IT SKILLS AND DATA ANALYSIS - 1

DWAIPYAN SINGHA

ROLL NO. : 20232711

COURSE - B.VOC SOFTWARE DEVELOPMENT
SEMESTER - IV

EXPERIMENT-1

AIM

To Construct the Discrete frequency table.

EXPERIMENT

Following numbers give the marks of 64 student in class. Prepare frequency table.

THEORY

A discrete frequency table is a way of organizing and presenting data when dealing with discrete variables-values that can only take specific, distinct numbers (like the number of students in a class or the number of cars in a parking lot).

DATA

10, 15, 12, 12, 15, 20, 20, 25, 10, 30, 35, 40, 40, 35, 45, 25, 12, 20, 20, 12, 15, 10, 30, 20, 25, 30, 10, 12, 15, 20, 25, 18, 18, 45, 18, 18, 50, 50, 50, 45, 45, 45, 45, 12 , 20, 30, 35, 40, 50, 50, 12, 12, 12, 18, 18, 20, 20, 22, 22, 25, 30, 22, 22, 25, 30.

CALCULATION

Observations	Frequency
10	4
12	9
15	4
18	6
20	9
22	4
25	6
30	6
35	3
40	3
45	5
50	5
	64

EXPERIMENT-2

AIM

To Construct the frequency table.

EXPERIMENT

Following numbers give the weights of 55 students of a class. Prepare suitable frequency table.

THEORY

A frequency distribution table for continuous data is used when dealing with values that fall within a range rather than being distinct, separate numbers. Continuous data can take any value within a given interval (like height, weight, temperature, or time).

DATA

42	74	40	60	82	115	41	61	75	83	63
53	110	76	84	50	67	65	78	77	56	95
68	69	104	80	79	79	54	73	59	81	100
66	49	77	90	84	76	42	86	69	70	80
72	50	79	52	103	96	51	64	78	94	71

FORMULA

- =MIN(data range)
- =MAX(data range)
- Range = MAX – MIN
- Number of classes = $1 + 3.322 * \log_N 10$
- Class Interval = Range / number of classes
- Frequency = COUNTIFS(data range,”>=LowerLimit”,data range,”<UpperLimit”)

CALCULATION

Class Interval	Frequency
40-51	7
51-62	8
62-73	11
73-84	17
84-95	5
95-106	5
106-117	2

Minimum	**40**
Maximum	**115**
Range	**75**
No. of Classes	**6.78**
Intervals	**11.06**
Count data	**55**

RESULTS

Total number of observations = 55

Minimum observation =40

Maximum observation = 115

Range = 75

Number of classes = 6.78

Class interval = 11.06

EXPERIMENT-3

AIM

To Construct a Histogram.

EXPERIMENT

Draw histogram of following Data.

THEORY

A **histogram** is a type of bar graph that visually represents the frequency distribution of numerical data. It shows how data values are grouped into intervals (or bins) and how many observations fall into each interval.

