Om Sakthi
ADHIPARASAKTHI COLLEGE OF ARTS AND SCIENCE
(Autonomous)
G.B. NAGAR, KALAVAI – 632 506. RANIPET DISTRICT, TAMILNADU
Permanently Affiliated to Thiruvalluvar University,
Re-accredited by NAAC – 'B' Grade (CGPA - 2.83)

STATISTICAL METHODS AND THEIR APPLICATIONS-I

II B.SC Computer Science

Handled by

Dr.P.RAJAKUMARI

ASSISTANT PROFESSOR

DEPARTMENT OF MATHEMATICS

ADHIPARASAKTHI COLLEGE OF ARTS AND SCIENCE

AUTONOMOUS (AUTONOMOUS)

SEMESTER – III (ALLIED – II)

STATISTICAL METHODS AND THEIR APPLICATIONS – I

SUBJECT CODE – U3CS22AT

Objective:

To understand and computing statistical aspects.

UNIT– I (11 Hrs)

Introduction – Scope and limitations of statistical methods– Classification of data–Tabulation of data – Diagrammatic and graphical representation of data – Diagrams: Line,Bar and Pie diagrams– Graphs: Histogram, Frequency Polygon, Frequency, Ogive and Laurents Curves.

UNIT – II (10 Hrs)

Measures of location: Arithmetic Mean, Median, Mode, Geometric mean and Harmonic mean and their properties.

UNIT – III (10 Hrs)

Measures of dispersion: Range, Quartile deviation, Mean deviation, Standard deviation, Combined Standard deviation, Co-efficient of variation.

UNIT – IV (11 Hrs)

Measures of Skewness : Karl Pearson’s, Bowley’s, Kelly’s and coefficient of skewness and Kurtosis based on moments.

UNIT – V (10 Hrs)

Correlation – Karl Pearson – Spearman’s rank correlation – Concurrent deviation methods. Regression Analysis : Simple Regression Equations.

OUT COMES:

1. To represent a statistical data using Diagrams & Graphs.

2. To find Arithmetic Mean, Median, Mode, Co-efficient of variation, Correlation,

Regression.

Text Book :

1. Dr. S.P. Gupta-Sultan chand & sons,Statistical Methods .

Reference Books :

1. S.C. Gupta and V.K. Kapoor- Sultan Chand, Fundamental of Mathematical statistics-

2. Snedecor G. W. & Cochran W.G. Oxford ,Statistical Methods .

3. Mode – E.B. – Prentice Hall, Elements of Statistics.

1.Diagrammatic and graphical representation

Definition of statistical:

A branch of mathematic dealing with the collection, analysis interpretation and presentation of masses of numerical data.

A collection of quantitative data.

The statistical data collected for analysis in some cases may be exact

Example

1.The number of train accident in a year

2.The number of students passing IIT entrance examination in a year.

In some cases it may not be possible to give exact statistical data.

Example

The number of persons who witnessed the test match between India and Pakistan at Madras.

Statistical Methods:

Scope:

- Classification and tabulation of row data for the purpose of easy interpretation and analysis.
- Various measure of averages for simplifying and condensing the numerical data.
- Various measures of dispersions to study the spreading of the observations and deviations from an average.
- Coefficient of dispersion for comparing two different sets of numerical data.
- Coefficient of correlation to study the degree of relationship between variables.
- Regression analysis gives the relationship between the variables which is also used for prediction of the value of one variable when the other is known.
- Various measures of time series for forecasting and for studying seasonal trend.

Limitations of statistics:

Statistics can be used only to study numerically valued data, quantitative phenomena like honesty, intelligence, poverty, etc

Statistics deals only with aggregate and not with individuals. Individual items taken separately play importance in statistics. The Collection of items contribute statistics.

Statistical data collected for a given purpose cannot be applied to any situation.

For example: The data collected to study the economic conditions of middle-class people an give all information about expenditure on food, entertainments, health care, etc.,

Classification:

Definition:

“The process of arranging or bringing together all the enumerated individuals or items under separate heads or classes according to some common characteristics possessed by them.”

1.Classification according to qualitative basis

2. Classification according to quantitative basis

3. Classification according to Chronological basis

4. Classification according to Geographical basis

1.Qualitative Classification

If the statistical data are collected numerical facts about the qualities like male, female, employed, Indian, foreigner., etc. The classification of data is done according to these characters.

Example

Country	Bombay		Calcutta		Delhi		Madras	
	M	F	M	F	M	F	M	
USA								
Canada								
Australia								
Total								

2.Quantitative Classification

The arrangement of statistical data according to numerical measurements such as age, height, weight, number of members in a family come under quantitative classification.

Example

Age	20-30	31-50	51-60	61-70
Total	10	15	20	30

3.Chronlogical Classification

Statistical data arranged according to the time of occurrence come under this classification.

Example: (i) Production of wheat from the year 1980-85

(ii) Sales of imported cars after independence, are classified chronologically.

Example: Population of India

Year	1940	1950	1960	1970	2000
Population (in million)					

4.Geographical classification

Statistical data classified according to different areas like states, districts, villages, etc., come under the category of Geographical classification.

For Example: The production of fertilizer from different parts of the country comes under this classification.

State	Production in million tons
Bombay	
Calcutta	
Punjab	

Tabulation

Tabulation is defined as “The orderly or systematic presentation of numerical data in rows and columns designed to facilitate the comparison between the figures.

Example:

The following are weight in kg of 50 college students. Construct the frequency table

42	42	46	54	41	37	54	44	38	45
47	50	58	49	51	42	46	37	42	39
54	39	51	58	47	51	43	48	49	48
49	39	41	40	58	49	49	59	57	52
56	41	45	52	46	40	51	51	41	41

Solution

Class	Tally Mark	Frequency
36-40	|||| |||	8
41-45	|||| |||| || |	13
46-50	|||| |||| || |	13
51-55	|||| ||||	10
56-60	|||| |	6
	Total	50

Diagrammatic Representation

1.Bar diagram (One dimensional)

2.Area diagram (Two dimensional)

3.Volume diagram (Three dimensional)

4. Histogram, frequency polygon and frequency curve

5.Cummulative frequency curve (or) Ogive

Bar diagram (One dimensional)

Bar diagram is a popular form of diagrammatic representation. This diagram consists of a series of rectangular bars standing on a common base. Bar diagrams are of two types:

(i) Vertical bar diagrams

(ii) Horizontal bar diagrams

The bar diagrams can be classified as

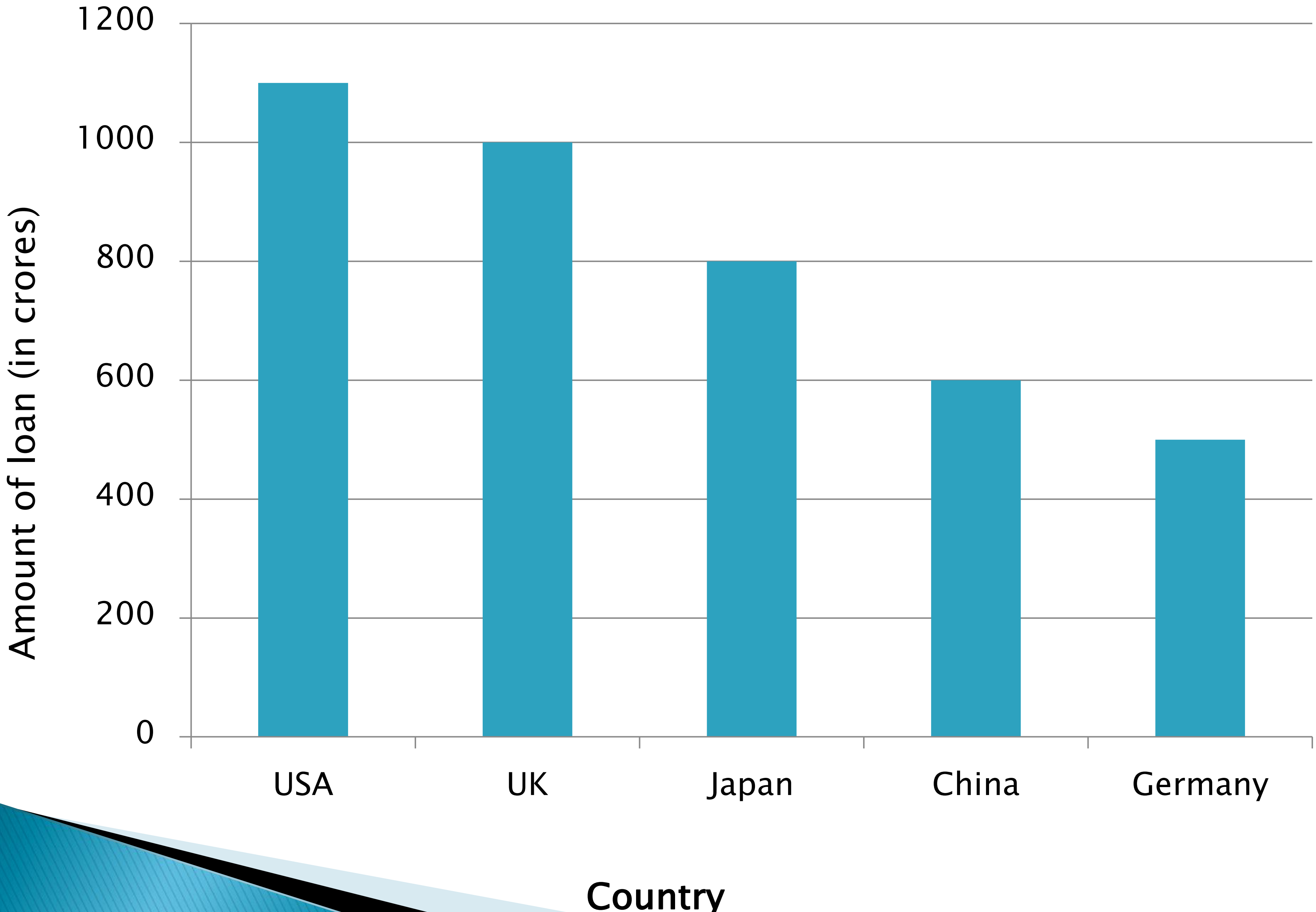
1. Simple Bar diagram
2. Multiple Bar diagram
3. Component or sub-divided Bar diagram
4. Percentage Bar diagram

Example :1

Prepare a bar diagram for the following data

Source of Borrow	Amount of loan (in crores)
USA	1100
UK	1000
Japan	800
China	600
Germany	500

Solution:



1200
1000
800
600
400
200
0
Amount of loan (in crores)
USA
UK
Japan
China
Germany
Country

1.Simple bar diagram

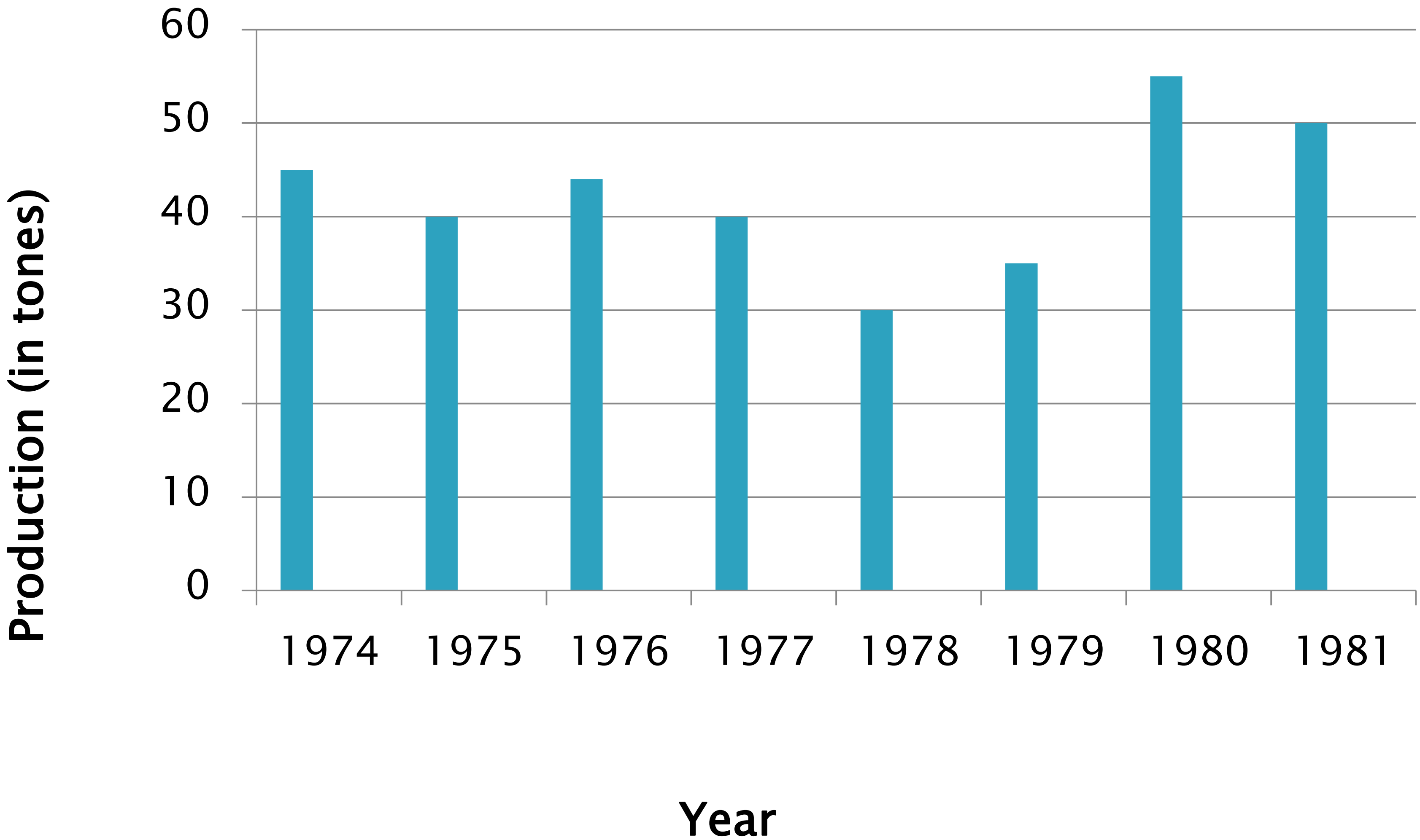
A simple bar diagram represents the magnitude of single variable like sales, production, profits, etc

Example

Represent the following data by a simple diagram

Year	Production (in tones)
1974	45
1975	40
1976	44
1977	40
1978	30
1979	35
1980	55
1981	50

Solution:


60
50
40
30
20
10
0
Production (in tones)
1974
1975
1976
1977
1978
1979
1980
1981
Year

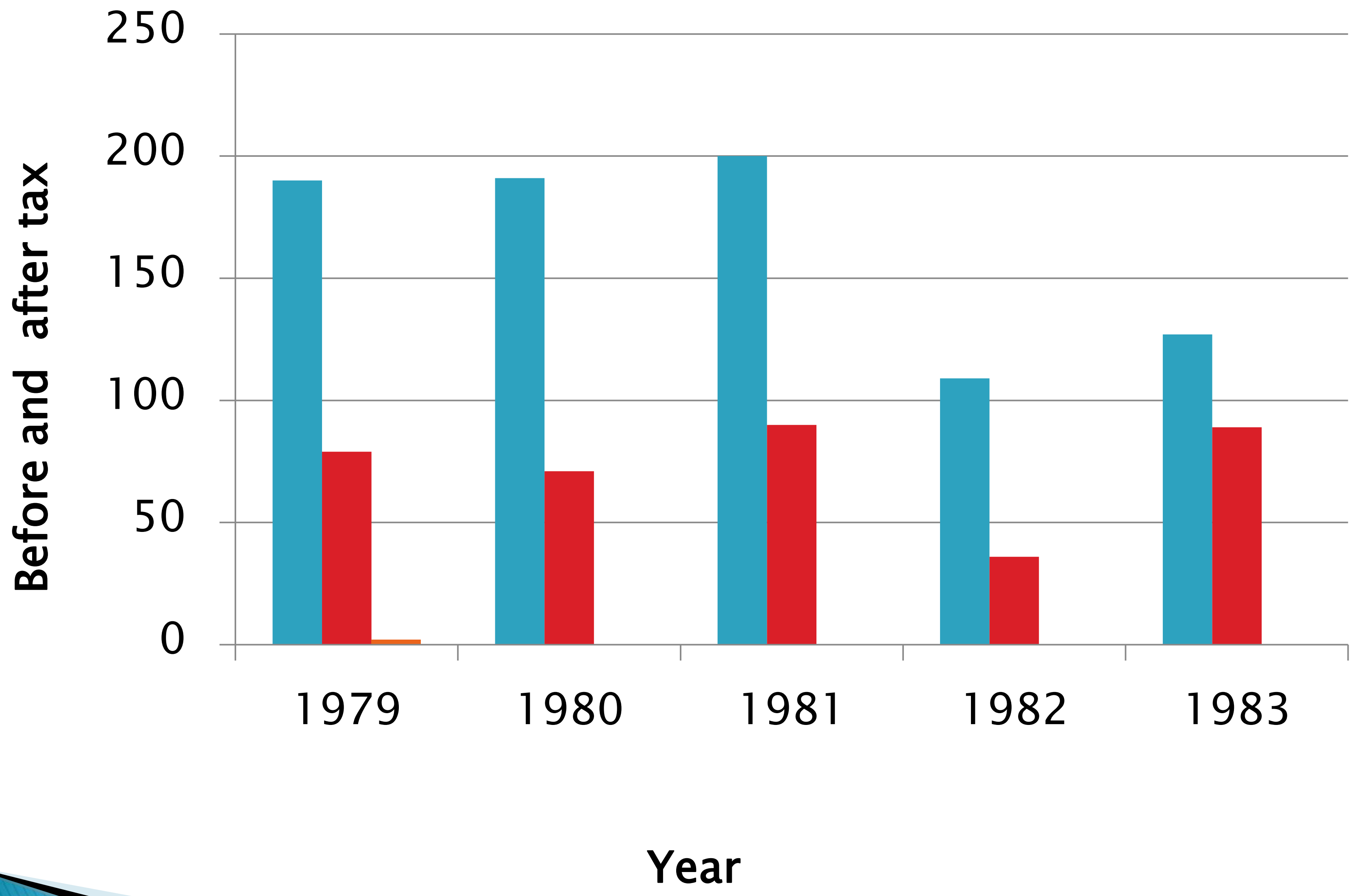
Multiple Bar Diagram

Multiple bar diagram contains two or more bars drawn side by side. It is also called component bar diagram

Example: Draw the Multiple bar diagram for the following data

Year	Before tax	After tax
1979	190	79
1980	191	71
1981	200	90
1982	109	36
1983	127	89

Solution:



250
200
150
100
50
0
Before and after tax
1979
1980
1981
1982
1983
Year

Component (or) Sub-divided bar diagram

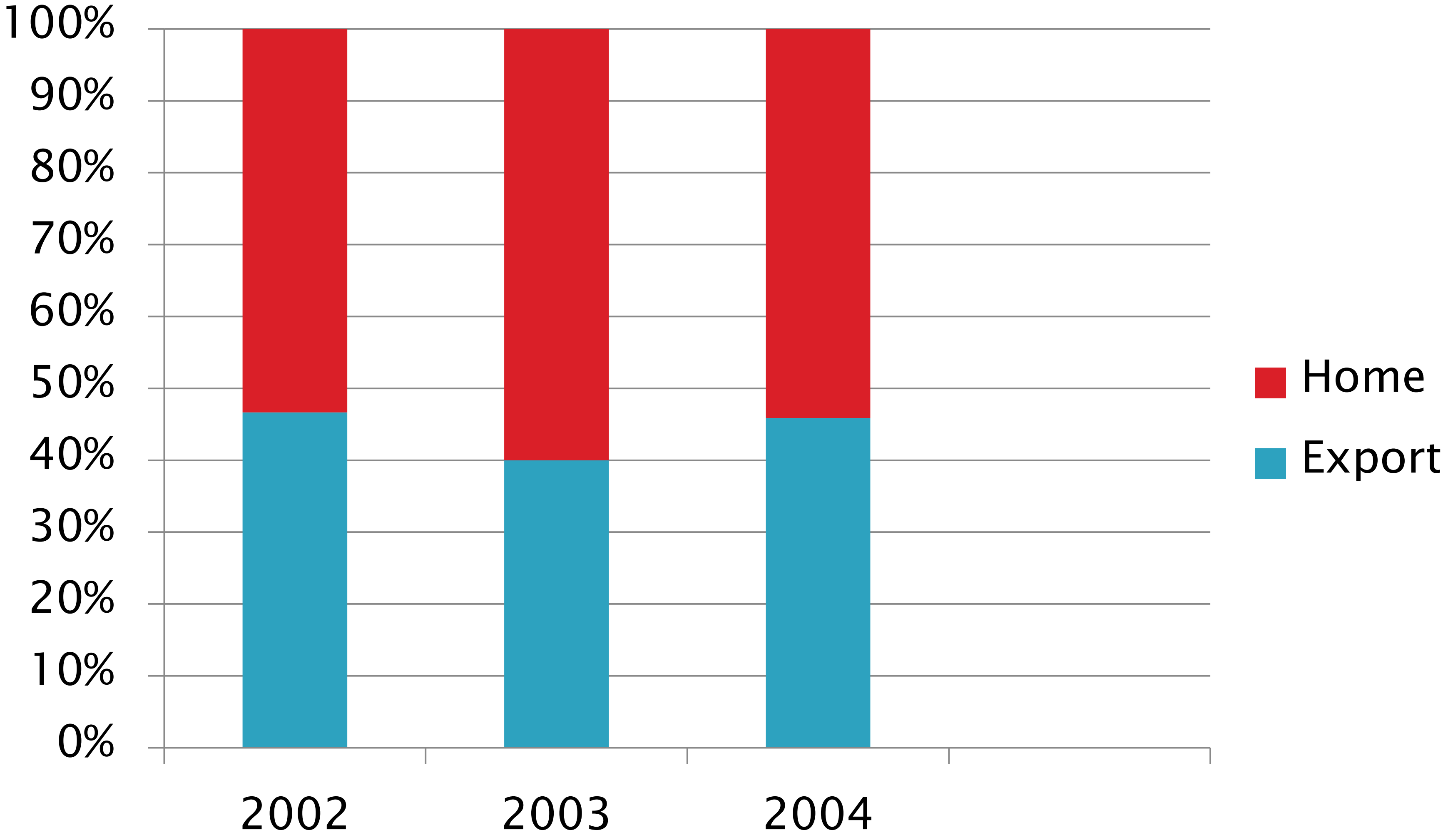
Here each bar representing the total value is sub-divided into its different component parts. This enables comparison between different components and also between a component and the whole.

