

Business Mathematics & Statistics (BMS)

Mr. Pasan Randeer

DATA ANALYSIS

Descriptive Statistics

Scientific survey of data can be described as descriptive statistics. It contains the following four stages.

- (i) Collection of data
- (ii) Organization of data
- (iii) Presentation of data
- (iv) Analyzing of data

BASIC DEFINITIONS INVOLVING IN STATISTICS

Population

The set of all observations which we are considering is known as a population.

Example

In a geographical study about rivers in Sri Lanka we can consider "The set of all rivers in Sri Lanka" as our population.

Sample

A limited number of items drawn from a population is a sample. A sample must be a correct representative of a population.

DATA CLASSIFICATION

Main raw material we are using in Data Analysis is data.

Variable

Any characteristic that can be measured is a variable

Examples

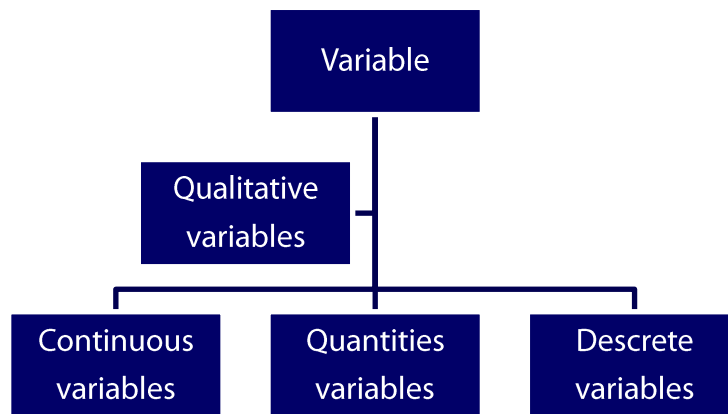
- (i) Height (ii) Weight (iii) Age

Data

In statistics you collect observations or measurements of some variable. Such observations are known as data.

Examples

- Heights of students
- Weights of packages
- Ages of persons



Qualitative Variables

Variables associated with non-numerical observations are called qualitative variables.

Examples

- Nationality of a person.
- Customer satisfaction regarding a service of a bank.

Quantitative Variables

Variables associated with numerical observations are called quantitative variables.

Examples

- Heights of students
- No. of defective items in a load of goods.
- Rainfall of a particular region
- No. of children in a family.

Quantitative variables can be further divided into two categories as:

- (i) Continuous variables
- (ii) Discrete variables

Continuous Variables

A variable that can take any value in a given range is a continuous variable.

Examples

- Marks of students
- Life time of an electric bulb
- Weights of packages

Discrete Variables

A variable that can take only specific values in a given range is a discrete data.

Example

- No of children in a family
- No of printing errors per page in a particular books
- No of vehicles passes through a particular junction

Raw Data

A data that is collected and not arranged in any way is raw data.

Example

As an example consider the data giving details about no: employees working in a small scale industries in a particular area.

21	28	26	26	23
26	28	24	28	25
24	23	24	21	26
28	26	28	23	25
22	21	28	25	21

Array

This is simplest way of arranging data. It arranges it in ascending or descending order. It has many advantages over the raw data.

Advantages

- We can quickly notice the lowest and highest value in the data
- We can see whether any value appears more than once
- We can always divide the data into various sections

Disadvantage

But an ordered array does not summarize data in any way.

Un-grouped Frequency Distribution

Large number of discrete data can be written as a table what is known as an un-grouped frequency distribution.

No of employees	Tally mark	Frequency (f)
21	IIII	04
22	I	01
23	III	03
24	III	03
25	III	03
26	IIII	05
27	IIII I	06
		25

Total Frequency (Σf)

Note

Sometimes continuous data can be summarized as an un-grouped frequency distribution.

Grouped Frequency Distribution

For continuous data we make non-overlapping sub classes as class intervals. Then count number of observations belonging to the each class, by using tally marks and summarize the data as a table. Such a table is known as a grouped frequency table.

Example

The doctor's office staff has studied the waiting times for the patients who arrive at the office for emergency service. The following data collected over one month period. Times are to the nearest minute.

02	10	04	05	11
08	21	08	13	03
05	2	04	07	08
12	06	07	08	09

Prepare a grouped frequency distribution by taking first class interval as 0 – 4.

Time (Minutes)	Tally mark	Frequency (f)
0 – 4	IIII	04
5 – 9	IIII IIII	09
10 – 14	IIII	05
15 – 19	I	01
20 – 24	I	01
		$\Sigma f = 25$

Note (i)

Sometimes discrete data can be summarized as grouped frequency distribution.

