CHAPTER 4
4. Measures of Dispersion (Variation)

Introduction and objectives of measuring Variation

-The scatter or spread of items of a distribution is known as dispersion or variation. In other words the degree to which numerical data tend to spread about an average value is called dispersion or variation of the data.
-Measures of dispersions are statistical measures which provide ways of measuring the extent in which data are dispersed or spread out.

Objectives of measuring Variation:

- To judge the reliability of measures of central tendency
- To control variability itself.
- To compare two or more groups of numbers in terms of their variability.
- To make further statistical analysis.

Absolute and Relative Measures of Dispersion

The measures of dispersion which are expressed in terms of the original unit of a series are termed as absolute measures. Such measures are not suitable for comparing the variability of two distributions which are expressed in different units of measurement and different average size. Relative measures of dispersions are a ratio or percentage of a measure of absolute dispersion to an appropriate measure of central tendency and are thus pure numbers independent of the units of measurement. For comparing the variability of two distributions (even if they are measured in the same unit), we compute the relative measure of dispersion instead of absolute measures of dispersion.

Types of Measures of Dispersion

Various measures of dispersions are in use. The most commonly used measures of dispersions are:

1) Range and relative range
2) Quartile deviation and coefficient of Quartile deviation
3) Mean deviation and coefficient of Mean deviation
4) Standard deviation and coefficient of variation.

The Range (R)

The range is the largest score minus the smallest score. It is a quick and dirty measure of variability, although when a test is given back to students they very often wish to know the range of scores. Because the range is greatly affected by extreme scores, it may give a distorted picture of the scores. The following two distributions have the same range, 13, yet appear to differ greatly in the amount of variability.

Distribution 1: 32 35 36 36 37 38 40 42 42 43 43 45

Distribution 2: 32 32 33 33 33 34 34 34 34 34 35 45

For this reason, among others, the range is not the most important measure of variability.

$$R = L - S \qquad , L = l\arg est\ observation$$
$$S = smallest\ observation$$

Range for grouped data:

If data are given in the shape of continuous frequency distribution, the range is computed as:

$$R = UCL_k - LCL_1 , \qquad UCL_k \ \ is\ upperclass \lim it\ of\ the\ last\ class.$$
$$UCL_1 \quad is\ lower\ class \lim it\ of\ the\ first\ class.$$

This is some times expressed as:

$$R = X_k - X_1 , \qquad X_k \ \ is\ class\ mark\ of\ the\ last\ class.$$
$$X_1 \quad is\ classmark\ of\ the\ first\ class.$$

Merits and Demerits of range

Merits:

- It is rigidly defined.
- It is easy to calculate and simple to understand.

Demerits:

- It is not based on all observation.
- It is highly affected by extreme observations.
- It is affected by fluctuation in sampling.
- It is not liable to further algebraic treatment.
- It can not be computed in the case of open end distribution.
- It is very sensitive to the size of the sample.

Relative Range (RR)

It is also some times called coefficient of range and given by:

$$RR = \frac{L - S}{L + S} = \frac{R}{L + S}$$

Example:

1. Find the relative range of the above two distribution. (Exercise!)
2. If the range and relative range of a series are 4 and 0.25 respectively. Then what is the value of: a) Smallest observation b) Largest observation

Solution: (2)

$R = 4 \Rightarrow L - S = 4$ ________________(1)

$RR = 0.25 \Rightarrow L + S = 16$ _____________(2)

Solving (1) *and* (2) *at the same time, one can obtain the following value*

$L = 10$ *and* $S = 6$

The Quartile Deviation (Semi-inter quartile range), Q.D

The inter quartile range is the difference between the third and the first quartiles of a set of items and semi-inter quartile range is half of the inter quartile range.

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

Coefficient of Quartile Deviation (C.Q.D)

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

- It gives the average amount by which the two quartiles differ from the median.

Example: Compute Q.D and its coefficient for the following distribution.