Example

The data on sales (Rs in Million) of a company are given below.

	2002	2003	2004
Export	1.4	1.8	2.29
Home	1.6	2.7	2.7

Solution



Percentage bar diagram

This is another form of component bar diagram. Here the components are not the actual but percentages of the whole.

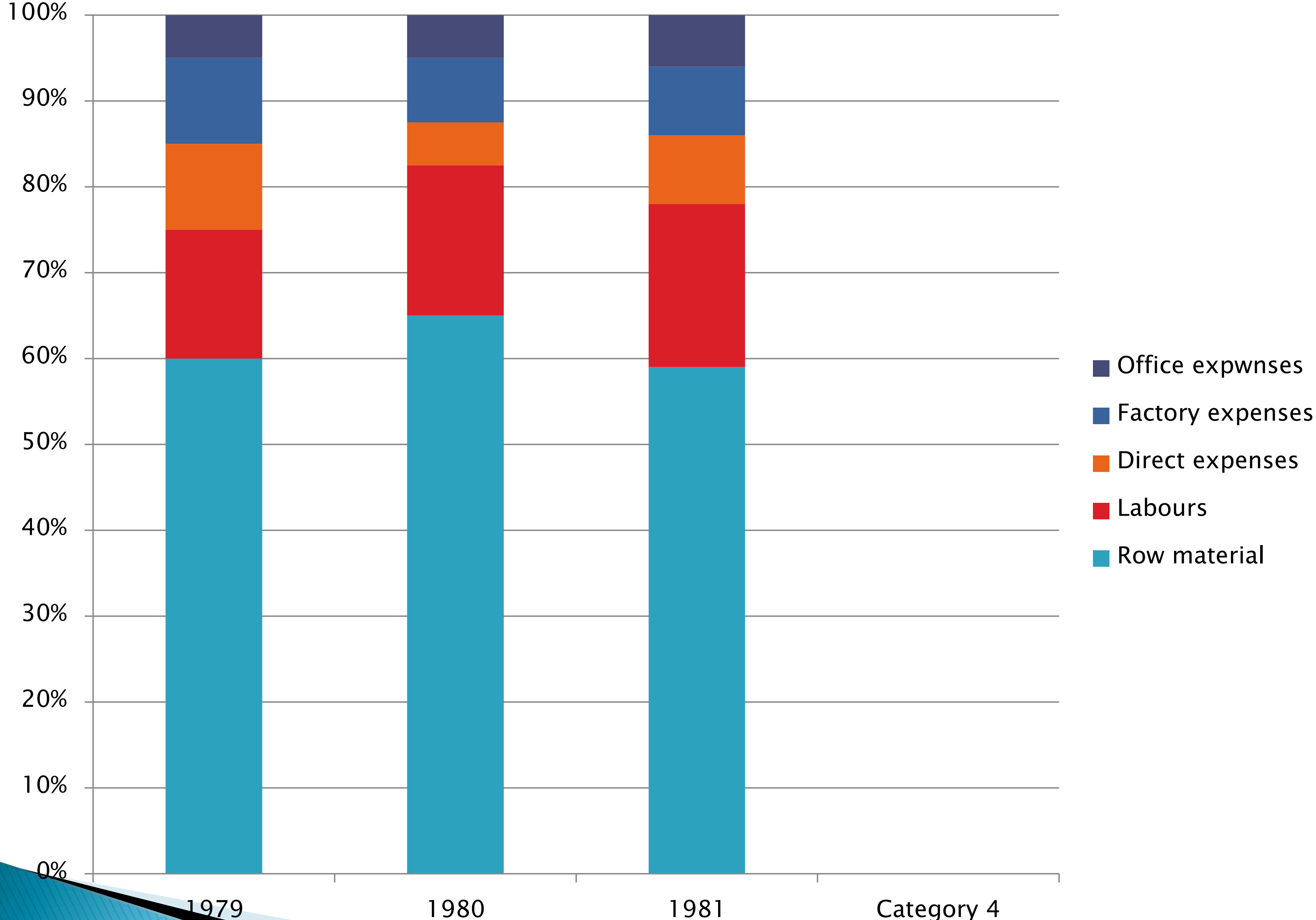
Example 1

Represent the following data by mean of percentage sub-divided bar diagram.

Particular	1979	1980	1981
Raw material	2160	2600	2700
Labour	540	700	870
Direct expenses	360	200	360
Factory expenses	360	300	360
Office expenses	180	200	270
TOTAL	3600	4000	4560

Solution

Particular	1979	1980	1981
Raw material	$\frac{2160}{3600} \times 100 = 60$	$\frac{2600}{4000} \times 100 = 65$	$\frac{2700}{4560} \times 100 = 59$
Labour	$\frac{540}{3600} \times 100 = 15$	$\frac{700}{4000} \times 100 = 17.5$	$\frac{870}{4560} \times 100 = 19$
Direct expenses	$\frac{360}{3600} \times 100 = 10$	$\frac{200}{4000} \times 100 = 5$	$\frac{360}{4560} \times 100 = 8$
Factory expenses	$\frac{360}{3600} \times 100 = 10$	$\frac{300}{4000} \times 100 = 7.5$	$\frac{360}{4560} \times 100 = 8$
Office expenses	$\frac{180}{3600} \times 100 = 5$	$\frac{200}{4000} \times 100 = 5$	$\frac{270}{4560} \times 100 = 6$
	100	100	100


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1979
1980
1981
Category 4
Office expwnses
Factory expenses
Direct expenses
Labours
Row material

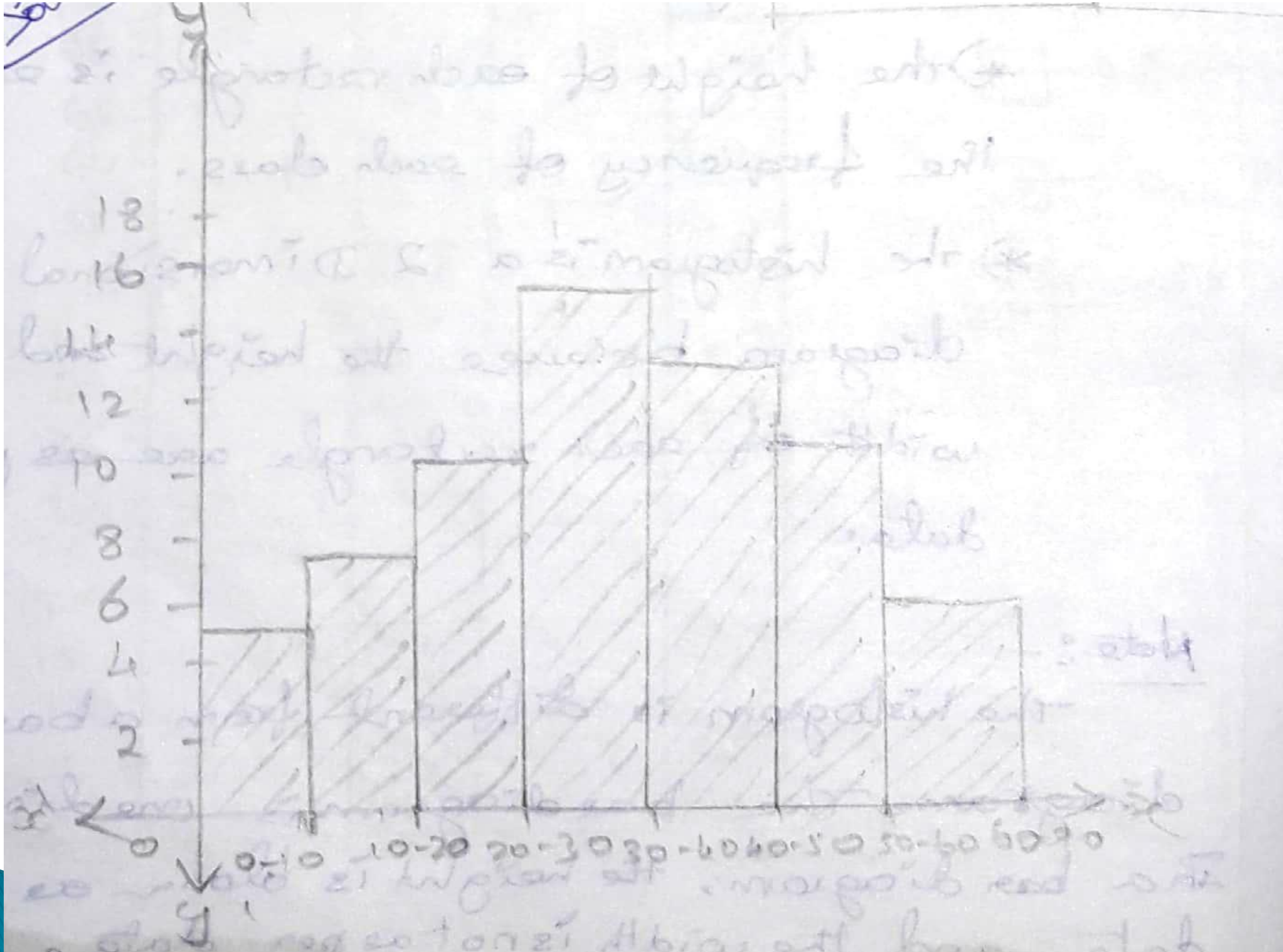
Histogram

A two dimensional representation of a continuous distribution is called a histogram

Example

Draw a histogram for the following data

Marks	0-10	10-20	20-30	30-40	40-50	50-60
Students	5	10	15	20	25	12



Y
18
16
12
10
8
6
4
2
O
0-10
10-20
20-30
30-40
40-50
50-60
60-70
Y'

Example ②:—

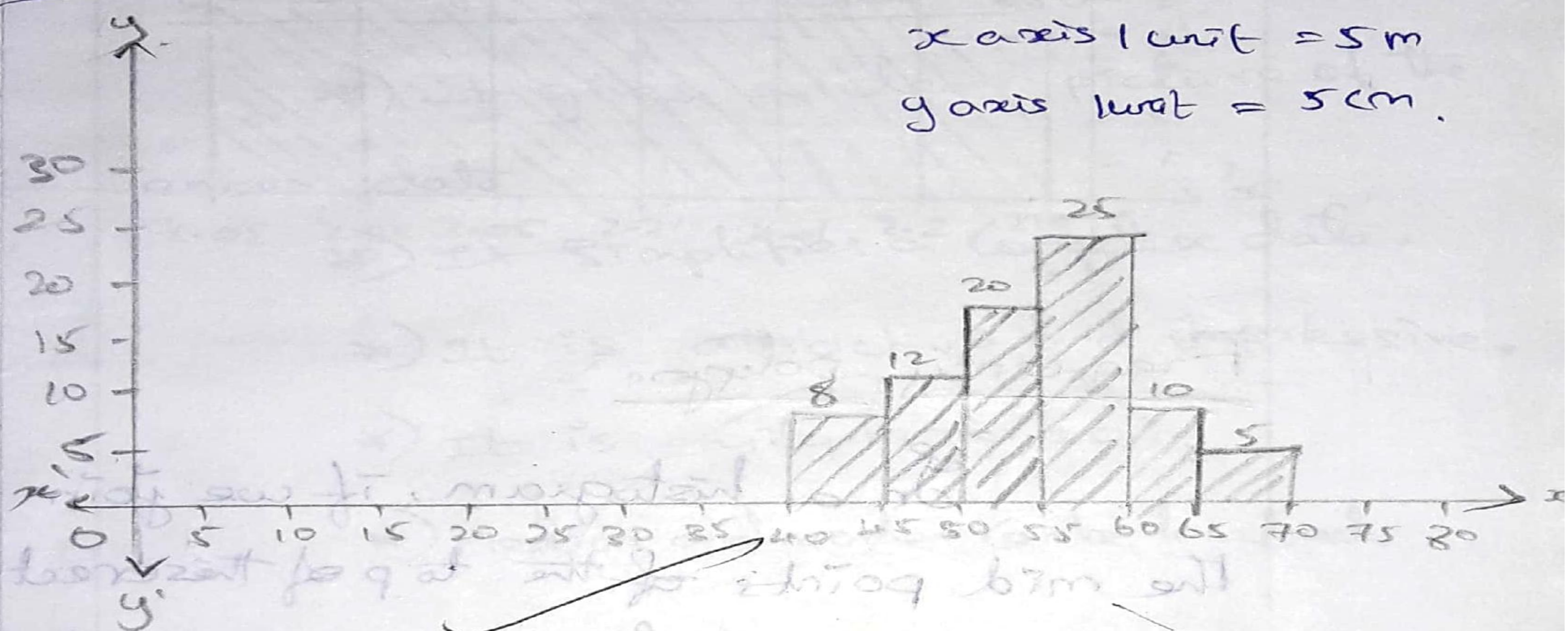
Draw a histogram for the following data.

Marks	40-45	45-50	50-55	55-60	60-65	65-70
students	8	12	20	25	10	5.

Solution:-

x axis 1 unit = 5 m

y axis 1 unit = 5 cm.

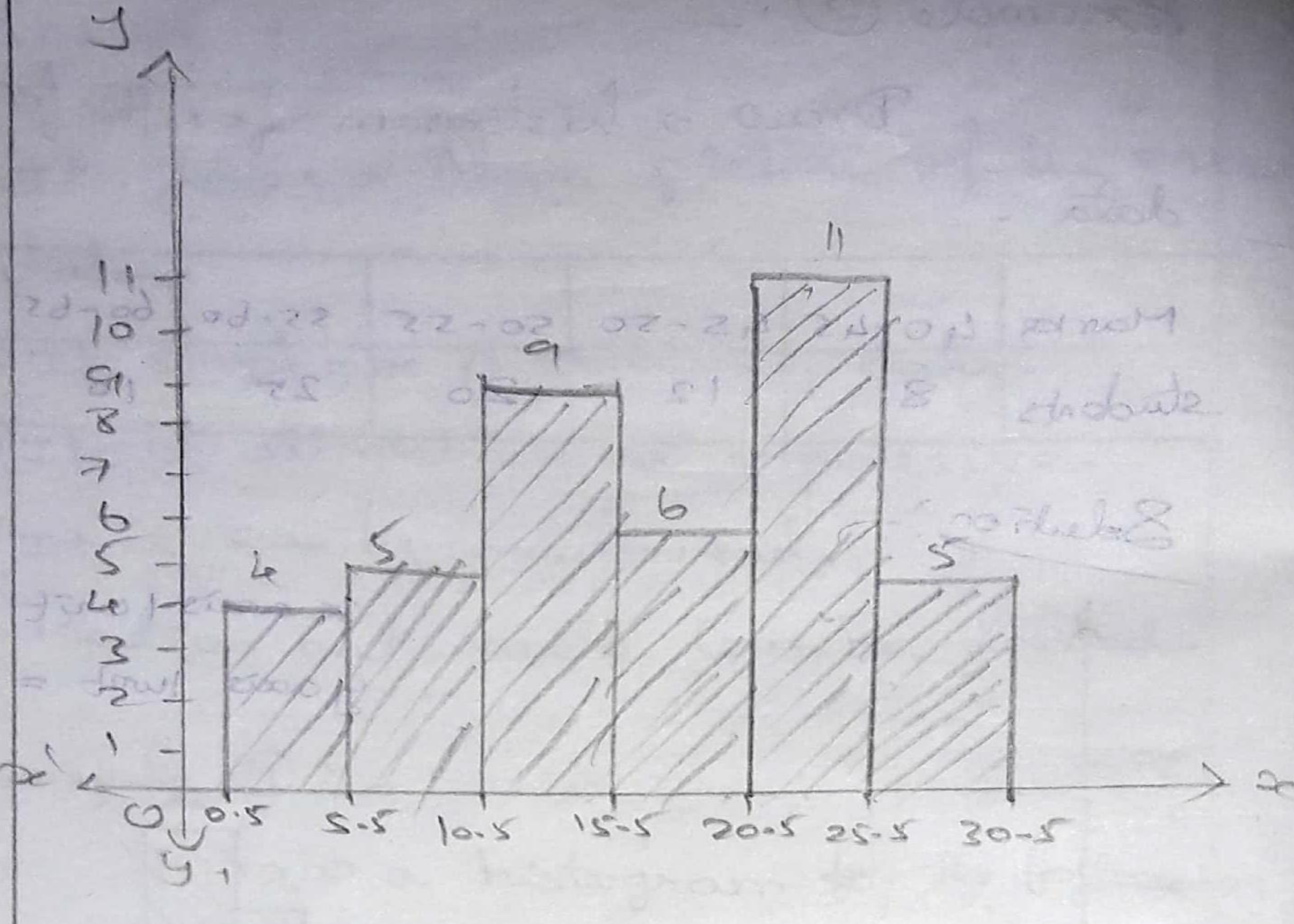


Draw a histogram for the following data

Weight in kg	1–5	6–10	11–15	16–20	21–25	26–30
Number of mean	4	5	9	6	11	5

Solution

True class limits	Frequency
0.5–5.5	4
5.5–10.5	5
10.5–15.5	9
15.5–20.5	6
20.5–25.5	11
25.5–30.5	5



Y
11
10
9
8
7
6
5
4
3
2
1
O
X'
X
Y'
4
5
9
6
11
5
0.5
5.5
10.5
15.5
20.5
25.5
30.5

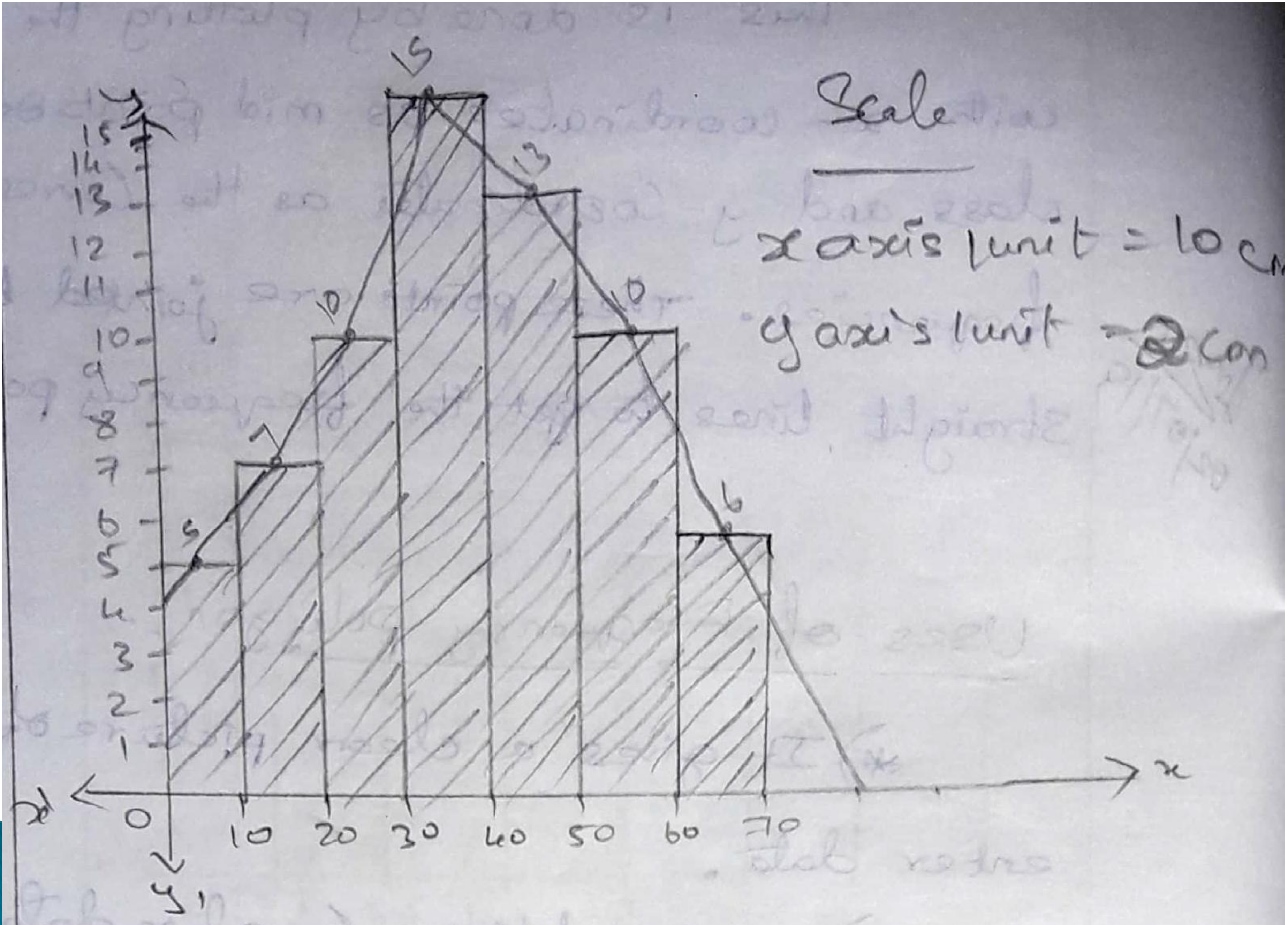
Frequency Polygon

In a histogram, if we join the mid points of the top there rectangular bars by straight lines, we get a polygon such a polygon is called a frequency polygon.