Note (ii)

If first class interval is not given, following formula can be used to decide the size of a class interval.

$$\text{Size} = \frac{(\text{Highest Value} - \text{Lowest Value})}{\text{No of Groups}}$$

Note (iii)

If raw data available with decimal numbers, following methods can be used to summarize data.

Method I

Time (Minutes)	f
Greater than or equal zero, but less than 5	04
Greater than or equal five but less than 10	09
Greater than or equal ten but less than 15	05
Greater than or equal 15 but less than 20	01
Greater than or equal 20 but less than 25	01
	$\Sigma f = 20$

Method II

Time (Minutes)	f
Greater than 0 but less than or equal 5	04
Greater than 5 but less than or equal 10	09
Greater than 10 but less than or equal 15	05
Greater than 15 but less than or equal 20	01
Greater than 20 but less than or equal 25	01
	$\Sigma f = 20$

When data is presented as a grouped frequency table, the specific data values are last. There for we use following properties of class intervals for further statistical calculations.

Class Limit

Extreme values of class intervals are class limits.

Example (1)

0 – 4	(Lower class limit 0 – Upper class limit 4)
5 – 9	(Lower class limit 5 – Upper class limit 9)
10 – 14	(Lower class limit 10 – Upper class limit 14)

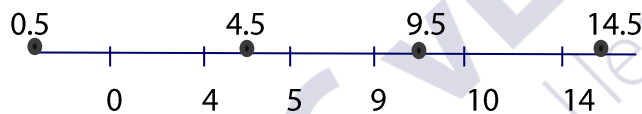
Example (2)

$> 0 < 5$	(Lower class limit 0 – Upper class limit 5)
$> 5 < 10$	(Lower class limit 5 – Upper class limit 10)
$> 10 < 15$	(Lower class limit 10 – Upper class limit 15)

Class Boundary

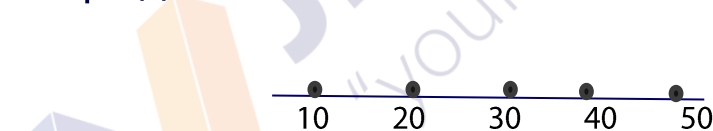
Theoretical limit of a class interval is class boundary.

Example (1)



- Class Limit
- Class Boundary

Example (1)



- Class Limit
- Class Boundary

Class Mark

Midpoint of a class interval is class mark. It can be evaluate by using following formulae.

$$\text{Class Mark} = \frac{\text{Lower Class Limit} + \text{Upper Class Limit}}{2}$$

$$\text{Class Mark} = \frac{\text{Lower Class Boundary} + \text{Upper Class Boundary}}{2}$$

Class Width

Size of a class interval is class width. It can be evaluate by using following formula.

$$\text{Class Width} = \text{Upper Class Boundary} - \text{Lower Class Boundary}$$

Example (1)

Evaluate (i) Class boundary
(ii) Class mark
(iii) Class width of following sets of class intervals

01. 71 - 80
81 - 90
91 - 100

02. Greater than or equal 50 but less than 100 ($> 50 < 100$)
Greater than or equal 100 but less than 150 ($> 100 < 150$)
Greater than or equal 150 but less than 200 ($> 150 < 200$)

03. 5,000 - 9,999
10,000 - 14,999
15,000 - 19,999

04. Greater than 20 but less than or equal 40 ($> 20 <= 40$)
 Greater than 40 but less than or equal 60 ($> 40 <= 60$)
 Greater than 60 but less than or equal 80 ($> 60 <= 80$)

MEASURE OF LOCATION

A set of data can be summarized by giving a single number to describe its centre. This number is called a measure of location. There are three alternative measures of location that you can use namely.

- Arithmetic mean
- Median
- Mode

THE ARITHMETIC MEAN

This is a measure what is known as common average.

Arithmetic Mean for an Un-grouped Data

Definition 01

The arithmetic mean of the set of numbers $x_1, x_2, x_3, \dots, x_n$ is given by a following formula. It is denoted by \bar{x}

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

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

Definition 02

The arithmetic mean for the set of observations $x_1, x_2, x_3, \dots, x_n$ occurs with frequencies $f_1, f_2, f_3, \dots, f_n$ respectively is given by a following formula.

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

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

Example (1)

- (a) Evaluate arithmetic mean of following set of observations.
10 20 30 40 50
- (b) A child at a junior school records the maximum temperature C_0 for seven days at school. The results are given below.

