UNIT - 1

Date:
Page:

Chapter 1 → Introduction to Statistics.

Central Tendency

→ The statistical measure that identifies a single value as representative of entire distribution.
→ It aims to provide an accurate description of the entire data.
→ It is the single value that is most typical or the representative of the collected data.

The following are the 3Ms of central tendency.

i Mean
ii Median
iii Mode

Mean

→ It is also known as arithmetic mean.
→ It is the sum of all the observations divided by the total number of observations.
→ denoted by $\bar{x}$.

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

→ It represents the typical value in the dataset
→ It is sensitive to outliers, which can heavily influence the result
→ It is widely used in Statistics, Economics and Science.

Eg 1. Consider dataset :- 10, 15, 20, 25, 30. Calculate its mean

$$\text{Mean} = \frac{10+15+20+25+30}{5}$$

$$= \frac{100}{5} = 20.$$

Eg 2. Find mean from the frequency table.

x	f	fx
1	5	5
2	9	18
3	12	36
4	17	68
5	14	70
6	10	60
7	6	42
	73	299

$$\bar{x} = \frac{1}{N} \Sigma fx \quad \text{where } N = \Sigma f.$$

$$= \frac{1}{73} \times 299$$

$$= 4.09$$

→ For the frequency table, one need to calculate fx and then find the mean by using the formula

$$\bar{x} = \frac{1}{N} \Sigma fx \qquad N = \Sigma f.$$

Eg 3 Calculate mean by the given intervals.

Marks	f	x	fx
0 - 10	12	5	60
10 - 20	18	15	270
20 - 30	27	25	675
30 - 40	20	35	700
40 - 50	17	45	765
50 - 60	6	55	330
	100		2800

$$\bar{x} = \frac{1}{N} \Sigma fx$$

$$= \frac{2800}{100} = 28$$

→ In the interval range, first calculate the middle term in the interval as 'x' and then calculate fx to find the mean.

Eg 4. Find the value of x when the mean is 18.

Allowance	f	x	xf
11 - 13	7	12	84
13 - 15	6	14	84
15 - 17	9	16	144
17 - 19	13	18	234
19 - 21	x	20	20x
21 - 23	5	22	110
23 - 25	4	24	96
	44 + x		752 + 20x

$$\bar{x} = \frac{752 + 20x}{44 + x} = 18$$

$$792 + 18x = 752 + 20x$$
$$-2x = -40$$
$$x = 20$$

Median

→ It is measure of central tendency that represents the middle value of a dataset.
→ It is not affected by outliers
→ It is robust measure for skewed distribution
→ It is ordered from least to greatest.

Odd number's median

→ Arrange the data in ascending order
→ Median is the middle value.

Even number's median

→ Arrange the data in ascending order
→ Median is the average of two middle values

Eg 1. Consider dataset :- 3, 1, 7, 5, 9

1, 3, 5, 7, 9
Median = 5.

Eg 2 Consider dataset :- 2, 8, 4, 6

2, 4, 6, 8
$$\text{Median} = \frac{4+6}{2} = 5.$$

Date: Page:

Eg 3. Find the median from the frequency table

x	f	c.f
1	8	8
2	10	18
3	11	29
4	16	45
5	20	65
6	25	90
7	15	105
8	9	114
9	6	120
	120	

Now, calculate $\frac{N}{2} = \frac{120}{2} = 60$.

Search for c.f just greater than N/2 i.e. 65 which is of 5.

∴ Median is 5.

Eg 4. Calculate median for class interval

CI	f	c.f
10 - 20	5	5
20 - 30	8	13
30 - 40	12	25
40 - 50	10	35
50 - 60	7	42

$\frac{N}{2} = \frac{42}{2} = 21$.

