PROBALITY AND STATSTICS

BMAT202L

INTRODUCTION

- In the modern world of computers and information technology, the importance of statistics is very well recogonised by all the disciplines.
- Statistics has originated as a science of statehood and found applications slowly and steadily in Agriculture, Economics, Commerce, Biology, Medicine, Industry, planning, education and so on.

The majority of students think that why they are studying statistics and what are the **uses of statistics in our daily life**.

They also want to know the importance of statistics in our daily life. Today, I am going to show you what are the uses of statistics in our daily life.

They also have a concern about what is the job scope of statistics. But first, we need to clear the information about what statistics.

There are millions of definitions of statistics. But we would like to give a short and simple definition of statistics.

DEFINATIONS

- Statistics is defined differently by different authors over a period of time
- Statistics are numerical statement of facts in any department of enquiry placed in relation to each other. - A.L. Bowley
- Statistics may be defined as the science of collection, presentation analysis and interpretation of numerical data from the logical analysis. It is clear that the definition of statistics by Croxton and Cowden is the most scientific and realistic one. According to this definition there are four stages:Collection of Data,Presentation of data,Analysis of data and Interpretation of data. - Croxton and Cowden:

These **statistical problems in real life** are usually based on facts and figures. Sir Ronald Aylmer Fisher, is known as the father of the modern science of statistics.

To understand **what is statistics** better; let's have a look at the example below:-

In this pandemic time of covid-19, statistics is used widely. If we need to determine the number of people who got vaccinated and who are left, we use statistics to obtain this data.

While doing the survey, the surveyors collect data from people to people and then convert this data into useful form with the help of statistical calculations.

DATA ANALYSIS

- Any statistical data can be classified under two categories depending upon the sources utilized.
- Primiary data
- Secondary data.

Primary Data

Primary data is a type of data that is collected by researchers directly from main sources through interviews, surveys, experiments, etc. Primary data are usually collected from the source-where the data originally originates from and regarded as the best kind of data in research.

Example

An organization doing market research about a new product (say phone) they are about to release will need to collect data like purchasing power, feature preferences, daily phone usage, etc. from the target market. The data from past surveys are not used because the product differs.

Primary Data Collection Methods

Primary data collection methods are different ways in which primary data can be collected.

1. Interviews

2. Surveys and questionnaires

3. Observation

4. Questioning

5. Focus groups

6. Experiments

Secondary data

Secondary data are those data which have been already collected and analyzed by some earlier agency for its own use; and later the same data are used by a different agency.

Basic Steps in a Statistical Study:

For any statistical study, there are some basic steps to be followed once we draw a sample. These are:

- **Step 1:** Gather first-hand information from the sample and this is called the raw data.
- **Step 2:** Tabular representation of the raw data, i.e., represent the raw data in a table.
- **Step 3:** Pictorial representation of the data, i.e., draw diagrams with the organized data in a table.
- **Step 4:** Numerically summarize the data, i.e., describe the entire data set with some key numbers.
- **Step 5:** Analyze the data using mathematical formulae.
- **Step 6:** Draw the final inference or conclusion about the population under study.

FREQUENCY DISTRIBUTION

The **frequency** of a value is the number of times it occurs in a dataset. A **frequency distribution** is the pattern of frequencies of a variable. It's the number of times each possible value of a variable occurs in a dataset.

Types of frequency distributions:

1. **Ungrouped frequency distributions:**
2. **Grouped frequency distributions:**

RAW DATA OR UNGROUPED DATA

The statistical data collected are generally raw data or ungrouped data.

Example

Let us consider the daily wages (in Rs) of 30 labourers in a factory. 800, 700, 550, 500, 600, 650, 400, 300, 800, 900, 750, 450, 350, 650, 700, 800, 820, 550, 650, 800, 600, 550, 380, 650, 750, 850, 900, 650, 450, 750.

Discrete (or) Ungrouped frequency distribution

In this form of distribution, the frequency refers to discrete value. Here the data are presented in a way that exact measurement of units are clearly indicated.

Example

In a survey of 40 families in a village, the number of children per family was recorded and the following data obtained.

1	0	3	2	1	5	6	2
2	1	0	3	4	2	1	6
3	2	1	5	3	3	2	4
2	2	3	0	2	1	4	5
3	3	4	4	1	2	4	5

Discrete frequency distribution.

Represent the data in the form of a discrete frequency distribution.

Number of Childern	Frequency
0	3
1	7
2	10
3	8
4	6
5	4
6	2
Total	40

GROUPED OR CONTINUOUS FREQUENCY DISTRIBUTION

Grouped frequency distribution or Continuous frequency distribution:

Again we can arrange it for the class intervals. For this situation, it is called as Grouped frequency distribution of the variable.

Example

Wage distribution of 100 employees

Weekly wages (Rs)	Number of employees
50-100	4
100-150	12
150-200	22
200-250	33
250-300	16
300-350	8
Total	100

Examples:

x	f
15	2
17	3
18	5
20	4
22	7
25	9
30	3

(i)

Frequency distribution
Or
Discrete frequency distribution

x	f
1 - 9	3
10 - 19	5
20 - 29	10
30 - 39	4
40 - 49	7
50 – 59	6
60 - 69	3

(ii)

Grouped Frequency distribution

x	f
5 - 10	2
10 - 15	3
15 -20	5
20 - 25	4
25 - 30	7
30 -35	9
35 - 40	3

(iii)

Continuous Frequency distribution

Special case in grouped frequency distribution:

If "d" is the gap between the upper limit of any class and the lower limit of the succeeding class, the class boundaries for any class are then given by:

$$\textbf{Upper class boundary} = \textbf{Upper class limit} + \frac{d}{2}$$

$$\textbf{Lower class boundary} = \textbf{Lower class limit} - \frac{d}{2}$$

Example:-

x	f
15-19	9
20-24	11
25-29	10
30-34	30
35-39	40

$\Rightarrow$

x	f
14.5-19.5	9
19.5-24.5	11
24.5-29.5	10
29.5-34.5	30
34.5-39.5	40

Measures of Central Tendency

A measure of central tendency is a numerical value around which the measurements have a tendency to cluster.

1. Arithmetic Mean
2. Median
3. Mode

Arithmetic Mean or Average

- For ungrouped data,

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

- If $x_i | f_i, i = 1, 2, \ldots, n$ is the frequency distribution, then Arithmetic Mean (AM) or Average $\bar{x}$ is given by

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

where $N = \sum_{i=1}^{n} f_i$. In case of continuous frequency distribution x_i is taken as the middle value of the corresponding interval.

Problem: Calculate the mean height of the following 10 measurements

Height (in cms): 120, 115, 140, 141, 125, 124, 127, 130, 130, 133

Solution:

$$\sum X = 1285$$

Number of measurements: n = 10

$$\bar{X} = \frac{\sum X}{n} = \frac{1285}{10} = 128.5$$

The mean height is 128.5 cms

Problem: Compute the arithmetic mean of daily wages of workers in a factory.

Worker :	A	B	C	D	E	F	G	H	I	J	K	L
Daily Wages:(in Rs.)	75	60	90	95	80	75	70	65	65	60	75	70

Solution:

We have $\sum X = 880$

n = 12

$$\bar{X} = \frac{\sum X}{n} = \frac{880}{12} = 73.33$$

The arithmetic mean of daily wages of workers is Rs. 73.33.

Problem: The following data gives the number of children born to 350 women.

No. of children :	0	1	2	3	4	5	6
No. of women :	171	82	50	25	13	7	2

Calculate the mean number of children born per woman.

Solution:

No. of children (x)	No. of women (f)	fx
0	171	0
1	82	82
2	50	100
3	25	75
4	13	52
5	7	35
6	2	12
Total	350	356

From the table, we have,

$\Sigma fx = 356 \quad N = 350$

$$\bar{X} = \frac{\sum fx}{N} = \frac{356}{350}$$

$\bar{X} = 1.017$

$\therefore$ The mean no. of children born to a woman =1.017

Problem: The following data relates to the marks of 100 students in statistics. Calculate the A.M. marks of students.

Marks	20-30	30-40	40-50	50-60	60-70	70-80
No. of students	7	13	20	30	18	12

Solution: **(a) Direct Method :**

Marks	No. of students (f)	Mid value (x)	fx
20-30	7	25	175
30-40	13	35	455
40-50	20	45	900
50-60	30	55	1650
60-70	18	65	1170
70-80	12	75	900
Total	100	-	5250

From table we have, $\sum fx = 5250$, $N = 100$

$$x = \frac{\sum fx}{N} = \frac{1520}{100} = 52.5$$

Step Deviation Method for Computing Arithmetic Mean

It may be pointed out that the above can be used conveniently if the values of X or/and f are small. However, if the values of X or/and f are large, the calculation of mean by the above method is quite tedious and time consuming. In such a case calculations can be reduced to a great extend by using the step deviation method which consist in taking deviations (differences) of given observations from any assumed mean A.

