INTRODUCTION

In the modern world of information and communication technology, the importance of statistics is very well recognised by all the disciplines. Statistics has originated as a science of statehood and found applications slowly and steadily in Agriculture, Economics, Commerce, Biology, Medicine, Industry, planning, education and so on. As of today, there is no other human walk of life, where statistics cannot be applied.

Statistics is concerned with the scientific method of collecting, organizing, summarizing, presenting and analyzing statistical information (data) as well as drawing valid conclusion on the basis of such analysis. It could be simply defined as the "science of data". Thus, statistics uses facts or numerical data, assembled, classified and tabulated so as to present significant information about a given subject. Statistic is a science of understanding data and making decisions in the face of randomness.

The study of statistics is therefore essential for sound reasoning, precise judgment and objective decision in the face of up- to- date accurate and reliable data. Thus many researchers, educationalists, business men and government agencies at the national, state or local levels rely on data to answer operations and programs. Statistics is usually divided into two categories, which is not mutually elution namely: Descriptive statistics and inferential statistics.

DESCRIPTIVE STATISTICS

This is the act of summarizing and given a descriptive account of numerical information in form of reports, charts and diagrams. The goal of descriptive statistics is to gain information from collected data. It begins with collection of data by either counting or measurement in an inquiry. It involves the summary of specific aspect of the data, such as averages values and measure of dispersion (spread). Suitable graphs, diagrams and charts are then used to gain understanding and clear interpretation of the phenomenon under investigation

keeping firmly in mind where the data comes from. Normally, a descriptive statistics should:

i. be single – valued
ii. be algebraically tractable
iii. consider every observed value.

INFERENTIAL STATISTICS

This is the act of making deductive statement about a population from the quantities computed from its representative sample. It is a process of making inference or generalizing about the population under certain conditions and assumptions. Statistical inference involves the processes of estimation of parameters and hypothesis testing.

SOURCES OF STATISTICAL DATA

1. **Primary data**: These are data generated by first hand or data obtained directly from respondents by personal interview, questionnaire, measurements or observation. Statistical data can be obtained from:
   (i) Census – complete enumeration of all the unit of the population
   (ii) Surveys – the study of representative part of a population
   (iii) Experimentation – observation from experiment carried out in laboratories and research center.
   (iv) Administrative process e.g. Record of births and deaths.

   ADVANTAGES

   - ✓ Comprises of actual data needed
   - ✓ It is more reliable with clarity
   - ✓ Comprises a more detail information

   DISADVANTAGES

   - ➢ Cost of data collection is high
   - ➢ Time consuming
   - ➢ There may larger range of non response

2. **Secondary data**: These are data obtained from publication, newspapers, and annual reports. They are usually summarized data used for purpose other than the intended one. These could be obtain from the following:

   (i) Publication e.g. extract from publications

   (ii) Research/Media organization

   (iii) Educational institutions

ADVANTAGES

- ✓ The outcome is timely
- ✓ The information gathered more quickly
- ✓ It is less expensive to gather.

DISADVANTAGES

- ➢ Most time information are suppressed when working with secondary data
- ➢ The information may not be reliable

METHODS OF COLLECTION OF DATA

There are various methods we can use to collect data. The method used depends on the problem and type of data to be collected. Some of these methods include:

1. Direct observation
2. Interviewing
3. Questionnaire
4. Abstraction from published statistics.

DIRECT OBSERVATION

Observational methods are used mostly in scientific enquiry where data are observed directly from controlled experiment. It is used more in the natural

sciences through laboratory works than in social sciences. But this is very useful studying small communities and institutions.

INTERVIEWING

In this method, the person collecting the data is called the interviewer goes to ask the person (interviewee) direct questions. The interviewer has to go to the interviewees personally to collect the information required verbally. This makes it different from the next method called questionnaire method.

QUESTIONNAIRE

A set of questions or statement is assembled to get information on a variable (or a set of variable). The entire package of questions or statement is called a questionnaire. Human beings usually are required to respond to the questions or statements on the questionnaire. Copies of the questionnaire can be administered personally by its user or sent to people by post. Both interviewing and questionnaire methods are used in the social sciences where human population is mostly involved.

ABSTRACTIONS FROM THE PUPLISHED STATISTICS

These are pieces of data (information) found in published materials such as figures related to population or accident figures. This method of collecting data could be useful as preliminary to other methods.

Other methods includes: Telephone method, Document/Report method, Mail or Postal questionnaire, On-line interview method, etc.

PRESENTATION OF DATA

When raw data are collected, they are organized, numerically by distributing them into classes or categories in order to determine the number of individuals belonging to each class. Most cases, it is necessary to present data in tables, charts and diagrams in order to have a clear understanding of the data, and to illustrate the relationship existing between the variables being examined.

FREQUENCY TABLE

This is a tabular arrangement of data into various classes together with their corresponding frequencies.

Procedure for forming frequency distribution

Given a set of observation $x_1, x_2, x_3, \ldots, x_n$, for a single variable.

1. Determine the range (R) = L – S where L = largest observation in the raw data; and S = smallest observation in the raw data.
2. Determine the appropriate number of classes or groups (K). The choice of K is arbitrary but as a general rule, it should be a number (integer) between 5 and 20 depending on the size of the data given. There are several suggested guide lines aimed at helping one decided on how many class intervals to employ. Two of such methods are:
   (a) $K = 1 + 3.322 (\log_{10} n)$
   (b) $K = \sqrt{n}$ where n = number of observations.
3. Determine the width (w) of the class interval. It is determined as $w = \frac{R}{K}$
4. Determine the numbers of observations falling into each class interval i.e. find the class frequencies.

NOTE: With advent of computers, all these steps can be accomplishes easily.

SOME BASIC DEFINITIONS

Variable: This is a characteristic of a population which can take different values. Basically, we have two types, namely: continuous variable and discrete variable.

A **continuous variable** is a variable which may take all values within a given range. Its values are obtained by measurements e.g. height, volume, time, exam score etc.

A **discrete variable** is one whose value change by steps. Its value may be obtained by counting. It normally takes integer values e.g. number of cars, number of chairs.

Class interval: This is a sub-division of the total range of values which a (continuous) variable may take. It is a symbol defining a class E.g. 0-9, 10-19 etc. there are three types of class interval, namely: Exclusive, inclusive and open-end classes method.

Exclusive method:

When the class intervals are so fixed that the upper limit of one class is the lower limit of the next class; it is known as the exclusive method of classification. E.g. Let some expenditures of some families be as follows:

0 – 1000, 1000 – 2000, etc. It is clear that the exclusive method ensures continuity of data as much as the upper limit of one class is the lower limit of the next class. In the above example, there are so families whose expenditure is between 0 and 999.99. A family whose expenditure is 1000 would be included in the class interval

1000-2000.

Inclusive method:

In this method, the overlapping of the class intervals is avoided. Both the lower and upper limits are included in the class interval. This type of classification may be used for a grouped frequency distribution for discrete variable like members in a family, number of workers in a factory etc., where the variable may take only integral values. It cannot be used with fractional values like age, height, weight

etc. *In case of continuous variables, the exclusive method should be used. The inclusive method should be used in case of discrete variable.*

Open end classes:

A class limit is missing either at the lower end of the first class interval or at the upper end of the last class interval or both are not specified. The necessity of open end classes arises in a number of practical situations, particularly relating to economic and medical data when there are few very high values or few very low values which are far apart from the majority of observations.

Class limit: it represents the end points of a class interval. {Lower class limit & Upper class limit}. A class interval which has neither upper class limit nor lower class limit indicated is called an open class interval e.g. "less than 25", '25 and above"

Class boundaries: The point of demarcation between a class interval and the next class interval is called boundary. For example, the class boundary of 10-19 is 9.5 – 19.5

Cumulative frequency: This is the sum of a frequency of the particular class to the frequencies of the class before it.

Example 1: The following are the marks of 50 students in STS 102:

48 70 60 47 51 55 59 63 68 63 47 53 72 53 67 62 64 70 57 56 48 51 58 63 65 62 49 64 53 59 63 50 61 67 72 56 64 66 49 52 62 71 58 53 63 69 59 64 73 56.

(a) Construct a frequency table for the above data.

(b) Answer the following questions using the table obtained:

(i) how many students scored between 51 and 62?

(ii) how many students scored above 50?

(iii) what is the probability that a student selected at random from the class will score less than 63?

Solution:

(a) Range (R) = $73 - 47 = 26$

No of classes $(k) = \sqrt{n} = \sqrt{50} = 7.07 \approx 7$

Class size $(w) = {}^{26}/_{7} = 3.7 \approx 4$

Frequency Table

Mark	Tally	frequency
47 – 50	𝍸 ||	7
51- 54	𝍸 ||	7
55 - 58	𝍸 ||	7
59 – 62	𝍸 |||	8
63 – 66	𝍸 𝍸 |	11
67 – 70	𝍸 |	6
71 – 74	||||	4
		$\sum f = 50$

(b) (i) 22 (ii) 43 (iii) 0.58

Example 2: The following data represent the ages (in years) of people living in a housing estate in Abeokuta.

18 31 30 6 16 17 18 43 2 8 32 33 9 18 33 19 21 13 13 14
14 6 52 45 61 23 26 15 14 15 14 27 36 19 37 11 12 11
20 12 39 20 40 69 63 29 64 27 15 28.

Present the above data in a frequency table showing the following columns; class interval, class boundary, class mark (mid-point), tally, frequency and cumulative frequency in that order.

Solution:

Range (R)= $69 - 2 = 67$

No of classes $(k) = \sqrt{n} = \sqrt{50} = 7.07 \approx 7.00$

Class width $(w) = {}^{R}/_{k} = {}^{67}/_{7} = 9.5 \approx 10$

Class interval	Class boundary	Class mark	Tally	Frequency	Cum.freq
2 – 11	1.5 – 11.5	6.5	𝍸 |	7	7
12 – 21	11.5 – 21.5	16.5	𝍸 𝍸 𝍸 𝍸 |	21	28
22 – 31	21.5 – 31.5	26.5	𝍸 |||	8	36
32 – 41	31.5 – 41.5	36.6	𝍸 ||	7	43
42 – 51	41.5 – 51.5	46.5	||	2	45
52 – 61	51.5 – 61.5	56.5	||	2	47
62 – 71	61.5 – 71.5	66.5	|||	3	50

Observation from the Table

The data have been summarized and we now have a clearer picture of the distribution of the ages of inhabitants of the Estate.

Exercise 1

Below are the data of weights of 40students women randomly selected in Ogun state. Prepare a table showing the following columns; class interval, frequency, class boundary, class mark, and cumulative frequency.

96 84 75 80 64 105 87 62 105 101 108 106 110 64 105 117
103 76 93 75 110 88 97 69 94 117 99 114 88 60 98 77
96 96 91 73 82 81 91 84

Use your table to answer the following question

i. How many women weight between 71 and 90?

ii. How many women weight more than 80?

iii. What is the probability that a woman selected at random from Ogun state would weight more than 90?

(10)

MEASURES OF LOCATION

These are measures of the centre of a distribution. They are single values that give a description of the data. They are also referred to as measure of central tendency. Some of them are arithmetic mean, geometric mean, harmonic mean, mode, and median.

THE ARITHMETIC MEAN (A.M)

The arithmetic mean (average) of set of observation is the sum of the observation divided by the number of observation. Given a set of a numbers $x_1, x_2 \ldots, x_n$, the arithmetic mean denoted by $\bar{X}$ is defined by

$$\bar{X} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} = \sum_{i=1}^{n} \frac{x_i}{n}$$

Example 1: The ages of ten students in STS 102 are 16,20,19,21,18,20,17,22,20,17, determine the mean age.

Solution: $\bar{X} = \sum_{i=1}^{n} \frac{x_i}{n}$

$$= \frac{16+20+19+21+18+20+17+22+20+17}{10}$$

$$= \frac{190}{10} = 19 years.$$

If the numbers $x_1, x_2, \ldots, x_n$ occur $f_1, f_2, f_3, \ldots f_n$ times respectively, the

$\bar{X} = \frac{f_1x_1+f_2x_2+\cdots+f_nx_n}{f_1+f_2+\cdots+f_n} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$ (or $\frac{\Sigma fx}{n}$ for short.)

Example 2: Find the mean for the table below

Scores (x)	2	5	6	8
Frequency (f)	1	3	4	2

Solution $\bar{x} = \frac{\Sigma fx}{\Sigma f} = \frac{(1\times2)+(3\times5)+(4\times6)+(2\times8)}{1+3+4+2}$

$$= \frac{57}{10} = 5.7$$

Calculation of mean from grouped data

If the items of a frequency distribution are classified in intervals, we make the assumption that every item in an interval has the mid-values of the interval and we use this midpoint for x.

Example 3: The table below shows the distribution of the waiting items for some customers in a certain petrol station in Abeokuta.

Waiting time(in mins)	1.5 – 1.9	2.0 – 2.4	2.5 – 2.9	3.0 – 3.4	3.5 – 3.9	4.0 – 4.4
No. of customers	3	10	18	10	7	2

Find the average waiting time of the customers.

Solution:

Waiting (in min)	No of customers	Class mark mid-value(X)	fx
1.5 – 1.9	3	1.7	5.1
2.0 – 2.4	10	2.2	22
2.5 – 2.9	18	2.7	48.6
3.0 – 3.4	10	3.2	32
3.5 – 3.9	7	3.7	25.9
4.0 – 4.4	2	4.2	8.4
	$\sum f = 50$		$\sum fx = 142$

$$\bar{X} = \frac{\sum fx}{\sum f}$$

$$= \frac{142}{50} = 2.84$$

Use of Assume mean

Sometimes, large values of the variable are involve in calculation of mean, in order to make our computation easier, we may assume one of the values as the mean. This if A= assumed mean, and d= deviation of x from A, i.e. $d = x - A$

Example 5: Consider the data in example 3, using a suitable assume mean, compute the mean.