∴ Median class will be 30 - 40.

$$\text{Median} = l + \frac{\left(\frac{n}{2} - cf\right)}{f} \times h.$$

l = lower limit of median class
n = no. of observation {total}
f = frequency of median class
cf = c.f of class preceding median class
h = class size / gap.

$$\therefore \text{Median} = 30 + \frac{\left(\frac{42}{2} - 13\right)}{12} \times 10$$

$$= 30 + \frac{8^2}{12_3} \times 10$$

$$= 30 + \frac{20}{3}$$

$$= 36.66$$

Mode

→ It represents the most frequently occuring value in a dataset.
→ It is not necessarily unique
→ A dataset may have more than one mode.
→ It is not affected by outliers
→ In a perfectly symmetrical distribution, the mean, median and mode are all equal.

Date: Page:

Types of Mode.

1. Unimodal – A dataset with one mode
2. Bimodal – A dataset with two modes
3. Multimodal – A dataset with more then two modes.
4. No mode – A dataset with no clear mode (all value occur with same frequency).

E1. Calculate mode of the dataset.

3, 5, 2, 7, 3, 8, 3, 2, 6, 5, 3

Mode = 3. {as it occurs 4 times}

E2 Find mode of 4, 4, 4, 9, 15, 15, 15, 27, 48.

Mode = 4, 15 Bimodal

E3. Find mode 3, 6, 9, 16, 27, 37, 48

No mode found

E4. In class of 30 students marks obtained by students in Maths out of 50 are given below

Marks	no. of students
10 – 20	5
20 – 30	12
30 – 40	8
40 – 50	5

Modal class is 20 – 30 as freq. is 12.

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

l :- lower limit of modal class
f_1 :- freq. of modal class
f_0 :- freq of preceding modal class
f_2 :- freq of succeeding modal class
h :- class interval / gap.

$$\text{Mode} = 20 + \left(\frac{12-5}{2\times12-5-8}\right)\times 10$$

$$= 20 + \left(\frac{7}{24-13}\right)\times 10$$

$$= 20 + \frac{7}{11}\times 10$$

$$= 20 + 6.364$$

$$= 26.364.$$

Empirical Relationship.
Mode, Median, Mean.

$$\boxed{\text{Mode} = 3\times\text{Median} - 2\times\text{mean}.}$$

E5 If ratio of mode to median is 2:3 then find ratio of mode to mean

Let mode be $2x$, median be $3x$ then;

$$2x = 3\times 3x - 2\times\text{Mean}$$

$$2\times\text{Mean} = 9x - 2x$$

$$2\times\text{Mean} = 7x$$

$$\text{Mean} = \frac{7x}{2}$$

$$\text{Mode} : \text{Mean} = \frac{2x}{\frac{7x}{2}} = \frac{4}{7} = 4:7$$

Date:
Page:

Measures of Dispersion

→ It quantify the extent to which data values deviate or spread out from central tendency
→ It is measure of central tendency that provides insights into the variability of a dataset.
→ It assess spread and consistency in data
→ Identify outliers and extreme values
→ Facilitate decision making by understanding data variability.

- Common measures of dispersions

i Range
ii Interquartile Range
iii Variance
iv Standard Deviation

i Range.

→ It is calculated by subtracting highest value from the lowest value.

Eg. 150, 160, 175, 190, 200. Calculate Range

Range = Highest value − Lowest value
= 200 − 150
= 50.

Date: Page:

Quartiles.

→ They are divided into 4 equal parts

→ 3 Quartiles :- Q1, Q2, Q3.

Five Number Summary

Every dataset can be described using these five numbers

i Lowest Value

ii Q1 — 25% ile

iii Q2 — Median

iv Q3 — 75% ile

v Highest Value.

Eg. IIIg Interquartile Range. (IQR)

→ It is defined as the range between 75 percentile (Q3) and 25 percentile (Q1)

$$IQR = Q3 - Q1$$

E1. Let there are 8 numbers between 10 and 90. which are equally distributed. Define five number summary and find IQR.

