STATISTCS FOR ENGINEERS (MAT 2001)

Module: I Introduction to Statistics

Introduction to statistics and data analysis - Measure of central tendency - Measure of variability - [moment - Skewness - Kurtosis (concepts only)).

Measure of central Tendency

(i) Arithmetic mean
(ii) Median
(iii) Mode

This is the central value for a probability distribution.

(1) Arithmetic mean (Direct method)

Arithmetic mean of a set of observations is their sum divided by the number of observations. i.e. The arithmetic mean $\bar{x}$ of n observation $x_1, x_2, \dots, x_n$ is given by

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \dots + x_n) = \frac{\sum_{j=1}^{n} x_j}{n}$$

If frequency distribution f_i is given for x_i, $i = 1, 2, \dots n$ where f_i is the frequency of x_i

$$\bar{x} = \frac{f_1 x_1 + f_2 x_2 + \dots + f_n x_n}{f_1 + f_2 + \dots + f_n} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$$

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

* In case of grouped or continuous frequency distribution, x is taken as the mid-value of the corresponding class.

Ex. (1) Find the arithmetic mean of the following frequency distribution

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

Soln

x	f	fx
1	5	5
2	9	18
3	12	36
4	17	68
5	14	70
6	10	60
7	6	42
Sum	73	299

Here $N = \Sigma f = 73$

$\Sigma fx = 299$

$\bar{x} = \frac{\Sigma fx}{N} = \frac{299}{73}$

$= 4.09$

Note * Data is often described as ungrouped or grouped

* Ungrouped data is data given as individual data pts.

Ex (1) Ungrouped data without a frequency distribution

1, 3, 6, 4, 5, 6, 3, 4, 6, 3, 6

(2) Ungrouped data with a frequency distribution

No. of TV sets	freq
0	2
1	13
2	18
3	0
4	10
5	2
Total	45

* Grouped data is data given in intervals

Ex

Exam. Score	Frequency
90-99	7
80-89	5
70-79	15
60-69	4
50-59	5
40-49	0
30-39	1
Total	37

Ex- Calculate the arithmetic mean of the marks from the following table

Marks:	0-10	10-20	20-30	30-40	40-50	50-60
No of Students:	12	18	27	20	17	6

Soln

Marks	No of Students (f)	Mid pt (x)	fx
0-10	12	5	60
10-20	18	15	270
20-30	27	25	675
30-40	20	35	700
40-50	17	45	765
50-60	6	55	330
Total	N = 100		2800

Arithmetic mean $(\bar{x}) = \frac{1}{N} \Sigma fx = \frac{2800}{100} = 28$

Mean of the composit series

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + \cdots + n_k\bar{x}_k}{n_1 + n_2 + \cdots + n_k} = \sum_i n_i\bar{x}_i \Big/ \sum_i n_i$$

Q.N The average salary of male employees in a firm was Rs. 5200 and that of females was Rs. 4200. The mean salary of all the employees was Rs 5000. Find the percentage of male and female employees

Soln $n_1 \rightarrow$ No. of male emp

$n_2 \rightarrow$ no. of female emp

$\bar{x}_1$ & $\bar{x}_2 \rightarrow$ Average salary in rupees

$\bar{x} \rightarrow$ Average salary of all workers in the firm

$\bar{x}_1 = 5200, \quad \bar{x}_2 = 4200 \quad \& \quad \bar{x} = 5000$

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

$$5000(n_1 + n_2) = 5200n_1 + 4200n_2$$

$$200n_1 = 800n_2$$

$$\Rightarrow \frac{n_1}{n_2} = \frac{4}{1}$$

$\therefore$ The percentage of male employees in the firm

$$= \frac{4}{4+1} \times 100 = 80\%$$

& the percentage of female employees in the firm

$$= \frac{1}{4+1} \times 100 = 20\%$$

Airthmetic mean - Short-cut method

$$\bar{x} = A + \frac{\Sigma fd}{N}$$ For discrete series

where A = Assumed mean

$d = X - A$

N = Total no of observations ie- Σf

Step (i) Take an assumed mean

(ii) Take the deviations of variable X from assumed mean ($i.e\ d = X - A$)

(iii) Multiply deviations with the respective frequency and take the total Σfd

(iv) Divide the total obtained in (iii) by total frequency

Ex Find the airthmetic mean

X	No of Student f	$(X - A)$ 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
	$N = 60$		$\Sigma fd = 60$

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

Arithmetic mean – continuous data (Short cut method)

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

A = assumed mean, d = deviation of X from assumed mean
i.e. (m − A)
↓
mid pt.

