30 MAT2001 - Statistics for Engineers - SST 508

Module-I

Introduction to Statistics

Introduction to Statistics & Data Analytics - Measures of Central tendency - Measures of Variability - Moments - Skewness - Kurtosis

Measures of Central Tendency (or) Averages

Five measures of central tendency are given below,

i) Arithmetic mean (or) Simple mean
ii) Median
iii) Mode
iv) Geometric mean
v) Harmonic mean

Central Tendency is a central value for a probability distribution. Called Average or just center of distribution. MCT is an average.

i) Arithmetic mean

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 observations $x_1, x_2 \dots x_n$ is given by

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

In case of frequency distribution $x_i \mid f_i$ for $i=1,2\cdots n$ where f_i is the frequency of x_i,

$$\bar{x} = \frac{f_1x_1+f_2x_2+\cdots+f_nx_n}{f_1+f_2+\cdots+f_n} = \frac{\sum_{i=1}^{n} f_ix_i}{\sum_{i=1}^{n} f_i}$$

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

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

Sums

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

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
Total	73	299

Here $N = \Sigma f = 73$

$\Sigma fx = 299$

$$\bar{x} = \frac{1}{N}\Sigma fx$$

$$= \frac{299}{73} = 4.09$$

Note

* Data is often described as Ungrouped (or) Grouped.
* Ungrouped data is data given as individual data points

For example:

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/r of Tv sets	Frequency
0	2
1	13
2	18
3	0
4	10
5	2
Total	45

* Grouped data is data given in intervals

For Example,

Exam Score	Frequency
90-99	7
80-89	5
70-79	15
60-69 (59.5 - 69.5)	4
50-59 (49.5 - 59.5)	5
40-49 (39.5 - 49.5)	0
30-39 (29.5 - 39.5)	1
Total	37

$l/2$

$(L - l/2, U + l/2)$

Sums

(2) Find the arithmetic mean of the following data :

1, 3, 5, 7, 8

Sol

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{24}{5} = 4.8$$

(3) 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

Sol

Marks	No. of Students (f)	Midpoint (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		2800

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

$= 28$

Mean of the Composite Series

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

$$= \sum_i n_i \bar{x}_i \Big/ \sum_i n_i$$

(1) The average Salary of male employees in a firm was Rs. 5200 and that of females was Rs. 4,200. The mean Salary of all the employees was Rs. 5000. Find the percentage of male and female employees.

Sol:

$n_1 \rightarrow$ No/- of male employees

$n_2 \rightarrow$ No/- of female employees

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

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

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

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

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

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

$$\Rightarrow 20n_1 = 80n_2$$

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

The percentage of male employees in the firm

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

& The percentage of female employees in the firm $= \frac{1}{4+1} \times 100 = 20\%$

Median

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

* Median is a Positional Average
* In case of ungrouped data, if the no/- of observations is odd ~~There are two middle~~ then median is the middle value after the values have been arranged in ascending (or) descending order of magnitude.

* In case of even no/. of observations, there are two middle terms and median is obtained by taking the arithmetic mean of the middle terms.

Example

(1) Find Median of 25, 20, 15, 35, 18

Sol

Ascending order:

15, 18, 20, 25, 35.

∴ Middle most value : 20 = Median.

(2) Find the median of 8, 20, 50, 25, 15, 30

Sol

Ascending order:

8, 15, 20, 25, 30, 50

Median = $\frac{1}{2}(20+25) = 22.5$

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

i) Find $\frac{1}{2}N$, where $N = \Sigma f_i$

ii) See the (less than) Cumulative frequency (cf) Just greater than $\frac{1}{2}N$.

iii) The corresponding value of x is median

Example

① 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

Sol

x	f	cf
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
Total	$N=120$	

Here $N = 120$

$N/2 = 60$

∴ The cumulative frequency (cf) just greater than $\frac{1}{2}N$ is 65 & the value of x corresponding to 65 is 5.