Example

Draw a frequency polygon for the following data

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70
Students	5	7	10	15	13	10	6


Scale
x axis 1 unit = 10 cm
y axis 1 unit = 2 cm
0
10
20
30
40
50
60
70
x
x'
y
y'
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

Frequency Curve

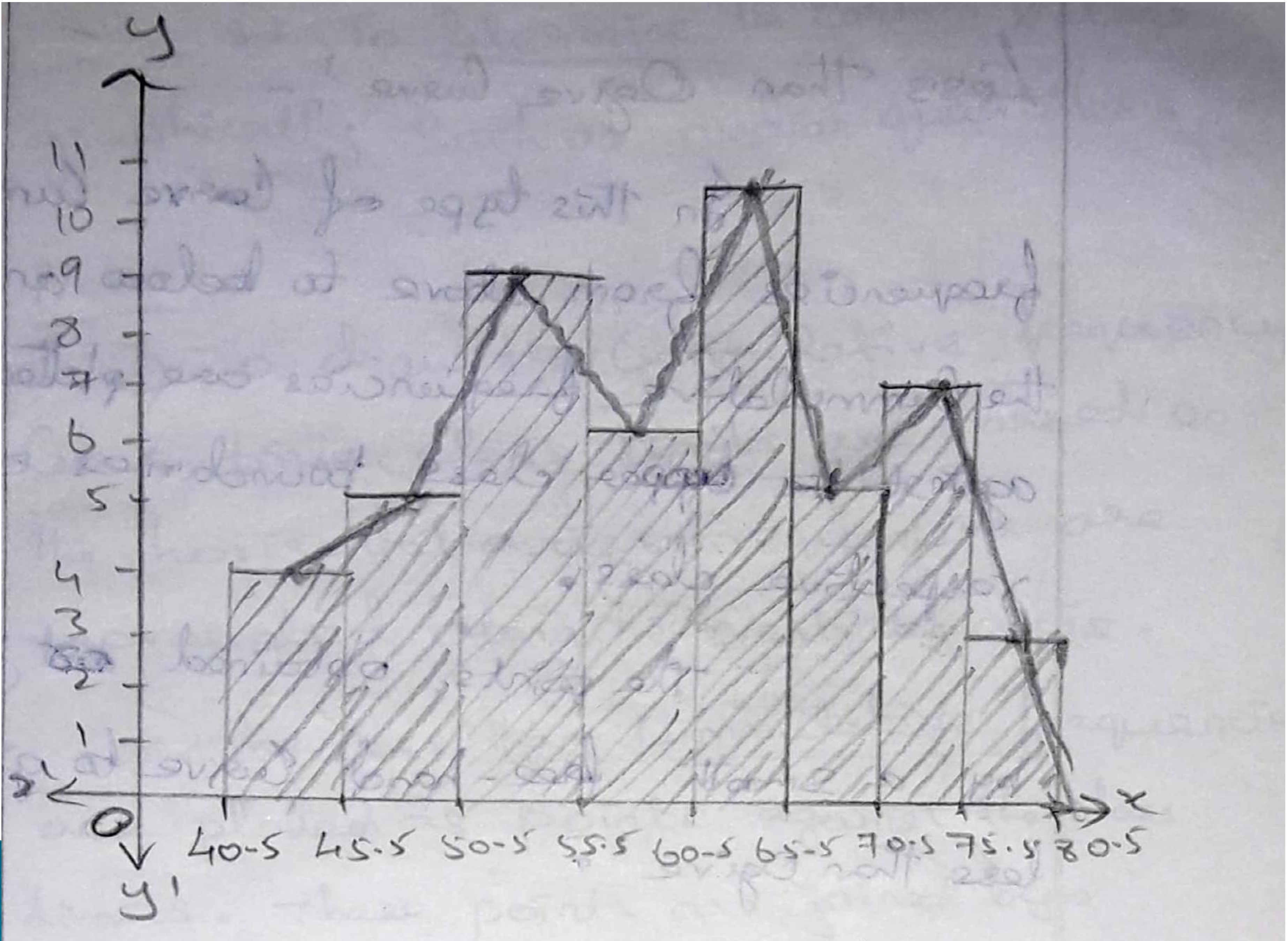
Frequency Curve is drawn by smoothing the frequency polygon. It is smoothened in a such a way that the sharp turns are avoided. It is also called smoothen frequency.

Example

Construct a histogram with frequency curve for the following data

Weight	41-45	46-50	51-55	56-60	61-65	66-70	71-75	76-80
Number of item	44	5	9	6	11	5	7	3

True class limit	f
40.5-45.5	4
45.5-50.5	5
50.5-55.5	9
55.5-60.5	6
60.5-65.5	11
65.5-70.5	5
70.5-75.5	7
75.5-80.8	3



y
11
10
9
8
7
6
5
4
3
2
1
x'
O
40.5
45.5
50.5
55.5
60.5
65.5
70.5
75.5
80.5
x
y'

Ogive curve or Cumulative curve

(i) Less than

(ii) More than

(i) Less than ogive

In this type of curve cumulative frequency from above to below and cumulative frequency are plotted against the upper class boundaries of the respective class. The points so obtained or joined by smooth free hand curve to give less than ogive.

(ii) More than ogive

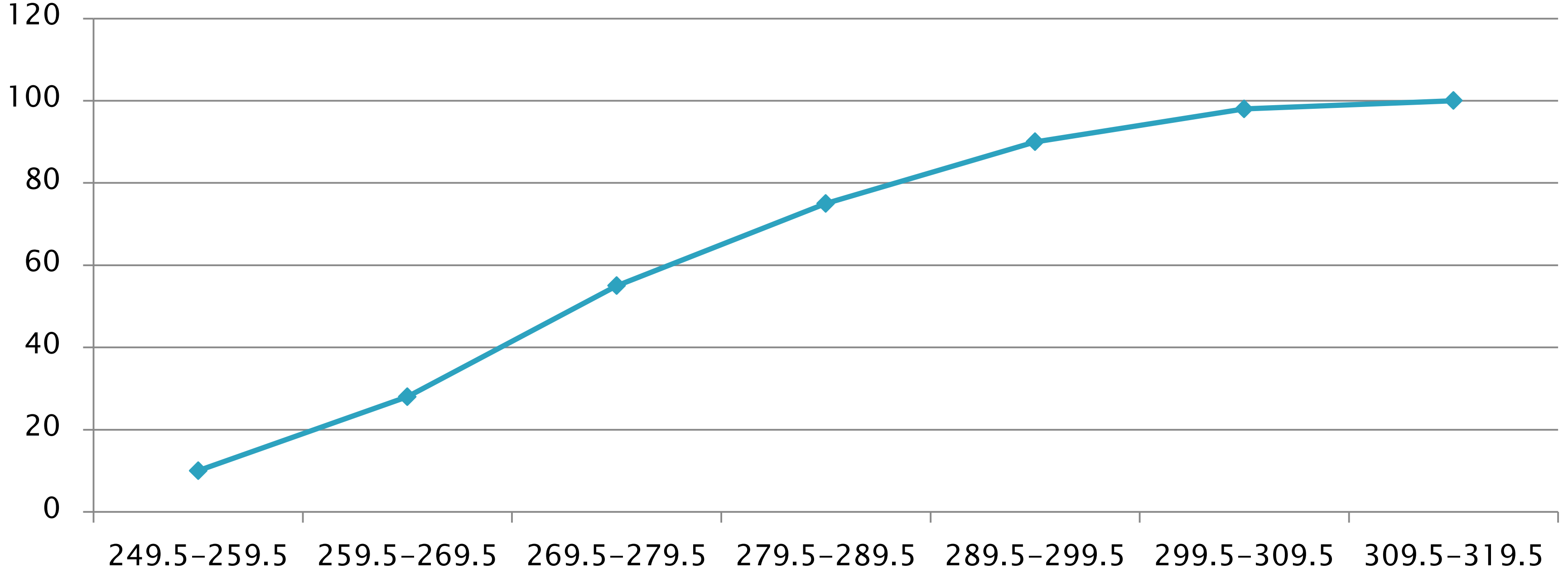
In this type of curve cumulative frequency from below to above and the cumulative frequencies are plotted against the lower class boundaries or the respective class. The points so obtained or joined by a smooth free hand curve to give more than ogive.

1.Draw less than Ogive from the following data

Wages	250-259	260-269	270-279	280-289	290-299	300-309	310-319
No.of workers	10	18	27	20	15	8	2

Solution

True class limit	No.of workers (f)	Less than (cf)
249.5-259.5	10	10
259.5-269.5	18	28
269.5-279.5	27	55
279.5-289.5	20	75
289.5-299.5	15	90
299.5-309.5	8	98
309.5-319.5	2	100



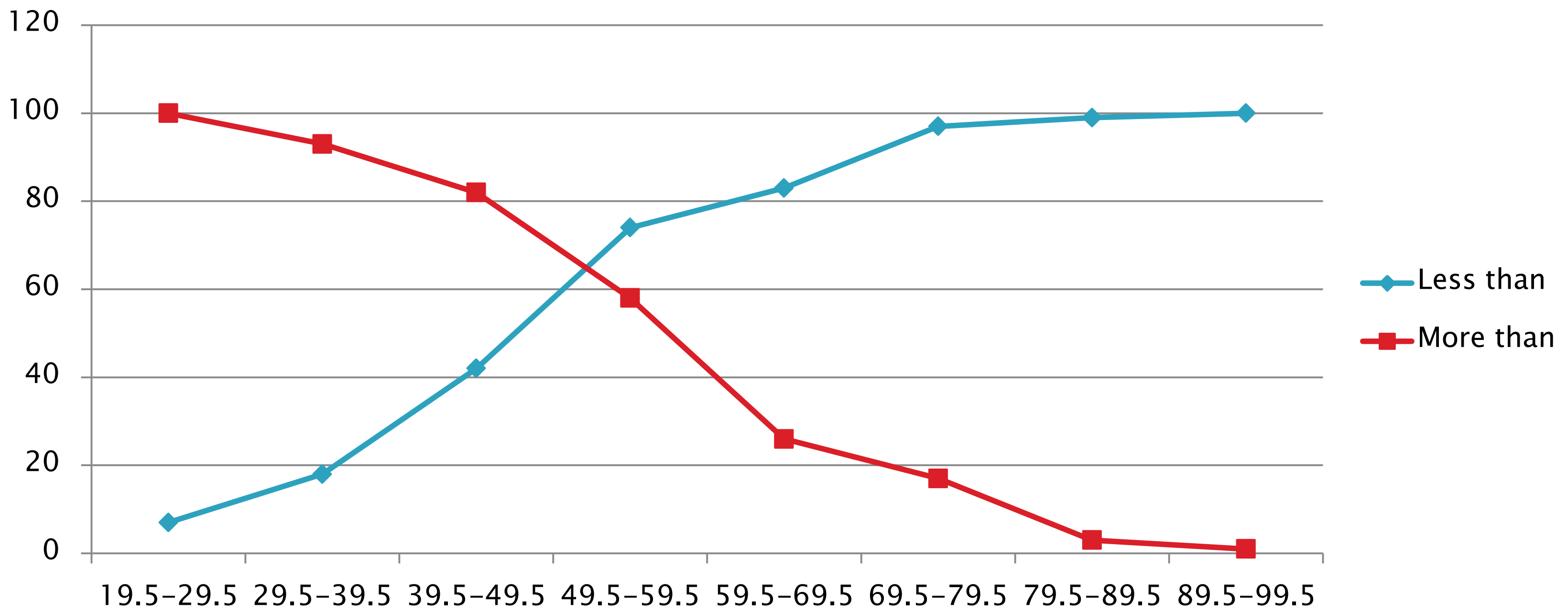
120
100
80
60
40
20
0
249.5–259.5
259.5–269.5
269.5–279.5
279.5–289.5
289.5–299.5
299.5–309.5
309.5–319.5

2.Draw less than and more than ogive for the following data

Marks	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
No.of students	7	11	24	32	9	14	2	1

Solution

True class limits	frequency	Less than	More than
19.5-29.5	7	7	100
29.5-39.5	11	18	93
39.5-49.5	24	42	82
49.5-59.5	32	74	58
59.5-69.5	9	83	26
69.5-79.5	14	97	17
79.5-89.5	2	99	3
89.5-99.5	1	100	1



120
100
80
60
40
20
0
19.5-29.5
29.5-39.5
39.5-49.5
49.5-59.5
59.5-69.5
69.5-79.5
79.5-89.5
89.5-99.5
Less than
More than

Pie diagram

In pie diagram the data are presented in the form of a circle. Pie diagram is also called circular diagram.

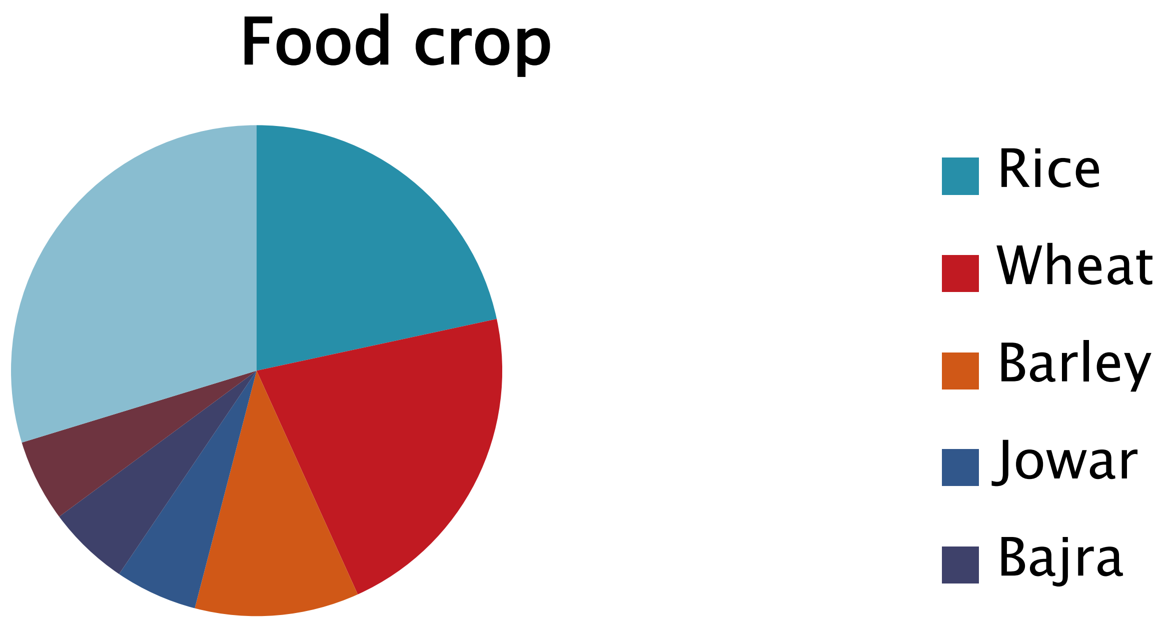
The total area of a circle is 360°. So data is converted into degrees. According to the circle is partioned as the data is drawn as per angles (degrees).Pie diagram is also called angular diagram.

1.Draw Pie diagram for the following data

Food crop	Area
Rice	8
Wheat	8
Barley	4
Jowar	2
Bajra	2
Maize	5
Others	11

Solution

Food crop	Area	Degree
Rice	8	$\frac{8}{40} \times 360^\circ = 72^\circ$
Wheat	8	$\frac{8}{40} \times 360^\circ = 72^\circ$
Barley	4	$\frac{4}{40} \times 360^\circ = 36^\circ$
Jowar	2	$\frac{2}{40} \times 360^\circ = 18^\circ$
Bajra	2	$\frac{2}{40} \times 360^\circ = 18^\circ$
Maize	5	$\frac{5}{40} \times 360^\circ = 45^\circ$
Others	11	$\frac{11}{40} \times 360^\circ = 99^\circ$
Total	40	360°



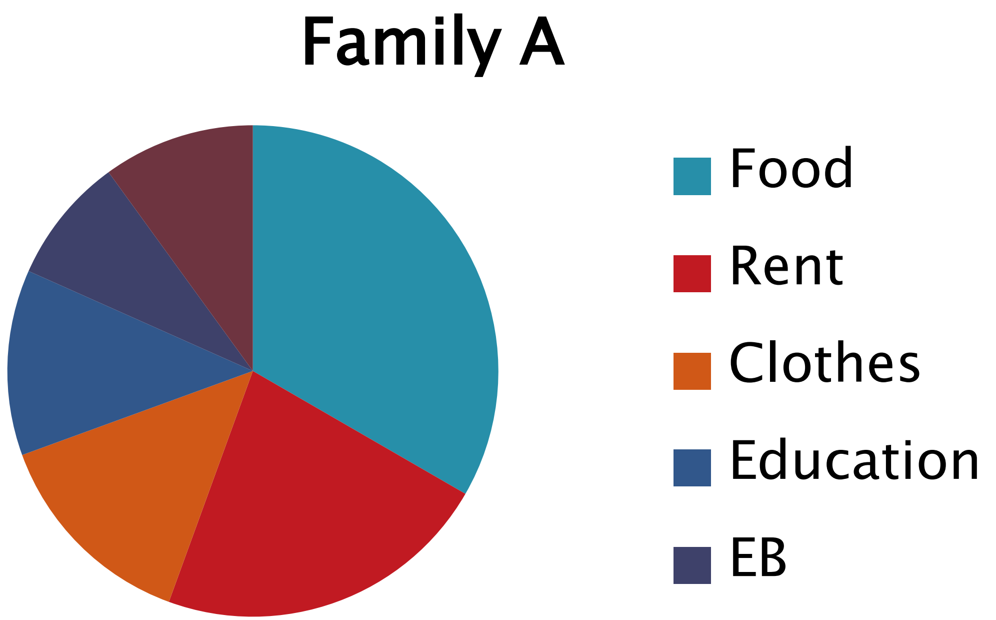
Pie diagram

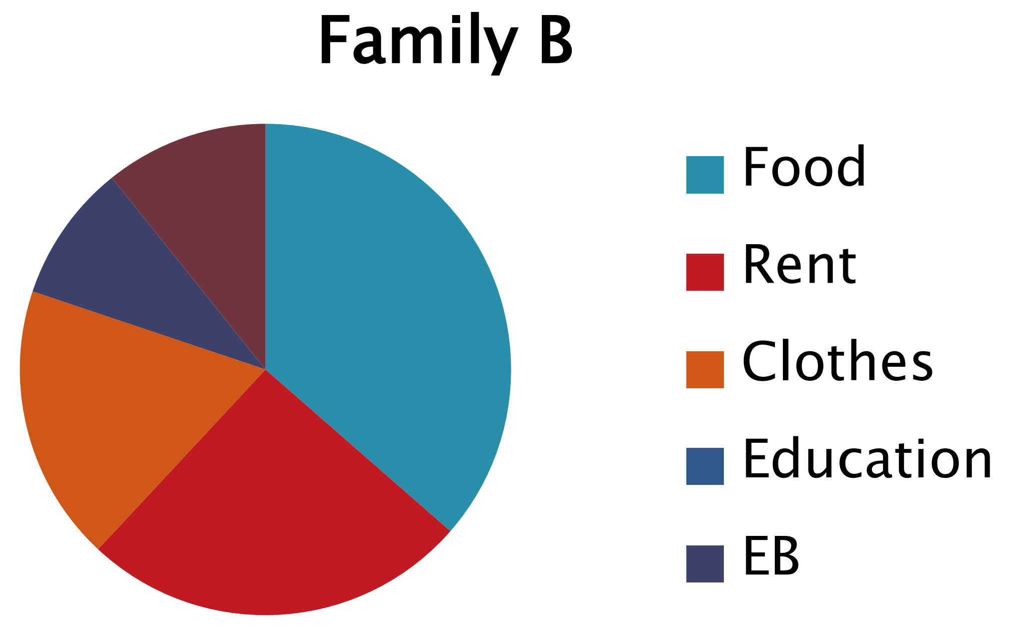
2.Draw a circular diagram from the following data

Type of commodity	Expenditure in Rupees	
	Family A	Family B
Food	300	500
Rent	200	350
Clothes	125	250
Education	110	225
EB	75	125
Savings	90	150

Solution

Type of commodity	Expressed in Degrees	
	Family A	Family B
Food	$\frac{300}{900} \times 360^{\circ} = 120^{\circ}$	$\frac{500}{1600} \times 360^{\circ} = 112.5^{\circ}$
Rent	$\frac{200}{900} \times 360^{\circ} = 80^{\circ}$	$\frac{350}{1600} \times 360^{\circ} = 78.75^{\circ}$
Clothes	$\frac{125}{900} \times 360^{\circ} = 50^{\circ}$	$\frac{250}{1600} \times 360^{\circ} = 56.25^{\circ}$
Education	$\frac{110}{900} \times 360^{\circ} = 44^{\circ}$	$\frac{225}{1600} \times 360^{\circ} = 50.6^{\circ}$
EB	$\frac{75}{900} \times 360^{\circ} = 30^{\circ}$	$\frac{125}{1600} \times 360^{\circ} = 28.1^{\circ}$
Savings	$\frac{90}{900} \times 360^{\circ} = 36^{\circ}$	$\frac{150}{1600} \times 360^{\circ} = 33.7^{\circ}$


Family A
Food
Rent
Clothes
Education
EB


Family B
Food
Rent
Clothes
Education
EB

Thank You

Om Sakthi
ADHIPARASAKTHI COLLEGE OF ARTS AND SCIENCE
(Autonomous)
G.B. NAGAR, KALAVAI – 632 506, RANIPET DISTRICT, TAMILNADU
Permanently Affiliated to Thiruvalluvar University,
Re-accredited by NAAC – 'B' Grade (CGPA - 2.83)

STATISTICAL METHODS AND THEIR APPLICATIONS-I

II B.SC Computer Science

Handled by

Dr.P.RAJAKUMARI

ASSISTANT PROFESSOR

DEPARTMENT OF MATHEMATICS

ADHIPARASAKTHI COLLEGE OF ARTS AND SCIENCE

AUTONOMOUS (AUTONOMOUS)

UNIT-II

MEASURES OF LOCATION

The various measures of averages commonly used are

(i) Mean

(ii) Median

(iii)Mode

There are 3 types of mean that are used they are,

(i) Arithmetic mean (AM)

(ii)Geometric mean (GM)

(iii)Harmonic mean (HM)

(i) Arithmetic Mean (AM)

The arithmetic mean of a set of n observations $x_1, x_2, \ldots\ldots x_n$ is defined

$$\bar{X} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

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

1.Find the AM of the following set of observations
25,32,28,34,24,31,36,27,29,30
Solution

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

$$= \frac{25 + 32 + 28 + 34 + 24 + 31 + 36 + 27 + 29 + 30}{10}$$

$$= \frac{296}{10}$$

$$\bar{X} = 29.6$$

2,Find the AM of the following set of observations 15,20,10,35,32 (HW)