Solution:

Waiting time	f	x	$d = x - A$	fd
1.5 – 1.9	3	1.7	-1	-3
2.0 – 2.4	10	2.2	-0.5	-5
2.5 – 2.9	18	2.7 A	0	0
3.0 – 3.4	10	3.2	0.5	5
3.5 – 3.9	7	3.7	1	7
4.0 – 4.4	2	4.2	1.5	3
	50			7

$$\bar{X} = A + \frac{\sum fd}{\sum f}$$

$$= 2.7 + \frac{7}{50}$$

$$= 2.7 + 0.14$$

$$= 2.84$$

NOTE: It is always easier to select the class mark with the longest frequency as the assumed mean.

ADVANTAGE OF MEAN

The mean is an average that considers all the observations in the data set. It is single and easy to compute and it is the most widely used average.

DISAVANTAGE OF MEAN

Its value is greatly affected by the extremely too large or too small observation.

THE HARMONIC MEAN (H.M)

The H.M of a set of numbers $x_1, x_2, \ldots, x_n$ is the reciprocal of the arithmetic mean of the reciprocals of the numbers. It is used when dealing with the rates of the type x per d (such as kilometers per hour, Naira per liter). The formula is expressed thus:

$$H.M = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}} = \frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}}$$

If x has frequency f, then

$$H.M = \frac{n}{\sum\frac{f}{x}}$$

Example: Find the harmonic mean of 2,4,8,11,4.

Solution:

$$H.M = \frac{5}{\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\frac{1}{11}+\frac{1}{4}} = \frac{5}{\frac{107}{80}} = 4\frac{12}{107} = 4.112\,.$$

Note:

(i) Calculation takes into account every value

(ii) Extreme values have least effect

(iii) The formula breaks down when "o" is one of the observations.

THE GEOMETRIC MEAN(G.M)

The G.M is an analytical method of finding the average rate of growth or decline in the values of an item over a particular period of time. The geometric mean of a set of number $x_1, x_2, \ldots, x_n$ is the nth root of the product of the number. Thus

$$G.M = \sqrt[n]{(x_1 \times x_2 \times \ldots \times x_n)}$$

If f_i is the frequency of x_i, then

$$G.M = \sqrt[n]{(x_1 f_1 \times x_2 f_2 \times \ldots \times x_i f_i)}$$

Example: The rate of inflation in ~~five~~ five successive year in a country was 5%, 8%, 12%, 25% and 34%. What was the average rate of inflation per year?

Solution:

$$G.M = \sqrt[5]{(1.05) \times (1.08) \times (1.12) \times (1.25) \times (1.34)}$$

$$= \sqrt[5]{2.127384}$$

$$= 1.16$$

∴ Average rate of inflation is 16%

Note: (1) Calculate takes into account every value.

(2) It cannot be computed when "0" is on of the observation.

Relation between Arithmetic mean, Geometric and Harmonic

In general, the geometric mean for a set of data is always less than or equal to the corresponding arithmetic mean but greater than or equal to the harmonic mean.

That is, $H.M \leq G.M \leq A.M$

The equality signs hold only if all the observations are identical.

THE MEDIAN

This is the value of the variable that divides a distribution into two equal parts when the values are arranged in order of magnitude. If there are n (odd) observation, the median $\tilde{X}$ is the center of observation in the ordered list. The location of the median is $\tilde{X} = \frac{(n+1)}{2}$ th item.

But if n is even, the median $\tilde{X}$ is the average of the two middle observations in the ordered list.

i.e. $$\tilde{X} = \frac{X_{\left(\frac{n}{2}\right)th} + X_{\left(\frac{n}{2}+1\right)th}}{2}$$

Example 1: The values of a random variable x are given as 8, 5, 9, 12, 10, 6 and 4. Find the median.

Solution: In an array: 4,5,6,8,9,10,12. n is odd, therefore

The median, $\tilde{X} = X_{\left(\frac{n+1}{2}\right)^{th}}$

$= X_{4^{th}}$

$= 8$

Example 2: The value 0f a random variable x are given as 15, 15, 17, 19, 21, 22, 25, and 28. Find the median.

Solution: n is ~~odd.~~ even

Median, $\tilde{X} = X_{\left(\frac{n}{2}\right)^{th}} + X_{\left(\frac{n}{2}+1\right)^{th}}$

$= \frac{X_4 + X_5}{2}$

$= \frac{19+21}{2}$

$= 20$

Calculation of Median from a grouped data

The formula for calculating the median from grouped data is defined as

$$\tilde{X} = L_1 + \left(\frac{\frac{n}{2} - Cf_b}{f_m}\right) w$$

Where: L_1 = Lower class boundary of the median class

$N = \sum f$ = Total frequency

Cf_b = Cumulative frequency before the median class

f_m = Frequency of the median class.

w = Class size or width.

Example3: The table below shows the height of 70 men randomly selected at Sango Ota.

Height	118-126	127-135	136-144	145-153	154-162	163-171	172-180
No of rods	8	10	14	18	9	7	4

Compute the median.

Solution

Height	Frequency	Cumulative frequency
118 - 126	8	8
127 - 135	10	18
136 - 144	14	32
145 - 153	18	50
154 - 162	9	59
163 - 171	7	66
172 - 180	4	70
	70	

$\frac{n}{2} = \frac{70}{2} = 35$. The sum of first three classes frequency is 32 which therefore means that the median lies in the fourth class and this is the median class. Then

$L_1 = 144.5,\ n = 70, cf_b = 32, w = 9$

$$\tilde{X} = L_1 + \left(\frac{\frac{n}{2} - cf_b}{f_m}\right)w$$

$$= 144.5 + \left[\frac{35-32}{18}\right] \times 9$$

$$= 144.5 + \left(\frac{3\times 9}{18}\right)$$

$$= 144.5 + 1.5 = 136.$$

ADVANTAGE OF THE MEDIAN

(i) Its value is not affected by extreme values; thus it is a resistant measure of central tendency.

(ii) It is a good measure of location in a skewed distribution

DISAVANTAGE OF THE MEDIAN

1) It does not take into consideration all the value of the variable.

THE MODE

The mode is the value of the data which occurs most frequently. A set of data may have no, one, two or more modes. A distribution is said to be uni-model, bimodal and multimodal if it has one, two and more than two modes respectively.

E .g: The mode of scores 2, 5, 2, 6, 7 is 2.

Calculation of mode from grouped data

From a grouped frequency distribution, the mode can be obtained from the formula.

$$\text{Mode, } \hat{X} = L_{mo} + \left(\frac{\Delta_1}{\Delta_1+\Delta_2}\right) w$$

Where: L_{mo} = lower class boundary of the modal class

Δ_1 = Difference between the frequency of the modal class and the class before it.

Δ_2 = Difference between the frequency of the modal class and the class after it.

w = Class size.

Example: For the table below, find the mode.

Class	11 – 20	21 – 30	31 – 40	41 - 50	51 – 60	61 – 70
frequency	6	20	12	10	9	9

Calculate the modal age.

Solution: $L_{mo} = 20.5, \Delta_1 = 20 - 6 = 14, \Delta_2 = 20 - 12 = 8,$

$w = 30.5 - 21.5 = 10$

$$\text{Mode, } \hat{X} = L_{mo} + \left(\frac{\Delta_1}{\Delta_1+\Delta_2}\right) w$$

$$= 20.5 + \left(\frac{14}{14+8}\right) 10$$

$$= 20.5 + \left(\frac{14}{22}\right) 10$$

$$= 20.5 + (0.64)10$$

$$= 20.5 + 6.4$$

$$= 26.9$$

ADVANTAGE OF THE MODE

1) It is easy to calculate.

DISADVANTAGE OF THE MODE

(i) It is not a unique measure of location.

(ii) It presents a misleading picture of the distribution.

(iii) It does not take into account all the available data.

Exercise 2

1. Find the mean, median and mode of the following observations: 5, 6,10,15,22,16,6,10,6.
2. The six numbers 4, 9,8,7,4 and Y, have mean of 7. Find the value of Y.
3. From the data below

Class	21 – 23	24 – 26	27 – 29	30 – 32	33 – 35	36 – 38	39 – 41
Frequency	2	5	8	9	7	3	1

Calculate the (i)Mean (ii)Mode (iii) Median

MEASURES OF PARTITION

From the previous section, we've seen that the median is an average that divides a distribution into two equal parts. So also these are other quantity that divides a set of data (in an array) into different equal parts. Such data must have been arranged in order of magnitude. Some of the partition values are: the quartile, deciles and percentiles.

THE QUARTILES

Quartiles divide a set of data in an array into four equal parts.

For ungrouped data, the distribution is first arranged in ascending order of magnitude.

Then

First Quartiles: $Q_1 = \left(\frac{N+1}{4}\right)^{th}$

Second Quartile: $Q_2 = 2\left(\frac{N+1}{4}\right) th = median$

Third Quartile: $Q_3 = 3\left(\frac{N+1}{4}\right) th$ member of the distribution

For a grouped data

$$Q_l = L_{ql} + \left(\frac{\frac{iN}{4} - Cf_{ql}}{f_{ql}}\right) \times w$$

Where

l = The quality in reference

L_{ql} = Lower class boundary of the class counting the quartile

N = Total frequency

Cf_{ql} = Cumulative frequency before the Q_l class

f_{ql} = The frequency of the Q_l class

w = Class size of the Q_l class.

DECILES

The values of the variable that divide the frequency of the distribution into ten equal parts are known as deciles and are denoted by $D_1, D_2, \ldots, D_9$. the fifth deciles is the median.

For ungrouped data, the distribution is first arranged in ascending order of magnitude. Then

$D_1 = \left(\frac{n+1}{10}\right) th$ member of the distribution

$D_2 = 2\left(\frac{n+1}{10}\right) th \ member \ of \ the \ distribution$

$D_9 = 9\left(\frac{n+1}{10}\right) th \ member \ of \ the \ distribution$

For a grouped data

$$D_i = L_{DI} + \left(\frac{\frac{iN}{10} - Cf_{DI}}{F_{DI}}\right) w \qquad i = 1, 2, \ldots, 9$$

Where $i = Decile \ in \ reference$

$L_{DI} = lower \ class \ boundary \ of \ the \ class \ counting \ the \ decile$

$N = Total \ frequency$

$Cf_{DI} = cumulative \ frequency \ up \ to \ the \ low \ boundary \ of \ the \ D_i \ class$

$F_{DI} = the \ frequency \ of \ the \ D_i \ class$

$w = Class \ size \ of \ the \ D_i \ class.$

PERCENTILE

The values of the variable that divide the frequency of the distribution into hundred equal parts are known as percentiles and are generally denoted by $P_{1,\ldots,}P_{99}$.

The fiftieth percentile is the median.

For ungrouped data, the distribution is first arranged in ascending order of magnitude. Then

$P_1 = \left(\frac{n+1}{100}\right) th$ member of the distribution

$P_2 = \frac{2(n+1)}{100} th$ member of the distribution

$P_{99} = \frac{99(n+1)}{100} th$ member of the distribution

For a grouped data,

$$P_i = L_{pi} + \left(\frac{\frac{iN}{100} - Cf_{pi}}{f_{pi}}\right) \times w \qquad i = 1, \ldots, 99$$

Where

$i = percentile\ in\ reference$

$L_{pi} = Lower\ class\ boundary\ of\ the\ class\ counting\ the\ percentile$

$N = Total\ frequency$

$Cf_{pi} = Cumulative\ frequency\ up\ to\ the\ lower\ class\ boundary\ of\ the\ P_i\ class$

$f_{pi} = Frequency\ of\ the\ p_i\ class.$

Example: For the table below, find by calculation (using appropriate expression)

(i) Lower quartile, Q_1

(ii) Upper Quartile, Q_3

(iii) 6th Deciles, D_6

(iv) 45th percentile of the following distribution

Mark	20 – 29	30 – 39	40 – 49	50 – 59	60 – 69	70 – 79	80 – 89	90 – 99
Frequency	8	10	14	26	20	16	4	2

23

Solution

Marks	frequency	cumulative frequency
20 - 29	8	8
30 - 39	10	18
40 - 49	14	32
50 - 59	26	~~32~~ 58
60 - 69	20	~~58~~ 78
70 - 79	16	~~78~~ 94
80 - 89	4	98
90 - 99	2	100
	100	

(i) Lower quartile, $Q_1 = L_{q1} + \left(\frac{\frac{iN}{4} - Cf_{q1}}{f_{q1}}\right)w$

$\frac{iN}{4} = \frac{1 \times 100}{4} = 25, Cf_{q_1} = 18, f_{q_1} = 14, w = 10, L_{q_1} = 39.5$

$Q_1 = 39.5 + \left(\frac{25 - 18}{14}\right)10$

$= 44.5$

(ii) Upper Quartile, $Q_3 = L_{q_3} + \left(\frac{\frac{3N}{4} - Cf_q}{f_{q3}}\right)w$

$\frac{3N}{4} = \frac{3 \times 100}{4} = 75, L_{q_3} = 59.5, Cf_{q_3} = 58, F_{q_3} = 20, w = 10$

$Q_3 = 59.5 + \left(\frac{75 - 58}{20}\right)10 = 68$

(iii) $D_6 = L_{D_6} + \left(\frac{\frac{6N}{10} - Cf_{D_6}}{f_{D_6}}\right)w$

$\frac{6N}{10} = \frac{6 \times 100}{10} = 60, L_{D_6} = 59.5, Cf_{D_6} = 58, f_{D_6} = 20, w = 10$

$D_6 = 59.5 + \left(\frac{60 - 58}{20}\right)10 = 60.5$

(iv) $P_{45} = L_{p_{45}} + \left(\frac{\frac{45N}{100} - Cf_{p45}}{f_{p45}}\right) w$

$\frac{45N}{100} = \frac{45 \times 100}{100} = 45, L_{p_{45}} = 49.5, Cf_{p45} = 32, f_{p45} = 26, w = 10$

$$P_{45} = 49.5 + \left(\frac{45-32}{26}\right) 10$$

$$= 49.5 + 5$$

$$= 54.5$$

MEASURES OF DISPERSION

Dispersion or variation is degree of scatter or variation of individual value of a variable about the central value such as the median or the mean. These include range, mean deviation, semi-interquartile range, variance, standard deviation and coefficient of variation.