Lowest Value :- 10

Q1 :- 25

Q2 :- 50

Q3 :- 75

Highest Value :- 90.

$$IQR = Q3 - Q1$$
$$= 75 - 25$$
$$= 50.$$

Quartile Deviation

→ It measures the deviation of the data from the average value.

$$Q.D = \frac{Q3 - Q1}{2}$$

$$\text{Coefficient of } QD = \frac{(Q3 - Q1)}{(Q3 + Q1)}$$

E1. Find QD and coefficient of QD from the data
23, 8, 5, 16, 33, 7, 24, 5, 30, 33, 37, 30, 9
11, 26, 32

5, 5, 7, 8, 9, 11, 16, 23, 24, 26, 30, 30,
32, 33, 33, 37

$$Q1 = \frac{8+9}{2} = 8.5$$

$$Q3 = \frac{30+32}{2} = 31$$

$$QD = \frac{31 - 8.5}{2} = 11.25$$

$$\text{coefficiency of } QD = \frac{31 - 8.5}{31 + 8.5}$$

$$= \frac{22.5}{39.5}$$

$$= 0.57$$

Date: Page:

E2 Find QD of the marks scored by 50 students

CI	frequency	C.f.
45 - 50	7	7
50 - 55	5	12
55 - 60	12	24
60 - 65	11	35
65 - 70	9	44
70 - 75	6	50

Calculate $\frac{N}{4}$ and $\frac{3N}{4}$

$= 12.5$ $= 37.5$

Q1 class is 55 - 60
Q3 class is 65 - 70

$$Q_1 = l_1 + \frac{n(N/4) - c}{f}(l_2 - l_1)$$

n :- quartile
N :- Total frequency
f :- frequency of that class
c :- c.f of preceding class
l_1 :- lower bound of class
l_2 :- Upper bound of class

$$Q_1 = 55 + \frac{1 \cdot (50/4) - 12}{12}(60 - 55)$$

$$= 55.21$$

$$Q_3 = 65 + \frac{3 \cdot (50/4) - 35}{9}(70 - 65)$$

$$= 65.83$$

Quartile Deviation $= \frac{Q_3 - Q_1}{2}$

$= \frac{65.83 - 55.21}{2}$

$= 5.31$

iii Variance

→ Defined as average of squared difference from the mean.

→ measures how far each data point in datasets from the mean.

→ Denoted by σ^2

$$\sigma^2 = \frac{\Sigma(x_i - \mu)^2}{n}$$

$i = 1, 2, 3, 4, 5, \ldots\ldots, n$

μ = population mean

n = no. of datapoints

iv Standard Deviation

SD is the square root of variance

$$SD = \sqrt{\text{variance}}$$

$$= \sqrt{\sigma^2}$$

$$= \sqrt{\frac{\Sigma(x_i - \mu)^2}{n}}$$

n :- when there is population variance

$n-1$:- when there is sample variance.

Date:
Page:

E1. Let there be 5 students of height 1m, 2m, 3m, 4m, 5m. Calculate SD.

$$\text{Mean} = \frac{1+2+3+4+5}{5}$$

$$= 3.$$

$$\text{Variance} = \frac{(1-3)^2+(2-3)^2+(3-3)^2+(4-3)^2+(5-3)^2}{5-1}$$

$$= \frac{4+1+0+1+4}{4}$$

$$= \frac{10}{4}$$

$$= 2.5.$$

$$S.D = \sqrt{2.5}$$

$$= 1.58$$

Chapter 2 → Introduction to Probability.

Theory of Probability

→ It is a branch of mathematics which has been developed to deal with situations involving uncertainty.

→ It has begin in 16th centuary

→ Eg :- It may rain today
Train is likely to be late
I doubt that he will win the race.