∴ 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 as the median class & the value of median is obtained by

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

Where $l \rightarrow$ lower limit of the median class

$f \rightarrow$ frequency of the median class

$h \rightarrow$ the magnitude of the median class

$c \rightarrow$ c.f of the class preceding the median class.

$$N = \sum f$$

Example

Find the median wage of the following distribution:

Wages (in Rs) :	2000-3000	3000-4000	4000-5000
No. of workers :	3	5	20

Wages (in Rs) :	5000-6000	6000-7000
No. of workers :	10	5

Sol

Wages (in Rs.)	No. of employees (f)	cf
2000 – 3000	3	3
3000 – 4000	5	8
4000 – 5000	20	28
5000 – 6000	10	38
6000 – 7000	5	43
Total	$N = \Sigma f = 43$	

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

Cumulative frequency just greater than 21.15 is 28 & the corresponding class is 4000 – 5000.

i.e., Median class : 4000 – 5000

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

$l = 4000$, $f = 20$, $h = 1000$, $\frac{N}{2} = 21.5$

$c = 8$

$\therefore$ Median $= 4000 + \frac{1000}{20}(21.5 - 8)$

$= 4000 + 675$

$= 4675$

(2) In a factory employing 3000 persons in a day 5 percent work less than 3 hours 58 work from 3.01 to 4.50 hrs, 30 percent work from

4.51 to 6.00 hrs, 500 work from 6.01 to 7.50 hours, 20 percent work from 7.51 to 9.00 hrs and the rest work 9.01 or more hours. What is the median hours of work?

Sol:

Work hours	No. of. employees (f)	Cf	Class boundaries
Less than 3	$\frac{5}{100} \times 3000 = 150$	150	Below 3.005
3.01 – 4.50	580	730	3.005 – 4.505
4.51 – 6.00	$\frac{30}{100} \times 3000 = 900$	1630	4.505 – 6.005
6.01 – 7.50	500	2130	6.005 – 7.505
7.51 – 9.00	$\frac{20}{100} \times 3000 = 600$	2730	7.505 – 9.005
9.01 & above	3000 – 2730 = 270	3000	9.005 and above

Here $N = \Sigma f = 3000$ & $N/2 = 1500$

* The cf just greater than 1500 is 1630.
& The corresponding class is 4.51 – 6.00 whose boundaries are 4.505 – 6.005 is the median class

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

$$= 4.505 + \frac{1.5}{900}\left(1500 - 730\right)$$

$$= 4.505 + 1.283 \simeq 5.79 \text{ (Median hours of work)}$$

$l = 4.505$
$h = 1.5$, $f = 900$
$N/2 = 1500$,
$c = 730$

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 frequency distribution:

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

Example

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

Maximum frequency is 25 & Corresponding x is 4

$\therefore$ Mode is 4.

Note In some cases the above method is not possible.

* If the maximum frequency is repeated.
* If the maximum frequency occurs in the very beginning or at the end of the distribution
* If there are irregularities in the distribution

then the value of mode is determined by the method of grouping.

For 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

Sol:

Here given distribution is not regular since the frequencies are increasing steadily upto 40 & then decrease but the frequency 45 after 20 does not seem to be consistent with the distribution.

So we cannot say that maximum frequency is 45 & mode is 10.

∴ we will apply grouping method

Size (x)	Frequency (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	80	93	79
6	40					
7	32	60	48	65		
8	28					
9	20	65	59			
10	45					
11	14	20				
12	6					

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

By observing Column(3) in analysis table, 6 occurs maximum no. of times.

∴ Mode is 6

In case of continuous frequency distribution:

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

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

Where $l \rightarrow$ lower limit of CI

$h \rightarrow$ Magnitude

$f_1 \rightarrow$ frequency of the model class,

f_0 & $f_2 \rightarrow$ frequencies of the classes preceding & succeeding of the model 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

Sol

Here maximum frequency is 28.

Corresponding class is 40-50 (Modal class)

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

$l = 40$, $h = 10$, $f_1 = 28$, $f_0 = 12$

$f_2 = 20$

$$\text{Mode} = 40 + \frac{10(28-12)}{2\times 28 - 12 - 20} = 40 + 6.666$$

$$= 46.67 \text{ (app)}$$

Note

* For a symmetrical distribution, mean, median & mode coincide.

* If the distribution is moderately asymmetrical the mean, median & mode obey the following empirical relationship

$$\text{Mean} = 3\,\text{median} - 2\,\text{mode}.$$

Geometric mean

Geometric mean of a set of n observations is the nth root of their product.