DATA

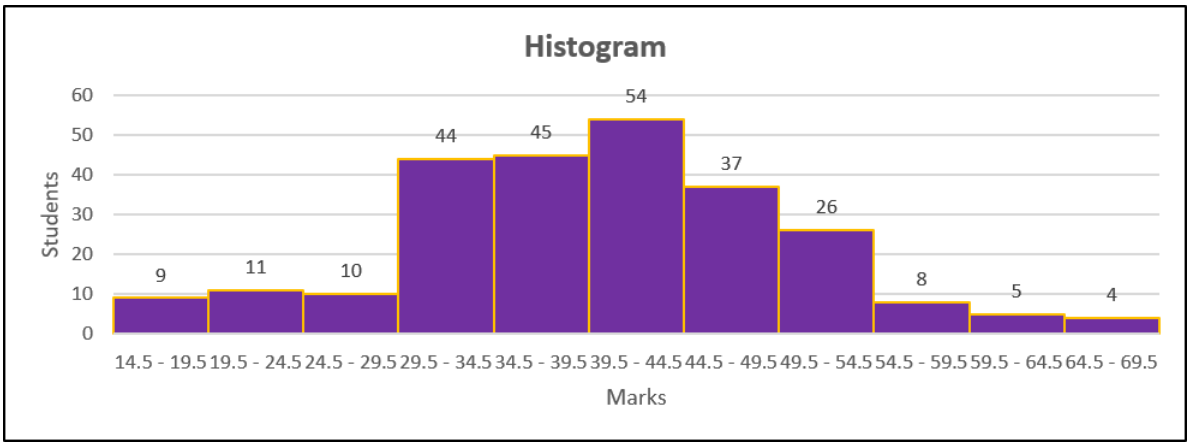
Marks (CI)	Students
14.5 - 19.5	9
19.5 - 24.5	11
24.5 - 29.5	10
29.5 - 34.5	44
34.5 - 39.5	45
39.5 - 44.5	54
44.5 - 49.5	37
49.5 - 54.5	26
54.5 - 59.5	8
59.5 - 64.5	5
64.5 - 69.5	4

STEPS

1. Enter data in excel sheet.
2. Select entire data and click on "insert tab".
3. Click on recommended charts and then clustered column chart.

4. A bar graph will appear.
5. Select graph and go to "chart design " tab and select the appropriate design like histogram.

OUTPUT



EXPERIMENT-4

AIM

To complete year wise total average and plot multiple bar and pie chart.

EXPERIMENT

Consider sales of woolen clothes in particular period of time, complete yearly total average and compute multiple bar and pie chart.

THEORY

A **histogram** is a type of bar graph that visually represents the frequency distribution of numerical data. It shows how data values are grouped into intervals (or bins) and how many observations fall into each interval.

DATA

Seasons	Years				
	1998	1999	2000	2001	2002
Summer	30	33	56	42	67
Monsoon	81	4	173	153	201
Winter	119	171	235	221	302
Autumn	62	86	129	99	136

FORMULA

TOTAL = SUM(data range)

AVERAGE = AVERAGE(data range)