Let $d = X - A$

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

Ex Find the airthmetic mean

X	No of Student f	$(X - A)$ → 40 d	fd
20	8	-20	-160
30	12	-10	-120
40 ← A	20	0	0
50	10	10	100
60	6	20	120
70	4	30	120
	$N = 60$		$\Sigma fd = 60$

$$\bar{x} = A + \frac{\Sigma fd}{N} = 40 + \frac{60}{60} = 41$$

Ex

Marks	mid-point m	No of Stud f	(m-A) = d	fd
0-10	5	5	-30	-150
10-20	15	10	-20	-200
20-30	25	25	-10	-250
30-40	35 → A	30	0	0
40-50	45	20	10	200
50-60	55	10	20	200
		$\Sigma f = N = 100$		$\Sigma fd = -200$

$$\bar{x} = A + \frac{\Sigma fd}{N} = 35 + \frac{(-200)}{100} = 33$$

Simplified Formula

In case of grouped or continuous frequency distribution, with class intervals of equal magnitude, the calculations are further simplified by taking :

$$d = \frac{X - A}{h}$$

where X is the mid-value of the class and h is the common magnitude of the class intervals. Then

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

Example 5·3. *Calculate the mean for the following frequency distribution :*

Marks :	*0—10*	*10—20*	*20—30*	*30—40*	*40—50*	*50—60*	*60—70*
Number of students :	*6*	*5*	*8*	*15*	*7*	*6*	*3*

(i) By the direct formula. ; (ii) By the step deviation method.

Solution.

COMPUTATION OF ARITHMETIC MEAN

Marks	*Mid-value (X)*	*Number of Students (f)*	fX	$d = \frac{X-35}{10}$	fd
0—10	5	6	30	–3	–18
10—20	15	5	75	–2	–10
20—30	25	8	200	–1	–8
30—40	35	15	525	0	0
40—50	45	7	315	1	7
50—60	55	6	330	2	12
60—70	65	3	195	3	9
		$N = \sum f = 50$	$\sum fX = 1670$		$\sum fd = -8$

(*i*) Direct Formula : Mean $(\bar{X}) = \frac{\sum fX}{\sum f} = \frac{1670}{50} = 33{\cdot}4$ marks.

(*ii*) Step Deviation Method : In the usual notations we have $A = 35$ and $h = 10$.

$$\therefore \quad \bar{X} = A + \frac{h\sum fd}{N} = 35 + \frac{10 \times (-8)}{50} = 35 - 1{\cdot}6 = 33{\cdot}4 \text{ marks.}$$

(b) Step Deviation Method :

Let A = 55 and C = Length of class = 10

Marks	No: of students	Mid value x	$d = \frac{X - A}{C}$	fd
20-30	7	25	-3	-21
30-40	13	35	-2	-26
40-50	20	45	-1	-20
50-60	30	55 = A	0	0
60-70	18	65	1	18
70-80	12	75	2	24
Total	100	-	-	-25

From table we have $\sum fd = -25$, $C = 10$, A = 55, $N = 100$

$$\bar{X} = A + \left(\frac{\sum fd}{N}\right)C$$

$$= 55 + \left(\frac{-25}{100}\right)10 = 55 - 2.5 = 52.5$$

Average Marks of Students = 52.5

Problem: compute arithmetic mean for the following frequency distribution:

49.5 - 59.5 59.5-69.5

Class :	50-59	60-69	70-79	80-89	90-99	100-109	110-119
Frequency :	1	3	8	17	35	4	2

Solution: Given, C =Common length of class internals = 10

Class	Frequency (f)	Mid value X	$d = \frac{X-A}{C}$	fd
50-59	1	54.5	-3	-3
60-69	3	64.5	-2	-6
70-79	8	74.5	-1	-8
80-89	17	84.5 = A	0	0
90-99	35	94.5	1	35
100-109	4	104.5	2	8
110-119	2	114.5	3	6
Total	70	-	-	32

From the table, we have, $\sum fd = 32$, $C = 10$, $A = 84.5$, $N = 70$

$$\bar{X} = A + \left(\frac{\sum \text{fd}}{N}\right)C = 84.5 + \left(\frac{32}{70}\right)10$$

$$= 84.5 + 4.5714$$

$\therefore$ Arithmetic Mean = $\bar{X}$ = 89.0714

* If the class of intervals are unequal we can simplify calculations by taking a common factor. In such case we should use $\frac{m-A}{c}$ instead of $\frac{m-A}{h}$ while calculating

Ex

Marks	Mid-point	f	$\frac{m-45}{c}=d$	fd
0-10	5	5	-8	-40
10-30	20	12	-5	-60
30-60	45 → A	25	0	0
60-100	80	8	7	56
		N=50		$\Sigma fd = -44$

$$\bar{x} = A + \frac{\Sigma fd}{N} \times c = 45 - \frac{44}{50} \times 5 = 40.6$$

Problem: For a certain frequency table, which has been partly reproduced here, the mean was found to be 1.46.

No. of Accidents	0	1	2	3	4	5	Total
Frequency	46	?	?	25	10	5	200

Find the missing frequencies.

Let f_1 and f_2 be missing frequencies, then

No. of accidents	*Frequency*	*fx*
0	46	0
1	f_1	f_1
2	f_2	$2f_2$
3	25	75
4	10	40
5	5	25
Total	$86+f_1+f_2$	$140+f_1+2f_2$

From table we have, $\sum f = N = (86 + f_1 + f_2)$

$\sum fx = (140 + f_1 + 2f_2)$

But given that

$\sum f = N = 200$

Also $\bar{X} = \frac{\sum fx}{N} = 1.46$

$\therefore \sum fx = N\,\bar{X} = (200)(1.46) = 292$

$\therefore$ We have $86 + f_1 + f_2 = 200$ (1)

and $140 + f_1 + 2f_2 = 292$ (2)

$f_1 = 76$ and $f_2 = 38$

Problem: Find the missing frequency from the following data, given the average mark is 16.82

Marks	Frequency
0-5	10
5-10	12
10-15	16
15-20	f_4
20-25	14
25-30	10
30-35	8

MEDIAN

- Median of distribution is the value of the variable which divides it into two equal parts
- It is the value which exceeds and is exceeded by the same number of observations
- Median is the value such that the number of observations above it is equal to the number of observations below it
- The median is thus a positional average

- In case of ungrouped data, if the number of observations is odd then median is the middle value after the values have been arranged in ascending or descending order of magnitude.
- In case of even number of observations, there are two middle terms and median is obtained by taking the arithmetic mean of the middle terms.
- For example, the median of the value 25, 20, 15, 35, 18, i.e., 15, 18, 20, 25, 35 is 20
- The median of 8, 20, 50, 25, 15, 30, i.e., of 8, 15, 20, 25, 30, 50 is $\frac{1}{2}(20 + 25) = 22.5$.

Median of Discrete Frequency Distribution

In case of discrete frequency distribution median is obtained by considering the cumulative frequencies. The steps for calculating median are given below:

1. Find $N/2$, where $N = \sum_i f_i$
2. See the (less than) cumulative frequency (cf.) just greater than $N/2$
3. The corresponding value of x is median

Example

Obtain the median for the following frequency distribution:

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

Solution:

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

Hence, $N = 120 \implies N/2 = 60$. Cumulative frequency (cf.) just greater than $N/2$ is 65 and the value of x corresponding to 65 is 5. Therefore, median is 5.

Median of continuous frequency distribution

In the case of continuous frequency distribution, the class corresponding to the cf. just greater than $N/2$ is called the median class and the value of median is obtained by the following formula :

$$Median = l + \frac{h}{f}\left(\frac{N}{2} - c\right)$$

where l is the lower limit of the median class, f is the frequency of the median class, h is the magnitude of the median class, c is the cf. of the class preceding the median class, and $N = \sum_i f_i$

Example

Find the median wage of the following distribution:

Wages (in Rs.) :	20–30	30–40	40–50	50–60	60–70
No. of labourers :	3	5	20	10	5

Solution:

Wages (in Rs.)	No. of labourers	c.f.
20–30	3	3
30–40	5	8
40–50	20	28
50–60	10	38
60–70	5	43

Here $N/2 = 43/2 = 21.5$.

Cumulative frequency just greater than 21.5 is 28 and the corresponding class is 40–50. Thus median class is 40–50.

$$Median = 40 + (10/20)(21.5 - 8) = 40 + 6.75 = 46.75.$$

Thus median wage is Rs. 46.75.