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

Note

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

* In case of frequency distribution x_i / f_i for $i = 1, 2 \ldots n$

$$G = \left(x_1^{f_1} \cdot x_2^{f_2} \cdots x_n^{f_n}\right)^{1/N}$$

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

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

* In case of grouped (or) continuous frequency distribution,

x is taken to be the value corresponding to the midpoint of the class intervals.

Harmonic Mean

$$H = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}(1/x_i)} \quad \text{for } i = 1, 2 \ldots n.$$

* For frequency distribution $x_i | f_i$

for $i = 1, 2, 3 \cdots n$.

$$H = \frac{1}{\frac{1}{N}\sum_{i=1}^{n}(f_i/x_i)} \quad \text{Where } N = \sum_{i=1}^{n} f_i$$

Measures of Dispersion (or) Variability:

Relative measure — Absolute measure

Co-eff of range * Range

Co-eff of Quartile * Quartile deviation

Co-eff. . . * Mean deviation

Co-efficient * Standard deviation & Root mean Square deviation of variance.

Variability refers how spread out a group of scores is. a_1: 5-9, a_2: 4-10

Variance $= \frac{\Sigma(x-M)^2}{N} = \sigma^2$

* Range

If A & B are the greatest & smallest observation respectively in a distribution, then its range is given by Range $= A - B$

(Mid value) highest class $U - L$

M.V. U - Upper boundary, L - Lower boundary, lowest class

Note: It is not a reliable measure of dispersion

Co. eff. of range $= \frac{U-L}{U+L}$

* Quartile deviation

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

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

$$\text{Co-efficient q.d} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Note: 384, 391, 407, 522, 591, 622, 733, 777, 1480, 2489

* Quartile deviation is definitely a better measure than the range as it makes use of 50% of the data & it ignores the remaining 50% of data.

* It is not a reliable measure.

* Mean deviation:

Individual series : $MD = \frac{\sum |x - \bar{x}|}{n}$

Discrete series $MD = \frac{\sum f|x - \bar{x}|}{N}$, $\bar{x} = A + \frac{\sum fd}{N}$, $d = x - A$, $N = \sum f$

If $x_i | f_i$ for $i = 1, 2 \ldots n$ is the frequency distribution then mean deviation from the Average (Mean (or) Median (or) mode) is given by

Continuous series MD: $m \rightarrow$ mid value, $d = m - A$, $\bar{x} = A + \frac{\sum fd}{N} \times c$

$$A = \frac{1}{N} \sum_{i=1}^{n} f_i |x_i - A| \quad \text{Where } N = \sum_{i=1}^{n} f_i$$

& $|x_i - A| \rightarrow$ Modulus (or) absolute value of the deviation $(x_i - A)$

Note

* It includes all the data for measure. So it is better measure of dispersion than range (or) quartile deviation.

Co-efficient of mean deviation $= \frac{\text{Mean deviation}}{\text{Mean or median or mode}}$

Example

① Calculate: (i) Quartile deviation (Q.D) & (ii) Mean deviation (M.D.) from mean for the following data:

$Q_1 = 2.75^{th} \Rightarrow 2nd\ value + 0.75(3rd - 2nd)$

Marks:	0-10	10-20	20-30	30-40	40-50	50-60	60-70
No/- of Students:	6	5	8	15	07	6	3

Calculation for Q.D & M.D. from mean:

Marks	Midvalue of Marks (x) m	No/- of Students f	cf	$d = \frac{x-35}{10}$	fd	d^2	fd^2	$\lvert x-\bar{x}\rvert$ (m) $\lvert x-33.4\rvert$	$f\lvert x-\bar{x}\rvert$ (m)	Less than Cf
0-10	5	6	6	-3	-18	9	54	28.4	170.4	6
10-20	15	5	11	-2	-10	4	20	18.4	92.0	11
20-30	25	8	19	-1	-8	1	8	8.4	67.2	19
30-40	35	15	34	0	0	0	0	1.6	24.0	34
40-50	45	7	41	1	7	1	7	11.6	81.2	41
50-60	55	6	47	2	12	4	24	21.6	129.6	47
60-70	65	3	50	3	9	9	27	31.6	94.8	50
Total		$N = 50$			-8		142		659.2	