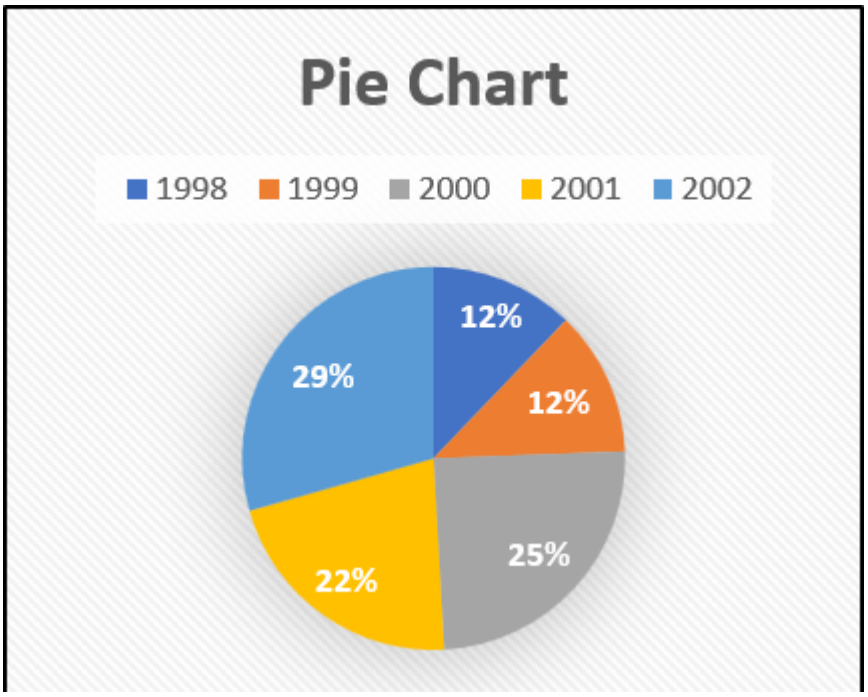
STEPS

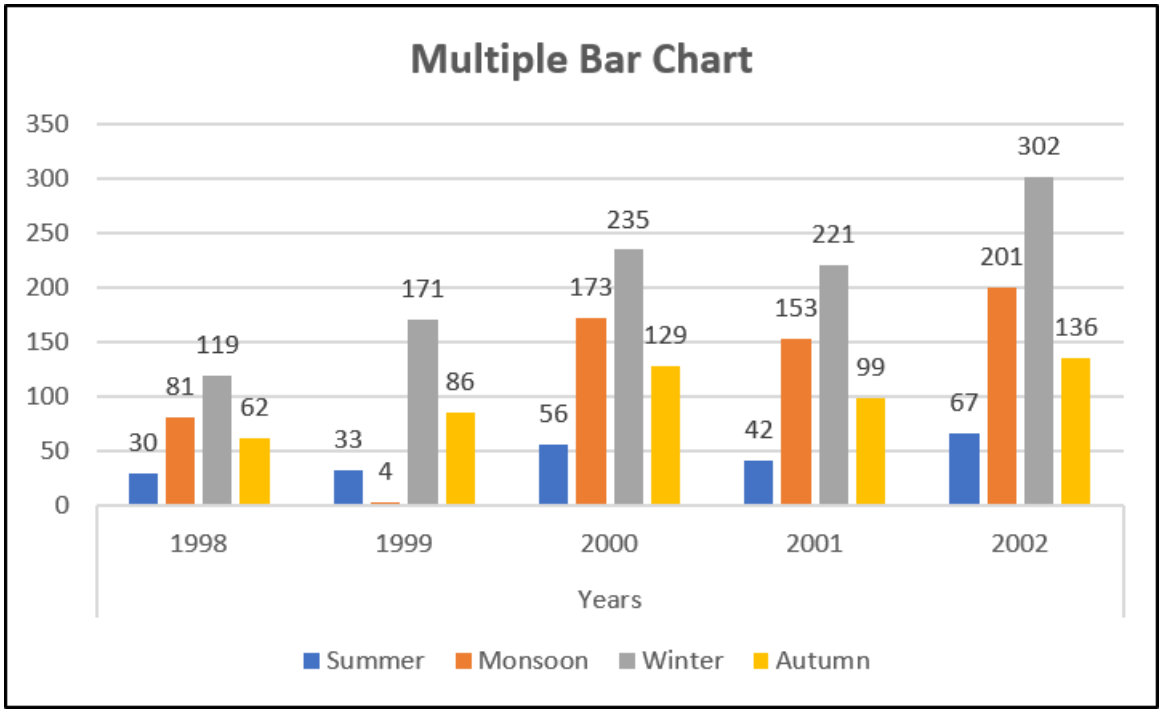
To prepare multiple bar and pie chart :-

- **Enter Data** – Organize your data in a table format.
- **Select Data** – Highlight the relevant cells for visualization.
- **Insert Charts** – Go to the **Insert** tab, choose **Bar Chart** and **Pie Chart**.
- **Customize** – Adjust titles, labels, colors, and layouts using the **Chart Design** tab.
- **Position Charts** – Resize and arrange them neatly in your worksheet.
- **Save s Share** – Export or share your charts for presentations.

CALCULATIONS

Seasons	Years				
	1998	1999	2000	2001	2002
Summer	30	33	56	42	67
Monsoon	81	4	173	153	201
Winter	119	171	235	221	302
Autumn	62	86	129	99	136
Total	**292**	**294**	**593**	**515**	**706**
Average	**73**	**73.5**	**148.25**	**128.75**	**176.5**




Multiple Bar Chart
350
300
250
200
150
100
50
0
30
81
119
62
33
4
171
86
56
173
235
129
42
153
221
99
67
201
302
136
1998
1999
2000
2001
2002
Years
Summer
Monsoon
Winter
Autumn

EXPERIMENT-5

AIM

To Calculate arithmetic mean and geometric mean.

