UNIT – I

1.1 INTRODUCTION

Statistics has been defined differently by different statisticians from time to time. These definitions emphasize precisely the meaning, scope and limitations of the subject. The reasons for such a variety definitions may be stated as follows:

- The field of utility of statistics has been increasing steadily.
- The word statistics has been used to give different meaning in singular (the science of statistical methods) and plural (numerical set of data) sense.

Definitions:

Webster:

"Statistics are the classified facts representing the conditions of the people in a State... specially those facts which can be stated in number or in tables of numbers or in any tabular or classified arrangement".

A.L. Bowley:

i. The science of counting

ii. The science of averages

iii. The science of measurements of social phenomena, regarded as a whole in all its manifestations.

iv. A subject not confined to any one science.

Yule and Kendall:

"By Statistics we mean quantitative data affected to a marked extent by multiplicity if causes"

A.M. Tuttle:

"Statistics are measurements, enumerations or estimates of natural phenomenon, usually systematically arranged, analysed and presented as to exhibit important inter-relationships among them".

Prof. Horace Secrist:

"Statistics may be defined as the aggregate of facts affected to a marked extent by multiplicity of causes, numerically expressed, enumerated of estimated according to a reasonable standared of accuracy, collected in a systematic manner, for a predetermined purpose and placed in relation to each other".

Croxton and Cowden:

Statistics may be defined as the science of collection, presentation analysis and interpretation of numerical data from the logical analysis. It is clear that the definition of statistics by Croxton and Cowden is the most scientific and realistic one. According to this definition there are four stages:

Collection of Data:

It is the first step and this is the foundation upon which the entire data set. Careful planning is essential before collecting the data. There are different methods of collection of data such as census, sampling, primary, secondary, etc., and the investigator should make use of correct method.

Presentation of data:

The mass data collected should be presented in a suitable, concise form for further analysis. The collected data may be presented in the form of tabular or diagrammatic or graphic form.

Analysis of data:

The data presented should be carefully analysed for making inference from the presented data such as measures of central tendencies, dispersion, correlation, regression etc.,

Interpretation of data:

The final step is drawing conclusion from the data collected. A valid conclusion must be drawn on the basis of analysis. A high degree of skill and experience is necessary for the interpretation.

Every day, we come across the different types of quantitative information in newspapers, magazines, over radio and television. For example, we may hear or read that population of India had increased at the rate of 2.5% per year (per annum) during the period 1981-1991, number of admission in National Open School had gone up by say 20% during 1996-97 as compared to 1995-96 etc. We would like to know that what these figures mean. These quantitative information or expression called statistical data or statistics.

Statistics is concerned with scientific methods for collecting, organising, summarising, presenting and analysing data as well as deriving valid conclusions and making reasonable decisions on the basis of this analysis. Statistics is concerned with the systematic collection of numerical data and its interpretation. The word 'statistic' is used to refer to

1. Numerical facts, such as the number of people living in particular area.

2. The study of ways of collecting, analysing and interpreting the facts.

1.2 ORIGIN

The science of statistics developed gradually and its field of application widened day by day. The subject of Statistics, as it seems, is not a new discipline but it is as old as the human society itself. It origin can be traced to the old days when it was regarded as the 'Science of State-craft' and was the by-product of the administrative activity if the State. The word 'Statistics' seems to have been derived from the latin word 'status' or the Italian word 'statista' or the German word 'statistik each of which means a 'political state'.

In ancient times, the government used to collect the information regarding the population and 'property or wealth' of the country-the former enabling the government to have an idea of the man-power of the country, and the latter providing it a basis for introducing new taxes and levies.

In India, an efficient system of collecting official and administrative statistics existed even more than 2,000 years ago, in particular, during the reign of Chandra Gupta Maurya (324-300 B.C). From Kautilya's

Arthshastra it is known that even before 300 B.C. a very good system of collecting 'Vital Statistics' and registration of births and deaths was in vogue. During Akbar's reign (1556-1605 A. D.), Raja, Todarmal, the then land and revenue minister, maintained good records of land and agricultural statistics.

In Alina-e-Akbari written by Abul Fazl (in 1596-97), one of the nine gems of Akbar, we find detailed accounts of the administrative and statistical surveys conducted during Akbar's reign. In Germany, the systematic collection of official statistics originated towards the end of the 18th century when, in order to have an idea of the relative strength of different German States, information regarding population and output-industrial and agricultural was collected.

