CHAPTER FOUR

MEASUREMENT OF VARIATION /DISPERSION

Specific objectives

At the end of this topic, the trainee should be able to:-

- State the characteristics of a good measure of dispersion.
- Differentiate between the absolute and relative measures.
- Calculate and interpret the measures of dispersion.

Introduction

Measures of variation help us in studying the important characteristics of a distribution. The measures of dispersion are very useful in statistical work because they indicate whether the rest of the data are scattered around the mean or away from the mean. If the data is approximately dispersed around the mean then the measure of dispersion obtained will be small therefore indicating that the mean is a good representative of the sample data. But on the other hand, if the figures are not closely located to the mean then the measures of dispersion obtained will be relatively big indicating that the mean does not represent the data sufficiently.

Significance of measuring variation or dispersion

i) to determine the reliability of an average
ii) to serve as a basis for the control of the variability
iii) to compare two or more series with regard to their variability

iv) to facilitate the use of other statistical measures

Property of a good measure of variation/dispersion.
A good measure of variation should possess as far as possible

i) It should be simple to understand
ii) It should be easy to compute
iii) It should be rigidly d3efined
iv) It should be based on every observation of the distribution
v) It should be amenable to further algebraic treatment
vi) It should have sampling stability
vii) It should not be unduly affected by extreme observation

Methods of studying variation
The following Aretha important methods of studying variation

- The range
- The Mean deviation
- The interquartile range or quartile deviation
- The standard deviation
- The Lorenz curve

Absolute and relative measures of variation
Measures of variation may be either absolute or relative. Absolute measures of variation are expressed in the same statistical unit in which the original data are given. Such shillings, kilograms, tones, etc
These values may be used to compare the variation in two or more than two distributions provided the variables are expressed in the same units and have almost the same average value.
In case the two sets of data are expressed in different units such as manager's salary versus workers salary, the absolute measures of variation are not comparable. In such cases measures of relative variation should be used.
A measure of relative variation is the ratio of a measure of absolute variation to an average. It is sometimes called a coefficient of variation, because coefficient means a pure number that is independent of the unit of measurement.

a) The Range

The range is defined as the difference between the highest and the smallest values in a frequency distribution. This measure is not very efficient because it utilizes only 2 values in a given frequency distribution. However the smaller the value of the range, the less dispersed the observations are from the arithmetic mean and vice versa

The range is not commonly used in business management because 2 sets of data may yield the same range but end up having different interpretations regarding the degree of dispersion.
Range is the simplest method of studying variation. It is defined as the difference between the value of the smallest observation and the value of the largest observation

Range = largest value- smallest value

R= L-S

Coefficient of Range = $\frac{L-S}{L+S}$

Example

The following are the prices of shares of a company from Monday to Saturday

days	Prices
Monday	200
Tuesday	210
Wednesday	208
Thursday	160
Friday	220
Saturday	250

Calculate the range and coefficient of range.
Solution
Range =L-S
L=250
S= 160
Range =250-160
= 90
Coefficient of range = $\frac{L-S}{L+S} = \frac{250-160}{250+160} = \frac{90}{410} = 0.219$

Merits of range

- It is the simplest to understand and easiest to compute
- It takes minimum time to calculate the value of range

Limitation

- Range is not based on each and every observation of the distribution
- It is subject to fluctuation of considerable magnitude from sample to sample
- Range cannot be computed in case of open –end distribution
- Range cannot tell us anything about the character of the distribution within two extreme observations

Uses of range

1. Quality control; the objective of quality control is to keep a check on the quality of the products without 100% inspection control charts are prepared.
2. fluctuation in the shares price; range is useful in studying the variation in the prices in stocks and shares and other commodities
3. Weather forecast; the meteorological department does make use of the range in determining the difference between the minimum temperature and maximum temperature.

(b) Mean Deviation

This is a useful measure of dispersion because it makes use of all the values given. The average deviation is obtained by calculating the absolute deviation of each observation from median (or mean) and then averaging this deviation by taking their arithmetic mean. The formula for average deviation may be written as;

$$A.D = \frac{\sum /x - med/}{N}$$

Incase deviation are taken from mean the formula shall be written as;

$$A.D\ x = \frac{\sum /x - x/}{N}$$

Computation of average deviation

The formula for computing average deviation is

$$A.D\ (med) = \frac{\sum /x - med/}{N}$$

If average deviation is small the distribution is highly compact or uniform since more than half of the cases are concentrated within a small range around the mean.

