MAT2001
Statistics for Engineers

Module 1
Introduction to Statistics

Syllabus

Introduction to Statistics:
Introduction to statistics and data analysis-Measures of central tendency – Measures of variability- [Moments-Skewness-Kurtosis (Concepts only)].

Statistics

- **Introduction**
- **Data Science**
- **Data Analysis**

Frequency Distribution

1. Discrete Frequency Distribution
2. Continuous Frequency Distribution

Measures of Central Tendency

List of Measures of Central Tendency:

1. Arithmetic Mean or Average
2. Median
3. Mode
4. Geometric Mean
5. Harmonic Mean

1. Arithmetic Mean or Average (M)

Arithmetic Mean. Arithmetic mean of a set of observastions is their sum divided by the number of observations, *e.g.*, the arithmetic mean $\bar{x}$ of n observations $x_1, x_2, \ldots, x_n$ is given by

$$\bar{x} = \frac{1}{n}\left(x_1 + x_2 + \ldots + x_n \right) = \frac{1}{n} \sum_{i=1}^{n} x_i$$

In case of frequency distribution $x_i \mid f_i$, $i = 1, 2, \ldots, n$, where f_i is the frequency of the variable x_i;

$$\bar{x} = \frac{f_1 x_1 + f_2 x_2 + \ldots + f_n x_n}{f_1 + f_2 + \ldots f_n} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} = \frac{1}{N} \sum_{i=1}^{n} f_i x_i, \quad \left[\sum_{i=1}^{n} f_i = N \right]$$

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

Arithmetic Mean or Average

x_i :	x_1	x_2	. . .	x_n
f_i :	f_1	f_2	. . .	f_n

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

where $N = \sum_{i=1}^{n} f_i$.

Example:

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

Solution:

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

$$\bar{x} = \frac{1}{N} \Sigma fx = \frac{299}{73} = 4{\cdot}09$$

Example:

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*

Solution:

Marks	*No. of students* (*f*)	*Mid - point* (*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	100		2,800

Arithmetic mean or $\bar{x} = \frac{1}{N} \Sigma f x = \frac{1}{100} \times 2{,}800 = 28$

Arithmetic Mean or Average

It may be noted that if the values of x or (and) f are large, the calculation of mean by formula is quite time-consuming and tedious. The arithmetic is reduced to a great extent by taking the deviations of the given values from any arbitrary point 'A', as explained below.

Let $d_i = x_i - A$, then $f_i d_i = f_i (x_i - A) = f_i x_i - A f_i$

Summing both sides over i from 1 to n, we get

$$\sum_{i=1}^{n} f_i d_i = \sum_{i=1}^{n} f_i x_i - A \sum_{i=1}^{n} f_i = \sum_{i=1}^{n} f_i x_i - A \cdot N.$$

$$\Rightarrow \quad \frac{1}{N} \sum_{i=1}^{n} f_i d_i = \frac{1}{N} \sum_{i=1}^{n} f_i x_i - A = \bar{x} - A$$

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

$$\therefore \quad \bar{x} = A + \frac{1}{N} \sum_{i=1}^{n} f_i d_i$$

Arithmetic Mean or Average

In case of grouped or continuous frequency distribution, the arithmetic is reduced to a still greater extent by taking

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

where A is an arbitrary point and h is the common magnitude of class interval. In this case, we have

$$h\, d_i = x_i - A,$$

and proceeding exactly similarly as above, we get

$$\bar{x} = A + \frac{h}{N} \sum_{i=1}^{n} f_i d_i$$

Example:

Calculate the mean for the following frequency distribution.

Class-interval :	*0–8*	*8–16*	*16–24*	*24–32*	*32–40*	*40–48*
Frequency :	*8*	*7*	*16*	*24*	*15*	*7*

Solution:

Class-interval	*mid-value* (x)	*Frequency* (f)	$d = (x - A)/h$	fd
0–8	4	8	–3	–24
8–16	12	7	–2	–14
16–24	20	16	–1	–16
24–32	28	24	0	0
32–40	36	15	1	15
40–48	44	7	2	14
		77		–25

Here we take $A = 28$ and $h = 8$.

$$\therefore \quad \bar{x} = A + \frac{h \Sigma f d}{N} = 28 + \frac{8 \times (-25)}{77} = 28 - \frac{200}{77} = 25{\cdot}404$$

2. Median (M_d)

Median of a 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, *i.e.*, it 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.