Values	Freq.
140- 150	17
150- 160	29
160- 170	42
170- 180	72
180- 190	84
190- 200	107
200- 210	49
210- 220	34
220- 230	31
230- 240	16
240- 250	12

Solutions:

In the previous chapter we have obtained the values of all quartiles as:

Q_1= 174.90, Q_2= 190.23, Q_3=203.83

$$\Rightarrow \quad Q.D = \frac{Q_3 - Q_1}{2} = \frac{203.83 - 174.90}{2} = 14.47$$

$$C.Q.D = \frac{2 * Q.D}{Q_3 + Q_1} = \frac{2 * 14.47}{203.83 + 174.90} = 0.076$$

Remark: Q.D or C.Q.D includes only the middle 50% of the observation.

The Mean Deviation (M.D):

The mean deviation of a set of items is defined as the arithmetic mean of the values of the absolute deviations from a given average. Depending up on the type of averages used we have different mean deviations.

a) Mean Deviation about the mean

- Denoted by M.D($\overline{X}$) and given by

$$M.D(\overline{X}) = \frac{\sum_{i=1}^{n} \left| X_i - \overline{X} \right|}{n}$$

- For the case of frequency distribution it is given as:

$$M.D(\overline{X}) = \frac{\sum_{i=1}^{k} f_i \left| X_i - \overline{X} \right|}{n}$$

Steps to calculate M.D ($\overline{X}$):

1. Find the arithmetic mean, $\overline{X}$
2. Find the deviations of each reading from $\overline{X}$.
3. Find the arithmetic mean of the deviations, ignoring sign.

b) Mean Deviation about the median.

- Denoted by M.D($\widetilde{X}$) and given by

$$M.D(\widetilde{X}) = \frac{\sum_{i=1}^{n} \left| X_i - \widetilde{X} \right|}{n}$$

- For the case of frequency distribution it is given as:

$$M.D(\widetilde{X}) = \frac{\sum_{i=1}^{k} f_i \left| X_i - \widetilde{X} \right|}{n}$$

Steps to calculate M.D ($\widetilde{X}$):

1. Find the median, $\widetilde{X}$
2. Find the deviations of each reading from $\widetilde{X}$.
3. Find the arithmetic mean of the deviations, ignoring sign.

c) Mean Deviation about the mode.

- Denoted by M.D($\hat{X}$) and given by

$$M.D(\hat{X}) = \frac{\sum_{i=1}^{n} \left| X_i - \hat{X} \right|}{n}$$

- For the case of frequency distribution it is given as:

$$M.D(\hat{X}) = \frac{\sum_{i=1}^{k} f_i \left| X_i - \hat{X} \right|}{n}$$

Steps to calculate M.D ($\hat{X}$):

1. Find the mode, $\hat{X}$
2. Find the deviations of each reading from $\hat{X}$.
3. Find the arithmetic mean of the deviations, ignoring sign.

Examples:

1. The following are the number of visit made by ten mothers to the local doctor's surgery.
   8, 6, 5, 5, 7, 4, 5, 9, 7, 4
   Find mean deviation about mean, median and mode.
   Solutions:
   First calculate the three averages
   $$\overline{X} = 6, \ \widetilde{X} = 5.5, \quad \hat{X} = 5$$

   Then take the deviations of each observation from these averages.

X_i	4	4	5	5	5	6	7	7	8	9	total
$\lvert X_i - 6\rvert$	2	2	1	1	1	0	1	1	2	3	14
$\lvert X_i - 5.5\rvert$	1.5	1.5	0.5	0.5	0.5	0.5	1.5	1.5	2.5	3.5	14
$\lvert X_i - 5\rvert$	1	1	0	0	0	1	2	2	3	4	14

$$\Rightarrow M.D(\overline{X}) = \frac{\sum_{i=1}^{10} |X_i - 6)|}{10} = \frac{14}{10} = 1.4$$

$$M.D(\widetilde{X}) = \frac{\sum_{i=1}^{10} |X_i - 5.5|}{10} = \frac{14}{10} = 1.4$$

$$M.D(\hat{X}) = \frac{\sum_{i=1}^{10} |X_i - 5)|}{10} = \frac{14}{10} = 1.4$$

2. Find mean deviation about mean, median and mode for the following distributions.(**exercise**)

Class	**Frequency**
40-44	**7**
45-49	**10**
50-54	**22**

55-59	**15**
60-64	**12**
65-69	**6**
70-74	**3**

Remark: Mean deviation about the median is always the smallest.