EXAMPLE

Ex Find the median marks (4)

Marks	45-50	40-45	35-40	30-35	25-30
No of Studen	10	15	26	30	42

Marks.	20-25	15-20	10-15	5-10
No. of Studen	31	24	15	7

Soln First arrange the data in ascending order

Marks	No of Stude (f)	cf
~~45-50~~		
5 - 10	7	7
10 - 15	15	22
15-20	24	46
20-25	31	77
25-30	42	119 ←
30 - 35	30	149
35-40	26	175
40-45	15	190
45-50	10	200
	N=200	

medium = Size of $\frac{N}{2}$th item

$= \frac{200}{2} = 100^{th}$ item

medium lies in the class 25-30

$$\text{Medium} = L + \frac{h}{f}\left(\frac{N}{2} - cf\right)$$

$L = 25,\ h = 5,\ f = 42$

$N = 200 \quad cf = 77$

$$\text{medium} = 25 + \frac{5}{42}\left(\frac{200}{2} - 77\right)$$

$$= 25 + \frac{5}{42}(23)$$

$$= 27.74$$

EXAMPLE

Q. Calculate median for the following data

Weight (gm)	No. of Apples	Weight (gm)	No. of Apples
410 - 419	14	450 - 459	45
420 - 429	20	460 - 469	18
430 - 439	42	470 - 479	7
440 - 449	54		

Weight	f	cf
409.5 – 419.5	14	14
419.5 – 429.5	20	34
429.5 – 439.5	42	76
439.5 – 449.5	54	130 ←
449.5 – 459.5	45	175
459.5 – 469.5	18	193
469.5 – 479.5	7	200
	N = 200	

Med = Size of $\frac{N}{2}$ th item

$= \frac{200}{2} = 100^{th}$

⇛ median lies in the class 439.5 – 449.5

$L = 439.5, \quad h = 10$

$cf = 76, \quad f = 54$

$$Med = L + \frac{h}{f}\left(\frac{N}{2} - cf\right)$$

$$= 439.5 + \frac{10}{54}\left(\frac{200}{2} - 76\right)$$

$$= 439.5 + 4.44$$

$$= 443.94$$

EXAMPLE

Ex. An incomplete distribution is given by (6)

Variable	0-10	10-20	20-30	30-40	40-50	50-60
Freq	10	20	?	40	?	25

Variable	60-70
Frq.	15

Find out missing frq. ?

Total freq. is 170 (Given)

Median " 35 (Given)

Solⁿ Let missing freq of class 20-30 is f_1

& " " " " 40-50 is f_2

$\Rightarrow 10 + 20 + f_1 + 40 + f_2 + 25 + 15 = 170$

$\Rightarrow f_1 + f_2 = 60$

$\text{median} = L + \frac{h}{f}\left(\frac{N}{2} - cf\right)$

median = Size of $\frac{N}{2}$th item $= \frac{170}{2} = 85^{th}$ item

Median is 35 which lies in the class 30-40

$\Rightarrow L = 30, \ \frac{N}{2} = 85, \ cf = 10 + 20 + f_1, \ h = 10$

$f = 40$

$35 = 30 + \frac{10}{40}(85 - 10 - 20 - f_1)$

$f_1 = 35$

$\Rightarrow f_2 = 60 - 35$

$= 25$

Example

An incomplete frequency distribution is given as follows

Variable	Frequency	Variable	Frequency
10–20	12	50-60	?
20–30	30	60 70	25
30–40	?	70-80	18
40–50	65	Total	229

Given that the median value is 46, determine the missing frequencies using the median formula.

Solution: Let the frequency of the class 30–40 be f_1 and that of 50–60 f_2. Then,

$$f_1 + f_2 = 229 - (12 + 30 + 65 + 25 + 18) = 79.$$

Since median is given to be 46, the class 40–50 is the median class. Hence using median formula

we get

$$46 = 40 + \frac{114.5 - (12 + 30 + f_1)}{60} x10$$

which gives $f_1 = 34$ and $f_2 = 45$, since frequency never be fractional and $f_1 + f_2 = 79$.

Examples

Example 5·24. *Find the missing frequency from the following distribution of daily sales of shops, given that the median sale of shops is Rs. 2,400.*

Sale in hundred Rs. :	0–10	10–20	20–30	30–40	40–50
No. of shops :	5	25	–	18	7

Example 5·25. *In the frequency distribution of 100 families given below, the number of families corresponding to expenditure groups 20–40 and 60–80 are missing from the table. However, the median is known to be 50. Find the missing frequencies.*

Expenditure :	*0–20*	*20–40*	*40–60*	*60–80*	*80–100*
No. of families :	*14*	*?*	*27*	*?*	*15*

MODE

- Mode is the value which occurs most frequently in a set of observations and around which the other items of the set cluster densely.
- Mode is the value of the variable which is predominant in the series.

Definition (Discrete frequency distribution)

In the case of discrete frequency distribution mode is the value of x corresponding to maximum frequency.

EXAMPLES

For example, in the following frequency distribution:

x:	1	2	3	4	5	6	7	8
f :	4	9	16	25	22	15	7	3

the value of x corresponding to the maximum frequency, viz., 25 is 4. Hence mode is 4.

But in anyone (or-more) of the following cases :

- if the maximum frequency is repeated
- if the maximum frequency occurs in the very beginning or at the end of the distribution and
- if there are irregularity in the distribution, the value of mode is determined by the method of grouping. which is illustrated below by an example.

DISCRETE FREQUENCY DISTRIBUTION

Example

Find the mode of the following frequency distribution:

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

Size (x)	Frequency (i)	(ii)	(iii)	(iv)	(v)	(vi)
1	3	11		26		
2	8		23		46	
3	15	38				73
4	23		58	98		
5	35	75			107	
6	40		72			100
7	32	60		80		
8	28		48		93	
9	20	65				79
10	45		59	65		
11	14	20				
12	6					

The frequencies two by two after leaving the first two frequencies results - in a repetition of column *(U)*. Hence. we proceed to combine the frequencies three by three. Thus getting column *(vi)*. The combination of frequencies three by three after leaving the first frequency results in column *(v)* and after leaving the first two frequencies results in column *(vi)*

The maximum frequency in each column is given in black type. To find mode we form the following table :

ANALYSIS TABLE

Column Number (1)	*Maximum Frequency (2)*	*Value or combination of values of x giving max. frequency in (2) (3)*
(i)	45	10
(ii)	75	5, 6
(iii)	72	6, 7
(iv)	98	4, 5, 6,
(v)	107	5, 6, 7
(vi)	100	6, 7, 8

On examining the values in column (3) above, we find that the value 6 is repeated the maximum number of times and hence the value of mode is 6 and not 10 which is an irregular item.

CONTINUOUS FREQUENCY DISTRIBUTION

In case of continuous frequency distribution. Mode is given by the formula :

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

Here l is the lower limit, h the magnitude and f_1 the frequency of the modal class, f_0 and f_2 are the frequencies of the classes preceding and succeeding the modal class respectively.

Example 2·10. *Find the mode for the following distribution :*

Class - interval :	0–10	10–20	20–30	30–40	40–50	50–60	60–70	70–80
Frequency :	5	8	7	12	28	20	10	10

Solution: Maximum frequency is 28. Thus the class 40-50 is the modal class.

- $l = 40$, the lower limit of the modal class
- $h = 10$, the magnitude
- $f_1 = 28$, the frequency of the modal class
- $f_0 = 12$ and $f_2 = 20$

Answer=46.67 (approx.).

$$\text{Mode} = 40 + \frac{10(28-12)}{(2\times 28 - 12 - 20)} = 40 + 6{\cdot}666 = 46{\cdot}67$$

Example 2·11. *The Median and Mode of the following wage distribution are known to be Rs. 33·50 and Rs. 34 respectively. Find the values of f_3, f_4 and f_5.*

Wages : (in Rs.)	*0–10*	*10–20*	*20–30*	*30–40*	*40–50*	*50–60*	*60–70*	*Total*
Frequency :	*4*	*16*	f_3	f_4	f_5	*6*	*4*	*230*

Solution.

CALCULATIONS FOR MODE AND MEDIAN

Wages (in Rs.)	*Frequency (f)*	*Less than c.f.*
0—10	4	4
10—20	16	20
20—30	f_3	$20 + f_3$
30—40	f_4	$20 + f_3 + f_4$
40—50	f_5	$20 + f_3 + f_4 + f_5$
50—60	6	$26 + f_3 + f_4 + f_5$
60—70	4	$30 + f_3 + f_4 + f_5$
Total	$230 = 30 + f_3 + f_4 + f_5$	

Since median is 33.5, which lies in the class 30-40, 30-40 is the median class. Using the median formula $Median = l + \frac{h}{f}\left(\frac{N}{2} - c\right)$, we get

$$f_3 = 95 - 0.35 f_4$$

Mode being 34, the modal class is also 30-40. Using mode formula $Mode = l + \frac{h(f_4 - f_3)}{2f_4 - f_3 - f_5}$, we get

$$34 = 30 + \frac{10(f_4 - f_3)}{2f_4 - f_3 - f_5}.$$

By applying $f_3 = 95 - 0.35 f_4$ and $200 - f_4 = -f_3 - f_5$, we have $f_4 = 100$ $f_3 = 60, f_5 = 40$

Ex Calculate the mode

Marks	No of Students	Marks	No. of Students
Above 0	80	Above 60	28
Above 10	77	Above 70	16
Above 20	72	Above 80	10
Above 30	65	Above 90	8
Above 40	55	Above 100	0
Above 50	43		

Sol. Since cumulative freq. is given, therefore first convert into a simple freq. distribution

Marks	No. of Students	Marks	No. of Students
0 - 10	3	50-60	15
10 -20	5	60-70	12
20-30	7	70-80	6
30-40	10	80-90	2
40-50	12	90-100	8

⇒ modal class is 50-60

$$\text{mode} = 50 + \frac{15-12}{2\times15-12-12} \times 10 = 55$$

3. If the method of grouping gives the modal class which does not correspond to the maximum frequency f_1 *i.e.*, the frequency of modal class is not the maximum frequency, then in some situations we may get $2f_1 - f_0 - f_2 = 0$. [This will not be possible if f_1 is maximum and f_0 and f_2 are less than f_1]. In such a situation *viz.*, $2f_1 - f_0 - f_2 = 0$, the value of mode cannot be computed by the formula.