Example:

For example, the median of the values 25, 20, 15, 35, 18, *i.e.*, 15, 18, 20, 25, 35 is 20 and the median of 8, 20, 50, 25, 15, 30, *i.e.*, of 8, 15, 20, 25, 30, 50 is $\frac{1}{2}(20 + 25) = 22{\cdot}5$.

Median for 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:

(i) Find $N/2$, where $N = \sum_i f_i$.

(ii) See the (less than) cumulative frequency (*c.f.*) just greater than $N/2$.

(iii) 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
	120	

Hence $N = 120 \Rightarrow N/2 = 60$

Cumulative frequency (*c.f.*) just greater than $N/2$, is 65 and the value of x corresponding to 65 is 5. Therefore, median is 5.

Median for Continuous Frequency Distribution

In the case of continuous frequency distribution, the class corresponding to the *c.f.* just greater than $N/2$ is called the *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 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 *c.f.* of the class *preceding* the median class,

and $N = \Sigma f$.

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 $N/2 = 43/2 = 21{\cdot}5$. Cumulative frequency just greater than 21·5 is 28 and the corresponding class is 40–50. Thus median class is 40–50.

$$\text{Median} = 40 + \frac{10}{20}(21{\cdot}5 - 8) = 40 + 6{\cdot}75 = 46{\cdot}75$$

Thus median wage is Rs. 46·75.

3. Mode (M_o)

Mode is the value which occurs most frequently in a set of observastions and around which the other items of the set cluster densely. In other words, mode is the value of the variable which is predominant in the series.

Mode for Discrete Frequency Distribution

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

Example:

For the following discrete 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.

Mode for Discrete Frequency Distribution

But in any one (or more) of the following cases :

(i) if the maximum frequency is repeated,

(ii) if the maximum frequency occurs in the very beginning or at the end of the distribution, and

(iii) if there are irregularities in the distribution,

the value of mode is determined by the *method of grouping*, which is illustrated below by an example.

Mode for Discrete Frequency Distribution

Method of Grouping

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

Solution:

Here we see that the distribution is not regular since the frequencies are increasing steadily up to 40 and then decrease but the frequency 45 after 20 does not seem to be consistent with the distribution. Here we cannot say that since maximum frequency is 45, mode is 10. Here we shall locate mode by the method of grouping as explained below :

The frequencies in column *(i)* are the original frequencies. Column *(ii)* is obtained by combining the frequencies two by two. If we leave the first frequency and combine the remaining frequencies two by two we get column *(iii)*. Combining the frequencies two by two after leaving the first two frequencies results in a repetition of column *(ii)*. Hence, we proceed to combine the frequencies three by three, thus getting column *(iv)*. 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)*.

Solution (Continued):

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

Solution (Continued):

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.

Mode for Continuous Frequency Distribution

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

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

where 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:

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:

Here maximum frequency is 28. Thus the class 40–50 is the modal class.

the value of mode is given by

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

4. Geometric Mean (G)

Geometric mean of a set of n observations is the nth root of their product. Thus the geometric mean G, of n observations x_i, $i = 1, 2, ..., n$ is

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

The computation is facilitated by the use of logarithms. Taking logarithm of both sides, we get

$$\log G = \frac{1}{n}(\log x_1 + \log x_2 + \ldots + \log x_n) = \frac{1}{n} \sum_{i=1}^{n} \log x_i$$

$$\therefore \quad G = \text{Antilog}\left[\frac{1}{n} \sum_{i=1}^{n} \log x_i\right]$$

4. Geometric Mean (Continued)

In case of frequency distribution $x_i | f_i$, $(i = 1, 2, ..., n)$ geometric mean, G is given by

$$G = \left[x_1^{f_1} . x_2^{f_2} ..... x_n^{f_n} \right]^{\frac{1}{N}}, \text{ where } N = \sum_{i=1}^{n} f_i$$

Taking logarithms of both sides, we get

$$\log G = \frac{1}{N} (f_1 \log x_1 + f_2 \log x_2 + \ldots + f_n \log x_n)$$

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

Thus we see that logarithm of G is the arithmetic mean of the logarithms of the given values.

$$G = \text{Antilog} \left(\frac{1}{N} \sum_{i=1}^{n} f_i \log x_i \right)$$

In the case of grouped or continuous frequency distribution, x is taken to be the value corresponding to the mid-point of the class-intervals.