In England, statistics were the outcome of Napoleonic wars. The wars necessitated the systematic collection of numerical data to enable the government to assess the revenues and expenditure with greater precision and then to levy new taxes in order to meet the cost of war. Seventeenth century saw the origin of the 'Vital Statistics'. Captain John grant of London (1960-1674), known as the 'father' of Vital Statistics, was the first man to study the statistics of births and deaths.

The theoretical development of the so-called, modern statistics came during the mid-seventeenth century with the introduction of *'Theory of Probability'* and *'Theory of Games and Chance'*. The chief contributors being mathematicians and gamblers of France, Germany and England. Statistics is an old science, originated during the time of Mahabharat. For the last few countries, it has remained a part of mathematicians like Pascal (1623-1662), James Bernouli (1654-1705), De Moivre (1667-1754), Laplace (1749-1827), gauss (1777-1855), lagrange, Bayes, Markoff, Euler etc. These mathematicians were mainly interested in the development of the theory of probability as applied to the theory of games and other chance phenomena. Till the early nineteenth century, statistics was mainly concerned population and area of land under cultivation, etc., of a state or kingdom.

A Ronald. A. Fisher (1890-1962) who applied statistics to a variety of diversified fields such as genetics, biometry, psychology and education, agriculture, etc and which is rightly termed as the Father of Statistics. His contributors to the subject of Statistics are described in the following words:

'R.A. Fisher is the real giant in the development of the theory of Statistics'

The varied and outstanding contributions of R.A. fisher put the subject of Statistics on a very firm footing and earned for it the status of fully fledged science.

Indian statisticians have also made notable contributions to the development of Statistics in various diversified fields. The valuable contributions of P.C. Mahalanobis and P.V. Sukhatme (Sample Surveys); R.C. Bose, Panse, J.N. Srivatsva (Design of experiments in Agriculture); S.N. Roy (Multivariate Analysis); C.R. Rao (Statistical Inference); Parthasarathy (Theory of Probability), to mention only a few, have earned for India a high position in the world map of Statistics.

1.3 FUNCTIONS OF STATISTICS

Statistics is viewed not as a more device for collecting numerical data but as a means of developing sound techniques for their handling and analysis and drawing valid inferences from them.

We know discuss briefly the functions of statistics. Let us consider the following important functions.

It simplifies facts in a definite Form:

Any conclusions stated numerically are definite and hence more convincing than conclusions stated qualitatively. Statistics presents facts in a precise and definite form and thus helps for a proper comprehension of what is stated.

Condensation:

The generally speaking by the word 'to condense', we mean to reduce or to lessen. Condensation is mainly applied at embracing the

understanding of a huge mass of data by providing only few observations. If in a particular class in Chennai School, only marks in an examination are given, no purpose will be served. Instead if we are given the average mark in that particular examination, definitely it serves the better purpose. Similarly the range of marks is also another measure of the data. Thus, Statistical measures help to reduce the complexity of the data and consequently to understand any huge mass of data.

It facilitates Comparison:

The various statistical methods facilitate comparison and enable useful conclusions to be drawn. The classification and tabulation are the two methods that are used to condense the data. They help us to compare data collected from different sources. Grand totals, measures of central tendency measures of dispersion, graphs and diagrams, coefficient of correlation etc provide ample scope for comparison. If we have one group of data, we can compare within itself. If the rice production (in Tonnes) in Tanjore district is known, then we can compare one region with another region within the district. Or if the rice production (in Tonnes) of two different districts within Tamilnadu is known, then also a comparative study can be made. As statistics is an aggregate of facts and figures, comparison is always possible and in fact comparison helps us to understand the data in a better way.

It helps in the formation of policies:

Scientific analysis of statistical data constitutes the starting point in all policy making. Decisions relating to import and export of various commodities, production of particular products etc., are all based on statistics.

It helps in Forecasting:

Plans and policies of organizations are invariably formulated well in advance of the time of their implementation. The word forecasting, mean to predict or to estimate beforehand. Given the data of the last fifteen years connected to rainfall of a particular district in Tamilnadu, it is possible to

predict or forecast the rainfall for the near future. In business also forecasting plays a dominant role in connection with production, sales, profits etc. Analysis of time series and regression analysis plays an important role in forecasting.

