UNIT - I

Descriptive statistics and methods for data Science.

* Measures of central tendency :-

→ measures of central tendency are sometimes needed to make meaningful interpretation of the data.

→ Generally it is found that in any distribution values of the variables tends to congregate around the central value of the distribution.

→ This tendency is known as measures of central tendency.

* The following are the five measures of central tendency.

1. Arithmetic mean {mean}
2. median
3. mode
4. Geometric mean
5. Harmonic mean

* Arithmetic mean [Direct method]:-

→ If $x_1, x_2, x_3 \ldots x_n$ are a set of n variables then arithmetic mean is given by

i.e. $$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}$$

$$= \frac{\Sigma x_i}{n}$$

→ In a frequency distribution if $x_1, x_2, x_3 \ldots x_n$ be the mid values of the class intervals having frequencies $f_1, f_2, f_3 \ldots f_n$ respectively then

i.e. $$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + x_3 f_3 + \ldots + x_n f_n}{f_1 + f_2 + f_3 + \ldots + f_n}$$

$$= \frac{\Sigma x_i f_i}{\Sigma f_i}$$

Problems:-

1. Find the arithmetic mean of the following
7, 6, 8, 10, 13, 14.

Sol:- Give data

7, 6, 8, 10, 13, 14 , $n = 6$

$$\text{mean} = \bar{x} = \frac{7+6+8+10+13+14}{6}$$

$$= \frac{58}{6}$$

$$= 9.66$$

2. Find the arithmetic mean of the following distribution.

x	1	2	3	4	5	6	7
f	5	9	12	17	14	10	6

Sol: Now

x_i	f_i	$x_i f_i$
1	5	5
2	9	18
3	12	36
4	17	68
5	14	70
6	10	60
7	6	42
	$\Sigma f_i = 73$	$\Sigma x_i f_i = 299$

Arithmetic mean

$$\bar{x} = \frac{\Sigma x_i f_i}{\Sigma f_i}$$

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

$$= 4.095$$

3. Calculate the arithmetic mean of the marks of the following.

marks	0-10	10-20	20-30	30-40	40-50	50-60
no. of students	12	18	27	20	17	6

Sol:- now

marks	no. of students f_i	mid value x_i	$x_i f_i$
0-10	12	$\frac{0+10}{2} = 5$	60
10-20	18	$\frac{10+20}{2} = 15$	270
20-30	27	$\frac{20+30}{2} = 25$	675
30-40	20	$\frac{30+40}{2} = 35$	700
40-50	17	$\frac{40+50}{2} = 45$	765
50-60	6	$\frac{50+60}{2} = 55$	330
	$\Sigma f_i = 100$		$\Sigma f_i x_i = 2800$

we know that

$$\text{Arithmetic mean} = \bar{x} = \frac{\Sigma f_i x_i}{\Sigma f_i}$$

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

$$= 28$$

Scanned with CamScanner

Note:-
→ Direct method of computing especially when applied to a grouped data involves heavy calculations.
→ Inorder to avoid this, the following formulae are generally used.

- Shortcut method:-

$$\bar{x} = A + \frac{\Sigma f_i d_i}{\Sigma f_i}$$

where
A → Assumed mean
$d = x_i - A$

- Stepdeviation method:-

$$\bar{x} = A + \frac{h \Sigma f_i d_i}{\Sigma f_i}$$

where
A → Assumed mean
h → length of class interval

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

Problems:-

1. Calculate the mean of the following.

class interval	0 – 8	8 – 16	16 – 24	24 – 32	32 - 40	40 – 48
frequency	8	7	16	24	15	7

Sol:- Now

class interval	frequency f_i	mid values x_i	$d_i = x_i - A$	$f_i d_i$
0 – 8	8	4	–24	–192
8 – 16	7	12	–16	–112
16 – 24	16	20	–8	–128
24 – 32	24	28 → A	0	0
32 – 40	15	36	8	120
40 – 48	7	44	16	112
	$\Sigma f_i = 77$			$\Sigma f_i d_i = -200$

Take A = 28.

$$\bar{x} = A + \frac{\Sigma f_i d_i}{\Sigma f_i}$$

$$= 28 + \frac{-200}{77}$$

$$= 25.402.$$

2. The following is the age distribution of 1000 persons working in a large industrial house

Age group	20-25	25-30	30-35	35-40	40-45	45-50	50-55	55-60	60-6[illegible]
No. of persons	30	160	210	180	145	105	70	60	40

Due to continuous heavy loses the management decides to bring down the strength to 30% of the present number according to the following scheme.

1. To reach the first 15% from lower age group
2. To absorb the next 45% in other branches
3. To make 10% from the highest age group retire permanently if necessary

calculate the age limits of the person retained and those to be transformed to other departments also find the average age of those retained.

Sol:- Total no. of persons in the industrial house = 1000

According to conditions of the problem

1. The no. of persons to be retrenched from the lower age group = 15% of 1000

$$= \frac{15}{100} \times 1000$$

$$= 150$$

→ Now 30 of these will be from 1st age group 20-25 and remaining 120 from the next age group 25-30.

→ Now 160 members will be from group 25-30.

i.e. 160 − 120 = 40

→ In the second group 25-30 we have 40 members.

2. Now the no. of persons absorbed 45% of 1000

$$= \frac{45}{100} \times 1000$$

$$= 450$$

→ From 25-30 we have to eliminate 40 members.

→ we have 410 members.

→ Now from 30-35 & 35-40 we have to eliminate 390 members out of 410 & remaining we have 20 members.

→ From the class 40-45 we have 145 members and we have to eliminate 20 members & remaining we have 125 members in the class 40-45.

3. Now The no. of persons retained from the highest age group is 10% of 1000 $= \frac{10}{100} \times 1000 = 100$

→ we have to eliminate 40 from 60-65 group & 60 from 55-60 group.

- From 1,2,3, the frequency distribution of the no. of persons retained in the industrial house as shown below

Agegroup	no. of persons
40-45	125
45-50	105
50-55	70

Now we have to find the mean of the distribution.