Coefficient of Mean Deviation (C.M.D)

$$C.M.D = \frac{M.D}{Average\ \ about\ which\ deviations\ are\ taken}$$

$$\Rightarrow C.M.D(\overline{X}) = \frac{M.D(\overline{X})}{\overline{X}}$$

$$C.M.D(\widetilde{X}) = \frac{M.D(\widetilde{X})}{\widetilde{X}}$$

$$C.M.D(\hat{X}) = \frac{M.D(\hat{X})}{\hat{X}}$$

Example: calculate the C.M.D about the mean, median and mode for the data in example 1 above.

Solutions:

$$C.M.D = \frac{M.D}{Average\ \ about\ which\ deviations\ are\ taken}$$

$$\Rightarrow C.M.D(\overline{X}) = \frac{M.D(\overline{X})}{\overline{X}} = \frac{1.4}{6} = 0.233 \qquad C.M.D(\widetilde{X}) = \frac{M.D(\widetilde{X})}{\widetilde{X}} = \frac{1.4}{5.5} = 0.255$$

$$C.M.D(\hat{X}) = \frac{M.D(\hat{X})}{\hat{X}} = \frac{1.4}{5} = 0.28$$

Exercise: Identify the merits and demerits of Mean Deviation

The Variance

Population Variance

If we divide the variation by the number of values in the population, we get something called the population variance. This variance is the "average squared deviation from the mean".

$$Population\ Varince = \sigma^2 = \frac{1}{N}\sum (X_i - \mu)^2, \ \ i = 1,2,.....N$$

For the case of frequency distribution it is expressed as:

$$Population\ Varince = \sigma^2 = \frac{1}{N}\sum f_i(X_i - \mu)^2, \quad i = 1,2,.....k$$

Sample Variance

One would expect the sample variance to simply be the population variance with the population mean replaced by the sample mean. However, one of the major uses of statistics is to estimate the corresponding parameter. This formula has the problem that the estimated value isn't the same as the parameter. To counteract this, the sum of the squares of the deviations is divided by one less than the sample size.

$$Sample\ Varince = S^2 = \frac{1}{n-1}\sum (X_i - \overline{X})^2, \quad i = 1,2,.....,n$$

For the case of frequency distribution it is expressed as:

$$Sample\ Varince = S^2 = \frac{1}{n-1}\sum f_i(X_i - \overline{X})^2, \quad i = 1,2,.....k$$

We usually use the following short cut formula.

$$S^2 = \frac{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}{n-1}, \ for\ raw\ data.$$

$$S^2 = \frac{\sum_{i=1}^{k} f_i X_i^2 - n\overline{X}^2}{n-1}, \ for\ frequency\ distribution.$$

Standard Deviation

There is a problem with variances. Recall that the deviations were squared. That means that the units were also squared. To get the units back the same as the original data values, the square root must be taken.

$$Population\ s\tan dard\ deviation = \sigma = \sqrt{\sigma^2}$$
$$Sample\ s\tan dard\ deviation = s = \sqrt{S^2}$$

The following steps are used to calculate the sample variance:

1. Find the arithmetic mean.
2. Find the difference between each observation and the mean.
3. Square these differences.
4. Sum the squared differences.

5. Since the data is a sample, divide the number (from step 4 above) by the number of observations minus one, i.e., n-1 (where n is equal to the number of observations in the data set).

Examples: Find the variance and standard deviation of the following sample data

1. 5, 17, 12, 10.
2. The data is given in the form of frequency distribution.