Estimation:

One of the main objectives of statistics is drawn inference about a population from the analysis for the sample drawn from that population. The four major branches of statistical inference are

- Estimation theory
- Tests of Hypothesis
- Non Parametric tests
- Sequential analysis

In estimation theory, we estimate the unknown value of the population parameter based on the sample observations.

1.4 SCOPE OF STATISTICS

There are many scopes of statistics. Statistics is not a mere device for collecting numerical data, but as a means of developing sound techniques for their handling, analysing and drawing valid inferences from them. Statistics is applied in every sphere of human activity – social as well as physical – like Biology, Commerce, Education, Planning, Business Management, Information Technology, etc. It is almost impossible to find a single department of human activity where statistics cannot be applied. We now discuss briefly the applications of statistics in other disciplines.

Statistics and Industry:

Statistics is widely used in many industries. In industries, control charts are widely used to maintain a certain quality level. In production engineering, to find whether the product is conforming to specifications or not, statistical tools, namely inspection plans, control charts, etc., are of

extreme importance. In inspection plans we have to resort to some kind of sampling – a very important aspect of Statistics.

Statistics and Commerce:

Statistics are lifeblood of successful commerce. Any businessman cannot afford to either by under stocking or having overstock of his goods. In the beginning he estimates the demand for his goods and then takes steps to adjust with his output or purchases. Thus statistics is indispensable in business and commerce. As so many multinational companies have invaded into our Indian economy, the size and volume of business is increasing. On one side the stiff competition is increasing whereas on the other side the tastes are changing and new fashions are emerging. In this connection, market survey plays an important role to exhibit the present conditions and to forecast the likely changes in future.

Statistics and Agriculture:

Analysis of variance (ANOVA) is one of the statistical tools developed by Professor R.A. Fisher, plays a prominent role in agriculture experiments. In tests of significance based on small samples, it can be shown that statistics is adequate to test the significant difference between two sample means. In analysis of variance, we are concerned with the testing of equality of several population means.

For an example, five fertilizers are applied to five plots each of wheat and the yield of wheat on each of the plots are given. In such a situation, we are interested in finding out whether the effect of these fertilisers on the yield is significantly different or not. In other words, whether the samples are drawn from the same normal population or not. The answer to this problem is provided by the technique of ANOVA and it is used to test the homogeneity of several population means.

Statistics and Economics:

Statistics data and techniques of statistical analysis have proved immensely useful in solving a variety of economic problems.

Statistical methods are useful in measuring numerical changes in complex groups and interpreting collective phenomenon. Nowadays the uses of statistics are abundantly made in any economic study. Both in economic theory and practice, statistical methods play an important role.

Alfred Marshall said, "Statistics are the straw only which I like every other economists have to make the bricks". It may also be noted that statistical data and techniques of statistical tools are immensely useful in solving many economic problems such as wages, prices, production, distribution of income and wealth and so on. Statistical tools like Index numbers, time series Analysis, Estimation theory, Testing Statistical Hypothesis are extensively used in economics.

Statistics and Education:

Statistics is widely used in education. Research has become a common feature in all branches of activities. Statistics is necessary for the formulation of policies to start new course, consideration of facilities available for new courses etc. There are many people engaged in research work to test the past knowledge and evolve new knowledge. These are possible only through statistics.

Statistics and Planning:

Statistics is indispensable in planning. In the modern world, which can be termed as the "world of planning", almost all the organisations in the government are seeking the help of planning for efficient working, for the formulation of policy decisions and execution of the same. In order to achieve the above goals, the statistical data relating to production, consumption, demand, supply, prices, investments, income expenditure etc and various advanced statistical techniques for processing, analysing and interpreting such complex data are of importance. In India statistics play an important role in planning, commissioning both at the central and state government levels.

Statistics and Medicine:

Statistical tools are widely used in Medical sciences. In order to test the efficiency of a new drug or medicine, t-test is used to compare the efficiency of two drugs or two medicines; t-test for the two samples is used. More and more applications of statistics are at present used in clinical investigation.

1.5 LIMITATIONS AND MISUSES OF STATISTICS

Although statistics is indispensable to almost all sciences: social, physical and natural and very widely used in most of spheres of human activity. It suffers from the following limitations. Statistics with all its wide application in every sphere of human activity has its own limitations. Some of them are given below.