Age group	Frequency f_i	mid value x_i	$x_i - A$	$d_i = \frac{x_i - A}{h}$	$f_i d_i$
40-45	125	42.5	-5	-1	-125
45-50	105	47.5 → A	0	0	0
50-55	70	52.5	5	1	70
	$\Sigma f_i = 300$				$\Sigma f_i d_i = -55$

Take $A = 47.5$; $h = 5$

$$\bar{x} = A + \frac{h \Sigma f_i d_i}{\Sigma f_i}$$

$$= 47.5 + \frac{5 \times (-55)}{300}$$

$$= 46.584$$

* Merits and demerits of arithmetic mean :-

→ merits :-

→ It is rigidly defined

→ It is easy to understand and easy to calculate

→ It is based on all observations

→ of all the averages, arithmetic mean is effected least by fluctuation of sampling.

→ Demerits :-

→ It cannot be determined by inspection nor it can be located ~~graphically~~ graphically.

→ Arithmetic mean cannot be used if we are dealing with qualitative characteristics which cannot be measured quantitatively such as intelligence, honestly, beauty etc..

→ Arithmetic mean cannot be obtained if a single observation is missing.

→ Arithmetic mean is affected very much by extreme values

→ In extremely assymetrical distribution arithmetic mean is not suitable measure of distribution.

Scanned with CamScanner

* Median :-

→ In case of ungrouped data if no. of observations is odd then median is the middle value after the values have been arranged in descending (or) ascending order of magnitude

→ In case of even no. of observations there are two middle no's and the median is obtained by taking the arthimetic mean of the middle terms.

Example:-

i) The median of the values

25, 20, 15, 35, 18,

Ascending order: 15, 18, 20, 25, 35

∴ 20 is median

ii) The median of the values 5, 4, 3, 2, 6, 1

Ascending order: 1, 2, 3, 4, 5, 6

∴ 3.5 is median

Note:-

→ In case of discrete frequency distribution median is obtained by considering the cumulative frequency (c.f)

→ The steps for calculating median is given below

→ Find the value of $N/2$ where $N = \Sigma f_i$

→ See the cumulative frequency just greater than $N/2$

→ The corresponding value of 'x' is median.

Problems :-

1. Obtain the median of the following distribution

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

Sol:- Now

x	f	c.f
1	8	8
2	10	18
3	11	29
4	16	45
$\boxed{5}$	$\boxed{20}$	$\boxed{65}$
6	25	90
7	15	105
8	9	114
9	6	120
	$N = 120$	

Here $\frac{N}{2} = \frac{120}{2} = 60$

→ The cumulative frequency just greater than 60 is 65

→ The corresponding x value of 65 is 5.

→ median of given distribution is 5

median for continuous frequency distribution:

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

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

where

l = lower limit of median class

f = frequency of median class

h = length of median class

c = c.f of class preceding the median class

$N = \Sigma f_i$

*Problems:

1. Find the median of the following distribution

wages (Rs)	2000-3000	3000-4000	4000-5000	5000-6000	6000-7000
no. of workers	3	5	20	10	5

Sol: Now

wages (Rs) x_i	no. of workers f_i	c.f
2000-3000	3	3
3000-4000	5	8 → c
l $\boxed{4000}$ - 5000	20 → f	28
5000-6000	10	38
6000-7000	5	43

Now $N = 43$, $\frac{N}{2} = 21.5$

→ The c.f just greater than 21.5 is 28

$l = 4000$, $f = 20$, $c = 8$, $h = 1000$

$$\text{median} = l + \frac{h}{f}\left[\frac{N}{2} - c\right]$$

$$= 4000 + \frac{100\cancel{0}}{2\cancel{0}}\left[21.5 - 8\right]$$

$$= 4675$$

∴ median of wages = 4675

2. ~~The~~ In a factory employing 3000 persons in a day 5% work less than 3hrs, 580 work from 3.01 – 4.50, 30% work from 4.51 to 6.00, 500 work from 6.01 – 7.50, 20% work from 7.51 to 9.00 & the rest work 9.01 or more hours. What is the median hours of work.

Sol: The given information can be expressed in tabular form as follows

work hours	No. of workers	c.f	equal boundaries
lessthan 3	5% of 3000 = 150	150	lessthan 3.005
3.01-4.50	580	730 → c	3.005 - 4.505
4.51-6.00	30% of 3000 = 900 → f	1630	l $\boxed{4.505}$ - 6.005
6.00-7.50	500	2130	6.005 - 7.505
7.51-9.00	20% of 3000 = 600	2730	7.505 - 9.005
9.01 - above	270	3000	9.005 - above.
	N = 3000.		

Now $\frac{N}{2} = \frac{3000}{2} = 1500$

The median class is $4.505 - 6.005$

$l = 4.505$, $N = 3000$, $c = 730$, $f = 900$, $h = 1.5$

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

$$= 4.505 + \frac{1.5}{900}[1500 - 730]$$

$$= 5.788$$

$\therefore$ median hours of work $= 5.788$

3. An incomple frequency distribution given as follows

variables	10-20	20-30	30-40	40-50	50-60	60-70	70-80
frequency	12	30		65		25	18

Given that median is 46. Determine the mixing frequency using median formula.

Sol: Let the missing frequencies are f_1 corresponding class 30-40, f_2 corresponding class 50-60.

variables	frequency	cf
10-20	12	12
20-30	30	42
30-40	f_1	$42+f_1$
40-50	65	$107+f_1$
50-60	f_2	$107+f_1+f_2$
60-70	25	$132+f_1+f_2$
70-80	18	$150+f_1+f_2$
	N = 229.	