The relative measure corresponding to the average deviation called the coefficient of average deviation is obtained by dividing average by the measure of central tendency used, i.e. mean or median.

Example 1

In a given exam the scores for 10 students were as follows

Student	Mark (x)	$\lvert x - \bar{x} \rvert$
A	60	1.8
B	45	16.8
C	75	13.2
D	70	8.2

E	65	3.2
F	40	21.8
G	69	7.2
H	64	2.2
I	50	11.8
J	80	18.2
Total	**618**	**104.4**

Required

Determine the absolute mean deviation

Mean, $\bar{x} = \frac{618}{10} = 61.8$

Therefore AMD = $\frac{\Sigma\left|X - \bar{X}\right|}{N} = \frac{104.4}{10} = 10.44$

Example 2

The following data was obtained from a given financial institution. The data refers to the loans given out in 1996 to several firms

Firms (f)	Amount of loan per firm (x)	fx	$\left\|x-\bar{x}\right\|$	$\left\|x-\bar{x}\right\|.f$
3	20000	60000	4157.9	12473.70
4	60000	240000	35842.1	143368.40
1	15000	15000	9157.9	9157.9
5	12000	60000	12157.9	60789.50
6	14000	84000	10157.9	60947.40
Σf = 19		Σfx = 459000		286736.90

Required

Calculate the mean deviation for the amount of items given

$$\overline{X} = \frac{\sum fx}{\sum f} = \frac{459,000}{19} = 24157.9$$

$$\therefore AMD = \frac{\sum \left| X - \overline{X} \right|}{\sum f} = \frac{286736.90}{19}$$

$$= Shs\, 15,091.40$$

NB if the absolute mean deviation is relatively small it implies that the data is more compact and therefore the arithmetic mean is a fair sample representative.

(b) <u>The interquartile range or quartile deviation</u>

This is a measure of dispersion which involves the use of quartile. A quartile is a mark or a value which lies at the boundary of a division when any given set of data is divided into four equal divisions

Each of such divisions normally carries 25% of all the observations

The semi interquartile range is a good measure of dispersion because it shows how the rest of the data are generally spread around the mean

The quartiles normally used are three namely;

i. The lower quartile (first quartile Q1) this usually binds the lower 25% of the data
ii. The median (second quartile Q2)
iii. The upper quartile (third quartile Q3)

Interquartile range = Q3-Q1

Very often the interquartile range is reduced to the form of the semi-interquartile range or quartile deviation by dividing it by 2

Q.D = quartile deviation

Q.D = Q3-Q1/2

Quartile deviation gives the average amount by which the quartile differs from the median. Quartile deviation is an absolute measure of variation. The relative measure corresponding to the measure called the coefficient of quartile deviation.

Coefficient of Q.D = Q3-Q1/Q3+Q1

The semi-interquartile range,

$$\text{SIR} = \frac{\text{Q3} - \text{Q1}}{2}$$

<u>Example 1</u>

The weights of 15 parcels recorded at the GPO were as follows:
16.2, 17, 20, 25(Q1) 29, 32.2, 35.8, 36.8(Q2) 40, 41, 42, 44(Q3) 49, 52, 55 (in kgs)

Required

Determine the semi interquartile range for the above data

$$SIR = \frac{Q3 - Q1}{2} = \frac{44 - 25}{2} = \frac{19}{2} = 8.5$$

Example 2 (Grouped Data)

The following table shows the levels of retirement benefits given to a group of workers in a given establishment.