Class	**Frequency**
40-44	7
45-49	10
50-54	22
55-59	15
60-64	12
65-69	6
70-74	3

Solutions:

1. $\overline{X} = 11$

X_i	5	10	12	17	Total
$(X_i - \overline{X})^2$	36	1	1	36	74

$$\Rightarrow S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1} = \frac{74}{3} = 24.67.$$

$$\Rightarrow S = \sqrt{S^2} = \sqrt{24.67} = 4.97.$$

2. $\overline{X} = 55$

X_i(C.M)	42	47	52	57	62	67	72	Total
$f_i(X_i - \overline{X})^2$	1183	640	198	60	588	864	867	4400

$$\Rightarrow S^2 = \frac{\sum_{i=1}^{n} f_i(X_i - \overline{X})^2}{n-1} = \frac{4400}{74} = 59.46.$$

$$\Rightarrow S = \sqrt{S^2} = \sqrt{59.46} = 7.71.$$

Special properties of Standard deviations

1. $\sqrt{\dfrac{\sum(X_i-\overline{X})^2}{n-1}} < \sqrt{\dfrac{\sum(X_i-A)^2}{n-1}} \quad , A \neq \overline{X}$

2. For normal (symmetric) distribution the following holds.

- Approximately 68.27% of the data values fall within one standard deviation of the mean. i.e. with in $(\overline{X}-S,\ \overline{X}+S)$
- Approximately 95.45% of the data values fall within two standard deviations of the mean. i.e. with in $(\overline{X}-2S,\ \overline{X}+2S)$
- Approximately 99.73% of the data values fall within three standard deviations of the mean. i.e. with in $(\overline{X}-3S,\ \overline{X}+3S)$

3. **Chebyshev's Theorem**

For any data set ,no matter what the pattern of variation, the proportion of the values that fall with in k standard deviations of the mean or $(\overline{X}-kS,\ \overline{X}+kS)$ will be at least $1-\dfrac{1}{k^2}$, where k is a number greater than 1. i.e. the proportion of items falling beyond k standard deviations of the mean is at most $\dfrac{1}{k^2}$

Example: Suppose a distribution has mean 50 and standard deviation 6. What percent of the numbers are:

a) Between 38 and 62
b) Between 32 and 68
c) Less than 38 or more than 62.
d) Less than 32 or more than 68.

Solutions:

a) 38 and 62 are at equal distance from the mean,50 and this distance is 12

$\Rightarrow ks = 12$

$$\Rightarrow k = \frac{12}{S} = \frac{12}{6} = 2$$

➔ Applying the above theorem, at least $(1-\dfrac{1}{k^2})*100\% = 75\%$ of the numbers lie between 38 and 62.

b) Similarly done.

c) It is just the complement of a) i.e. at most $\frac{1}{k^2} * 100\% = 25\%$ of the numbers lie less than 32 or more than 62.

d) Similarly done.

Exercise: The average score of a special test of knowledge of wood refinishing has a mean of 53 and standard deviation of 6. Find the range of values in which at least 75% the scores will lie.

4. If the standard deviation of $X_1, X_2, ...X_n$ *is* S, then the standard deviation of

a) $X_1 + k,\ X_2 + k, ...X_n + k$ *will also be* S

b) $kX_1,\ kX_2, ...kX_n$ *would be* $|k| S$

c) $a + kX_1,\ a + kX_2, ...a + kX_n$ *would be* $|k|\ S$

Exercise: Verify each of the above relation ship, considering *k* and *a* as constants.

Examples:

1. The mean and standard deviation of n Tetracycline Capsules $X_1, X_2, ...X_n$ are known to be 12 gm and 3 gm respectively. New set of capsules of another drug are obtained by the linear transformation $Y_i = 2X_i - 0.5$ (i = 1, 2, …, n) then what will be the standard deviation of the new set of capsules.

2. The mean and the standard deviation of a set of numbers are respectively 500 and 10.
   a) If 10 are added to each of the numbers in the set, then what will be the variance and standard deviation of the new set?
   b) If each of the numbers in the set are multiplied by -5, then what will be the variance and standard deviation of the new set?

Solutions:

1. Using c) above the new standard deviation $= |k| S = 2 * 3 = 6$
2. a. They will remain the same.

   b. New standard deviation $= |k| S = 5 * 10 = 50$

<u>Coefficient of Variation (C.V)</u>

- Is defined as the ratio of standard deviation to the mean usually expressed as percents.

$$C.V = \frac{S}{\overline{X}} * 100$$

- The distribution having less C.V is said to be less variable or more consistent.