1) Here $N = 50$, $N/4 = 12.75$, $3/4\,N = 37.25$

* The Cf is just greater than $N/4 = 12.75$ is 19.
* Corresponding class 20-30 contains Q_1,

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

$l \rightarrow$ Lower limit of the class
$h \rightarrow$ width of the CI
$f \rightarrow$ Corresponing frequency
$c \rightarrow$ Cf. of the preceding class

(11)

$l = 20, h = 10, f = 8, N/4 = 12.75, c = 11$

$$Q_1 = 20 + \left(\frac{10}{8}\right)(12.75 - 11) = 22.19$$

* The cf just greater than $3N/4 = 37.25$ is 41.
* The corresponding class 40-50 contains Q_3.

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

$l = 40, h = 10, f = 7, 3N/4 = 37.25, c = 34$

$$Q_3 = 40 + 10/7(37.25 - 34) = 44.64$$

Hence $Q.D = 1/2(Q_3 - Q_1) = 1/2(44.64 - 22.9)$

$= 11.23$

(c)

ii) Mean $\bar{x} = A + h\frac{\Sigma fd}{N}$

$$= 35 + \frac{10 \times -8}{50} = 33.4$$

M.d from mean $= 1/N \Sigma f |x - \bar{x}|$

$$= \frac{659.2}{50} = 13.184$$

Question Calculate mean deviation from median

Coefficient of M.D

Karl pearson (1893)

Standard deviation & Root mean Square deviation

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

* Square of the S.D is called the variance which is given by

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

where $\bar{x} \rightarrow$ Arithmetic mean

$$N = \sum_i f_i$$

* Root mean square deviation:

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

$A \rightarrow$ arbitrary number

$S^2 \rightarrow$ mean square deviation.

Example:

Calculate the mean & 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

Sol

$$\text{let } A = 55 \text{ \& } d = \frac{x - A}{h} = \frac{x - 55}{10}$$

Age Group	Mid value x (m)	Frequency f	$d = \frac{x-55}{10}$ (m)	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
Total		$\Sigma f = N = 542$		-15	765

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

$$= 55 - 0.28$$

$$= 54.72 \text{ years}$$

i) Variance $= \sigma^2 = h^2 \left[\frac{1}{N} \Sigma fd^2 - \left(\frac{1}{N} \Sigma fd \right)^2 \right]$ (c^2)

$$= 100 \left[\frac{765}{542} - (0.028)^2 \right] = 100 \times 1.4107$$

$$= 141.07$$

ii) Standard deviation:

$$S.D = \sqrt{141.07} = 11.88 \text{ years}$$

Note

Coefficient of Variation

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

Moments

* r^{th} moment of a variable x about any point $x = A$, usually denoted by μ_r' is it is given by

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

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

* r^{th} moment of a variable x about mean $\bar{x}$, usually denoted by μ_r is given by,

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

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

Note

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

$$\& \quad \mu_1 = \frac{1}{N}\sum_i f(x_i - \bar{x}) = 0$$

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

Relation between Moments about mean in terms of moments about any point & vice versa.

$$\mu_2 = \mu_2' - (\mu_1')^2$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4$$

Pearson's β & γ Co-efficients:

Karl Pearson defined the following four Co efficients based upon the first four moments about mean.

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}, \quad \gamma_1 = +\sqrt{\beta_1}$$

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

Example

① The first four moments of a distribution about the value 4 of the variable are -1.5, 17, -30 & 108. Find the moments about mean, β_1 & β_2. Find also the moments about i) The origin & ii) The point $x=2$.