5. Harmonic Mean (H)

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

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

In case of frequency distribution $x_i \mid f_i$, $(i = 1, 2, ..., n)$,

$$H = \frac{1}{\frac{1}{N}\sum_{i=1}^{n}(f_i/x_i)}, \quad \left[N = \sum_{i=1}^{n} f_i\right]$$

Measures of Variablity or Measures of Dispersion

Dispersion - Variations or Scatteredness

List of Measures of Dispersion:

1. Range
2. Quartile Deviation or Semi-interquartile Range
3. Mean Deviation and
4. Standard Deviation

1. Range

Range = Maximum Value - Minimum Value

The range is the difference between two extreme observations of the distribution. If A and B are the greatest and smallest observations respectively in a distribution, then its range is $A - B$.

Range is the simplest but a crude measure of dispersion. Since it is based on two extreme observations which themselves are subject to chance fluctuations, it is not at all a reliable measure of dispersion.

2. Quartile Deviation or Semi-interquartile Range

Partition Values

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

Quartiles

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

Quartile Deviation or Semi-interquartile Range (Q)

Q is given by

$$Q = \tfrac{1}{2}(Q_3 - Q_1),$$

where Q_1 and Q_3 are the first and third quartiles of the distribution respectively.

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

Quartile Deviation (Q) = (1/2)(Q_3-Q_1).

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

Example:

Eight coins were tossed together and the number of heads resulting was noted. The operation was repeated 256 times and the frequencies (f) that were obtained for different values of x, the number of heads, are shown in the following table. Calculate median, quartiles.

x :	0	1	2	3	4	5	6	7	8
f :	1	9	26	59	72	52	29	7	1

Solution:

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 : Here $N/2 = 256/2 = 128$. Cumulative frequency (*c.f.*) just greater than 128 is 167. Thus, median = 4.

Q_1 : Here $N/4 = 64$. *c.f.* just greater than 64 is 95. Hence, $Q_1 = 3$.

Q_3 : Here $3N/4 = 192$ and *c.f.* just greater than 192 is 219. Thus $Q_3 = 5$.

Quartile Deviation (Q) = (1/2)(Q_3-Q_1).

3. Mean Deviation

If $x_i \mid f_i$, $i = 1, 2, \ldots, n$ is the frequency distribution,
then mean deviation from the average A, (usually mean, median or mode), is given by

$$\text{Mean deviation} = \frac{1}{N} \sum_i f_i \mid x_i - A \mid, \quad \Sigma f_i = N$$

where $\mid x_i - A \mid$ represents the modulus or the absolute value of the deviation $(x_i - A)$, when the –ive sign is ignored.

Since mean deviation is based on all the observations, it is a better measure of dispersion than range or quartile deviation. But the step of ignoring the signs of the deviations $(x_i - A)$ creates artificiality and renders it useless for further mathematical treatment.

It may be pointed out here that mean deviation is least when taken from median.

$$MD = \frac{1}{N} \sum_{i=1}^{n} f_i \mid x_i - A \mid, \quad N = \sum_{i=1}^{n} f_i$$

4. Standard Deviation (σ)

Standard deviation, usually denoted by the Greek letter small sigma (σ), is the positive square root of the arithmetic mean of the squares of the deviations of the given values from their arithmetic mean. For the frequency distribution $x_i \mid f_i$, $i = 1, 2, \ldots, n$,

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

Mean

where $\bar{x}$ is the arithmetic mean of the distribution and $\sum_i f_i = N$.