Retirement benefits £ '000	No of retirees (f)	UCB	cf
20 - 29	50	29.5	50
30 - 39	69	39.5	119
40 - 49	70	49.5	189
50 - 59	90	59.5	279
60 - 69	52	69.5	331
70 - 79	40	79.5	371
80 - 89	11	89.5	382

Required

i. Determine the semi interquartile range for the above data
ii. Determine the minimum value for the top ten per cent.(10%)
iii. Determine the maximum value for the lower 40% of the retirees

Solution

The lower quartile (Q1) lies on position

$$\frac{N+1}{4} = \frac{382+1}{4}$$

$$= 95.75$$

$$\therefore \text{the value of } Q1 = 29.5 + \frac{(95.75 - 50)}{69} \times 10$$

= 29.5 + 6.63

= £36.13

The upper quartile (Q3) lies on position

$$= 3\left(\frac{N + 1}{4}\right)$$

$$= 3\left(\frac{382 + 1}{4}\right)$$

= 287.25

∴ the value of Q3 $= 59.5 + \frac{(287.25 - 279)}{52} \times 10$

= 61.08

The semi interquartile range = $\frac{Q3 - Q1}{2}$

$$= \frac{61.08 - 36.13}{2}$$

= 12.475

= £12,475

ii. The top 10% is equivalent to the lower 90% of the retirees

The position corresponding to the lower 90%

$$= \frac{90}{100}(n + 1) = 0.9\,(382 + 1)$$

= 0.9 x 383

= 344.7

∴ the benefits (value) corresponding to the minimum value for top 10%

= 69.5 + $\frac{(344.7 - 331)}{40}$ x 10

= 72.925

= £ 72925

iii. The lower 40% corresponds to position

$$= \frac{40}{100}(382 + 1)$$

$$= 153.20$$

∴ Retirement benefits corresponding to its position

$$= 39.5 + \frac{(153.2 - 119)}{70} \times 10$$

$$= 39.5 + 4.88$$

$$= 44.38$$

$$= £\ 44380$$

The 10^{th} - 90^{th} percentile range

This is a measure of dispersion which uses percentile. A percentile is a value which separates one division from the other when a given data is divided into 100 equal divisions.
This measure of dispersion is very important when calculating the co-efficient of skewness.

Example

Using the above data for retirees calculate the 10^{th} - 90^{th} percentile. The tenth percentile 10^{th} percentile lies on position

$$\frac{10}{100}(382 + 1) = 0.1 \times 383$$

$$= 38.3$$

∴ the value corresponding to the tenth percentile

$$= 19.5 + \frac{(38.3 \times 10)}{50}$$

$$= 19.5 + 7.66$$

$$= 27.16$$

The 90^{th} percentile lies on position

$$\frac{90}{100}(382 + 1) = 0.9 \times 383$$

= 344.7

∴ the value corresponding to the 90^{th} percentile

$$= 69.5 + \frac{(344.7 - 331)}{40} \times 10$$

$$= 69.5 + 3.425$$

$$= 72.925$$

∴ the required value of the 10^{th} – 90^{th} percentile = 72.925 – 27.16 = 45.765

Merits of quartile deviation

i) In certain respect it is superior to range as a measure of variation
ii) It has a special utility in measuring variation in cases of open – ended distribution
iii) It has also useful in erratic or highly skewed distribution where the other measure of variation would be warped by extreme values. It is not affected by the presence of extreme values.

Limitation

i) Quartile deviation ignores 50% item as the value of quartile deviation does not depend upon every observation it cannot be regarded as a god method of measuring variation.
ii) It is not capable of mathematical manipulation
iii) Its value sis very much affected by sampling fluctuation
iv) It is in fact not a measure of variation as it really does not show the scatter around an average but rather a distance on scale.

(d) Standard deviation

The standard deviation measures the absolute dispersion (or variability of distribution, the greater amount of dispersion or variability). The greater the standard deviation the greater will be the magnitude of the deviations of the value from their mean.

A small standard deviation means a high degree of uniformity of the observation as well as homogeneity of a series.

A large standard deviation means just the opposite.

Difference between deviation and standard deviation

Both these measure of dispersion are based on each and very item of the distribution. But they differ in the following respects.

- Algebraic signs are ignores while calculating mean deviation whereas in the calculation of standard deviation signs are taken into account.

- Mean deviation can be computed either from median or mean. The standard deviation on the other hand is always computed from the arithmetic mean because the sum of the squares of the deviation of items from arithmetic mean is least.