THE RANGE

This is the simplest method of measuring dispersions. It is the difference between the largest and the smallest value in a set of data. It is commonly used in statistical quality control. However, the range may fail to discriminate if the distributions are of different types.

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

SEMI – INTERQUARTILE RANGE

This is the half of the difference between the first (lower) and third quartiles (upper). It is good measure of spread for midrange and the quartiles.

$$S.I.R = \frac{Q_3 - Q_1}{2}$$

THE MEAN/ABSOLUTE DEVIATION

Mean deviation is the mean absolute deviation from the centre. A measure of the center could be the arithmetic mean or median.

Given a set of $x_1, x_2, \ldots, x_n$, the mean deviation from the arithmetic mean is defined by:

$$MD = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{N}$$

In a grouped data

$$MD_{\bar{x}} = \frac{\sum_{i=1}^{n} f|x_i - \bar{x}|}{\sum_{i=1}^{n} f_i}$$

Example1: Below is the average of 6 heads of household randomly selected from a country. 47, 45, 56, 60, 41, 54 .Find the

(i) Range

(ii) Mean

(iii) Mean deviation from the mean

(iv) Mean deviation from the median.

Solution:

(i) Range $= 60 - 41 = 19$

(ii) Mean $(\bar{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i$

$$= \frac{47+45+56+60+41+54}{6}$$

$$= \frac{303}{6} = 50.5$$

(iii) Mean Deviation $(MD_{\bar{x}}) = \frac{\sum_{i=1}^{n}|x_i-\bar{x}|}{n}$

$$= \frac{|47-50.5|+|45-50.5|+|56-50.5|+|60-50.5|+|41-50.8|+|54-50.5|}{6}$$

$$= \frac{|-3.5|+|-5.5|+|5.5|+|9.5|+|-9.5|+|3.5|}{6}$$

$$= \frac{37}{6}$$

$$= 6.17$$

(iv) In array: 41, 45, 47, 54, 56, 60

$$\text{Median} = \frac{X_{\left(\frac{n}{2}\right)th} + X_{\left(\frac{n}{2}+1\right)th}}{2}$$

$$= \frac{X_3+X_4}{2} = \frac{47+54}{2} = 50.5$$

$$MD_{\tilde{x}} = \frac{|47-50.5|+|45-50.5|+|56-50.5|+\cdots+|54-50.5|}{6}$$

$$= 6.17$$

Example2: The table below shown the frequency distribution of the scores of 42 students in MTS 201

Scores	0 - 9	10 - 19	20 - 29	30 - 39	40 - 49	50 - 59	60 - 69
No of student	2	5	8	12	9	5	1

Find the mean deviation from the mean for the data.

Solution:

Classes	midpoint x	f	fx	$x-\bar{x}$	$\lvert x-\bar{x}\rvert$	$f\lvert x-\bar{x}\rvert$
0 - 9	4.5	2	9	-29.52	29.52	59.04
10 - 19	14.5	5	72.5	-19.52	19.52	97.60
20 - 29	24.5	8	196	-9.52	9.52	76.16
30 - 39	34.5	12	414	0.48	0.48	5.76
40 - 49	44.5	9	400.5	10.48	10.48	94.32
50 - 59	54.5	5	272.5	20.48	20.48	102.4
60 - 69	64.5	1	64.5	30.48	30.48	30.48
		42	1429			465.76

$$\bar{X} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} = \frac{1429}{42} \approx 34.02$$

$$MD_{\bar{x}} = \frac{\sum_{i=1}^{n} f|x_i - \bar{x}|}{\sum_{i=1}^{n} f_i} = \frac{465.70}{42}$$

$$= 11.09$$

THE STANDARD DEVIATION

The standard deviation, usually denoted by the Greek alphabet σ (small signal) (for population) is defined as the "positive square root of the arithmetic mean of the squares of the deviation of the given observation from their arithmetic mean". Thus, given $x_1, \ldots, x_n$ as a set of n observations, then the standard deviation is given by:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} \quad where\ \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

(Alternatively, $\sigma = \sqrt{\frac{\sum_{i=1}^{n} X_i^2}{n} - \left(\frac{\sum_{i=1}^{n} X_i}{n}\right)^2}$.)

For a grouped data

The standard deviation is computed using the formula

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} f_1 (X_i - \bar{X})^2}{\sum_{i=1}^{n} f_i}} \qquad or \quad \sigma = \sqrt{\frac{\sum_{i=1}^{n} f_i X_i}{\sum f_i} - \left(\frac{\sum f_i x_i}{\sum f_i}\right)^2}$$

If A= assume mean and $d = x - A$ is deviation from the assumed mean, then

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

Note: We use $S = \sqrt{\frac{\sum f(X_i - \bar{X})^2}{\sum f - 1}}$ when using sample instead of the population to obtain the standard deviation.

MERIT

(i) It is well defined and uses all observations in the distribution.

(ii) It has wider application in other statistical technique like skewness, correlation, and quality control e.t.c

DEMERIT

(i) It cannot be used for computing the dispersion of two or more distributions given in different unit.

THE VARIANCE

The variance of a set of observations is defined as the square of the standard deviation and is thus given by σ^2

COEFFICIENT OF VARIATION/DISPERSION

This is a dimension less quantity that measures the relative variation between two servers observed in different units. The coefficients of variation are obtained by dividing the standard deviation by the mean and multiply it by 100. Symbolically

$$CV = \frac{\sigma}{\bar{X}} \times 100\ \%$$

The distribution with smaller C.V is said to be better.

EXAMPLE3: Given the data 5, 6, 9, 10, 12. Compute the variance, standard deviation and coefficient of variation

SOLUTION

$$\bar{X} = \frac{5+6+9+10+12}{5} = 8.4$$

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{N}$$

$$= \frac{(5-8.4)^2+(6-8.4)^2+(9-8.4)^2+(10-8.4)^2+(12-8.4)^2}{5}$$

$$= \frac{11.56+5.76+0.36+2.56+12.96}{5}$$

$$= 33.2/5$$

$$= 6.64$$

$$\therefore \sigma = \sqrt{\sigma^2}$$

$$= \sqrt{6.64}$$

$$= 2.58$$

Hence $C.V = \frac{2.58}{8.4} \times 100$

$$= 30.71\%$$

EXAMPLE4: Given the following data. Compute the

(i) Mean

(ii) Standard deviation

(iii) Coefficient variation.

30

Ages(in years)	50 – 54	55 – 59	60 – 64	65 – 69	70 – 74	75 – 79	80 – 84
Frequency	1	2	10	12	18	25	9

SOLUTION

Classes	x	F	fx	$x-\bar{x}$	$f(x-\bar{x})^2$
50 – 54	52	1	52	-20.06	402.40
55 – 59	57	2	114	-15.06	453.61
60 – 64	62	10	620	-10.06	1012.04
65 – 69	67	12	804	-5.06	307.24
70 – 74	72	18	1296	-0.06	0.07
75 – 79	77	25	1925	4.94	610.09
80 – 84	82	9	738	9.94	889.23
		77	5549		3674.68

$$\bar{X} = \frac{\Sigma fx}{\Sigma f} = \frac{5549}{77} = 72.06$$

$$\sigma = \sqrt{\frac{\Sigma f(x-\bar{x})^2}{\Sigma f}}$$

$$= \sqrt{\frac{3674.68}{77}}$$

$$= \sqrt{47.7231}$$

$$= 6.9082$$

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

$$= \frac{7.14}{72} \times 100$$

$$= 9.917\%$$

Exercise 3

The data below represents the scores by 150 applicants in an achievement text for the post of Botanist in a large company:

Scores	10 - 19	20 - 29	30 - 39	40 - 49	50 - 59	60 - 69	70 - 79	80 - 89	90 - 99
Frequency	1	6	9	31	42	32	17	10	2

Estimate

(i) The mean score

(ii) The median score

(iii) The modal score

(iv) Standard deviation

(v) Semi – interquartile range

(vi) D_4

(vii) P_{26}

(viii) coefficient of variation

Index:

An index is a statistical measure of changes in a representative group of individual data points. These data may be derived from any number of sources, including company performance, prices, productivity, and employment.

Index number: an index number is an economic data figure reflecting price or quantity compared with a standard or base value. The base usually equal 100 and the index number is usually expressed as 100 times the ratio to the base value.

Index numbers are used especially to compare business activities, the cost of living, and employment. They enable economists to reduce bulky business data into easily understood terms.

Some of the problems which make it difficult to construct an index number on regular basis include base year (what should be the base the year), commodities to be included (which commodities should be included), allocation of weights to each commodity) etc.

Defining purpose: in the construction of an index number, the defining purpose is the important step and indicates the purpose of measure to be calculated. Since the choice of weights, commodities, and base year all depends on the purpose of measure, therefore: to define it before the construction of an index number, for example we are constructing two indices, i.e. cost of living index and wholesale price index.

Here the weights, commodities, and base year of cost-of-living index will be totally different from the weight of wholesale price index.

Commodities to be included: selection of the commodities is an important problem in the construction of the index number. The commodities should be relevant, comparable, representative, and reliable. That is if we are measuring the cost of living of a particular class, then only those commodities should be included which are consumed by the class. Similarly, most common, or most

popular commodities should be used, and the quantity of these commodities should not vary from each other.

Determination of weights: when the selection process of the commodities is completed, the next step is the determination of weights to each commodity. The determination of weights depends upon the relative importance attached to each commodity. But the importance of a commodity varies with place, time, and the habits of those who use or trade in it. Similarly, it is also important to decide whether to keep the weights constant or change it periodically. Generally, weights are kept constant over a long period of time to facilitate the comparisons.

Choice of base year: a base year should be a normal year because it is the year to which all are compared. The base year should be free from all hazards that affect the economy of the country (like floods, war, earthquakes etc.) if such things happen in a year, then the economy of country will be affected which will ultimately affect the price in the year.

Now because the year is not a normal year, therefore it cannot be taken as base year. It is there necessary to take care while selecting the base year. There are two types of base year.

i. **Fixed base year:** it is the year which remains fixed for the comparison of prices of all years. For example, 2005 is taken as base year for the comparison of prices of 2006, 2007 and 2008. Now because the prices of three years are compared with the prices of 2005 therefore 2005 is the base year.
ii. **Chain base year:** it is the year which does not remain fixed for the comparison of prices of all years. In case of chain base year every preceding year is taken as base year for every preceding year. For example, 2005 is taken as base year for 2006 and 2006 is take as base year for 2007, similarly 2007 is taken as base year for 2008 and so on.

SIMPLE OR UNWEIGHTED INDEX NUMBERS:

In unweighted index numbers, the weights are not assigned to the variables, i.e. consideration is not given tot the importance of each variable.

The following are different methods followed to calculate the unweighted index number:

I. Simple Aggregative Method:

In this method of computing the price index, we express the total price of the commodity in a given year as a percentage of the total of the commodity in the base year.

Hence, $P_{01}\frac{current\ year\ total}{base\ year\ total} \times 100 = \frac{\sum P_1}{\sum P_0} \times 100$

Where: $P_{01} \longrightarrow price\ index$

$P_1 \longrightarrow price\ of\ the\ commodity\ in\ the\ given\ year.$

$P_0 \longrightarrow price\ of\ the\ commodity\ in\ the\ base\ year.$

Note: the index for the base year is always taken as 100.

Example 1:

Commodity	1995	1996	1997
A	13	14	15.4
B	2	2.9	3.2
C	7	8	6.7
Total	22	24.9	25.3

Compute the simple aggregative index for 1996 and 1997 over 1995.

Solution.

Simple aggregative index for 1996 over 1995 is:

$P_{01} = \frac{24.9}{22} \times 100$ (since $P_{01} = \frac{\sum P_1}{\sum P_0} \times 100$)

$\Longrightarrow P_{01} = 1.1318 \times 100$

$\Longrightarrow P_{01} = 113.18$

Simple aggregative index for 1997 over 1995 is:

$P_{01} = \frac{25.3}{22} \times 100$ (since $P_{01} = \frac{\sum P_1}{\sum P_0} \times 100$)

$\Longrightarrow P_{01} = 1.15 \times 100$

$\Longrightarrow P_{01} = 115$

II. Simple Average of Relatives Method:

It can be derived using unweighted geometric mean:

$$P_{01} = antilog \left[\frac{\sum log \left(\frac{P_1}{P_0} \right) \times 100}{N} \right]$$

Where: $P_{01} \longrightarrow index\ number\ of\ the\ current\ year.$

$\sum P_1 \longrightarrow total\ of\ the\ current\ year's\ price\ of\ all\ items.$

$\sum P_0 \longrightarrow total\ of\ the\ base\ year's\ price\ of\ all\ items.$

$N \longrightarrow total\ number\ of\ commodities$

Example: use the data given tin example (1) and fine the simple average of relatives for the year 1997, taking 1995 as the base year.