Evaluate arithmetic mean

15.7 16.1 16.2 17.6 17.4 18.6 16.7

Combining Means

$$\frac{\bar{X}_A}{n_A}$$

A

$$\frac{\bar{X}_B}{n_B}$$

B

If set A of size n_A has mean \bar{X}_A and set B of size n_B has mean \bar{X}_B then the mean of combined set of A and B is given as follows.

$$\bar{X}_{A+B} = \frac{n_A \bar{X}_A + n_B \bar{X}_B}{n_A + n_B}$$

Example (2)

The mean of a sample of 25 observations is 6.4. The mean of a second sample of 30 observations is 7.2. Evaluate arithmetic mean of combined set.

Example (3)

The frequency table shows the number of breakdowns per month recorded by a firm over a certain period time. Calculate mean number of break downs.

Number of breakdowns	0	1	2	3	4	5
Frequency	08	11	12	03	01	01

Example (4)

Evaluate arithmetic mean of following data set.

X	37	38	39	40	41	42
F	10	15	20	30	14	11

Arithmetic Mean for Grouped Data

When data is presented as a grouped frequency table, specific data values are lost. Therefore we work out estimate for an arithmetic mean by using midpoint of a class interval as its representative.

The arithmetic mean of a data that is summarized by a grouped frequency distribution is given by the following formula.

$$\bar{X} = \frac{\sum fx}{\sum f} \quad x = \text{mid point of a class interval}$$

Example (5)

The life time of 80 batteries to the nearest hour, is shown in a table below. Calculate mean life time of

Length of TP Call (min)	No. of Occasions
6 – 10	02
11 – 15	10
16 – 20	18
21 – 25	45
26 – 30	05

Example (6)

Estimate the mean length of a telephone call given the data in table below.

Length of TP Call (min)	No. of Occasions
0 – 5	04
5 – 10	15
10 – 15	05
15 – 20	02
20 – 60	00
60 – 70	01

MEASURE OF VARIATION

By comparing several data sets average may be same, but variables may highly differ in magnitude.

Therefore, the central tendency calculated from such variables may not be the most typical or representative, in many cases. To know the extent of spread about these averages or the variations of items, we have to resort to some other measures. Such measures known as measures of variation. The following are the important methods of measuring dispersion.

- (i) Range
- (ii) Quartile deviation
- (iii) Mean deviation
- (iv) Standard deviation

Standard Deviation

The standard deviation is found by adding the squares of the deviations of the individual values from the mean of a distribution, dividing the sum by the number of times in the distribution, and then finding the square root of the quotient.

Note (i)

The notation σ used for standard deviation of the population and S used for standard deviation of sample.


Note (ii)

Square of standard deviation is variance.

Standard Deviation for an Un-grouped Data

Definition 01

The set of observations standard deviation is given by the following formula



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

A more friendly version of formula is given by

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

Definition 02

The set of observations occurs with frequencies respectively. The standard deviation is given by:

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

A more friendly version of this formula is given by:

$$\sigma = \sqrt{\frac{\sum fx^2}{\sum f} - (\bar{x})^2}$$

Example (1)

- (a) Evaluate mean, standard deviation and variance for following set of observations.
10 20 30 40 50
- (b) Heights of 8 students is given below. Evaluate mean and standard deviation.
165 170 190 175 185 176 184

Example (2)

Evaluate mean and standard deviation for following data set.

X	5	10	15	20	25
F	03	08	18	08	03

Example (3)

Time (Minutes)	35	36	37	38	39
No. of Students	03	17	29	34	26

Standard Deviation for Grouped Data

When data presented as a grouped frequency distribution specific data values are lost. Therefore we calculate estimate for standard deviation by using mid point of a class interval as its representative. It is given by a following formula.

$$\sigma = \sqrt{\frac{\sum f(x - \bar{x})^2}{\sum f}} \quad X = \text{mid point of a class interval}$$

A more friendly version of formula is given by:

$$\sigma = \sqrt{\frac{\sum fx^2}{\sum f} - (\bar{x})^2} \quad X = \text{mid point}$$

Example (5)

The life time of 70 light bulbs is shown in a table below. Estimate mean and standard for the data.

Life Time Hours	<i>f</i>
20 – 22	03
22 – 24	12
24 – 26	40
26 – 28	10
28 – 30	03

Example

Marks of 60 students is given below. Evaluate mean and standard deviation for the data.

Marks	<i>f</i>
100 – 107	08
108 – 114	13
115 – 121	24
122 – 128	11
129 – 135	04

Coefficient of Variation

When we express the variation of a set of data relative to its mean, we use coefficient of variation, to measure the variability of data. Coefficient of variation is given by a following formula.

$$CV = \frac{\text{Standard Deviation}}{\text{Mean}} \times 100$$

Note:

Decision makers can use coefficient of variation to compare the variability of two or more distributions, specially when the units of measurements are different in the distribution.

