tosses three coins simultaneously. He gets Rs. 8 if three heads appear, Rs. 4 if two heads appear and Rs. 2 if one head appears. What penalty should be charged if no head appears in order that game is fair?

[Ans: 1.25]

2. Continuous probability distribution

Continuous data is the data that can be of any value. Over time, some continuous data can change. It may take any numeric value, within a potential value range of finite or infinite. If the variable which take all possible values within a given range and its value not countable than it is called continuous data. The continuous data can be broken down into fractions and decimals, i.e. according to measurement accuracy, it can be significantly subdivided into smaller sections.

Examples: Measurement of height and weight of a student, Daily temperature measurement of a place, Wind speed measured daily, etc.

Normal Distribution

Normal distribution is also known as normal probability distribution which is very useful for continuous random variables. Many statistical data concerned with business and economic problems are displayed in the form of normal distribution. Normal distribution is the cornerstone of the modern biostatistics. It is important for the reason that it plays a vital role in the theoretical and applied statistics. In many natural processes, random variation matches to a particular probability distribution which is known as the normal distribution. In 1733, English mathematicians deMoivre and Laplace first discovered normal distribution. Later 11 in 1812, German mathematician Gauss rediscovered it to analyse astronomical data, and it consequently became to be known as the Gaussian distribution.

Here we study Normal Distribution which Continuous probability distribution.

- **Definition**

A continuous random variable x is said to have random distribution with parameters of mean (μ) and standard deviation (σ) if its density function is,

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

Here,

e and π are mathematical constants
μ is mean
σ is standard deviation
If μ=0 and σ=1 then the variate is called as standard normal variate

- **Properties of normal distribution:**

1. This is a distribution of a continuous variable.
2. μ and σ are the parameters of this distribution.
3. The curve of the normal distribution is symmetrical about mean and it is bell shaped.
4. Mean, median and mode are equal in this distribution.
5. Quartiles are equidistant from median.
6. Its skewness is zero.
7. The total area under the normal curve is 1. The following are some of the important areas of the normal curve:

   Area under μ ± σ is 68.27%
   Area under μ ± 1.96σ is 95%
   Area under μ ± 2σ is 95.45%
   Area under μ ± 2.58σ is 95%
   Area under μ ± 3σ is 99.73%
8. The tails of the normal curve do not meet x axis. i.e., The curve is asymptotic to x axis.
9. Mean deviation about mean= $\frac{4}{5}$ × Standard deviation (approximately)
10. Quartile deviation = $\frac{2}{3}$ × Standard deviation (approximately)
11. Median deviation = standard deviation
12. The sum of two independent normal variates is also a normal variate.
13. When n is very large, and p and q are not very small, Binomial distribution tends to Normal distribution.
14. In fact, most of the distributions in statistics follow normal distribution when n is large.
15. The spread of a normal distribution is controlled by the standard deviation, σ.
16. The smaller the standard deviation, the more concentrated the data.

- **Questions**

1. Find the value of the average deviation if the standard deviation in a normal distribution is 5.
2. If the average deviation in the normal distribution is 12, find its Quartile deviation.
3. Find $Q1$, if the normal distribution has $Q3$ = 40 and mode = 2.
4. The Quartiles of a normal distribution are 8.64 and 14.32 respectively. So, find its mean and standard deviation.
5. If $Q3$ = 60 and the standard deviation is 3 in normal distribution, state the value of $Q1$ as well as the Mode.
6. If X is normal variable whose mean is 10 and S.D. is 2, p($x \leq 10$).
7. If X:N(25, 25), 1) p($20 \leq x \leq 35$) 2) p(x < 35).

8. The probability function of normal distribution is
   $P(x) = \frac{1}{\sqrt{2\pi 5}} \times e^{-\frac{1}{50}(x-50)^2}$, Find p(x ≥ 60).
9. The probability function of normal variable x is $P(x) = \frac{1}{\sqrt{2\pi 5}} \times e^{-\frac{1}{50}(x-50)^2}$, find
   A. p(x ≤ 45)
   B. p(x ≥ 55).
10. The probability function of normal distribution is $P(x) = \frac{1}{\sqrt{2\pi 8}} \times e^{-\frac{1}{128}(x-40)^2}$ then find
   A. p(32 ≤ x ≤ 40)
   B. p(x ≤ 48).
11. The mean and standard deviation of the wages of 6000 workers engaged in a factory are Rs. 4800 and Rs. 400 respectively. Assuming the distribution to be normally distributed estimate
   A. Percentage of workers getting wages more than Rs. 5400.
   B. Number of workers getting wages between Rs. 4200 and Rs. 4600.
12. An aptitude test for selecting officers in a bank was conducted on 1000 candidates. The average score is 42 and the standard deviation of scores is 24. Assuming normal distribution for the scores, find
   A. The number of candidates whose scores exceeds 58.
   B. The number of candidates whose scores lies between 30 and 60.
13. The average the weights of 4000 girls of a university are normally distributed. The mean and S.D. of weights are 95 lbs and 7.5 lbs. How many girls weigh between 100 and 110 lbs?
14. The income distribution of officers follows a normal distribution. The average income of an officer was Rs. 1,50,000 and the S.D. was Rs. 5000. If there were 242 officers drawing salary then find above 1.62,000 Rs. how many officers were there in the company?
15. In an intelligence test administered to 1000 children the average score is 42 and its s.d. is 24. Assuming that the scores are normally distributed find.
   A. The number of children exceeding score 50
   B. The number of children getting score between 30 and 54
   C. The minimum score of the most intelligent 100 students.
16. A normal distribution has mean 77, find the variance if 20% of the area under normal curve lies to the right of 90.
17. In a normal distribution 7% of the observations are less than 35 and 89% of the observations are less than 63. Find mean and standard deviation of the distribution.

❖ Sampling Techniques

Unit -3 : Introduction to R and working with Data

3.1. Overview of R and its applications in data analysis and statistics.

R is a programming language and software environment used for statistical computing, data analysis, and graphical representation. It was developed by statisticians and is widely used by data scientists, analysts, and statisticians for its extensive package ecosystem and powerful tools for data manipulation, visualization, and statistical modeling. R is open-source and offers a wide range of applications across different industries.

Key Features of R:

1. *Statistical Analysis*: R provides a wide range of statistical techniques, including linear and nonlinear modeling, time-series analysis, classification, clustering, etc.

2. *Data Manipulation*: R has powerful packages like dplyr, tidyr, and data.table for efficient data manipulation and cleaning.

3. *Visualization*: R offers excellent graphical capabilities with packages like ggplot2 for creating complex and publication-quality visualizations.

4. *Machine Learning*: R can be used for machine learning applications with packages such as caret, randomForest, and xgboost.

5. *Reproducible Research*: Tools like RMarkdown and Sweave allow researchers to combine code, text, and results in a single document for reproducibility.

6. *Extensive Libraries*: CRAN (Comprehensive R Archive Network) offers thousands of packages for various statistical and machine learning techniques.

Applications of R in Data Analysis and Statistics:

1. *Exploratory Data Analysis (EDA)*:

- R allows for detailed exploration of data through descriptive statistics and visualization techniques like histograms, box plots, and scatterplots.

- Useful libraries: dplyr, ggplot2, psych

2. *Statistical Modeling*:

- R supports regression analysis, hypothesis testing, ANOVA, and other advanced statistical models.

- It’s widely used for generalized linear models, time-series analysis, and multivariate statistics.

- Useful libraries: stats, lm(), forecast

3. *Machine Learning*:

- R is popular for developing and implementing machine learning algorithms for classification, regression, clustering, and more.

- Useful libraries: caret, randomForest, e1071

4. *Data Visualization*:

- R is famous for producing high-quality plots and charts, from basic to advanced visualizations like heatmaps and 3D plots.

- Useful libraries: ggplot2, lattice, plotly

5. *Time Series Analysis*:

- R is highly regarded for time-series forecasting and analysis, making it valuable in fields like finance and economics.

- Useful libraries: forecast, tseries, zoo

6. *Bioinformatics*:

- R is used extensively in biological and medical fields for genomic data analysis, microarray analysis, and clinical data analysis.

- Useful libraries: Bioconductor, edgeR, limma

7. *Text Mining and Natural Language Processing (NLP)*:

- R is used for processing text data, sentiment analysis, and building NLP models.

- Useful libraries: tm, text2vec, tidytext

8. *Big Data and Integration with Other Tools*:

- While R primarily handles data in memory, it can interface with big data tools like Hadoop and Spark.

- It can also integrate with other programming languages like Python, SQL, and Java.

9. *Econometrics and Financial Analysis*:

- R is widely used in econometrics for time-series analysis, econometric modeling, and financial risk analysis.

- Useful libraries: quantmod, PerformanceAnalytics

Strengths of R:

- Extensive ecosystem of packages

- High-quality visualization capabilities

- Comprehensive support for statistical analysis

- Active and supportive community

Limitations of R:

- Can be slower compared to languages like Python for large datasets

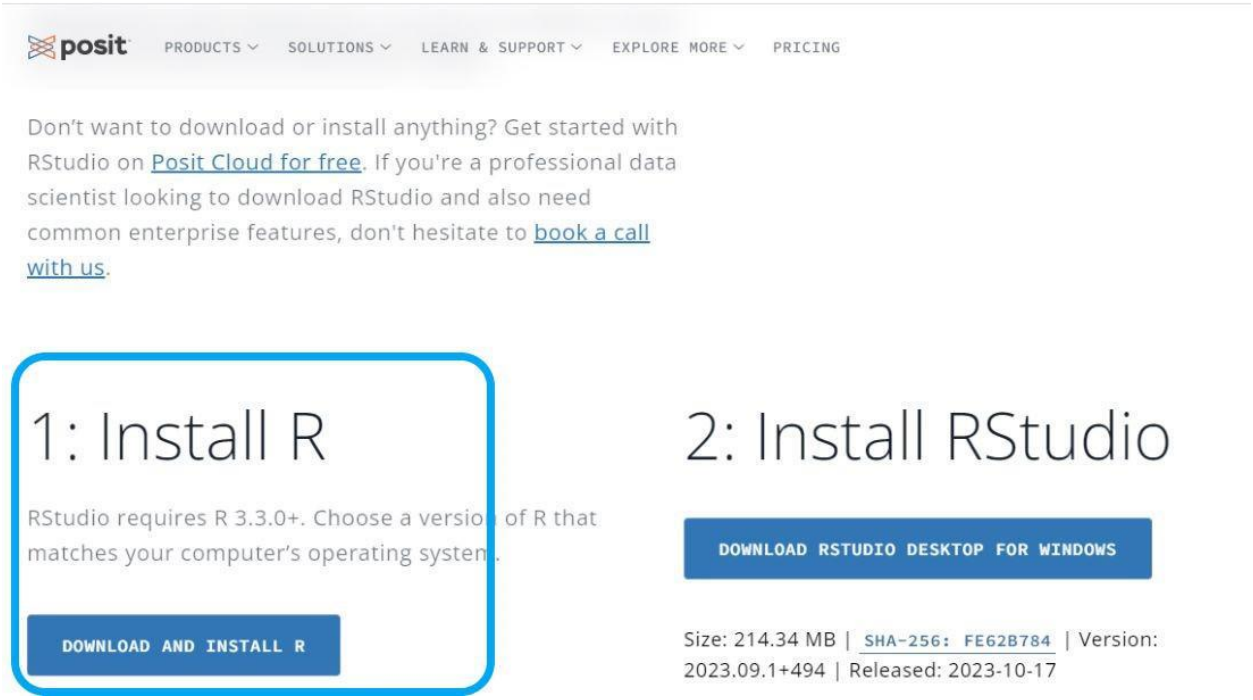
- High memory consumption for large datasets