Variance

The square of standard deviation is called the *variance* and is given by

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

Root Mean Square Deviation

Root mean square deviation, denoted by 's' is given by

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

→ Any Number

where A is any arbitrary number. s^2 is called mean square deviation.

Difference Formulae for Calculating Variance

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

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

We know that if $d_i = x_i - A$ then $\bar{x} = A + \frac{1}{N}\sum_i f_i d_i$

$$\therefore \qquad A - \bar{x} = -\frac{1}{N}\sum_i f_i d_i$$

Hence

$$\sigma_x^2 = \frac{1}{N}\sum f_i d_i^2 + \left(-\frac{1}{N}\sum_i f_i d_i\right)^2 + 2\left(-\frac{1}{N}\sum_i f_i d_i\right)\left(\frac{1}{N}\sum_i f_i d_i\right)$$

$$= \frac{1}{N}\sum_i f_i d_i^2 - \left(\frac{1}{N}\sum_i f_i d_i\right)^2$$

$$\Rightarrow \qquad \sigma_x^2 = \sigma_d^2$$

Hence variance and consequently standard deviation is independent of change of origin.

Difference Formulae for Calculating Variance

If we take $d_i = (x_i - A)/h$ so that $(x_i - A) = hd_i$, then

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

$$= \frac{1}{N}\sum_i f_i (hd_i + A - \bar{x})^2$$

$$= h^2 \frac{1}{N}\sum_i f_i d_i^2 + (A - \bar{x})^2 + 2(A - \bar{x}) \cdot h \cdot \frac{1}{N}\sum_i f_i d_i$$

Using $\bar{x} = A + h\,\dfrac{\sum f_i d_i}{N}$, we get

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

$\sigma_x^2 = h^2 \sigma_d^2$

which shows that *variance is not independent of change of scale.*

Hence variance is independent of change of origin but not of scale.

Co-efficient of Dispersion

Whenever we want to compare the variability of the two series which differ widely in their averages or which are measured in different units, we do not merely calculate the measures of dispersion but we calculate the co-efficients of dispersion which are pure numbers independent of the units of measurement. The co-efficients of dispersion (C.D.) based on different measures of dispersion are as follows :

1. C.D. based upon range $= \frac{A - B}{A + B}$, where A and B are the greatest and the smallest items in the series.

2. Based upon quartile deviation :

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

3. Based upon mean deviation :

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

4. Based upon standard deviation :

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

Co-efficient of Variation

100 times the co-efficient of desperison based upon standard deviation is called co-efficient of variation (C.V.),

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

According to Professor Karl Pearson who suggested this measure, *C.V. is the percentage variation in the mean, standard deviation being considered as the total variation in the mean.*

For comparing the variability of two series, we calculate the co-efficient of variations for each series. The series having greater C.V. is said to be more variable than the other and the series having lesser C.V. is said to be more consistent (or homogenous) than the other.

Example:

Calculate the mean and standard deviation for the following table giving the age distribution of 542 members.

Age in years :	20—30	30—40	40—50	50—60	60—70	70—80	80—90
No. of members :	3	61	132	153	140	51	2

Solution:

Also find CD & CV.

Here we take $d = \frac{x - A}{h} = \frac{x - 55}{10}$

Age group	*Mid-value* (x)	*Frequncy* (f)	$d = \frac{x - 55}{10}$	fd	fd^2
20 — 30	25	3	–3	–9	27
30 — 40	35	61	–2	–122	244
40 — 50	45	132	–1	–132	132
50 — 60	55	153	0	0	0
60 — 70	65	140	1	140	140
70 — 80	75	51	2	102	204
80 — 90	85	2	3	6	18
		$N = \Sigma f = 542$		$\Sigma fd = -15$	$\Sigma fd^2 = 765$

$$\bar{x} = A + h\frac{\Sigma fd}{N} = 55 + \frac{10 \times (-15)}{542} = 55 - 0{\cdot}28 = 54{\cdot}72 \text{ years.}$$

$$\sigma^2 = h^2\left[\frac{1}{N}\Sigma fd^2 - \left(\frac{1}{N}\Sigma fd\right)^2\right] = 100\left[\frac{765}{542} - (0{\cdot}28)^2\right]$$

$$= 100 \times 1{\cdot}333 = 133{\cdot}3$$

σ (standard deviation) = 11·55 years

$CD = \frac{\sigma}{\bar{x}}$

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

Moments

The rth moment of a variable x about any point $x = A$, usually denoted by μ_r' is given by