Example

Distribution A gives the heights of students in 'cm' and distribution B gives weights of students in kg. Evaluate coefficient of variation in each distribution and state which distribution has more variation.

A (cm)	140	145	148	150	152	155	160
B (kg)	51	56	59	60	61	64	69

DATA PRESENTATION

For presentation of data, we use

- (i) Histogram
- (ii) Frequency Polygon
- (iii) Cumulative Frequency Curve
- (iv) Bar Charts
- (v) Pie Charts

HISTOGRAM

Histogram is a graph of a frequency distribution. It consists of vertical rectangles which are attached to each other. Area of rectangle proportional to class frequency.

Example (1)

The following frequency distribution shows distances travelled by 60 sales representatives, during a given period. Represent the data as a form of histogram.

Distance (Km)	No. of People
300 – 399	03
400 – 499	09
500 – 599	15
600 – 699	23
700 – 799	11
800 – 899	05

Example (2)

Following frequency distribution shows ages of 160 workers in a factory. Represent the data as a form of a histogram.

Age (Years)	Frequency <i>f</i>
20 – 24	10
25 – 29	16
30 – 34	22
35 – 39	48
40 – 44	36
45 – 49	20
50 – 54	08

Example (3)

Following freq: table shows boundries of 6 cricket matches. Represent the data as a form of histogram.

Boundaries	Frequency
20 – 30	04
30 – 40	06
40 – 50	02
50 – 70	08
70 – 90	06
90 – 120	06

FREQUENCY POLYGON

A histogram a graph of rectangles. Instead of rectangles it may be preferable to show a single curve rising and failing.

There are two types of frequency polygons can be constructed namely.

- By using histogram
- With-out using histogram

Example (1)

Distances in Km recorded by 120 sales people is given below. Represent the data as a histogram and construct frequency polygon on it.

Distance (Km)	No. of Sales People
400 - 420	12
420 - 440	27
440 - 460	34
460 - 480	24
480 - 500	15
500 - 520	08

Example (2)

Weights of 100 students is given below. Construct a frequency polygon for the data without using histogram.

Weights (Kg)	No. of Students
30 – 39	10
40 – 49	14
50 – 59	26
60 – 69	20
70 – 79	18
80 – 89	12

CUMULATIVE FREQUENCY CURVE

When cumulative frequencies of a distribution are graphed the name ogive (or cumulative frequency curve) is given to the curve obtained. There are two types of ogives can be constructed namely

- (i) Less than method
- (ii) More than method

Example

(Less than method)

The following frequency distribution shows the distances travelled by 70 sales people.

Distance (km)	<i>f</i>
300 – 399	03
400 – 499	09
500 – 599	15
600 – 699	23
700 – 799	11
800 – 899	09

- (i) Construct cumulative frequency distribution (less)
- (ii) Draw a less than ogive
- (iii) Using the above graph find the number of sales people who have travelled less than 550 Km.

Example (2)

The times it took a random sample of runners to complete a race is given below.

House (Min)	f
20 – 29	05
30 – 39	10
40 – 49	36
50 – 59	20
60 – 69	09

- Construct a cumulative frequency distribution (less)
- Draw a less than ogive.

Example (03)

(More than method)

Marks of students from a particular examination is given below.

Exam Marks	20 – 29	30 – 39	40 – 49	50 – 59	60 – 69	70 – 79	80 – 89
No. of Students	01	03	06	06	11	10	08

- Construct a cumulative frequency distribution (more)
- Draw a more than ogive.
- By using the above graph find number of students who scored more than 75 marks.

Example (4)

The table summarizes distances travelled by 150 students to college each day.

Distance (Km)	f
0 – 2	14
3 – 5	24
6 – 8	70
9 – 11	32
12 – 14	08
15 – 17	02

- Construct a cumulative frequency distribution (more)
- Draw a more than ogive.