- Steeper learning curve for beginners

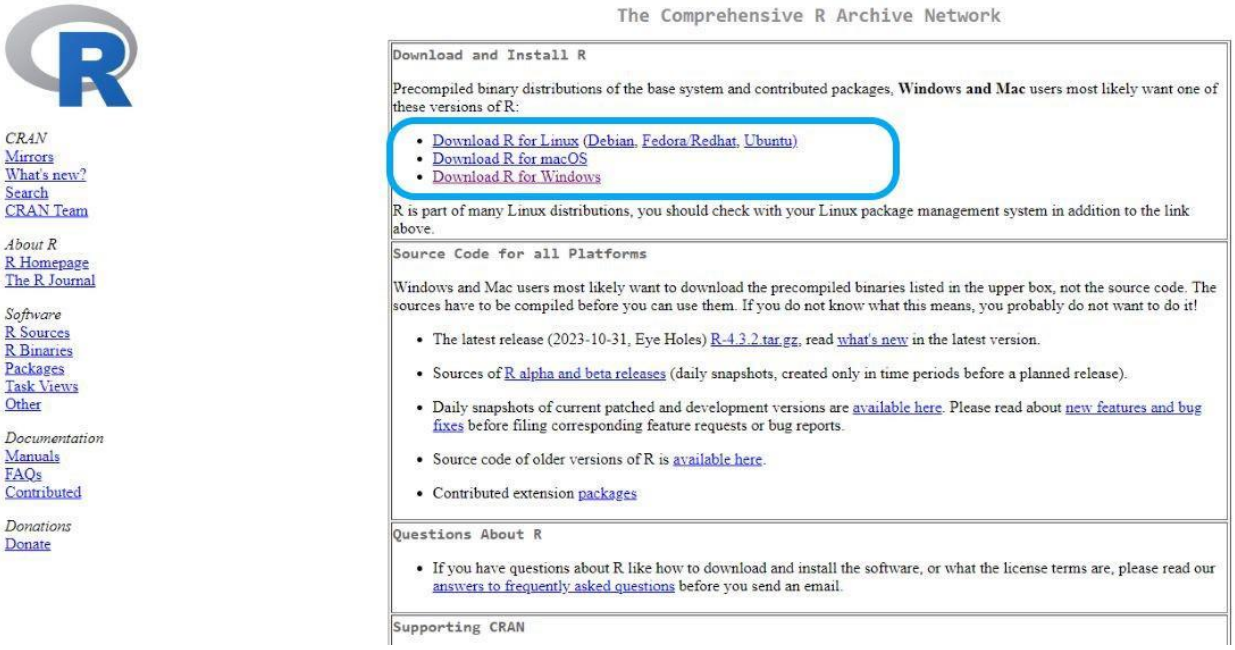
3.2. Installing R and RStudio

To Install R and R Studio on Windows we will have to download R and R Studio with the following steps.

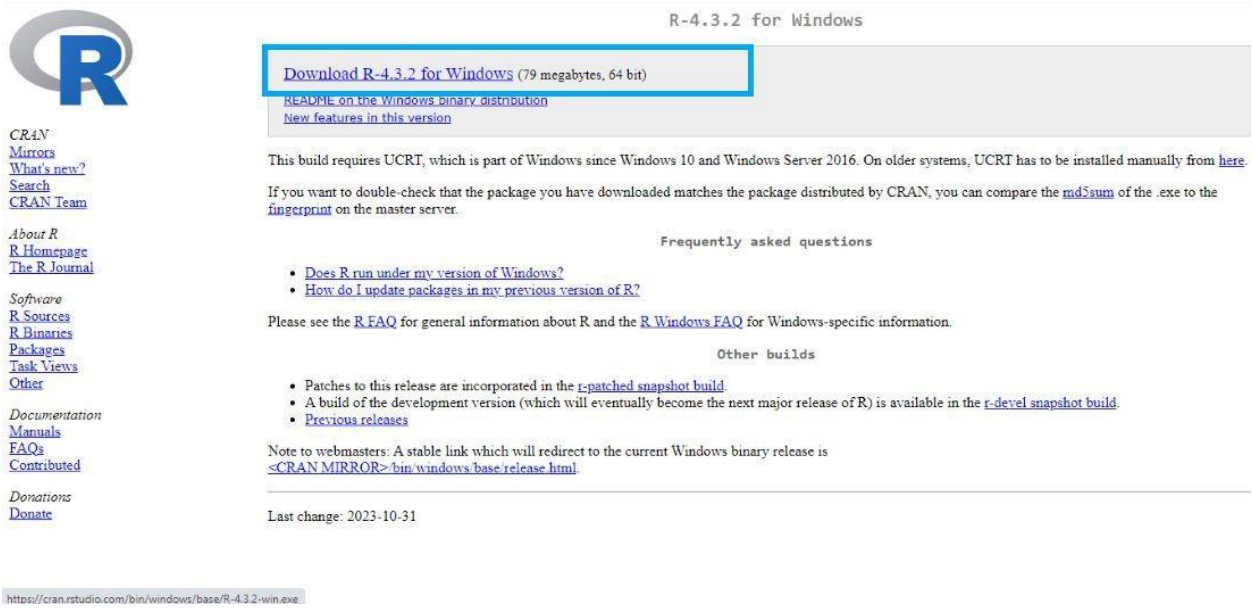
Step 1: *First, you need to set up an R environment in your local machine. You can download the same from r-project.org.*



You have to download both the applications first go with R Base and then install RStudio. after click on install R you will get a new page like this.



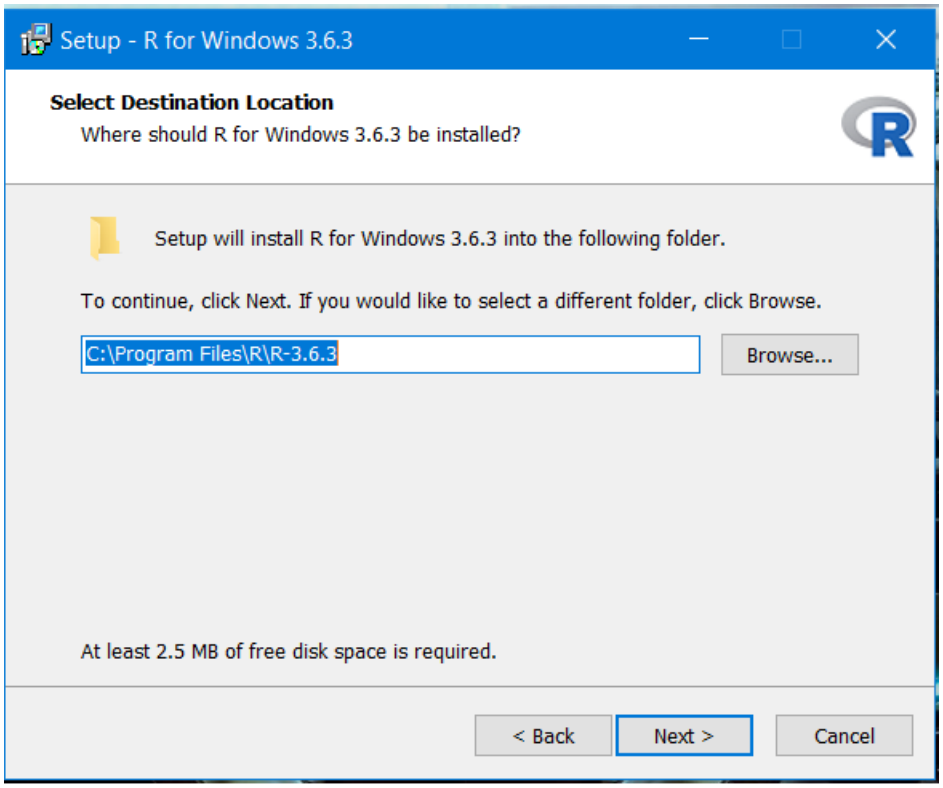
Here we can select the linux,mac or windows any one according to users system. you have to click on for which you want to install.



now click on the link show above in image so R base start downloading and after again go to main page and download and click on Install RStudio.