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

$$= \frac{1}{N} \sum_i f_i d_i^r,$$

where $d_i = x_i - A$.

$$\mu_r' = \frac{1}{N} \sum_{i=1}^{n} f_i (x_i - A)^r$$

$$A = \bar{x}$$

The rth moment of a variable about the mean $\bar{x}$, usually denoted by μ_r is given by

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

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

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

Particular Cases

In particular

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

and $\mu_1 = \frac{1}{N} \sum_i f_i (x_i - \bar{x}) = 0$, being the algebraic sum of deviations from the mean. Also

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

These results, viz., $\mu_0 = 1$, $\mu_1 = 0$, and $\mu_2 = \sigma^2$, are of fundamental importance and should be committed to memory.

Pearson's Co-efficients

Karl Pearson defined the following

four coefficients, based upon the first four moments about mean :

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}, \quad \gamma_1 = +\sqrt{\beta_1} \quad \text{and} \quad \beta_2 = \frac{\mu_4}{\mu_2^2}, \quad \gamma_2 = \beta_2 - 3$$

Skewness

Literally, skewness means '*lack of symmetry*'. We study skewness to have an idea about the shape of the curve which we can draw with the help of the given data. A distribution is said to be skewed if

(*i*) Mean, median and mode fall at different points,

i.e., Mean ≠ Median ≠ Mode,

(*ii*) Quartiles are not equidistant from median, and

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

Measures of Skewness

Measures of Skewness. Various measures of skewness are

(1) $S_k = M - M_d$ (2) $S_k = M - M_0$.

where M is the mean, M_d, the median and M_0, the mode of the distribution.

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

These are the absolute measures of skewness. As in dispersion, for comparing two series we do not calculate these absolute measures but we calculate the relative measures called the *co-efficients of skewness* which are pure numbers independent of units of measurement.

Co-efficients of Skewness

1. *Prof. Karl Pearson's Coefficient of Skewness.*

$$S_k = \frac{(M - M_0)}{\sigma}$$

II. *Prof. Bowley's Coefficient of Skewness.* Based on quartiles,

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

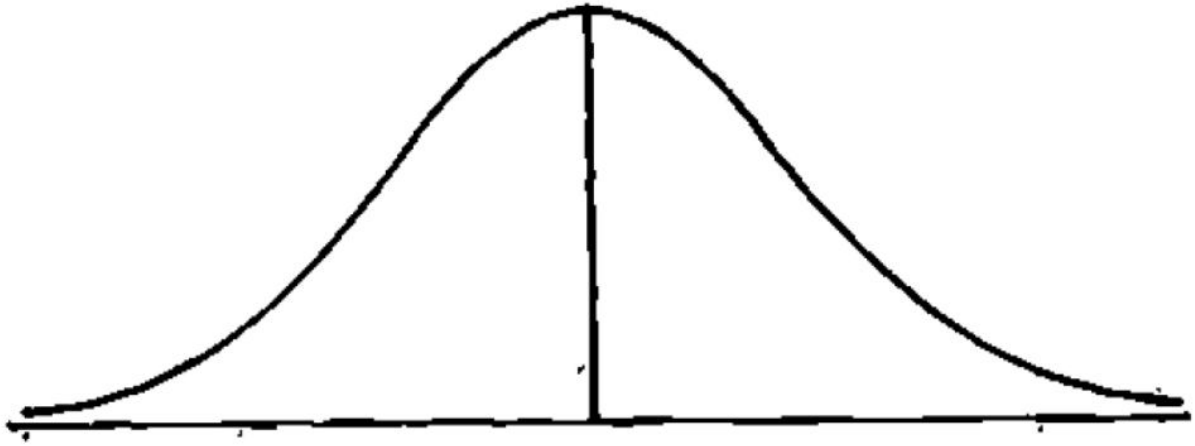
III. Based upon moments, co-efficient of skewness is