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

as it gives $\quad \text{Mode} = l + \infty = \infty. \qquad [\because 2f_1 - f_0 - f_2 = 0]$

In such cases, the value of mode can be obtained by the formula :

$$\text{Mode} = l + \frac{h|f_1 - f_0|}{|f_1 - f_0| + |f_1 - f_2|} \qquad \ldots(5\cdot 18a)$$

where $|A|$ represents the absolute (positive) value of A.

Formula ($5\cdot 18a$) is only an approximate formula and does not give very correct result because further grouping of classes, say, 4 at a time may give different value of the modal class and as such a different result.

As an illustration, for the following data :

X :	10–20	20–30	30–40	40–50	50–60	60–70	70–80	80–90	90–100	100–110
f :	4	6	5	10	20	22	24	6	2	1

the usual method of grouping (up to 3 classes at a time) will give 60–70 as the modal class such that : $f_1 = 22, f_0 = 20, f_2 = 24$ and therefore, $2f_1 - f_0 - f_2 = 44 - 20 - 24 = 0$. Hence, usual formula for mode cannot be applied. Using (5·18*a*), an approximate value of mode may be obtained as :

$$Mo = 60 + \frac{10\,|22-20|}{|22-20|+|22-24|} = 60 + \frac{10 \times 2}{2+2} = 60 + 5 = 65$$

Karl Pearson Relationship

Sometimes mode is estimated from the mean and the median. For a symmetrical distribution, mean, median and mode coincide. If the distribution is moderately asymmetrical, the mean, median and mode obey the following empirical relationship (due to Karl Pearson) :

The distance between mean and median is about one-third of the distance between the mean and mode

$$Mean - Median = \frac{1}{3}(Mean - Mode)$$

which gives $Mode = 3Median - 2Mean$

Relation between Mean, Median, Mode

1 In symmetrical distribution $Mean = Median = Mode$

2 In positively skewed distribution $Mode < Median < Mean$

3 In negatively skewed distribution $Mean < Median < Mode$

RELATIONSHIP BETWEEN ARITHMETIC MEAN, MEDIAN AND MODE

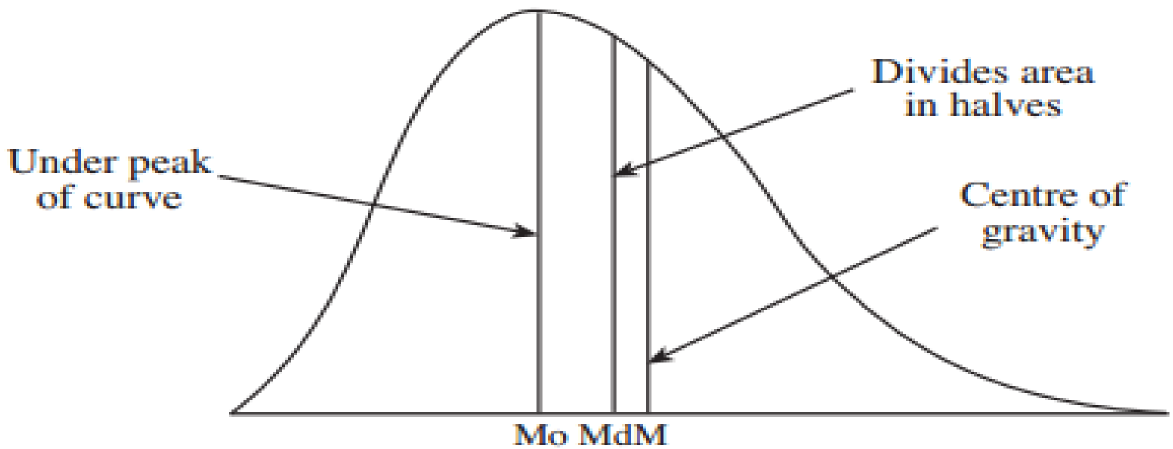


Fig. 5·3.

Problem: In a moderately skewed distribution (Asymmetrical distribution) A.M. = 15 and Mode = 12. Find the value of the Median of the given distribution.

Solution : Consider the empirical relationship between mean, median and mode.

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

Hence, 15 - 12 = 3 (15 - Median)

Therefore, 3 Median = 45 – 3 = 42

$$\text{Median} = \frac{42}{3} = 14$$

Problem: Find the Mode for the following data

2, 5, 3, 2, 1, 4, 6, 3, 7

- ***Solutions:*** Since 2 and 3 have maximum frequencies. So the given data is bio-modal data. Therefore we use empirical relationship of mean median and mode.

$$A.M = \bar{x} = \frac{\sum x}{n} = \frac{33}{9} = 3.6667$$

To find median arrange the data in ascending order:

1, 2, 2, 3, 3, 4, 5, 6, 7

Median = 3

$$3.6667 - \text{Mode} = 3\ (3.667-3)$$

$$3.6667 - \text{Mode} = 2.0001$$

$$\therefore \text{Mode} = 3.667 - 2.0001 = 1.6666$$

Problem: Compute Mode for the following data :

Size	0-4	4-8	8-12	12-16	16-20
Frequency	10	20	30	35	35

Solution: The highest frequency is 35 and it corresponds the two bottom most class intervals frequency table. Hence, the given distributions is a bimodal. In this case, we use empirical relationship between A.M., Median and Mode.

Class	Frequency f	Mid values x	fx	Less than cumulative frequency
0-4	10	2	20	10
4-8	20	6	120	30
8-12	30	10	300	60 =m
12-16	35 = f	14	490	95
16-20	35	18	630	130
Total	130	-	1560	-

$$\text{A.M.} = \bar{x} = \frac{\sum fx}{N} = \frac{1560}{130} = 12$$

$$\text{Median} = \text{L} + \left(\frac{\frac{N}{2} - m}{f}\right)C$$

$$= 12 + \left[\frac{65 - 60}{35}\right]4 = 12 + \frac{20}{35} = 12 + 0.5714 = 12.5714$$

Median = 12.5714

Consider the empirical relationship

(A.M. - Mode) = 3 (Mean - Median)

12 – Mode = 3 (12 – 12.5714)

12 – Mode = 3 (- 0.5714)

12 – Mode = - 1.7142

$\therefore$ Mode = 12 + 1.7142

Mode = 13.7142

Ex Determine the modal weight

Weight	No. of persons	Weight	No. of persons
100 – 110	4	140 – 150	33
110 – 120	6	150 – 160	17
120 – 130	20	160 – 170	8
130 – 140	32	170 – 180	2

Grouping method [Since it is difficult to say which is the modal class, hence use grouping method]

Weight	f(I)	II	III	IV	V	VI
100 - 110	4	10		30		
110 - 120	6		26		58	
120 - 130	20	52				85
130 - 140	32		65	82		
140 - 150	33	50			58	
150 - 160	17		25			27
160 - 170	8	10				
170 - 180	2					

Analysis table

Colm	120 - 130	130 - 140	140 - 150
I	—	—	1
II	1	1	—
III	—	1	1
IV	—	1	1
V	1	1	1
VI	1	1	1
Total	3	5	5

This is a bi-modal series.

Hence use the formula to determine mode

$$\text{Mode} = 3\ \text{Median} - 2\ \text{Mean}$$

GEOMETRIC MEAN AND HARMONIC MEAN

Definition (Geometric Mean)

Geometric mean of a set of n observations is the n^{th} root of their product. Thus the geometric mean GM of n observations $x_i, i = 1, 2, \ldots, n$ is

$$GM = (x_1 x_2 \ldots x_n)^{1/n}$$

Definition (Harmonic Mean)

Harmonic mean of n number of observations is the reciprocal of the arithmetic mean of the reciprocals of the given values. Thus the harmonic mean HM of n observations $x_i, i = 1, 2, \ldots, n$ is

$$HM = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}(1/x_i)}.$$

Partition Values

These are the values which divide the series into a number of equal parts.