Steps to Install R and R Studio

Step 1: *After downloading R for the Windows platform, install it by double-clicking it.*



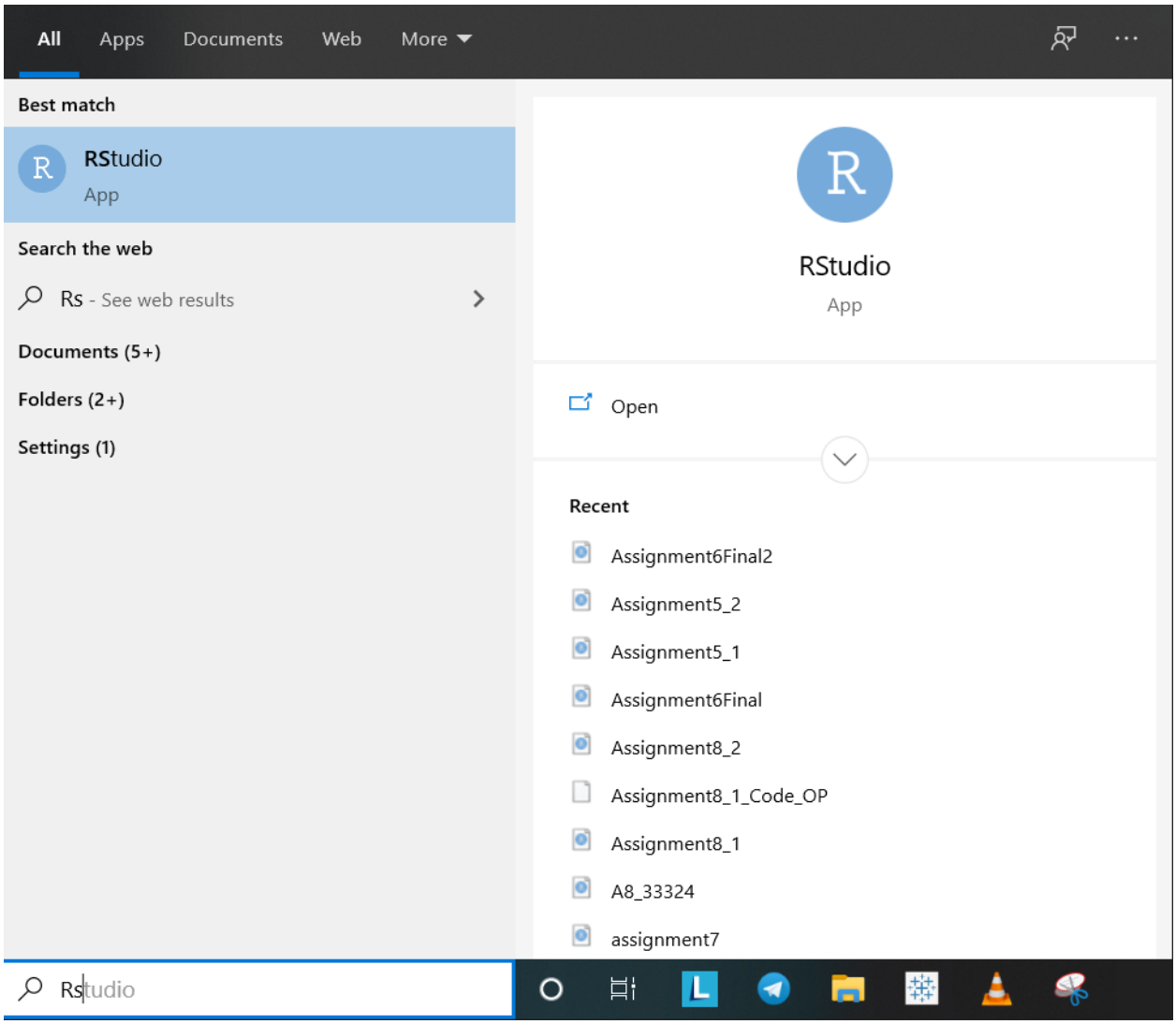
Step 2: *Download R Studio from their official page. Note: It is free of cost (under AGPL licensing).*

Step 3: *After downloading, you will get a file named "RStudio-1.x.xxxx.exe" in your Downloads folder.*

Step 4: *Double-click the installer, and install the software.*

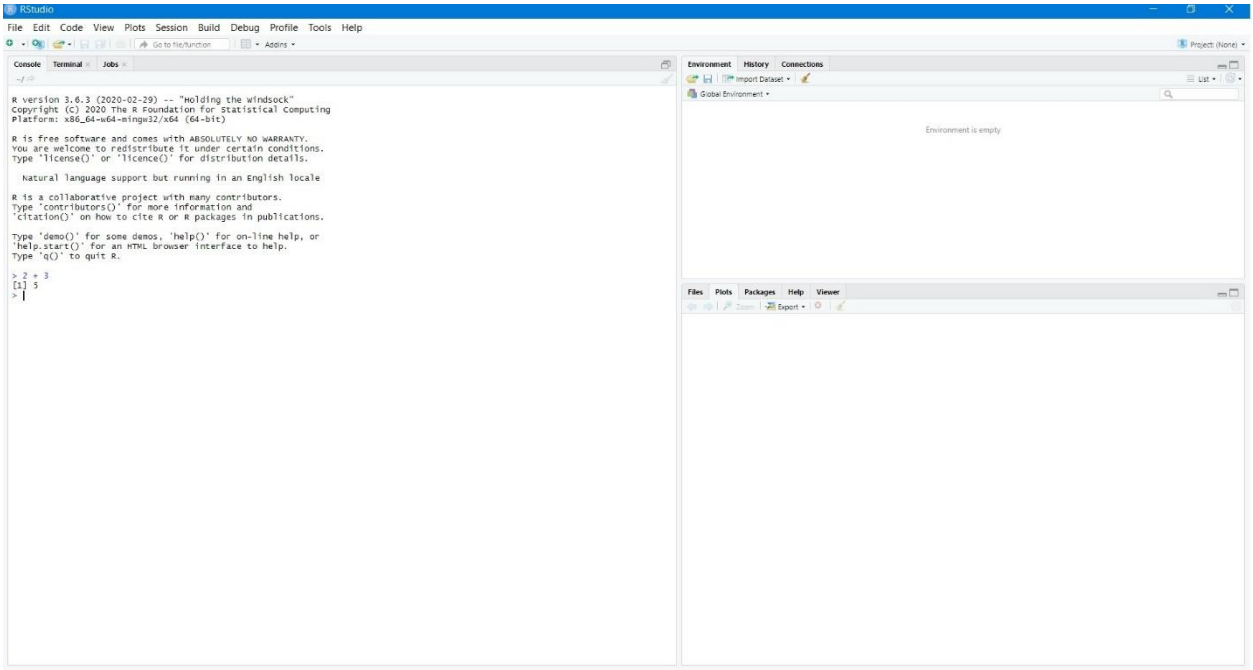
Step 5: Test the R Studio installation

- Search for RStudio in the Window search bar on Taskbar.



- Start the application.
- Insert the following code in the console.

Input : print('Hello world!')
Output : [1] "Hello world!"



Step 6: *Your installation is successful.*

3.3 Basic R syntax, variables and data types.

Basic of R- syntax:

- ‘ # ’ sign use for comment
- getwd ( ) # Get information about working directory
- setwd (“write file path”) # Set working directory

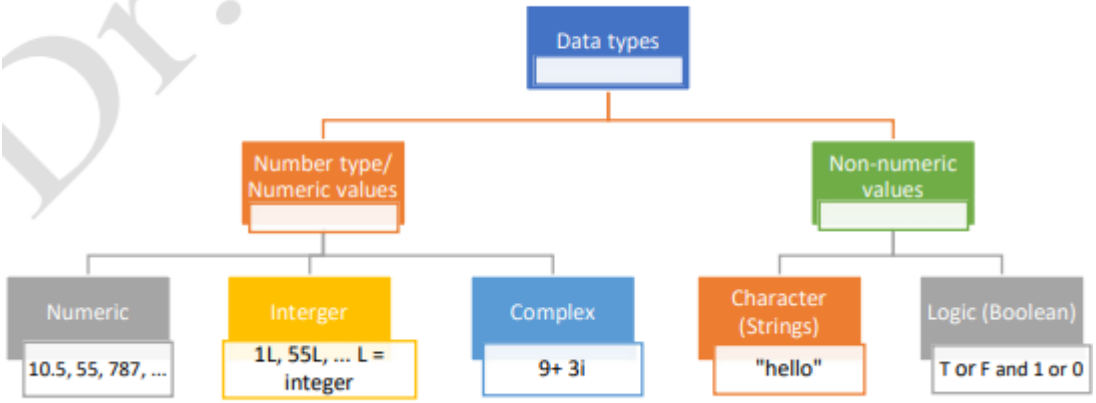
  For example, setwd(“C:/user/dell/Destop/New Folder”)
- install.package (“package name”) # Install package
- update.package (“package name”) # Update package
- ? defined problem # help

  For example; ? mean #get information about mean

  ? plot # get information about plot
- library (“package name”) # loading package
- q() ↲ # Quick way to exit

  then, Y/N appears. For exit, press Y ↲

Variable and Data types:



➔ class () # for check data types

For examples;

(1) Numeric x <- 10.5 class (x) ↲ [1] "numeric" y <- 55 class (y) ↲ [1] "numeric"	(2) Integer x <- 10L y <- 5 L class (x) ↲ [1] "integer" class (y) ↲ [1] "integer"	(3) Complex number x <- 3+5i x ↲ [1] "3+5i" class(x) [1] "complex"
(4) Character (Strings) x <- "hello" class(x) [1] "character"	(5) Logic (Boolean) x <- 5 y <- 3 x > y [1] TRUE x < y [1] FALSE	

You can convert from one type to another with the following function:

♣ as.numeric ( ) # convert into numeric

♣ as.integer ( ) # convert into integer

♣ as.complex ( ) # convert into complex number

x <- 1L # integer
y <- 2 # numeric
x ↲
[1] 1 L
y ↲
[1] 2
a <- as.numeric (x)
a ↲
	x <- 3 + 5i #complex number
	x ↲
	[1] 3 + 5i
	class (x)
	[1] "complex"
	a <- as.number (x)
	a ↲
	[1] 3 # imaginary parts discarded in correction

[1] 1 class (a) [1] "numeric" b <- as.integer (y) class(b) [1] "integer"	class (a) [1] "numeric" b <- as.integer (x) b ↲ [1] 3 class (b) [1] "integer"

3.4.Importing data into R from different file formats(CSV, Excel,etc..)

The collection of facts is known as data. Data can be in different forms. To analyze data using R programming Language, data should be first imported in R which can be in different formats like txt, CSV, or any other delimiter-separated files. After importing data then manipulate, analyze, and report it.

Import Data from a File in R Programming Language

In this article, we are going to see how to import different files in the R Programming Language.

=>CSV Files

Use the `read.csv()`

function to import CSV

files. Import

CSV file data Syntax: read.csv(path, header = TRUE, sep = ",")
Arguments :

path : The path of the file to be imported

header : By default : TRUE . Indicator of whether to import column headings.

sep = “,” : The separator for the values in each row.

=>Importing Excel Files

Use the `readxl` package to import Excel files. Install the package using `install.packages("readxl")` and then use the `read_excel()` function.

Install and load readxl package

install.packages("readxl")

library(readxl)

Import_Excel_file_data <- read_excel("path/to/your/file.xlsx")

=>Importing Other Formats

R supports importing data from various formats, including text files, JSON, and SQL databases. For example, to import data from a text file, you can use the `read.table()` function.