Statistics deals only with aggregate of facts and not with individuals:

Statistics does not give any specific importance to the individual items; in fact it deals with an aggregate of objects. Individual items, when they are taken individually do not constitute any statistical data and do not serve any purpose for any statistical enquiry.

Statistics does not study of qualitative phenomenon:

Since statistics is basically a science and deals with a set of numerical data, it is applicable to the study of only these subjects of enquiry, which can be expressed in terms of quantitative measurements. As a matter of fact, qualitative phenomenon like honesty, poverty, beauty, intelligence etc, cannot be expressed numerically and any statistical analysis cannot be directly applied on these qualitative phenomenons. Nevertheless, statistical techniques may be applied indirectly by first reducing the qualitative expressions to accurate quantitative terms. For example, the intelligence of a group of students can be studied on the basis of their marks in a particular examination.

Statistics laws are tone only on an average:

It is well known that mathematical and physical sciences are exact. But statistical laws are not exact and statistical laws are only approximations. Statistical conclusions are not universally true. They are true only on an average.

Statistics table may be misused:

Statistics must be used only by experts; otherwise, statistical methods are the most dangerous tools on the hands of the inexpert. The use of statistical tools by the inexperienced and untraced persons might lead to wrong conclusions. Statistics can be easily misused by quoting wrong figures of data.

Statistics is only, one of the methods of studying a problem:

Statistical method do not provide complete solution of the problems because problems are to be studied taking the background of the countries culture, philosophy or religion into consideration. Thus the statistical study should be supplemented by other evidences.

Statistics is only inappropriate information:

Unskilled, idle and inexperienced person often collect data. As a result, erroneous, puzzling and partial information is collected. As a result, very often improper decision is taken.

Statistics is purposive Misuses:

The most total limitation of statistics is that its purposive misuse. Very often erroneous information may be collected. But sometimes some institutions use statistics for self interest and puzzling other organizations.

1.6 COLLECTION OF DATA

Everybody collects, interprets and uses information, much of it in numerical or statistical forms in day-to-day life. It is a common practice that people receive large quantities of information everyday through conversations, televisions, computers, the radios, newspapers, posters, notices and instructions. It is just because there is so much information

available that people need to be able to absorb, select and reject it. In everyday life, in business and industry, certain statistical information is necessary and it is independent to know where to find it how to collect it. As consequences, everybody has to compare prices and quality before making any decision about what goods to buy. As employees of any firm, people want to compare their salaries and working conditions, promotion opportunities and so on. In time the firms on their part want to control costs and expand their profits.

One of the main functions of statistics is to provide information which will help on making decisions. Statistics provides the type of information by providing a description of the present, a profile of the past and an estimate of the future. The following are some of the objectives of collecting statistical information.

- To consider the status involved in carrying out a survey.
- To analyse the process involved in observation and interpreting.
- To describe the methods of collecting primary statistical information.
- To define and describe sampling.
- To analyse the basis of sampling.
- To describe a variety of sampling methods.

Statistical investigation is a comprehensive and requires systematic collection of data about some group of people or objects, describing and organizing the data, analyzing the data with the help of different statistical method, summarizing the analysis and using these results for making judgements, decisions and predictions. The validity and accuracy of final judgement is most crucial and depends heavily on how well the data was collected in the first place. The quality of data will greatly affect the conditions and hence at most importance must be given to this process and every possible precaution should be taken to ensure accuracy while collecting the data.

Nature of data:

It may be noted that different types of data can be collected for different purposes. The data can be collected in connection with time or geographical location or in connection with time and location. The following are the three types of data:

- Time series data.
- Spatial data.
- Spacio-temporal data.

Time series data:

It is a collection of a set of numerical values, collected over a period of time. The data might have been collected either at regular intervals of time or irregular intervals of time.

Example

The following is the data for the three types of expenditures in rupees for a family for the four years 2011,2012,2013,2014.

Year	Food	Education	Others	Total
2011	2000	3000	3000	8000
2012	2500	3500	3500	9500
2013	3000	2500	4000	9500
2014	3000	4000	5000	12000

Spatial Data:

If the data collected is connected with that of a place, then it is termed as spatial data. For example, the data may be

- Number of runs scored by a batsman in different test matches in a test series at different places

- District wise rainfall in Tamil Nadu
- Prices of silver in four metropolitan cities
- State wise population in Tamil Nadu

Example

The population of the southern states of India in 2011.