N = total no. of observations

Ex

Marks	Mid-point m	No of Students 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

divide the deviation (m − A) by the class intervals.
i.e. calculate (m − A)/h and then multiply by h.

$$\bar{x} = A + \frac{\Sigma fd}{N} \times h$$

EX

Marks	mid point (m)	No of Studs (f)	(m-35)	(m-35)/h = d	fd
0-10	5	5	-30	-3	-15
10-20	15	10	-20	-2	-20
20-30	25	25	-10	-1	-25
30-40	(35) ← A	30	0	0	0
40-50	45	20	10	1	20
50-60	55	10	20	2	20
		N = 100			$\Sigma fd = -20$

$$\bar{x} = A + \frac{\Sigma fd}{N} \times h$$

$$= 35 - \frac{20}{100} \times 10 = 33$$

* 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		$\Sigma fd = -44$

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

Merits of Arithmetic mean

1. Simplest average to understand and easiest compute.
2. It is affected by the value of every item in series
3. It is defined by a rigid formula.
4. It is relatively reliable in the sense that it does not vary too much when repeated samples are taken from one and the same population.
5. mean is typical in the sense that it is the centre of gravity, balancing the values on either side of it
6. It is a calculated value, and not based on position in the series

Limitations

* Since the value of mean depends upon each item of the series, extreme items i.e. very small and very large items, unduly affect the value of average The smaller the no. of observations, the greater is likely to be the impact of extreme values.
* In a distribution with open-end classes, mean cannot be computed without making assumptions. If such classes contain a large proportion of the values, then mean may be subject to substantial error.
* Arithmetic mean is not always a good measure of central tendency. The mean provides a characteristic value, in the sense of indicating where most of the values lie; only when the distribution of the variable is reasonably normal.

Median

Median of a distribution is the value of the variable which divides it into two equal parts.

* median is a positional average

calculation of median - Individual observation

step - ① Arrange data in ascending or descending order

② If no. of data is odd then median is the middle value

If no. of data is even, then airthemetic mean of the middle value is median.

Ex ① Data of wages of 7 workers (odd)

Wages 14,100, 14,150, 16,080, 17,120, 15,200, 16,160, 17,400

Soln Step-1

S.N	Wages (ascending)
1	14,100
2	14,150
3	15,200
4	16,080
5	16,160
6	17,120
7	17,400

Here no. of data is odd

$\text{median} = \frac{N+1}{2} = \frac{7+1}{2} = 4^{th}$ item

$= 16,080$ Rs

Ex ② monthly income of 10 employee (even)

14,391, 15,384 15,407 16,672 26,522 16,772 26,753 27850 37490

Soln

S.no.	Income (Ascending order)
1	14391
2	15384
3	15407
4	16672
5	16777
6	25591
7	26522
8	26753
9	27850
10	37490

(2)

Median = Size of $\frac{N+1}{2}$ th item

$= \frac{10+1}{2} = 5.5^{th}$ item

Size of 5.5^{th} item

$= \frac{5^{th} \text{ item} + 6^{th} \text{ item}}{2}$

$= \frac{16777 + 25591}{2}$

$= \frac{42368}{2} = 21,184$

Computation of Median - Discrete Series

Step ① Arrange data in ascending or descending order

② Calculate median = Size of $\frac{N+1}{2}$

③ Find the cumulative freqs

④ Look at the cumulative freq. and find that total which either equal to $\frac{N+1}{2}$ or next higher to that and determine the value of the corresponding variable

Ex

Income	15000	15500	16800	18000	18500	17800
No of persons	24	26	16	20	6	30

76)