Import text file data <- read.table("path/to/your/file.txt", header = TRUE, sep = "\t")

3.5 Read, Write and view data using data frames.

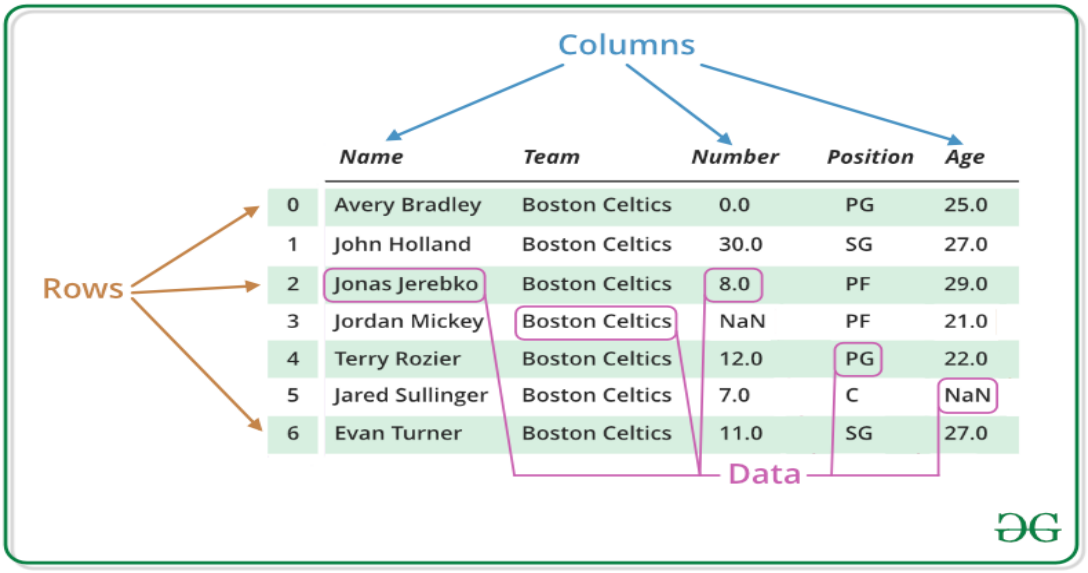
R Programming Language is an open-source programming language that is widely used as a statistical software and data analysis tool. **Data Frames in R Language** are generic data objects of R that are used to store tabular data.

Data frames can also be interpreted as matrices where each column of a matrix can be of different data types. R DataFrame is made up of three principal components, the data, rows, and columns.

R Data Frames Structure

As you can see in the image below, this is how a data frame is structured.

The data is presented in tabular form, which makes it easier to operate and understand.



Create Dataframe in R Programming Language

To create an R data frame use **data.frame()** function and then pass each of the vectors you have created as arguments to the function.

```
data <- data.frame(
                    Id=(1:4),
                    name=c('Sachin','Sourav','Ram','shyam')
                    )
> data
```

Output:

```
  Id   name
1  1 Sachin
2  2 Sourav
3  3    Ram
4  4  shyam
```

Summary of Data in the R data frame

In the R data frame, the statistical summary and nature of the data can be obtained by applying summary() function.

It is a generic function used to produce result summaries of the results of various model fitting functions. The function invokes particular methods which depend on the class of the first argument.

```
data <- data.frame(
                    Id=(1:4),
                    name=c('Sachin','Sourav','Ram','shyam')
) data <- data.frame(Id=(1:4),name=c('Sachin','Sourav','Ram','shyam'))
> summary(data)
```

Output:

```
     Id              name
 Min.   :1.00    Length:4
 1st Qu.:1.75    Class :character
 Median :2.50    Mode  :character
 Mean   :2.50
 3rd Qu.:3.25
 Max.   :4.00
```

Extract Data from Data Frame in R

Extracting data from an R data frame means that to access its rows or columns. One can extract a specific column from an R data frame using its column name.

```
result <- data.frame( data$name)
> print(result)
```

Output:

```
  data.name
1   Sachin
2   Sourav
3      Ram
4    shyam
```

Expand(add column) Data Frame in R Language

A data frame in R can be expanded by adding new columns and rows to the already existing R data frame.

```
data$sal <- c(20000,30000,23000,24000)

> data
```

Output:

```
  Id  name  sal
1  1 Sachin 20000
2  2 Sourav 30000
3  3   Ram 23000
4  4 shyam 24000
```

In R, one can perform various types of operations on a data frame like **accessing rows and columns, selecting the subset of the data frame, editing data frames, delete rows and columns in a data frame**, etc.

Please refer to DataFrame Operations in R to know about all types of operations that can be performed on a data frame.

Access Items in R Data Frame

We can select and access any element from data frame by using single $,brackets [ ] or double brackets [[]] to access columns from a data frame.

```
#Access Item using []

> data[1]
```

Output:

```
  Id
1  1
2  2
3  3
4  4
```

```
# Access Item Using [ [  ] ]

data[["name"]]
```

Output:

```
[1] "Sachin" "Sourav" "Ram"   "shyam"
```

```
# Access Item Using $

Data $ sal
```

Output:

```
[1] 20000 30000 23000 24000
```

Amount of Rows and Columns

We can find out how many rows and columns parsant in our dataframe by using dim function.

```
#Creating a data frame
data <- data.frame(
        Id=(1:4),
        name=c('Sachin','Sourav','Ram','shyam'),
        sal= c(20000,30000,23000,24000)
        )

        # find out the number of rows and columns
        dim(data)
```

Output:

```
[1] 4 3
```

Add Rows and Columns in R Data Frame

You can easily add rows and columns in a R DataFrame. Insertion helps in expanding the already existing DataFrame, without needing a new one.

Let's look at how to add rows and columns in a DataFrame ? with an example:

Add Rows in R Data Frame

To add rows in a Data Frame, you can use a built-in function **rbind().**

Following example demonstrate the working of rbind() in R Data Frame.

```
# Creating a dataframe
data <- data.frame(
Id=c(1:4),
name=c ('Sachin','Sourav','Ram','shyam'),
sal= c (20000,30000,23000,24000)
)

#Printing Existing DataFrame
print(data)
  Id   name   sal
1  1 Sachin 20000
2  2 Sourav 30000
3  3    Ram 23000
4  4  shyam 24000

#Adding New row
new_row  <-  c(5 , 'sita' , 25500)
data <- rbind ( data , new_row )
print(data)
```

Output:

```
   Id   name   sal
1   1 Sachin 20000
2   2 Sourav 30000
3   3    Ram 23000
4   4  shyam 24000
5   5   sita 25500
```

Add Columns in R Data Frame

To add columns in a Data Frame, you can use a built-in function **cbind().**

Following example demonstrate the working of cbind() in R Data Frame**.**

```
#Creating a data frame
data <- data.frame(
Id=(1:4),
name=c('Sachin','Sourav','Ram','shyam'),
sal= c(20000,30000,23000,24000)
)

# Adding New Column
gender <- c ( "male" , "male" , "male" , "male" , "female" )
data <- cbind(data,gender)

print ( data )
```

Output:

```
  Id   name   sal gender
1   1 Sachin 20000   male
2   2 Sourav 30000   male
3   3    Ram 23000   male
4   4  shyam 24000   male
5   5   sita 25500 female
```

Combine R Data Frame Vertically

If you want to combine 2 data frames vertically, you can use rbind() function. This function works for combination of two or more data frames.

```
# Creating Two Data Frames.
data <- data.frame(
Id=(1:4),
name=c('Sachin','Sourav','Ram','shyam'),
sal= c(20000,30000,23000,24000)
)
```

```
data2 <- data.frame(
Id=c( 6 ),
name= c ( 'ABC' ),
sal= c ( 40000 ),
gender= c ( "female" )
)
combined_df <- rbind ( data , data2 )
print ( combined_df )
```

```
  Id  name  sal gender
1  1 Sachin 20000   male
2  2 Sourav 30000   male
3  3   Ram 23000   male
4  4 shyam 24000   male
5  5  sita 25500 female
6  6   ABC 40000 female
```

Combine R Data Frame Horizontally:

If you want to combine 2 data frames horizontally, you can use cbind() function. This function works for combination of two or more data frames.

 # Creating Two Data Frames.

```
data <- data.frame(
Id=(1:4),
name=c('Sachin','Sourav','Ram','shyam'),
sal= c(20000,30000,23000,24000)
)

data2 <- data.frame(
marks= c ( 30,40,60,70,80 )
 )
combined_df <- cbind ( data , data2 )
> print(combined_df)
```

	Id	name	sal	gender	marks
1	1	Sachin	20000	male	30
2	2	Sourav	30000	male	40
3	3	Ram	23000	male	60
4	4	shyam	24000	male	70
5	5	sita	25500	female	80

functions of dataframe.

1)head():

it prints records from starts.

by default it prints 1st 6 records.

for ex:

head(dataframe,number_of_records)

➔head(data,3)

	Id	name	sal	gender
1	1	Sachin	20000	male
2	2	Sourav	30000	male
3	3	Ram	23000	male

2)tail():

it prints records from bottom.

by default it prints last 6 records.

for ex:

tail(dataframe,number_of_records)

➔tail(data,2)

	Id	name	sal	gender
4	4	shyam	24000	male
5	5	sita	25500	female

3)nrow():

it display total number of rows.

for ex:

nrow(dataframe)

➔nrow(data)

[1] 5

4)ncol():

it display total number of columns.

for ex:

ncol(dataframe)

➔ncol(data)

[1] 4

5)View():

it display data in tabular format.

for ex:

View(dataframe)

➔View(data)

6)rbind():

rbind() stands for row bind

rbind() is used to add new row in dataframe

for ex:

rbind(dataframe,rowName)

7)cbind():

cbind() stands for column bind

cbind() is used to add new column in dataframe

for ex:

cbind(dataframe,col_name)

Unit-4: Data Filtering and Clearing.

1. Data Filtering and Cleaning

Data filtering and cleaning are essential steps in preparing data for analysis. Filtering helps in selecting specific rows and columns of interest, while cleaning ensures that the data is free from errors, inconsistencies, and missing values. **The `dplyr` package in R provides a comprehensive set of functions to facilitate these processes.**

Key Functions and Examples:

`filter()`: Used to filter rows based on specific conditions.

`select()`: Used to select specific columns.

`mutate()`: Used to add or modify columns.

`arrange()`: Used to sort data by columns.

library(dplyr) # Sample data data <-

data.frame(name = c("Alice", "Bob",

"Charlie", "David"), age = c(25, 30, 35, 40),

salary = c(50000, 60000, 70000, 80000)

)

 # Filtering rows where age is greater than 30 filtered_data

<- data %>% filter(age > 30)

 # Selecting specific columns selected_data

<- data %>% select(name, age)

 # Adding a new column with mutated data

data <- data %>% mutate(salary_in_thousands = salary / 1000)

 # Arranging data by age

arranged_data <- data %>% arrange(age)

2. Subsetting and Filtering Data

Subsetting involves selecting a portion of data based on certain conditions. This can be done for both rows and columns.

Key Functions and Examples:

`subset()`: A base R function for subsetting data.

`filter()`: For row-wise subsetting.

`select()`: For column-wise subsetting.

 # Base R subsetting subset_data <- subset(data, age > 30)

 # dplyr subsetting filtered_data <- data %>% filter(age > 30) selected_data <- data %>% select(name, age)

Subsetting and filtering data are essential operations in data manipulation using R. Below are examples of how to subset and filter data using various techniques in R programming.