EXPERIMENT

For the following frequency distribution calculate AM and GM.

THEORY

AM : It's calculated by summing all the values in a dataset and then dividing by the total number of values.

GM: The geometric mean is particularly useful when dealing with data that involves multiplicative relationships, growth rates, or ratios. It tends to dampen the effect of outliers compared to the arithmetic mean.

DATA

CI	15-19	19-23	23-27	27-31	31-35	35-39	39-43	43-47	47-51	51-55	55-59
fi	9	11	10	44	45	54	37	26	8	5	1

FORMULA

Xi = avg(upper limit + lowerlimit)

$$AM = \frac{\sum_{i=1}^{n}(f_i \cdot x_i)}{\sum_{i=1}^{n} f_i}$$

$$GM = \text{antilog}\left(\frac{\sum_{i=1}^{n}(f_i \cdot \log(x_i))}{\sum_{i=1}^{n} f_i}\right)$$

CALCULATIONS

AIM-5: To calculate AM and GM for the following frequency distribution table:-

Range		Class Interval	Frequency (Fi)	Xi	Fi*Xi	log(Xi)	Fi*log10(Xi)
15	19	15-19	9	17	153	1.2	11.1
19	23	19-23	11	21	231	1.3	14.5
23	27	23-27	10	25	250	1.4	14.0
27	31	27-31	44	29	1276	1.5	64.3
31	35	31-35	45	33	1485	1.5	68.3
35	39	35-39	54	37	1998	1.6	84.7
39	43	39-43	37	41	1517	1.6	59.7
43	47	43-47	26	45	1170	1.7	43.0
47	51	47-51	8	49	392	1.7	13.5
51	55	51-55	5	53	265	1.7	8.6
55	59	55-59	1	57	57	1.8	1.8
			250		**8794**		**383.5**

Arithmetic Mean → **35.2** —— **Fi*Xi/Fi**

Geometric Mean → **34.2** —— **Fi*log10(Xi)/Fi**

RESULTS

Arithmetic mean is 35.28

Geometric mean is 34.20

EXPERIMENT-6

AIM

To obtain mode using grouping function.

EXPERIMENT

For the following discrete frequency distribution calculate mode.

THEORY

Mode is the value from a data set that has occurred the most number of times.

It is the **method of calculating mode in cases where more than one value has the highest frequency and it becomes impossible to find the mode.**

DATA

x	1	2	3	4	5	6	7	8	9	10	11	12
f	3	8	15	23	35	40	32	28	20	45	14	6

CALCULATION

Data (X)	Frequency	Column C	Column D	Column E	Column F	Column G	Column H
1	3	3	11		26		
2	8	8		23		46	
3	15	15	38				73
4	23	23		58			
5	35	35	**75**		**98**		
6	40	40		**72**		**107**	
7	32	32	60				**100**
8	28	28		48			
9	20	20	65		80		79
10	45	**45**		59		93	
11	14	14	20				
12	6	6			65		
Highest Frequency		45	75	72	98	107	100
Repeated Values		10	5,6	6,7	4,5,6	5,6,7	6,7,8
Mode		**6**					

STEPS:

- **Original Frequency Check (Column I):**
  Find the x (class or value) with the highest frequency → **x = 10** (f = 45).
- **Group Frequencies (Columns II to VI):**
  Add adjacent frequencies (2 at a time, 3 at a time, etc.) to form new groups.
- **Find Maximum in Each Group**:
  Check which x appears where the grouped sum is highest.
- **Frequency Counting**:
  Count how many times each x appears among the maximums in Columns II–VI.
- **Identify Modal Value**:
  The x that appears most often in max groups **and** has the highest original frequency is the **mode**.