- The three points which divide the series into four equal parts are called quartiles
- The nine points which divide the series into ten equal parts are called deciles
- The ninety-nine points which divide the series into hundred equal parts are called percentiles

Quartiles

First Quartile (Q_1) = Size of $\frac{N+1}{4}$ th item (Discrete series)

Q_1 = Size of $\frac{N}{4}$ th item (Continuous series)

$$Q_1 = l + \frac{\frac{N}{4} - c.f.}{f} \times i$$

Third Quartile (Q_3) = Size of $3\left(\frac{N+1}{4}\right)$ th item (Discrete series)

Q_3 = Size of $\frac{3N}{4}$ th item. (Continuous series)

$$Q_3 = l + \frac{\frac{3N}{4} - c.f.}{f} \times i$$

In quartiles, The first, second and third points are known as the first, second and third quartiles respectively. The first quartile, Q_1 is the value which exceed 25% of the observations and is exceeded by 75% of the observations. The second quartile, Q_2, coincides with median. The third quartile, Q_3 , is the point which has 75% observations before it and 25% observations after-it.

- $$Q_1 = l + \frac{h}{f}\left(\frac{N}{4} - c\right)$$
- $$Q_2 = l + \frac{h}{f}\left(\frac{N}{2} - c\right)$$
- $$Q_3 = l + \frac{h}{f}\left(\frac{3N}{4} - c\right)$$

where l is the lower limit of the Q_i class, f is the frequency of the Q_i class, h is the magnitude of the Q_i class, c is the cf. of the class preceding the Q_i class, and $N = \sum_i f_i$

Example 3: Calculate Q_1 and Q_3 for the following data.

Roll No.	1	2	3	4	5	6	7
Marks	20	28	40	12	30	15	50

Solution: Marks in ascending order 12 15 20 28 30 40 50

Q_1 = Size of $\frac{N+1}{4}$ th item = Size of $\frac{7+1}{4}$ = 2nd item.

Size of 2nd item is 15. Hence $\boldsymbol{Q_1 = 15}$

Q_3 = Size of $3\left(\frac{N+1}{4}\right)$ th item = Size of $3\left(\frac{7+1}{4}\right)$ = 6th item.

Size of 6th item is 40. Hence $\boldsymbol{Q_3 = 40}$.

Example 4: Compute the value of Q_1 and Q_3 for following data:

C.I.	10-20	20-30	30-40	40-50	50-60	60-70	70-80
f	12	19	5	10	9	6	6

Solution:

Marks	***Frequency***	***Cumulative Frequency***
10-20	12	12
20-30	19	31
30-40	5	36
40-50	10	46
50-60	9	55
60-70	6	61
70-80	6	67
	$N = 67$	

Q_1 = Size of $\frac{N}{4}th$ item = Size of $\frac{67}{4}$ = 16.75th item.

Q_1 - lies in the interval **20-30**

$$Q_1 = l + \frac{\frac{N}{4} - c.f.}{f} \times i$$

$l = 20,\ N/4 = 16.75,\ c.f. = 12$

$f = 19,\ i = 10$

$$Q_1 = 20 + \frac{\frac{67}{4} - 12}{19} \times 10 = 20 + 2.5 = 22.5$$

Hence $\mathbf{Q_1 = 22.5}$

Q_3 = Size of $\frac{3N}{4}$ th item = Size of $\frac{3\times 67}{4}$ = 50.25th item.

Q_3 - lies in the class **50-60**.

$$Q_3 = l + \frac{\frac{3N}{4} - c.f.}{f} \times i$$

$l = 50,\ 3N/4 = 50.25,\ c.f. = 46$

$f = 9,\ i = 10$

$$Q_3 = 50 + \frac{50.25-46}{9} \times 10 = 50+4.72 = 54.72$$

Hence **Q_3 = 54.72**

$\boldsymbol{D_4}$ = Size of $\frac{4N}{10}$ th item = $\frac{4\times67}{10}$ = 26.8 th item.

$\boldsymbol{D_4}$ lies in the interval of 20-30.

$$\boldsymbol{D_4} = l + \frac{\frac{4N}{10} - c.f.}{f} \times i$$

$l = 20$, $4N/10 = 26.8$, $c.f. = 12$

$f = 19$, $i = 10$

$\boldsymbol{D_4}$ = 27.79

$\boldsymbol{P_{60}}$= Size of $\frac{60N}{100}$ th item = $\frac{60\times67}{100}$ = 40.2 th item

$\boldsymbol{P_{60}}$ lies in the interval of 40-50.

$$\boldsymbol{P_{60}} = l + \frac{\frac{60N}{100} - c.f.}{f} \times i$$

$l = 40$, $4N/10 = 40.2$, $c.f. = 36$

$f = 10$, $i = 10$

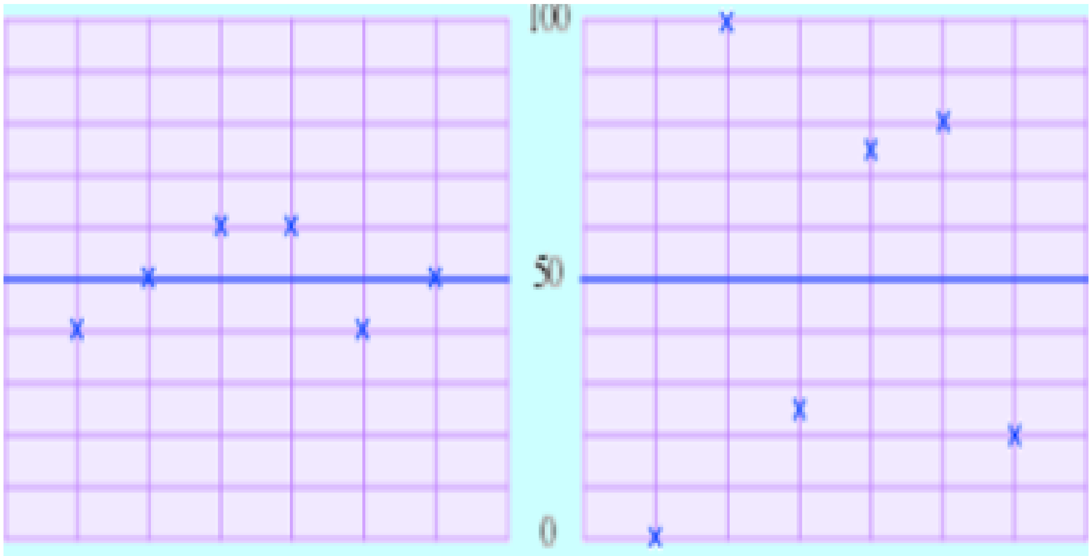
$\boldsymbol{P_{60}}$ = 44.2

MEASURES OF VARIATION/DISPERSION

Example

Consider the following two sets of scores:Set 1: 40, 50, 60, 60, 40, 50Set 2: 0, 100, 25, 75, 80, 20

- Both these sets have the same mean (50),
- but the second set is a lot more widely dispersed ("scattered") than the first....



MEASURES OF VARIATION/DISPERSION

- The scatter or spread of items of a distribution is known as dispersion or variation.
- In other words the degree to which numerical data tend to spread about an average value is called dispersion or variation of the data.

- Measures of dispersion are statistical measures which provide ways of measuring the extent in which data are dispersed or spread out.

Objective of measuring variation

- To determine the reliability of an average by pointing out as how far an average is representative of the entire data.
- To determine the nature and cause of variation in order to control the variation itself.
- Enable comparison of two or more distribution with regard to their variability.
- Measuring variability is of great importance to other statistical analysis. E.g., it is the basis of statistical quality control

A GOOD MEASURE OF VARIATION

- It should be easy to compute and understand.
- It should be based on all observations.
- It should be Uniquely defined
- It should be capable of further algebraic treatment.
- It should be as little as affected by extreme values

TYPES OF MEASURE OF VARIATION

Absolute measure:

- Range
- Quartile deviation
- Mean deviation
- Variance
- Standard deviation

Relative measures:

- Relative range
- Coefficient of quartile deviation
- Coefficient of mean deviation
- Coefficient of variation
- Standard scores

RANGE

- The difference between the largest (maximum) and smallest (minimum) values.

$$Range = Maximum - Minimum \quad (1)$$

For frequency distributed data, the range is:

- The difference between the upper class boundary of the last class and the lower class boundary of the first class.

Measure of variation-dispersionn

- Find the Range of 54.5, 55.0, 55.7, 51.8, 54.2, 52.4 Solution:
- range(R) = 55.7- 51.8 = 3.9cm

Given the following frequency distribution. Find the range

Class	frequency
52.5-63.5	6
63.5-74.5	12
74.5-85.5	25
85.5-96.5	18
96.5-107.5	14
107.5-118.5	5

Solution: Range $= UCB_l - LCB_f = 118.5 - 52.5 = 66$

RANGE

Definition: Difference between the value of the smallest item and the value of the largest item in the distribution.

$$\textbf{Range} = L - S$$

L – Largest Value, S- Smallest Value

The *relative measure* corresponding to range is called the ***coefficient of range***,

$$\textbf{Coefficient of Range} = \frac{L-S}{L+S}$$

Example 1: The following are the prices of shares of a company from Monday to Saturday:

Days	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
Price(Rs.)	200	210	208	160	220	250

Calculate the range and its coefficient.