→ Probability ranges b/w 0 and 1

Terminology Used in probability

1. Event – It is conclusion/outcome of an experiment
2. Sample Space – It denotes all possible outcomes or events
3. Mutually Exclusive Events – All possible events of an experiment that cannot occur simultaneously or together
4. Mutually Exhaustive Events – It represents all those possible events that can happen simultaneously
5. Independent Events – Don't rely on occurrence of the other events

Probability of an event 'E' is

$$P(E) = \frac{\text{no of outcomes favorable to } E}{\text{no. of all possible outcomes}}$$

Ex A coin is tossed once. Find probability of getting

(i) Head
$$P(H) = \frac{1}{2}$$

(ii) tail
$$P(T) = \frac{1}{2}$$

E2 A dice is thrown once. Probability of getting a number other than 3.

$$P(F) = \frac{5}{6}$$

E3 A dice is thrown once. Probability of getting number greater than seven

$$P(F) = \frac{0}{6} = 0$$

Mathematical Definition

It is defined as a number between 0 and 1, inclusive of 0 and 1. Probability of an event A is denoted by P(A).

Statistical Definition.

It is defined based on the relative frequency of occurrence of an event in repeated trials or observations.

Law of Addition

If A and B are any two events and are not disjoint then

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

When A and B are mutually exclusive events, then $P(A \cap B) = 0$.

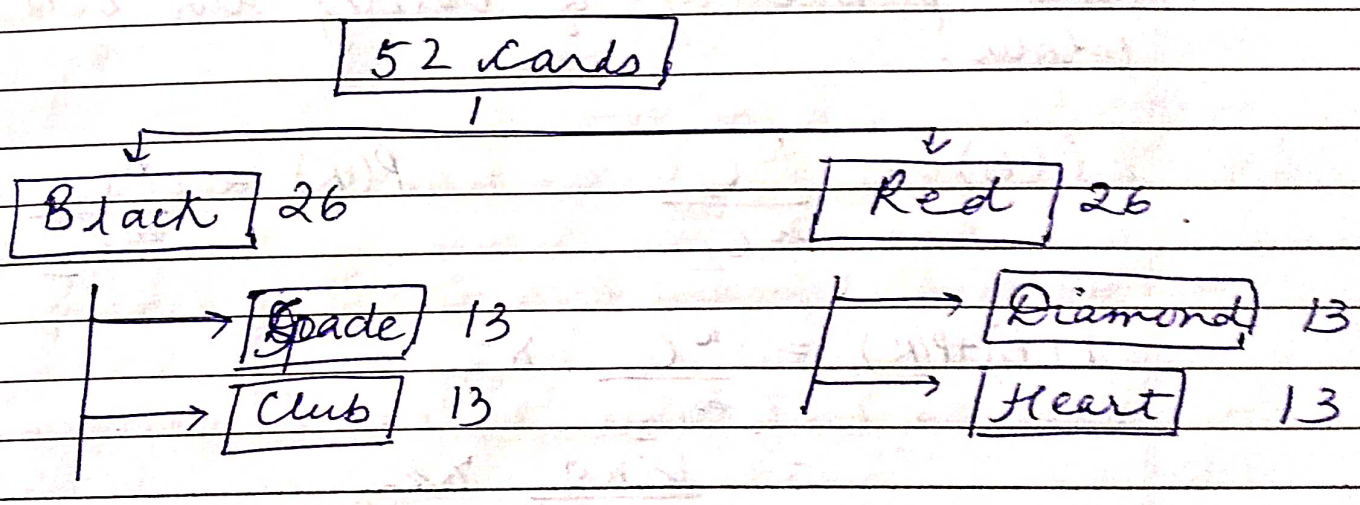
Law of Multiplication

It states that whenever an event is the intersection of two other events, that is event A and B need to occur simultaneously then

Independant $P(A \text{ and } B) = P(A) \cdot P(B)$

Depandant $P(A \text{ and } B) = P(A) \cdot P(B/A)$

52 deck of cards.