RESULT

MODE is 6

EXPERIMENT-7

AIM

To obtain mode for continuous data.

EXPERIMENT

For the following continous frequency distribution calculate mode.

THEORY

Mode is the value from a data set that has occurred the most number of times.

DATA

ci	f
0-10	5
10-20	8
20-30	7
30-40	12
40-50	28
50-60	20
60-70	10
70-80	10

FORMULA

$$M_o = l_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

M_o = Mode

l_1 = Lower limit of modal class

f_1 = Frequency of modal class

f_0 = Frequency of class preceding the modal class

f_2 = Frequency of the class succeeding the modal class

i = Class interval of the modal class

CALCULATION

Data		Class Interval	Frequency		
0	10	0-10	5		
10	20	10-20	8	h	10
20	30	20-30	7	l	40
30	40	30-40	12	f_1	12
40	50	40-50	28	f_0	28
50	60	50-60	20	f_2	20
60	70	60-70	10		
70	80	70-80	10		

Mode	**46.7**

By looking at the data, it is evident that the modal class is 40-50 because the frequency of this class is the maximum; i.e., 28.

l=40 f0=12 f=28 f2=20 i=10

=40+((28-12)/(2*28-12-20))*10

=46.67

RESULT

Mode of following continuous frequency distribution is 46.67

EXPERIMENT-8

AIM

To obtain median, P70, Q1 for discrete data.

EXPERIMENT

For the following frequency distribution calculate median, percentile 70, quartile 1.

THEORY

1. The median is the middle value in a sorted data set.

2. A percentile indicates the relative standing of a value in a dataset. The **Nth percentile** means that **N% of the data falls below this value**. For example:

- The 90th percentile means the value is **higher than G0% of the data**.

3. Quartiles divide data into **four equal parts**:

1. **Q1 (First Quartile, 25th percentile)** – The value below which 25% of the data falls.
2. **Q2 (Second Quartile, 50th percentile)** – The **median** of the dataset.
3. **Q3 (Third Quartile, 75th percentile)** – The value below which 75% of the data falls.
4. **Q4 (Fourth Quartile)** – Represents the highest values in the dataset.

DATA

x	f	cf
2	5	5
4	6	11
6	7	18
8	8	26
10	5	31
12	4	35
13	3	38
14	2	40

FORMULA

N = total number of frequencies.

Qr = rN/4

M=N/2

Pr = rN/100

CALCULATION

Data (X)	Frequency	CF			
2	5	5	Median	8	4
4	6	11			
6	7	18	Q_1	4	2
8	8	26			
10	5	31	P_{70}	12	6
12	4	35			
13	3	38	Count	8	
14	2	40			
40					

N = 40

Q1 = 40/4 = 10

Median = 40/2 = 20

P70 = 70*40/100 = 28

RESULTS

First Quartile is 4

Median is 8

70th percentile is 10

EXPERIMENT-9

AIM

To obtain median and quartile deviation.

EXPERIMENT

Find median and quartile deviation for the following data.

THEORY

1. The median is the middle value in a sorted data set.

2. Quartile deviation is a measure of **spread or dispersion** in a dataset. It helps to understand how data is distributed around the middle values.