Given that sum of frequency $= 150 + f_1 + f_2 = 229$

$f_1 + f_2 = 79$

Scanned with CamScanner

Given that median is 46 & corresponding median class is 40-50

$\therefore l = 40, \; c = 42 + f_1, \; f = 65, \; \frac{N}{2} = \frac{229}{2} = 114.5, \; h = 10.$

$$\text{median} = l + \frac{h}{f}\left[\frac{N}{2} - c\right]$$

$$46 = 40 + \frac{10}{65}\left[114.5 - (42 + f_1)\right]$$

$$6 = \frac{10}{65}\left[72.5 - f_1\right]$$

$$39 = 72.5 - f_1$$

$$f_1 = 72.5 - 39$$

$$f_1 = 33.5$$

$$\boxed{f_1 = 34}$$

from (1) $\Rightarrow f_1 + f_2 = 79$

$$\boxed{f_2 = 45}$$

* merits & demerits of median:-

→ merits:-

→ It is rigidly define

→ It is easy to understand & easy to calculate

→ In some cases it can be located nearly by inspection

→ It is not at all affected by extreme values

→ It can be calculated ~~by~~ for distribution with opened distribution.

→ Demerits:-

→ In case of even no. of observations median cannot be determined exactly.

→ we necoily estimate it by taking the mean of two middle items.

→ It is not based on all observations.

→ It is not a meanable to algebraic treatment

→ As compared to mean it is affected much by fluctuations by sampling.

* mode:-

→ It is the value which occurs most frequently in set of observations.

Example:- 2, 4, 5, 2, 3, 2, 3, 4, 2, 1.

mode = 2

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

Example:-

x	1	2	3	4	5
f	4	7	24	38	10

$\therefore$ Max frequency = 38, mode = 4

Scanned with CamScanner

· Note:-

In any one of the following cases

i, If the maximum frequency is repeated

ii, If the maximum frequency occurs in very beginning or end of distribution.

iii, If irregularities in the distribution, then the value of mode is defined by using method of grouping.

1. Find the mode of the following 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

Sol:- From the given data we observed that distribution is not regular because the frequencies are increasing steadly upto 40 and then decreasing but the frequency 45 after 20 does not seem to be consistent. So in this case we have to find mode. by using method of grouping we form 6 columns.

i, The frequencies in column I are original frequencies.

ii, column II is obtained by combining the frequency by two by two

iii, column III is obtained by leaving the first frequency and combaining the remaining frequency by two by two.

iv, column IV is obtained by combining the frequencies three by three

v, column V is obtained by leaving the first frequency and combining the remaining frequency by three by three

vi, column VI is obtained by leaving the first two frequencies and combine the remaining frequencies three by three.

→ The maximum frequency in each column is marked and prepare the analysis table to find the exact value of mode.

→ Let us prepare the table under the above conditions.

Scanned with CamScanner

Scanned with CamScanner

Size(x)	I	II	III	IV	V	VI
1	3					
2	8	11		26		
3	15	[illegible]	23		46	
4	23	38				
5	35		58	$\boxed{98}$		73
6	40	$\boxed{75}$			$\boxed{107}$	
7	32		$\boxed{72}$			
8	28	60		80		$\boxed{100}$
9	20		48			
10	$\boxed{45}$	65			93	
11	14		59	65		79
12	6	20				

we prepare the analysis table

No. of columns	max	combining the values of x to give max. frequency
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

From the analysis table we find out that the value '6' is repeated maximum number of times, we have mode is '6'.

* mode for continuous frequency distribution:-

→ In case of continuous frequency distribution mode is given by the formula

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

where

$l \rightarrow$ lower limit of modal class

$h \rightarrow$ length (or) magnitude of interval

$f_1 \rightarrow$ frequency of modal class

$f_0 \rightarrow$ frequency of preceeding modal class

$f_2 \rightarrow$ frequency of succeeding modal class.

Problems:

1. Find the mode of the following distribution

cls interval	0-10	10-20	20-30	30-40	40-50	50-60	60-70
frequency	5	8	7	12	28	20	10

Sol: In the given table maximum frequency is 28,

Let it be f_1 and corresponding modal class is 40-50

i.e $f_1 = 28$, $f_0 = 12$, $h = 10$

$l = 40$, $f_2 = 20$

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

$$= 40 + \frac{10(28-12)}{2(28)-12-20}$$

$$= 40 + \frac{160}{24}$$

mode = 46.66

* Merits and Demerits of mode :-

→ Merits :-

→ Mode is easy to calculate

→ mode is not at all affected by extreme values.

→ mode can be conveniently located even if frequency distributio has class intervals of unequal magnitudes provided the modal class and the classes of preceeding & succeeding for the same magnitude.

→ Demerits:-

→ It is not always possible to find a clearly defined mode In some cases we may occur distribution with 2 modes

→ It is not based upon all observations

→ It is not compatible of further mathematical treatment.

→ As compared to mean, mode is affected to a greater extent by fluctuations of sampling.

Note:-

For symmetrical distribution mean, median and mode coincide.

In general

mode = 3(median) - 2(mean)

Geometric mean:

→ Geometric mean is denoted by G which is n^{th} root of product of 'n' observations.

i.e. If $x_1, x_2, x_3 \ldots x_n$ then $G = (x_1 \cdot x_2 \cdot x_3 \cdots \cdot x_n)^{1/n}$

→ In case of frequency distribution x_i/f_i for $i = 1, 2, 3 \ldots n$ then geometric mean $G = (x_1^{f_1} \cdot x_2^{f_2} \cdot x_3^{f_3} \cdots \cdot x_n^{f_n})^{1/N}$

where $\boxed{\Sigma f_i = N}$

Note:

→ For the class intervals if you find the geometric mean we take the observations $x_1, x_2, x_3 \ldots x_n$ as middle values of each class.

Problems:-

1. Find the Geometric mean of 2, 4, 7, 9.

Sol: Given observations are 2, 4, 7, 9

$\therefore n = 4$

$G = (2(4)7(9))^{1/4}$

$= 4.738$

2. Find the geometric mean of the following distribution

X	1	2	3	4	5	6
Y	2	3	2	1	2	2

Sol- we know that geometric mean for frequency distribut is

$G = (x_1^{f_1} \cdot x_2^{f_2} \cdot x_3^{f_3} \cdots x_n^{f_n})^{1/N}$ $\therefore N = \Sigma f$

$= (1^2 (2^3) 3^2 (4^1) 5^2 (6^2))^{1/12}$ $\therefore N = \Sigma Y = 12$

$= 2.825$

3. If a person receives 25% raise after 1 year of service and 15% raise after 2 year of service. Find the average raise per year.