1. Subsetting Data

Subsetting involves selecting specific rows and columns from a data frame based on certain conditions.

 # Example: Subsetting by Rows and Columns

Suppose you have the following data frame:

 # Sample data frame df <- data.frame(

Name = c("John", "Alice", "Bob", "Carol"),

Age = c(28, 34, 22, 45),

Salary = c(50000, 60000, 45000, 70000),

Department = c("HR", "Finance", "IT", "Admin")

)

Display the original data frame

print(df)

Output:

	Name	Age	Salary	Department
1	John	28	50000	HR
2	Alice	34	60000	Finance
3	Bob	22	45000	IT
4	Carol	45	70000	Admin

Subsetting Specific Columns:

You can subset specific columns by their names or indices.

Subset columns by name

df_subset_cols <- df[, c("Name", "Salary")]

Display the subsetted data frame

print(df_subset_cols)

Output:

	Name	Salary
1	John	50000
2	Alice	60000
3	Bob	45000
4	Carol	70000

Subsetting Specific Rows:

You can subset specific rows based on row indices.

Subset rows by index

```
df_subset_rows <- df[1:2, ]
# Display the subsetted data frame
print(df_subset_rows)
```

Output:

	Name	Age	Salary	Department
1	John	28	50000	HR
2	Alice	34	60000	Finance

Subsetting Both Rows and Columns:

You can combine row and column subsetting.

```
# Subset specific rows and columns
df_subset <- df[1:3, c("Name", "Age")]
# Display the subsetted data frame
print(df_subset)
```

Output:

	Name	Age
1	John	28
2	Alice	34
3	Bob	22

2. Filtering Data

Filtering involves selecting rows that meet specific conditions.

#Example: Filtering Rows Based on a Condition

You can filter rows where a specific column meets a condition, such as selecting employees with a salary greater than 50,000.

```
# Filter rows where Salary is greater than 50,000
df_filtered <- df[df$Salary > 50000, ]
 # Display the filtered data frame
print(df_filtered)
```

Output:

```
  Name Age Salary Department
2 Alice  34   60000   Finance
4 Carol  45   70000    Admin
```

#Example: Filtering with Multiple Conditions

You can also filter data based on multiple conditions, such as selecting employees from the IT department with a salary less than 60,000.

```
# Filter rows where Department is 'IT' and Salary is less than 60,000
df_filtered_multiple <- df[df$Department == "IT" & df$Salary < 60000, ]
# Display the filtered data frame
print(df_filtered_multiple)
```

Output:

```
  Name Age Salary Department
3  Bob  22   45000      IT
```

3. Using `dplyr` for Subsetting and Filtering

The `dplyr` package provides more intuitive functions for subsetting and filtering.

#Example: Using `select()` to Subset Columns

```
# Install and load dplyr package install.packages("dplyr")
 library(dplyr)
# Subset specific columns using select()
df_select <- select(df, Name, Salary)
# Display the subsetted data frame
print(df_select)
```

```
Name Salary
```

1 **John 50000**

2 **Alice 60000**

3 **Bob 45000**

4 **Carol 70000**

#Example: Using `filter()` to Filter Rows

 # Filter rows where Age is greater than 30 using filter()

df_filter <- filter(df, Age > 30)

 # Display the filtered data frame

print(df_filter)

Output:

Name Age Salary Department

2 Alice 34 60000 Finance

4 Carol 45 70000 Admin

3. Adding, Removing, and Renaming Variables/Columns

Adding, removing, and renaming columns are common operations during data manipulation.

Key Functions and Examples:

`mutate()`: For adding or modifying columns.

`select()`: With the `-` sign for removing columns.

`rename()`: For renaming columns.

 # Adding a new column data <- data %>% mutate(new_column = age + 10)

 # Removing a column data <- data %>% select(-salary) # Renaming a column data <- data %>% rename(new_age = age)

Here are examples of adding, removing, and renaming variables (columns) in a data frame using R programming:

1. Adding Variables/Columns

You can add a new column to a data frame by simply assigning a vector of values to a new column name.

#Example: Adding a New Column

 # Sample data frame df <- data.frame(

Name = c("John", "Alice", "Bob", "Carol"),

Age = c(28, 34, 22, 45),

Salary = c(50000, 60000, 45000, 70000)

)

 # Add a new column 'Department' df$Department <- c("HR", "Finance", "IT", "Admin")

 # Display the updated data frame print(df)

Name Age Salary Department 1 John 28 50000 HR

2 Alice 34 60000 Finance

3 Bob 22 45000 IT

4 Carol 45 70000 Admin

2. Removing Variables/Columns

To remove a column, you can either set it to `NULL` or use the `select()` function from the `dplyr` package.

#Example: Removing a Column Using `NULL`

 # Remove the 'Salary' column by setting it to NULL df$Salary <- NULL

 # Display the updated data frame print(df)

Output:

Name Age Department

1 John 28 HR

2 Alice 34 Finance

3 Bob 22 IT

4 Carol 45 Admin

#Example: Removing a Column Using `dplyr`

 # Install and load dplyr package install.packages("dplyr") library(dplyr)

 # Remove the 'Age' column using select() function df <- select(df, -Age)

 # Display the updated data frame print(df)

Output:

	Name	Department
1	John	HR
2	Alice	Finance
3	Bob	IT
4	Carol	Admin

3. Renaming Variables/Columns

Renaming columns can be done using the `names()` function or the `rename()` function from the `dplyr` package.

#Example: Renaming Columns Using `names()` # Rename the 'Department' column to 'Dept' names(df)[names(df) == "Department"] <- "Dept"

 # Display the updated data frame print(df)

	Name	Dept
1	John	HR
2	Alice	Finance
3	Bob	IT
4	Carol	Admin

#Example: Renaming Columns Using `dplyr`

 # Rename the 'Dept' column back to 'Department' using rename() function df <- rename(df, Department = Dept) # Display the updated data frame print(df)

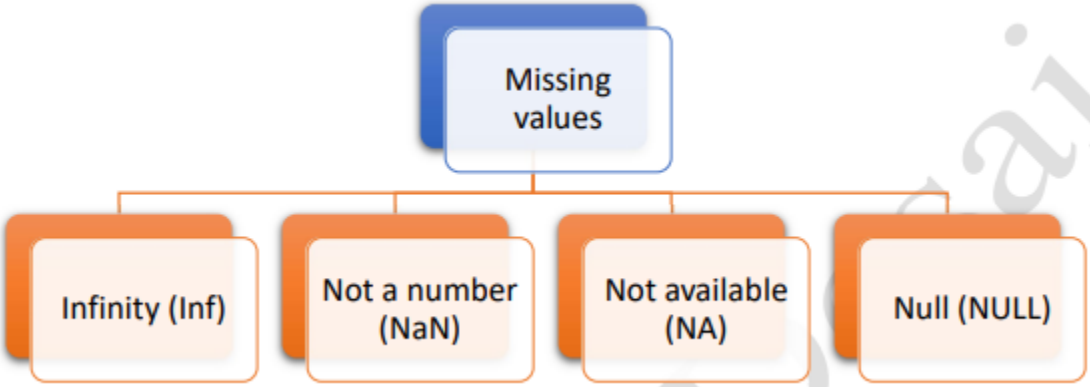
Output:

	Name	Department

1	**John**	**HR**
2	**Alice**	**Finance**
3	**Bob**	**IT**

4.4 Identfying and handling missing values.

Identify missing values



 # Create data frame

data <- data.frame (ID = 1:5,

Gender = c (“M”, “F”, NA, “M”, “F”),

Age = c (25, NA, 30, 40, NA))

is.na (data) ↲ # Check NA value in data frame: TRUE for NA value

ID gender Age

[1,] FALSE FALSE FALSE

[2,] FALSE FALSE TRUE

[3,] FALSE TRUE FALSE

[4,] FALSE FALSE FALSE

[5,] FALSE FALSE TRUE

is. null (data) ↲ # Check NULL value in data frame: TRUE for null value

is. infinite (data) ↲ # Check infinite value in data frame: TRUE for infinite value

sum (is.na (data)) ↲ # Count number of missing value ‘NA’

[1] 3

na.omit (data) ↲ # Show only non-NA column and row

ID gender Age

1 1 M 25

4 4 M 40

colSums (is.na (data)) ↲ # Number of NA, column-wise

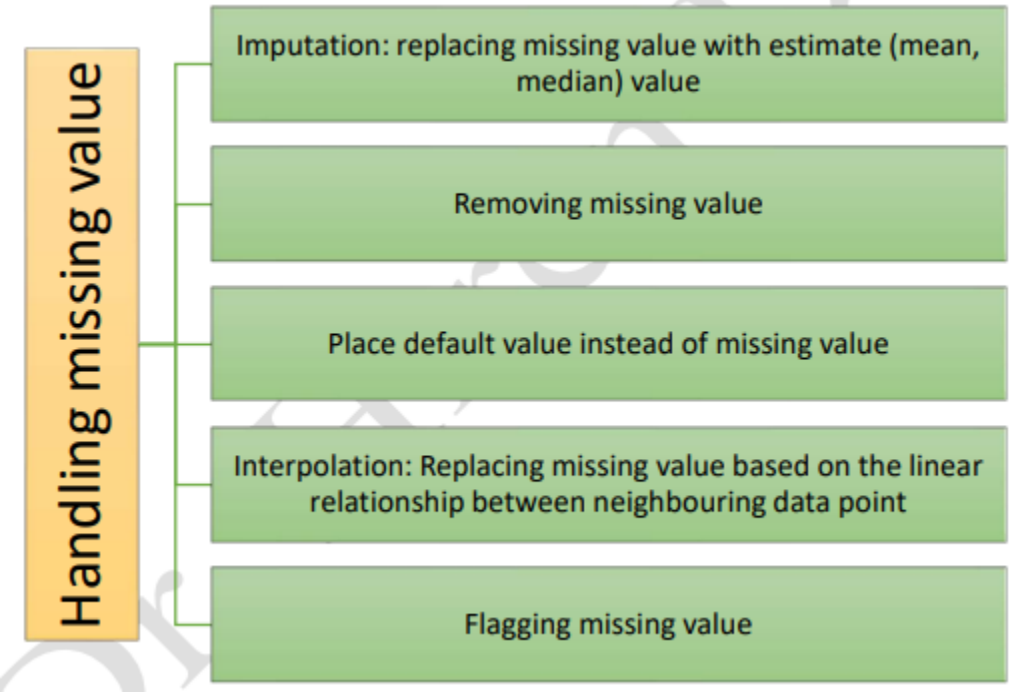
ID gender Age

0 1 2

rowSums (is.na (data)) ↲ # Number of NA, row-wise

[1] 0 1 1 0 1

Handling missing values



 # Create data frame

data <- data.frame (ID = 1:6,

name = c (“A”, “B”, “C”, “D”, “E”, “F”),

age = c (25, 35, NA, 30, 40, NA),

salary = c (5000, NA, 6000, 7000, NA, 8000))

data ↲

	ID	name	age	salary
1	1	A	25	5000
2	2	B	35	NA
3	3	C	NA	6000
4	4	D	30	7000
5	5	E	40	NA
6	6	F	NA	8000