Income ascending order	No of person (f)	Cumulative freq
15000	24	24
15500	26	50
16800	20	70
17800	30	100
18000	16	116
18500	6	122

Medicm = Size of $\frac{N+1}{2}$ th item = $\frac{122+1}{2}$ = 61.5 th item

Size of 61.5th item = 16800 Rs

Calculation of Medicm - continous Series

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

l = lower limit of the medicm class i.e. the class in which the middle item of the distribution lie,

c.f. = cumulative freq. of the class preceding the medicm class or sum of the freq. of all classes lower than the medicm class.

f = Simple freq. of the medicm class

h = The class interval of the medicm class

$N = \Sigma f$

Ex Find the medicm marks

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 Studen (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	

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

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

Medicm lies in the class 25-30

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

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

$N = 200 \quad cf = 77$

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

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

$= 27.74$

Q Calculate median for the following data (5)

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

Since the data given inclusive class interval, we should first convert it into exclusive by deducting 0.5 from lower limit and adding 0.5 to upper limit

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	

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

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

⇒ Median lies in the class 439.5 - 449.5

$L = 439.5, \quad h = 10$

$cf = 76, \quad f = 54$

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

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

$= 439.5 + 4.44$

$= 443.94$

Ex. An incomplete distribution is given by

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 frq 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

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

$f = 40$

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

$$f_1 = 35$$

$\Rightarrow$ $f_2 = 60 - 35$

$= 25$

Merits of Median

1. Useful for the open-end classes also recommended if distribution has unequal class
2. Extreme values do not affect the median as strongly they do the mean.
3. In markedly skewed distributions where the arithmetic mean would be distorted by extreme values, the median is especially useful.
4. most appropriate average in dealing with qualitative data
5. Value of median can be determined graphically.

Limitations

1. It is necessary to arrange the data in order.
2. It is a positional avg, its value is not determined by each and every observation
3. Median is more affected by sampling fluctuations than the arithmetic mean
4. In some cases cannot be computed exactly as the mean. (For even data it gives approxi-mately mid-point of the two middle items).

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.

In case of Discrete freqn distribution

Mode is the value of x corresponding to the maximum frequency.

Ex

x:	1	2	3	4	5	6	7	8
f:	6	9	11	19	17	15	9	7

max freq is 19 & corresponding x is 4

$\Rightarrow$ mode is 4.

* For some cases the above method is not possible
  - ↳ If the maximum freq is repeated
  - ↳ If the max. freq occurs in the very begining or at the end of the distribution
  - ↳ If there are irregularities in the distribution

Then mode is determined by the method of grouping.

Ex Calculate the value of mode for the following data.

Marks	10	15	20	25	30	35	40
Freq.	8	12	36	35	28	18	9

Soln

Grouping Table

X	(f) I	II	III	IV	V	VI
10	8	20				
15	12		48			
20	(36)	(71)		56		
25	35		63		83	
30	28	46		81		99
35	18		27		55	
40	9					

Analysis Table

Col no.	20	25	30
I	1	—	—
II	1	1	—
III	—	1	1
IV	—	1	1
V	1	1	—
VI	1	1	1
	4	(5) max	3

⇒ modal value is 25

Calculation of Mode - Continuous Series

$$\text{Mode} = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times h$$

where L = Lower limit of the modal class

f_1 = Freq. of the modal class

f_0 = " " " class preceeding the modal class

f_2 = " " " " succeeding the modal class

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^n 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

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

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				
110 - 120	6		26			
120 - 130	20	52		30		
130 - 140	32		65		58	
140 - 150	33	50		82		85
150 - 160	17		25		58	
160 - 170	8	10				27
170 - 180	2					

Analysis table

Col. No	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

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

Weight	m	f	c.f.	$\frac{m-A}{h} = d$	fd
100-110	105	4	4	-3	-12
110-120	115	6	10	-2	-12
120-130	125	20	30	-1	-20
130-140	135 (A)	32	62	0	0
140-150	145	33	95	1	33
150-160	155	17	112	2	34
160-170	165	8	120	3	24
170-180	175	2	122	4	8
		$N = 122$			$\Sigma fd = 55$

$$\bar{x} = A + \frac{\Sigma fd}{N} \times h \qquad [A=135,\ N=122,\ h=10,\ \Sigma fd=55]$$

$$= 135 + \frac{55}{122} \times 10 = 139.51$$

$$\text{Median} = \frac{N}{2}\text{th item} = \frac{122}{2} = 61^{st}\text{ item}$$