Sol:- At end of 1^{st} year the salary of person is 125%.

At the end of 2^{nd} year the salary of person is 115%.

Now by using Geometric mean the average % of his salary $= (125 \times 115)^{1/2}$

$= 119.895$

$\therefore$ The average raise per year $= 119.895 - 100$

$= 19.895\%$.

4. The geometric mean of 10 observations on a certain variable was calculated as 16.2. It was later discovered that one observation was wrongly recorded as 12.9 (In-fact it was 21.9). Apply appropriate correction and calculate the correct Geometric mean.

Sol:- The geometric mean G of n observations $x_1, x_2 \dots x_n$ is

$$G = (x_1 \cdot x_2 \cdot x_3 \cdots x_n)^{1/n}$$

→ Let x_1 be the observation recorded wrongly instead of correct value. Let it be x_1'

∴ The correct geometric mean

$$G' = (x_1' \cdot x_2 \cdot x_3 \cdots x_n)^{1/n}$$

$$= \left(\frac{x_1'}{x_1} \cdot x_1 \cdot x_2 \cdot x_3 \cdots x_n\right)^{1/n}$$

$$= \left(\frac{x_1'}{x_1}\right)^{1/n} G$$

we have $x_1 = 12.9$, $x_1' = 21.9$ $n = 10$.

$$G' = \left(\frac{21.9}{12.9}\right)^{1/10} \cdot (16.2)$$

$$= 17.08.$$

* merits and Demerits :-

→ Merits :-

→ It is rigidly defined

→ It is based on all values

→ It is very suitable for average ratios, rate & percentage.

→ It is capable of further mathematical treatment

→ Unlike arithmetic mean, it is not effected much by the prescence of extreme values.

→ Demerits :-

→ It cannot be used when the values are negative (or) if any one of the observation is zero.

→ It is difficult to calculate particularly when the values are very large

→ It brings out the property of ratio of changes not absolute difference of the change as the case in arithmetic mean.

→ The geometric mean may not be actual value of the series.

Harmonic mean:-

→ It is the reciprocal of the arithmetic mean (or) reciprocal of the given values. Thus harmonic mean of 'n' observations $x_1, x_2, x_3, \ldots x_n$ where none of which is zero is

$$H.m = \frac{1}{\frac{1}{n}\left(\frac{1}{x_1}+\frac{1}{x_2}+\frac{1}{x_3}+\frac{1}{x_4}\cdots+\frac{1}{x_n}\right)}$$

i.e $$H.m = \frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\frac{1}{x_3}+\cdots\frac{1}{x_n}}$$

$$= \frac{n}{\Sigma \frac{1}{x_i}}$$

→ In the case of frequency distribution x_i/f_i for $i = 1, 2, 3, \ldots n$ then

$$H.m = \frac{1}{\frac{1}{N}\left[\frac{f_1}{x_1}+\frac{f_2}{x_2}\cdots+\frac{f_n}{x_n}\right]}$$

$$H.m. = \frac{N}{\Sigma \frac{f_i}{x_i}}$$ where $N = \Sigma f_i$

Note:-

→ For continuous class intervals we take $x_1, x_2, x_3 \ldots x_n$ are mid values of class intervals.

→ The harmonic mean is often confused with arithmetic mean.

→ The harmonic mean is best suited data that involves rates such as miles per hour (or) miles per litre.

Problems:

1. milk is sold at the rates of 8, 10, 12 and 15 rupees per litre in four different months. Assuming that equal amount are spent on milk by a family in the 4 months. Find the average price rupees for month.

Sol: Since equal amounts of money are spent by the family for each of four months.

Now the average price of the milk per month is given by the harmonic mean of 8, 10, 12 & 15 is

$$H.m. = \frac{4}{\frac{1}{8}+\frac{1}{10}+\frac{1}{12}+1/15}$$

$$= 10.66$$

$$\simeq 11$$

The average price rupees for month is 11/-

2. Three cities A, B, C are equi-distance from each other. A moterist travels from A to B at 3km/hr. From B to C at 40 km/hr from C to A at 50 km/hr determine the average speed.

Sol: Given A to B → 30 km/hr

B to C → 40 km/hr

C to A → 50 km/hr

The average speed can be obtained by H.M

$$H.m = \frac{3}{\frac{1}{30}+\frac{1}{40}+\frac{1}{50}}$$

Avg speed. = 38.29 km/hr.

3. The following is a frequency distibution find H.M.

Ages	15-25	25-35	35-45	45-55	55-65	65-75	75-85
Frequency	8	15	20	25	15	28	18.

Sol: Let us prepare the following table.

Age	Frequency f_i	mid values x_i	f_i/x_i
15-25	8	20	0.4
25-35	15	30	0.5
35-45	20	40	0.5
45-55	25	50	0.5
55-65	15	60	0.25
65-75	28	70	0.4
75-85	18	80	0.225
	$\Sigma f_i = 129$		$\Sigma f_i/x_i = 2.775$

$$H.m = \frac{N}{\Sigma f_i/x_i} = \frac{129}{2.775}$$

$$= 46.48.$$

*merits and Demerits:

merits:

→ It is regidly defined

→ It is defined on all observations

→ It is capable to further algebraic treatment

→ Like geometric mean it is not affected much by fluctuation of sampling.

→ It gives greater importance to small items & it is useful only when small items have to be given greater weightage.

→ Demerits:-
→ It is not easily understood
→ It is difficult to compute.