This is one of the most accurate measures of dispersion. It has the following advantages;

i. It utilizes all the values given
ii. It makes use of both negative and positive values if they occur
iii. The standard deviation reflects an accurate impression of how much the sample data varies from the mean. This is because its suitability can also be tested using other statistical methods

Example

A sample comprises of the following observations; 14, 18, 17, 16, 25, 31
Determine the standard deviation of this sample
Observation.

	x	$\left(x-\bar{x}\right)$	$\left(x-\bar{x}\right)^2$
	14	-6.1	37.21
	18	-2.1	4.41
	17	-3.1	9.61
	16	-4.1	16.81
	25	4.9	24.01
	31	10.9	118.81
Total	**121**		**210.56**

$\overline{X} = \frac{121}{6} = 20.1$

$\therefore$ Standard deviation, $\sigma = \sqrt{\frac{\Sigma\left(x-\bar{x}\right)^2}{n}} = \sqrt{\frac{210.56}{6}}$

= 5.93

Alternative method

	x	X^2
	14	196
	18	324
	17	289
	16	256

	25	625
	31	961
Total	**121**	**2651**

$$\sigma = \sqrt{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2} = \sqrt{\frac{2651}{6} - \left(\frac{121}{6}\right)^2}$$

= 5.93

Example 2

The following table shows the part-time rate per hour of a given no. of laborers in the month of June 1997.

Rate per hr (x) Shs	No. of laborers (f)	fx	fx^2
230	7	1610	370300
400	6	2400	960000
350	2	700	245000
450	1	450	202500
200	8	1600	320000
150	11	1650	247500
Total	**35**	**8410**	**2345300**

Calculate the standard deviation from the above table showing how the hourly payment were varying from the respective mean

$\therefore$ Standard deviation, $\sigma = \sqrt{\frac{\sum fx^2}{\sum f} - \left(\frac{\sum fx}{\sum f}\right)^2}$

$$= \sqrt{\frac{2345300}{35} - \left(\frac{8410}{35}\right)^2}$$

$$= \sqrt{67008.6 - 577372}$$

$$= \sqrt{9271.4}$$

= 96.29

Example 3 – Grouped data

In business statistical work we usually encounter a set of grouped data. In order to determine the standard deviation from such data, we use any of the three following methods

i. The long method
ii. The shorter method
iii. The coded method

The above methods are used in the following examples

Example 3.1

The quality controller in a given firm had an accurate record of all the iron bars produced in May 1997. The following data shows those records

i. Using long method

Bar lengths (cm)	**No. of bars(f)**	**Class mid point (x)**	**fx**	**fx^2**
201 – 250	25	225.5	5637.5	1271256.25
251 – 300	36	275.5	9918	2732409
301 – 350	49	325.5	15949.5	5191562.25
351 – 400	80	375.5	30040	11280020
401 – 450	51	425.5	21700.5	9233562.75
451 – 500	42	475.5	19971	9496210.50
501 - 550	30	525.5	15765	8284507.50
	313		118981.50	47489526

Calculate the standard deviation of the lengths of the bars

$$\therefore \text{Standard deviation, } \sigma = \sqrt{\frac{\sum fx^2}{\sum f} - \left(\frac{\sum fx}{\sum f}\right)^2}$$

$$= \sqrt{\frac{47489526}{313} - \left(\frac{118981.50}{313}\right)^2}$$

$$= 84.99 \text{ cm}$$

ii. Using the shorter method

Bar lengths (cm)	No. of bars(f)	mid point (x)	x-A = d	fd	Fd^2
201 – 250	25	225.5	-150	-3750	562500
251 – 300	36	275.5	-100	-3600	360000
301 – 350	49	325.5	-50	-2450	122500
351 – 400	80	375.5 (A)	0	0	0

401 – 450	51	425.5	50	2550	127500
451 – 500	42	475.5	100	4200	420000
501 - 550	30	525.5	150	4500	675000
Total	**313**			**1450**	**2267500**

Calculate the standard deviation using the shorter method quagmire

∴ Standard deviation, σ $= \sqrt{\frac{\sum fd^2}{\sum f} - \left(\frac{\sum fd}{\sum f}\right)^2}$