Solution:

Commodity	$\boldsymbol{P_0}$	$\boldsymbol{P_1}$	$\boldsymbol{\frac{P_1}{P_0} \times 100}$
A	13	15.4	118.4615
B	2	3.2	160
C	7	6.7	95.71429
Total	22	25.3	115

$$\text{Simple average of relatives} = P_{01} = \frac{374.16}{3}\left[since\ P_{01} = \frac{\sum\left(\frac{P_1}{P_0}\right)\times 100}{N}\right]$$

$$\Longrightarrow P_{01} = 124.72$$

WEIGHTED INDEX NUMBER:

In this method, each variable is assigned a weighted depending on its relative importance. Often the quality or the volume of the commodity sold during the base year, or some typical year may also be taken as the weight.

i. Weighted aggregative index method:

a. Laspeyres's Index Number:

$$P_L = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$

Here, weights assigned are base year quantities.

Note: Laspeyres's price index number is same as consumer price index number.

b. Paasche's Index Number:

$$P_p = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100$$

Here, weights assigned are current year quantities.

Note: Dobish-Bowley index number is the A.M of Laspeyres's and Paasche's index number. (A.M = Arithmetic mean).

i.e. $P_{DB} = \frac{P_L + P_P}{2}$

c. Marshall-Edgeworht Index Number:

$$P_{ME} = \frac{\sum p_1 q_0 + \sum p_1 q_1}{\sum p_0 q_0 + \sum \mathrm{p_0 q_1}} \times 100$$

Here, the weights assigned are the average of base year and current year quantities.

d. Fisher's Ideal Index Number:

This formular is the geometric mean of Laspeyres's and Paasche's index numbers.

$$P_F = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times 100 \Longrightarrow P_F = \sqrt{P_L \times P_P} \times 100$$

Example: Compute the Laspeyres's, Paasches's, Marshall-Edgeworth and Fisher's index number for the following data:

Commodities	Base year		Current year	
	Price (p_0)	Quantity (q_0)	Price (p_1)	Quantity (q_1)
A	4	3	6	2
B	5	4	0	4
C	7	2	9	2
D	2	3	1	5

Solution:

Commodities	p_0q_0	p_1q_0	p_0q_1	p_1q_1
A	12	18	8	12
B	20	0	20	0
C	14	18	14	18
D	6	3	10	5
Total	52	39	52	35

$$Laspeyres's\ index\ = P_L = \frac{39}{52} \times 100 \left(since\ P_L = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100 \right)$$

$$= 0.75 \times 100 = \underline{\underline{75}}$$

$$Paasche's\ index\ = P_P = \frac{35}{52} \times 100 \left(since\ P_P = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100 \right)$$

$$= 0.673 \times 100 = \underline{\underline{67.3}}$$

$$Marshall-Edgewoth\ index\ = P_{ME} = \frac{39+35}{52+52} \times 100$$

$$\left(since\ P_{ME} = \frac{\sum p_1 q_0 + \sum p_1 q_1}{\sum p_0 q_0 + \sum \mathrm{p_0 q_1}} \times 100 \right)$$

$$= \frac{74}{104} \times 100$$

$$= 0.711 \times 100 = 71.1\ \therefore P_{ME} = \underline{\underline{71.1}}$$

$$Fisher's\ index\ = P_F = \sqrt{75 \times 67.3}\ \left(since\ P_F = \sqrt{P_L \times P_P}\ \right)$$

$$\Longrightarrow P_F = \sqrt{5047.5} = \underline{\underline{71.04}}$$

ii. Weighted aggregative of relatives method:

The weighted arithmetic-mean of price relatives using base year value weights is represented by:

$$P_{01} = \frac{\sum w \left(\frac{P_1}{P_0} \times 100 \right)}{\sum w}$$

Example:

Calculate the weighted arithmetic mean index number from the following data:

Commodities	Price		Weight (Quintal)
	Base Year	Current Year	
A	2	3	10
B	4	5	4
C	10	10	5
D	12	9	1

Solution:

Items	P_0	P_1	w	$p = \frac{P_1}{P_0} \times 100$	**wp**
A	2	3	10	150	1500
B	4	5	4	125	500
C	10	10	5	100	500
D	12	9	1	75	75
Total			20		2575

The weighted arithmetic mean is: $P_{01} = \frac{\sum w\left(\frac{P_1}{P_0} \times 100\right)}{\sum w}$

$$= \frac{2575}{20} = \underline{\underline{128.75}}$$

Quantity Index Number:

When we want to measure and compare quantities, we resort to quality index number. Quantity indices are used as indicators of the level of output in economy. These are calculated by adopting price as weight, i.e., changing 'p' into 'q' and 'q' into 'p' in all formulae for price index number.

The various types of quality indices are as follows:

- Simple aggregative of qualities: this is expressed as: $\frac{\sum q_n}{\sum q_0} \times 100$
- Simple average of quantity relatives: this is expressed as: $\frac{\sum\left(\frac{q_n}{q_0} \times 100\right)}{N}$
- Weighted aggregate quantity indices:
    a. Laspeyres's quantity index number: $Q_L = \frac{\sum q_1 p_0}{\sum q_0 p_0} \times 100$
    b. : Paasche's index number: $Q_P = \frac{\sum q_1 p_1}{\sum q_0 p_1} \times 100$
    c. Marshall-Edgeworth index number: $Q_{ME} = \frac{\sum q_1 p_0 + \sum q_1 p_1}{\sum q_0 p_0 + \sum q_0 p_1} \times 100$
    d. Fisher's ideal index number:

This formula is the geometric mean of the Laspeyres's and Paasche's index numbers.

$$Q_F = \sqrt{\frac{\sum q_1 p_0}{\sum q_0 p_0} \times \frac{\sum q_1 p_1}{\sum q_0 p_1}} \times 100 \quad \Longrightarrow Q_F = \sqrt{Q_L \times Q_P} \times 100$$

Note: base year weighted average of quantity relative is given by: $\frac{\sum\left(\frac{q_1}{q_0} \times p_0 q_0\right)}{\sum p_0 q_0}$

Example: Calculate:

i. Laspeyres's quantity index number.
ii. Paasches's quality index number.
iii. Dorbish-Bowley quantity index number.
iv. Marshall-Edgeworth quality index number.
v. Fisher's ideal quality index number for the given date:

	Price		**Quantity sold (quintals)**	
Items	**Base year**	**Current year**	**Base year**	**Current year**
Rice	400	850	100	120
Wheat	320	690	20	60
Sugar	720	1600	10	10
Dal	720	2100	10	20

Solution:

Items	p_0	p_1	q_0	q_1	p_0q_0	p_0q_1	p_1q_0	p_1q_1
Rice	400	850	100	120	40000	48000	85000	102000
Wheat	320	690	20	60	6400	19200	13800	41400
Sugar	720	1600	10	10	7200	7200	16000	16000
Dal	720	2100	10	20	7200	14400	21000	42000
Total					60,800	88,000	135,800	201,400

The Laspeyres's quantity index number is:

$$Q_L = \frac{\sum q_1 p_0}{\sum q_0 p_0} \times 100 = \frac{88800}{60800} \times 100 = \underline{\underline{146.05}}$$

The Paasche's quality index number is:

$$Q_P = \frac{\sum q_1 p_1}{\sum q_0 p_1} \times 100 = \frac{201400}{135800} \times 100 = \underline{\underline{148.31}}$$

The Dorbish-Bowley quality index number is:

$$Q_{DB} = \frac{Q_L + Q_P}{2} \Longrightarrow \frac{146.05 + 148.31}{2} = \underline{\underline{147.18}}$$

The Marshall-Edgeworht quality index number is:

$$Q_{ME} = \frac{\sum q_1 p_0 + \sum q_1 p_1}{\sum q_0 p_0 + \sum q_0 p_1} \times 100$$

$$\Longrightarrow Q_{ME} = \frac{88800 + 201400}{60800 + 135800} \times 100 = \underline{\underline{147.61}}$$

The Fisher's quantity index number is: $Q_F = \sqrt{\frac{\sum q_1 p_0}{\sum q_0 p_0} \times \frac{\sum q_1 p_1}{\sum q_0 p_1}} \times 100$

$$= \sqrt{\frac{88800}{60800} \times \frac{201400}{135800}} \times 100 = \underline{\underline{147.18}}$$

Value Indices:

Value is the product of price and quality. Thus, a value index equals the total sum of the values of given year divided by the sum of the values of base year.

i.e. value index number $= \frac{\sum V_1}{\sum V_0} \times 100$

$$= \frac{\sum p_1 q_1}{\sum p_0 q_0} \times 100$$

Example:

Calculate the value index number for the given data:

	Base year		Current year	
Items	**Price**	**Total value**	**Price**	**Total value**
A	50	100	60	180
B	40	120	40	200
C	100	100	120	120
D	20	80	25	100

Solution:

Items	P_0	V_0	P_1	V_1
A	50	100	60	180
B	40	120	40	200
C	100	100	120	120
D	20	80	25	100
Total		400		600

$$\text{Value Index Number} = \frac{\sum V_1}{\sum V_0} \times 100$$

$$= \frac{600}{400} \times 100$$

$$= \underline{\underline{150}}$$

CHAPTER FIVE

MOMENT, SKEWNESS AND KURTOSIS

5.1 INTRODUCTION

In the previous chapter, we have looked at the measure of central tendency which tells us the concentration of observations about the middle position. We have also looked at measure of dispersion or variation which tells us about the spread of observations about its central position. However, in true statistical measures, this does not reveal the exact features of a frequency distribution most especially if we come across a frequency distribution which differ widely in their nature and composition but have the same measure of central tendency and dispersion. It follows that, the two measures (measures of central tendency and measures of dispersion) are inadequate to characterize a distribution completely; therefore, they need to be supported by other measures such as skewness, moment and kurtosis.

5.2 MOMENT

In mathematics, any distribution can be characterized by a number of features (such as the mean, the variance, the skewness, etc.) and the moments of a random variable's probability distribution are related to these features. A moment can simply be defined as a quantitative measure of the shape of a set of points. Moment can also be defined as the mean value of the power of the variant. Moment about the origin can be obtained by using the formula;

$$\bar{x}^1 = \text{First moment about the origin} = \sum \frac{x^1}{n}$$

$$\bar{x}^2 = \text{Second moment about the origin} = \sum \frac{x^2}{n}$$

$$\bar{x}^3 = \text{Third moment about the origin} = \sum \frac{x^3}{n}$$

$$\bar{x}^4 = \text{Four moment about the origin} = \sum \frac{x^4}{n}$$

$$\bar{x}^r = r^{th}\text{moment about the origin} = \sum \frac{x^r}{n}$$

Example 5.1

Find the (a) first (b) second (c) third and (d) fourth moment about the origin of 1, 2, 3, 4 and 5

$$\bar{x} = \sum \frac{x^r}{n}$$

SOLUTION

(a) $\bar{x}^1$ = first moment = $\sum \frac{x1}{n}$

$$= \frac{1^1 + 2^1 + 3^1 + 4^1 + 5^1}{5} = \frac{15}{5} = 3$$

(b) $\bar{x}^2 = \frac{1^2 + 2^2 + 3^2 + 4^2 + 5^2}{5} = \frac{55}{5} = 11$

(c) $\bar{x}^3 = \frac{1^3 + 2^3 + 3^3 + 4^3 + 5^3}{5} = \frac{225}{5} = 45$

(d) $\bar{x}^4 = \frac{1^4 + 2^4 + 3^4 + 4^4 + 5^4}{5} = \frac{979}{5} = 195.8$

If $x_1, x_2, x_3, \ldots x_n$ occur with the frequencies $f_1, f_2, f_3, \ldots f_n$

Then, the r^{th} moment about the origin $= \frac{\sum f x^r}{\sum f}$

The first moment about the origin $= \frac{\sum f x^1}{\sum f}$

The second moment about the origin $= \frac{\sum f x^2}{\sum f}$

The third moment about the origin $= \frac{\sum f x^3}{\sum f}$

The fourth moment about the origin $= \frac{\sum f x^4}{\sum f}$

Example 5.2

Find the (a) first (b) second (c) third and (d) fourth moment about the origin of the distribution below;

X	1	2	3	4	5
F	3	4	10	6	2

SOLUTION

Using

The r^{th} moment about the origin $= \frac{\sum f x^r}{\sum f}$

Represent the above distribution with the aid of

i. A bar chart ii. A histogram iii. An Ogi

iv. A frequency polygon

36) Briefly discuss each of the following graphical representation

a. Bar chart b. Pie chart c. Histogram

d. Frequency polygon e. Cumulative frequency curve

37) Draw (i) Bar chart (ii) Histogram

(iii) Frequency polygon and (iv) Cumulative frequency curve to represent the distribution below.

Class Interval	2 – 4	5 – 7	8 – 10	11 – 13	14 – 16	17 – 19	20 – 22
Frequency	1	3	6	12	9	6	3

38) Draw a histogram and a frequency polygon to represent the table below

Class Interval	1 – 5	6 – 10	11 – 15	16 – 20	21 – 25
Frequency	3	5	7	6	4

39) Draw up a frequency polygon and a cumulative frequency curve to represent the distribution below

Class Interval	1 – 3	4 – 6	7 – 9	10 – 12	13 – 15	16 – 18
Frequency	3	5	12	16	10	4

40) The percentage marks of 100 students in a school certificate examination are:

Class Interval	0 – 9	10 – 19	20 – 29	30 – 39	40 – 49	50 – 59	60 – 69	70 – 79	80 – 89	90 – 99
Frequency	1	2	5	17	23	25	18	5	3	1

Draw a histogram and the frequency polygon of the distribution.