Example: An analysis of the monthly wages paid (in Birr) to workers in two firms A and B belonging to the same industry gives the following results

Value	Firm A	Firm B
Mean wage	52.5	47.5
Median wage	50.5	45.5
Variance	100	121

In which firm A or B is there greater variability in individual wages?

Solutions:
Calculate coefficient of variation for both firms.

$$C.V_A = \frac{S_A}{\overline{X}_A} * 100 = \frac{10}{52.5} * 100 = 19.05\%$$

$$C.V_B = \frac{S_B}{\overline{X}_B} * 100 = \frac{11}{47.5} * 100 = 23.16\%$$

Since $C.V_A < C.V_B$, in firm B there is greater variability in individual wages.

Exercise: A meteorologist interested in the consistency of temperatures in three cities during a given week collected the following data. The temperatures for the five days of the week in the three cities were

City 1	25	24	23	26	17
City2	22	21	24	22	20
City3	32	27	35	24	28

Which city have the most consistent temperature, based on these data?

Standard Scores (Z-scores)

- If X is a measurement from a distribution with mean $\overline{X}$ and standard deviation S, then its value in standard units is

$$Z = \frac{X - \mu}{\sigma}, \ \textit{for population.}$$

$$Z = \frac{X - \overline{X}}{S}, \ \textit{for sample}$$

- Z gives the deviations from the mean in units of standard deviation
- Z gives the number of standard deviation a particular observation lie above or below the mean.
- It is used to compare two observations coming from different groups.

Examples:

1. Two sections were given introduction to statistics examinations. The following information was given.

Value	Section 1	Section 2

Mean	78	90
Stan.deviation	6	5

Student A from section 1 scored 90 and student B from section 2 scored 95.Relatively speaking who performed better?

Solutions:

Calculate the standard score of both students.

$$Z_A = \frac{X_A - \overline{X}_1}{S_1} = \frac{90 - 78}{6} = 2$$

$$Z_B = \frac{X_B - \overline{X}_2}{S_2} = \frac{95 - 90}{5} = 1$$

➔ Student A performed better relative to his section because the score of student A is two standard deviations above the mean score of his section while, the score of student B is only one standard deviation above the mean score of his section.

2. Two groups of people were trained to perform a certain task and tested to find out which group is faster to learn the task. For the two groups the following information was given:

Value	Group one	Group two
Mean	10.4 min	11.9 min
Stan.dev.	1.2 min	1.3 min

Relatively speaking:

a) Which group is more consistent in its performance
b) Suppose a person A from group one take 9.2 minutes while person B from Group two take 9.3 minutes, who was faster in performing the task? Why?

Solutions:

a) Use coefficient of variation.

$$C.V_1 = \frac{S_1}{\overline{X}_1} * 100 = \frac{1.2}{10.4} * 100 = 11.54\%$$

$$C.V_2 = \frac{S_2}{\overline{X}_2} * 100 = \frac{1.3}{11.9} * 100 = 10.92\%$$

Since $C.V_2 < C.V_1$, group 2 is more consistent.

b) Calculate the standard score of A and B

$$Z_A = \frac{X_A - \bar{X}_1}{S_1} = \frac{9.2 - 10.4}{1.2} = -1$$

$$Z_B = \frac{X_B - \bar{X}_2}{S_2} = \frac{9.3 - 11.9}{1.3} = -2$$

➔Child B is faster because the time taken by child B is two standard deviations shorter than the average time taken by group 2, while the time taken by child A is only one standard deviation shorter than the average time taken by group 1.

Moments

If X is a variable that assume the values $X_1, X_2, \ldots, X_n$ then

1. The r[th] moment is defined as:

$$\bar{X}^r = \frac{X_1^r + X_2^r + \ldots + X_n^r}{n}$$

$$= \frac{\sum_{i=1}^{n} X_i^r}{n}$$

- For the case of frequency distribution this is expressed as:

$$\bar{X}^r = \frac{\sum_{i=1}^{k} f_i X_i^r}{n}$$

- If $r = 1$, it is the simple arithmetic mean, this is called the first moment.