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

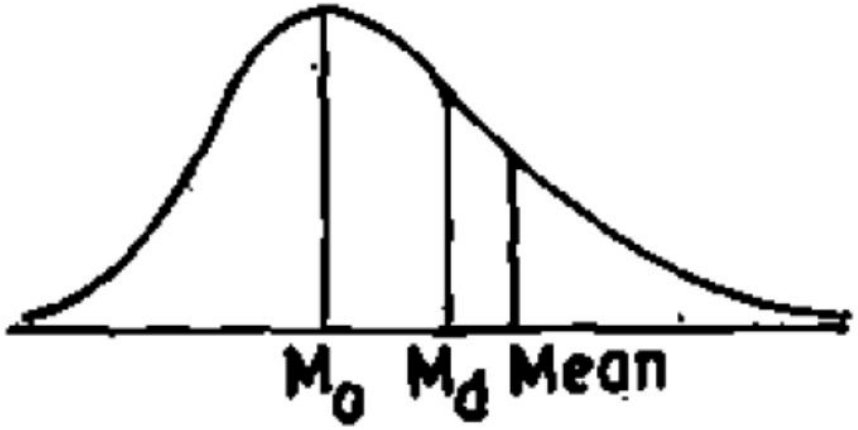
where symbols have their usual meaning. Thus $S_k = 0$ if either $\beta_1 = 0$ or $\beta_2 = -3$. But since $\beta_2 = \mu_4/\mu_2^2$, cannot be negative, $S_k = 0$ if and only if $\beta_1 = 0$. Thus for a symmetrical distribution $\beta_1 = 0$. In this respect β_1 is taken to be a measure of skewness.

The skewness is positive if the larger tail of the distribution lies towards the higher values of the variate (the right), i.e., if the curve drawn with the help of the given data is tretched more to the right than to the left and is negative

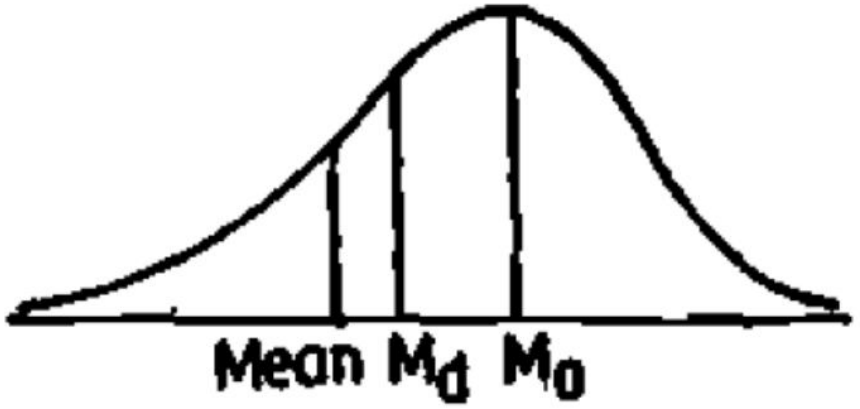
Graphical Representation of Skewness



$\bar{x}$ (*Mean*) = M_0 = M_d
(Symmetrical Distribution)



(Positively Skewed Distribution)