* Partition values :-
Partition values are the values which divide the series into equal parts
• Quartiles:-
The three points which divide the series into four equal parts they are called quartiles and they are denoted by Q_1, Q_2, Q_3 and which is defined as

$$Q_1 = l + \frac{h}{f}\left(\frac{N}{4} - c\right) \quad \{\text{lower quartile}\}$$

$$Q_2 = l + \frac{h}{f}\left(\frac{N}{8} - c\right) \quad \{\text{medium quartile}\}$$

$$Q_3 = l + \frac{h}{f}\left(\frac{3N}{4} - c\right) \quad \{\text{upper quartile}\}$$

Note:-
→ For individual observations $Q_1 = \left(\frac{n+1}{4}\right)^{th}$ ~~obs~~ observation
$Q_2 = \left(\frac{n+1}{2}\right)^{th}$ obeservation, $Q_3 = \frac{3(n+1)}{4}^{th}$ observation.

• First we arrange in

* measures of dispersion:-
→ The measurement of the scattered of the given data about the average is said to be a measure of dispersion.
→ The measure of dispersion is commonly used

1. Range (R):-
Range is the difference between the highest and lowest values in the given data
Range (R) = max value − min value

2. Quartile deviation:-
Quartile deviation or semi inter quartile range is given by

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

where Q_1 & Q_3 are 1^{st} & 3^{rd} quartile's from the given data respectively.

3. mean deviation:-
mean deviation is arithmetic mean of absolute deviations from their mean.

$$M.D = \frac{1}{N} \Sigma f_i |x_i - \bar{x}|$$

where
$N = \Sigma f_i$
$\bar{x}$ is mean

Scanned with CamScanner

4. Standard deviation :-

Standard deviation is denoted by 'σ' which is defined as (+ve) positive squareroot of the arithmetic mean of the squares of the deviation of the given values from their arithmetic mean.

For the frequency distribution x_i/f_i for $i = 1,2,3 \dots n$

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

where

$N = \Sigma f_i$

$\bar{x} \rightarrow$ mean

• Note:-

1. The square of standard deviation is called variance

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

(or)

$$\sigma^2 = \frac{1}{N} \Sigma f_i x_i^2 - \left(\frac{\Sigma f_i x_i}{N}\right)^2$$

2. For continuous frequency distribution variance

$$\sigma^2 = h^2 \left[\frac{1}{N} \Sigma f_i d_i^2 - \left(\frac{\Sigma f_i d_i}{N}\right)^2\right]$$

where $d_i = \frac{x_i - A}{h}$

$N = \Sigma f_i$

Problems:-

1. Find the range of the observations 8, 2, 5, 1, 9, 15, 20, 7, 3

Sol:- Range (R) = MAX value − MIN value

= 20 − 1

= 19.

2. Find the quartile deviation of the following data
8, 2, 5, 1, 9, 15, 20, 7, 3, 25, 27

Sol:- Given that
8, 2, 5, 1, 9, 15, 20, 7, 3, 25, 27

Now Ascending order
1, 2, 3, 5, 7, 8, 9, 15, 20, 25, 27

$Q_1 = \left(\frac{n+1}{4}\right)^{th}$ observation $= \left(\frac{11+1}{4}\right)^{th} = 3^{rd}$

$\boxed{\therefore Q_1 = 3}$

$Q_2 = \left(\frac{n+1}{2}\right)^{th}$ observation $= \left(\frac{11+1}{2}\right)^{th} = 6^{th}$

$\boxed{Q_2 = 8}$

$Q_3 = \left[\frac{3}{4}(n+1)\right]^{th}$ observation $= \left(\frac{3}{4}(12)\right)^{th} = 9^{th}$

$\boxed{Q_3 = 20}$

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

$= \frac{20-3}{2}$

$= 8.5$

3. calculate quartile deviation, mean deviation from mean to the following data

marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70
No. of students	6	5	8	15	7	6	3

Sol. Now

marks	No. of students	c.f	mid values	$d_i = \frac{x_i - A}{h}$	$f_i d_i$	$\lvert x_i - \bar{x}\rvert$	$f_i \lvert x_i - \bar{x}\rvert$
0-10	6	6	5	-3	-18	28.4	170.4
10-20	5	11	15	-2	-10	18.4	92
20-30	8	19	25	-1	-8	8.4	62.4
30-40	15	34	35 → A	0	0	1.6	24
40-50	7	41	45	1	7	11.6	81.2
50-60	6	47	55	2	12	21.6	129.6
60-70	3	50	65	3	9	31.6	94.8
	50				-8		659.2

$\therefore N = 50$; $\Sigma f_i d_i = -8$; $\Sigma f_i \lvert x_i - \bar{x}\rvert = 659.2$.

i, Quartile deviation:

Now $\frac{N}{4} = \frac{50}{4} = 12.5$, $\frac{3N}{4} = \frac{3(50)}{4} = 37.5$,

Now c.f just greater than $\frac{N}{4}$ is 19

$f = 8$, $l = 20$, $h = 10$, $C = 11$

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

$= 20 + \frac{10}{18}(12.5 - 11)$

$\boxed{Q_1 = 21.87}$

Scanned with CamScanner

Now. c.f just greater than $\frac{3N}{4}$ is 41

$l = 40,\ f = 7,\ c = 34,\ h = 10$

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

$$= 40 + \frac{10}{7}(37.5 - 34)$$

$$\boxed{Q_3 = 45}$$

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

$$= \frac{45 - 21.87}{2}$$

$$\boxed{Q = 11.56}$$

ii, Mean deviation:

we know that mean $\bar{x} = A + \frac{h \Sigma f_i d_i}{\Sigma f_i}$

where

A = assumed value of $x = 35$

now $\bar{x} = 35 + \frac{10(-8)}{50}$

$$\boxed{\bar{x} = 33.4}$$

Mean deviation $= \frac{1}{N} \Sigma f_i |x_i - \bar{x}|$

$$= \frac{1}{50}(659.2)$$

$$= 13.184$$

4. Calculate mean & standard deviation of the following distribution

marks	0–10	10–20	20–30	30–40	40–50	50–60	60–70
No. of students.	6	5	8	15	7	6	3

Sol:

marks	No. of students	mid values x_i	$d_i = \frac{x_i - A}{h}$	$f_i d_i$	$f_i d_i^2$
0–10	6	5	-3	-18	54
10–20	5	15	-2	-10	20
20–30	8	25	-1	-8	8
30–40	15	35 → A	0	0	0
40–50	7	45	1	7	7
50–60	6	55	2	12	24
60–70	3	65	3	9	27
	$\Sigma f_i = 50$			$\Sigma f_i d_i = -8$	$\Sigma f_i d_i^2 = 140$