State	Population
Andhra Pradesh	84665533
Karnataka	61130704
Kerala	33387677
Pondicherry	1244464
Tamil Nadu	72138958

Spacio Temporal Data:

If the data collected is connected to the time as well as place then it is known as spacio temporal data.

Example

State	Population	
	1981	1991
Andhra Pradesh	5,34,03,619	6,63,04,854
Karnataka	3,70,43,451	4,48,17,398
Kerala	2,54,03,217	2,90,11,237
Pondicherry	6,04,136	7,89,416
Tamil Nadu	4,82,97,456	5,56,38,318

Categories of data:

Any statistical data can be classified under two categories depending upon the sources utilized. These categories are,

a) Primary data

b) Secondary data

a) Primary data

Primary data is the one, which is collected by the investigator himself for the purpose of a specific inquiry or study. Such data is original in character and is generated by survey conducted by individuals or research institution or any organisation.

The primary data can be collected by the following five methods.

1. Direct personal interviews.

2. Indirect Oral interviews.

3. Mailed questionnaire method.

4. Information from correspondents.

5. Schedules sent through enumerators.

b) Secondary Data

Secondary data are those data which have been already collected and analysed by some earlier agency for its own use; and later the same data are used by a different agency. According to W.A. Neiswanger, a primary source is a publication in which the data are published by the same authority which gathered and analysed them. A secondary source is a publication, reporting the data which have been gathered by other authorities and for which others are responsible'.

Sources of Secondary data

In most of the studies the investigator finds it impracticable to collect first-hand information on all related issues and as such he makes use of the data collected by others. There is a vast amount of published information

from which statistical studies may be made and fresh statistics are constantly in a state of production. The sources of secondary data can broadly be classified under two heads:

a) Published sources

b) Unpublished sources

a) Published Sources:

Generally, published sources are international, national, govt., semi-Govt, private corporate bodies, trade associations, expert committee and commission reports and research reports. They collect the statistical data in different fields like national income, population, prices, employment, wages, export, import etc. These reports are published on regular basis i.e., annually, quarterly, monthly, fortnightly, weekly, daily and so on. These published sources of the secondary data are given below:

i) Govt. Publications

The Central Statistical Organization (CSO) and various state govt. collect compile and publish data on regular basis. Some of the important such publications are:

- Indian Trade Journals
- Reports on Currency and Finance
- Indian Customs and Central Excise Tariff
- Statistical Abstract of India
- Reserve Bank of India Bulletin
- Labour Gazette
- Agricultural Statistics of India
- Bulletin of Agricultural Prices
- Indian Foreign Statistics
- Economic Survey and so on.

ii) International Bodies

All foreign Governments and international agencies publish regular reports of international significance. These reports are regularly published by the agencies like;

- United Nations Organization
- World Health Organization
- International Labour Organization
- Food and Agriculture Organization
- International Bank for Reconstruction and Development
- World Meteorological Organization.

iii) Semi Govt. Publications

Semi govt, organizations municipalities, District Boards and others also publish reports in respect of birth, death and education, sanitation and many other related fields.

iv) Reports of Committee and Commissions

Central Govt, or State Govt, sometimes appoints committees and commissions on matters of great importance. Reports of such committees are of great significance as they provide invaluable data. These reports are like, Shah Commission Report, Sarkaria Commission Report and Finance Commission Reports etc.

v) Private Publications

Some commercial and research institutes publish reports regularly. They are like Institutes of Economic Growth, Stock Exchanges, National Council of Education Research and Training (NCERT), National Council of Applied Economic Research (NCAER) etc.

vi) Newspapers and Magazines

Various newspapers as well as magazines also do collect data in respect of many social and economic aspects. Some of them are as:

- Economic Times
- Financial Express
- Hindustan Times
- Indian Express
- Business Standard
- Economic and Political Weekly
- Main-stream
- Kurukshetra
- Yojna etc.

vii) Research Scholars:

Individual research scholars collect data to complete their research work which further is published with their research papers.

b) Unpublished Source

There are certain records maintained properly by the govt, agencies, private offices and firms. These data are not published.