BAR CHARTS

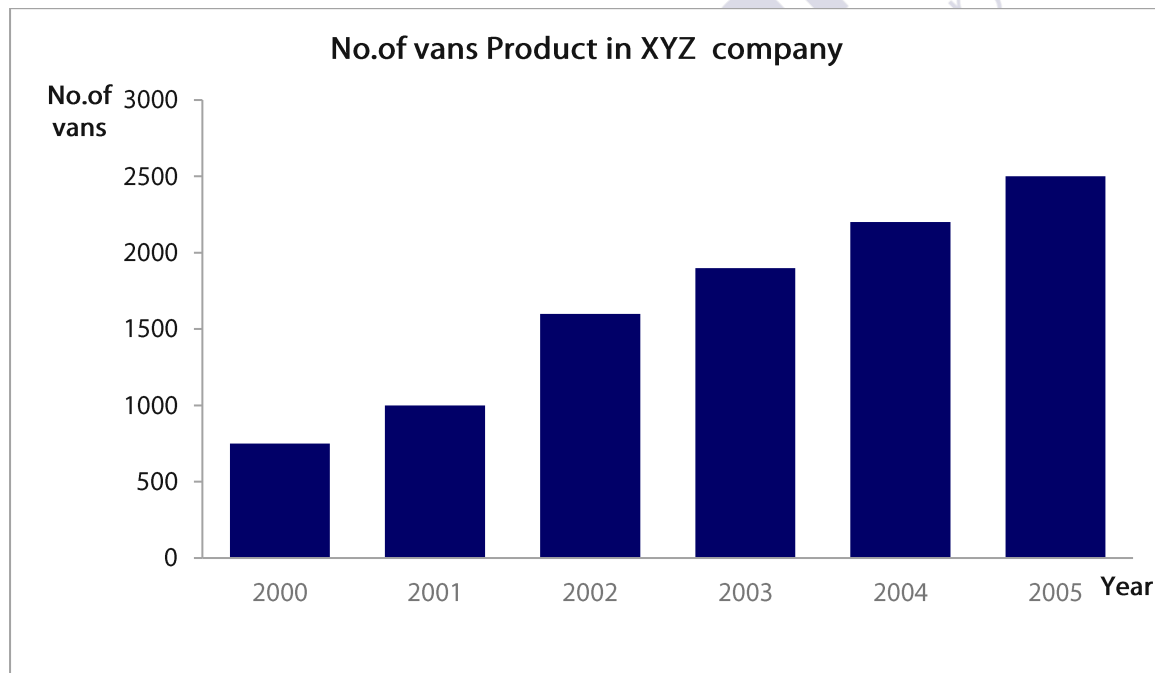
Simple Bar Chart

It consists of vertical bars which are separated from each other. Height of each bar indicating the size of the figure indicated. Simple bar charts should be used where changes in totals only are required.

Worked Example

The number of vans produced each year in XYZ company during the period 2000 – 2005 is given below. Represent the data as a form of simple bar chart.

Year	2000	2001	2002	2003	2004	2005
No. of Vans	750	1000	1600	1900	2200	2500



Example

No. of students registered in a particular institute during 2005 – 2009 is given below. Represent the data as a form of simple bar chart.

Year	2005	2006	2007	2008	2009
No. of Students	200	250	300	400	600

COMPONENT BAR CHART

A component bar chart shows the breakdown of each total into its components. It should be used where changes in totals and indication of the size of each component figure are required.