(1) Imputation:

```
# Replace with mean value
data$age [is.na (data$age)] <- mean
(data$age, na.rm = TRUE)
data ↲
```

	ID	name	age	salary
1	1	A	25.0	5000
2	2	B	35.0	NA
3	3	C	32.5	6000
4	4	D	30.0	7000
5	5	E	40.0	NA
6	6	F	32.5	8000

```
# Replace with median value
data$salary [is.na (data$salary)] <- median
(data$salary, na.rm = TRUE)
data ↲
```

	ID	name	age	salary
1	1	A	25.0	5000
2	2	B	35.0	6500
3	3	C	32.5	6000
4	4	D	30.0	7000
5	5	E	40.0	6500
6	6	F	32.5	8000

(2) Removing missing value:

```
na.omit (data)                    # Remove NA valued row
```

```
  ID name age salary
1  1    A  25   5000
4  4    D  30   7000
```

(3) Place default value:

```
default_age <- 30                 # You can use any numerical value
default_salary <- 3000
data$age <- [is.na (data$age)] <- default_age
```

OR

```
data$age [is.na (data$age)] <- 30      # You are not defined any code, then used this code directly
data$age [is.na (data$salary)] <- 3000
```

(4) Interpolation:

```
install.package(zoo)              # Install 'zoo' package

library (zoo)                     # Loading 'zoo' package
data$salary_interp <- na.approx (data$salary)      # Approximate value in place of NA value
data ↲
```

```
  ID name age salary salary_interp
1  1    A  25   5000         5000
2  2    B  35     NA         5500
3  3    C  NA   6000         6000
4  4    D  30   7000         7000
5  5    E  40     NA         7500
6  6    F  NA   8000         8000
```

(5) Flagging missing value:

```
ifelse   (is.na (data$age), 1, 0)          # 1 for NA flagging & 0 for non- NA flagging.
```

here you can write any numerical or logical values (e.g. 5, 10, T, F, …)

```
data$age <- ifelse (is.na (data$age), 1, 0)
data ↲
```

	ID	name	age	salary
1	1	A	0	5000
2	2	B	0	NA
3	3	C	1	6000
4	4	D	0	7000
5	5	E	0	NA
6	6	F	1	8000

```
data$age <- ifelse (is.na (data$age), T, F)
data ↲
```

	ID	name	age	salary
1	1	A	FALSE	5000
2	2	B	FALSE	NA
3	3	C	TRUE	6000
4	4	D	FALSE	7000
5	5	E	FALSE	NA
6	6	F	TRUE	8000

4.5 Data type conversion and recoding variables.

Recoding Variable

(1) Grading

```
data <- data.frame (ID = 101:106,
                    name = c ("A", "B", "C", "D", "E", "F"),
                    mark = c (20, 40, 34, 50, 34, 35))

data ↲
```

	ID	name	mark
1	101	A	20
2	102	B	40
3	103	C	34
4	104	D	50
5	105	E	34
6	106	F	35

```
data$grade <- ifelse (data$mark >= 35, "PASS", "FAIL")

data ↲
```

	ID	name	mark	grade
1	101	A	20	FAIL
2	102	B	40	PASS
3	103	C	34	FAIL
4	104	D	50	PASS
5	105	E	34	FAIL
6	106	F	35	PASS

(2) Sorting:

```
data <- data.frame (ID= 101:105,
        name = c (“A”, “B”, “C”, “D”, “E”),
        title = c (“clark”, “peon”, “teacher”, “TA”, “principal”))

data ↲
```

ID	name	title
1	101	A Clark
2	102	B peon
3	103	C teacher
4	104	D TA
5	105	E principal

```
data$category <- ifelse (data$title %in% c ("clark", "peon"),
"non-teaching", "teaching")

data ↲
```

	ID	name	title	category
1	101	A	clark	non-teaching
2	102	B	peon	non-teaching
3	103	C	teacher	teaching
4	104	D	TA	teaching
5	105	E	principal	teaching

(3) double if, for grading

```
customers <- data.frame (ID=1:10, review= c( 2, 3.5, 5, 2.5, 4, 6,
1.5, 3, 7, 10))

customers$rating <- ifelse (customers$review <= 2.5, "poor",
                    ifelse (customers$review <= 3.5, "medium",
"good"))

customers ↲
```

	ID	review	rating
1	1	2.0	poor
2	2	3.5	medium
3	3	5.0	good
4	4	2.5	poor
5	5	4.0	good
6	6	6.0	good
7	7	1.5	poor
8	8	3.0	medium
9	9	7.0	good
10	10	10.0	good

(4) grading with 'dplyr'

```
result <- data.frame (ID= 101:110,
                      score= c (80, 65, 90, 75, 85, 25, 37, 42, 55, 17))

result ↲
```

```
   ID score
1  101    80
2  102    65
3  103    90
4  104    75
5  105    85
6  106    25
7  107    37
8  108    42
9  109    55
10 110    17
```

```
result <- result %>%
        mutate (grade= case_when(
           score >= 90 ~ "A+",
           score >= 80 ~ "A",
           score >= 70 ~ "B+",
           score >= 60 ~ "B",
           score >= 50 ~ "C+",
           score >= 35 ~ "C",
              TRUE ~ "F" ))

result ↲
```

```
   ID score grade
1  101    80     A
2  102    65     B
3  103    90    A+
4  104    75    B+
5  105    85     A
6  106    25     F
7  107    37     C
8  108    42     C
9  109    55    C+
10 110    17     F
```

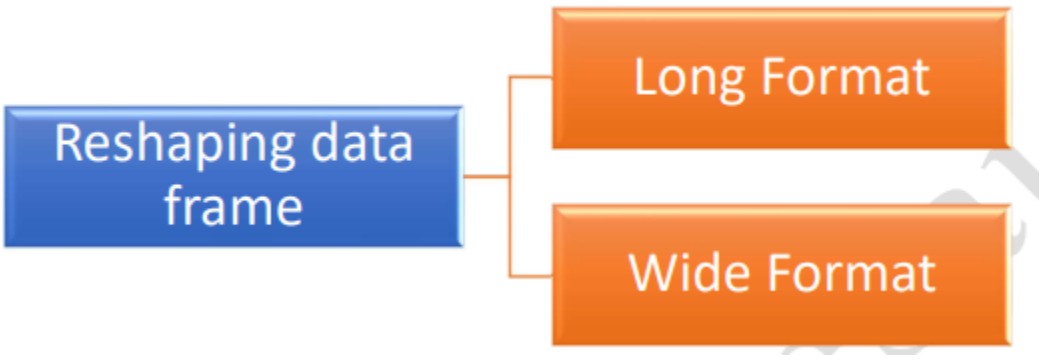
Unit-5: Working with Data in R.

Reordering and reshaping data frames.

->Re-ordering

```
# re-ordering using “dplyr”
reo_data1 <- data %>% select (ID, age, name, city)
                        OR
reo_data1 <- select (data, ID, age, name, city)
reo_data ↲
ID age name city
# re-ordering using “base package”
reo_data2 <- df [ , c (“name”, “city”, “age”, “ID”)]
reo_data2 ↲
name city age ID
asc_order <- data %>% arrange (age)                    # Ascending order
des_order <- data %>% arrange (desc (age))             # Descending order
                        OR
des_order <- data %>% arrange (-age)
order1 <- data %>% arrange (age, .na_last = TRUE)      # NA palace in last column
```

-> Re-shaping




```
install.package (“tidyverse”)          # Install package
library (“tidyr”)                      # Loading package
```

```
# Construction of data frame
data <- data.frame (ID = 101:105,
                    name = c (“AB”, “CD”, “EF”, “GF”, “IJ”),
                    sex = c (“M”, “F”, “M”, “F”, “M”),
                    mark = c (12, 15, 18, 15, 20))
data ↲
   ID name sex mark
1  101  AB  M  12
2  102  CD  F  15
3  103  EF  M  18
4  104  GH  F  15
5  105  IJ  M  20
```

```
# Wider format
wide_data <- pivot_wider (data,
                          names_from = sex,
                          values_from = mark)
wide_data ↲
    ID name    M    F
1   101 AB     12   NA
2   102 CD     NA   15
3   103 EF     18   NA
4   104 GH     NA   15
5   105 IJ     20   NA
```

```
# Longer format
long_data <- pivot_longer (wide_data,
                           cols= c (“M”, “F”),
                           names_to = “sex”,
                           values_to = “mark”)
long_data ↲
    id name  sex   mark
1   101 AB   M     12
2   101 AB   F     NA
3   102 CD   M     NA
4   102 CD   F     15
5   103 EF   M     18
6   103 EF   F     NA
7   104 GH   M     NA
8   104 GH   F     15
9   105 IJ   M     20
10  105 IJ   F     NA
```

Merging and joining Data Frames.

data1 <- data.frame (serial = c("A", "B", "C", "D"), score1 = c(1, 2, 3, 4))

data2 <- data.frame (serial = c("B", "D", "E", "F"), score2 = c(5, 6, 7, 8))

data1 ↲

	serial	score1
1	A	1
2	B	2
3	C	3
4	D	4

data2 ↲

	serial	score2
1	B	5
2	D	6
3	E	7
4	F	8

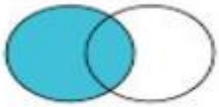
inner join/ merge

inner_data <- merge(data1, data2, by= "serial", all= FALSE)

inner_data ↲

	serial	score1	score2
1	B	2	5
2	D	4	6

left join/ merge

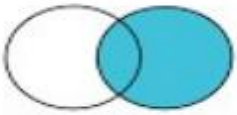


Left Join

```
left_data <- merge (data1, data2, by= "serial", all.x = TRUE)

left_data ↲
```

serial	score1	score2
1	A	1	NA
2	B	2	5
3	C	3	NA
4	D	4	6

right join/ merge

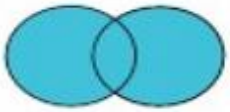


Right Join

```
right_data <- merge (data1, data2, by= "serial", all.y = TRUE)

right_data ↲
```

serial	score1	score2
1	B	2	5
2	D	4	6
3	E	NA	7
4	F	NA	8

outer join/ merge



Full Outer Join

```
outer_data <- merge (data1, data2, by= "serial", all= TRUE)

outer_data ↲
```

serial	score1	score2
1	A	1	NA
2	B	2	5
3	C	3	NA
4	D	4	6
5	E	NA	7
6	F	NA	8

- Using “dplyr” package

```
data_inner <- inner_join (data1, data2, by= “serial”)        # inner join

data_left <- left_join (data1, data2, by= “serial”)          # left join

data_right <- right_join (data1, data2, by= “serial”)        # right join

data_outer <- full_join (data1, data2, by= “serial”)         # outer join
```

Generating Frequency Tables and Cross-Tabulation.