Limitations of Secondary Data

One should not use the secondary data without care and precautions. As such, secondary data suffers from pitfalls and limitations as stated below:

- No proper procedure is adopted to collect the data.
- Sometimes, secondary data is influenced by the prejudice of the investigator.
- Secondary data sometimes lacks standard of accuracy.
- Secondary data may not cover the full period of investigation.

1.7 CLASSIFICATION

Classification defined as: "the process of arranging things in groups or classes according to their resemblances and affinities and gives expression to the unity of attributes that may subsist amongst a diversity of individuals".

The Collected data, also known as raw data or ungrouped data are always in an un organised form and need to be organised and presented in meaningful and readily comprehensible form in order to facilitate further statistical analysis. It is, therefore, essential for an investigator to condense a mass of data into more and more comprehensible and assailable form. The process of grouping into different classes or sub classes according to some characteristics is known as classification, tabulation is concerned with the systematic arrangement and presentation of classified data. Thus classification is the first step in tabulation. For Example, letters in the post office are classified according to their destinations viz., New Delhi, Mumbai, Bangalore, Chennai etc.,

Objects of Classification:

The following are main objectives of classifying the data:

- It eliminates unnecessary details.
- It facilitates comparison and highlights the significant aspect of data.
- It enables one to get a mental picture of the information and helps in drawing inferences.
- It helps in the statistical treatment of the information collected.

Types of classification:

Statistical data are classified in respect of their characteristics. Broadly there are four basic types of classification namely

- Qualitative classification
- Quantitative classification
- Chronological classification

- Geographical classification

Qualitative classification

Qualitative classification is done according to attributes or non-measurable characteristics; like social status, sex, nationality, occupation, etc.

For example, the population of the whole country can be classified into four categories as married, unmarried, widowed and divorced.

When only one attribute, e.g., sex, is used for classification, it is called simple classification.

When more than one attributes, e.g., deafness, sex and religion, are used for classification, it is called manifold classification.

Quantitative classification

If the data are classified on the basis of phenomenon which is capable of quantitative measurements like age, height, weight, prices, production, income expenditure, sales, profits, etc., it s itemed as quantitative variable.

For example the daily incomes of different retail shops in a town may be classified as under.

Daily earnings in rupees of 100 retail shop in a town

Daily earnings	No. of retail shops
Upto 100	9
101 - 200	25
201 - 300	33
301 - 400	28
401 - 500	2
Above 500	8

In the above classification, the daily earnings of the shops are termed as variable and the number of shops in each class or group as the frequency. This classification is called grouped frequency distribution. Hence this classification is often called 'classification by variables'.

Variable:

A variable in statistics means any measurable characteristic or quantity which can assume a range of numerical values within certain limits, e.g., income, height, age, weight, wage, price, etc. A variable can be classified as either a) Discrete, b) Continuous.

a) Discrete variable

A variable which can take up only exact values and not any fractional values, is called a 'discrete' variable. Number of workmen in a factory, members of a family, students in a class, number of births in a certain year, number of telephone calls in a month, etc., are examples of discrete-variable.

b) Continuous variable

A variable which can take up any numerical value (integral/fractional) within a certain range is called a 'continuous' variable. Height, weight, rainfall, time, temperature, etc., are examples of continuous variables. Age of students in a school is a continuous variable as it can be measured to the nearest fraction of time, i.e., years, months, days, etc.

Chronological classification

When the data are classified on the basis of time then it is known as chronological classification. Such series are also known as time series because one of the variables in them is time. If the population of India during the last eight censuses is classified it will result in a time series or chronological classification.

The following table would give an idea of chronological classification:

Production of Washing Machine by Company 'X'

1966	2600
1967	3400
1968	4800
1969	5100
1970	6900
1971	7300
1972	8600
1973	9800

Geographical classification

When the data are classified by geographical regions or location, like states, provinces, cities, countries etc..., this type of classification is based on geographical or location differences between various items in the data like states, cities, regions, zones etc. For eg. The Rainfall output per Millimetre for different states of India in some given period may be presented as follows:

Rainfall output of different countries in 2015 (Millimetre (mm)

States	Tamil Nadu	Andhra Pradesh	Kerala	Karnataka	Pondicherry
Avg. Output (mm) (Approximate)	70.13	65.33	85.21	75.66	6.23