Solution:

Given $A=4$, $M_1' = -1.5$, $M_2' = 17$, $M_3' = -30$ & $M_4' = 108$

* Moments about mean

$$M_2 = M_2' - M_1'^2 = 17 - (-1.5)^2 = 17 - 2.25 = 14.75$$

$$M_3 = M_3' - 3M_2'M_1' + 2M_1'^3$$

$$= 30 - 3 \times (17) \times (-1.5) + 2(-1.5)^3 = 39.75$$

$$M_4 = M_4' - 4M_3'M_1' + 6M_2'M_1'^2 - 3M_1'^4$$

$$= 108 - 4 \times 30 \times 1.5 + 6(17)(-1.5)^2 - 3(-1.5)^4$$

$$= 142.3125$$

Hence, $\beta_1 = \frac{M_3^2}{M_2^3} = \frac{(39.75)^2}{(14.75)^3} = 0.4926$

$$\beta_2 = \frac{M_4}{M_2^2} = \frac{142.2125}{(14.75)^2} = 0.6543$$

Also mean $= \bar{x} = A + M_1' = A + (-1.5) = 2.5$

* Moments about the origin:

$\bar{x} = 2.5$, $M_2 = 14.75$, $M_3 = 39.75$,

$M_4 = 142.31$ (app)

We know that, $\bar{x} = A + M_1'$, where M_1' is the first moment about the point $x = A$

Let $A = 0$, we have M_1' = mean $= \bar{x} = 2.5$

= first moment about origin.

M_2' = 2nd moment about origin

$= M_2 + M_1'^2 = 14.75 + (2.5)^2$

$= 14.75 + 6.25 = 21$

$M_3' = M_3 + 3M_2 M_1' + M_1'^3 = 39.75 + 3(14.75)(2.5) + (2.5)^3 = 166$

$M_4' = M_4 + 4M_3 M_1' + 6M_2 M_1'^2 + (M_1')^4$

$= 142.3125 + 4(39.75)(2.5) + 6(14.75)(2.5)^2 + (2.5)^4 = 1132.$

* Moment about the point $x = 2$.

W.k.t, $\bar{x} = A + M_1'$, Put $A = 2$

M_1' = Ist moment about $A = 2$.

$= \bar{x} - A = \bar{x} - 2 = 2.5 - 2 = 0.5$

$\therefore M_2' = M_2 + M_1'^2 = 14.75 + 0.25 = 15$

$M_3' = M_3 + 3M_2M_1' + M_1'^3$

$= 39.75 + 3(14.75)(0.5) + 0.5)^3 = 62$

$M_4' = M_4 + 4M_3M_1' + 6M_2M_1'^2 + M_1'^4$

$= 142.3125 + 4(39.75)(0.5) + 6(14.75)(0.5)^2 + (0.5)^4 = 244.$

(2) Calculate the first four Moments of the following distribution about the mean & hence find β_1 & β_2

x	0	1	2	3	4	5	6	7	8
f	1	8	28	56	70	56	28	8	1

Sol

x	f	$d=x-4$	fd	fd^2	fd^3	fd^4
0	1	-4	-4	16	-64	256
1	8	-3	-24	72	-216	648
2	28	-2	-56	112	-224	448
3	56	-1	-56	56	-56	56
4	70	0	00	0	0	0
5	56	1	56	56	56	56
6	28	2	56	112	224	448
7	8	3	24	72	216	648
8	1	4	4	16	64	256
Total	256	0	0	512	0	2816

Moments about the point $x=4$ are

$$\mu_1' = \frac{1}{N}\Sigma fd = 0$$

$$\mu_2' = \frac{1}{N}\Sigma fd^2 = \frac{512}{256} = 2$$

$$\mu_3' = \frac{1}{N}\Sigma fd^3 = 0$$

$$\mu_4' = \frac{1}{N}\Sigma fd^4 = \frac{2816}{256} = 11$$

Moment about the mean are

$$\mu_1 = 0 \quad , \mu_2 = \mu_2' - \mu_1'^2 = 2$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 = 0$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4 = 11$$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0 \quad , \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{11}{4} = 2.75$$

Skewness

* Skewness '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 diven data
* A distribution is said to be skewed if
  i) Mean, median & mode fall at different points.
     i.e., Mean ≠ Median ≠ 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 oneside than to the other.

Measures of Skewness

Various measures of Skewness are,

i) $S_k = M - M_d$

ii) $S_k = M - M_o$

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

$= Q_3 - 2M_d + Q_1$

Where $M \rightarrow$ Mean

$M_d \rightarrow$ Median

$M_0 \rightarrow$ Mode

$Q_1 \rightarrow$ First Quartile & $Q_3 \rightarrow 3^{rd}$ Quartile

* 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 coefficient of skewness which are pure numbers independent of units of measurement.