$$\text{mean} = \bar{x} = A + h\frac{\Sigma f_i d_i}{\Sigma f_i}$$

$$= 35 + \frac{10(-8)}{50}$$

$$= 33.4$$

Now variance $= \sigma^2 = h^2\left[\frac{1}{N}\Sigma f_i d_i^2 - \left(\frac{1}{N}\Sigma f_i d_i\right)^2\right]$

$$= 100\left[\frac{1}{50}(140) - \left(\frac{1}{50}\times(-8)\right)^2\right]$$

$$= 277.44$$

standard deviation

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

$$= \sqrt{277.44}$$

$$= 16.65$$

* Standard deviation of combination of two groups :-

→ If $\bar{x}_1, \sigma_1$, be the mean and standard deviation of sample size n_1 and $\bar{x}_2, \sigma_2$ be the mean and standard deviation of of sample size n_2 then mean and standard deviation of combined sample size n_1+n_2 is given by

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1+n_2}$$

$$\sigma = \sqrt{\frac{n_1\sigma_1^2 + n_2\sigma_2^2 + n_1D_1^2 + n_2D_2^2}{n_1+n_2}}$$

where. $D_1 = \bar{x}_1 - \bar{x}$

$D_2 = \bar{x}_2 - \bar{x}$

1. The number examined by mean weight and S.D in each group of examination by three examiners are given below. Find mean weight and S.D of entire data when grouped together.

Sol:-

medical exam	No. of examined	mean weight	S.D
A	50	113	6
B	60	120	7
C	90	115	8.

Sol:- From given table $\bar{x}_1 = 113$; $\bar{x}_2 = 120$; $\bar{x}_3 = 115$;

$\sigma_1 = 6$; $\sigma_2 = 7$; $\sigma_3 = 8$;

$n_1 = 50$; $n_2 = 60$; $n_3 = 90$;

Now

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + n_3\bar{x}_3}{n_1+n_2+n_3}$$

$$= \frac{50(113)+60(120)+90(115)}{50+60+90}$$

$$= \frac{5650 + 7200 + 10350}{200}$$

$$\boxed{\bar{x} = 116}$$

$\therefore$ combined mean $\bar{x} = 116$

Now $D_1 = \bar{x}_1 - \bar{x} = 113 - 116 = -3 \quad \therefore D_1^2 = 9$

$D_2 = \bar{x}_2 - \bar{x} = 120 - 116 = 4 \quad \therefore D_2^2 = 16$

$D_3 = \bar{x}_3 - \bar{x} = 115 - 116 = -1 \quad \therefore D_3^2 = 1.$

$$S.D = \sigma = \sqrt{\frac{n_1\sigma_1^2 + n_2\sigma_2^2 + n_3\sigma_3^2 + n_1D_1^2 + n_2D_2^2 + n_3D_3^2}{n_1+n_2+n_3}}$$

$$= \sqrt{\frac{50(6)^2 + 60(7)^2 + 90(5)^2 + 50(9) + 60(16) + 90(1)}{50+60+90}}$$

$$= 7.745$$

$\therefore$ Combined $S.D = \sigma = 7.745.$

* Movements:-

→ The r^{th} movement about the mean $\bar{x}$ of distribution is denoted by μ_r and which is defined as

$$\mu_r = \frac{1}{N} \Sigma f_i (x_i - \bar{x})^r$$

→ The r^{th} movement any point A is denoted by μ'_r and which is defined as

$$\mu'_r = \frac{1}{N} \Sigma f_i (x_i - A)^r$$

• Note:-

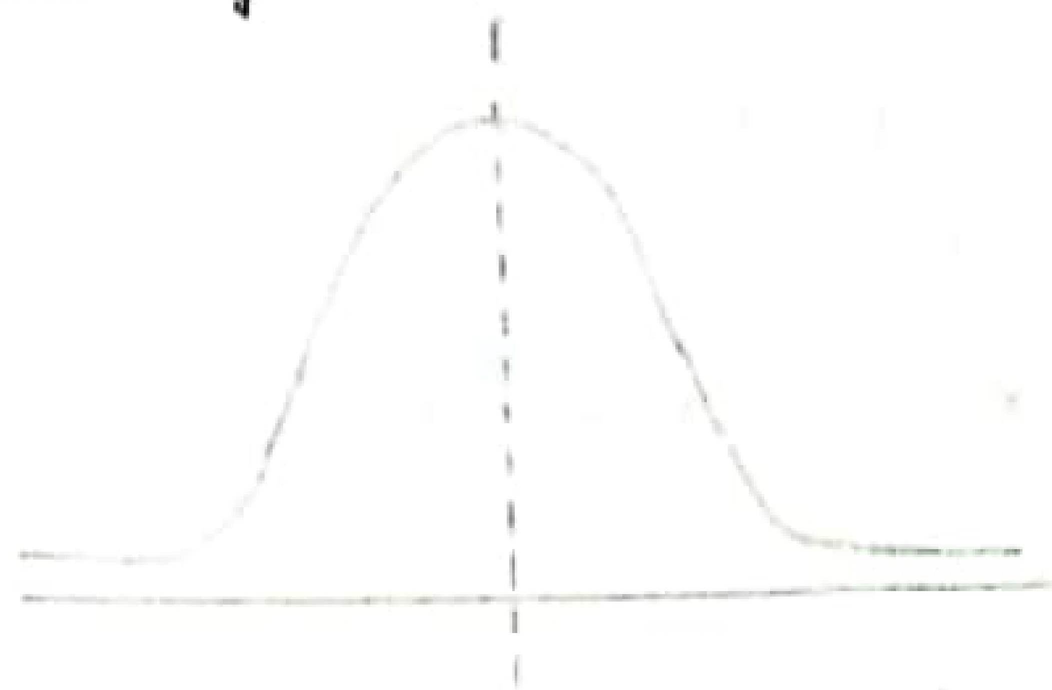
→ 1st and 2nd movements are known as mean & variance

* Skewness:-

Some distribution of data such as the bell curve or normal distribution are symmetric. This means that right & left of the distribution are perfect mirror image of one another.