$= \sqrt{\frac{2267500}{313} - \left(\frac{1450}{313}\right)^2}$

$= \sqrt{7244.40 - 21.50}$

$= \sqrt{7222.90}$

= 84.99 cm

iii. Using coded method

Bar lengths (cm)	(f)	mid point (x)	x-A = d	d/c = u	fu	fu^2
201 – 250	25	225.5	-150	-3	-75	225
251 – 300	36	275.5	-100	-2	-72	144
301 – 350	49	325.5	-50	-1	-49	49
351 – 400	80	375.5 (A)	0	0	0	0
401 – 450	51	425.5	50	1	51	51
451 – 500	42	475.5	100	2	84	168
501 - 550	30	525.5	150	3	90	270
	313				29	907

C = 50 *where c is an arbitrary number, try picking a different figure say 45 the answer should be the same.*

Standard deviation using the coded method. This is the most preferable method among the three methods

$$\sigma = c \times \sqrt{\frac{\sum fu^2}{\sum f} - \left(\frac{\sum fu}{\sum f}\right)^2}$$

$$= 50 \times \sqrt{\frac{907}{313} - \left(\frac{29}{313}\right)^2}$$

= 50 × 1.6997

= 84.99

Variance

Square of the standard deviation is called variance.

(e) Lorenz curve

The Lorenz curve derived by Max O. Lorenz, famous economic statistician, is a graphic method of studying dispersion. This curve was used by him for the first time to measure the distribution off wealth and income. Now the curve is also used to study the distribution of profits, wages, turnover, etc.

Illustration

In the following table is given the number of companies belonging to two areas A and B according to the amount of profits earned by them. Draw in the same diagram their Lorenz curve and interpret them.

Profits earned shs (000s)	Number of companies	
	Area A	Area B
6	6	2
25	11	38
60	13	52
84	14	28
105	15	38
150	17	26
170	10	12
400	14	4

Solution

Relative measures of dispersion

Def: A relative measure of dispersion is a statistical value which may be used to compare variations in 2 or more samples.

The measures of dispersion are usually expressed as decimals or percentages and usually they do not have any other units

Example

The average distance covered by vehicles in a motor rally may be given as 2000 km with a standard deviation of 5 km.
In another competition set of vehicles covered 3000 km with a standard deviation of 10 kms
NB: The 2 standard deviations given above are referred to as absolute measures of dispersion. These are actual deviations of the measurements from their respective mean
However, these are not very useful when comparing dispersions among samples.
Therefore the following measures of dispersion are usually employed in order to assess the degree of dispersion.

i. Coefficient of mean deviation

$$= \frac{\text{Mean deviation}}{\text{mean}}$$

ii. Coefficient of quartile deviation

$$= \frac{1/2(Q_3 - Q_1)}{Q_2}$$

Where Q_1 = first quartile

Q_3 = third quartile

iii. Coefficient of standard deviation

$$= \frac{\text{Standard deviation}}{\text{mean}}$$

iv. Coefficient of variation

$$= \frac{\text{standard deviation}}{\text{mean}} \times 100$$

Example (see information above)

First group of cars: mean = 2000 kms

Standard deviation = 5 kms

$\therefore$ C.O.V = $\frac{5}{2000} \times 100$

= 0.25%

Second group of cars: mean = 3000 kms

Standard deviation = 10kms

$\therefore$ C.O.V = $\frac{10}{3000} \times 100$

= 0.33%

Conclusion

Since the coefficient of variation is greater in the 2nd group, than in the first group we may conclude that the distances covered in the 1st group are much closer to the mean that in the 2nd group.

Example 2

In a given farm located in the UK the average salary of the employees is £ 3500 with a standard deviation of £150
The same firm has a local branch in Kenya in which the average salaries are Kshs 8500 with a standard deviation of Kshs.800
Determine the coefficient of variation in the 2 firms and briefly comment on the degree of dispersion of the salaries in the 2 firms.
First firm in the UK

C.O.V = $\frac{150}{3500} \times 100$

= 4.29%

Second firm in Kenya

C.O.V = $\frac{800}{} \times 100$

$$\frac{8500}{}$$

$$= 9.4\%$$

Conclusively, since 4.29% < 9.4% then the salaries offered by the firm in UK are much closer to the mean given them in the case to the local branch in Kenya

COMBINED MEAN AND STANDARD DEVIATION

Sometimes we may need to combine 2 or more samples say A and B. It is therefore essential to know the new mean and the new standard deviation of the combination of the samples.