Coefficient of Skewness

i) Karl Pearson's coefficient of Skewness:

$$S_k = \frac{M - M_0}{\sigma}$$ where $\sigma \rightarrow$ standard deviation of the distribution

If mode is ill-defined, then using the empirical relation

$$M_0 = 3M_d - 2M$$

we get, $$S_k = \frac{3(M - M_d)}{\sigma}$$

Note: ① If $M = M_0 = M_d$ then $S_k = 0$

(2) For a symmetrical distribution, mean, median & mode coincide

(3) Skewness is +ve if $M > M_o$ (or) $M > M_d$ & -ve if $M < M_o$ (or) $M < M_d$.

(4) Limits for Karl-pearson's coefficient of Skewness :

$$|S_k| \leq 3$$

i.e., $-3 \leq S_k \leq 3$

ii) Bowley's Coefficient of Skewness

Bowley's coefficient of Skewness based on Quartiles.

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

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

Note:

(1) It is based on Quartiles & Median

② when the mode is ill-defined and extreme observations are present in the data.

③ when the distribution has open end Classes (or) Unequal class intervals.

In these situations pearson's coefficient of Skewness cannot be used.

④ $S_k = 0$ if $Q_3 - M_d = M_d - Q_1$

$\Rightarrow$ For a Symmetrical distribution ($S_k = 0$) median is equidistant from the upper & lower Quartiles

i) Skewness is +ve if $Q_3 + Q_1 > 2M_d$

ii) Skewness is -ve if $Q_3 + Q_1 < 2M_d$

⑤ limits of Bowley's Coefficient of Skewness

$$-1 \leq S_k \leq 1$$

iii) Based upon moments, Coefficient of Skewness is

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

Note

* If $\beta_1 = 0$ (or) $\beta_2 = -3$ then $S_K = 0$

Since $\beta_2 = \frac{\mu_4}{\mu_2^2}$, Cannot be negative

$\therefore$ $S_K = 0$ if and only if $\beta_1 = 0$

Kurtosis (Convexity of the frequency curve)

Measures of central tendancy, dispersion & Skewness cannot form a complete idea about the distribution.

So, in addition to that we should know one more measure "Kurtosis" which will enable us to have an idea about the "flatness (or) peakness" of the frequency curve.

* It is measured by the coefficient of $\beta_2 = \frac{\mu_4}{\mu_2'^2}$ (or) its derivation $\gamma_2 = \beta_2 - 3$.

Types of curves:

* If $\beta_2 = 3$ (or) $\gamma_2 = 0$ then the curve is said to be normal curve (or) mesokurtic curve.

* If $\beta_2 < 3$ (or) $\gamma_2 < 0$ then the normal curve is known as platykurtic curve.

* If $\beta_2 > 3$ (or) $\gamma_2 > 0$ then the normal curve is called as leptokurtic curve

Example:

For a distribution, the mean is 10, variance is 16, γ_1 is +1 & β_2 is 4. Obtain the first four moments about the origin. (i.e., zero) Comment upon the nature of the distribution.

Sol

Given, mean $= M_1' =$ First moment about origin

$= 10$

Variance $M_2 = M_2' - M_1'^2 = 16 = \sigma^2$

$\therefore \sigma = 4$, $\gamma_1 = 1$

& $\beta_1 = 4$

To find: first 4 moments about the origin

$$\mu_1' = 10$$

$$\mu_2 = \mu_2' - \mu_1'^2$$

$$\therefore \mu_2' = \mu_2 + \mu_1'^2 = 16 + 10^2 = 116$$

we have

$$\gamma_1 = 1$$

$$\frac{\mu_3}{\mu_2^{3/2}} = 1$$

$$\frac{\mu_3}{\sigma^3} = 1$$

$$\mu_3 = \sigma^3 = 4^3 = 64 \quad (\because \mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3)$$

$$\Rightarrow \mu_3' = \mu_3 + 3\mu_2'\mu_1' - 2\mu_1'^3$$

$$= 64 \times 3 + 3 \times 116 \times 10 - 2 \times 1000 = 1544$$

Now $$\beta_2 = \frac{\mu_4}{\mu_2^2} = 4$$

$$\Rightarrow \mu_4 = 4 \times 16^2 = 1024$$

$$\& \quad \mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4$$

$$\therefore M_4 = 1024 + 4 \times 1544 \times 10 - 6 \times 116 \times 100 + 3 \times 10000$$

$$= 23184$$

Comment on the Nature of the distribution:

* Since $\gamma_1 = 1$

$\therefore$ The distribution is moderately +vely Skewed.