Discrete series

$$\bar{X} = \frac{\sum fx}{n}$$

Where x=Random value

f=frequency

N=Total frequency

1.Calculate the AM for the following data

Age in year	8	10	12	15	18
No.of workers	5	7	12	6	10

x	f	fx
8	5	40
10	7	70
12	12	144
15	6	90
18	10	180
	N=40	$\sum fx = 524$

$$\bar{X} = \frac{\sum fx}{n}$$

$$\bar{X} = \frac{524}{40}$$

$$\bar{X} = 13.1$$

2. Calculate the AM for the following data

x	61	64	67	70	73
f	15	54	96	81	24

3. Calculate the AM for the following data

x	14.5	24.5	34.5	44.5	54.5	64.5	74.5	84.5
f	5	9	14	20	25	15	8	4

4. Calculate the AM for the following data

No.of components	0	1	2	3	4	5	6
Number of workers	2	36	40	20	10	5	6

Continuous series

$$\bar{X} = \mathrm{A} + \frac{\sum fd}{N} \times \mathrm{C}$$

Where $\mathrm{A} = \mathrm{Assumed\ mean}$

$f = frequency$

$d = \frac{x - A}{C}$

$C = Class\ interval$

$N = Total\ frequency$

1.Calculate the AM for the following data

x	0-10	10-20	20-30	30-40	40-50
f	5	12	15	20	25

Solution

x	f	Mid x	$d = \frac{x - A}{C}$	fd
0-10	5	5	$\frac{5-25}{10} = \frac{-20}{10} = -2$	-10
10-20	12	15	$\frac{15-25}{10} = \frac{-10}{10} = -1$	-12
20-30	15	$25 \rightarrow A$	$\frac{25-25}{10} = \frac{0}{10} = 0$	0
30-40	20	35	$\frac{35-25}{10} = \frac{10}{10} = 1$	20
40-50	25	45	$\frac{45-25}{10} = \frac{20}{10} = 2$	50
	N=77			$\sum fd = 48$

$A = 25\ ,\ \ C = 10,\ \ N = 77, \sum fd = 48$

$$\bar{X} = A + \frac{\sum fd}{N} \times C$$

$$\bar{X} = 25 + \frac{48}{47} \times 10$$

$$= 25 + \frac{480}{47}$$

$$= 25 + 6.23$$

$$\bar{X} = 31.23$$

2. **Calculate the AM for the following data**

x	0-10	10-20	20-30	30-40	40-50	50-60	60-70
f	32	25	30	24	16	15	8

3. Calculate the AM for the following data

x	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
f	7	11	24	32	9	14	2	1

Solution

X	f	Mid x	$d = \frac{x - A}{C}$	fd
19.5-29.5	7	24.5	$\frac{24.5 - 59.5}{10} = \frac{-35}{10} = -3.5$	-24.5
29.5-39.5	11	34.5	$\frac{34.5 - 59.5}{10} = -2.5$	-2.5
39.5-49.5	24	44.5	$\frac{44.5 - 59.5}{10} = -1.5$	-36
49.5-59.5	32	54.5	$\frac{54.5 - 59.5}{10} = -0.5$	-16
59.5-69.5	9	64.5	$\frac{64.5 - 59.5}{10} = 0.5$	4.5
69.5-79.5	14	74.5	$\frac{74.5 - 59.5}{10} = 1.5$	21
79.5-89.5	2	84.5	$\frac{84.5 - 59.5}{10} = 2.5$	5
89.5-99.5	1	94.5	$\frac{94.5 - 59.5}{10} = 3.5$	3.5
	N=100			$\sum fd = -70$

$$\boldsymbol{A = 59.5\ ,\ C = 10,\quad N = 100, \sum fd = -70}$$

$$\bar{X} = \mathbf{A} + \frac{\sum fd}{N} \times \mathbf{C}$$

$$\bar{X} = 59.5 + \frac{-70}{100} \times 10$$

$$\bar{X} = 52.5$$

Geometric mean (GM)

$$\mathbf{G.M} = (x_1 . x_2 . x_3 \dots \dots . . x_n)^{\frac{1}{n}}$$

1.Find GM for the following data 3,6,24,48

Solution

n=4

$$\mathbf{G.M} = (x_1 . x_2 . x_3 \dots \dots . . x_n)^{\frac{1}{n}}$$

$$\mathbf{G.M} = (\mathbf{3} \times \mathbf{6} \times \mathbf{24} \times \mathbf{48})^{\wedge} \left(\frac{\mathbf{1}}{\mathbf{4}}\right)$$

$$\mathbf{G.M} = \mathbf{12}$$

2. Find GM for the following data 7,12,35,52,64

Discrete series

$$G.M = Anti\ log\left(\frac{\sum f.\log x}{N}\right)$$

Where N=Total frequency ($\sum f$)

1. Find GM for the following data

Age	8	10	12	15	18
No.of workers	5	7	12	6	10

Solution

x	$log x$	f	$f log x$
8	0.9031	5	4.5155
10	1	7	7
12	1.0792	12	12.9504
15	1.1761	6	7.0566
18	1.2553	10	12.553
		$N = 40$	$\sum f.\log x = 44.0755$

$N = 40$, $\sum f . \log x = 44.0755$

$$G.M = Anti\ log \left(\frac{\sum f . \log x}{N} \right)$$

$$G.M = Anti\ log \left(\frac{44.0755}{40} \right)$$

$$G.M = Anti\ log\ (1.1019)$$

$$G.M = 12.6445$$

2. Find GM for the following data

x	35	40	45	50	55	60
f	12	18	24	16	6	4

Continuous series

$$G.M = Anti\ log\left(\frac{\sum f.\log m}{N}\right)$$

1.Calculate GM for the following data

x	40-50	50-60	60-70	70-80	80-90	90-100
f	19	25	36	72	51	43

Solution

x	$Mid\ x\ (or)\ m$	$log\ m$	f	$f\ logm$
40-50	45	1.6532	19	31.4108
50-60	55	1.7404	25	43.51
60-70	65	1.8129	36	65.2644
70-80	75	1.8751	72	135.0072
80-90	85	1.9294	51	98.3994
90-100	95	1.9777	43	85.0411
			N=246	$\sum f\,.log\ m = 458.6329$

N=246 $\sum f.\log m = 458.6329$

$$\boldsymbol{G.M = Anti\ log\left(\frac{\sum f.\log m}{N}\right)}$$

$$G.M = Anti\ log\left(\frac{458.6329}{246}\right)$$

$$G.M = Anti\ log\ (1.8644)$$

$$G.M = 73.1813$$

2.Calculate GM for the following data

Marks	4-8	8-12	12-16	16-20	20-24	24-28	28-32	32-36	36-44
Frequency	6	10	14	18	22	26	30	34	38

Harmonic Mean (HM)

Individual data (or) Individual series

$$HM = \frac{n}{\sum\left(\frac{1}{x}\right)} \quad (or) \quad \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}}$$

1.Calculate the HM from the following data 100,200,300,400

Solution

$$n = 4, x_1 = 100, x_2 = 200, x_3 = 300, x_4 = 400$$

$$HM = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}}$$

$$= \frac{4}{\frac{1}{100} + \frac{1}{200} + \frac{1}{300} + \frac{1}{400}}$$

$$= \frac{4}{0.01 + 0.005 + 0.003 + 0.0025}$$

$$= \frac{4}{0.0205}$$

$$HM = 195.1219$$

2.Calculate the HM from the following data 60,25,350,25

Discrete series

$$HM = \frac{N}{\sum(fx)} \ (or)\ HM = \frac{N}{\sum f\left(\frac{1}{x}\right)}$$

1.Find the HM for the following data

Size	3	5	7	9
Frequency	20	40	30	10

x	$\frac{1}{x}$	f	$\frac{f}{x}$
3	0.3333	20	6.666
5	0.20	40	3
7	0.1429	30	4.2857
9	0.1111	10	1.1111
		N=100	$\sum\left(\frac{f}{x}\right)$=20.0634

N=100, $\sum\left(\frac{f}{x}\right)$=20.0634

$$HM = \frac{N}{\Sigma\left(\frac{f}{x}\right)}$$

$$HM = \frac{100}{20.0634}$$

$$HM = 4.9842$$

2.Find the HM for the following data

Size	4	5	6	7	8	9
Frequency	8	10	9	6	4	3

3.Find the HM for the following data

Size	5	10	15	20	25	30
Frequency	7	13	35	27	19	9

Continuous series

$$HM = \frac{N}{\sum f\left(\frac{1}{m}\right)}$$

Where

m=mid value of a class

N=Total frequency

1.Find HM for the following data

Class	1-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80	81-90	91-100
f	3	7	13	17	23	27	18	12	6	4

Solution

x	f	$mid\ m$	$\frac{1}{m}$	$\frac{f}{m}$
0.5-10.5	3	5.5	0.1818	0.5454
10.5-20.5	7	15.5	0.0645	0.4515
20.5-30.5	13	25.5	0.0392	0.5096
30.5-40.5	17	35.5	0.0282	0.4794
40.5-50.5	23	45.5	0.022	0.506
50.5-60.5	27	55.5	0.018	0.486
60.5-70.5	18	65.5	0.0153	0.2754
70.5-80.5	12	75.5	0.0132	0.1584
80.5-90.5	6	85.5	0.0117	0.0702
90.5-100.5	4	95.5	0.0105	0.042
	N=130			$\sum\left(\frac{f}{m}\right)$=3.5239

$$HM = \frac{N}{\sum\left(\frac{f}{m}\right)}$$

$$HM = \frac{130}{3.5239}$$

$$HM = 37.8909$$

2.Find HM for the following data

x	20-30	30-40	40-50	50-60	60-70	70-80	80-90
f	8	12	23	31	9	13	2

3.Find HM for the following data

x	1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40	41-45	46-50
f	2	8	12	18	24	14	10	6	4	2

Median

If the observations are arranged in ascending order or descending order of magnitude, the middle most item is called the median.

1.Find the Median for the following data 8,10,11,16,20,25,15,9,6

Solution

The observations are arranged in ascending order 6,8,9,10,11,15,16,20,25

Number of observation=9

Median=11

2.Find the Median for the following data 27,36,28,18,35,26,20,35,40,26

Solution

The observations are arranged in ascending orders are 18,20,26,26,27,28,35,35,36,40

Number of observation=10

$$\text{Median}=\frac{27+28}{2}=27.5$$

3.Find the Median for the following data 18,20,26,27,28,35,40

Discrete series

Median= $\frac{N}{2}$

1.Find the Median for the following data

Solution

Daily wages (in Rs)	No.of persons
5	7
10	12
15	37
20	25
25	22
30	11

x	f	Cf
5	7	7
10	12	19
15	37	56
20	25	81
25	22	103
30	11	114

Median= $\frac{N}{2}$

Median= $\frac{114}{2}$

Median=57

Median=20

2. Find the median for the following data

x	4	5	6	7	8	9
f	10	15	22	16	12	15

3. Find the median for the following data

x	0	1	2	3	4	5	6	7	8
f	8	10	11	16	20	25	15	9	6

Continuous series

Median= $l + \frac{\frac{N}{2} - m}{f} \times c$

$l = lower\ limit$

N=Total frequency

m-Cumulative frequency of the pre-median class

f-frequency of the median class

C-class interval

1.Find the median for the following frequency distribution

Solution

Class	0-10	10-20	20-30	30-40	40-50	50-60	60-70
frequency	4	16	100	60	40	6	4

x	f	cf
0-10	4	4
10-20	16	20
20-30	100	120
30-40	60	180
40-50	40	220
50-60	6	226
60-70	4	230

N=230,

$\frac{N}{2} = \frac{230}{2} = 115,$

m=20,

c=10,

f=100

Median= $l + \frac{\frac{N}{2} - m}{f} \times c$

$= 20 + \frac{115 - 20}{100} \times 10$

$= 20 + \frac{95}{10}$

$= 20 + 9.5$

$Median = 29.5$

2. Find the median for the following data

Class	0-10	10-20	20-30	30-40	40-50
f	22	38	46	34	20

3. Find the median for the following data

x	1-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80	81-90	91-100
f	3	7	13	17	12	10	8	8	6	6

Mode

Mode is defined to be the value that occurs most often

Row data

1.Find the mode for the following data 3,7,2,5,7,3

Solution

7 occurs 3 times

Mode=7

2.Find the mode for the following data 18,21,25,23,27,25,29,25,29,25

Discrete series

Here mode is defined as the value that corresponds to the maximum frequency in the given data.

1.Find the mode for the following data

x	10	11	12	13	14	15	16
f	8	4	12	24	26	7	11

Solution

Maximum frequency=26

Mode=14

2.Find the mode for the following data

Height	58	60	61	62	64	65	66	77
Plant	4	6	5	10	20	24	6	2

Continuous series

Mode $Z = l + \frac{f_1 - f_0}{2f_1 - (f_0 + f_2)} \times c$

1.Find the mode for the following data

x	0-10	10-20	20-30	30-40	40-50	50-60	60-70
f	4	9	13	15	12	8	3

Solution

x	f
0-10	4
10-20	9
20-30	13
30-40	15
40-50	12
50-60	8
60-70	3

$C = 10,\ l = 30,\ \ f_0 = 13,\ f_1 = 15,\ f_2 = 12$

$$\text{Mode } Z = l + \frac{f_1 - f_0}{2f_1 - (f_0 + f_2)} \times c$$

$$\text{Mode } Z = 30 + \frac{15 - 13}{2(15) - (13 + 12)} \times 10$$

$$= 30 + \frac{2}{30 - 25} \times 10$$

$$= 30 + \frac{2}{5} \times 10$$

$$= 30 + 4$$

$$Mode\ Z = 34$$

2.Find the Mode for the following data

x	46-50	51-55	56-50	61-65	66-70	71-75	76-80	81-85	86-90	91-95
f	2	3	5	7	9	11	17	2	3	1

1.Find the mean , median, mode for the following data and also verify empirical relation

x	130-134	135-139	140-144	145-149	150-154	155-154	160-164
f	5	15	28	24	17	10	1

Solution

x	f	Mid x	$d = \frac{x - A}{c}$	cf	fd
129.5-134.5	5	132	-3	5	-15
134.5-139.5	15	137	-2	20	-30
139.5-144.5	28	142	-1	48	-28
144.5-149.5	24	147	0	72	0
149.5-154.5	17	152	1	89	17
154.5-159.5	10	157	2	99	20
159.5-164.5	1	162	3	100	3
	N=100				$\sum fd = -33$

A=147, C=5, $\sum f\, d = -33, N = 100$

$$\bar{X} = A + \frac{\sum f d}{N} \times C$$

$$\bar{X} = 147 + \frac{(-33)}{100} \times 5$$

$$\bar{X} = 147 - 165$$

$$\bar{X} = 145.35$$

N=100, $\frac{N}{2} = \frac{100}{2} = 50$, m=48, c=5, f=24

Median= $l + \frac{\frac{N}{2} - m}{f} \times c$

$= 144.5 + \frac{50-48}{24} \times 5$

$= 144.5 + \frac{2}{24} \times 5$

$= 144.5 + \frac{5}{12}$

$= 144.5 + 0.416$

$Median = 144.916$

$$C = 5,\ l = 139.5,\ f_0 = 15,\ f_1 = 28,\ f_2 = 24$$

$$\text{Mode } Z = l + \frac{f_1 - f_0}{2f_1 - (f_0 + f_2)} \times c$$

$$\text{Mode } Z = 139.5 + \frac{28 - 15}{2(18) - (15 + 24)} \times 5$$

$$= 139.5 + \frac{28 - 15}{2(18) - (15 + 24)} \times 5$$

$$= 139.5 + \frac{13}{56 - 39} \times 5$$

$$= 139.5 + \frac{13}{17} \times 5$$

$$= 139.5 + 3.8235$$

$$Mode\ Z = 143.3235$$

Empirical relation

Mean-mode=3(mean-median)

145.35-143.3235=3(145.35-144.916)

2.0265=3(0.0432)

2.0265≠1.302

Empirical relation is not true

2.Find the mean , median, mode for the following data and also verify empirical relation

x	1-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80	81-90	91-100
f	3	7	13	17	12	10	8	8	6	6

3.Find the mean , median, mode for the following data

x	58	60	61	62	63	64	65	66	68	70
f	4	6	5	10	20	22	24	6	2	1

Solution

x	f	fx	cf
58	4	232	4
60	6	360	10
61	5	305	15
62	10	620	25
63	20	1260	45
64	22	1408	67
65	24	1560	91
66	6	396	97
68	2	136	99
70	1	70	100
	N=100	$\sum fx = 6347$	

$$\text{Mean} = \frac{\sum fx}{N}$$

$$= \frac{6347}{100}$$

Mean = 63.47

$$\text{Median} = \frac{N}{2}$$

$$= \frac{100}{2} = 50$$

Median = 64

Mode = High frequency value

= 24

Mode = 65

4.Find the mean , median, mode for the following data

x	5	15	25	35	45	55	65	75	85	95
f	2	5	14	30	52	45	27	15	10	7

Missing frequency

1.Estimate the missing frequency from the given data. Total frequency 100 and median is 44.

x	10-20	20-30	30-40	40-50	50-60	60-70	70-80
f	5	12	-	20	-	10	4

Solution

x	f	cf
10-20	5	5
20-30	12	17
30-40	x	17+x
40-50	20	37+x
50-60	x	37+x+y
60-70	10	47+x+y
fd70-80	4	51+x+y

$51+x+y=100$

$x+y=100-51$

$x+y=49$ ______________________ (1)

$$\text{Median}= l + \frac{\frac{N}{2}-m}{f} \times c$$

$$44=40 + \frac{50-(17+x)}{20} \times 10$$

$$44 - 40 = \frac{50-(17+x)}{20} \times 10$$

$$4=\frac{33-x}{2}$$

$8=33-x$

$x=33-8$

$x=25$ ______________________ (2)

$x + y = 49$

$25 + y = 49$

$y = 49 - 25$

$y = 24$

The frequency class value 30-40 and 50-60 is 25 and 24

2.The arithmetic mean is calculated from the following frequency distribution is known to be 67.5 inches. Find the missing frequency.

Wight in inches	60-62	63-65	66-68	69-71	72-74
frequency	15	54	-	81	24

3.The expenditure of 1000 families is given as under

Expenditure in Rs	40-57	60-79	80-99	100-119	120-139
No.of families	50	-	500	-	50

The median and mean for the distributions are both Rs.87.5 respectively. Calculate missing frequency.

Om Sakthi
ADHIPARASAKTHI COLLEGE OF ARTS AND SCIENCE
(Autonomous)
G.B. NAGAR, KALAVAI – 632 506. RANIPET DISTRICT, TAMILNADU
Permanently Affiliated to Thiruvalluvar University,
Re-accredited by NAAC – 'B' Grade (CGPA - 2.83)

STATISTICAL METHODS AND THEIR APPLICATIONS-I

II B.SC Computer Science

Handled by

Dr.P.RAJAKUMARI

ASSISTANT PROFESSOR

DEPARTMENT OF MATHEMATICS

ADHIPARASAKTHI COLLEGE OF ARTS AND SCIENCE

AUTONOMOUS (AUTONOMOUS)

UNIT-3

MEASURES OF DISPERSION

The measurement of the scatterness of the mass of figures in a series about an average is called measures of dispersion.

The various measures of dispersion

(i) Range

(ii)Quartile deviation (QD)

(iii) Mean deviation (MD)

(iv) Standard deviation (SD)

Range

Range is the difference between the greatest and the smallest value of the distribution.

Range=L-S

Where L-Largest value

S-Smallest value

Co-efficient of Range

The relative measure of range is called co-efficient of range is obtained by applying the following formula

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

Where L-Largest value

S-Smallest value

1.Find range and co-efficient of range for the following observation

35,40,52,29,51,46,27,30,23

Solution

Largest value=52

Smallest value=23

Range=L-S

=52-23

Range=29

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

Co-efficient of Range= $\frac{52-23}{52+23}$

$= \frac{29}{75}$

Co-efficient of Range=0.3867

1.Find the range and coefficient of range for the following data

x	35-45	45-55	55-65	65-75	75-85
f	18	22	30	6	4

Solution

Largest value=85

Smallest value=35

Range=L-S

=85-35

Range=50

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

$$=\frac{85-35}{85+35}$$

$$=\frac{50}{120}$$

Co-efficient of Range= 0.4167

Merits and Demerits of Range

1. It is simple to compute and understand
2. It gives a rough and quick answer
3. It is used in statistical quality control
4. It is not based on all the observation
5. It is unaffected by all other items except the smallest and largest
6. It is affected by the presence of an extremely high or low item.

Quartile deviation (QD)

$$QD = \frac{Q_3 - Q_1}{2}$$

$$Q_1 = l_1 + \frac{\frac{N}{4} - m_1}{f_1} \times c_1$$

$$Q_3 = l_3 + \frac{\frac{3N}{4} - m_3}{f_3} \times c_3$$

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

Merits of QD

1.It is simple to understand

2.It is easy to compute

3.It is a useful measure when the extreme classes in a frequency distribution are not well-defined.