(a) The first moment about the origin $= \frac{\sum fx^1}{\sum f}$

$$= \frac{(3\times1^1)+(4\times2^1)+(10\times3^1)+(6\times4^1)+(2\times5^1)}{3+4+10+6+2}$$

$$= \frac{3+8+30+24+10}{25} = \frac{75}{25} = 3$$

(b) The second moment about the origin $= \frac{\sum fx^2}{\sum f}$

$$= \frac{(3\times1^2)+(4\times2^2)+(10\times3^3)+(6\times4^4)+(2\times5^5)}{3+4+10+6+2}$$

$$= \frac{3+16+90+96+50}{25} = \frac{255}{25} = 10.2$$

(c) The third moment about the origin $\frac{\sum fx^1}{\sum f}$

$$= \frac{(3\times1^3)+(4\times2^3)+(10\times3^3)+(6\times4^3)+(2\times5^3)}{3+4+10+6+2}$$

$$= \frac{3+32+270+384+450}{25} = \frac{1139}{25} = 45.56$$

(d) The fourth moment about the origin $= \frac{\sum fx^4}{\sum f}$

$$= \frac{(3\times1^4)+(4\times2^4)+(10\times3^4)+(6\times4^4)+(2\times5^4)}{3+4+10+6+2}$$

$$= \frac{3+64+810+1536+2250}{25} = \frac{4663}{25} = 186.52$$

Similarly, the moment about the mean can be obtained by using the formula;

$M_r = \bar{x}^r = \frac{\sum(x-\bar{x})^r}{n}$ = the r^{th} moment about the mean

The first moment about the mean $= M_1 = \frac{\sum(x-\bar{x})^1}{n}$

The second moment about the mean $= M_2 = \frac{\sum(x-\bar{x})^2}{n}$

The third moment about the mean $= M_3 = \frac{\sum(x-\bar{x})^3}{n}$

The fourth moment about the mean = $M_4 = \frac{\sum(x-\bar{x})^4}{n}$

Example 5.3
Find (a) first (b) second (c) third and (d) fourth moment about the mean of 2, 3, 7, 8 and 10

SOLUTION

$$\text{Mean} = \bar{x} = \frac{2+3+7+8+10}{5} = \frac{30}{5} = 6$$

(a) The first moment about the mean = $M_1 = \frac{\sum(x-\bar{x})^1}{n}$

$$\frac{(2-6)^1 + (3-6)^1 + (7-6)^1 + (8-6)^1 + (10-6)^1}{5}$$

$$\frac{(-4)^1 + (-3)^1 + (1)^1 + (2)^1 + (4)^1}{5} = \frac{-4-3+1+2+4}{5} = \frac{0}{5} = 0$$

(b) The second moment about the mean = $M_2 = \frac{\sum(x-\bar{x})^2}{n}$

$$\frac{(2-6)^2 + (3-6)^2 + (7-6)^2 + (8-6)^2 + (10-6)^2}{5}$$

$$\frac{16+9+1+4+16}{5} = \frac{46}{5} = 9.2$$

(c) The third moment about the mean = $M_3 = \frac{\sum(x-\bar{x})^3}{n}$

$$\frac{(2-6)^3 + (3-6)^3 + (7-6)^3 + (8-6)^3 + (10-6)^3}{5}$$

$$\frac{-64-27+1+8+64}{5} = \frac{-18}{5} = -3.6$$

(d) The fourth moment about the mean = $M_4 = \frac{\sum(x-\bar{x})^4}{n}$

$$\frac{(2-6)^4 + (3-6)^4 + (7-6)^4 + (8-6)^4 + (10-6)^4}{5}$$

$$\frac{256+81+1+16+256}{5} = \frac{610}{5} = 122$$

4.2.1 Moment about the Mean for Group Data

The moment about the mean of a group data can be obtained by using;

First moment about the mean $= M_1 = \frac{\sum(x-\bar{x})^1}{n}$

Second moment about the mean $= M_2 = \frac{\sum(x - \bar{x})^2}{n}$

Third moment about the mean $= M_3 = \frac{\sum(x - \bar{x})^3}{n}$

r^{th} moment about the mean $= M_r = \frac{\sum(x - \bar{x})^r}{n}$

Example 5.4

Find the (a) first (b) second (c) third and (d) fourth moment about the mean of the distribution below;

Class interval	1-3	4-6	7-9	10-12	13-15
Frequency	2	6	10	4	3

SOLUTION

Class interval	X	F	Fx
1-3	2	2	4
4-6	5	6	30
7-9	8	10	80
10-12	11	4	44
13-15	14	3	42
		25	200

$\bar{x} = 200/_{25} = 8$

First moment about the mean $= M_1 = \frac{\sum(x - \bar{x})^1}{n}$

$$\frac{2(2 - 8)^1 + (3 - 8)^1 + (7 - 8)^1 + (8 - 8)^1 + (10 - 8)^1}{5}$$

$$= \frac{2(-6)^1 + 6(-5)^1 + 10(-1)^1 + 4(0)^1 + 3(2)^1}{25} = \frac{-12 - 30 - 10 + 0 + 6}{25} = \frac{-46}{25}$$

Second moment about the mean $= M_2 = \frac{\sum(x - \bar{x})^2}{n}$

$$= \frac{2(2 - 8)^2 + 6(3 - 8)^2 + 10(7 - 8)^2 + 4(8 - 8)^2 + 3(10 - 8)^2}{25}$$

$$= \frac{2(36) + 6(25) + 10(+1) + 4(0) + 3(2)}{25} = \frac{72 + 150 + 10 + 0 + 6}{25} = \frac{238}{25}$$

Third moment about the mean $= M_3 = \frac{\sum(x - \bar{x})^3}{n}$

$$= \frac{2(2 - 8)^3 + 6(3 - 8)^3 + 10(7 - 8)^3 + 4(8 - 8)^3 + 3(10 - 8)^3}{25}$$

$$= \frac{2(-216) + 6(-125) + 10(-1) + 4(0) + 3(8)}{25} = \frac{-432 - 750 - 10 + 0 + 24}{25} = \frac{-1168}{25}$$

Fourth moment about the mean $= M_4 = \frac{\sum(x - \bar{x})^4}{n}$

$$= \frac{2(2 - 8)^4 + 6(3 - 8)^4 + 10(7 - 8)^4 + 4(8 - 8)^4 + 3(10 - 8)^4}{25}$$

$$= \frac{2(-6)^4 + 6(-5)^4 + 10(-1)^4 + 4(0)^4 + 3(2)^4}{25}$$

$$= \frac{2(1296) + 6(625) + 10(1) + 4(0) + 3(16)}{25} = \frac{2592 + 3750 + 10 + 0 + 48}{25} = \frac{6400}{25}$$

5.3 SKEWNESS

In probability theory and statistics, **skewness** is a measure of the extent to which a probability distribution of a real-valued random variable "leans" to one side of the mean. The skewness value can be positive or negative, or even undefined. For a unimodal distribution, negative skew indicates that the *tail* on the left side of the probability density function is longer or fatter than the right side. In this situation, the mean and the median are both less than the mode.

In summary, for a data set skewed to the left:

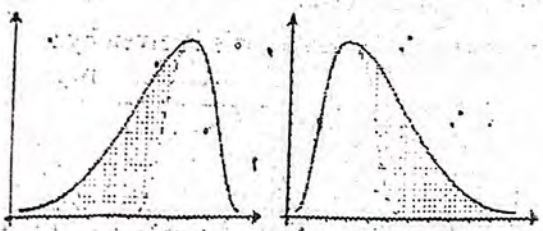
Always: mean < mode
Always: median < mode
Most of the time: mean < median < mode

Conversely, positive skew indicates that the tail on the right side is longer or fatter than the left side. In this situation, the mean and the median are both greater than the mode.

In summary, for a data set skewed to the right:

Always: mode < mean
Always: mode < median
Most of the time: mode < median < mean

A zero value indicates that the tails on both sides of the mean balance out, which is the case both for a symmetric distribution. Consider the distribution in the figures below;



Negative Skew Positive Skew

5.4 MEASURES OF SKEWNESS

It's one thing to look at two set of data and determine that one is symmetric while the other is asymmetric. It's another to look at two sets of asymmetric data and say that one is more skewed than the other. It can be very subjective to determine which is more skewed by simply looking at the graph of the distribution. This is why there are ways to numerically calculate the measure of skewness. One measure of skewness called Pearson's first coefficient of skewness is to subtract the mean from the mode and then divide this difference by the standard deviation of the data. The reason for dividing the difference is to have a dimensionless quantity. This explains why data skewed to the right has positive skewness. If the data set is skewed to the right, the mean is greater than the mode and so subtracting the mode from the mean gives a positive number. A similar argument explains why data skewed to the left has negative skewness.

Pearson's first coefficient of skewness is given by;

$$Skewness = \frac{\bar{x} - mode}{\sigma}$$

Pearson's second coefficient of skewness is also used to measure the asymmetry of a data set. For this quantity we subtract the mode from the median, multiply this number by three and then divide by the standard deviation.

Pearson's second coefficient of skewness is derived from the empirical relationship between the mean and the median and is given by;

$$Skewness = \frac{3(\bar{x} - median)}{\sigma}$$

The following are other measures of skewness:

1. **Quartile coefficient of skewness:** This is given by;

$$Qcoefficientofskewness = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1}$$

1. **Percentile coefficient of skewness:** This is given by ;

$$\text{P Coefficient of skewness} = \frac{(P_{90} - P_{50}) - (P_{50} - P_{10})}{P_{90} - P_{10}}$$

5.4.1 Moment as a Measure of Skewness

Moment can be used to measure the degree of Skewness of a distribution. It is the most accurate and widely used measure of skewness. It is given by the formula:

$$a_3 = \frac{m^3}{s^3} = \frac{m^3}{(\sqrt{m^3})^2}$$

Where

M_2 = the variance

M_3 = third moment about the mean

a_3 = moment coefficient of skewness

When

$a_3 > 0$, the distribution is positively skewed

$a_3 < 0$, the distribution is negatively skewed

$a_3 = 0$, the distribution is symmetrical

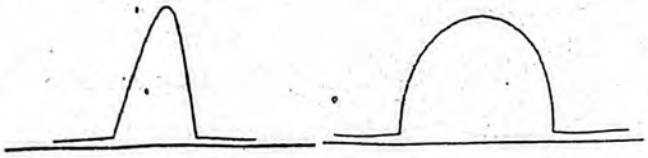
5.4.2 Applications

Skewness has benefits in many areas. Many models assume normal distribution; i.e., data are symmetric about the mean. A normal distribution has a skewness of zero. But in reality, data points may not be perfectly symmetric. So, an understanding of the skewness of a dataset indicates whether deviations from the mean are going to be positive or negative.

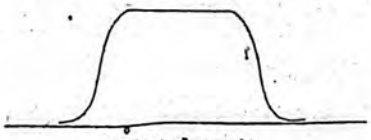
Skewed data arises quite naturally in various situations. Incomes are skewed to the right because even just a few individuals who earn millions of dollars can greatly affect the mean and there are no negative incomes. Similarly, data involving the lifetime of a product, such as a brand of light bulb, are skewed to the right. Here, the smallest that a lifetime can be is zero and long-lasting light bulbs will impart a positive skewness to the data.

5.5 KURTOSIS

In probability theory and statistics, **kurtosis** (from the Greek word , *kyrtos* or *kurtos*, meaning curved or arching) is any measure of the "Peakedness" of a probability distribution of a real-valued random variable. In a similar way to the concept of skewness, *kurtosis* is a descriptor of the shape of a probability distribution. Pearson's kurtosis, provide a comparison of the shape of a given distribution to that of the normal distribution. A distribution that is highly peaked is called a leptokurtic distribution. A distribution that is flatly peaked is called a Platykurtic distribution. While a distribution that is neither highly nor flatly peaked is called Mesokurtic distribution.



Leptokurtic Mesokurtic



Platykurtic

5.5.1 Measures of Kurtosis

One measure of kurtosis is based on both quartile and percentile and it is called the percentile coefficient of kurtosis. It is given by

$$K = \frac{Q}{P_{90} - P_{10}}$$

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

$P_{90} = 90^{th}$ *percentileand*

$P_{10} = 10^{th}$ *percentile*

5.5.2 Moment as Measures of Kurtosis

The fourth moment about the mean is the most frequent moment used in the measurement of kurtosis. It is given by

$$\beta_2 = a_2 = \frac{m_4}{S^4} = \frac{m_4}{m_2^2}$$

Where

M_4 = the fourth moment about the mean

M_2 = the variance

$\beta_2 = a_4$ = moment coefficient of kurtosis

When

$a_4 > 3$, the distribution is leptokurtic

$a_4 < 3$, the distribution is platykurtic and

$a_4 = 3$, the distribution is mesokurtic

Exercises 5

(1) The following shows the distribution of marks obtained by certain group of students.

Marks	0-5	5-10	0-15	15-20	20-25	25-30	30-35	35-40
Frequency	2	5	7	13	21	16	8	3

Compute mean, standard deviation and the coefficient of variation of the distribution.

(2) The following table shows the weight of the students of two classes.
Calculate the coefficient of variation of the two classes and which one is more viable.

Weight (in kg)	Class A	Class B
20-30	7	5
30-40	10	9
40-50	20	21
50-60	18	15
60-70	7	6

(3) Calculate the person's coefficient of skewness of the following distribution.

Wages	0-10	10-20	20-30	30-40	40-50
No. of Workers	15	20	30	25	10

(4) Calculate the person's coefficient of skewness of the distribution below:

Class Interval	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
Frequency	10	28	50	37	29	17	16	10	3

(5) Assume that a firm has selected a random sample of 100 from its production line and has obtained the data shown in the table below.

Class Interval	130-134	135-139	140-144	145-149	150-154	155-159	160-164
Frequency	3	12	21	28	19	12	5

Compute the following:

(a) Arithmetic mean (b) Standard deviation (c) Mode and

(d) Pearson's coefficient of skewness.