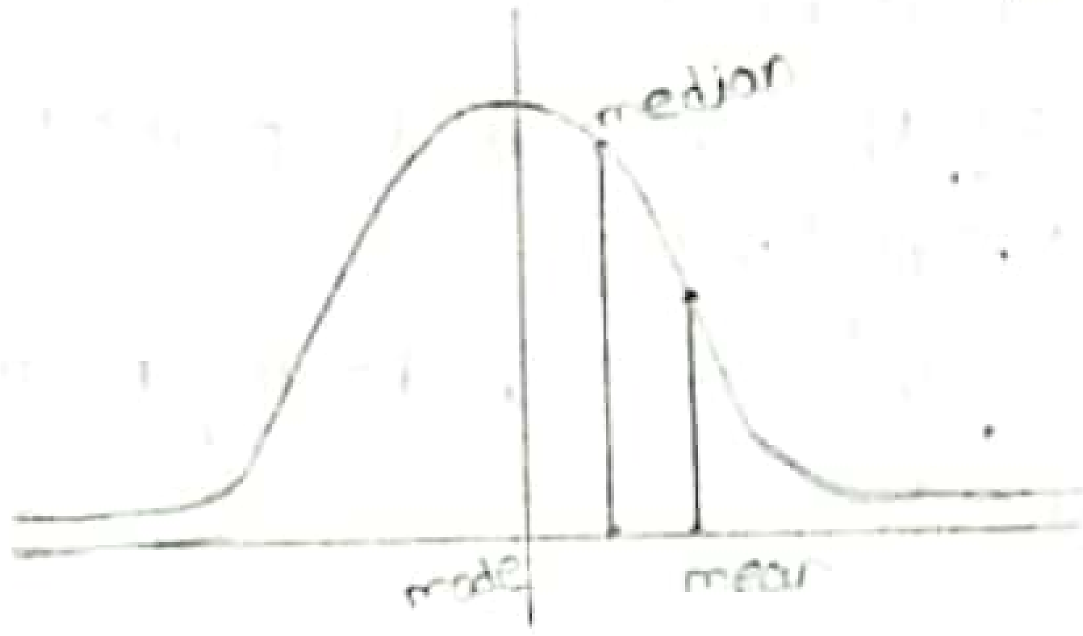
Scanned with CamScanner

→ In symmetrical distribution, mean, median, mode are equal as shown in the figure.

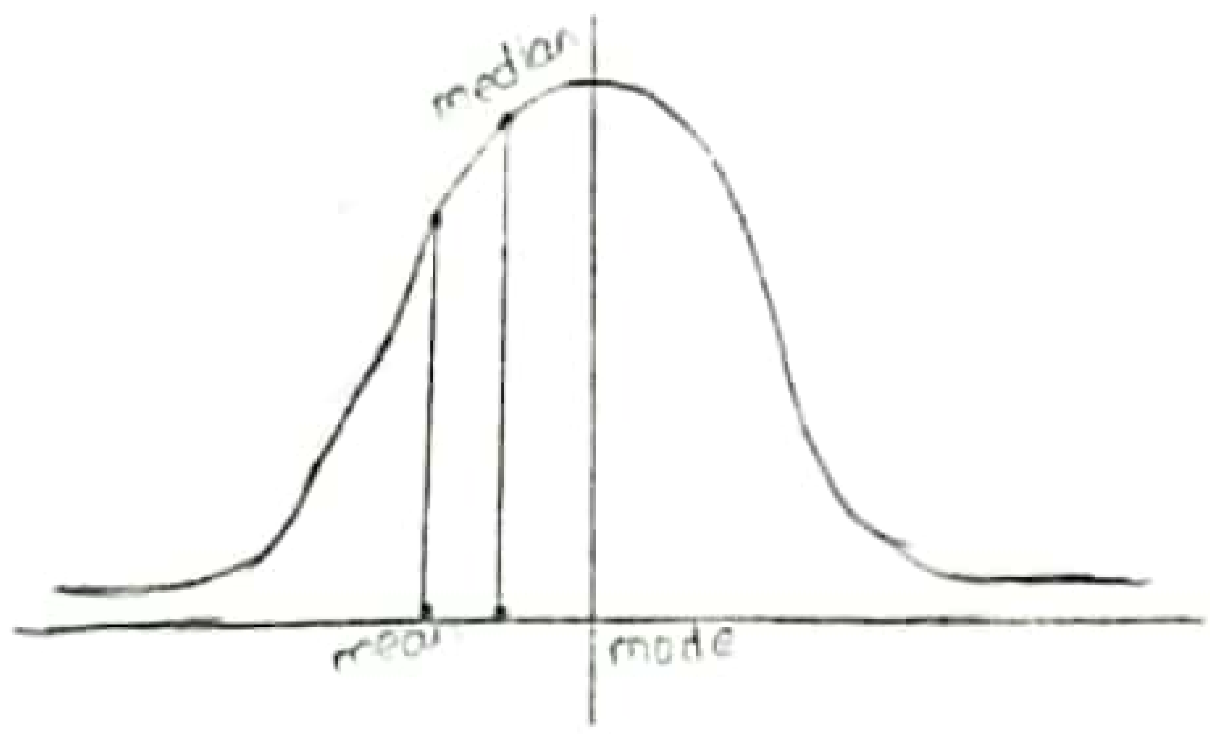


→ Not every distribution of data is symmetric.
→ set of data that are not symmetric are said to be asymmetric.
→ The measures of how asymmetric a distribution can be called as skewness.
→ There are two types of skewness
i, Positive skewness ii, negative skewness

* Positive skewness:-
→ Data that are skew to right have a long tail that extends to right is called positively skewed.
→ In this situation mean and median are greater than mode



* negative skewness:-
→ Data that are skew to left have a long tail that extends to left is called negative skewness (or) negative skewed.
→ In this situation mean and median are less than mode.