Ace, 2, 3, 4, 5, 6, 7, 8, 9, 10, Jack, Queen, King

Face Card – 3 {J, Q, K}

1 Ace

9 no. card – 2, 3, 4, 5, 6, 7, 8, 9, 10

E1. If you take out a single card from a regular pack of cards, what is probability that the card is either ace or spade.

$$P(\text{Ace or Spade}) = P(\text{Ace}) + P(\text{Spade}) - P(\text{Ace} \cap \text{Spade})$$
$$= \frac{4}{52} + \frac{13}{52} - \frac{1}{52}$$
$$= \frac{16}{52} = \frac{4}{13}$$

E2 A bag contains 10 black and 10 white balls. find probability of drawing two ball of same colour.

$$P(B) = \frac{{}^{10}C_2}{{}^{20}C_2} \qquad P(R) = \frac{{}^{10}C_2}{{}^{20}C_2}$$

$$P(B)+P(R) = \frac{{}^{10}C_2}{{}^{20}C_2} \times 2$$
$$= \frac{10 \times 9}{20 \times 19} \times 2$$
$$= \frac{9}{19}$$

E5.

Bag A	Bag B
4 white balls	3 white ball
2 black balls	3 black ball

A bag is selected and a ball is drawn random. $P(W)$ is ?

$P(\text{White ball drawn from A bag}) = \frac{1}{2} \times \frac{4}{6}$

$= \frac{1}{3}$

$P(\text{White ball drawn from B bag}) = \frac{1}{2} \times \frac{3}{6}$

$= \frac{1}{4}$

$\therefore P(\text{white ball}) = \frac{1}{3} + \frac{1}{4}$

$= \frac{7}{12}$

E 6

	Machine I	Machine II	Machine III
manufact	0.4	0.5	0.1
Defective	2%	4%	1%

An item is choosen at random. $P(\text{defective}) = ?$

$P(\text{Machine I}) = \frac{0.4 \times 2}{100} = \frac{0.8}{100}$

$P(\text{Machine II}) = \frac{0.5 \times 4}{100} = \frac{2}{100}$

$P(\text{Machine III}) = \frac{0.1 \times 1}{100} = \frac{0.1}{100}$

$P(\text{defective item}) = \frac{0.8}{100} + \frac{2}{100} + \frac{0.1}{100}$

$= \frac{2.9}{100}$

$= 0.029$

Date:
Page:

E7 Two cards are selected without replacing the first card from the deck. Find the probability of selecting a king and then a queen

$$P(\text{King}) = \frac{4}{52}$$

$$P(\text{Queen}) = \frac{4}{51}$$

$$P(\text{king} \cap \text{Queen}) = \frac{4}{52_{13}} \times \frac{4}{51}$$

$$= \frac{4}{663}$$

E8. The probability of machine A performing in 5 years time is 1/4 while of machine B is 1/3. Find the probability in

(i) both machines will perform usual function.

$$\frac{1}{4} \times \frac{1}{3} = \frac{1}{12}$$

(ii) neither will be operating

$$P(\text{A not operating}) = 1 - \frac{1}{4} = \frac{3}{4}$$

$$P(\text{B not operating}) = 1 - \frac{1}{3} = \frac{2}{3}$$

$$P(\text{Neither}) = \frac{3}{4} \times \frac{2}{3} = \frac{1}{2}$$

(iii) Only machine B will be operating

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

(iv) Atleast one of the machine will be operating

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

Conditional Probability

→ It measures the likelihood of an event occuring, given that event second has already occurred.

→ It is denoted by $P(A|B)$ read as "the probability of event A given event B".

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad ; P(B) > 0$$

E1 Two dice are thrown and the sum of no's obtained is found to be 7. What is the probability that the number 3 has appeared at least once?

Sample space = $6 \times 6 \Rightarrow 36$ events

A :- occurrence of 3 atleast once.

B :- sum of dice no is 7.