(6) Calculate the first four moments about the mean of the distribution Below;

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70
Frequency	5	12	18	40	15	7	3

(7) Given the distribution below.

Class Interval	20-30	30-40	40-50	50-60	60-70	70-80	80-90
Frequency	5	14	20	25	17	11	8

Compute the first four moments about the mean.

(8) Given the following set of data

2, 3, 5, 7, 5, 3, 8, 5, 7, 5, 7, 3

Obtain:

(a) The first moment about the origin.
(b) The second moment about the origin.
(c) The third moment about the origin.
(d) The second moment about the mean.
(e) The third moment about the mean.

(9) Given the distribution below;

X	2	3	5	7	8
Frequency	1	3	4	3	1

Obtain:

(i) The first moment about the origin.
(ii) The second moment about the origin.
(iii) The third moment about the origin.
(iv) The second moment about the mean.
(v) The third moment about the mean.

(10) Given the following distribution.

Class Interval	20-29	30-39	40-49	50-59	60-69
Frequency	2	4	10	6	3

Find the:

(a) Second moment about the origin
(b) Second Moment about the mean
(c) Third Moment about the mean
(d) Fourth moment about the mean.

(11) Given the following set of numbers

2 3 5 8 6 2 6 4 8 6

Obtain the:

(I) First moment (ii) Second moment and

(iii) Fourth moment.

(12) Given the distribution below.

Class Interval	20-24	25-29	30-34	35-39	40-44	45-49
Frequency	3	5	10	7	4	3

Calculate the:

(a) Second moment
(b) Third moment and
(c) Fourth moment

(13). Given the distribution below:

Class Interval	60-64	65-69	70-74	75-79	80-84	85-89	90-94	95-9
Frequency	4	3	2	4	6	8	1	5

Calculate:

(a) Pearson's first coefficient of skewness
(b) Pearson's second coefficient of skewness.
(c) Quartile coefficient of skewness.
(d) Percentile coefficient of kurtosis.

(14). Given

X	10-14	15-19	20-24	25-29	30-34	35-39	40-4
F	8	2	1	5	10	2	16

Calculate:

(a) Pearson's first coefficient of skewness.
(b) Pearson's second coefficient of skewness.
(c) Quartile coefficient of skewness.
(d) Percentile coefficient of skewness.

(15). Given the distribution below.

Class Interval	1-19	20-29	30-39	40-49	50-59	60-
Frequency	2	4	5	10	2	4

Calculate:

(a) Pearson's First coefficient of skewness
(b) Pearson's second coefficient of skewness
(c) Quartile coefficient of skewness
(d) Percentile coefficient of skewness

(16). Given the distribution below

Class Interval	6-8	9-11	12-14	15-17	18-20
Frequency	5	0	20	17	2

Calculate:

(a) Pearson's First coefficient of skewness
(b) Pearson's second coefficient of skewness

CHAPTER TWO

DESCRIPTIVE STATISTICS

2.1 INTRODUCTION

As earlier stated, descriptive statistics also called quantitative statistics is used to describe or summarize data. Descriptive statistics describe Statistics; it summarizes a collection of data as a description rather than using the data to learn about the field in which the data represents. Average, standard deviation, frequency and percentage are all examples of descriptive statistics. For example, the sports sections of newspapers tend to give description on the performance of many players such as the amount of goals scored compared with the amount of goal attempts. This is a descriptive statistic as it summarizes the data in a quantitative manner.

2.2 PRESENTATION OF STATISTICAL DATA

In statistics, data are raw information collected from the field which are not yet organized or processed. It may be grouped data or ungrouped data. The singular of data is datum.

Basically, there are three main methods of representing data and they are:

2.2.1 Arrangement of the Data

This is an orderly way of arranging statistical data and it can take any of the following forms; ascending order (from lowest to the highest) or descending order (from highest to the lowest).

Example 2.1

Arrange the following statistical data in:

a. Ascending and

b. Descending order

3, 7, 1, 10, 5, 2, 8, 6, 4, 9

SOLUTION

a. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10

b. 10, 9, 8, 7, 6, 5, 4, 3, 2, 1

Example 2.2
Arrange the following statistic data in
a. Ascending order
b. Descending order
7, 10, 0, -1, 3 -4, 5, 9,

SOLUTION
a. -4, -1, 0, 3, 5, 7, 9, 10
b. 10, 9, 7, 5, 0, -1, -4

Example 2.3
Arrange the following statistical data in an array

3,	8,	-5,	1;	-13,
2,	-7,	17,	-9,	20
-14,	18,	-25,	23,	-19

Ascending order
-25, -19, -14, -13, -9, -7, -5, 1, 2, 3, 8, 17, 18, 20, 23,

Descending order
23, 20, 18, 17, 8, 3, 2, 1, -5, -7, -9, -13, -14, -19, -25

2.3 TABULATION OF STATISTICAL DATA

This is the representation of statistical data in the form of a table. When data are tabulated, they are put in a table called frequency table. In a frequency table, we collect like quantities and display them by writing down the number of times each quantity appears (frequency). When the data is large, we group it so as to be contained in a table and they are called group data.

2.3.1 Frequency Distribution Table

This is the tabular representation of statistical data to include set of numbers or class interval tally representation and frequency.

Example 2.4
20 sweets are shared among some students; Musa got 5 sweets, Abu got 3 sweets, John got 7 sweets, Yusuf got 4 sweets and Tunde got 1 sweet. Represent the above information on a frequency table.

SOLUTION

Table 2.1

Name	Number of Sweets (Frequency)
Musa	5
Abu	3
John	7
Yusuf	4
Tunde	1

Example 2.5

Represent the following data on a frequency table; 3, 1, 2, 2, 1, 3, 2, 1, 4, 2, 3, 5, 2, 3, 4

SOLUTION

By arranging the numbers in order

1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 3, 4, 4, 5

Numbers	Tally	Number of Time it Occur (Frequency)
1	III	3
2	~~IIII~~	5
3	IIII	4
4	II	2
5	I	1

Example 2.6

Represent the following numbers in a frequency table

2	3	6	4	4	4	2	5	1	3
3	6	6	4	5	5	2	2	5	1
1	6	4	3	3	2	2	4	4	5
5	3	3	2	2	3	1	5	3	4
1	6	2	1	1	4	4	6	5	3

SOLUTION

Table 2.2

Frequency Distribution Table

Numbers	Tally	Number of Time it Occur (Frequency)
1	~~IIII~~ II	7
2	~~IIIIII~~	9
3	~~IIIIII~~	10
4	~~IIIIII~~	10
5	~~IIII~~ III	8
6	~~IIII~~	6

Note that when representing data that is large in a frequency table, it shall be grouped so as to be contained by the table or else it will be cumbersome. For instance, representing hundred data may require about 100 spaces which cannot be contained by the paper itself: instead we group it inform of class interval. *A class interval* is the total number of items that belongs to each class. This is illustrated in the example below.

Example 2.7

Construct a frequency distribution table to represent the following data using the class intervals 15-19, 20-24…etc.

26,	37,	31,	21,	29,	18,	39,	33,	28,	22,
33,	22,	21,	16,	36,	28,	29,	20,	24,	32,
34,	24,	25,	27,	32,	23,	27,	26,	28,	38,
25,	28,	30,	31,	29,	27,	20,	27,	29,	26

Table 2.3

Frequency Distribution Table

Class Interval	Tally	Frequency
15 – 19	II	2
20 – 24	~~IIIIII~~	9
25 – 29	~~IIIIIIIIII~~ II	17
30 – 34	~~IIII~~ III	8
35 – 39	IIII	4
		40

Example 2.8

Construct a frequency distribution table using the class intervals 1–5, 6-10, etc. to represent the following set of numbers:

14,	28,	1,	3,	12,	11,	10,	21,	2,	6,
6,	4,	21,	16,	11,	14,	3,	9,	22,	5,
13,	19,	19,	24,	14,	13,	16,	14,	26,	21,
4,	16,	25,	22,	12,	15,	3,	14,	21,	6,

Table 2.4

Frequency Distribution Table

Class Interval	Tally	Frequency
1 – 5	~~IIII~~ III	8
6 – 10	~~IIII~~	5
11 – 15	~~IIIIIII~~ II	12
16 – 20	~~IIII~~	5
21 – 25	~~IIII~~ III	8
25 – 30	II	2
		40

Example 2.9

The height of the members of a family is measured in centimeter and is given below:

150	100	111	136	80	140	90	107	146	85
119	120	150	120	93	132	97	85	127	130
85	103	126	125	90	121	85	87	80	149

Use a class interval of 80 – 89, 90 – 99, 100 – 109, etc. to represent their heights in a frequency table.

SOLUTION

Table 2.5

Frequency Distribution Table

Height	Tally	Number of Time it Occur (Frequency)
80 – 89	𝍸 II	7
90 – 99	IIII	4
100 – 109	III	3
110 – 119	II	2
120 – 129	𝍸 I	6
130 – 139	III	3
140 – 149	III	3
150 – 159	II	2

2.4 DIAGRAMMATIC REPRESENTATION OF STATISTICAL DATA

This is the representation of statistical data using diagrams such as pictogram, bar chart, histogram, pie chart etc.

2.4.1 Bar Chart

This is the representation of statistical data using bars or rectangles. In this representation, the bars or rectangles are separated by equal space, with the height or vertical axis corresponding to the class frequency and the horizontal axis corresponding to the numbers or class intervals.

Example 2.10

Represent the information below in a bar chart

Items	Omo	Sugar	Magi	Soap	Salt
Price (₦)	200	400	150	100	50

SOLUTION

Figure 2.1 of Price Against Items

Numbers	1	2	3	4	5	6
Frequency	3	7	9	8	5	1

SOLUTION

Figure 2.2 of Number against Frequency

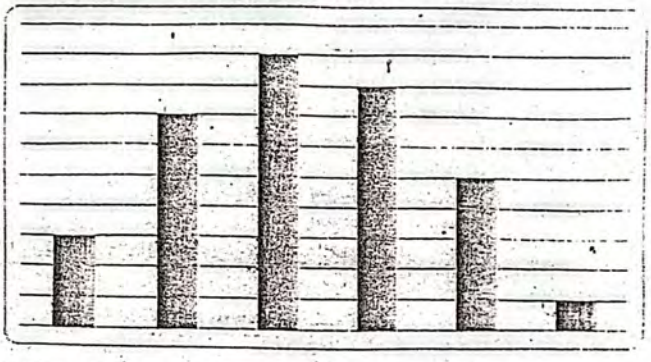


Figure 2.1 Showing a Bar Chart of Essential Commodities

Example 2.11

Draw a bar chart to represent the following data

SOLUTION

Figure 2.3 of Class Interval against Frequency

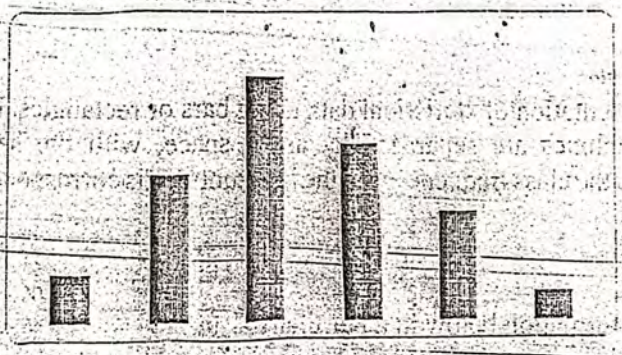


Figure2.3 Showing a Bar Chart of Class Interval against Frequency

Example 2.13

Draw a bar chart to represent the distribution below

Class interval	1 – 10	11 – 20	21 – 30	31 – 40	41 - 50
Frequency	5	35	50	20	10

SOLUTION

Figure 2.4 Class Intervals against Frequency

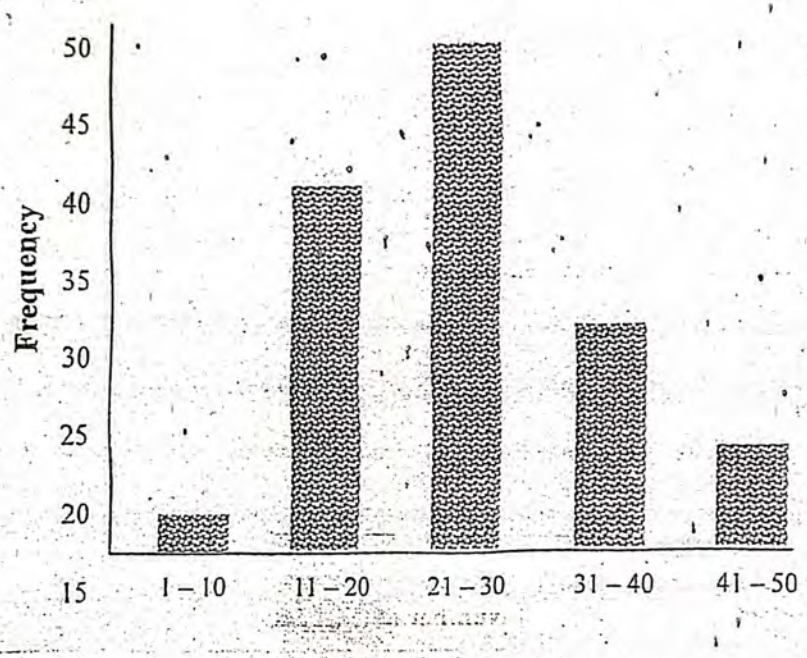


Figure 2.4 Showing a Bar Chart of Class Interval against Frequency

2.4.2 Pictogram

Pictogram is the representation of statistical data using pictures. It is considered to be the simplest method of data representation. Pictograms are commonly used by newspaper agency, journalist and advertisers among others.