$\Rightarrow$ median class = 130-140

$$\text{Median} = L + \frac{h}{f}\left(\frac{N}{2} - c.f.\right) \qquad [L=130,\ h=10,\ f=32,\ N=122,\ c.f.=30]$$

$$= 130 + \frac{10}{32}(61-32) = 139.69$$

$\therefore$ Mode $= 3 \times$ Median $- 2$ Mean

$= 3 \times 139.69 - 2 \times 139.51$

$= 140.05$

Merits of mode

1. Like median, mode is not unduly affected by extreme values.
2. Its value can be determined in open-end distributions without ascertaining the class limits.
3. It can be used to describe qualitative phenomenon.
4. The value of mode can also be determined graphically.

Limitations

1. It can not not always be determined. In some cases we may have a bimodal series
2. It is not capable of algebraic manipulations
3. mode is not based on each & every item of the series
4. While dealing with quantitative data, the disadvantages of the mode outweighs its good features and hence it is seldom used.

Measures of Variability

1. Range
2. Quartile Deviation or Semi-interquartile Range
3. Mean Deviation
4. Standard Deviation & Root mean square deviation

Range

$$\text{Range} = \text{Maximum value} - \text{Minimum value}$$

Range is the difference between two extreme observations of the distribution.

* Since it is based on two extreme observations which themselves are subject to chance fluctuations. Therefore, it is not at all a reliable measure of dispersion.

Quartile deviation

The three points which divide the series into four equal parts are called quartiles.

1st quartile Q_1 is the value which exceed 25% of the observations and is exceeded by 75% of the observations

2nd quartile, Q_2, coincides with median.

3rd quartile, Q_3, is the pt which has 75% observations before it and 25% observations after it

$$\text{Quartile deviation} \quad Q = \frac{1}{2}(Q_3 - Q_1)$$

where Q_1 & Q_3 are 1st & 3rd quartiles

Quartile deviation is a better measure than the range as it makes use of 50% of the data. But ignores the other 50% of the data, it cannot be regarded as a reliable measure.

Ex. Calculate median, quantiles

x	0	1	2	3	4	5	6	7	8
f	1	9	26	59	72	52	29	7	1
c.f	1	10	36	95	167	219	248	255	256

Median $\frac{N}{2} = \frac{256}{2} = 128$

c.f greater than 128 is 167

$\Rightarrow$ median = 4

Q_1: $\frac{N}{4} = 64$, c.f greater than 64 is 95

$\Rightarrow Q_1 = 3$

Q_3: $\frac{3N}{4} = 192$, c.f greater than 192 is 219

$\Rightarrow Q_3 = 5$

$\therefore$ quartile deviation $= \frac{1}{2}(Q_3 - Q_1) = \frac{5-3}{2} = 1$

Mean Deviation

If $x_i | f_i$, $i = 1, 2, \cdots, n$ is the freq. distribution, then mean deviation from the average A, (Usually mean, median, or mode) is given by

$$\frac{1}{N} \sum_i f_i |x_i - A|, \quad \Sigma f_i = N$$

where $|x_i - A|$ represents the modulus of the absolute value of the deviation $(x_i - A)$, when the $-ve$ sign is ignored. It is a better measure of dispersion, since it is based on all the observations.

But ignoring the signs of the deviations $(x_i - A)$ creates artificiality and renders it useless for further mathematical treatment.

* Mean deviation is least when taken from median.

Standard deviation (σ)

For the freq. distribution $x_i | f_i$, $i = 1, 2, \dots, n$

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

(It measures the absolute dispersion or variability of distribution)

where $\bar{x}$ is the arithmetic mean.

Variance

Square of standard deviation is ~~variation~~ variance.