(1) # to generate frequency table

```
data <- c ("M", "F", "F", "M", "F", "M", "F", "M", "M")
freq_data <- table(data)
freq_data ↲
```

```
data
F M
4 5
```

(2) # to generate cross-tabulation (contingency table) or flat-contingency table

```
d1 <- c ("A", "B", "A", "C", "A", "B", "B", "C", "A", "C")
d2 <- c ("X", "Y", "Y", "X", "X", "Y", "X", "Y", "X", "Y")

cross_table <- table (d1, d2)   OR   flat_table <- ftable (d1, d2)
cross_table ↲                   OR   flat_table ↲
```

```
   d2
d1  X Y
 A 3 1
 B 1 2
 C 1 2
```

(3) # marginal totals

```
marginal_data <- addmargins (table (d1, d2))

marginal_data ↲
```

```
d2
d1    X  Y Sum
 A    3  1  4
 B    1  2  3
 C    1  2  3
 Sum  5  5  10
```

(4) # generate cross tabulation by 'xtabs'

```
gender <- c ("male", "female", "male", "female", "male")
education <- c ("yes", "no", "yes", "yes", "no")
freq_data <- xtabs (~ gender + education)
freq_data ↲
```

```
          education
gender    no   yes
 female   1    1
 male     1    2
```

(5) # sorting/ grouping with 'xtabs'

```
data <- data.frame (gender= c ("M", "F", "M", "F", "M"),
   smoking = c ("Y", "Y", "N", "N", "Y"),
   age_group = c ("10-20", "20-30", "20-30", "10-20", "20-30"))
data ↲
```

```
  gender smoking age_group
1    M       Y    10-20
2    F       Y    20-30
3    M       N    20-30
4    F       N    10-20
5    M       Y    20-30
```

```
sort_data <- xtabs (~ gender + smoking + age_group, data=data)

sort_data ↲
```

```
, , age_group = 10-20
      smoking
gender N Y
   F   1 0
   M   0 1

, , age_group = 20-30
      smoking
gender N Y
   F   0 1
   M   1 1
```

```
sort_d1 <- xtabs(~ age_group + gender + smoking, data=data)

sort_d1 ↲
```

```
, , smoking = N
          gender
age_group F M
    10-20 1 0
    20-30 0 1

, , smoking = Y
          gender
age_group F M
    10-20 0 1
    20-30 1 1
```

Commands to measures of central tendency and dispersion.

Measurement of central tendency

(1) Mean: Arithmetic average of the set values

(2) Median: middle value of the data set, from least to greatest

For even observation; Two middle values observed

For odd observation; one middle value observed

(3) Mode: most frequently occurring value

Create data frame

```
data <- data.frame (no.= 1:9,
Division = c ("SY1", "SY2", "SY3", "SY4",
"SY5", "SY6", "SY7", "SY8", "SY9"),
students = c (80, 85, 82, 90, 92, 94, 89, 70,
75))

data ↲
```

	no.	Division	students
1	1	SY1	80
2	2	SY2	85
3	3	SY3	82
4	4	SY4	90
5	5	SY5	92
6	6	SY6	94
7	7	SY7	89
8	8	SY8	70
9	9	SY9	75

```
mean_data <- mean(data$students)                    # Calculate mean value
 mean_data ↲
[1] 84.11111
 media_data <- median (data$students)               # Calculate median value
media_data ↲
[1] 85
```

- "DescTools" package used for mode

```
 install.package ("DescTools")                      # Install package
library ("DescTools")                               # Loading package
series <- c (12, 18, 12, 13, 15, 18, 12, 20, 22, 18)
mode_data <- Mode (series)                          # Calculate Mode value
mode_data ↲
 [1] 12 18
 attr(,"freq")
 [1] 3
```

Measurement of dispersion

(1) Range: difference between maximum and minimum value

(2) Variance: Average square difference from mean

(3) Standard Deviation: square-root of variance

(4) Inter-quartile-Range (IQR): the range between first quartile, Q1 (25%) and third quartile, Q3 (75%)

```
data <- data.frame (no.= 1:9,
Division = c ("SY1", "SY2", "SY3", "SY4",
"SY5", "SY6", "SY7", "SY8", "SY9"),
students = c (80, 85, 82, 90, 92, 94, 89, 70,
75))

data ↲
```

```
  no. Division students
1  1      SY1       80
2  2      SY2       85
3  3      SY3       82
4  4      SY4       90
5  5      SY5       92
6  6      SY6       94
7  7      SY7       89
8  8      SY8       70
9  9      SY9       75
```

```
summary(data$students)                    # create summary of data frame

 Min.  1st Qu.  Median  Mean 3rd Qu. Max.

70.00   80.00   85.00  84.11  90.00  94.00
```

```
# Calculate manually:
data_mean <- mean(data$students)

data_sqr <- (data$students-data_mean)^2

variance <- sum(data_sqr)/

length(data$students)

variance ↲

[1] 58.09877

sd_data <- sqrt(variance)
sd_data ↲
[1] 7.622255
```

```
# Calculate directly with code

variance <- var (data$students)

variance ↲

[1] 65.36111

sd_data <- sd(data$students)
sd_data ↲
[1] 8.084622
```

```
IQR(data$students)                        # code for IQR

 [1] 10

range(data$students)                      # code for range

 [1] 70 94
```

Commands to explore view data distributions graphically (Bell Curve).

Graphical exploration of data distributions helps in understanding the shape and spread of the data. The bell curve, or normal distribution, is a common way to visualize continuous data.

➔ Graphical parameters

- type = plot the supplied coordinate (eg., dot, lines, ...)
- main = plot title
- xlab, ylab = horizontal and vertical axis label, respectively
- col = colour to use for plotting point and lines
- pch = "point character"
- lty = line type (eg., solid, dotted, dashed, ...)
- lwd = line width- thickness of plot line
- xlim, ylim= limit for horizontal range and vertical range, respectively

"p" – scatter plot	"l" – lines	"b" – both point & lines	"o" – overplotted point & lines
"h"- histogram like vertical lines	"s" – step function	"n" – no plotting	

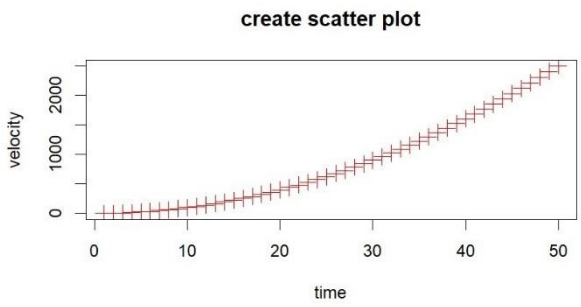
pch

"1" – circle	"2" – triangle point up	"3" – square	"4" - cross
"5" – x-symbol	"6" – diamond	"." – small filled circle	"*"- star
"A" to "Z" – uppercase letter			

- Scatter plot:

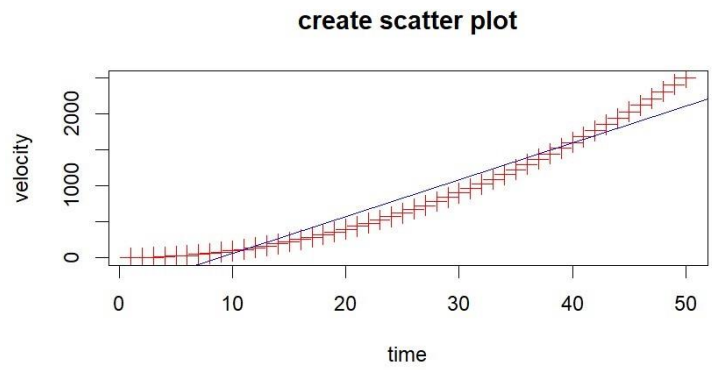
Create single plot

```
x <- 1:50
y <- x^2
plot (x, y, main= "create scatter plot",
      xlab="time", ylab="velocity",
      col="red", pch=3, cex=1.5, lwd=1)
```



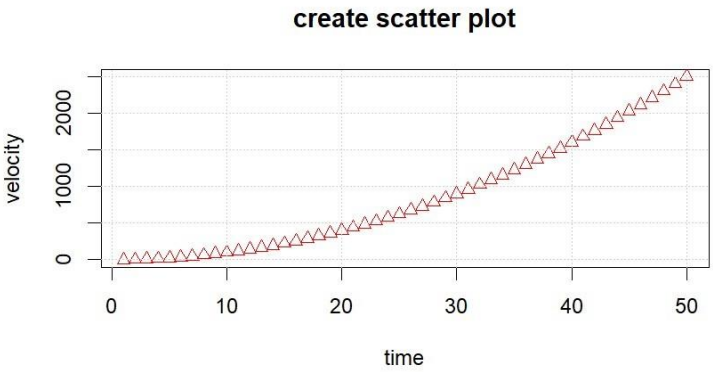
adding a trend line

```
abline (lm (y~x), col= “blue”)
```