Demerits of QD

1.It is ignore the first 75% of the value and the last 75% of the last value

2.It is not based on all the observations

3.It is not suitable for mathematical treatment

1.Find QD and it’s co-efficient of QD for the following data

x	**0-10**	**10-20**	**20-30**	**30-40**	**40-50**	**50-60**	**60-70**
f	8	20	34	46	28	14	10

Solution

x	f	cf
0-10	8	8
10-20	20	28
20-30	34	62
30-40	46	108
40-50	28	136
50-60	14	150
60-70	10	160

N=160 ,

$\frac{N}{4}$=40,

$\frac{3N}{4} = 3(40)$

$=120$

$$Q_1 = l_1 + \frac{\frac{N}{4} - m_1}{f_1} \times c_1$$

$$= 20 + \frac{40 - 28}{34} \times 10$$

$$= 20 + \frac{12}{34} \times 10$$

$$= 20 + \frac{60}{17}$$

$$= 20 + 3.5294$$

$$Q_1 = 23.5294$$

$$Q_3 = l_3 + \frac{\frac{3N}{4} - m_3}{f_3} \times c_3$$

$$= 40 + \frac{120 - 108}{28} \times 10$$

$$= 40 + \frac{12}{28} \times 10$$

$$= 40 + 4.2857$$

$$Q_3 = 49.2857$$

$$QD = \frac{Q_3 - Q_1}{2}$$

$$QD = \frac{44.2857 - 23.5294}{2}$$

$$QD = 10.3782$$

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

$$= \frac{44.2857 - 23.5294}{44.2857 + 23.5294}$$

$$= \frac{20.7503}{67.8151}$$

Coefficient of QD=0.3061

2.Find QD and it's co-efficient of QD for the following data

x	0-10	10-20	20-30	30-40	40-50
f	13	26	32	44	55

Mean Deviation (MD)

Individual series

MD about mean$=\frac{\sum|x-\bar{X}|}{n}$

Where $\bar{X} = \frac{\sum x}{n}$, n=number of observation

1.Find MD about mean for the following observations

27,36,28,35,26,20,35,40,26,26

Solution

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

$$\bar{X} = \frac{27 + 36 + 28 + 35 + 26 + 20 + 35 + 40 + 26 + 26}{10}$$

$$= \frac{299}{10}$$

$$\bar{X} = 29.9$$

x	$x-\bar{X}$	$\lvert x-\bar{X}\rvert$
27	27-29.9=-2.9	2.9
36	36-29.9=6.1	6.1
28	28-29.9=-1.9	1.9
35	35-29.9=5.1	5.1
26	26-29.9=-3.9	3.9
20	20-29.9=-9.9	9.9
35	35-29.9=5.1	5.1
40	40-29.9=10.1	10.1
26	26-29.9=3.9	3.9
26	26-29.9=3.9	3.9
		$\sum \lvert x-\bar{X}\rvert = 52.8$

MD about mean $= \frac{\sum|x-\bar{X}|}{n}$

$= \frac{52.8}{10}$

MD about mean $= 5.28$

2.Find MD about mean for the following observations

18,20,12,14,19,22,26,16,19,24

Discrete frequency distribution

MD about mean$=\frac{\sum f\,|x-\bar{X}|}{N}$

Where $\bar{X} = \frac{\sum f\,x}{N}$

1.Find MD about mean for the following data

x	10	11	12	13	14
f	3	12	18	12	3

Solution

$$\bar{X} = \frac{\sum fx}{N}$$

$$= \frac{576}{48}$$

$$= 12$$

x	f	fx	$x - \bar{X}$	$\lvert x - \bar{X} \rvert$	$f\lvert x - \bar{X} \rvert$
10	3	30	10-12=-2	2	6
11	12	132	11-12=-1	1	12
12	18	216	12-12=0	0	0
13	12	136	13-12=1	1	12
14	3	42	14-12=2	2	6
	N=48	$\sum fx = 576$			$\sum f\lvert x - \bar{\boldsymbol{X}} \rvert = 36$

MD about mean $= \frac{\sum f\,|x-\bar{X}|}{N}$

$$= \frac{36}{48}$$

MD about mean $= 0.75$

2.Find MD about mean for the following data

x	2	4	6	8	10	12	14	16
f	2	2	4	5	3	2	1	1

Continuous frequency distribution

MD about mean $= \frac{\sum f\,|x - \bar{X}|}{N} \times C$

$$\bar{X} = A + \frac{\sum fd}{N} \times C$$

1.Find MD about mean for the following data

x	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
f	3	8	9	15	20	13	8	4

Solution

x	f	Mid x	$d = \frac{x - A}{C}$	fd	$x - \bar{X}$	$\lvert x - \bar{X} \rvert$	$f\lvert x - \bar{X} \rvert$
20-30	3	25	-3.5	-10.5	-36.5	36.5	109.5
30-40	8	35	-2.5	-20	-26.5	26.5	212
40-50	9	45	-1.5	-13.5	-16.5	16.5	148.5
50-60	15	55	-0.5	-7.5	-6.5	6.5	97.5
60-70	20	65	0.5	10	3.5	3.5	70
70-80	13	75	1.5	19.5	13.5	13.5	175.5
80-90	8	85	2.5	20	23.5	23.5	188
90-100	4	95	3.5	14	33.5	33.5	134
	N=80			$\sum fd = 12$			$\sum f\lvert x - \bar{\boldsymbol{X}} \rvert = 1135$

$$\bar{X} = \mathrm{A} + \frac{\sum fd}{N} \times \mathrm{C}$$

$$\bar{X} = 60 + \frac{12}{80} \times 10$$

$$= 60 + 1.5$$

$$\bar{X} = 61.5$$

$$\text{MD about mean} = \frac{\sum f\ |x - \bar{X}|}{N} \times C$$

$$= \frac{1135}{80} \times 10$$

$$= 14.1875$$

4.Find MD about mean for the following data

x	0-10	10-20	20-30	30-40	40-50	50-60	60-70
f	8	12	10	8	3	2	7

MD about Median

Individual data

MD about Median$=\frac{\sum|x-M|}{n}$

1. Find MD about median for the following observation

42,63,88,33,24,54,9,18,29,43

Solution

Ascending order 9,18,24,29,33,42,43,54,63,88

Median$=\frac{33+42}{2}$

$=37.5$

x	$x - M$	$\|x - M\|$
9	-28.5	28.5
18	-19.5	19.5
24	-13.5	13.5
29	-8.8	8.8
33	-4.5	4.5
42	4.5	4.5
43	5.5	5.5
54	16.5	16.5
63	25.5	25.5
88	50.5	50.5
		$\sum \|x - M\| = 177$

MD about Median $= \frac{\sum |x - M|}{n}$

MD about Median $= \frac{177}{10}$

MD about Median $= 17.7$

2. Find MD about median for the following observation27,36,28,18,35,26,20,35,40,26

3.Find MD about median for the following observation27,36,28,18,35,26,20,33,42,21

Discrete frequency distribution

MD about Median$=\frac{\sum f\ |x-\boldsymbol{M}|}{N}$

1.Find MD about Median for the following data

x	18	25	32	39	46	53	60
f	10	15	32	42	26	12	9

Solution

x	f	cf	$x - M$	$\lvert x - M \rvert$	$f\lvert x - M \rvert$
18	10	10	-21	21	210
25	15	25	-14	14	210
32	32	57	-7	7	224
39	42	99	0	0	0
46	26	125	7	7	182
53	12	137	14	14	168
60	9	146	21	21	189
	N=146				$\sum f \lvert x - \boldsymbol{M} \rvert = 1183$

$$\frac{N}{2} = \frac{146}{2} = 73$$

$$M = 39$$

MD about Median$=\frac{\sum f\ |x-\boldsymbol{M}|}{N}$

MD about Median$=\frac{1183}{146}$

MD about Median$=8.1027$

2.Find MD about Median for the following data

x	5	15	25	35	45	55	65	75
f	18	16	15	12	10	5	2	2

3.Find MD about Median for the following data

x	8	11	12	15	17	20
f	3	4	3	5	9	7

Continuous frequency distribution

MD about Median= $\frac{\sum f\ |x-\boldsymbol{M}|}{N} \times C$

M= Median= $l + \frac{\frac{N}{2}-m}{f} \times c$

N=total frequency

C=class interval

1.Find MD about Median for the following data

x	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
f	18	16	15	12	10	5	2	2

Solution

x	f	Cf
0-10	18	18
10-20	16	34
20-30	15	49
30-40	12	61
40-50	10	71
50-60	5	76
60-70	2	78
70-80	2	80

$$\frac{N}{2} = \frac{80}{2} = 4, m = 34, l = 20, c = 10$$

Median$= l + \frac{\frac{N}{2} - m}{f} \times c$

$= 20 + \frac{40-34}{15} \times 10$

$= 20 + \frac{6}{15} \times 10$

$= 20 + 4$

Median=24

x	f	$Mid\ x$	$x - M$	$\lvert x - M \rvert$	$f\lvert x - M \rvert$
0-10	18	5	-19	19	342
10-20	16	15	-9	9	144
20-30	15	25	1	1	15
30-40	12	35	11	11	132
40-50	10	45	21	21	210
50-60	5	55	31	31	155
60-70	2	65	41	41	82
70-80	2	75	51	51	102
	N=80				$\sum f\ \lvert x - \boldsymbol{M} \rvert = 1182$

MD about Median$=\frac{\sum f\,|x-\boldsymbol{M}|}{N}\times C$

$$=\frac{1182}{80}\times 10$$

MD about Median=147.75

2.Find MD about Median for the following data

x	3-7	8-12	13-17	18-22	23-27	28-32	32-38	38-42
f	12	18	35	42	50	45	20	8

3.Find MD about Median for the following data

x	0-10	10-20	20-30	30-40	40-50	50-60
f	8	7	11	13	15	6

MD about Mode

Individual data

MD about Mode$=\frac{\sum|x-z|}{n}$

1. Find MD about Mode for the following observation 3,7,2,7,5,7,3

Solution

Mode Z=7

x	$x - Z$	$\|x - Z\|$
3	-4	4
7	0	0
2	5	5
7	0	0
5	-2	2
7	0	0
3	-4	4
		$\sum \|x - Z\| = 15$

MD about Mode$=\frac{\sum|x-Z|}{n}$

$=\frac{15}{7}$

$=2.1429$

2.Find MD about Mode for the following observation

18,21,25,23,27,25,29,25,29,25

Discrete frequency distribution

MD about Mode$=\frac{\sum f\ |x-\boldsymbol{Z}|}{N}$

1.Find MD about Mode for the following data

x	10	11	12	13	14	15	16
f	8	4	12	24	26	7	11

Solution

Maximum frequency=26

Mode (Z)=14

x	cf	$x-Z$	$\lvert x-Z \rvert$	$f\lvert x-Z \rvert$
10	8	-4	4	32
11	4	-3	3	12
12	12	-2	2	24
13	24	-1	1	24
14	26	0	0	0
15	7	1	1	7
16	11	2	2	22
	N=92			$\sum f\lvert x-\boldsymbol{Z} \rvert = \mathbf{121}$

MD about Mode= $\frac{\sum f\ |x-\boldsymbol{Z}|}{N}$

MD about Mode= $\frac{121}{92}$

MD about Mode=1.3152

2.Find MD about Mode for the following data

x	58	60	61	62	63	64	65	66	68	70
f	4	6	5	10	20	22	24	6	2	1

Continuous frequency distribution

MD about Mode= $\frac{\sum f\ |x - \boldsymbol{Z}|}{N} \times C$

1.Find MD about Mode for the following data

x	0-10	10-20	20-30	30-40	40-50	50-60	60-70
f	4	9	13	15	12	8	3

Solution

x	f	$Mid\ x$	$x - Z$	$\lvert x - Z \rvert$	$f \lvert x - Z \rvert$
0-10	4	5	-29	29	116
10-20	9	15	-19	19	171
20-30	13	25	-9	9	117
30-40	15	35	1	1	15
40-50	12	45	11	11	132
50-60	8	55	21	21	168
60-70	3	65	31	31	93
	N=64				$\sum f \lvert x - \boldsymbol{Z} \rvert = \mathbf{812}$

$$\text{Mode } Z = l + \frac{f_1 - f_0}{2f_1 - (f_0 + f_2)} \times c$$

$$\text{Mode } Z = 30 + \frac{15 - 13}{2(15) - (13 + 12)} \times 10$$

$$= 30 + \frac{2}{30 - 25} \times 10$$

$$= 30 + \frac{2}{5} \times 10$$

$$= 30 + 4$$

$$\text{Mode } Z = 34$$

MD about Mode$=\frac{\sum f\ |x-\mathbf{Z}|}{N} \times C$

$=\frac{812}{64} \times 10$

$=12.6875 \times 10$

MD about Mode $= 126.875$

2. Find MD about Mode for the following data

70-80	80-90	90-100	100-110	110-120	120-130	130-140	140-150
12	18	35	42	50	45	20	8

Standard Deviation (SD)

It is defined as the positive square root of the arithmetic mean of the square of the deviation of the observations from their arithmetic mean. It is denoted by σ.

Individual series

$$\sigma = S.D = \sqrt{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2} \quad (small\ values)$$

Where n= number of observation

$$\sigma = S.D = \sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2} \quad (large\ values)$$

1.Find SD for the following observation
32,12,15,9,7,22,16,5,28,19
Solution

x	x^2
32	1024
12	144
15	225
9	81
7	49
22	484
16	256
5	25
28	784
19	361
$\sum x = 165$	$\sum x^2 = 3433$

$$\sigma = S.D = \sqrt{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2} \quad (small\ values)$$

$$\sigma = \sqrt{\frac{3433}{10} - \left(\frac{165}{10}\right)^2}$$

$$\sigma = \sqrt{343.3 - (16.5)^2}$$

$$\sigma = \sqrt{343.3 - 272.25}$$

$$\sigma = \sqrt{71.05}$$

$$\sigma = 8.4291$$

2.Find SD for the following observation 3,8,6,10,12,9,11,10,12,7

$(Ans\ \sigma = 2.7129)$

3.Find SD for the following observation 45,36,40,37,39,42,45,35,40,39

Solution

A=40; $d = x - A = x - 40$

x	$d = x - A$	d^2
45	5	25
36	-4	16
40	0	0
37	-3	9
39	-1	1
42	2	4
45	5	25
35	-5	25
40	0	0
39	-1	1
	$\sum d = -2$	$\sum d^2 = 106$

$$\sigma = S.D = \sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2}$$

$$\sigma = S.D = \sqrt{\frac{106}{10} - \left(\frac{-2}{10}\right)^2}$$

$$= \sqrt{10.6 - 0.04}$$

$$= \sqrt{10.56}$$

$$\sigma = 3.25$$

4.Find SD for the following observation 60,60,61,62,63,63,63,64,64,70

Discrete frequency distribution

$$\sigma = S.D = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}$$

Where, N=total frequency; $d = x - A$

1.Find SD for the following

Size	10	11	12	13	14	15	16	17
f	2	7	10	15	10	4	1	1

Solution

A=13.5

A=13.5

x	f	$d = x - A$	d^2	fd	fd^2
10	2	-3.5	3.5	-7	24.5
11	7	-2.5	2.5	-17.5	43.75
12	10	-1.5	1.5	-15	22.5
13	15	-0.5	0.5	-7.5	3.75
14	10	0.5	0.5	5	2.5
15	4	1.5	1.5	6	9
16	1	2.5	2.5	2.5	6.25
17	1	3.5	3.5	3.5	12.25
	N=50			$\sum fd = -30$	$\sum fd^2 = 124.5$

$$\sigma = S.D = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}$$

$$\sigma = S.D = \sqrt{\frac{124.5}{50} - \left(\frac{30}{50}\right)^2}$$

$$= \sqrt{2.49 - (0.6)^2}$$

$$\sigma = 1.4595$$

2. Find SD for the following data

x	1	2	3	4	5	6	7
f	8	12	18	28	16	10	8

Continuous frequency distribution

$$\sigma = S.D = \left(\sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}\right) \times C$$

1.Calculate SD for the following data

X	0-10	10-20	20-30-	30-40	40-50	50-60	60-70	70-80
f	5	10	20	40	30	20	10	4

Solution

A=40,C=10

x	f	Mid x	$d = \frac{x-A}{c}$	d^2	fd	fd^2
0-10	5	5	-3.5	12.25	-17.5	61.25
10-20	10	15	-2.5	6.25	-25	62.5
20-30	20	25	-1.5	2.25	-30	45
30-40	40	35	-0.5	0.25	-20	10
40-50	30	45	0.5	0.25	15	7.5
50-60	20	55	1.5	2.25	30	45
60-70	10	65	2.5	6.25	25	62.5
70-80	4	75	3.5	12.25	14	49
	N=139				$\sum fd = -8.5$	$\sum fd^2 = 342.75$

$$\sigma = S.D = \left(\sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}\right) \times C$$

$$\sigma = \left(\sqrt{\frac{342.75}{139} - \left(\frac{-8.5}{139}\right)^2}\right) \times 10$$

$$= \left(\sqrt{2.4658 - 0.0037}\right) \times 10$$

$$= 1.5691 \times 10$$

$$\sigma = 15.691$$

2. Calculate SD for the following data

x	10-15	15-20	20-25	25-30	30-35	35-40
f	2	8	20	35	20	15

Co-efficient of variation (C.V)

$$CV = \frac{\sigma}{\bar{x}} \times 100$$

Where $\sigma = S.D = \sqrt{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2}$

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

1.Calculate CV for the following data 25,32,28,34,24,31,36,27,29,30

x	x^2
25	625
32	1024
28	784
34	1156
24	576
31	961
36	1296
27	729
29	841
30	900
$\sum x = 296$	$\sum x^2 = 8892$

To find $\sigma = S.D = \sqrt{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2}$

$$= \sqrt{\frac{8892}{10} - \left(\frac{296}{10}\right)^2}$$

$$= \sqrt{889.2 - (296)^2}$$

$$= \sqrt{889.2 - 876.16}$$

$$= \sqrt{13.04}$$

$$\sigma = 3.6111$$

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

$$\bar{x} = \frac{296}{10}$$

$$\bar{x} = 29.6$$

$$CV = \frac{\sigma}{\bar{x}} \times 100$$

$$= \frac{3.6111}{29.6} \times 100$$

$$= 0.1219 \times 100$$

CV= 12.19

2.The Scores of two player A and B in rounds are given below

A	74	75	78	72	78	77	79	81	79	76	72	71
B	81	84	80	88	89	85	86	82	82	79	86	80

Solution

A=78 A=85.5

x_A	$d = x_A - A$	d^2	x_B	$d = x_B - A$	d^2
74	-4	16	81	-4.5	20.25
75	-3	9	84	-1.5	2.25
78	0	0	80	-5.5	30.25
72	-6	36	88	2.5	6.25
78	0	0	89	3.5	12.25
77	-1	1	85	-0.5	0.25
79	1	1	86	0.5	0.25
81	3	9	82	-3.5	12.25
79	1	1	82	-3.5	12.25
76	-2	4	79	-6.5	42.25
72	-6	36	86	0.5	0.25
71	-7	49	80	5.5	30.25
	$\sum d = -24$	$\sum d^2 = 162$		$\sum d = -24$	$\sum d^2 = 169$

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

$$= \sqrt{\frac{162}{12} - \left(\frac{-24}{12}\right)^2}$$

$$= \sqrt{13.5 - 4}$$

$$= \sqrt{9.5}$$

$$\sigma_A = 3.0822$$

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

$$= \sqrt{\frac{169}{12} - \left(\frac{-24}{12}\right)^2}$$

$$\sigma_B = 3.1754$$

$$\bar{x}_B = \frac{\sum x_A}{n}$$

$$= \frac{1002}{12}$$

$$\bar{x}_B = 83.5$$

$$\text{CV for A} = \frac{\sigma_A}{\bar{x}_A} \times 100$$

$$= \frac{3.0822}{76} \times 100$$

$$= 4.0555$$

$$\text{CV for B} = \frac{\sigma_B}{\bar{x}_B} \times 100$$

$$= \frac{3.1754}{83.5} \times 100$$

$$\text{CV for B} = 3.8029$$

As $\bar{x}_A < \bar{x}_B$

$$76 < 83.5$$

B is better

As CV for A> CV for B

$$4.0555 > 3.8029$$

A is consistent

3.Given below find which series is more consistent variable

Variable	10-20	20-30	30-40	40-50	50-60	60-70
Series A	10	16	30	40	26	18
Series B	22	18	32	34	18	16

Solution

	f_A	Mid	$d = \frac{x-A}{C}$	d^2	f_{Ad}	f_{Ad^2}	f_B	f_{Bd}	f_{Bd^2}
)	10	15	-2.5	6.25	-25	62.5	22	-55	137.5
)	16	25	-15	-2.25	-24	36	18	-22	40.5
)	30	35	0.5	0.25	-15	7.5	32	-16	8
)	40	45	0.5	0.25	20	10	34	17	8.5
)	26	55	1.5	2.25	39	58.5	18	27	40.5
)	18	65	2.5	6.25	75	112.5	16	40	100
	N=140				$\sum f_A d = 40$	$\sum f_A d^2 = 287$	N=140	$\sum f_B d = 14$	$\sum f_B d^2 = 287$

$$\sigma_A = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \times 10$$

$$= \sqrt{\frac{287}{140} - \left(\frac{40}{140}\right)^2} \times 10$$

$$= \sqrt{2.05 - (0.2857)^2} \times 10$$

$$= \sqrt{2.05 - 0.0816} \times 10$$

$$= \sqrt{1.9684} \times 10$$

$$\sigma_A = 14.03$$

$$\sigma_B = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \times 10$$

$$= \sqrt{\frac{335}{140} - \left(\frac{-14}{140}\right)^2} \times 10$$

$$= \sqrt{2.3929 - (0.1)^2} \times 10$$

$$= \sqrt{2.3929 - 0.01} \times 10$$

$$= \sqrt{2.3829} \times 10$$

$$= 1.5437 \times 10$$

$$\sigma_B = 15.437$$

$$\bar{X}_A = \mathrm{A} + \frac{\sum fd}{N} \times \mathrm{C}$$

$$= 40 + \frac{40}{140} \times 10$$

$$= 40 + 0.2857 \times 10$$

$$= 40 + 2.857$$

$$\bar{X}_A = 42.257$$

$$\bar{X}_B = \mathrm{A} + \frac{\sum fd}{N} \times \mathrm{C}$$

$$= 40 + \frac{(-14)}{140} \times 10$$

$$= 40 - 0.1 \times 10$$

$$= 40 - 1$$

$$\bar{X}_B = 39$$

$\text{CV for A} = \frac{\sigma_A}{\bar{x}_A} \times 100$

$$= \frac{14.03}{42.857} \times 100$$

$$= 0.3274 \times 100$$

$$= 32.74$$

$\text{CV for B} = \frac{\sigma_B}{\bar{x}_B} \times 100$

$$= \frac{15.437}{39} \times 100$$

$$= 0.3952 \times 100$$

$$\text{CV for B} = 39.58$$

$As\ \bar{X}_A > \bar{X}_B$

$42.857 > 39$

A is better

As CV for A< CV for B

$32.74 < 39.58$

B is consistent

Home work

3.Compare the variability of the two varieties

Length of life	No.of bulbs	
	A	B
500-700	5	4
700-900	11	30
900-1100	26	12
1100-1300	10	8
1300-1500	8	6

4.The following table gives the frequency distribution of expenditure on education per family per month among middle class of families in two towns.