Combined mean
Let m be the combined mean
Let x_1 be the mean of first sample
Let x_2 be the mean of the second sample
Let n_1 be the size of the 1st sample
Let n_2 be the size of the 2nd sample
Let s_1 be the standard deviation of the 1st sample
Let s_2 be the standard deviation of the 2nd sample

$$\therefore \text{ combined mean } = \frac{n_1x_1 + n_2x_2}{n_1 + n_2}$$

$$\text{combined standard deviation } = \sqrt{\frac{n_1s_1^2 + n_1\left(m - x_1\right)^2 + n_2s_2^2 + n_2\left(m - x_2\right)^2}{n_1 + n_2}}$$

Example

A sample of 40 electric batteries gives a mean life span of 600 hrs with a standard deviation of 20 hours.
Another sample of 50 electric batteries gives a mean lifespan of 520 hours with a standard deviation of 30 hours.
If these two samples were combined and used in a given project simultaneously, determine the combined new mean for the larger sample and hence determine the combined or pulled standard deviation.

Size	x	s
40(n_1)	600 hrs(x_1)	20hrs (s_1)
50 (n_1)	520 hrs (x_2)	30 hrs (s_2)

$$\textit{Combined mean} \quad = \frac{40(600)+50(520)}{40+50} \quad = \frac{50,000}{90} = 555.56$$

Combined standard deviation

$$= \sqrt{\frac{40(20^2)+40(555.56-660)^2+50(30)^2+50(555.56-520)^2}{40+50}}$$

$$= \sqrt{\frac{1600+78996.54+45000+63225.68}{90}}$$

$$= 47.52\ hrs$$

SKEWNESS

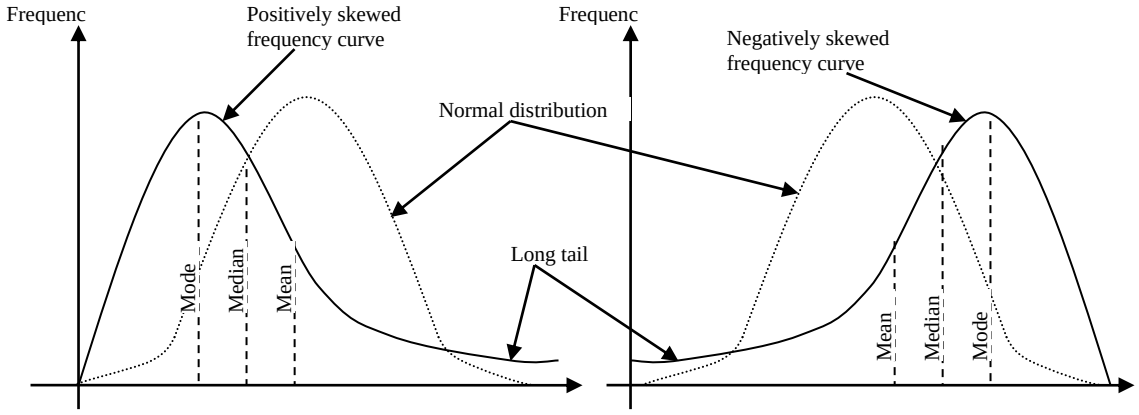
- This is a concept which is commonly used in statistical decision making. It refers to the degree in which a given frequency curve is deviating away from the normal distribution
  - There are 2 types of skew ness namely
    - i. Positive skew ness
    - ii. Negative skew ness

1. Positive Skewness
   - This is the tendency of a given frequency curve leaning towards the left. In a positively skewed distribution, the long tail extended to the right.