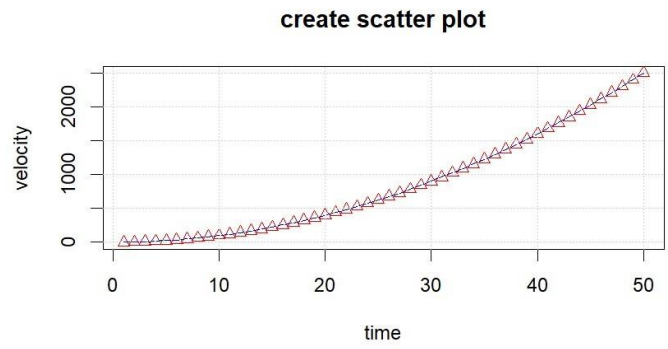


add grid and trend line

```
grid ()
```

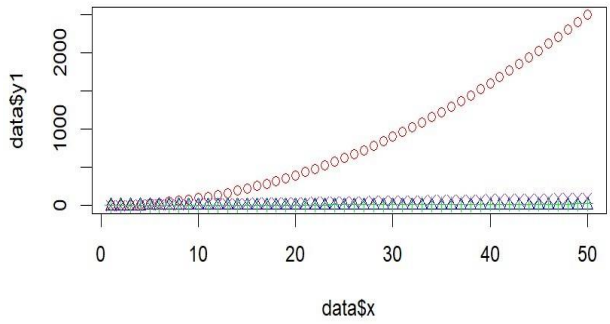


#add line in scatter plot

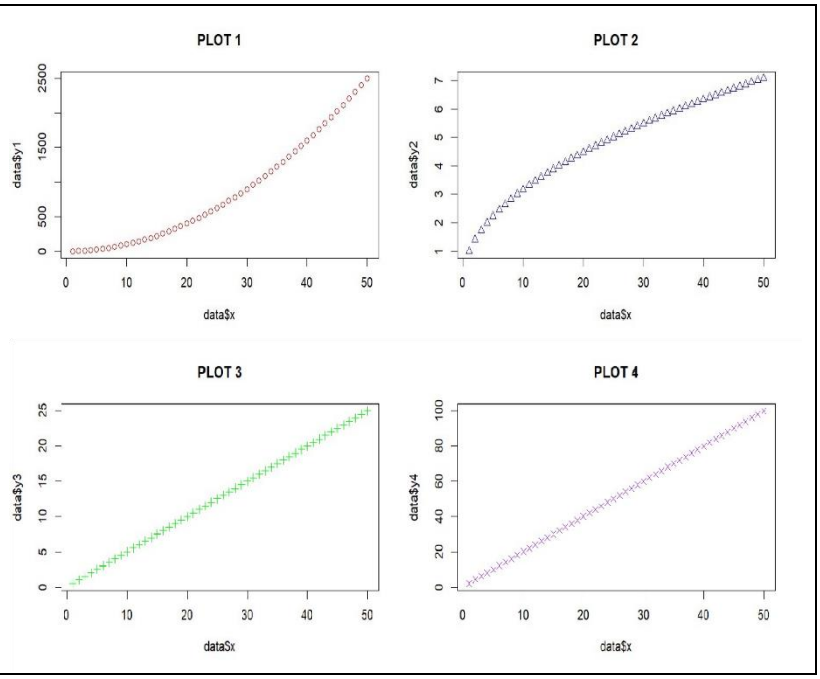
lines (x, y, col= "blue", lty=5)	

Create multi plot in one data frame

x <- 1:50 data <- data.frame(x, y1 = x^2, y2 = sqrt(x), y3 = x/2, y4 = 2*x) data ↲

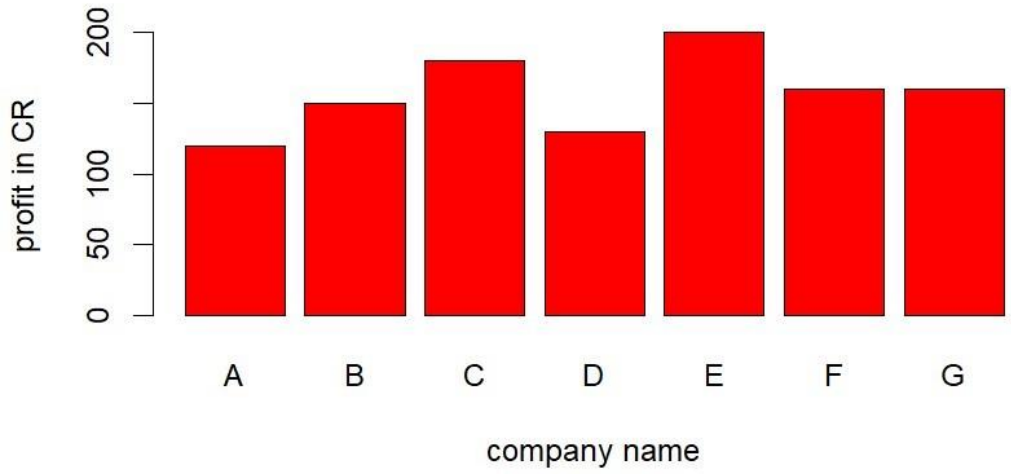
plot(data$x,data$y1, pch=1, col= "red") points(data$x,data$y2, pch= 2, col= "blue") points(data$x,data$y3, pch= 3, col= "green") points(data$x,data$y4, pch= 4, col= "purple")	

```
par (mfrow= c (2,2))

plot(data$x, data$y1, pch=1,
col= "red", main= "PLOT 1")

plot(data$x, data$y2, pch=2,
col= "blue", main= "PLOT 2")

plot(data$x, data$y3, pch=3,
col= "green", main= "PLOT
3")

plot(data$x, data$y4, pch=4,
col= "purple", main= "PLOT
4")
```



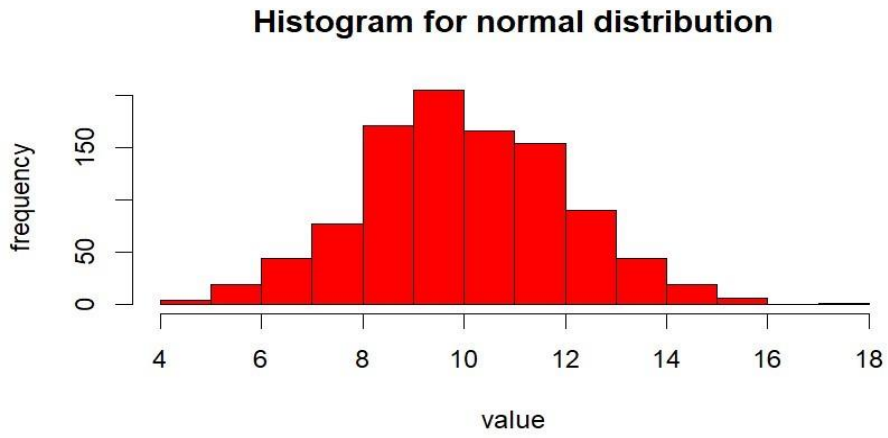
draw 'Bar plot'

```
barplot(data$profit_CR, names.arg = data$comp_name, col= "red", xlab= "company name", ylab= "profit in CR")
```



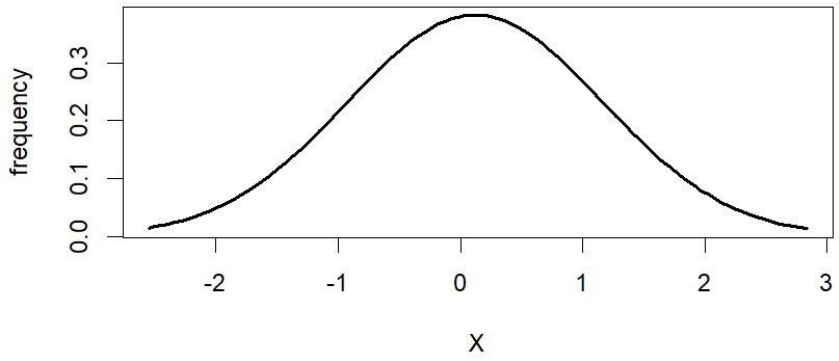
Histogram data <- rnorm(1000, mean=10, sd=2) # generate random data point

Create histogram hist(data, main= "Histogram for normal distribution", xlab = "value", ylab= "frequency", col= "red", breaks= 10)



Create Bell- curve

curve(dnorm(x, mean= mean(data), sd= sd(data)), col= "black", lwd= 2, from= min(data), to = max(data))



Various type of Distribution

Normal Distribution with Histogram:

Create random normal distribution data point

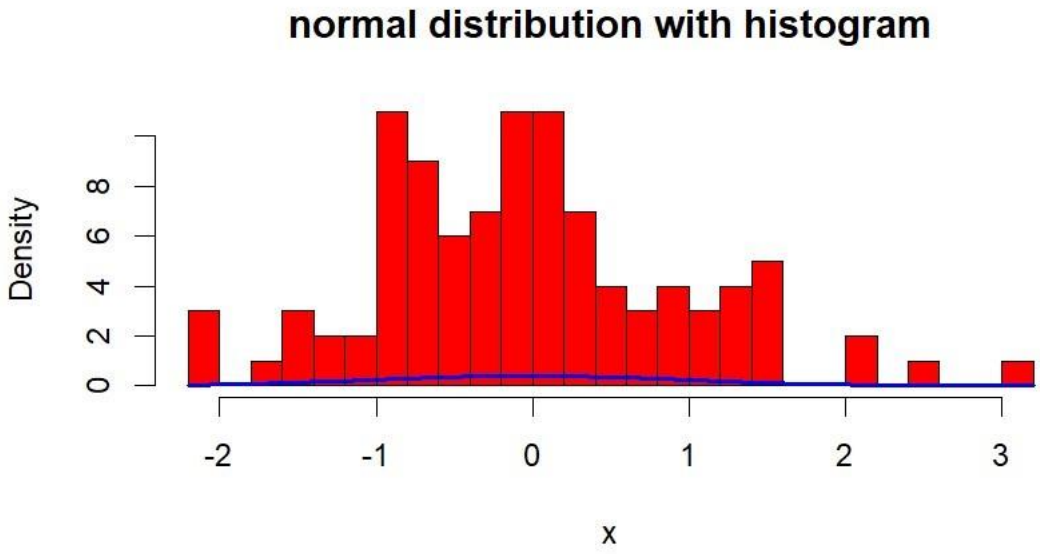
```
data <- rnorm (100, mean=0, sd=1)
```

Create Histogram

```
hist (data, breaks= 30, col= "red", main="normal distribution with histogram", xlab= "x", ylab= "Density")
```

Create bell curve over histogram

```
curve (dnorm (x, mean= mean (data), sd= sd (data)), col= "blue", lwd= 2, add= TRUE)
```



- Multiple Bell curve

```
# Generating Data Set  set.seed(123)
data1 <- rnorm(100, mean = 0, sd = 1)
data2 <- rnorm(100, mean = 0, sd = 2)
data3 <- rnorm(100, mean = 0, sd = 3)
# Compute density estimate dens1 <-
density(data1) dens2 <- density(data2)
dens3 <- density(data3) # Create an
empty plot
```

```
plot (dens1, main = "Multiple Bell Curves", xlab = "Value", ylab = "Density",
xlim = range(c(dens1$x, dens2$x, dens3$x)),        ylim = range(c(dens1$y,
dens2$y, dens3$y)),
      col = "blue", lwd = 2, type = "l")
# Add the other density curves lines
(dens2, col = "red", lwd = 2) lines
(dens3, col = "green", lwd = 2)
# Add a legend
legend ("topright", legend = c("Group 1", "Group 2", "Group 3"), col = c("blue", "red", "green"),
lwd = 2)
```