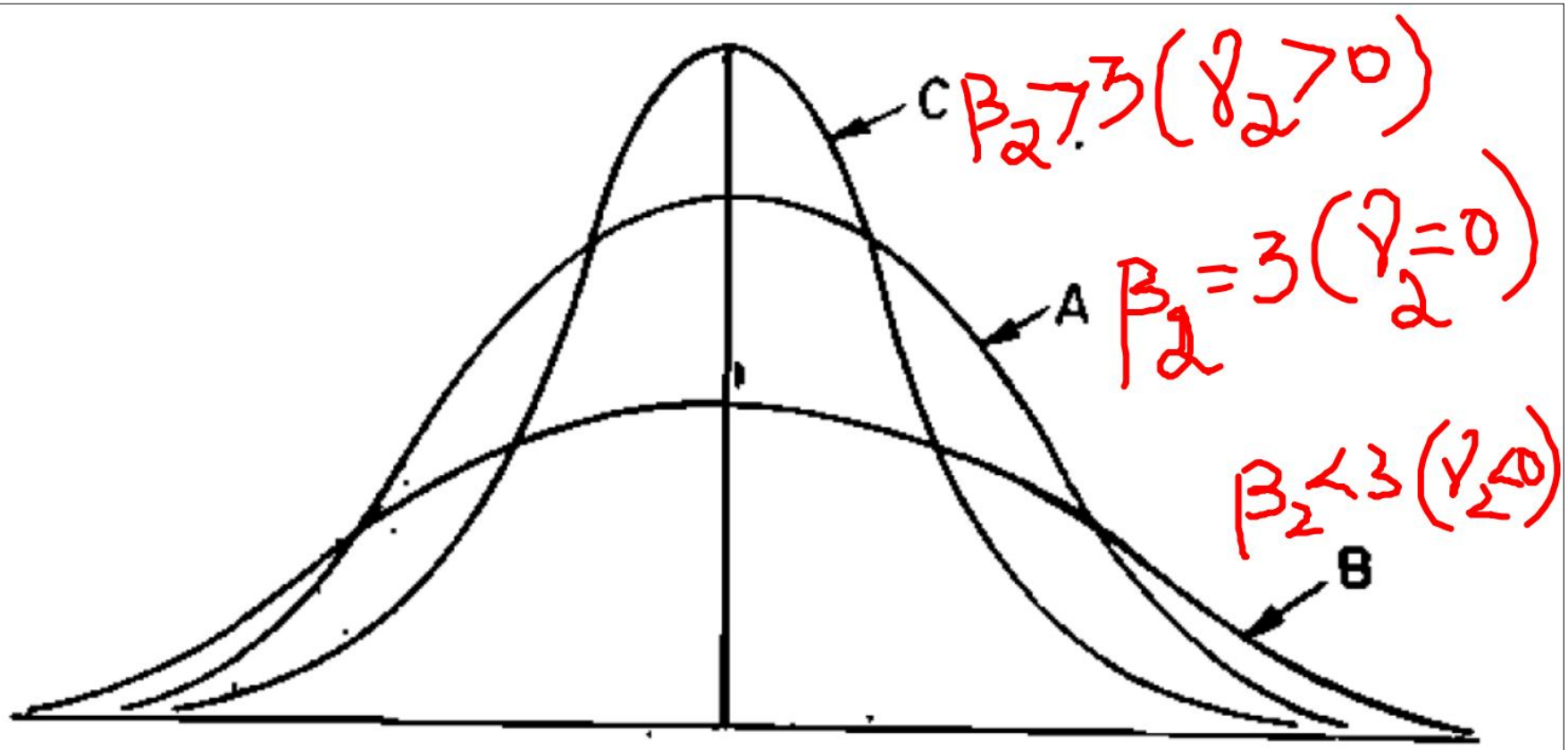
(Negatively Skewed Distribution)

Kurtosis

If we know the measures of central tendency, dispersion and skewness, we still cannot form a complete idea about the distribution as will be clear from the following figure in which all the three curves A, B and C are symmetrical about the mean 'm' and have the same range.

In addition to these measures we should know one more measure which Prof. Karl Pearson calls as the '*Convexity of curve*' or *Kurtosis*. Kurtosis enables us to have an idea about the flatness or peakedness of the curve. It is measured by the co-efficient β_2 or its derivation r_2 given by

$$\beta_2 = \mu_4/\mu_2^2, \quad \gamma_2 = \beta_2 - 3$$



Curve of the type '*A*' which is neither flat nor peaked is called the *normal curve or mesokurtic curve* and for such a curve $\beta_2 = 3$, i.e., $\gamma_2 = 0$. Curve of the type '*B*' which is flatter than the normal curve is known as *platykurtic* and for such a curve $\beta_2 < 3$, *i.e.*, $\gamma_2 < 0$. Curve of the type '*C*' which is more peaked than the normal curve is called *leptokurtic* and for such a curve $\beta_2 > 3$, i.e., $\gamma_2 > 0$.