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

Root mean Square Deviation

$$S = \sqrt{\frac{1}{N} \sum_i f_i (x_i - A)^2}$$

where A is any arbitary no. ∈

In a frequency distribution where deviations are taken from assumed mean variation may directly be computed as

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

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

Standard deviation

When deviations are taken from assumed mean

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

where $d = X - A$

For discrete series using assumed mean

$$\sigma = \sqrt{\frac{\sum f d^2}{N} - \left(\frac{\sum f d}{N} \right)^2}, \quad d = X - A$$

Calculation of mean deviation

$$M.D = \frac{1}{N}\sum_{i=1}^{n} f_i |x_i - A|$$

Ex Calculate mean deviation from the following series

x	f	c.f.	$\|x_i - 12\| = \|D\|$	$f\|D\|$
10	3	3	2	6
11	12	15	1	12
12	18	33	0	0
13	12	45	1	12
14	3	48	2	6
	$N = 48$			$\Sigma f\|D\| = 36$

$$\text{median} = \frac{N+1}{2}^{th} \text{ item} = \frac{48+1}{2} = 24.5^{th} \text{ item}$$

$\Rightarrow$ median $= 12$

$$\text{Mean deviation} = \frac{36}{48} = 0.75$$

Ex Mean deviation from the mean

x	f	fx	$\|x - \bar{x}\|$	$f\|x - \bar{x}\|$
2	2	4	6	12
4	2	8	4	8
6	4	24	2	8
8	5	40	0	0
10	3	30	2	6
12	2	24	4	8
14	1	14	6	6
16	1	16	8	8
	$N = 20$	$\Sigma fx = 160$		$\Sigma f\|D\| = 56$

$$\bar{x} = \frac{\Sigma fx}{N} = \frac{160}{20} = 8$$

$$M.D. = \frac{\Sigma f|D|}{N} = \frac{56}{20} = 2.8$$

Mean deviation – continuous series

Ex

x	f	c.f	middle	$\|\text{middle} - 43\| \rightarrow D$	$f\|D\|$
0-10	5	5	5	38	190
10-20	8	13	15	28	224
20-30	12	25	25	18	216
30-40	15	40	35	8	120
40-50	20	60	45	2	40
50-60	14	74	55	12	168
60-70	12	86	65	22	264
70-80	6	92	75	32	192
	$\Sigma f = 92$				$\Sigma f\|D\| = 1414$

$\text{Med} = \text{Size of } \frac{N}{2}^{th} = \frac{92}{2} = 46^{th}$

median class 40-50

$\text{median} = L + \frac{h}{f}\left(\frac{N}{2} - cf\right) = 40 + \frac{10}{20}(46 - 40) = 43$

$\text{mean deviation from median} = \frac{\Sigma f|D|}{N} = \frac{1414}{92} = 15.37$

$\text{Coeff of M.D.} = \frac{\text{M.D}}{\text{median}} = \frac{15.37}{43} = 0.357$

Ex. Calculate the mean & Standard deviation for the following table

Age in years	No. of members	x, mid value	$d = \frac{x-55}{10}$	fd	fd^2
20 - 30	3	25	-3	-9	27
30 - 40	61	35	-2	-122	244
~~40~~ 50	~~153~~ 132	45	-1	-132	132
~~140~~ 50-60	153	55	0	0	0
60-70	140	65	1	140	140
70-80	51	75	2	102	204
80-90	2	85	3	6	18
	$\Sigma f = 542$			$\Sigma fd = -15$	$\Sigma fd^2 = 765$

$$\bar{x} = A + h\frac{\Sigma fd}{N} = 55 + \frac{10(-15)}{542} = 55 - 0.28 = 54.72$$

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

$$= (10)^2\left\{\frac{1}{542} \times 765 - \left(\frac{1}{542} \times (-15)\right)^2\right\}$$

$$= 100 \times 1.333 = 133.3$$

Standard deviation $\sigma = 11.55$ year

Coefficient of Variation

$$C.V. = 100 \times \frac{\sigma}{\bar{X}}$$

It is a relative measure of dispersion

It is used for comparing the variability of two series.