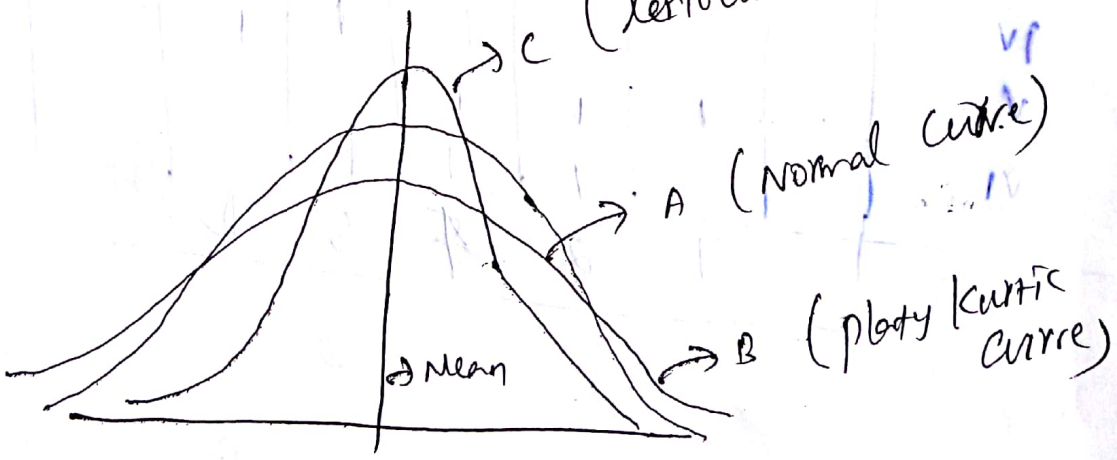
* Further, Since $\beta_2 = 4 > 3$

$\therefore$ The distribution is leptokurtic curve.

i.e., It will be slightly more peaked than the normal Curve.

— × — × —

Note:



Calculate Mode

Class Interval	0–5	5–10	10–15	15–20	20–25	25–30	30–35	35–40
Fre	9	12	15	16	17	15	10	13

CI	$f_{(v)}$	(ii)	(iii) one	(iv) three	V one	VI two
0–5	9	21		36		
5–10	12				43	
10–15	15	31				48
15–20	16		33	48		
20–25	17	32			42	
25–30	15		25			38
30–35	10	23				
35–40	13					

↑ modal class

Analysis table

Column	0–5	5–10	10–15	15–20	20–25	25–30	30–35	35–40
i					1			
ii				1	1	1		
iii				1	1			
iv				1	1	1		
v		1	1	1				
vi			1	1	1			
		1	2	4	5	2		

↓

Standard deviation

Coefficient of S.D

① Individual series

(a) $\sigma = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n}}$ — $\bar{x}$ – Actual mean. deviation from

(b) $\sigma = \sqrt{\frac{\Sigma d^2}{n} - \left(\frac{\Sigma d}{n}\right)^2}$ $d = x - A$ deviation from assumed mean

② Discrete series

(a) $\sigma = \sqrt{\frac{\Sigma f d^2}{\Sigma f}}$, $d = x - \bar{x}$ Actual mean

(b) $\sigma = \sqrt{\frac{\Sigma f d^2}{\Sigma f} - \left(\frac{\Sigma f d}{\Sigma f}\right)^2}$ $d = x - A$, $N = \Sigma f$

③ Step-deviation

If the variables ~~and~~ values are equal intervals, then.

$$\sigma = \sqrt{\frac{\Sigma f d'^2}{N} - \left(\frac{\Sigma f d'}{N}\right)^2} \times c$$

$d' = \frac{x - A}{c}$, c is the interval between each value.