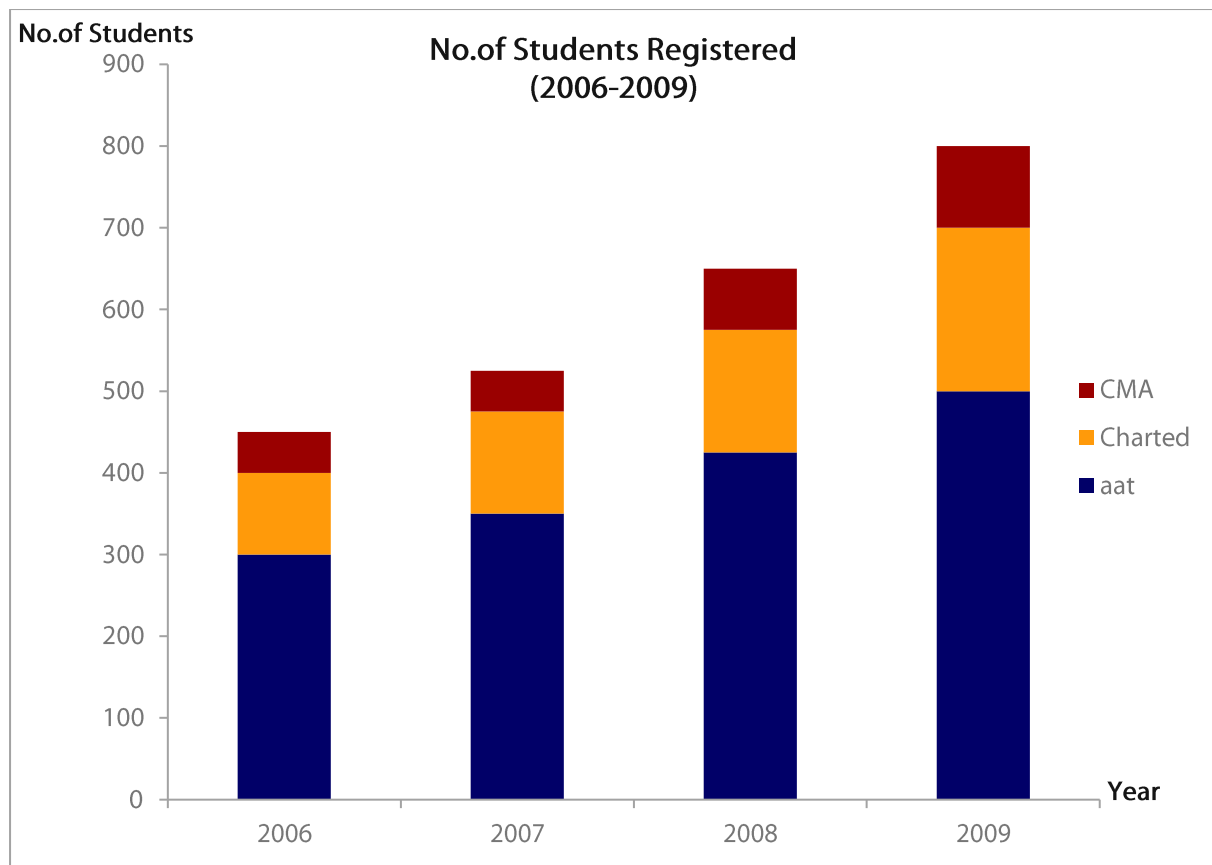
Worked Example

No. of students registered under each course in particular institute is given below. Represent the data as a form of component bar chart.

Course	2006	2007	2008	2009
aat	300	350	425	500
Chartered	100	125	150	200
CMA	50	50	75	100

Prepare a table as follows.

Course	2006		2007		2008		2009	
	No.	cum↑	No.	cum↑	No.	cum↑	No.	cum↑
aat	300	450	350	525	425	650	500	800
Chartered	100	150	125	175	150	225	200	300
CMA	50	50	50	50	75	75	100	100



Example

Sale of two products A and B of a firm for the first four months is given below. represent the data as a form of component bar chart.

Month	Product A	Product B
January	100	160
February	140	140
March	150	160
April	80	170

Percentage Component Bar Chart

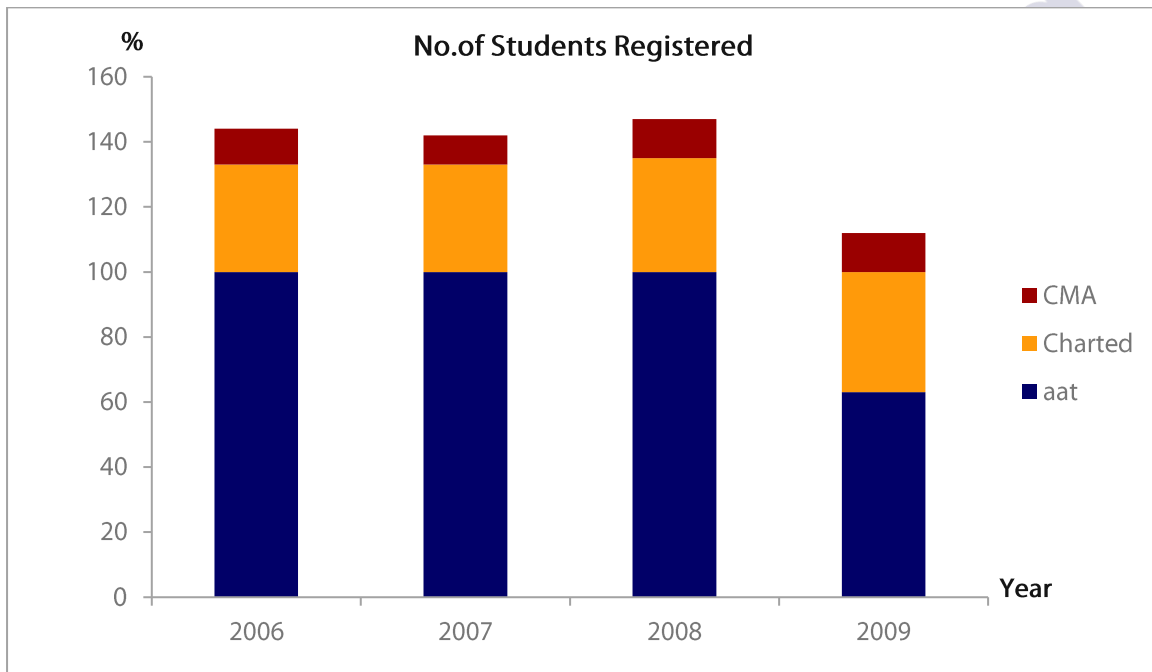
This is similar to the component bar chart, except that a percentage bar chart total magnitude not shown. The height of each bar is the same and this total height represents 100%. It should be used where changes in the relative size only of component figures are required.