DATA

wages	workers	CF
2k-3k	3	3
3k-4k	5	8
4k-5k	20	28
5k-6k	10	38
6k-7k	5	43

FORMULA

$$\text{Median} = L + \left(\frac{\frac{N}{2} - CF}{f} \right) \times h$$

Where:

- L = lower boundary of median class
- N = total frequency
- CF = cumulative frequency before the median class
- f = frequency of the median class
- h = class width (upper limit - lower limit).

$$Q_r = L + \left(\frac{rN}{4} - CF \right) \times \frac{h}{f}$$

- r = rth quartile

quartile deviation

$$Q.D. = \frac{Q_3 - Q_1}{2}$$

CALCULATION

Wages	Workers	Cf
2k-3k	3	3
3k-4k	5	8
4k-5k	20	28
5k-6k	10	38
6k-7k	5	43
43		

Median	**4675**
Quartile 1	**4137.5**
Quartile 3	**5425**
Quartile Dev.	**643.75**

N (total workers)/2	21.5
Lower Limit	4000
Class width	1000
N (total workers)/4	10.75
N (total workers)*3/4	32.25
Lower Limit	5000

N/2 = 43/2 = 21.5

Median class = 4k-5k

L = 4000

F = 20

CF = 8

H = 1000

Median = 4000 + (21.5-8 / 20) * 1000 = 4675

Q1

N/4 = 10.75

Q1 class is 4k-5k

L = 4000

F = 20

CF = 8

H = 1000

Q1 = 4000 + (10.75-8 /20) * 1000 = 4137.5

Q3

3N/4 = 32.25

Q3 class is 5k-6k

L=5000

f=10

CF=28

h=1000

Q3 = 5000 + (32.25-28 /10) * 1000 = 5425

QD

= 5425-4137.5/2 = 643.75

RESULTS

Median is 4675

Quartile deviation is 643.75

EXPERIMENT-10

AIM

To obtain Arithmetic mean, geometric mean and harmonic mean.

EXPERIMENT

For the following data calculate AM, GM, HM and verify relationship AM>GM>HM.

THEORY

AM : It's calculated by summing all the values in a dataset and then dividing by the total number of values.

GM: The geometric mean is particularly useful when dealing with data that involves multiplicative relationships, growth rates, or ratios. It tends to dampen the effect of outliers compared to the arithmetic mean.

HM: The **harmonic mean** is a type of average that's best used **when dealing with rates** (like speed, efficiency, ratios). It's especially useful when **all values contribute to a whole** in **reciprocal** terms.

DATA

Class interval	frequency
29-31	14
31-33	7
33-35	19
35-37	11
37-39	13
39-41	6

FORMULA

Xi = avg(upper limit + lowerlimit)

$$AM = \frac{\sum_{i=1}^{n}(f_i \cdot x_i)}{\sum_{i=1}^{n} f_i}$$

$$GM = \text{antilog}\left(\frac{\sum_{i=1}^{n}(f_i \cdot \log(x_i))}{\sum_{i=1}^{n} f_i}\right)$$

$$H.M = \sum fi / (\sum fi/xi)$$

Range		Class Interval	Frequency (Fi)	Xi	Fi*Xi	log(Xi)	Fi*log10(Xi)	Fi/Xi
15	19	15-19	9	17	153	1.2	11.1	0.5
19	23	19-23	11	21	231	1.3	14.5	0.5
23	27	23-27	10	25	250	1.4	14.0	0.4
27	31	27-31	44	29	1276	1.5	64.3	1.5
31	35	31-35	45	33	1485	1.5	68.3	1.4
35	39	35-39	54	37	1998	1.6	84.7	1.5
39	43	39-43	37	41	1517	1.6	59.7	0.9
43	47	43-47	26	45	1170	1.7	43.0	0.6
47	51	47-51	8	49	392	1.7	13.5	0.2
51	55	51-55	5	53	265	1.7	8.6	0.1
55	59	55-59	1	57	57	1.8	1.8	0.0
			250		**8794**		**383.5**	**7.5**

Arithmetic Mean	**35.2**	**Fi*Xi/Fi**
Geometric Mean	**34.2**	**Fi*log10(Xi)/Fi**
Harmonic Mean	**33.1**	**Fi/(Fi/Xi)**
Verify - AM>GM>HM →	**True**	

RESULT

A.M=35.2, G.M= 34.2, H.M = 33.1

Hence verified that A.M>G.M>H.M.