Ex From the prices of shares of X & Y below find out which is more stable

X	35	54	52	53	56	58	52	50	51	49
Y	108	107	105	105	106	107	104	103	104	101

Sol^n

X	$X-\bar{X}$	$(X-\bar{X})^2$	Y	$Y-\bar{Y}$	$(Y-\bar{Y})^2$
35	-16	256	108	3	9
54	+7	9	107	2	4
52	1	1	105	0	0
53	2	4	105	0	0
56	5	25	106	1	1
58	7	49	107	2	4
52	1	1	104	-1	1
50	-1	1	103	-2	4
51	0	0	104	-1	1
49	-2	4	101	-4	16
$\Sigma X = 510$	$\Sigma = 0$	$\Sigma(x-\bar{x})^2 = 350$	$\Sigma Y = 1050$	$\Sigma = 0$	$\Sigma(y-\bar{y})^2 = 40$

For X:

$$\bar{X} = \frac{\Sigma X}{N} = \frac{510}{10} = 51$$

$$\sigma = \sqrt{\frac{\Sigma(X-\bar{X})^2}{N}} = \sqrt{\frac{350}{10}} = 5.916$$

$$C.V. = 100 \times \frac{\sigma}{\bar{X}} = \frac{5.916}{51} \times 100 = 11.6$$

For Y

$$\bar{Y} = \frac{\Sigma Y}{N} = \frac{1050}{10} = 105$$

$$\sigma = \sqrt{\frac{40}{10}} = 2$$

$$C.V. = 100 \times \frac{2}{105} = 1.905$$

$\therefore$ CV for Y is less

$\Rightarrow$ Y is more stable

Coefficient of dispersion (C.D)

These are the coefficients i.e. pure numbers independent of the units of measurement.

(1) C.D. based on range $= \frac{A-B}{A+B}$, where A & B are the greatest & smallest items

(2) C.D based on quartile deviation

$$C.D. = \frac{(Q_3 - Q_1)/2}{(Q_3 + Q_1)/2} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

(3) C.D. based on mean deviation

$$C.D. = \frac{\text{Mean deviation}}{\text{Average from which it is calculated}}$$

(4) Based on Standard deviation

$$C.D = \frac{S.D.}{\text{Mean}} = \frac{\sigma}{\bar{x}}$$

Moments

In statistics it describes the various characters of a freq. distribution like central tendency, variation, skewness & kurtosis.

The rth moment of a variable x - about any point $x = A$, denoted by μ'_r

$$\mu'_r = \frac{1}{N} \sum_i f_i (x_i - A)^r, \quad \sum f_i = N$$

$$= \frac{1}{N} \sum_i f_i d_i^r, \quad d = x_i - A$$

rth moment about mean $\bar{x}$

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

$$z_i = x_i - \bar{x}$$

Particular case

$$\mu_0 = \frac{1}{N} \sum_i f_i (x_i - \bar{x})^0 = \frac{1}{N} \sum_i f_i = 1$$

$$\mu_1 = \frac{1}{N} \sum_i f_i (x_i - \bar{x}) = 0,$$ The Algebraic sum of deviations from mean zero.

$$\mu_2 = \frac{1}{N} \sum_i f_i (x_i - \bar{x})^2 = \sigma^2$$

$$\boxed{\mu_0 = 1, \quad \mu_1 = 0, \quad \mu_2 = \sigma^2}$$

---

4 Central moments

(1) 1st moment $r = 1$, sum of the difference of each observation from sample avg., is zero

(2) 2nd central moment $r = 2$, is variance

(3) 3rd $r = 3$, is Skewness

(4) 4th central $r = 4$ is kurtosis.

Pearson's co-efficients

Karl Pearson defined the following four coefficients based on first four moments about mean

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

$$\gamma_1 = +\sqrt{\beta_1}$$

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

$$\gamma_2 = \beta_2 - 3$$

Skewness (Lack of symmetry)

A distribution is said to be Skewed if

(i) Mean, median and mode fall at diff^{n} points

Mean $\neq$ median $\neq$ mode

(ii) Quartiles are not equidistant from median.