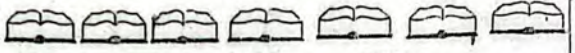
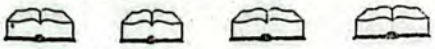
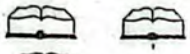
Example 2.14

A student has 5 Mathematics text books, 3 English text books, 7 Integrated Science textbooks, 4 Agric. Textbooks, 2 Health Education textbooks and 1 Social Studies textbook. Represent the text books using a pictogram.

SOLUTION

Table 2.6

Pictogram of Textbooks of Different Subjects owned by a Student

Subject	Number of Textbook
Mathematics	
English	
Integrated Science	
Agric. Science	
Health Science	
Social Studies	

Example 2.15

In a school, there are 7 English Teachers, 5 Mathematics Teachers, 3 Hausa Teachers, 9 Agric science Teachers, 2 Science Teachers and 1 Social Study Teacher. Represent the above information using a pictogram.

SOLUTION

Table 2.7

Pictogram of Teachers in a School

Subject	Number of teachers
English	
Mathematics	
Hausa	
Agric. Science	
Science	
Social Studies	

2.4.3 Histogram

This is the same as bar chart. But in the case of histogram, the bars or rectangles are not separated. It is plotted frequency against items. However, for group data it is plotted frequency against class boundary. The class boundary is obtained by subtracting 0.5 from the lowerclass interval and adding the same 0.5 to the upper class interval. For example, given a class interval 80 – 89, the lower class interval is 80 and the upper class interval is 89, then the class boundary is 79.5 – 89.5. The class boundary is what makes the bars of the histogram to be joined by common boundary (without any space).

Class interval	2 – 4	5 – 7	8 – 10	11 – 13	14 – 16
Frequency	2	8	12	5	4

SOLUTION

Table 2.8

Frequency Distribution Table

Class Interval	Class Boundary	Frequency
2 – 4	1.5 – 4.5	2
5 – 7	4.5 – 7.5	8
8 – 10	7.5 – 10.5	12
11 – 13	10.5 – 13.5	5
14 – 16	13.5 – 16.5	4

Figure 2.5 of Class Boundary against Frequency

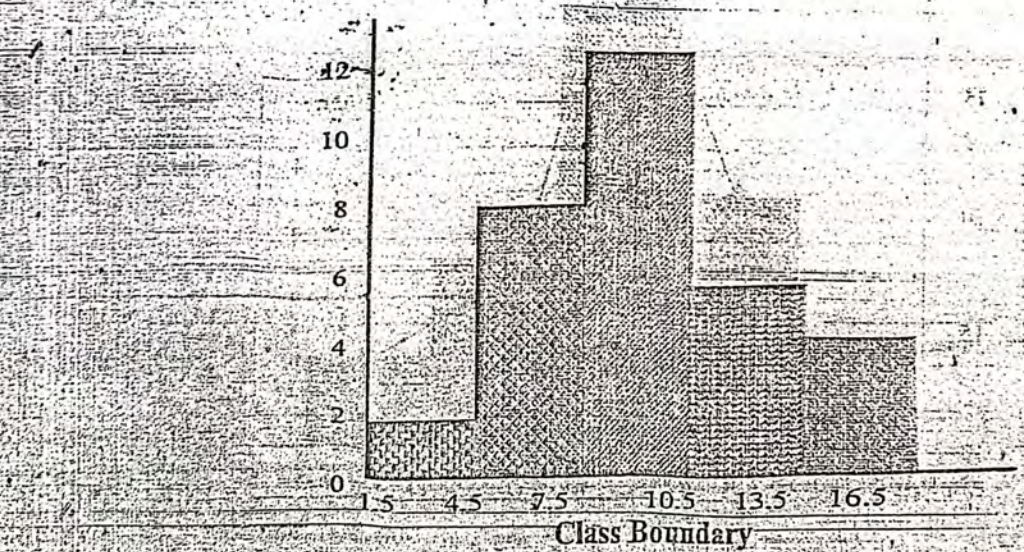


Figure 2.5 Showing Class against Frequency

2.4.4 Frequency Polygon

This is a line graph of class mark (mid points) against the class frequency. The other way of drawing a frequency polygon, is to construct a histogram, mark the mid points or the top of the bars or rectangles and join them together.

Example 2.17

Draw a frequency polygon to represent the following distribution.

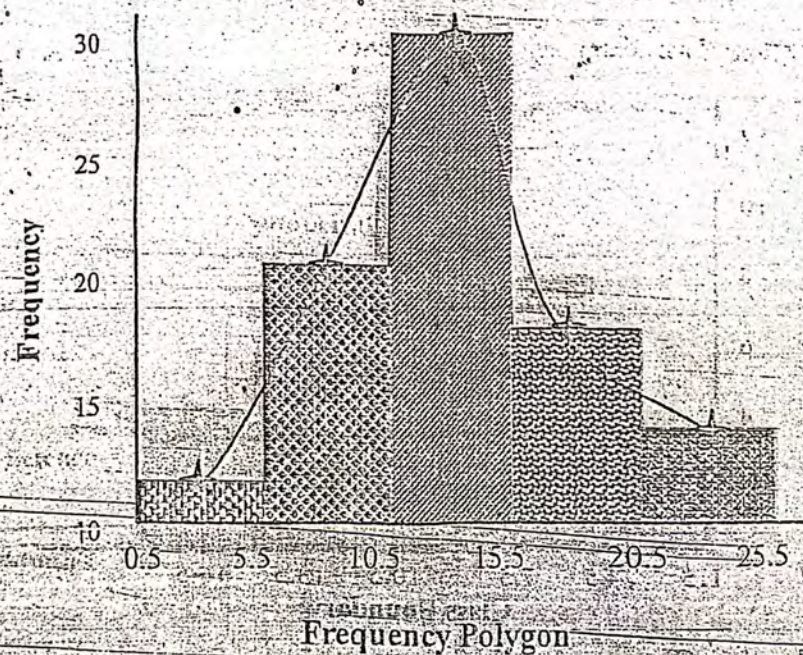
Class interval	1 – 5	6 – 10	11 – 15	16 – 22	21 – 25
Frequency	2	17	30	11	5

SOLUTION

Table 2.9 Frequency Distribution Table

1 – 5	0.5 – 5.5	2
6 – 10	5.5 – 10.5	17
11 – 15	10.5 – 15.5	30
16 – 20	15.5 – 20.5	11
21 – 25	20.5 – 25.5	5

Figure 2.6 of Class Mark against Class Frequency



Frequency Polygon

Example 2.18

Draw a frequency polygon to compare the age distribution of the teachers in a college.

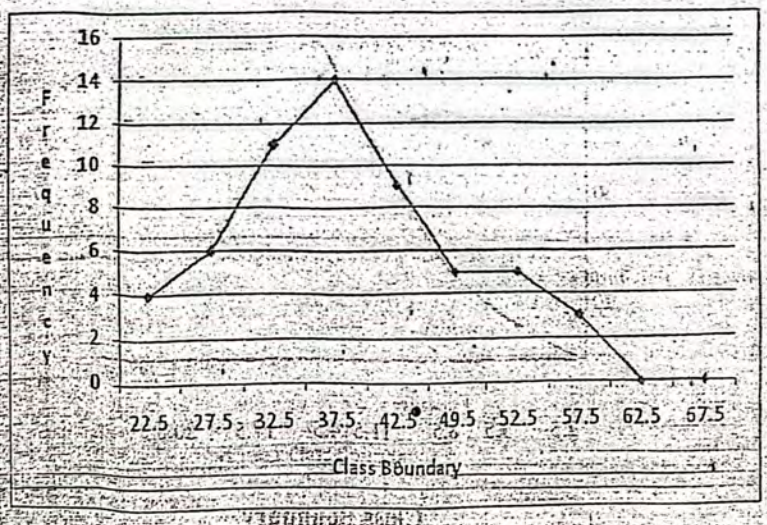
Age	20-25	25-30	30-35	35-40	40-45	45-50	50-55	55-60	60-65	65-70
Freq	4	6	11	14	9	5	5	3	0	0

SOLUTION

Table 2.10 of Class Mark against Frequency

Mid Value	Frequency
22.5	4
27.5	6
32.5	11
37.5	14
42.5	9
49.5	5
52.5	5
57.5	3
62.5	0
67.5	0

Figure 2.7 of Class Mark against Frequency



2.4.5 Cumulative Frequency Curve (ogive)

This is a line graph of cumulative frequency on the vertical axis against the class boundary on the horizontal axis. It is also called an ogive.

Example 19 Construct a cumulative frequency curve to represent the following distribution.

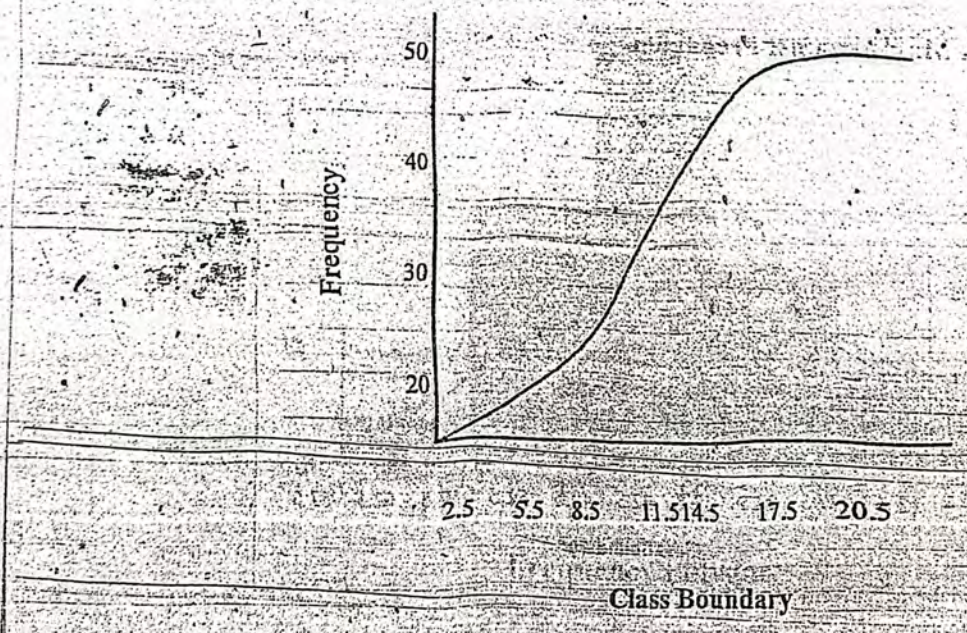
Class Interval	3 – 5	6 – 8	9 – 11	12 – 14	15 – 17	18 – 20
Frequency	3	10	18	10	7	2

SOLUTION

Table 2.11 Frequency Distribution Table

Class Interval	Class boundary	Frequency	Cumulative frequency
3 – 5	2.5 – 5.5	3	3
6 – 8	5.5 – 8.5	10	13
9 – 11	8.5 – 11.5	18	31
12 – 15	11.5 – 14.5	10	41
15 – 17	14.5 – 17.5	7	48
18 – 20	17.5 – 20.5	2	50

Figure 2.8 Cumulative Frequency Curve



Example 2.20

Six weeks after planting, the heights of 30 broad bean plants were measured and the frequency of the distribution formed are as shown below, represent the data using a cumulative frequency curve.

Height

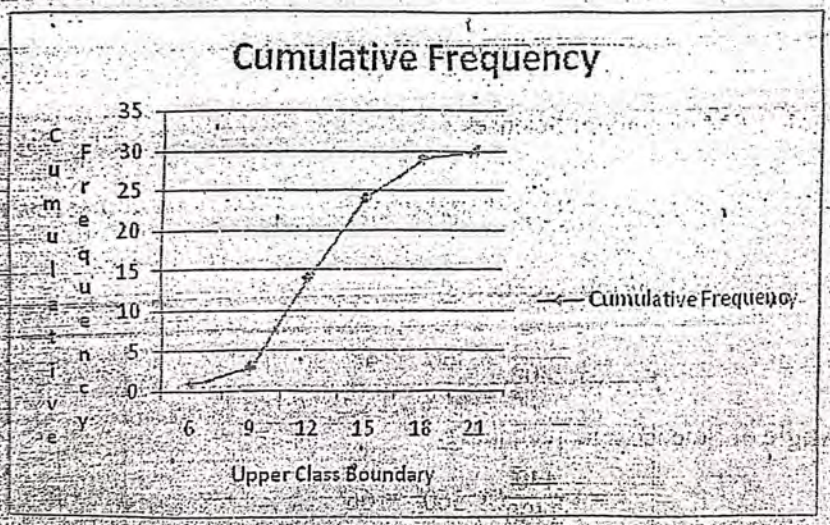
Height (Cm)	3-6	6-9	9-12	12-15	15-18	18-21
Frequency	1	2	11	10	5	1

SOLUTION

Table 2.12 Frequency Distribution Table

Class Boundary	Cumulative Frequency
3-6	1
6-9	3
9-12	14
12-15	24
15-18	29
18-21	30

Figure 2.9 Cumulative Frequency Curve



2.4.6 Pie Chart

This is a circular representation of a statistical data. The circle is subdivided into sectors and each sector corresponds to the class frequency. The sector angle corresponding to each sector can be obtained by using the formula.

$$\text{Sector Angle} = \frac{\text{class frequence}}{\text{total frequence}} \times 360^0 = - \times 360^0$$

Where a=Class frequency

n=Total frequency

Example 2.21

The following are the numbers of students admitted in a school.

Schools	No of Students Admitted
Management Science	780
Agricultural Science	520
Engineering Technology	300
Science and Technology	440.
Remedial	240
IJMB	120
Total	**2400**

Draw a pie chart to represent the above information.