2. The r[th] moment about the mean (the r[th] central moment)

- Denoted by M_r and defined as:

$$M_r = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^r}{n} = \frac{(n-1)}{n} \frac{\sum_{i=1}^{n} (X_i - \bar{X})^r}{n-1}$$

- For the case of frequency distribution this is expressed as:

$$M_r = \frac{\sum_{i=1}^{k} f_i (X_i - \bar{X})^r}{n}$$

- If $r = 2$, it is population variance, this is called the second central moment. If we assume $n - 1 \approx n$, it is also the sample variance.

3. The r[th] moment about any number A is defined as:

- Denoted by $M_r^{'}$ and

$$M_r' = \frac{\sum_{i=1}^{n}(X_i - A)^r}{n} = \frac{(n-1)}{n}\frac{\sum_{i=1}^{n}(X_i - A)^r}{n-1}$$

- For the case of frequency distribution this is expressed as:

$$M_r' = \frac{\sum_{i=1}^{k} f_i (X_i - A)^r}{n}$$

Example:

1. Find the first two moments for the following set of numbers 2, 3, 7
2. Find the first three central moments of the numbers in problem 1
3. Find the third moment about the number 3 of the numbers in problem 1.

Solutions:

1. Use the r[th] moment formula.

$$\overline{X}^r = \frac{\sum_{i=1}^{n} X_i^r}{n}$$

$$\Rightarrow \overline{X}^1 = \frac{2+3+7}{3} = 4 = \overline{X}$$

$$\overline{X}^2 = \frac{2^2 + 3^2 + 7^2}{3} = 20.67$$

2. Use the r[th] central moment formula.

$$M_r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^r}{n}$$

$$\Rightarrow M_1 = \frac{(2-4)+(3-4)+(7-4)}{3} = 0$$

$$M_2 = \frac{(2-4)^2 + (3-4)^2 + (7-4)^2}{3} = 4.67$$

$$M_3 = \frac{(2-4)^3 + (3-4)^3 + (7-4)^3}{3} = 6$$

3. Use the r[th] moment about A.

$$M_r = \frac{\sum_{i=1}^{n}(X_i - A)^r}{n}$$

$$\Rightarrow M_3{}' = \frac{(2-3)^3 + (3-3)^3 + (7-3)^3}{3} = 21$$

Skewness

- Skewness is the degree of asymmetry or departure from symmetry of a distribution.
- A skewed frequency distribution is one that is not symmetrical.
- Skewness is concerned with the shape of the curve not size.
- If the frequency curve (smoothed frequency polygon) of a distribution has a longer tail to the right of the central maximum than to the left, the distribution is said to be skewed to the right or said to have positive skewness. If it has a longer tail to the left of the central maximum than to the right, it is said to be skewed to the left or said to have negative skewness.
- For moderately skewed distribution, the following relation holds among the three commonly used measures of central tendency.
  $Mean - Mode = 3*(Mean - Median)$

Measures of Skewness

- Denoted by α_3
- There are various measures of skewness.

1. The Pearsonian coefficient of skewness

$$\alpha_3 = \frac{Mean - Mode}{Standard\ deviation} = \frac{\overline{X} - \hat{X}}{S}$$

2. The Bowley's coefficient of skewness (coefficient of skewness based on quartiles)

$$\alpha_3 = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1} = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$$

3. The moment coefficient of skewness

$$\alpha_3 = \frac{M_3}{M_2^{3/2}} = \frac{M_3}{(\sigma^2)^{3/2}} = \frac{M_3}{\sigma^3},\ Where\ \sigma\ is\ the\ population\ standard\ deviation.$$ The shape of the curve is determined by the value of α_3

- If $\alpha_3 > 0$ then the distribution is positively skewed.
- If $\alpha_3 = 0$ then the distribution is symmetric.
- If $\alpha_3 < 0$ then the distribution is negatively skewed.

Remark:

- In a positively skewed distribution, smaller observations are more frequent than larger observations. i.e. the majority of the observations have a value below an average.
- In a negatively skewed distribution, smaller observations are less frequent than larger observations. i.e. the majority of the observations have a value above an average.