Worked Example

Represent the above data as percentage component bar chart.

Prepare a table as follows.

Course	2006		2007		2008		2009	
	%	Cum %	%	Cum %	%	Cum %	%	Cum %
aat	67	100	67	100	65	100	63	63
Chartered	22	33	24	33	23	35	25	37
CMA	11	11	9	9	12	12	12	12
Total	100		100		100		100	



Example

No. of candidates sat for an examination during 2006 – 2009 given below. represent the data as form of a percentage component bar chart.

Year	Boy	Girls
2006	2600	2000
2007	2700	2400
2008	2950	2900
2009	3000	3250

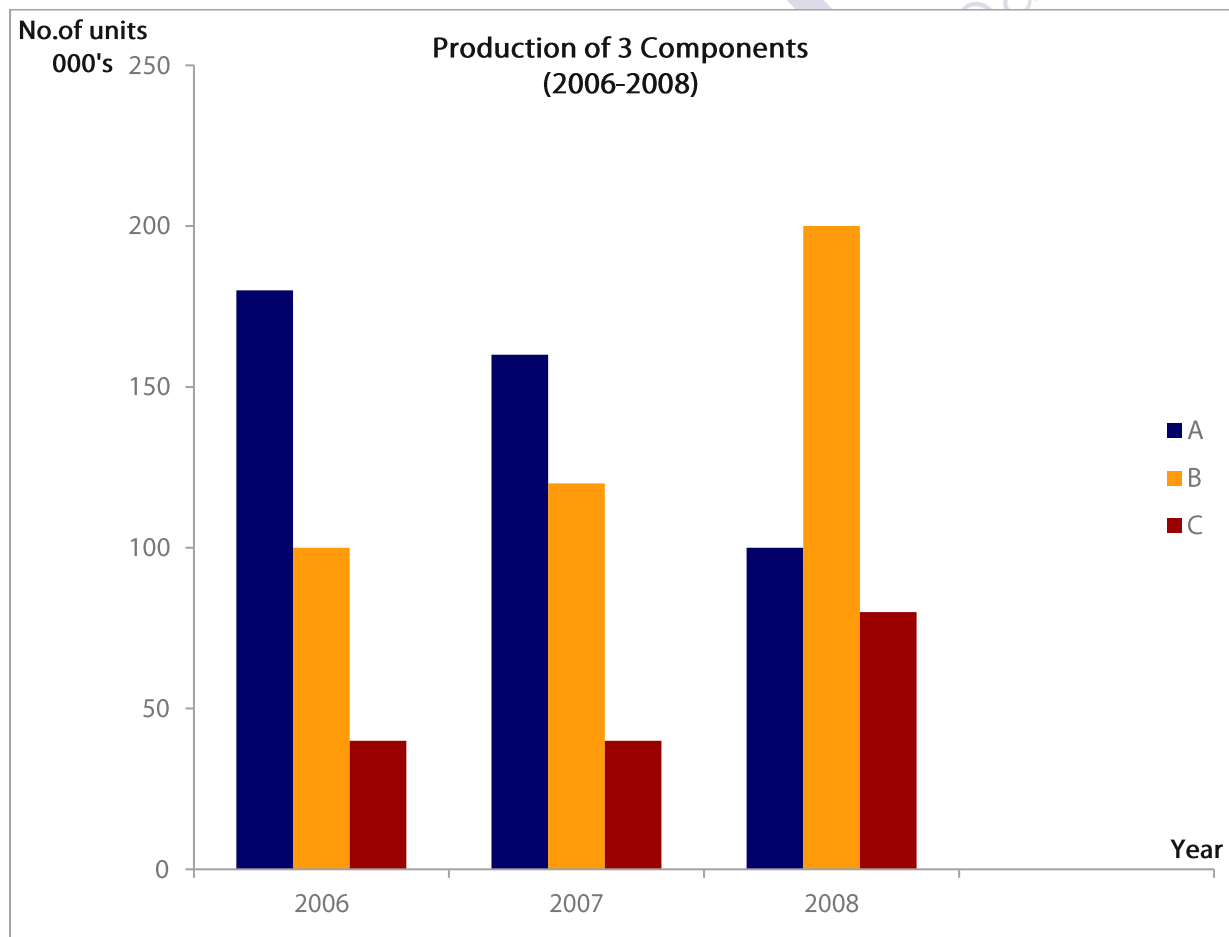
Multiple Bar Chart

In a multiple bar chart the component figures are shown as separate bars adjoining each other. The height of each bar represent the actual value of component. Multiple bar chart should be used where changes in the actual values of the components figures only are required.