(iii) The curve drawn with the help of the given data is not symmetrical but stretched more to one side than to the other.

Measure of Skewness

(1) $S_k = M - M_d$

(2) $S_k = M - M_o$

M = mean
M_o = Mode
M_d = median

(3) $S_k = (Q_3 - M_d) - (M_d - Q_1)$

These are all absolute measure of skewness.

For comparing two series we calculate the relative measures called coefficients of skewness.

1-) Kar Pearson's coeff^{n} of Skewness

$$S_k = \frac{M - M_o}{\sigma}$$

2) Bowley's coeff^{n} of skewness

$$S_k = \frac{(Q_3 - M_d) - (M_d - Q_1)}{(Q_3 - M_d) + (M_d - Q_1)} = \frac{Q_3 + Q_1 - 2M_d}{Q_3 - Q_1}$$

3) Based on moments

$$S_k = \frac{\sqrt{\beta_1}(\beta_2 + 3)}{2(5\beta_2 - 6\beta_1 - 9)}$$

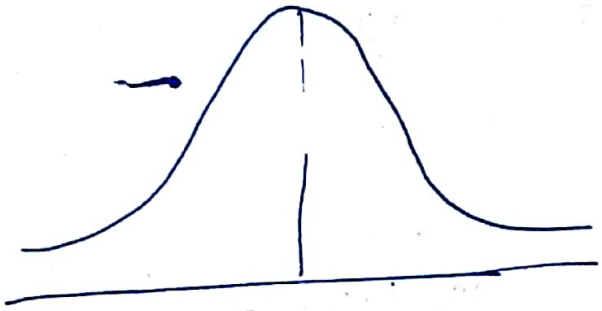
$S_k = 0$, if $\beta_1 = 0$, or $\beta_2 = -3$

$\therefore$ $\beta_2 = \frac{\mu_4}{\mu_2^2}$ can not be -ve

$\Rightarrow$ $S_k = 0$ iff $\beta_1 = 0$

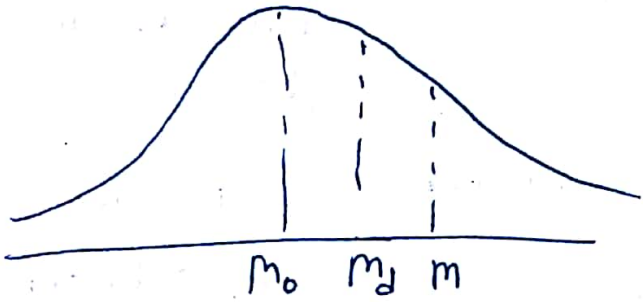
$\therefore$ For a symmetrical distribution $\beta_1 = 0$

Graphical representation

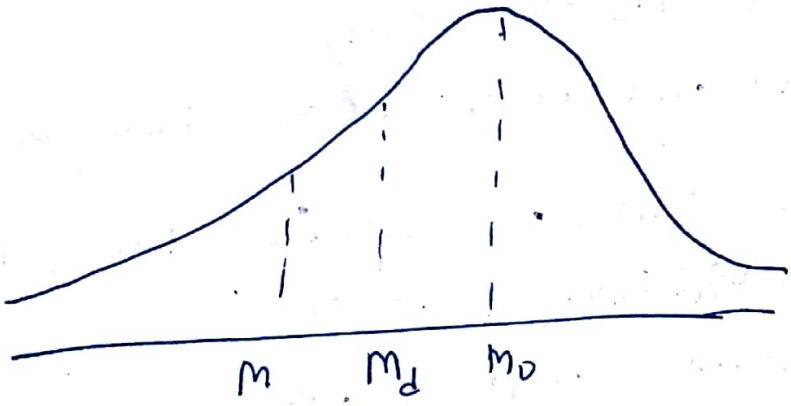


$\bar{x} = m_o = m_d$

(symmetrical distribution)



(+ve skewed)



(-ve skewed)

Kurtosis

It gives an idea about the flatness or peakedness of the curve.

It is measured by the coeff β_2

$$\beta_2 = \frac{\mu_4}{\mu_2^2}, \quad \gamma_2 = \beta_2 - 3$$