A – center value of the series or assumed mean A

$$\sigma^2 = 100\left(\frac{1}{50}^{\frac{1}{2}} - \left(\frac{-8}{50}\right)^2\right) \quad \frac{N}{50}\ \frac{}{25}$$

$$= 100\left(2.84 - \left(\frac{64}{2500}\right)\right)$$

$$-0.0256$$

$$= 100 \times 2.8144 = 281.44$$

2. Absolute measure — amount of variation in a set of values in terms of unit of observation.

Continuous series

$$\sigma = \sqrt{\frac{\Sigma f d'^2}{N} - \left(\frac{\Sigma f d'}{N}\right)^2} \times c$$

$d' = \frac{m - A}{c}$, c – class interval

m – mid value

A – assume as center value as an assumed mean

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

Combined S.D

series $x_1 \longrightarrow n_1 \quad \bar{x}_1, \sigma_1$

$x_2 \longrightarrow n_2 \quad \bar{x}_2, \sigma_2$

$$\bar{X}_{12} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$$

$$\sigma_{12} = \sqrt{\frac{n_1 \sigma_1^2 + n_2 \sigma_2^2 + n_1 d_1^2 + n_2 d_2^2}{n_1 + n_2}}$$

$d_1 = \bar{x}_{12} - \bar{x}_1 \qquad d_2 = \bar{x}_{12} - \bar{x}_2$

Eg: S.D is expressed in terms of units in which the original figures are collated and stated.

The S.D of heights of students cannot be compared with a S.D of weights of students, because both are of different units.

Prices of particular commodity in five years in two cities are given below.

	price in city A	price in city B	
	20	10	
	22	20	y
x	19	18	
	23	12	
	16	13	

Which city has more stable prices?

$\sigma_x = 2.45$

$\sigma_y = 3.69$

C.V of x = 12.25 %.

C.V of y = 24.6 %

∴ A has more stable price than city B.

C.V of A < C.V of B

Quartiles

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

Discrete	1
3	4.

Q_1 = lower quartile

Q_2 = ($\frac{1}{2}$ dis disc) median.

Q_3 = upper quartile.

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

Raw or ungroupe data

1. First arrange in increasing order

2. $Q_1 = \left(\frac{n+1}{4}\right)^{th}$ item

$Q_3 = 3\left(\frac{n+1}{4}\right)^{th}$ item.

For ema 5, 8, 10, 15, 16, 25, 30, 40, 47

$Q_1 = \left(\frac{n+1}{4}\right)^{th}$ iw $= 2.25^{th}$ iw

$= 2^{nd}$ item + 0.25 (3^{rd} iw $- 2^{nd}$ iw)

$Q_3 = 3\left(\frac{n+1}{4}\right)^{th}$ iw $= 8.25^{th}$

$= 8^{th}$ item + .25 (9^{th} valu $- 8^{th}$ valu)

Discrete series

1. Find c.f
2. Find $\frac{N+1}{4}$
3. see > $\left(\frac{N+1}{4}\right)$ valu, then the corresponding value of x is Q_1

A: Find $3\left(\frac{N+1}{4}\right)$

S. seq $> 3\left(\frac{N+1}{4}\right)$, then the corresponding

x is Q_3.

Ex Compute quartiles

x	5	8	12	15	19	24	30
f	4	3	2	4	5	2	4

$Q_1 = 6.25^{th}$ = 6^{th} value + 0.25 (7^{th} val − 6^{th} val)

= 8 + 0.25 (8 − 8) = 8

$Q_3 = 18.75 = 18^{th}$ val + .75 ($19^{th} - 18^{th}$)

= 24 + (0.75 (24 − 24)

= 24.

Continuous seq

1. C.f $N = \Sigma f x$
2. find $\frac{N}{4}$
3. see $> \frac{N}{4}$, the corresponding class interval is called a Q_1 class
4. Find $3\left(\frac{N}{4}\right)$, see $> 3\left(\frac{N}{4}\right)$ then the corresponding class is called a Q_3 class

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

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

l = lower limit of the quartile class

f - frequency of the quartile class

c - width of the quartile class

m → c.f preceding the quartile class