Solution: Range = $L - S$ = 250 – 160 = 90

Range = Rs. 90

$$\text{Coefficient of Range} = \frac{L-S}{L+S} = \frac{250-160}{250+160} = \frac{90}{410} = 0.22$$

In a frequency distribution, range is calculated by taking the difference between the lower limit of the lowest class and the upper limit of the highest class.

Example 2:

Marks	10-20	20-30	30-40	40-50	50-60	60-70
No. of Students	12	18	27	20	17	6

$$\text{Range} = L - S = 70 - 10 = 60$$

$$\text{Coefficient of Range} = \frac{L-S}{L+S} = \frac{70-10}{70+10} = \frac{60}{80} = 0.75$$

Quartile Deviation

Definition: Average amount by which the two quartiles differ from the median.

$$\textbf{Quartile Deviation (Q.D.)} = \frac{Q_3 - Q_1}{2}$$

- The Median ± Q.D. covers exactly 50 per cent of the observations.

- When Q.D. is very small, it describes high uniformity or small variation of the central 50% items, and a high Q.D. means that the variation among the central items is large.

$$\textbf{Coefficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

It can be used to compare the degree of variation in different distributions.

Example 3: Calculate the value of Q.D. and its coefficient of Q.D. from the following data.

Roll No.	1	2	3	4	5	6	7
Marks	20	28	40	12	30	15	50

Solution: Marks in ascending order 12 15 20 28 30 40 50

Q_1 = Size of $\frac{N+1}{4}th$ item = Size of $\frac{7+1}{4}$ = 2^{nd} item.

Size of 2^{nd} item is 15. Hence $\boldsymbol{Q_1 = 15}$

Q_3 = Size of $3\left(\frac{N+1}{4}\right)th$ item = Size of $3\left(\frac{7+1}{4}\right)$ = 6^{th} item.

Size of 6^{th} item is 40. Hence $\boldsymbol{Q_3 = 40.}$

$$\therefore \quad Q.D. = \frac{Q_3 - Q_1}{2} = \frac{40-15}{2} = 12.5$$

$$\textit{Coefficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{40-15}{40+15} = 0.455$$

Example 4: Compute the value of Q.D. and its coefficient from the following data.

Marks	10	20	30	40	50	60
No. of Students	4	7	15	8	7	2

Solution:

Marks	*Frequency*	*cumulative frequency*
10	4	4
20	7	11
30	15	26
40	8	34
50	7	41
60	2	43

Q_1 = Size of $\frac{N+1}{4}$ *th* item = Size of $\frac{43+1}{4}$ = 11^{th} item.

Size of 11^{th} item is 20. Hence $\boldsymbol{Q_1 = 20}$

Q_3 = Size of $3\left(\frac{N+1}{4}\right)$ *th* item = Size of $3\left(\frac{43+1}{4}\right) = 33^{rd}$ item.

Size of 33^{rd} item is 40. Hence $\boldsymbol{Q_3 = 40.}$

$$Q.D. = \frac{Q_3 - Q_1}{2} = \frac{40-20}{2} = 10$$

$$\textit{Coefficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{40-20}{40+20} = 0.333$$

Example 4: Compute the value of Q.D. and coefficient of Q.D. from the following data

C.I.	10-20	20-30	30-40	40-50	50-60	60-70	70-80
f	12	19	5	10	9	6	6

Solution:

Marks	*Frequency*	*Cumulative Frequency*
10-20	12	12
20-30	19	31
30-40	5	36
40-50	10	46
50-60	9	55
60-70	6	61
70-80	6	67
	$N = 67$	

Q_1 = Size of $\frac{N}{4}$ *th* item = Size of $\frac{67}{4}$ = 16.75th item.

Q_1 lies in the interval **20-30**

$$Q_1 = l + \frac{\frac{N}{4} - c.f.}{f} \times i \qquad l = 20,\ N/4 = 16.75,\ c.f. = 12\ \ f = 19,\ \ i = 10$$

$$Q_1 = 20 + \frac{\frac{67}{4} - 12}{19} \times 10 = 20 + 2.5 = 22.5$$

Hence $\mathbf{Q_1 = 22.5}$

Q_3 = Size of $\frac{3N}{4}$ *th* item = Size of $\frac{3 \times 67}{4}$ = 50.25th item.

Q_3 lies in the class **50-60**.

$$Q_3 = l + \frac{\frac{3N}{4} - c.f.}{f} \times i \qquad l = 50,\ 3N/4 = 50.25,\ c.f. = 46\ \ f = 9,\ \ i = 10$$

$$Q_3 = 50 + \frac{50.25 - 46}{9} \times 10 = 50 + 4.72 = 54.72$$

Hence $\mathbf{Q_3 = 54.72}$

$$\boldsymbol{Q.D.} = \frac{Q_3 - Q_1}{2} = \frac{54.72 - 22.5}{2} = \boldsymbol{16.11}$$

$$\textbf{\textit{Coefficient of Q.D.}} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{54.72 - 22.5}{54.72 + 22.5} = \boldsymbol{0.4172}$$

MEAN DEVIATION

- **Mean Deviation(M.D):**
  - The average deviation measures the scatter of the individual observations around a central value usually the mean or the median of a distribution.
  - The mean deviation is defined as the arithmetic mean of positive deviations of each observation from either the mean or the median of a distribution.
  - If the deviations are taken from the mean then it is called mean deviation about the mean. On the other hand, if the deviations are taken from the median we call it mean deviation about the median.

MEAN DEVIATION ABOUT THE MEAN

- The mean Deviation (M.D) is the arithmetic mean of the absolute deviations of the values from the mean.
- It is the "average absolute deviation of the values from the mean".

$$Mean\ deviation = \frac{\sum_{i=1}^{n} | X_i - \bar{x} |}{n} \qquad (4)$$

- **Note that:** while dealing with population values, it is adjusted accordingly
- **Mean Deviations for Grouped data (discrete or continuous)**

$$Mean\ deviation = \frac{\sum_{i=1}^{n} f_i | X_i - \bar{x} |}{n} \qquad (5)$$

- Where m = number of classes and x_i = *class mark of the* i^{th} *class*, n = *number of observation*

MEAN DEVIATION ABOUT THE MEDIAN

ungrouped data:

$$MD(\hat{X}) = \frac{\sum_{i=1}^{n} | X_i - \hat{x} |}{n}$$

grouped Frequency Distribution:

$$MD(\hat{X}) = \frac{\sum_{i=1}^{k} f_i | X_i - \hat{x} |}{n}$$

EXAMPLE

- The weights of a sample of six students from a class (in kilograms) is measured as: 53, 56, 57, 59, 63 and 66. Find the mean deviation about the mean and the mean deviation from the median.
- **solution:**First find the mean and the median. The mean is 59 kg and the median is 58 kg. Then take the deviations of each observation from these averages as shown below

weight X_i	Ad from mean$\| x_i - \bar{x} \|$	AD from median$\| x_i - \hat{x} \|$
53	6	5
56	3	2
57	2	1
59	0	1
63	4	5
66	2	8
Total	22	22

mean deviation a bout the mean:

$$MD(\bar{X}) = \frac{\sum_{i=1}^{n} f_i \mid X_i - \bar{x} \mid}{n}$$

$$= MD(\bar{x}) = \frac{22}{6}$$

$$= 3.67$$

mean deviation about median:

$$MD(\hat{X}) = \frac{\sum_{i=1}^{k} f_i \mid X_i - \hat{x} \mid}{n}$$

$$MD(\hat{x}) = \frac{22}{6} = 3.67$$

Example 4.4: Calculate the mean deviation from the mean and median for the following data.

Class interval	1-5	6-10	11-15	16-20
Frequency	4	1	2	3

Coefficients of Mean Deviation(C.M.D)

- C.M.D = M.D/Average about which deviations are taken
- Coefficient of mean deviation about the

$$CMD(\bar{X}) = \frac{MD\bar{X}}{\bar{X}}$$

- Coefficient of mean deviation about the median=

$$CMD(\hat{X}) = \frac{MD\hat{X}}{\hat{X}}$$

Variance and Standard Deviation

- The variance and standard deviation are the most superior and widely used measures of dispersion
- Both measures the average dispersion of the observations around the mean.
- The variance is defined as the average of the squared deviation from the mean.