A :- {(3,1) (3,2) (3,3) (3,4) (3,5) (3,6) (1,3) (2,3) (4,3) (5,3) (6,3)}

B :- {(1,6) (2,5) (3,4) (4,3) (5,2) (6,1)}

$$P(A) = \frac{11}{36} \qquad P(B) = \frac{6}{36}$$

$$A \cap B = 2$$

$$P(A \cap B) = \frac{2}{36}$$

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

$$= \frac{\frac{2}{36}}{\frac{6}{36}} = \frac{2 \times 36}{36 \times 6} 3$$

$$\Rightarrow \frac{1}{3}$$

E2. In a group of 100 computer buyers, 40 bought CPU, 30 bought monitor and 20 purchases CPU and monitors. If a computer buyer chose at random and bought a CPU, what is probability they also bought a monitor.

$$P(CPU) = \frac{40}{100} \qquad P(monitor) = \frac{30}{100}$$

$$P(CPU \cap monitor) = \frac{20}{100}$$

$$P(Monitor \mid CPU) = \frac{\frac{20}{100}}{\frac{40}{100}}$$

$$= \frac{1}{2} = 50\%$$

Eg. Consider drawing 2 cards without replacement. Let event A be drawing a red card on the first draw, and event B be drawing a face card on the second draw. Probability of drawing face card on second draw given that first is red

$$P(A) = \frac{26}{52} = \frac{1}{2}$$

$$P(B|A) = \frac{12}{51}$$

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

$$\frac{12}{51} = \frac{P(A \cap B)}{1/2}$$

$$P(A \cap B) = \frac{12}{51} \times \frac{1}{2}$$

$$= \frac{6}{51}.$$

Bayes Theorem

→ It describes the probability of occurrence of an event related to any condition.

→ Also known as probability of "causes".

If $E_1, E_2, E_3, \ldots, E_n$ are mutually disjoint events with $P(E_i) \neq 0$, $(i = 1, 2, 3, \ldots n)$ then for arbitary event A which is a subset of E such that $P(A) > 0$ then

$$P(E_i|A) = \frac{P(E_i)\, P(A|E_i)}{\sum_{i=1}^{n} P(E_i)\, P(A|E_i)} \qquad i = 1, 2, \ldots n$$

E1

Bag I	Bag II
4 white balls	4 white balls
6 black balls	3 black balls

One ball is drawn at random from one of the bag and is found to be black. Find probability that it was drawn from Bag I.

E_1 :- Choosing bag I
E_2 :- Choosing bag II
A :- drawing a black ball

$$P(E_1) = P(E_2) = \frac{1}{2}$$

$$P(A/E_1) = \frac{6}{10} = \frac{3}{5}$$

$$P(A/E_2) = \frac{3}{7}$$

$$P(E_1/A) = \frac{\frac{1}{2} \times \frac{3}{5}}{\frac{1}{2} \times \frac{3}{5} + \frac{1}{2} \times \frac{3}{7}} \Rightarrow \frac{\frac{3}{10}}{\frac{3}{10} + \frac{3}{14}}$$

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

E2 A man is known to speak the truth 2 out of 3 times. He throws a die and reports that 4 is obtained. Find probability that no. obtained is actually a four.

A :- no. Obtained is four
E_1 :- four has occured
E_2 :- four has not occured.

$P(E_1) = \frac{1}{6}$ $\quad P(E_2) = \frac{5}{6}$

$P(A/E_1) = \frac{2}{3}$ $\quad P(A/E_2) = \frac{1}{3}$

$$P(E_1/A) = \frac{\frac{1}{6} \times \frac{2}{3}}{\frac{1}{6} \times \frac{2}{3} + \frac{5}{6} \times \frac{1}{3}} \Rightarrow \frac{\frac{1}{9}}{\frac{1}{9} + \frac{5}{18}}$$

$$= \frac{1 \times 18^2}{9 \times 7}$$

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

Chapter 3 → Introduction to R.

- R basic knowledge.

→ Open source programming language for statistical computing & data analysis.
→ It has command line interface
→ Developed by Ross Ihaka and Robert Gentleman.
→ It is a interpreted computer programming language
→ It supports branching, looping, modular programming via functions.
→ For efficiency boost, R is integrated with procedure written in C, C++, .Net, Python

Date:
Page:

Reasons to use R.

1. Free Installation
2. Hottest Trend
3. Independent Platform
4. Latest cutting edge technology
5. Has large community of users.
6. Supports Cross language Integration

Applications of R.