Expenditure	Town A	Town B
3-6	28	39
6-9	292	284
9-12	389	401
12-15	212	202
15-18	59	48
18-21	18	21
21-24	2	5

Merits and Demerits for SD

(i) It is well defined and suitable for algebraic treatment

(ii) It is based on all the observation

(iii)It is the most commonly used measure of dispersion

(iv)It is less affected by sampling fluctuations

(v) It is useful for comparing the variability of two distribution

(vi) It is not easy to understand

Combined Mean and SD

The mean and SD of two groups are given in the following table

Group	Mean	SD	Size
I	$\overline{x_1}$	σ_1	n_1
II	$\overline{x_2}$	σ_2	n_2

Let $\bar{x}$ and σ be the mean and SD of the combined group of (n_1 and n_2)

Then $\bar{x}$ *and* σ are determined by the formula

$$\bar{x} = \frac{n_1\overline{x_1} + n_2\overline{x_2}}{n_1 + n_2}$$

$$\sigma^2 = \frac{n_1\sigma_1^2 + n_2\sigma_2^2 + n_1 d_1^2 + n_2 d_2^2}{n_1 + n_2}$$

Where

$d_1 = \overline{x_1} - \bar{x}, d_2 = \overline{x_2} - \bar{x},$

1.The following table gives the result of three centres of a public examination

Centre	No.of Candidates	SD	Mean Score
A	200	3	75
B	250	4	10
C	300	5	15

Solution

$n_1 = 200 \quad \sigma_1 = 3 \quad \overline{x_1} = 75$

$n_2 = 250 \quad \sigma_2 = 4 \quad \overline{x_2} = 10$

$n_3 = 300 \quad \sigma_3 = 5 \quad \overline{x_1} = 15$

$$\bar{x} = \frac{n_1\overline{x_1} + n_2\overline{x_2} + n_3\overline{x_3}}{n_1 + n_2 + n_3}$$

$$\bar{x} = \frac{(200 \times 75) + (250 \times 10) + (300 \times 15)}{200 + 250 + 300}$$

$$= \frac{15000 + 2500 + 4500}{750}$$

$$= \frac{22000}{750}$$

$$\bar{x} = 29.33$$

$$d_1 = \overline{x_1} - \bar{x},$$

$$= 75 - 29.33$$

$$= 45.67$$

$$d_2 = \overline{x_2} - \bar{x},$$

$$= 10 - 29.33$$

$$= -19.33$$

$$d_3 = \overline{x_3} - \bar{x},$$

$$= 15 - 29.33$$

$$= -14.33$$

$$\sigma^2 = \frac{n_1\sigma_1^2 + n_2\sigma_2^2 + n_3\sigma_3^2 + n_1 d_1^2 + n_2 d_2^2 + +n_3 d_3^2}{n_1 + n_2 + n_3}$$

$$= \frac{\begin{array}{c}(200 \times 9) + (250 \times 16) + (300 \times 25) + (200 \times 2085.7489) \\ +(250 \times 373.6489) + (300 \times 205.3489) + +n_3 d_3^2\end{array}}{200 + 250 + 300}$$

$$= \frac{585866.675}{750}$$

$$\sigma^2 = 781.1556$$

$$\sigma = 27.9492$$

Om Sakthi
ADHIPARASAKTHI COLLEGE OF ARTS AND SCIENCE
(Autonomous)
G.B. NAGAR, KALAVAI – 632 506, RANIPET DISTRICT, TAMILNADU
Permanently Affiliated to Thiruvalluvar University,
Re-accredited by NAAC – 'B' Grade (CGPA - 2.83)

STATISTICAL METHODS AND THEIR APPLICATIONS-I

II B.SC Computer Science

Handled by

Dr.P.RAJAKUMARI

ASSISTANT PROFESSOR

DEPARTMENT OF MATHEMATICS

ADHIPARASAKTHI COLLEGE OF ARTS AND SCIENCE

AUTONOMOUS (AUTONOMOUS)

UNIT IV

Measure of Skewness

Measure of Skewness is the statistical technique to indicate the direction and extend of skewness is the distribution of numerical values in the data

i) Karl pearson's co-efficient of skewness=$\frac{Mean - Mode}{SD}$

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

ii) Karl pearson's co-efficient of skewness=$\frac{3(Mean - Median)}{SD}$

$$= \frac{3(\bar{X} - M)}{\sigma}$$

iii) Bowley's co-efficient of skewness=$\frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1}$

1.Find the karl pearson's co-efficient o skewness for the following data

X	0-20	20-40	40-60	60-80	80-100	100-120
f	20	50	59	30	25	16

Solution

X	f	$Mid\ x$	$d = \frac{x-A}{C}$	d^2	fd	fd^2
0-20	20	10	-2.5	6.25	-50	125
20-40	50	30	-1.5	2.25	-75	112.5
40-60	59	50	-0.5	0.25	-29.5	147.5
60-80	30	70	0.5	0.25	15	7.5
80-100	25	90	1.5	2.25	37.5	56.25
100-120	16	110	2.5	6.25	40	100
	N=200				$\sum fd = -62$	$\sum fd^2 = 416$

$$\bar{X} = \mathrm{A} + \frac{\sum fd}{N} \times \mathrm{C}$$

A=60, **C=20,** **N=200,** $\boldsymbol{\sum fd = -62}$

$$\bar{X} = 60 + \frac{-62}{200} \times 20$$

$$= 60 - 0.31 \times 20$$

$$= 60 - 6.2$$

$$\bar{X} = 53.8$$

$$\text{Mode } Z = l + \frac{f_1 - f_0}{2f_1 - (f_0 + f_2)} \times c$$

l=40, $f_0 = 50, f_1 = 59, f_2 = 30, C = 20$

$$= 40 + \frac{59 - 50}{2(59) - (50 + 30)} \times 20$$

$$= 40 + \frac{9}{118 - 80} \times 20$$

$$= 40 + \frac{9}{38} \times 20$$

$$= 40 + 0.2368 \times 20$$

$$= 40 + 4.736$$

$$Z = 44.736$$

$$\sigma = S.D = \left(\sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}\right) \times C$$

$$= \left(\sqrt{\frac{416}{200} - \left(\frac{-62}{200}\right)^2}\right) \times 20$$

$$= \left(\sqrt{2.08 - (-0.31)^2}\right) \times 20$$

$$= \left(\sqrt{2.08 - 0.0961}\right) \times 20$$

$$= \left(\sqrt{1.9839}\right) \times 20$$

$$= 1.4085 \times 20$$

$$\sigma = 28.17$$

Karl pearson's co-efficient of skewness=$\frac{Mean-Mode}{SD}$

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

$$= \frac{53.8 - 44.736}{28.17}$$

$$= \frac{9.064}{28.17}$$

$$= 0.3218$$

Home work

2.Find the karl pearson's co-efficient of skewness for the following data

X	0-10	10-20	10-30	30-40	40-50	50-60	60-70
f	10	15	24	25	10	10	6

3.Find the Bowley's co-efficient of skewness for the following data

x	0-9	10-19	20-29	30-39	40-49
f	8	15	23	16	9

Solution

x	f	Cf
0.5-9.5	8	8
9.5-19.5	15	23
19.5-29.5	23	46
29.5-39.5	16	62
39.5-49.5	9	71
	N=71	

$$\frac{N}{2} = 35.5$$

$$\text{Median} = l + \frac{\frac{N}{2} - m}{f} \times c$$

$$= 19.5 + \frac{35.5 - 33}{23} \times 10$$

$$= 19.5 + \frac{12.5}{23} \times 10$$

$$= 19.5 + 0.5435 \times 10$$

$$= 19.5 + 5.435$$

$$\text{Median} = 24.935$$

$$Q_1 = l_1 + \frac{\frac{N}{4} - m_1}{f_1} \times c_1 \ ; \frac{N}{4} = 17.75$$

$$= 9.5 + \frac{17.75 - 8}{15} \times 10$$

$$= 9.5 + \frac{9.75}{15} \times 10$$

$$= 9.5 + 0.65 \times 10$$

$$= 9.5 + 6.5$$

$$Q_1 = 16$$

$$Q_3 = l_3 + \frac{\frac{3N}{4} - m_3}{f_3} \times c_3$$

$$\frac{3N}{4} = 53.25$$

$$= 29.5 + \frac{53.25 - 46}{16} \times 10$$

$$= 29.5 + \frac{7.25}{16} \times 10$$

$$= 29.5 + 0.4531 \times 10$$

$$= 29.5 + 4.531$$

$$Q_3 = 34.031$$

Bowley's co-efficient of skewness $= \frac{Q_3+Q_1-2M}{Q_3-Q_1}$

$$= \frac{34.031 + 16 - 2(24.935)}{34.031 - 16}$$

$$= \frac{50.031 - 49.869}{18.031}$$

$$= \frac{0.162}{18.031}$$

Bowley's co-efficient of skewness $= 0.0089$

3.Find the Quartile co-efficient of skewness for the following data

X	55-58	58-61	61-64	64-67	67-70
Group A	12	17	23	18	11
Group B	20	22	25	13	7

Solution

Group A

x	f	cf
55-58	12	12
58-61	17	29
61-64	23	52
64-67	18	70
67-70	11	81
	N=81	

Median= $l + \frac{\frac{N}{2} - m}{f} \times c$

$\frac{N}{2} = 40.5$

Median= $61 + \frac{40.5 - 29}{23} \times 3$

$= 61 + \frac{11.5}{23} \times 3$

$= 61 + 0.5 \times 3$

$=61 + 1.5$

M=62.5

$$Q_1 = l_1 + \frac{\frac{N}{4} - m_1}{f_1} \times c_1 \; ; \frac{N}{4} = 20.25$$

$$= 58 + \frac{20.25 - 12}{17} \times 3$$

$$= 58 + \frac{8.25}{17} \times 3$$

$$= 58 + 0.4853 \times 3$$

$$= 58 + 1.4559$$

$$Q_1 = 59.4559$$

$$Q_3 = l_3 + \frac{\frac{3N}{4} - m_3}{f_3} \times c_3$$

$$\frac{3N}{4} = 20.25 \times 3$$

$$\frac{3N}{4} = 60.75$$

$$= 64 + \frac{60.75 - 52}{18} \times 3$$

$$= 64 + \frac{8.75}{18} \times 3$$

$$= 64 + 0.48611 \times 3$$

$$= 64 + 1.4583$$

$$Q_3 = 65.4583$$

$$\text{Bowley's co} - \text{efficient of skewness} = \frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1}$$

$$= \frac{65.4583 + 59.4559 - 2(62.5)}{65.4583 - 59.4559}$$

$$= \frac{124.9142 - 125}{6.0024}$$

$$= \frac{-0.0858}{6.0024}$$

$$\text{Bowley's co} - \text{efficient of skewness} = -0.0143$$

Group B

x	f	cf
55-58	20	20
58-61	22	42
61-64	25	87
64-67	13	100
67-70	7	107
	N=107	

$$\text{Median} = l + \frac{\frac{N}{2} - m}{f} \times c$$

$$\frac{N}{2} = \frac{107}{2} = 43.5$$

$$\text{Median} = 61 + \frac{43.5 - 42}{25} \times 3$$

$$= 61 + \frac{1.5}{25} \times 3$$

$$= 61 + 0.06 \times 3$$

$$= 61 + 0.18$$

$$M = 61.18$$

$$Q_1 = l_1 + \frac{\frac{N}{4} - m_1}{f_1} \times c_1 \ ; \frac{N}{4} = 21.75$$

$$= 58 + \frac{21.75 - 20}{22} \times 3$$

$$= 58 + \frac{1.75}{22} \times 3$$

$$= 58 + 0.0795 \times 3$$

$$= 58 + 0.2385$$

$$Q_1 = 58.2385$$

$$Q_3 = l_3 + \frac{\frac{3N}{4} - m_3}{f_3} \times c_3$$

$$\frac{3N}{4} = 21.75 \times 3$$

$$\frac{3N}{4} = 65.25$$

$$= 61 + \frac{65.25 - 42}{25} \times 3$$

$$= 61 + \frac{23.25}{25} \times 3$$

$$= 61 + 0.093 \times 3$$

$$= 61 + 2.79$$

$$Q_3 = 63.79$$

$$\text{Bowley's co}-\text{efficient of skewness} = \frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1}$$

$$= \frac{63.79 + 58.2385 - 2(61.18)}{63.79 - 58.2385}$$

$$= \frac{122.0285 - 122.36}{5.5515}$$

$$= \frac{-0.3315}{5.5515}$$

$$\text{Bowley's co}-\text{efficient of skewness} = -0.0597$$

A is more skewed

Home work

4.Find the Bowley's co-efficient of skewness for the following data

Mid value	21	27	33	39	45	51	57
f	18	22	40	50	38	12	4

Solution

x	F	Cf
18-24	18	18
24-30	22	40
30-36	40	80
36-42	50	130
42-48	38	168
48-54	12	180
54-60	4	184

5.For a distribution Bowley's co-efficient of Skewness is -0.36, Lower quartile is 8.6 and Median is 12.3. What is its quartile co-efficient of dispersion?

Solution

Bowley's co-efficient of Skewness = -0.36, $Q_1 = 8.6, M = 12.3$

$$\textbf{Bowley's co-efficient of skewness} = \frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1}$$

$$-0.36 = \frac{Q_3 + 8.6 - 2(12.3)}{Q_3 - 8.6}$$

$$-0.36\,(Q_3 - 8.6) = Q_3 + 8.6 - 24.6$$

$$-0.36\,Q_3 + 3.096 = Q_3 - 16$$

$$3.096 + 16 = Q_3 + 0.36\,Q_3$$

$$19.096 = 1.36\,Q_3$$

$$\frac{19.096}{1.36} = Q_3$$

$$Q_3 = 14.0412$$

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

$$= \frac{14.0412 - 8.6}{14.0412 + 8.6}$$

$$= \frac{5.4412}{22.6412}$$

Quartile Coefficient of dispersion $= 0.2403$

5.In a distribution mean=65, median=70 and co-efficient of skewness=-0.6.Find (i) Mode (ii) Coefficient of variation

Solution

skewness=-0.6

mean $\bar{X}$ =65

median (M)=70

$$Mean - Mode = 3(Mean - Median)$$

$$65 - Mode = 3(65 - 70)$$

$$= 3(-5)$$

$$= -15$$

$$65 + 15 = Mode$$

$$Mode = 80$$

Karl pearson's co-efficient of skewness$= \frac{3(Mean - Median)}{SD}$

$$= \frac{3(\bar{X} - M)}{\sigma}$$

$$-0.6 = \frac{3(65 - 70)}{\sigma}$$

$$\sigma(-0.6) = -15$$

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

$$\sigma = 25$$

$$CV = \frac{\sigma}{\bar{x}} \times 100$$

$$= \frac{25}{65} \times 100$$

$$= 0.3846 \times 100$$

$$CV = 38.46$$

6.In a distribution the sum of two quartiles is 78.2 and their difference is 14.3 and if its median is 35.7.Find the Bowley's co-efficient of skewness.

Solution

$$Q_3 + Q_1 = 78.2, \qquad Q_3 - Q_1 = 14.3 \qquad Median = 35.7$$

$$\textbf{Bowley's co} - \textbf{efficient of skewness} = \frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1}$$

$$= \frac{78.2 - 2(35.7)}{14.3}$$

$$= \frac{78.2 - 71.4}{14.3}$$

$$= \frac{6.8}{14.3}$$

$$= 0.4755$$

7.In a distribution, Pearson's co-effficient of skewness is -0.7 and the value of the median and standard deviation are 12.8 and 6 respectively. Find mean.

Solution

Pearson's co-effficient of skewness = -0.7

Median=12.8

standard deviation=6

Karl pearson's co-efficient of skewness= $\frac{3(Mean-Median)}{SD}$

$$= \frac{3(\bar{X} - M)}{\sigma}$$

$$-0.7 = \frac{3(Mean - 12.8)}{6}$$

$$6 \times (-0.7) = 3(Mean - 12.8)$$

$$\frac{-4.2}{3} = Mean - 12.8$$

$$-1.4 = Mean - 12.8$$

$$12.8 - 64 = Mean$$

$$Mean = 11.4$$

Kurtosis

Definition of kurtosis

Kurtosis is a measure of flatness or peakness of a distribution

1.Definition of Platykurtic

The curve which is more plateopped than the normal curve is called Platykurtic.

2.Definition of Mesokurtic

The normal curve (or bell-shaped curves) is called mesokurtic.

3.Definition of Leptokurtic

The curve which is more peaked than the normal curve is called Leptokurtic.

Measure of Kurtosis

$$\gamma_2 = \frac{\mu_4}{\mu_2^3} - 3 \quad or\ \beta_2 = \frac{\mu^4}{\mu_2^2}$$

For the normal distribution $\gamma_2 = 0\ and\ \beta_2 = 3$

If $\gamma_2 = 0$ **the curve is called Mesokurtic**

If $\gamma_2 < 0$ **the curve is called Platykurtic**

If $\gamma_2 > 0$ **the curve is called Leptokurtic**

If $\beta_2 = 3$ **the curve is called Mesokurtic**

If $\beta_2 > 3$ **the curve is called Leptokurtic**

If $\beta_2 < 3$ **the curve is called Platykurtic**

Measure of Skewness

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

1.The first four central moments of a distribution are 0,2.5,0.7 and 18.75.Test the skewness and kurtosis of the distribution.

Solution

$$\mu_1 = 0, \mu_2 = 2.5, \mu_3 = 0.7, \mu_4 = 18.75$$

Skewness

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

$$= \frac{(0.7)^2}{(2.5)^2}$$

$$= 0.031$$

Since β_1 is positive, the distribution is positively skewed.

Kurtosis

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

$$= \frac{18.75}{(2.5)^2}$$

$$= \frac{18.75}{6.25}$$

$$\beta_2 = 3$$

Since $\beta_2 = 3$ the distribution is normal.

2.The first four central moments of a distribution are 0,6,12 and 120.Test the skewness and kurtosis

Solution

$$\mu_1 = 0, \mu_2 = 6, \mu_3 = 12, \mu_4 = 120$$

Skewness

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

$$= \frac{(12)^2}{(6)^2}$$

$$= \frac{144}{216}$$

$$= 0.6666$$

Since β_1 is positive, the distribution is positively skewed.

Kurtosis

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

$$= \frac{120}{(6)^2}$$

$$= \frac{120}{36}$$

$$\beta_2 = 3.3333$$

Since $\beta_2 > 3$ the distribution is Leptokurtic.

The first four Moments

$$\mu_1 = 0$$

$$\mu_2 = \mu_2' - \mu_1'^2$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2{\mu_1'}^3$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'{\mu_1'}^2 - 3{\mu_1'}^4$$

1.The first four moments of a distribution about the value 5 are 2,20,40 and 150.Find the measure of kurtosis

Solution

$$\mu_1' = 2, \quad \mu_2' \quad = 20, \quad \mu_3' = 40, \quad \mu_4' - 150$$

$$\mu_1 = 0$$

$$\mu_2 = \mu_2' - \mu_1'^2$$

$$= 20 - (2)^2$$

$$= 20 - 4$$

$$\mu_2 = 16$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^{\,3}$$

$$= 40 - (3 \times 20 \times 2) + 2(2)^3$$

$$= 40 - 120 + 16$$

$$= 56 - 120$$

$$\mu_3 = -64$$

$$\mu_4 = \mu'_4 - 4\mu'_3\mu'_1 + 6\mu'_2{\mu'_1}^2 - 3{\mu'_1}^4$$

$$= 150 - (4 \times 40 \times 2) + (6 \times 20 \times 4) - (3 \times 16)$$

$$= 150 - 320 + 480 - 48$$

$$= 630 - 368$$

$$\mu_4 = 262$$

Skewness

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

$$= \frac{4096}{4096}$$

$$\beta_1 = 1$$

Kurtosis

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

$$= \frac{262}{(16)^2}$$

$$= \frac{262}{256}$$

$$\beta_2 = 1.0234$$

Home Work

2.The first four moments of a distribution about the value 5 are 7,70,140 and 175.Find the measure of skewness and kurtosis

Moments

Individual series

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

$$\mu_1 = \frac{\sum(x - \bar{x})}{n}$$

$$\mu_2 = \frac{\sum(x - \bar{x})^2}{n}$$

$$\mu_3 = \frac{\sum(x - \bar{x})^3}{n}$$

$$\mu_4 = \frac{\sum(x - \bar{x})^4}{n}$$

1.Compute the first four central moments for the following data 8,10,11,12,14

Solution

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

$$= \frac{55}{5}$$

$$\bar{X} = 11$$

x	$x-\overline{x}$	$(x-\overline{x})^2$	$(x-\overline{x})^3$	$(x-\overline{x})^4$
8	-3	9	-27	81
10	-1	1	-1	1
11	0	0	0	0
12	1	1	1	1
14	3	9	27	81
	$\sum(x-\overline{x})=0$	$\sum(x-\overline{x})^2=20$	$\sum(x-\overline{x})^3=0$	$\sum(x-\overline{x})^4=164$

The four central moments are $\mu_1, \mu_2, \mu_3, \mu_4$

$$\mu_1 = \frac{\sum(x-\overline{x})}{n}$$

$$\mu_1 = 0$$

$$\mu_2 = \frac{\sum(x - \overline{x})^2}{n}$$

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

$$= 4$$

$$\mu_3 = \frac{\sum(x - \overline{x})^3}{n}$$

$$= 0$$

$$\mu_4 = \frac{\sum(x - \overline{x})^4}{n}$$

$$= \frac{164}{5}$$

$$= 32.8$$

2.Find the first moment about 10 for the following data 8,10,11,12,14

Solution

x	$d = x - \overline{x}$	$d^2 = (x - \overline{x})^2$	$d^3 = (x - \overline{x})^3$	$d^4 = (x - \overline{x})^4$
8	-2	4	-8	16
10	0	0	0	0
11	1	1	1	1
12	2	4	8	16
14	4	16	64	256
	$\sum(x - \overline{x}) = 5$	$\sum(x - \overline{x})^2 = 25$	$\sum(x - \overline{x})^3 = 65$	$\sum(x - \overline{x})^4 = 289$