VARIANCE AND STANDARD DEVIATION

For the frequency distribution $x_i \mid f_i$; i =1, 2, …, n ,

$$Variance = \sigma^2 = \frac{\sum(x_i - \bar{x})^2}{N} = \frac{1}{N}\sum x_i^2 - \left(\frac{1}{N}\sum x_i\right)^2 = \frac{1}{N}\sum x_i^2 - \bar{x}^2$$

(ii) Discrete or Continuous frequency distribution

$$\sigma^2 = \frac{1}{N}\sum_i f_i (x_i - \bar{x})^2$$

$$= \frac{1}{N}\sum f_i x_i^2 - \left(\frac{1}{N}\sum f_i x_i\right)^2 \textbf{ or } \frac{1}{N}\sum f_i d_i^2 - \left(\frac{1}{N}\sum f_i d_i\right)^2$$

$$Standard\ deviation = \sqrt{variance} \quad \sigma = \sqrt{\frac{1}{N}\sum (x_i - \bar{x})^2} \quad \textbf{or} \quad \sqrt{\frac{1}{N}\sum f_i (x_i - \bar{x})^2}$$

$\bar{x}$ – Arithmetic mean of the distribution

Coefficient of Variation

- In situations where either two series have different units of measurements, or their means differ sufficiently in size, the CV should be used as a measure of dispersion.

$$coefficient\ of\ variation(CV) = \frac{standard\ deviation}{mean} * 100\%$$

$$CV = \frac{S}{\bar{X}} * 100\% for\ sample\ and$$

- In spite of the fact that the C.V. is broadly applied, its disadvantage is that it's not useful when the mean is negative or zero or very close to zero.

- **Interpretation of the coefficient of variation:** the distribution having less CV is said to be less variable or more consistent

The score of two players A and B in ten innings during a certain season are:

A	32	28	47	63	71	39	10	60	96	14
B	19	31	48	53	67	90	10	62	40	80

Find which of the two players A, B is more consistent in scoring.

Solution: Calculation of Coefficient of Variation

X	$(X-\bar{X})$	$(X-\bar{X})^2$
32	-14	196
28	-18	324
47	+1	1
63	+17	289
71	+25	625
39	-7	49
10	-36	1296
60	+14	196
96	+50	2500
14	-32	1024
$\sum X = 460$	0	6500

Y	$(Y-\bar{Y})$	$(Y-\bar{Y})^2$
19	-31	961
31	-19	361
48	-2	4
53	+3	9
67	+17	289
90	+40	1600
10	-40	1600
62	+12	144
40	-10	100
80	+30	900
$\sum Y = 500$	0	5968

$$\bar{X} = \frac{460}{10} = 46 \qquad \bar{Y} = \frac{500}{10} = 50$$

$$\sigma_A^2 = \frac{\sum(x_i - \bar{x})^2}{N} = \frac{6500}{10} = 650 \qquad \sigma_B^2 = \frac{\sum(y_i - \bar{y})^2}{N} = \frac{5968}{10} = 596.8$$

$$\sigma_A = \sqrt{\frac{\sum(x_i - \bar{x})^2}{N}} = 25.5 \qquad \sigma_B = \sqrt{\frac{\sum(y_i - \bar{y})^2}{N}} = 24.43$$

$$\text{C.V.}_{(A)} = \frac{\sigma_A}{\bar{x}} \text{ x } 100 = 55.43 \qquad \text{C.V.}_{(B)} = \frac{\sigma_B}{\bar{y}} \text{ x } 100 = 48.86$$

$$\sum X = 460 \; ; \; \sum (X_i - \bar{x}) = 0 \; ; \sum (X_i - \bar{x})^2 = 6500$$

$$\sum Y = 500 \; ; \; \sum (Y_i - \bar{y}) = 0 \; ; \sum \left(Y_j - \bar{y}\right)^2 = 5968$$

$\sigma_A = 25.5$ $\sigma_B = 24.43$

$C.V._{(A)} = 55.43$ $C.V._{(B)} = 48.86$

Suppose that samples of polythene bags from two manufacturers, A and B, are tested by a prospective buyer for bursting pressure, with the following results:

Bursting Pressure (lb.)	***Number of Bags***	
	A	***B***
5.0 – 9.9	2	9
10.0 – 14.9	9	11
15.0 – 19.9	29	18
20.0 – 24.9	54	32
25.0 – 29.9	11	27
30.0 – 34.9	5	13

Which set of bags has the highest average burning pressure?
Which has more uniform pressure? If prices are the same,
which manufacturer's bags would be preferred by the buyer? **Why**?

For Manufacturer A

Bursting Pressure (lb.)	m	f	$\left(\frac{m-17.45}{5}\right)$ d	fd	fd^2
4.95-9.95	7.45	2	-2	-4	4
9.95-14.95	12.45	9	-1	-9	9
14.95-19.95	17.45	29	0	0	0
19.95-24.95	22.45	54	1	54	54
24.95-29.95	27.45	11	2	22	44
29.95-34.95	32.45	5	3	15	45
		N = 110		$\sum fd$=78	$\sum fd^2$=160

$$\bar{X}_A = A + \frac{\sum_{i=1}^{n} f_i d_i}{N} \times i$$

Here, $A = 17.45, \sum fd=78, N = 110, i = 5$

$$\bar{X}_A = 17.45 + \frac{78}{110} \times 5 = 21$$

$$\sigma_A = \sqrt{\frac{\sum f_i d_i^2}{N} - (\frac{\sum f_i d_i}{N})^2} \times i$$

$$= \sqrt{1.455 - 0.503} \times 5 = 4.88$$

$$\text{C.V.} = \frac{\sigma_A}{\bar{x}} \text{ x } 100 = 23.24\%$$

For Manufacturer B

Bursting Pressure (lb.)	m	f	$\left(\frac{m-17.45}{5}\right)$ d	fd	fd^2
4.95-9.95	7.45	9	-2	-18	36
9.95-14.95	12.45	11	-1	-11	11
14.95-19.95	17.45	18	0	0	0
19.95-24.95	22.45	32	+1	+32	32
24.95-29.95	27.45	27	+2	+54	108
29.95-34.95	32.45	13	+3	+39	117
		N = 110		$\sum fd = 96$	$\sum fd^2 = 304$

$$\bar{X}_B = A + \frac{\sum_{i=1}^{n} f_i d_i}{N} \times i$$

Here, $A = 17.45, \sum fd = 96, N = 110, i = 5$

$$\bar{X}_B = 17.45 + \frac{96}{110} \times 5 = 21.81$$

$$\sigma_B = \sqrt{\frac{\sum f_i d_i^2}{N} - \left(\frac{\sum f_i d_i}{N}\right)^2} \times i$$

$$= \sqrt{2.764 - 0.762} \times 5 = 7.075$$

$$\text{C.V.} = \frac{\sigma_B}{\bar{x}} \text{ x } 100 = 32.44\%$$

$\bar{X}_A = 21$ $\bar{X}_B = 21.81$

$\sigma_A = 4.88$ $\sigma_B = 7.07$

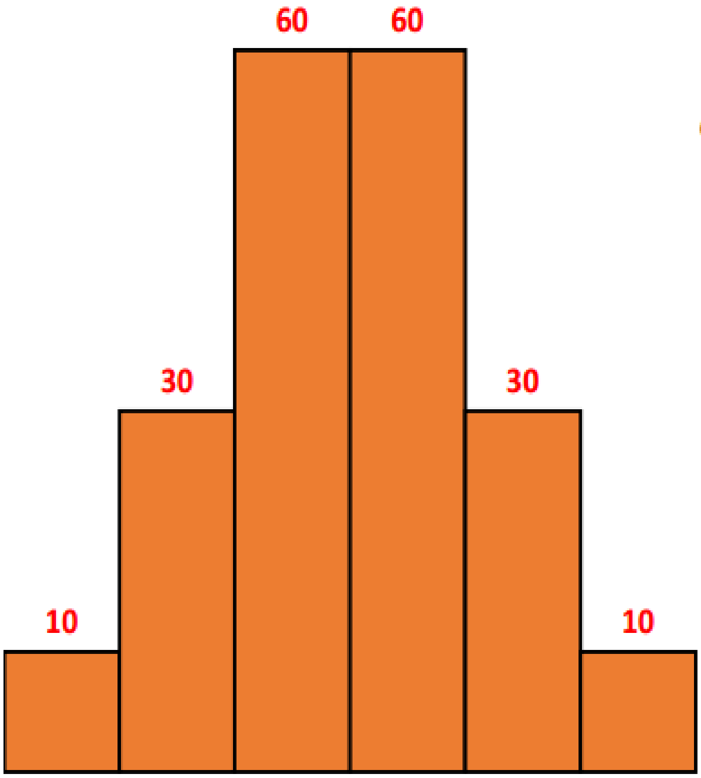
C.V.$_{(A)}$ = 23.24% C.V.$_{(B)}$ = 32.44%

Since the average bursting pressure is higher for manufacturer B, the bags of manufacturer B have higher bursting pressure.