Examples:

1. Suppose the mean, the mode, and the standard deviation of a certain distribution are 32, 30.5 and 10 respectively. What is the shape of the curve representing the distribution?

Solutions:

Use the Pearsonian coefficient of skewness

$$\alpha_3 = \frac{Mean - Mode}{S\tan dard\ deviation} = \frac{32 - 30.5}{10} = 0.15$$

$$\alpha_3 > 0 \Rightarrow The\ distribution\ is\ positively\ skewed.$$

2. In a frequency distribution, the coefficient of skewness based on the quartiles is given to be 0.5. If the sum of the upper and lower quartile is 28 and the median is 11, find the values of the upper and lower quartiles.

Solutions:

Given: $\alpha_3 = 0.5, \quad \widetilde{X} = Q_2 = 11$ Required: Q_1, Q_3

$$Q_1 + Q_3 = 28 ....(*)$$

$$\alpha_3 = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1} = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1} = 0.5$$

Substituting the given values, one can obtain the following

$$Q_3 - Q_1 = 12 ....(**)$$

Solving (*) *and* (**) *at the same time we obtain the following values*

$$Q_1 = 8 \quad and \quad Q_3 = 20$$

Exercises:

1. Some characteristics of annually family income distribution (in Birr) in two regions is as follows:

Region	Mean	Median	Standard Deviation
A	6250	5100	960
B	6980	5500	940

a) Calculate coefficient of skewness for each region

b) For which region is, the income distribution more skewed. Give your interpretation for this Region
c) For which region is the income more consistent?

3. For a moderately skewed frequency distribution, the mean is 10 and the median is 8.5. If the coefficient of variation is 20%, find the Pearsonian coefficient of skewness and the probable mode of the distribution.

4. The sum of fifteen observations, whose mode is 8, was found to be 150 with coefficient of variation of 20%
   (a) Calculate the pearsonian coefficient of skewness and give appropriate conclusion.
   (b) Are smaller values more or less frequent than bigger values for this distribution?
   (c) If a constant *k* was added on each observation, what will be the new pearsonian coefficient of skewness? Show your steps. What do you conclude from this?

Kurtosis

Kurtosis is the degree of peakdness of a distribution, usually taken relative to a normal distribution. A distribution having relatively high peak is called leptokurtic. If a curve representing a distribution is flat topped, it is called platykurtic. The normal distribution which is not very high peaked or flat topped is called mesokurtic.

Measures of kurtosis

The moment coefficient of kurtosis:

- Denoted by α_4 and given by

$$\alpha_4 = \frac{M_4}{M_2^{\ 2}} = \frac{M_4}{\sigma^4}$$

Where: M_4 *is the fourth moment about the mean.*

M_2 *is the second moment about the mean.*

σ *is the population standard deviation.*

The peakdness depends on the value of α_4.

- *If* $\alpha_4 > 3$ *then the curve is leptokurtic.*
- *If* $\alpha_4 = 3$ *then the curve is mesokurtic.*
- *If* $\alpha_4 < 3$ *then the curve is platykurtic.*

Examples:

1. If the first four central moments of a distribution are:
   $M_1 = 0,\ M_2 = 16,\ M_3 = -60,\ M_4 = 162$

a) Compute a measure of skewness
b) Compute a measure of kurtosis and give your interpretation.

Solutions:

a) $$\alpha_3 = \frac{M_3}{{M_2}^{3/2}} = \frac{-60}{16^{3/2}} = -0.94 < 0$$

$\Rightarrow$ *The distribution is negatively skewed.*

b) $$\alpha_4 = \frac{M_4}{{M_2}^2} = \frac{162}{16^2} = 0.6 < 3$$

$\Rightarrow$ *The curve is platykurtic.*

Exercise:

1. The median and the mode of a mesokurtic distribution are 32 and 34 respectively. The 4th moment about the mean is 243. Compute the Pearsonian coefficient of skewness and identify the type of skewness. Assume (n-1 = n).

2. If the standard deviation of a symmetric distribution is 10, what should be the value of the fourth moment so that the distribution is mesokurtic?