In this distribution one should note the following

i. The mean is usually bigger than the mode and median
ii. The median always occurs between the mode and mean
iii. There are more observations below the mean than above the mean

This frequency distribution as represented in the skewed distribution curve is characteristic of the age distributions in the developing countries



2. Negative Skewness

This is an asymmetrical curve in which the long tail extends to the left

NB: This frequency curve for the age distribution is characteristic of the age distribution in developed countries

- The mode is usually bigger than the mean and median
- The median usually occurs in between the mean and mode
- The no. of observations above the mean are usually more than those below the mean (see the shaded region)

MEASURES OF SKEWNESS

- These are numerical values which assist in evaluating the degree of deviation of a frequency distribution from the normal distribution.

- Following are the commonly used measures of skew ness.

1. Coefficient Skewness

$$= 3 \times \frac{(\text{mean} - \text{median})}{\text{Standard deviation}}$$

2. Coefficient of skewness

$$= \frac{\text{mean} - \text{mode}}{\text{Standard deviation}}$$

NB: These 2 coefficients above are also known as Pearsonian measures of skewness.

3. Quartile Coefficient of skewness

$$= \frac{Q3 + Q1 - 2Q2}{Q3 + Q1}$$

Where Q1 = 1st quartile

Q2 = 2nd quartile

Q3 = 3rd quartile

NB: The Pearsonian coefficients of skewness usually range between –ve 3 and +ve 3. These are extreme value i.e. +ve 3 and –ve 3 which therefore indicate that a given frequency is negatively skewed and the amount of skewness is quite high.

Similarly if the coefficient of skewness is +ve it can be concluded that the amount of skew ness of deviation from the normal distribution is quite high and also the degree of frequency distribution is positively skewed.

Example

The following information was obtained from an NGO which was giving small loans to some small scale business enterprises in 1996. the loans are in the form of thousands of Kshs.

Loans	Units (f)	Midpoints(x)	x-a=d	d/c= u	fu	Fu^2	UCB	cf
46 – 50	32	48	-15	-3	-96	288	50.5	32
51 – 55	62	53	-10	-2	-124	248	55.5	94
56 – 60	97	58	-5	-1	-97	97	60.5	191
61 –65	120	63 (A)	0	0	0	0	0	0
66 –70	92	68	5	+1	92	92	70.5	403
71 –75	83	73	10	+2	166	332	75.5	486
76 – 80	52	78	15	+3	156	468	80.5	538
81 – 85	40	83	20	+4	160	640	85.5	57.8
86 – 90	21	88	25	+5	105	525	90.5	599
91 – 95	11	93	30	+6	66	396	95.5	610
Total	**610**				**428**	**3086**		

Required

Using the Pearsonian measure of skew ness, calculate the coefficients of skew ness and hence comment briefly on the nature of the distribution of the loans.

Arithmetic mean = Assumed mean + $\frac{c(\sum fu)}{\sum f}$

$= 63 + \frac{(428 \times 5)}{610}$

$= 66.51$

It is very important to note that the method of obtaining arithmetic mean (or any other statistic) by misusing assumed mean (A) from X and then dividing by c can be abit confusing, if this is the case then just use the straight forward method of:

$Arithmetic\ mean = \frac{\sum f.x}{\sum f}$ *where x* is the midpoint, *the answers are the same.*

The standard deviation $= c \times \sqrt{\frac{\sum fu^2}{\sum f} - \left(\frac{\sum fu}{\sum f}\right)^2}$

$$=5 \times \sqrt{\frac{3086}{610} - \left(\frac{428}{610}\right)^2}$$

$$= 10.68$$

The Position of the median lies m = $\frac{n+1}{2}$

$$= \frac{610+1}{2} = 305.5$$

$$= 60.5 + \frac{(305.5 - 191)}{120} \times 5$$

$$= 60.5 + \frac{(114.4)}{120} \times 5$$

Median = 65.27

Therefore the Pearsonian coefficient

$$= 3\frac{(66.51 - 65.27)}{10.68}$$

$$= 0.348$$

Comment

The coefficient of skewness obtained suggests that the frequency distribution of the loans given was positively skewed

This is because the coefficient itself is positive. But the skewness is not very high implying the degree of deviation of the frequency distribution from the normal distribution is small