C.I.	*f*
0-5	10
5-10	30
10-15	60
15-20	60
20-25	30
25-30	10

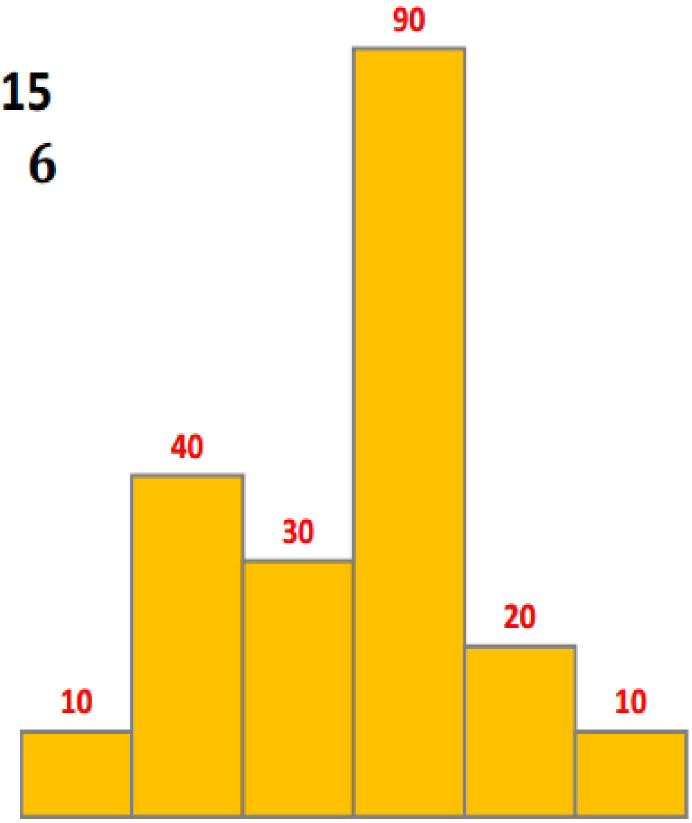
C.I.	*f*
0-5	10
5-10	40
10-15	30
15-20	90
20-25	20
25-30	10

$$\bar{X} = 15$$
$$\sigma = 6$$



$\bar{x} = 15$

$\sigma = 6$



Skewness

When a series is not symmetrical it is said to be asymmetrical or skewed.

- Skewness is the degree of asymmetry or departure from symmetry of a distribution.
- A skewed frequency distribution is one that is not symmetrical.
- Skewness is concerned with the shape of the curve not size
- If the frequency curve (smoothed frequency polygon) of a distribution has a longer tail to the right of the central maximum than to the left, the distribution is said to be skewed to the right or said to have positive skewness. If it has a longer tail to the left of the central maximum than to the right, it is said to be skewed to the left or said to have negative skewness.
- For moderately skewed distribution, the following relation holds among the three commonly used measures of central tendency.

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

SKEWNESS

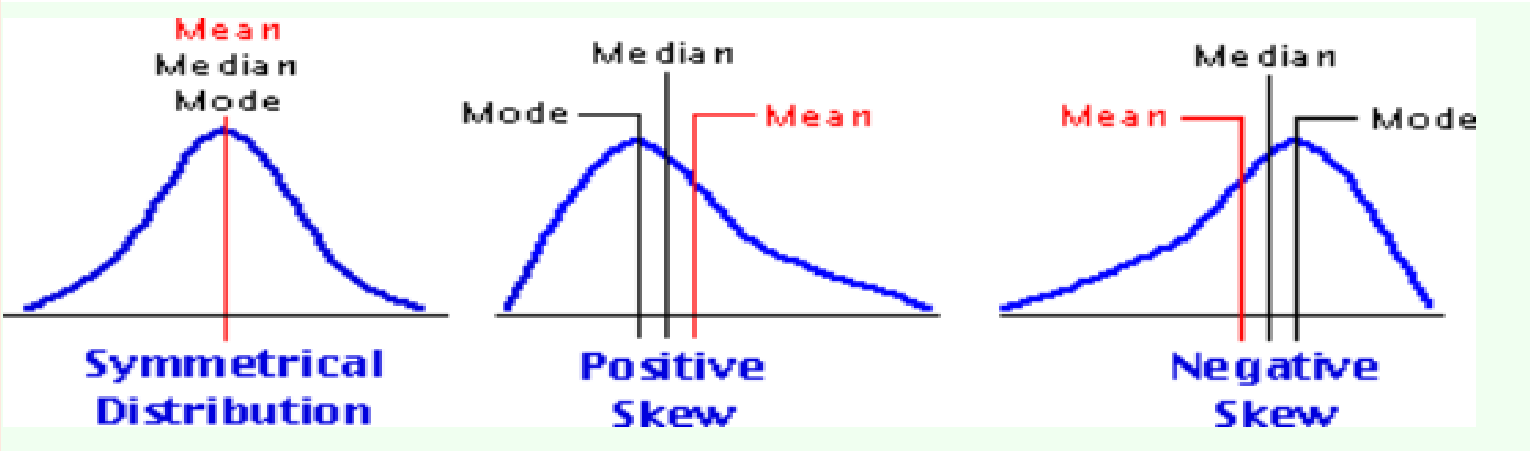
Dispersion is concerned with the amount of variation rather than with its direction.

Skewness tell us about the ***direction of the variation*** or the departure from symmetry.

Types of Skewness: (i) Symmetrical Distribution

(ii) Positively Skewed Distribution (**Mean > Mode**)

(iii) Negatively Skewed Distribution (**Mean < Mode**)



Measure of Skewness

Absolute measures of Skewness(Sk)

$$Sk = \bar{X} - \text{Mode}$$

Relative measures of Skewness:

(i) Karl pearson's coefficient of skewness.

(ii) Bowley's coefficient of skewness.

(iii) Measure of skewness based on moments.

1. Karl Pearson's coefficient of skewness

Coefficient of Skewness : $Sk = \frac{Mean - Mode}{Standard\ deviation} = \frac{\bar{X} - Mo}{\sigma}$

however, in practice, it is rare that the value of ***Sk*** exceed the limits of $\pm$**1.**

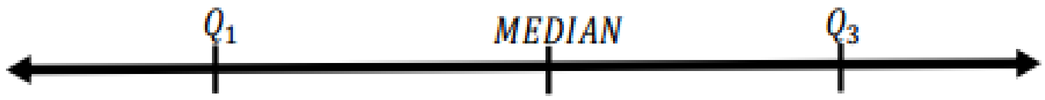
But using , $Mode = 3\ Median - 2\ Mean$

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

this measure can vary between $\pm$**3** ;

2. BOWLEY'S COEFFICIENT OF SKEWNESS

It is based on Quartiles. In a symmetrical distribution first and third quartiles are equidistant from the median :



In a symmetrical distribution the third quartile is the same distance above the median as the first quartile is below it, i.e.,

$$Q_3 - Median = Median - Q_1$$

$$or \qquad Q_3 + Q_1 - 2\,Median = 0$$

$$\therefore \qquad Sk = \frac{Q_3 + Q_1 - 2\,Median}{Q_3 - Q_1}$$

This measure is called the ***quartile measure of skewness*** and varies between ± 1.

MOMENTS

The r*th* *moment about the mean* (*the* r*th* *central moment*) *defined as* :

$$M_r = \frac{\sum(X_i - \bar{X})^r}{n}, r = 0, 1, 2..$$

for continuous grouped data it is given by:

$$M_r = \frac{\sum f_i(X_i - \bar{X})^r}{n}, \text{ where } X_i \text{ are class marks}$$

Central Moments (Moments about the Arithmetic Mean):

First Moment $\mu_1 = \frac{\sum(X_i - \bar{X})}{N}$: (sum of the deviations from A.M. is always zero. $\boldsymbol{\mu_1 = 0}$)

Second Moment $\mu_2 = \frac{\sum(X_i - \bar{X})^2}{N} = \sigma^2 =$ Variance

Third Moment $\mu_3 = \frac{\sum(X_i - \bar{X})^3}{N}$

Fourth Moment $\mu_4 = \frac{\sum(X_i - \bar{X})^4}{N}$

For a frequency distribution:

First Moment $\mu_1 = \frac{\sum f_i(X_i - \bar{X})}{N}$

Second Moment $\mu_2 = \frac{\sum f_i(X_i - \bar{X})^2}{N} = \sigma^2 =$ Variance

Third Moment $\mu_3 = \frac{\sum f_i(X_i - \bar{X})^3}{N}$ or $\frac{\sum f_i\, x_i^3}{N}$

Fourth Moment $\mu_4 = \frac{\sum f_i(X_i - \bar{X})^4}{N}$ or $\frac{\sum f_i\, x_i^4}{N}$

KURTOSIS

Kurtosis enables us to have an idea about the '*flatness*' or '*peakedness*' in the region about the mode of a frequency curve.

"*Convexity of the frequency curve*"

It is measured by the Coefficient of $\beta_2 = \frac{\mu_4}{\mu_2^2}$ or $\gamma_2 = \beta_2 - 3$

(i)	$\beta_2 = 3$, *i.e.*, $\gamma_2 = 0$	:	mesokurtic curve
(ii)	$\beta_2 < 3$, *i.e.*, $\gamma_2 < 0$	:	platykurtic curve
(iii)	$\beta_2 > 3$, *i.e.*, $\gamma_2 > 0$	:	leptokurtic curve

- Kurtosis is the degree of peakedness of a distribution, usually taken relative to a normal distribution.

- When the curve of a distribution is relatively:

  - flatter than normal it is known as **platykurtic** and
  - the distribution is more peaked than normal, it is called **leptokuric.**
  - The normal distribution which is not very high peaked or flat topped is called **mesokurtic.**