Worked Example

Production in thousands units components during 2006 – 2008 is given below. represent the data as a form of a multiple bar chart.

Product	2006	2007	2008
A	180	160	100
B	100	120	200
C	40	40	80



Example

Sales of two products A and B a firm for the first four months is given below. Represent the data as a form of multiple bar chart.

Month	Product A	Product B
January	100	160
February	140	140
March	150	160
April	80	170

PIE CHART

This is often used to present categorical distributions, where a circle is divided into sectors which are proportional in size to the frequencies of corresponding categories. Area of each sector is proportional to the value represented by it.

Worked Example

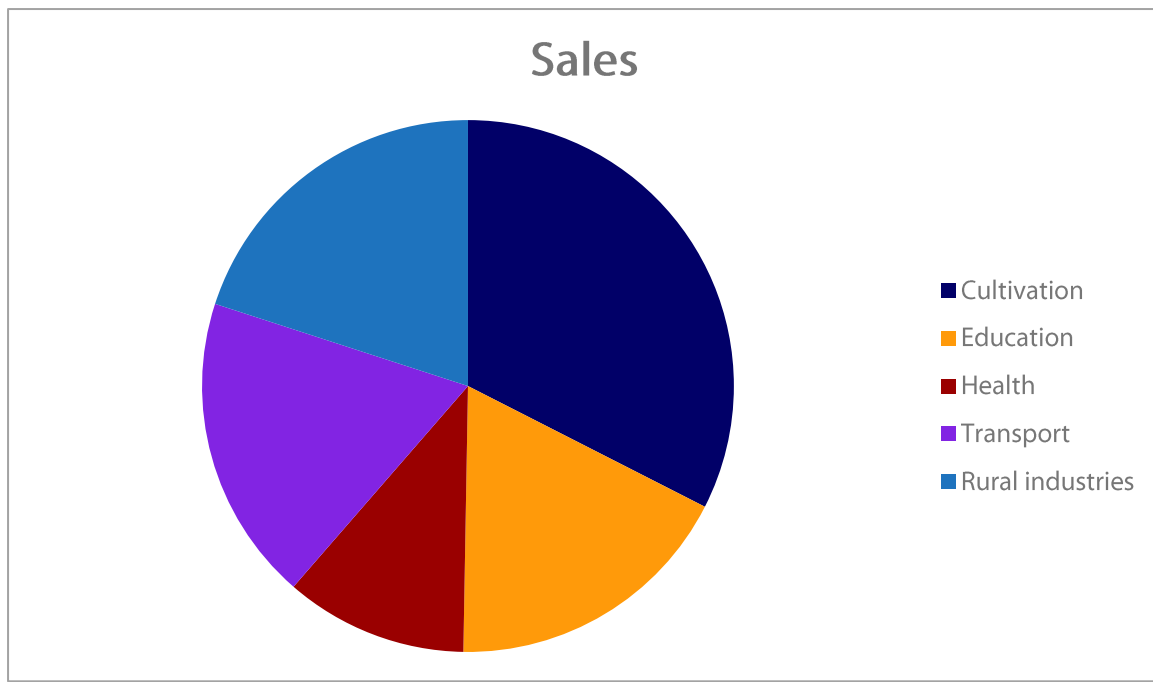
An analysis of expenditure in an area for three months is shown below. Represent the data as a form of pie chart.

Item	Expenditure Rs. '000
Cultivation	175
Education	96
Health	61
Transport	100
Rural industries	108

Prepare a table as follows

Item	Expenditure Rs. '000	Percentage %	No. of Degrees
Cultivation	175	32	117
Education	96	18	64
Health	61	11	40
Transport	100	09	67
Rural industries	108	20	72
Total	540	100	360

Expenditure in an Area



Example

Mean monthly expenditure of families in a certain district is given below. Represent the data as a form of pie chart.

Item	Percentage %
Food	51.2
Education	25.9
Light / Fuel	10.8
Rent	8.2
Other	3.9