Example 2

Using the above data calculate the quartile coefficient of skewness

Quartile coefficient of skewness = $\frac{Q3 + Q1 - 2Q2}{Q3 + Q1}$

The position of Q1 lies on = $\frac{610+1}{4} = 152.75$

$\therefore$ actual value Q1 =55. 5 + $\frac{(152.75-94)}{97} \times 5 = 58.53$

The position of Q3 lies on = $3\frac{(610+1)}{4} = 458.25$

$\therefore$ actual value Q_3 =70.55 + $\frac{(458.25-403)}{83} \times 5 = 73.83 \times 5$

Q_2 position: i.e. $2\frac{(610+1)}{4}$ = 305.5

Actual Q_2 value $= 60.5 + \frac{(305.5-191)}{120} \times 5 = 65.27$

The required coefficient of skew ness

$$= \frac{73.83+58.53-2(65.27)}{73.83+58.53} = 0.013$$

Conclusion

Same as above when the Pearsonian coefficient was used

KURTOSIS

- This is a concept, which refers to the degree of peaked ness of a given frequency distribution. The degree is normally measured with reference to normal distribution.
- The concept of kurtosis is very useful in decision making processes i.e. if is a frequency distribution happens to have either a higher peak or a lower peak, then it should not be used to make statistical inferences.
- Generally there are 3 types of kurtosis namely;-
  i. Leptokurtic
  ii. Mesokurtic
  iii. Platykurtic

Leptokurtic

a) A frequency distribution which is lepkurtic has generally a higher peak than that of the normal distribution. The coefficient of kurtosis when determined will be found to be more than 3. thus frequency distributions with a value of more than 3 are definitely leptokurtic

b) Some frequency distributions when plotted may produce a curve similar to that of the normal distribution. Such frequency distributions are referred to as mesokurtic. The degree of kurtosis is usually equal to 3

c) When the frequency curve contacted produces a peak which is lower that that of a normal distribution when such a curve is said to be platykurtic. The coefficient of such is usually less than 3

- It is necessary to calculate the numerical measure of kurtosis. The commonly used measure of kurtosis is the percentile coefficient of kurtosis. This coefficient is normally determined using the following equation

Percentile measure of kurtosis, K (Kappa) = $\frac{1}{2}\frac{(Q3 - Q1)}{P90 - P10}$

Example

Refer to the table above for loans to small business firms/units

Required

Calculate the percentile coefficient of Kurtosis

P90 = $\frac{90}{100}(n+1) = 0.9(610+1)$

= 0.9 (611)

= 549.9

The actual loan for a firm in this position

(549.9) = 80.5 + $\frac{(549.9 - 538)}{40}$ x 5 = 81.99

P10 = $\frac{10}{100}$ (n + 1) = 0.1 (611) = 61.1

The actual loan value given to the firm on this position is

$$50.5 + \frac{(61.1-32)}{62} \times 5 = 52.85$$

$= 0.9\ (611)$

$= 549.9$

∴ Percentile measure of kurtosis

$$K \text{ (Kappa)} = \frac{1}{2}\frac{(Q3 - Q1)}{P90 - P10}$$

$$= \frac{1}{2}\frac{(73.83 - 58.53)}{81.99 - 52.85}$$

$$= 0.26$$

Since 0.26 < 3, it can be concluded that the frequency distribution exhibited by the distribution of loans is platykurtic

Kurtosis is also measured by moment statistics, which utilize the exact value of each observation.

i. M1 the first moment = M1 = $\frac{\sum X}{n}$ = Mean M1 or M_1

$$M2 = \frac{\sum X^2}{n}$$

$$M3 = \frac{\sum X^3}{n}$$

$$M4 = \frac{\sum X^4}{n}$$

3. M2 second moment about the mean M2 or f^2
   $M2 = M_2 - M1^2$
4. M3 third moment about the mean M3 (a measure of the absolute skew ness)
   $M3 = M_3 - 3M_2M_1 + 2M1^3$
5. M4 fourth moment about the mean M4 (a measure of the absolute Kurtosis)
   $M4 = M_4 - 4M_3M_1 + 6M_2M_{12} + 3M_14$

An alternative formula

$$M4 = \frac{\sum (x-m)^4 f}{\sum f}$$ Where m is mean