SOLUTION

Using the formula

$$\text{Sector Angle} = \frac{\text{class frequence}}{\text{total frequence}} x\ 360^0$$

Sector Angle of Management Sciences

$$= \frac{780}{2400} x\ 360^0 = 117^0$$

Sector Angle of Engineering Technology

$$= \frac{300}{2400} x\ 360^0 = 45^0$$

Sector Angle of Science and Technology

$$= \frac{440}{2400} x\ 360^0 = 66^0$$

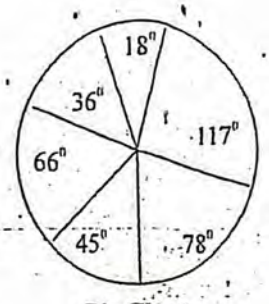
Sector Angle of Remedial

$$= \frac{240}{2400} \times 360^0 = 36^0$$

Sector Angle of IJMB

$$= \frac{120}{2400} \times 360^0 = 18^0$$

Figure 2.10 Pie Chart



Pie Chart

Exercise 2

1) In a survey the masses of 50 apples were noted and recorded in the following table.

Each value was given to the nearest gram

101,114,118,87,92,93,116,105,102,97,93,101,111,96,117,100,106,
118,101,107,96,101,102,104,92,99,107,98,105,113,100,103,108,92,
109,95,100,103,110,113,99,106,116,101,105,86,88,108,92,86

a) Construct a frequency distribution table using an appropriate class width
b) Construct a histogram
c) Construct a bar chart

2) Draw a histogram to show the masses, to the nearest kilogram of the 200 girls shown below

Mass(kg)	Frequency
41-50	21
51-55	62
56-60	55
61-70	50
71-75	12

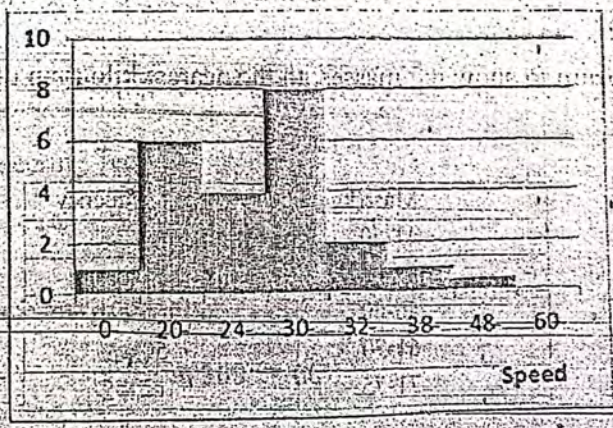
3) A researcher timed how long it took for each of 38 volunteers to perform a simple tas results are shown in the table below. Draw a histogram to illustrate the data

Time	Frequency
1-5	2
6-10	12
11-15	7
16-20	15
21-25	2
26-30	6

4) On a particular day the length of stay of each car at a city park was recorded, represe data by a histogram and state the modal class

Length of stay(min)	Frequency
$t<25$	62
$25<t<60$	70
$60<t<80$	88
$80<t<150$	280
$150<t<300$	30

5) The histogram below represents the data gotten from the speed of cars passing a 30miles per hour sign, write out the frequency distribution



6) In a competition to grow the tallest hollyhock, the heights recorded by 50 primary school children were as follows, represent the data on a histogram.

Height (cm)	Frequency
177 - 186	12
187 - 191	8
192 - 196	8
197 - 201	9
202 - 206	7
207 - 216	6

7) The number of times the letter 'e' appears in each sentence in an article called 'My kind of day' is shown below. Make a grouped frequency distribution and draw a histogram
15, 12, 8, 12, 3, 10, 14, 17, 5, 3, 8, 11, 7, 16, 5, 13, 12, 11, 6, 7, 4, 17, 8, 1

8) The following data represents the heights of a maize plant grown in a special resort. Represent the data using a histogram.
46 46 46 46 46 46 46 47 47 47 47 47 47 47 47 47 47 47 48 48 48 48 48 48 48 48 48 48 48 48 49 49 49 49 49 49 49 49 49 49 49 49 49 50 50 50 50 50 50 50 50 50 50 50 50 50 51 51 51 51 51 51 51 51 51 51 51 51 51 51 51 51 51 51 52 52 52 52 52 52 52 52 52 52 52 53 53 53 53 53 53 53 53 54 54 54 55

9) The length of a room was measured by 100 students using a ruler in an interactive class work activity; draw a histogram to represent the data below
51 53 53 54 54 54 54 54 55 55 55 55 55 55 55 55 55 56 56 56 56 56 56 56 56 56 56 56 56 56 56 56 56 56 56 57 57 57 57 57 57 57 57 57 57 57 57 57 57 57 57 57 57 57 57 57 57 57 57 58 58 58 58 58 58 58 58 58 58 58 58 58 58 58 58 58 58 59 59 59 59 59 59 59 59 59 59 59 59 59 59 60 60 60 60 60 60 60 60 60 60 60 60 60

10) The data below represent the weight of eggs (in grams) gotten from a poultry farm in Kaduna state; represent the data by a histogram
55 55 56 56 57 57 58 58 58 58 58 58 58 59 59 59 59 59 59 59 59 59 59 60 60 60 60 60 60 60 60 60 60 60 60 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 62 62 62 62 62 62 62 62 62 62

62 62 62 62 62 62 62 62 63 63 63 63 63 63 63 63 63 63 63 63 63 63 63 64 64 64 64 64 64 64
64 64 65 65 65 65 65

11)The table below shows the duration in minutes of 64 telephone calls made from a high street call box in a day.

Length of call (mins)	Frequency
0 - 1.5	3
1.5 -3.0	7
3 - 4.5	22
4.5 - 6.0	20
6.0 - 7.5	6
7.5 -9.0	6
9.0 - 10.5	0

Draw a frequency polygon to illustrate the data.

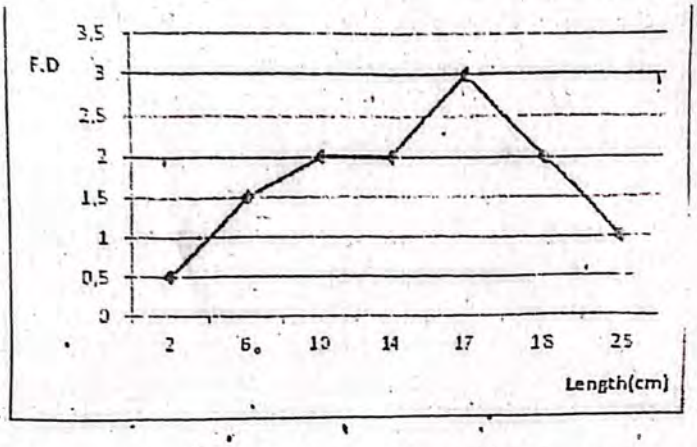
12)The table below shows the ages, in completed years of women who gave birth to a child at a Maternity Hospital during a particular year. Without drawing a histogram first, draw a frequency polygon to illustrate the information. Describe the distribution

Age (years)	Number of births
16 - 20	70
20 - 25	470
25 - 30	535
30 - 35	280
35 - 40	118
40 - 45	0

13) The patients at a chest clinic were asked to keep a record of the number of cigars the smoked each day. Construct a frequency polygon to represent the data

Number of cigars smoked per day	Frequency
0 - 9	5
10 - 14	8
15 - 19	32
20 - 29	41
30 - 39	16
40 and over	2

14)



Copy and complete the frequency distribution represented by the polygon shown above.

Length(cm)	Frequency
0-4	2
4-8	*
8-12	*
12-16	*
16-18	*
18-20	*
20-30	*

15) Lucy and Jack play a computer game every day and keep record of their scores. Lucy's scores are shown in the table below. Draw a frequency polygon to represent her scores.

Lucy's scores	50 - 99	100 - 149	150 - 199	200 - 249	250 - 299
Frequency	6	14	10	6	4

Jack's scores are as follows

Lucy's scores	50 - 99	100 - 149	150 – 199	200 - 249	250 - 299
Frequency	2	6	10	16	6

Draw a frequency polygon for Jack's scores on the same set of axes as Lucy's and use it to compare the two set of scores.

16) There are 28 pupils in Peter's class. He carried out a survey of how pupils in his class travelled to school. His results are shown in the table below

Method of Travel	Number of Pupils
Bus	12
Car	2
Bicycle	5
Walking	9

Draw a pie chart to represent the above information.

17) The following data summaries the expenditure by a country council during a particular year

Service	Expenditure ($m)
Education	160.2
Highway & Public Transport	25.7
Police	28.9
Social Services	27.9
Other	24.5

Draw the pie chart and calculate to the nearest degree, the angle corresponding to each of the five classifications.

18) Five companies form a group, the sales of each company during the year ending 5th April, 1988, are shown in the table below

Company	A	B	C	D	E
Sales (in $1000)	55	130	20	35	60

Draw a pie chart to illustrate this information.

19) The table below shows the sources and the corresponding amounts of income obtained from a charity

Construct a cumulative frequency table and the ogive curve.

22) The table below shows the frequency distribution of the masses of 52 women students at a college.

Mass (kg)	Frequency
40 - 44	3
45 - 49	2
50 - 54	7
55 - 59	18
60 - 64	18
65 - 69	3
70 - 74	1

Construct a cumulative frequency table and draw a cumulative frequency curve and use it to find how many students weighed less than 57kg. How many students weighed heavier than 61kg? If 20% were heavier than x kg, find the value of x.

23) Fifty soil samples were collected in an area of woodland and the pH value for each sample was found. The cumulative frequency distribution was constructed as shown in the table.

pH value	Cumulative Frequency
< 4.8	1
< 5.2	2
< 5.6	5
< 6	10
< 6.4	19
< 6.8	38
< 7.2	43
< 7.6	46
< 8	49
< 8.4	50

Draw a cumulative frequency curve.

24) In a quality control survey, the length of life in hours of 50 light bulbs is noted. The results are summarized in the table below, find the median and interquartile range.

Length of Life (h)	Frequency
650<h<670	3
670<h<680	7
680<h<690	20
690<h<700	17
700<h<710	3

25) The table below shows the orbital velocities of our solar system given to the nearest m/s. Graph the data using a bar chart.

Planet	Mercury	Venus	Earth	Mars	Jupiter	Saturn	Uranus	Nep	Pluto
V(m/s)	30	22	19	15	8	6	4	3	3

26) Shown below is the annual population of randomly selected countries. Draw a pie chart to represent the information.

Country	China	India	U.S	Indonesia	Brazil	Russia	UAE
Population (InMillions)	1222	968	268	210	165	186	132

27) In one month, a student recorded the length, to the nearest minute of each of the lectures she attended. The table below shows her data and the calculations she made. Calculate the values of a, b, c and the total number of lectures attended during the month.

Length of Lecture (min)	50 - 53	54 - 55	56 - 59	60 - 67
Number of Lectures	a	b	30	C
Freq. Density	5	13	7.5	1.5

28) Write short notes on the following
a. Class interval
b. Class boundary
c. Class mark
d. Class width
e. Frequency distribution table

29) The following are set of marks scored in a statistics test by students of Management Technology of a University

2	3	6	7	8	3	6	5	2	7	9	4
7	4	5	3	4	9	4	7	6	8	7	3
4	5	7	5	5	6	8	4	4	5	3	7
3	6	3	7	8	7	0	2	8	0	7	1
1	7	9	9	6	4	2	0	6	1	4	0

Construct a frequency distribution table for the set of data.

30) Given the distribution below

28	18	37	26	29	33	39	31	22
24	28	22	33	36	20	29	21	32
28	22	24	34	32	26	27	25	38

Construct a frequency distribution table for the set of data.

31) Arrange the following data in an array and construct a frequency distribution table using the class intervals 1 – 5, 6 – 10 etc. to represent the data.

6	28	1	12	10	11	21	2	3	14
5	4	21	11	3	14	9	22	16	6
21	19	19	14	16	13	14	26	24	13
6	16	25	12	3	15	14	21	22	4

Source	Income (N)
Advertising	30000
Donations	X
Fees	9000
Investments	3000
Sponsorship	10000

A pie chart was drawn to illustrate the data. Given that the angle of the sector representing Donations was 204o. Calculate the total income, the value of x, the angle of each of the remaining sectors

20) On a certain day, 125 people were seen buying one newspaper and were asked which newspaper they had bought. The results of the survey are shown in the table below

Newspaper	Number Bought
The Sun	10
Daily Trust	25
Vanguard	40
Arewa Daily	50

Calculate the angles and draw a pie chart to illustrate this data.

21) The frequency table below shows the number of goals scored in a netball game by Jemima in 25 games played

Number of Goals	Frequency
0	0
1	1
2	3
3	2
4	5
5	8
6	6

32) The scores of 50 students in a mathematics test are as follows:

61	62	69	63	61	63	70	68	67	62
62	64	64	65	65	64	69	66	65	64
65	66	61	66	70	66	64	70	62	67
67	70	68	67	69	67	63	61	70	68
69	61	70	62	62	68	61	69	68	65

Prepare a frequency distribution table to represent the set of numbers.

33) The following represent the votes cast in bye election of a local council in Nigeria.

Political Party		Votes
A	-	90
B	-	50
C	-	60
D	-	200
E	-	100
F	-	80
G	-	140

Represent the above data on a pie chat

34) In a recently concluded All Polytechnics Games competition the following are the number of students that represented the Excellent Polytechnic in the sporting activities;

Sport	Football	Boxing	Wrestling	Basketball	Table tennis
No. of Students	12	6	4	6	8

Represent the data with the aid of a pie chart and a bar chart.

35) Given the distribution below;

35) Given the distribution below;

Class Interval	1 - 10	11 - 20	21 - 30	31 - 40	41 - 50	51 - 60	61 - 70
Frequency	20	30	45	50	25	20	10