The four central moments are $\mu_1, \mu_2, \mu_3, \mu_4$

$$\mu_1 = \frac{\sum(x - \bar{x})}{n}$$

$$\mu_1 = \frac{5}{5}$$

$$= 1$$

$$\mu_2 = \frac{\sum(x - \bar{x})^2}{n}$$

$$= \frac{25}{5}$$

$$= 5$$

$$\mu_3 = \frac{\sum(x-\overline{x})^3}{n}$$

$$= \frac{65}{5}$$

$$= 13$$

$$\mu_4 = \frac{\sum(x-\overline{x})^4}{n}$$

$$= \frac{289}{5}$$

$$= 57.8$$

Discrete frequency distribution

$$\bar{X} = \frac{\sum fx}{N}$$

$$\mu_1 = \frac{\sum f(x - \bar{x})}{N}$$

$$\mu_2 = \frac{\sum f(x - \bar{x})^2}{N}$$

$$\mu_3 = \frac{\sum f(x - \bar{x})^3}{N}$$

$$\mu_4 = \frac{\sum f\,(x - \bar{x})^4}{N}$$

1.Find the first central moments for the following data

Solution

X	0	1	2	3	4	5	6	7	8
f	1	8	28	56	70	56	28	8	1

x	f	fx	$x-\bar{x}$	$(x-\bar{x})^2$	$(x-\bar{x})^3$	$(x-\bar{x})$	$(f(x-\bar{x})$	$f(x-\bar{x})^2$	$f(x-\bar{x})^3$	$f(x-\bar{x})^4$
0	1	0	-4	16	-64	256	-4	16	-64	256
1	8	8	-3	9	-27	81	-24	72	-216	648
2	28	56	-2	4	-8	16	-56	112	-224	448
3	56	168	-1	1	-1	1	-56	56	-56	56
4	70	280	0	0	0	0	0	0	0	0
5	56	280	1	1	1	1	56	56	56	56
6	28	168	2	4	8	16	56	112	224	448
7	8	56	3	9	27	81	24	72	216	648
8	1	8	4	16	64	256	4	16	64	236
	N= 256	$\sum fx = 1024$					$\sum f(x-\bar{x}) = 0$	$\sum f(x-\bar{x})^2 = 512$	$\sum f(x-\bar{x})^3 = 0$	$\sum f(x-\bar{x})^4 = 2816$

$$\bar{X} = \frac{\sum fx}{N}$$

$$= \frac{1024}{256}$$

$$= 4$$

$$\mu_1 = \frac{\sum f(x - \bar{x})}{N}$$

$$= \frac{0}{256}$$

$$= 0$$

$$\mu_2 = \frac{\sum f(x - \bar{x})^2}{N}$$

$$= \frac{512}{256}$$

$$= 2$$

$$\mu_3 = \frac{\sum f(x-\overline{x})^3}{N}$$

$$= \frac{0}{256}$$

$$= 0$$

$$\mu_4 = \frac{\sum f\,(x-\overline{x})^4}{N}$$

$$= \frac{2816}{256}$$

$$= 11$$

Home work

2.Find the first central moments for the following data

x	2	3	4	5	6
f	1	3	7	3	1

Continuous frequency distribution

$$\mu_1 = \frac{\sum fd}{N} \times c$$

$$\mu_2 = \frac{\sum fd^2}{N} \times c^2$$

$$\mu_3 = \frac{\sum fd^3}{N} \times c^3$$

$$\mu_4 = \frac{\sum f\, d^4}{N} \times c^4$$

1.Calculate the first four central moments for the following data

Mark less than	80	70	60	50	40	30	20	10
F	10	10	20	28	12	7	8	5

Solution

A=4

x	$Mid\ x$	$d = \frac{x-A}{c}$	d^2	d^3	d^4	f	fd	fd^2	fd^3	fd^4
0-10	**5**	**-3.5**	**12.25**	**-42.875**	**150.0625**	**5**	**-17.5**	**61.25**	**-214.375**	**750.3125**
10-20	**15**	**-2.5**	**6.25**	**-15.625**	**39.0625**	**8**	**-20**	**50**	**-125**	**312.5**
20-30	**25**	**-1.5**	**2.25**	**-3.375**	**5.0625**	**7**	**-10.5**	**15.75**	**-28.625**	**35.4375**
30-40	**35**	**-0.5**	**0.25**	**-0.125**	**0.0625**	**12**	**-6**	**3**	**-1.5**	**0.75**
40-50	**45**	**0.5**	**0.25**	**0.125**	**0.0625**	**28**	**14**	**7**	**3.5**	**1.75**
50-60	**55**	**1.5**	**2.25**	**3.375**	**5.0625**	**20**	**30**	**45**	**67.5**	**101.25**
60-70	**65**	**2.5**	**6.25**	**15.625**	**39.0625**	**10**	**25**	**62.5**	**156.25**	**390.625**
70-80	**75**	**3.5**	**12.25**	**42.875**	**150.0625**	**10**	**35**	**122.5**	**428.75**	**1500.625**
						100	**50**	**367**	**291.5**	**3093.25**

$$\sum fd = 50\ ; \sum fd^2 = 367; \sum fd^3 = 291.5\ ; \sum fd^4 = 3093.25$$

N=100 c=10

$$\mu_1 = \frac{\sum fd}{N} \times c$$

$$= \frac{50}{100} \times 10$$

$$= 0.5 \times 10$$

$$\mu_1 = 5$$

$$\mu_2 = \frac{\sum fd^2}{N} \times c^2$$

$$= \frac{367}{100} \times 10^2$$

$$= 3.67 \times 100$$

$$\mu_2 = 367$$

$$\mu_3 = \frac{\sum f d^3}{N} \times c^3$$

$$= \frac{291.5}{100} \times 10^3$$

$$\mu_3 = 2.95$$

$$\mu_4 = \frac{\sum f\, d^4}{N} \times c^4$$

$$= \frac{3093.25}{100} \times 10^4$$

$$= 30.6325 \times 10000$$

$$\mu_4 = 306325$$

Skewness

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

$$= \frac{(2.915)^2}{(367)^3}$$

$$= \frac{8.4972}{49430863}$$

$$=0.00000017$$

Kurtosis

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

$$= \frac{306325}{(367)^2}$$

$$= \frac{306325}{134689}$$

$$\beta_2 = 2.2743$$

1.Calculate the first four central moments for the following data

X	4-8	8-12	12-16	16-20	20-24	24-28	28-32	32-36	36-40	40-44
f	10	12	16	14	10	8	17	5	4	4

Skewness and kurtosis for continuous frequency distribution

1.Calculate skewness and kurtosis for the following data

Class	0-10	10-20	20-30	30-40	40-50
Frequency	10	20	40	20	10

Solution

A=25, C=10

$Mid\ x$	$d = \frac{x-A}{c}$	d^2	d^3	d^4	f	fd	fd^2	fd^3	fd^4
5	-2	4	-8	16	10	-20	40	-80	160
15	-1	1	-1	1	20	20	20	-20	20
25	0	0	0	0	40	0	0	0	0
35	1	1	1	1	20	20	20	20	20
45	2	4	8	16	10	40	40	80	160
					N=100	$\sum fd = 0$	$\sum fd^2 = 120$	$\sum fd^3 = 0$	$\sum fd^4 = 360$

$\sum fd = 0\,; \sum fd^2 = 120; \sum fd^3 = 0\,; \sum fd^4 = 360$

N=100 c=10

$$\mu_1{}' = \frac{\sum fd}{N} \times c$$

$$\mu_1{}' = 0$$

$$\mu_2{}' = \frac{\sum fd^2}{N} \times c^2$$

$$= \frac{120}{100} \times 10$$

$$\mu_2{}' = 120$$

$$\mu_3{}' = \frac{\sum fd^3}{N} \times c^3$$

$$\mu_3{}' = 0$$

$\sum fd = 0$; $\sum fd^2 = 120$; $\sum fd^3 = 0$; $\sum fd^4 = 360$

N=100 c=10

$$\mu_1' = \frac{\sum fd}{N} \times c$$

$$\mu_1' = 0$$

$$\mu_2' = \frac{\sum fd^2}{N} \times c^2$$

$$= \frac{120}{100} \times 10$$

$$\mu_2' = 120$$

$$\mu_3' = \frac{\sum fd^3}{N} \times c^3$$

$$\mu_3' = 0$$

$$\mu_4{}' = \frac{\sum f\, d^4}{N} \times c^4$$

$$= \frac{360}{100} \times 10^4$$

$$\mu_4{}' = 36000$$

$$\mu_1' = 0, \quad \mu_2' \quad = 120, \quad \mu_3' = 0, \quad \mu_4' = 36000$$

$$\mu_1 = 0$$

$$\mu_2 = \mu_2' - \mu_1'^2$$

$$= 120 - 0$$

$$= 120$$

$$\mu_2 = 16$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3$$

$$= 0 - (3 \times 120 \times 0) + 0$$

$$\mu_3 = 0$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'{\mu_1'}^2 - 3{\mu_1'}^4$$

$$= 36000 - (4 \times 0 \times 0) + (6 \times 120 \times 0) - (3 \times 0)$$

$$= 36000 - 0 + 0 - 0$$

$$= 36000$$

Skewness

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

$$= \frac{0}{120}$$

$$= 0$$

Kurtosis

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

$$= \frac{36000}{(120)^2}$$

$$= \frac{36000}{14400}$$

$$\beta_2 = 2.5$$

2.Calculate skewness and kurtosis for the following data

x	0.5-5.5	5.5-10.5	10.5-15.5	15.5-20.5	20.5-25.5	25.5-30.5	30.5-35.5
f	3	4	68	30	10	6	2

Om Sakthi
ADHIPARASAKTHI COLLEGE OF ARTS AND SCIENCE
(Autonomous)
G.B. NAGAR, KALAVAI – 632 506, RANIPET DISTRICT, TAMILNADU
Permanently Affiliated to Thiruvalluvar University,
Re-accredited by NAAC – 'B' Grade (CGPA - 2.83)

STATISTICAL METHODS AND THEIR APPLICATIONS-I

II B.SC Computer Science

Handled by

Dr.P.RAJAKUMARI
ASSISTANT PROFESSOR
DEPARTMENT OF MATHEMATICS
ADHIPARASAKTHI COLLEGE OF ARTS AND SCIENCE
AUTONOMOUS (AUTONOMOUS)

UNIT V

CORRELATION

Definition

The term of correlation refers to the relationship between two variables.

Formula

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}} \quad -1 \leq r \leq 1$$

This formula is used when deviations are measured from their mean

Method-1

1.Calculate the co-efficient of correlation between x and y from the following data.

x	1	3	5	8	9	10
y	3	4	8	10	12	11

Solution

$$r = \frac{\sum(x-\bar{x})\sum(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}}$$

$$\bar{x} = \frac{\sum x}{n} = \frac{36}{6} = 6$$

$$\bar{y} = \frac{\sum y}{n} = \frac{48}{6} = 8$$

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$\sum(x-\bar{x})\sum(y-\bar{y})$
1	3	-5	-5	25	25	25
3	4	-3	-4	9	16	12
5	8	-1	0	1	0	0
8	10	2	2	4	4	4
4	12	3	4	9	16	12
10	11	4	3	16	9	12
		$\sum(x-\bar{x}$ $=0$	$\sum(y-\bar{y})$ $=0$	$\sum(x-\bar{x})^2 = 64$	$\sum(y-\bar{y})^2 = 70$	$\sum xy = 65$

$$y = (y - \overline{y})$$

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$$

$$= \frac{65}{\sqrt{64 \times 70}}$$

$$= \frac{65}{\sqrt{4480}}$$

$$= \frac{65}{66.9328}$$

$$r = 0.9711$$

2.Find the co-efficient of correlation between x and y from the following data.

x	1	2	3	4	5	6	7	8	9
y	12	11	13	15	14	17	16	19	18

Solution

$$\bar{x} = \frac{\sum x}{n} = \frac{45}{9} = 5$$

$$\bar{y} = \frac{\sum y}{n} = \frac{135}{9} = 15$$

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})$	$\sum_{x}(x-\bar{x})\sum_{y}(y-\bar{y})$
1	12	-4	-3	16	9	12
2	11	-3	-4	9	16	12
3	13	-2	-2	4	4	4
4	15	-1	0	0	0	0
5	14	0	-1	1	1	0
6	17	1	2	4	4	2
7	16	2	1	1	1	2
8	19	3	4	16	16	12
9	18	4	3	9	9	12
		0	0	60	60	56

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$$

$$= \frac{56}{\sqrt{60 \times 60}}$$

$$= \frac{56}{60}$$

$$r = 0.9333$$

Method-II

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

This formula is used If no assumed average is taken for x and y series

1.Find the co-efficient of correlation between x and y from the following data.

Solution

x	5	10	5	11	12	4	3	2	7	1
y	1	6	2	8	5	1	4	6	5	2

x	y	x^2	y^2	xy
5	1	25	1	5
10	6	100	36	60
5	2	25	4	10
11	8	121	16	83
12	5	144	25	60
4	1	16	1	4
3	4	9	16	12
2	6	4	36	12
7	5	49	25	35
1	2	1	4	2
$\sum x = 60$	$\sum y = 40$	$\sum x^2 = 494$	$\sum y^2 = 494$	$\sum xy = 288$

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

N=10

$$r = \frac{288 - \frac{60 \times 40}{10}}{\sqrt{494 - \frac{(60)^2}{10}}\sqrt{212 - \frac{(40)^2}{10}}}$$

$$= \frac{288 - \frac{2400}{10}}{\sqrt{494 - \frac{3600}{10}}\sqrt{212 - \frac{1600}{10}}}$$

$$= \frac{288 - 240}{\sqrt{494 - 360}\sqrt{212 - 160}}$$

$$= \frac{48}{\sqrt{134}\sqrt{52}}$$

$$= \frac{48}{11.5758 \times 7.211}$$

$$= \frac{48}{83.4743}$$

$$\underline{r = 0.58}$$

Method-III

(Assumed Values)

$$r = \frac{\sum dx\, dy - \frac{\sum dx \sum dy}{N}}{\sqrt{\sum dx^2 - \frac{(\sum dx)^2}{N}}\sqrt{\sum dy^2 - \frac{(\sum dy)^2}{N}}}$$

Where $dx = x - A$

$dy = x - B$

1.Calculate the co-efficient of correlation between x and y from the following data.

x	10	12	13	16	17	20	25
y	19	22	26	27	29	33	37

Solution

$$\bar{x} = \frac{\sum x}{n} = \frac{113}{7} = 16.1429 \quad A = 16$$

$$\bar{y} = \frac{\sum y}{n} = \frac{193}{9} = 27.5714 \quad B = 27$$

x	y	$dx = x - A$	$dy = y - B$	$(dx)^2$	$(dy)^2$	$dxdy$
10	19	-6	-8	36	64	48
12	22	-4	-5	16	25	20
13	26	-3	-1	9	1	3
16	27	0	0	0	0	0
17	29	1	2	1	4	2
20	33	4	6	16	36	24
25	37	9	10	81	100	90
		$\sum dx = 1$	$\sum dy = 4$	$\sum dx^2 = 159$	$\sum dy^2 = 230$	$\sum dxdy = 187$

$$r = \frac{\sum dx\, dy - \frac{\sum dx \sum dy}{N}}{\sqrt{\sum dx^2 - \frac{(\sum dx)^2}{N}}\sqrt{\sum dy^2 - \frac{(\sum dy)^2}{N}}}$$

$$= \frac{187 - \frac{1 \ \times 4}{7}}{\sqrt{159 - \frac{(1^2)}{7}}\sqrt{230 - \frac{(4^2)}{7}}}$$

$$= \frac{187 - 0.5714}{\sqrt{159 - \frac{1}{7}}\sqrt{230 - \frac{16}{7}}}$$

$$= \frac{186.4286}{\sqrt{159 - 0.1429}\sqrt{230 - 2.2857}}$$

$$= \frac{186.4286}{\sqrt{158.8571}\sqrt{227.7143}}$$

$$= \frac{186.4286}{12.6039 \times 15.0902}$$

$$= \frac{186.4286}{12.6039 \times 15.0902}$$

$$= \frac{186.4286}{190.1954}$$

$$r = 0.98$$

Home work

2.Find the co-efficient of correlation between x and y from the following data.

X	**10**	**12**	**13**	**16**	**17**	**20**	**25**
Y	**19**	**22**	**26**	**27**	**29**	**33**	**37**

Karl Pearson's Co-efficient of correlation

$$r = \frac{\sum dx\, dy - \frac{\sum dx \sum dy}{N}}{\sqrt{\sum dx^2 - \frac{(\sum dx)^2}{N}}\sqrt{\sum dy^2 - \frac{(\sum dy)^2}{N}}} \quad [OR]$$

$$r = \frac{N \sum dx\, dy - \sum dx \sum dy}{\sqrt{N \sum dx^2 - (\sum dx)^2}\sqrt{N \sum dy^2 - (\sum dy)^2}}$$

1.Calculate the Pearson's co-efficient of correlation from the following data using 44 and 26 respectively as the origin of x and y.

x	43	44	46	40	44	42	45	42	38	40	42	57
Y	29	31	19	18	19	27	27	29	41	30	26	10

Solution

x	y	dx	dy	dx^2	dy^2	$dxdy$
43	29	-1	3	1	9	-3
44	31	0	5	0	25	0
46	19	2	-7	4	49	-14
40	18	-4	-8	16	64	32
44	19	0	-7	0	49	0
42	27	-2	1	4	1	-2
45	27	1	1	1	1	1
42	29	-2	3	4	9	-6
38	41	-6	15	36	225	-90
40	30	-4	4	16	16	-16
42	26	-2	0	4	0	0
57	10	13	-16	169	256	-208
		$\sum dx =$ -5	$\sum dy$ =-6	$\sum dx^2$ =255	$\sum dy^2$ =704	$\sum dxdy =$ -306

$$dx = x - 44 \qquad dy = y - 26$$

$$r = \frac{\sum dx\,dy - \frac{\sum dx \sum dy}{N}}{\sqrt{\sum dx^2 - \frac{\left(\sum dx\right)^2}{N}}\sqrt{\sum dy^2 - \frac{\left(\sum dy\right)^2}{N}}}$$

$$= \frac{-306 - \frac{(-5) \times (-6)}{12}}{\sqrt{255 - \frac{(-5)^2}{12}}\sqrt{704 - \frac{(-6)^2}{12}}}$$

$$= \frac{-306 - \frac{30}{12}}{\sqrt{255 - \frac{25}{12}}\sqrt{704 - \frac{36}{12}}}$$

$$= \frac{-306 - 2.5}{\sqrt{255 - 2.08}\sqrt{704 - 3}}$$

$$= \frac{-308.5}{\sqrt{252.92}\ \sqrt{701}}$$

$$= \frac{-308.5}{421.07}$$

$$r = -0.733$$

2.Calculate the Pearson's co-efficient of correlation from the following data using 65 and 70 as the assumed values of variables x and y.

x	45	55	56	58	60	65	68	70	75	80	85
y	56	40	48	60	62	64	65	70	74	82	90

Solution

$dx = x - 65$ $dy = y - 70$

x	y	dx	dy	dx^2	dy^2	$dxdy$
45	56	-20	-14	400	196	280
55	40	-10	-30	100	900	300
56	48	-9	-22	81	484	198
58	60	-7	-10	49	100	70
60	62	-5	-8	25	64	40
65	64	0	-6	0	36	0
68	65	3	-5	9	25	-15
70	70	5	0	25	0	0
75	74	10	4	100	16	40
80	82	15	12	225	144	180
85	90	20	20	400	400	400
		$\sum dx =$ 2	$\sum dx =$ -59	$\sum dx^2$ =1414	$\sum dx^2$ =2365	$\sum dxdy =$ 1493

$$r = \frac{\sum dx\, dy - \frac{\sum dx \sum dy}{N}}{\sqrt{\sum dx^2 - \frac{(\sum dx)^2}{N}}\sqrt{\sum dy^2 - \frac{(\sum dy)^2}{N}}}$$

$$= \frac{1493 - \frac{(2) \times (-59)}{11}}{\sqrt{1414 - \frac{(2)^2}{11}}\sqrt{2365 - \frac{(-59)^2}{11}}}$$

$$= \frac{1493 - \frac{(-118)}{11}}{\sqrt{1414 - \frac{4}{11}}\sqrt{2365 - \frac{3481}{11}}}$$

$$= \frac{1493 + 10.7273}{\sqrt{1414 - 0.3636}\sqrt{2365 - 316.4545}}$$

$$= \frac{1503.7273}{\sqrt{1413.6364}\sqrt{2048.5455}}$$

$$= \frac{1503.7273}{37.7046 \times 45.2609}$$

$$= \frac{1503.7273}{1701.5441}$$

$$r = -0.8812$$

3.The following table gives the age-distribution of the population and the number of unemployed in a town.

Age	No.of persons in '000	No.of unemployed
20-30	40	400
30-40	55	1100
40-50	32	960
50-60	20	1600
60-70	8	1600

Find the co-efficient of correlation r between the mid values of the age groups and percentage of unemployed in different age-constituents.

Solution

Let x represent the age (mid-value of classes) and y represent percentage employed.

x	y	dx	dy	dx^2	dy^2	$dxdy$
25	1	-2	-4	4	16	8
35	2	-1	-3	1	9	3
45	3	0	-2	0	4	0
55	8	1	3	1	9	3
65	20	2	15	4	225	30
		$\sum dx = 0$	$\sum dy = 9$	$\sum dx^2 = 10$	$\sum dy^2 = 263$	$\sum dxdy = 44$

Let $dx = \frac{x-45}{10}$ $\quad dy = y - 5$

$$r = \frac{\sum dx\, dy - \frac{\sum dx \sum dy}{N}}{\sqrt{\sum dx^2 - \frac{(\sum dx)^2}{N}}\sqrt{\sum dy^2 - \frac{(\sum dy)^2}{N}}}$$

$$= \frac{44 - \frac{0 \times 9}{11}}{\sqrt{10 - 0}\sqrt{263 - \frac{(9)^2}{5}}}$$

$$= \frac{44}{\sqrt{10}\sqrt{246.8}}$$

$$r = 0.88$$

1.Calculate the Karl Pearson's coefficient of correlation between x and y for the following information.

$$N = 12\ ; \sum(x-8)^2 = 150;\ \sum x = 120;\ \sum(y-10)^2 = 200;\ \sum y = 130;\ \sum(x-8)(y-10) = 50$$

Solution

$$r = \frac{\sum dx\,dy - \frac{\sum dx \sum dy}{N}}{\sqrt{\sum dx^2 - \frac{(\sum dx)^2}{N}}\ \sqrt{\sum dy^2 - \frac{(\sum dy)^2}{N}}} \quad [OR]$$

$$r = \frac{N\sum dx\,dy - \sum dx \sum dy}{\sqrt{N\sum dx^2 - (\sum dx)^2}\ \sqrt{N\sum dy^2 - (\sum dy)^2}}$$

Take $A = 8$; $B = 10$; $dx = x - 8$; $dy = y - 10$

$$\sum dx = \sum x - \sum dx$$

$$= 120 - 96$$

$$\sum dx = 24$$

$$\sum dy = \sum y - \sum dy$$

$$= 130 - 120$$

$$\sum dy = 10$$

$$r = \frac{N\sum dx\,dy - \sum dx \sum dy}{\sqrt{N\sum dx^2 - \left(\sum dx\right)^2}\sqrt{N\sum dy^2 - \left(\sum dy\right)^2}}$$

$$r = \frac{12 \times 50 - 24 \times 10}{\sqrt{12 \times 150 - 24^2}\sqrt{12 \times 200 - 10^2}}$$

$$= \frac{600 - 240}{\sqrt{1800 - 576}\sqrt{2400 - 100}}$$

$$= \frac{360}{\sqrt{1224}\sqrt{2300}}$$

$$= \frac{360}{1677.86}$$

$$r = 0.2146$$

Home work

1.Calculate the coefficient of correlation for the following data.

$$N = 10\ ; \sum x^2 = 290; \sum x = 50; \sum y = -30 \quad \sum xy = -115;$$

$$\sum y^2 = 300$$

2.Calculate the coefficient of correlation between x and y for the following data.

$$N = 10\ ; \sum x^2 = 400; \sum x = 60; \sum y = 60 \quad \sum xy = 305;$$

$$\sum y^2 = 580$$

3.Calculate the Karl Pearson's coefficient of correlation from the following data using 20 as the working mean for the price and 70 as the working mean for demand

Price	14	16	17	18	19	20	21	22	23
Demand	84	78	70	75	66	67	62	58	60

Rank Correlation

Method I

The rank correlation coefficient when there are n ranks in each variable is given by the formula (due to Spearman).

$$\rho = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

Where $d = x - y$ is the difference between ranks of corresponding pairs of x and y.