→ The following are the coefficients of skewness

i, Pearson's coefficient of skewness (S_k)

$$S_k = \frac{\text{mean} - \text{mode}}{S.D.}$$

ii, Quartile coefficient of skewness (S_k)

$$S_k = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$$

iii, Coefficient of skewness based on 3rd movement $\gamma_1 = \sqrt{\beta_1}$

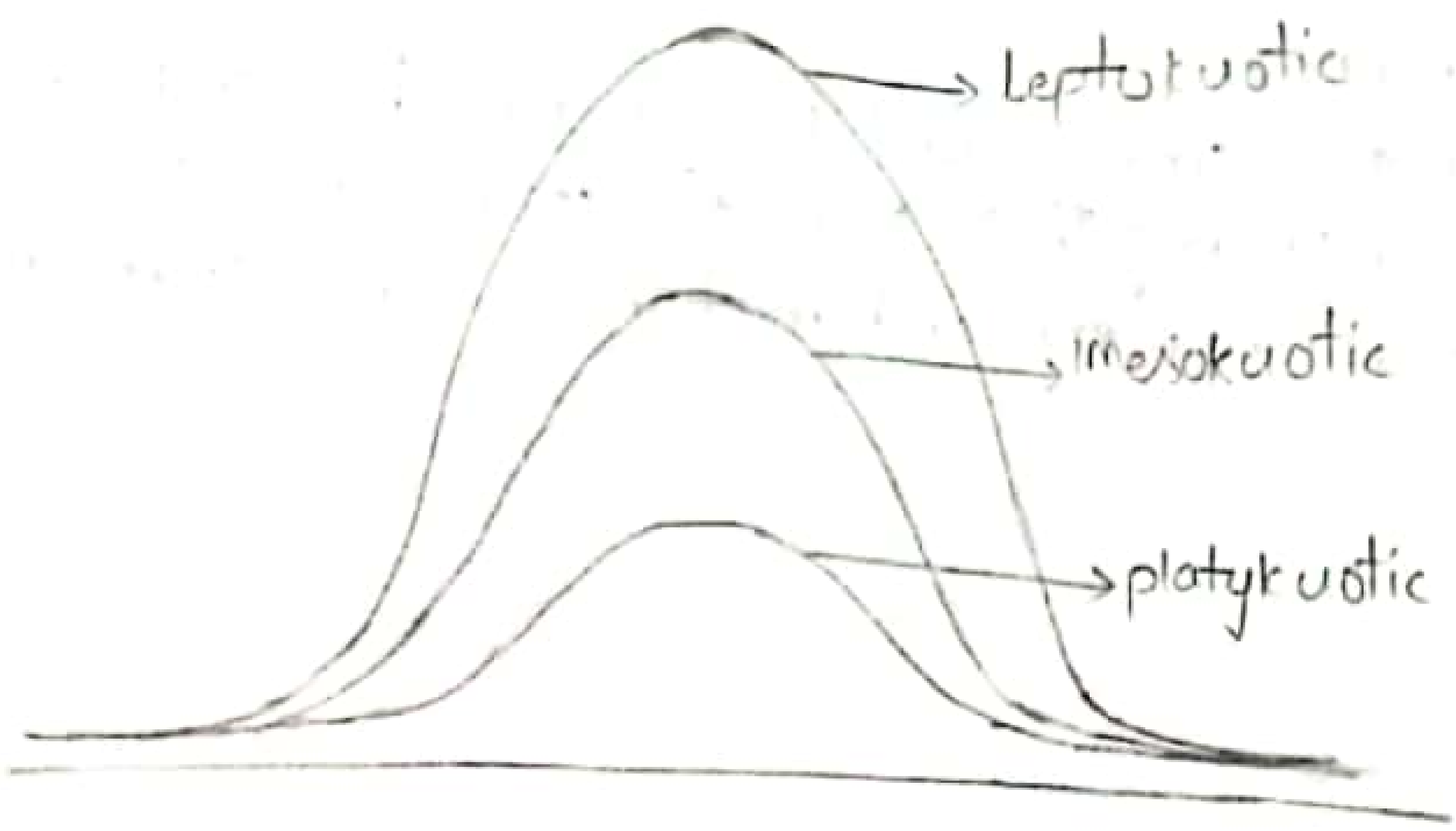
where $\beta_1 = \frac{\mu_3^2}{\mu_2^3}$

* Kurtosis:-

→ Kurtosis enables us to have an idea about flatness (or) Peak-ness of the frequency distribution

→ It is measured by the coefficient $\beta_2 = \frac{\mu_4}{\mu_2^2}$

- If the value $\boxed{\beta_2 > 3}$, the curve is more peaked than normal is called "leptokuotic."
- If the value $\boxed{\beta_2 < 3}$, the curve is less peaked than normal is called "platykuotic."
- If the value $\boxed{\beta_2 = 3}$ the curve is having normal peak is called "mesokuotic."



1. In a frequency distribution, the coefficient of skewness based upon quartile is 0.6, if the sum of upper and lower quartile is 10 and median is 38. Find the values of upper and lower quartiles.

Sol:- we know that Q_1 is lower quartile and Q_2 is middle quartile (median), Q_3 is upper quartile.

→ Given that coefficient of skewness based on quartile

$$S_K = 0.6; \quad Q_3 + Q_1 = 10; \quad Q_2 = 38.$$

we know that

$$S_K = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$$

$$0.6 = \frac{10 - 2(38)}{Q_3 - 10 + Q_3}$$

$$(0.6)(2Q_3 - 10) = 10 - 76$$

$$2Q_3 - 10 = \frac{-66}{0.6}$$

$$2Q_3 - 10 = -110$$

$$2Q_3 = -110 + 10$$

$$2Q_3 = -100$$

$$\boxed{Q_3 = -50}$$

we know that

$$Q_3 + Q_1 = 10$$

$$\boxed{Q_1 = 60}$$

Scanned with CamScanner