1. R is used for Data Science
2. Environmental Science
3. Data Visualization
4. Data Mining
5. Machine Learning
6. Data and Finance
7. Reproducible Research
8. Web Scraping

- Key Feature.

1. Open Source — It is freely available for anyone to use, modify and distribute.

2 Statistical Computing — It offers a rich set of statistical function and library that cover a wide range of techniques

3. Graphics and DV — Excellent tools for creating high quality graphics. The ggplot2 package is used for data visualization.

4. Data Manipulation — The library like dplyr, tidyr makes manipulation easy where user can efficiently clean, transform and reshape data.

5. Extensibility and Packages — Users can write their function and CRAN host thousands of user-contributed packages.

6. Reproducibility — Promotes reproducible research through the use of scripts and R Markdown. This ensures that analysis are easily shared, replicated and updated.

Design of R System

→ Primary R system is available from the CRAN (Comprehensive R Archive Network).
→ CRAN host many add-on package that can be used to extend functionality of R.
→ R system is divided into 2 parts
(i) Base (ii) Everything else.

Limitation of R.

1. Memory Management
2. Learning Curve
3. Performance
4. Error Handling
5. Deployment
6 Graphics Customization
7 Package Fragmentation.

Date:
Page: 8

R Studio

→ It is an integrated development environment (I.D.E) for R

→ It is available as open source and commerical software

→ It is for both desktop and server version

→ The R Studio has different .exe for Linux, Windows and macOs.

→ It allows enterprise ready professional software for data science to develop and share the work.

• Tabs in R Studio

1. Console – The place where R program is written and the results are generated by clicking the 'RUN' button.

2. Environment Tab – It shows the variable that are generated during course of programming in temporary workspace.

3. History Tab – It shows all the commands that are used till now from the start of usage of R Studio.

4. File Tab – It shows the files and directories that are available within the default workspace of R.

5. Plot Tab – It shows the graphs, charts

that are generated during programming.

6. Packages Tab – It shows the already installed package and also allows the user to install new packages.

7 Help Tab – It is most important from where the documentation tells about the built in functions of R.

8. Viewer Tab – It can be used to see the local web content that's generated using R

- R Syntax

→ A program in R is made up of 3 things
1 Variables
2 Comments
3 Keywords.

→ Variable are used to store data
→ Comments are used to improve code readability
→ Keywords are reserved words that holds a specific meaning to the compiler.

- Variables in R.

→ They are the name given to reserved memory location that can store any type of data
→ Assignment of variable is done by 3 types
(i) = (Simple Assignment)
(ii) ← (Leftward Assignment)
(iii) → (Rightward Assignment)

Date:
Page:

Code

```
name = "Yukti"
name1 ← "Gupta"
"24-A" → name2.

print (name)
print (name1)
print (name2)
```

Output

```
Yukti
Gupta
24-A
```

- Comments in R
  - → They are a way to improve code readability
  - → They are only meant for user and are ignored by interpreter.
  - → One line comment are available in R by using # before the sentence.
  - → Multi line are used by using the either single or a double quote within the sentence

```
# This is R Program
" I am doing GREAT "
' What is R Studio ? '
```

- Keywords in R
→ They are the words reserved by program
→ They have a special meaning to the compiler.
→ The name of the variable cannot be a keyword
→ Eg :- if, else, while, for, TRUE, FALSE

i Control Flow Keywords
if, else, repeat, while, function, for, in, next, break

ii Boolean Keywords
TRUE, FALSE

iii Undefined Keyword
NaN, NULL

iv Infinity Keyword
Inf.

- Data Types in R
→ Different forms of data that can be saved and manipulated are defined and categorized using data types in R.
→ They are used in computer programming to specify the kind of data that a variable can store
→ Each data type has its own set of regulation and restriction.