N=number of observations

1.The following are the rank obtained by 10 students in Statistics and Mathematics

Statistics	1	2	3	4	5	6	7	8	9	10
Mathematics	1	4	2	5	3	9	7	10	6	8

To what extent is the knowledge of students in the two subjects related?

Solution

x	y	$d = x - y$	d^2
1	1	0	0
2	4	-2	4
3	2	1	1
4	5	-1	1
5	3	2	4
6	9	-3	9
7	7	0	0
8	10	-2	4
9	6	3	9
10	8	2	4
			$\sum d^2 = 36$

The rank correlation is given by

$$\rho = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

$$= 1 - \frac{6 \times 36}{10(10^2 - 1)}$$

$$= 1 - \frac{6 \times 36}{10(99)}$$

$$= 1 - 0.219$$

$$\rho = 0.781$$

2.Ten competitors in a beauty contest are ranked by three judges in the following order

1st judge	1	4	6	3	2	9	7	8	10	5
2nd judge	2	6	5	4	7	10	9	3	8	1
3rd judge	3	7	4	5	10	8	9	2	6	1

Use the method of rank correlation coefficient to determine which pair of judges have the nearest approach to common taste in beauty?

Solution

Let x,y,z denote the ranks by 1st,2nd, and 3rd judges respectively.

x	y	z	$d_{xy} = x - y$	$d_{yz} = y - z$	$d_{zx} = z - x$	${d_{xy}}^2$	${d_{yz}}^2$	${d_{zx}}^2$
1	2	3	-1	-1	-2	1	1	4
4	6	7	-2	-1	-3	4	1	9
6	5	4	1	1	2	1	1	4
3	4	5	-1	-1	-2	1	1	4
2	7	10	-5	-3	-8	25	9	64
9	10	8	-1	2	1	1	4	1
7	9	9	-2	0	-2	4	0	4
8	3	2	5	1	6	25	1	36
10	8	6	2	2	4	4	4	16
5	1	1	4	0	4	16	0	16
						$\sum {d_{xy}}^2 = 82$	$\sum {d_{yz}}^2 = 22$	${d_{zx}}^2 = 158$

$$\rho_{xy} = 1 - \frac{6\sum {d_{xy}}^2}{n(n^2-1)}$$

$$= 1 - \frac{6 \times 82}{10(10^2-1)}$$

$$= 1 - \frac{6 \times 82}{10(99)}$$

$$\rho_{xy} = 0.503$$

$$\rho_{yz} = 1 - \frac{6\sum {d_{yz}}^2}{n(n^2-1)}$$

$$= 1 - \frac{6 \times 22}{10(99)}$$

$$\rho_{yz} = 0.867$$

$$\rho_{zx} = 1 - \frac{6\sum {d_{zx}}^2}{n(n^2-1)}$$

$$= 1 - \frac{6 \times 158}{10(99)}$$

$$\rho_{xy} = 0.04$$

Since the rank correlation coefficient between y and z is positive and highest among the three coefficients, judges y and z have the nearest approach for common taste in beauty.

Home work

A random sample of recent repair jobs was selected and estimated cost and actual cost were recorded

Estimated cost	**300**	**450**	**800**	**250**	**500**	**975**	**475**	**400**
Actual cost	**273**	**486**	**734**	**297**	**631**	**872**	**396**	**457**

Calculate the value of Spearman's correlation coefficient.

Method II

Note: Tie ranks

When the values of variables x and y are given, we can rank the values in each of the variables and determine the Spearman's rank correlation coefficient. If two or more observations have the same rank we assign to them the mean rank. In this care, there is a correlation factor in the formula for ρ. The formula for ρ is given by

$$\rho = 1 - \frac{6\left[\sum d^2 + \frac{\sum m\,(m^2-1)}{12} + \cdots\right]}{n(n^2-1)}$$

Where m stands for the number of times a rank is repeated.

Example

If a rank is repeated 2 times in $x -$series and $y -$series and 3 times in $y - series$.the correlation factor is

$$\frac{2(2^2-1)}{12} + \frac{3(3^2-1)}{12}$$

1.Find the rank correlation co-efficient for the following data

x	92	89	87	86	86	77	71	63	53	50
y	86	83	91	77	68	85	52	82	37	57

Solution

Let R_1 and R_2 denote the ranks in x and y respectively

x	y	R_1	R_2	$d = R_1 - R_2$	d^2
92	86	1	2	-1	1
89	83	2	4	-2	4
87	91	3	1	2	4
86	77	4.5	6	-1.5	2.25
86	68	4.5	7	-2.5	6.25
77	85	6	3	3	9
71	52	7	9	-2	4
63	82	8	5	3	9
53	37	9	10	-1	1
50	57	10	8	2	4
					$\sum d^2 = 44.5$

$$\rho = 1 - \frac{6\left[\sum d^2 + \frac{m(m^2-1)}{12}\right]}{n(n^2-1)}$$

$$\rho = 1 - \frac{6\left[44.5 + \frac{2(2^2-1)}{12}\right]}{10(10^2-1)}$$

$$= 1 - \frac{6\left[44.5 + \frac{2(4-1)}{12}\right]}{10(100-1)}$$

$$= 1 - \frac{6\left[44.5 + \frac{2(3)}{12}\right]}{10(99)}$$

$$= 1 - \frac{6\left[44.5 + \frac{6}{12}\right]}{990}$$

$$= 1 - \frac{6[44.5 + 0.5]}{990}$$

$$= 1 - \frac{6[45]}{990}$$

$$= 1 - \frac{270}{990}$$

$$= 1 - 0.2727$$

$$\rho = 0.7273$$

3.Find the rank correlation co-efficient for the following data

x	43	44	46	40	44	42	45	42	38	40	42	57
y	29	31	19	18	19	27	27	29	41	30	26	10

Solution

Solution

x	y	R_1	R_2	$d = R_1 - R_2$	d^2
43	29	6	4.5	1.5	2.25
44	31	4.5	2	2.5	6.25
46	19	2	9.5	-7.5	56.25
40	18	10.5	11	0.5	0.25
44	19	4.5	9.5	-5	25
42	27	8	6.5	1.5	2.25
45	27	3	6.5	-3.5	12.25
42	29	7	4.5	3.5	12.25
38	41	12	1	11	121
40	30	10.5	3	7.5	56.25
42	26	8	8	0	0
57	10	1	12	11	121
					$\sum d^2 = 415$

$$\rho = 1 - \frac{6\left[\sum d^2 + m\left(\frac{m^2-1}{12}\right) + m\left(\frac{m^2-1}{12}\right) + m\left(\frac{m^2-1}{12}\right) + m\left(\frac{m^2-1}{12}\right) + m\left(\frac{m^2-1}{12}\right) + m\left(\frac{m^2-1}{12}\right)\right]}{n(n^2-1)}$$

$$\rho = 1 - \frac{6\left[415 + 2\left(\frac{2^2-1}{12}\right) + 3\left(\frac{3^2-1}{12}\right) + 2\left(\frac{2^2-1}{12}\right) + 2\left(\frac{2^2-1}{12}\right) + 2\left(\frac{2^2-1}{12}\right)\right] + 2\left(\frac{2^2-1}{12}\right)}{12(12^2-1)}$$

$$= 1 - \frac{6\left[\begin{array}{c} 415 + 2\left(\frac{3}{12}\right) + 3\left(\frac{8}{12}\right) + 2\left(\frac{3}{12}\right) \\ +2\left(\frac{3}{12}\right) + 2\left(\frac{3}{12}\right) \end{array}\right] +2\left(\frac{3}{12}\right)}{12(144 - 1)}$$

$$= 1 - \frac{6\left[\begin{array}{c} 415 + 0.5 + 2 + 0.5 \\ +0.5 + 0.5 \end{array}\right] +0.5}{12(143)}$$

$$= 1 - \frac{6(419.5)}{1716}$$

$$= 1 - \frac{2517}{1716}$$

$$= 1 - 1.4668$$

$$\rho = 0.4668$$

Concurrent deviation

This method requires only a direction of change $(+ve\ to - ve\ or - ve\ to\ + ve)$ in the successive values of the variables

The coefficient of correlation is given by the formula

$$r_c = \pm\sqrt{\pm\frac{(2C-N)}{N}}\quad (-1 \le r_c \le 1)$$

Where $N = n - 1$

N is the number of pairs of deviation

Where C is the number of concurrent deviation

Merits

1.It is the simplest formula for calculating '*r*'

2.It is easily determined by this method whether the variables are dependent or not.

Demerits

1.This method does not differentiate small and big changes in the values of variables. Both small and big changes have the same weight when they are considered for the change in the direction of variables.

2.The value of '*r*' obtained by this method is only a rough indicator for the presence or absence of correlation.

1.Calculate coefficient of correlation by the method of concurrent deviation from the following data.

X	84	85	62	48	84	95	103	100	85	115
Y	20	23	19	21	25	25	28	27	26	30

Solution

X	Changes in direction of variable X Dx	Y	Changes in direction of variable Y Dy	$DxDy$
84		20		
85	$+$	23	$+$	$+$
62	$-$	19	$-$	$+$
48	$-$	21	$+$	$-$
84	$+$	25	$+$	$+$
95	$+$	25	*No change*	$-$
103	$+$	28	$+$	$+$
100	$-$	27	$-$	$+$
85	$-$	26	$-$	$+$
115	$+$	30	$+$	7
			No.of concurrent deviation	7
			Disagreement	2

$$r_c = \pm\sqrt{\pm\frac{(2C - N)}{N}}$$

$$= \pm\sqrt{\pm\frac{(2 \times 7 - 9)}{9}}$$

$$= \pm\sqrt{\pm\frac{5}{9}}$$

$$= +\sqrt{+0.5556}$$

$$r = 0.7454$$

1.Calculate coefficient of correlation by the method of concurrent deviation from the following data.

X	60	59	72	51	55	54	65
Y	23	36	10	38	33	44	33

Solution

X	Changes in direction of variable X Dx	Y	Changes in direction of variable Y Dy	$DxDy$
60		23		
59	$-$	36	$+$	$-$
72	$+$	10	$-$	$-$
51	$-$	38	$+$	$-$
55	$+$	33	$-$	$-$
54	$-$	44	$-$	$-$
65	$+$	33	$-$	$-$
			No.of concurrent deviation $\sum C$	0
			Disagreement	6

$$r_c = \pm\sqrt{\pm\frac{(2C - N)}{N}}$$

$$= \pm\sqrt{\pm\frac{(2 \times 0 - 6)}{9}}$$

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

$$= -\sqrt{1}$$

$$r = -1$$

3.If N is 11 and C is 2 find out concurrent deviation correlation.

Solution

$$r_c = \pm\sqrt{\pm\frac{(2C-N)}{N}}$$

$$r_c = \pm\sqrt{\pm\frac{(2\times 2-11)}{11}}$$

$$= \pm\sqrt{\pm\frac{(4-11)}{11}}$$

$$= +\sqrt{-\frac{-7}{11}}$$

$$= +\sqrt{-0.63}$$

$$= -0.7937$$

Regression

Definition

Regression is the study of the relationship between the variables. If Y is dependent variable and X is the independent variable the linear relationship suggested between the variables is called regression.

Method-I

The regression equation of *x* on *y*

$$x - \overline{x} = b_{xy}(y - \overline{y})$$

$$b_{xy} = \frac{n\sum xy - \sum x \sum y}{n\sum y^2 - (\sum y)^2}$$

The regression equation of *y* on *x*

$$y - \overline{y} = b_{yx}(x - \overline{x})$$

$$b_{yx} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

1.Find the regression equation for the following data

x	1	2	3	4	5	8	10
y	9	8	10	12	14	14	15

Solution

x	y	x^2	y^2	xy
1	9	1	81	9
2	8	4	64	16
3	10	9	100	30
4	12	16	144	48
5	14	25	196	70
8	16	64	256	128
10	15	100	225	150
$\sum x = 33$	$\sum y = 84$	$\sum x^2 = 219$	$\sum y^2 = 1066$	$\sum xy = 451$

$$\bar{x} = \frac{\sum x}{n} = \frac{33}{7} = 4.7143$$

$$\bar{y} = \frac{\sum y}{n} = \frac{84}{7} = 12$$

The regression equation of x on y

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

$$x - 4.7143 = b_{xy}(y - 12) \_\_\_\_\_\_\_\_ (1)$$

$$b_{xy} = \frac{n\sum xy - \sum x \sum y}{n\sum y^2 - (\sum y)^2}$$

$$= \frac{(7 \times 451) - (33 \times 84)}{(7 \times 1066) - (84)^2}$$

$$= \frac{3157 - 2772}{7462 - 7056}$$

$$= \frac{385}{406}$$

$$b_{xy} = 0.9483$$

From (1), $x - 4.7143 = 0.9483(y - 12)$

$$x = 0.9483(y - 12) + 4.7143$$

$$= 0.9483y - 11.3796 + 4.7143$$

$$x = 0.9483y - 6.6658$$

The regression equation of y on x

$$y - \overline{y} = b_{yx}(x - \overline{x})$$

$$y - 12 = b_{yx}(x - 4.7143) \_\_\_\_\_\_\_\_\_\_ (2)$$

$$b_{yx} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

$$= \frac{(7 \times 451) - (33 \times 84)}{(7 \times 219) - (33)^2}$$

$$= \frac{3157 - 2772}{1533 - 1089} = \frac{385}{444}$$

$$b_{yx} = 0.8671$$

From (2), $y - 12 = b_{yx}(x - 4.7143)$

$y = 0.8671(x - 4.4143) + 12$

$= 0.8671x - 4.0878 + 12$

$x = 0.8671x - 7.9122$

2.The following gives age x in years of cars and annual maintenance cost y.

X	1	3	5	7	9
Y	15	18	21	23	22

$x = 4$ finding the regression equation

Solution

x	y	x^2	xy
1	15	1	15
3	18	9	54
5	21	25	105
7	23	49	161
9	22	81	198
$\sum x = 25$	$\sum y = 99$	$\sum x^2 = 165$	$\sum xy = 533$

$$\bar{x} = \frac{\sum x}{n} = \frac{25}{5} = 5$$

$$\bar{y} = \frac{\sum y}{n} = \frac{99}{5} = 19.8$$

The regression equation of y on x $y - \bar{y} = b_{yx}(x - \bar{x})$

$$y - 19.8 = b_{yx}(x - 5) \_\_\_\_\_\_\_\_\_\_ (1)$$

$$b_{yx} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

$$= \frac{(5 \times 533) - (25 \times 99)}{(5 \times 165) - (25)^2}$$

$$= \frac{2665 - 2475}{825 - 625}$$

$$= \frac{190}{200}$$

$$b_{yx} = 0.95$$

From (1), $y - 19.8 = 0.95(x - 5)$

$$y = 0.95(x - 5) + 19.8$$

$$= 0.95x - 4.75 + 19.8$$

$$y = 0.95x + 15.05 \_\_\_\_\_\_\_\_\_\_ (2)$$

When $x = 4$, equation (2) gives $y = 0.95(4) + 15.05$

$$y = 3.8 + 15.05$$

$$y = 18.85$$

Method-II (Assumed method)

The regression equation of x on y

$$x - \overline{x} = b_{xy}(y - \overline{y})$$

$$b_{xy} = \frac{n\sum dx\,dy - \sum dx \sum dy}{n\sum dy^2 - (\sum dy)^2}$$

The regression equation of y on x

$$y - \overline{y} = b_{yx}(x - \overline{x})$$

$$b_{yx} = \frac{n\sum dx\,dy - \sum dx \sum dy}{n\sum dx^2 - (\sum dx)^2}$$

1.Find the equation of regression lines for the following data

x	25	28	35	32	36	36	29	38	34	32
y	43	46	49	41	36	32	31	30	33	39

Solution

x	y	$dx = x - A$	$dy = y - B$	dx^2	dy^2	$dxdy$
25	43	-7	5	49	25	-35
28	46	-4	8	16	64	-32
35	49	3	11	9	121	33
32	41	0	3	0	9	0
36	36	4	-2	16	4	-8
36	32	4	-6	16	36	-24
29	31	-3	-7	9	49	21
38	30	6	-8	36	64	-48
34	33	2	-5	4	25	-10
32	39	0	1	0	1	0
$\sum x = 325$	$\sum y = 380$	$\sum dx = 5$	$\sum dy = 0$	$\sum dx^2 = 155$	$\sum dy^2 = 398$	$\sum dx\,dy = -103$

Let $dx = x - 32$; $dy = y - 38$

$$\bar{x} = \frac{\sum x}{n} = \frac{325}{10} = 32.5; \quad A = 32$$

$$\bar{y} = \frac{\sum y}{n} = \frac{380}{10} = 38; \quad B = 38$$

The regression equation of x on y

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

$$x - 32.5 = b_{xy}(y - 38)$$

$$b_{xy} = \frac{n \sum dx\, dy - \sum dx \sum dy}{n \sum dy^2 - (\sum dy)^2}$$

$$= \frac{(10 \times (-103) - (5 \times 0)}{(10 \times 398) - (0)^2}$$

$$= \frac{-1030 - 0}{3980 - 0}$$

$$= \frac{-1030}{3980}$$

$$b_{xy} = -0.2588$$

$$x - 32.5 = -0.2588(y - 38)$$

$$x = -0.2588(y - 38) + 32.5$$

$$= -0.2588y + 9.8344 + 32.5$$

$$x = -0.2588y + 42.3344$$

The regression equation of y on x

$$y - \overline{y} = b_{yx}(x - \overline{x})$$

$$y - 38 = b_{yx}(x - 32.5)$$

$$b_{yx} = \frac{n \sum dx\, dy - \sum dx \sum dy}{n \sum dx^2 - (\sum dx)^2}$$

$$= \frac{(10 \times (-103)) - (5 \times 0)}{(10 \times 155) - (5)^2}$$

$$= \frac{-1030 - 0}{1550 - 25}$$

$$= \frac{-1030}{1525}$$

$$b_{yx} = -0.6754$$

x	y	$dx = x - A$	$dy = y - B$	dx^2	$dx\,dy$
65	67	-3	-2	9	6
66	68	-2	-1	4	2
67	69	-1	0	1	0
67	68	-1	-1	1	1
69	70	1	1	1	1
71	70	3	1	9	3
72	69	4	0	16	0
70	70	2	1	4	2
65	70	-3	1	9	-3
$\sum x = 612$	$\sum y = 621$	$\sum dx = 0$	$\sum dy = 0$	$\sum dx^2 = 54$	$\sum dy^2 = 12$

Let $dx = x - 68$; $dy = y - 69$

$$\bar{x} = \frac{\sum x}{n} = \frac{612}{9} = 68; \quad A = 68$$

$$\bar{y} = \frac{\sum y}{n} = \frac{621}{9} = 69; \quad B = 69$$

The regression equation of y on x

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$y - 69 = b_{yx}(x - 68)$$

$$b_{yx} = \frac{9(12) - 0}{9(54)}$$

$$= \frac{108}{486}$$

$$b_{yx} = 0.2222$$

$$y - 69 = 0.2222(x - 68)$$

$$y = 0.2222x - 15.1096 + 69$$

$$y = 0.2222x + 53.8904 \_\_\_\_\_\_\_\_\_\_ (1)$$

When $x = 64, from\ (1), y = 0.2222(64) + 53.8904$

$$y = 14.2208 + 53.8904$$

$$y = 68.1112$$

3.Obtain the two regression lines from the following data

$$n = 20, \sum x = 80, \sum y = 40, \sum x^2 = 1680, \sum y^2 = 320,$$

$$\sum xy = 480,$$

Solution

The regression line equation of x on y

$$x - \overline{x} = b_{xy}(y - \overline{y})$$

$$b_{xy} = \frac{n\sum xy - \sum x \sum y}{n\sum y^2 - (\sum y)^2}$$

$$= \frac{(20 \times 480) - (80 \times 40)}{(20 \times 320) - (40)^2}$$

$$= \frac{6400}{6400 - 1600}$$

$$= \frac{6400}{4800}$$

$$b_{xy} = 1.33$$

The regression line equation of x on y

$$x - 4 = 1.33(y - 2)$$

$$\underline{x = 1.33y + 1.34}$$

The regression line equation of y on x

$$y - \overline{y} = b_{yx}(x - \overline{x})$$

$$b_{yx} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

$$= \frac{(20 \times 480) - (80 \times 40)}{(20 \times 1680) - (8s0)^2}$$

$$= \frac{9600 - 3200}{33600 - 6400}$$

$$= \frac{6400}{27200}$$

$$b_{yx} = 0.2353$$

The regression line equation of y on x

$$y - 2 = 0.2353(x - 4)$$

$$\underline{y = 0.2353x + 1.0588}$$

Home work

1.Obtain the least square regression line y on x for the following data

X	1	2	3	4	5	6	7	8	9
Y	9	8	10	12	11	13	14	16	15

Also obtain an estimate of y which should be correspond on an average to x=6.4

2.Find the correlation coefficients between x and y and write down the equation of the regression lines from the following data.

$$n = 25, \sum x = 125, \sum y = 100, \sum x^2 = 650, \sum y^2 = 460,$$

$$\sum xy = 508,$$

Solution

Hint :find b_{yx} and b_{xy}

Sine b_{yx} and b_{xy} are positive, $r = \sqrt{b_{yx} \times b_{xy}}$

$$r = 0.2065$$

Then find regression lines for both x on y and y on x

THANK YOU