Basic data type

1. Numeric — set of real numbers
2. Integer — set of integers
3. Logical — boolean ie true / false
4. Complex — set of complex no. (i)
5. Character — set of words, symbol in " "
6. raw — binary data storage

Datatype of object
→ Use class() function
→ Syntax :- Class (object)
Eg → print (class (TRUE))
→ Logical

Type Verification
→ Use is datatype() function
→ Returns logical value ; true / false
→ Syntax :- is. data_type (object)
Eg → print (is. complex (2+4i))
→ TRUE

Type of the object
→ Use typeof() function
→ Syntax :- typeof (object)
Eg → print (typeof (2.41))
→ decimal

Date:
Page:

• Data structure in R
→ DS is a particular way of organizing data in a computer so that it can be used effectively
→ The idea is to reduce the space and time complexity of different task.
→ DS in R are tools for holding multiple values
→ Data structure in R are divided on the the dimensionality.
(i) 1D (ii) 2D (iii) 3D.
→ Moreover, on the data types appearing
(i) Homogeneous (ii) Heterogeneous.

Essential Data Structure
The most essential data structure in R are as follows :-
a Vectors
b Dataframes
c Matrices.

Detail Study of Each DS:

I VECTORS
→ It is an ordered collection of data type
→ Homogenous data structure
→ Vectors are of 1D structure
→ They are fundamental building blocks in R programming
→ They are extensively used in various operation and calculations.

Benefits

1. Efficient storage and manipulation of homogeneous data
2. Enables to perform operation on entire vector at once without explicit looping
3. Offers a compact and convenient representation for storing data.
4. It is easy to use and the fundamentals are straightforward.

II Data Frames.

→ They are 2D tabular data structure consisting of rows and columns.
→ They are heterogeneous data structure.
→ A data frame must have column names and every row should have a unique name.
→ Each column must have identical no. of items
→ Each item in a single column must have homogenous data type
→ Different columns may have different type.

Benefits

1. Provides a convenient way to organize and work with structured data.
2. They can integrate data of different types into a single structure
3. It also offers a wide range of functions and package for ~~multi~~ manipulating data frames, sorting, filtering data.

Date:
Page:

III Matrices

→ They are 2D arrays with rows and column
→ They are homogeneous data structure
→ They are useful for mathematical operations and linear algebraic computation.

Benefits

1. It supports wide range of mathematical operations including addition, substraction, multiplication, inversion
2. They are essential for performing linear algebraic computation such as eigenvalue analysis and matrix factorization.
3. Matrices provide efficient storage and computation for numerical data
4. Many statistical models and algorithms such as regression analysis rely on matrix operations for computation.

Code of every Data Structure

1. Vector

```
X = c (1, 3, 5, 7, 8)
print (X)
```

output :- 1 3 5 7 8

2. Dataframe

```
Name = c("Yukti", "Anushka", "Shreya")
Language = c("Java", "R", "Python")
Age = c(22, 19, 21)
df = data.frame(Name, Language, Age)
print(df)
```

output :-

	Name	Language	Age
1	Yukti	Java	22
2	Anushka	R	19
3	Shreya	Python	21

3. Matrix

```
A = matrix(c(1,2,3,4,5,6,7,8,9), nrow=3,
ncol=3, byrow=TRUE)
print(A)
```

output :-

	[,1]	[,2]	[,3]
[1,]	1	2	3
[2,]	4	5	6
[3,]	7	8	9

Application of R.

1. Finance
2. Banking
3. Healthcare
4. Social Media
5. Manufacturing

Date:
Page:

Real life use cases of R.

1. Facebook — It uses R to update status and its social network graph.
2. Ford Motor — Ford relies on Hadoop. It also relies on R for statistical analysis.
3. Google — It uses R to calculate ROI on advertsing campaigns and to predict activity
4. Microsoft — It uses R for Xbox matchmaking service
5. New York Times — R is used in the news cycle to crunch data and prepare graphics.
6. Mozilla — It is the foundation behind the Firefox web browser and uses R for visual
7. Twitter — R is a part of Twitter's Data Science toolbox for sophisticated modeling.
8. Foursquare — It uses R for famed recommendation engine