1.8 TABULATION

Tabulation may be defined as the systematic presentation of numerical data in rows or/and columns according to certain characteristics. It is the process of summarizing classified or grouped data in the form of a table so that it is easily understood and an investigator is quickly able to locate the desired information. A table is a systematic arrangement of classified data in columns and rows. Thus, a statistical table makes it possible for the investigator to present a huge mass of data in a detailed and

orderly form. It facilitates comparison and often reveals certain patterns in data which are otherwise not obvious. Classification and 'Tabulation', as a matter of fact, are not two distinct processes. Actually they go together. Before tabulation data are classified and then displayed under different columns and rows of a table.

Objectives of Tabulation:

The main objectives of tabulation are following

- To carry out investigation;
- To do comparison;
- To locate omissions and errors in the data;
- To use space economically;
- To study the trend;
- To simplify data;
- To use it as future reference.

Advantages of Tabulation:

The advantages of Tabulation are following

- It simplifies complex data and the data presented are easily understood.
- It facilitates comparison of related facts.
- It facilitates computation of various statistical measures like averages, dispersion, correlation etc.
- It presents facts in minimum possible space and unnecessary repetitions and explanations are avoided. Moreover, the needed information can be easily located.
- Tabulated data are good for references and they make it easier to present the information in the form of graphs and diagrams.

Table:

The making of a compact table itself an art. This should contain all the information needed within the smallest possible space. What the purpose of tabulation is and how the tabulated information is to be used are the main points to be kept in mind while preparing for a statistical table. An ideal table should consist of the following main parts are: (i)Table number; (ii) Title of the table; (iii) Captions or column headings; (iv) Stubs or row designation; (v) Body of the table; (vi) Footnotes; and (vii) Sources of data.

Types of Tables:

The tables can be classified according to their purpose, stage of enquiry, nature of data or number of characteristics used. On the basis of the number of characteristics, tables classified as follows: (i) Simple or one-way table; (ii) Two way table; and (iii) Manifold table.

A good statistical table is not merely a careless grouping of columns and rows but should be such that it summarizes the total information in an easily accessible form in minimum possible space. Thus while preparing a table, one must have a clear idea of the information to be presented, the facts to be compared and he points to be stressed.

1.9 FREQUENCY DISTRIBUTION:

A frequency distribution is an arrangement where a number of observations with similar or closely related values are put in separate groups, each group being n order of magnitudes in the arrangement based on magnitudes. It is a series when a number of observations with similar or closely related values are put in separate bunches or groups, each group being in order of magnitude in a series. It is simply a table in which the data are grouped into classes and the numbers of cases which fall in each class are recorded. It shows the frequency of occurrence of different values of a single Phenomenon.

The frequency distribution is constructed for three main reasons are: (i) To facilitate the analysis of data.

(ii) To estimate frequencies of the unknown population distribution from the distribution of sample data.

(iii) To facilitate the computation of various statistical measures.

Example

60	70	55	50	80	65	40	30	80	90
35	45	75	65	70	80	82	55	65	80
90	55	38	65	75	85	60	65	45	75

The above figures are nothing but raw or ungrouped data and they are recorded as they occur without any pre consideration. This representation of data does not furnish any useful information and is rather confusing to mind. A better way to express the figures in an ascending or descending order of magnitude and is commonly known as array. But this does not reduce the bulk of the data. The above data when formed into an array is in the following form:

30	35	38	40	45	45	50	55	55	55
60	60	65	65	65	65	65	65	70	70
75	75	75	80	80	80	80	85	90	90

The array helps us to see at once the maximum and minimum values. It also gives a rough idea of the distribution of the items over the range. When we have a large number of items, the formation of an array is very difficult, tedious and cumbersome. The Condensation should be directed for better understanding and may be done in two ways, depending on the nature of the data.

Discrete frequency distribution

Discrete frequency distribution shows the number of times each value and not to a range of values, of the variable occurs in the data set. Discrete frequency distribution is called ungrouped frequency distribution. In this form of distribution, the frequency refers to discrete value. Here the data are presented in a way that exact measurement of units is clearly indicated. There are definite differences between the variables of different groups of items. Each class is distinct and separate from the other class. Non-continuity from one class to another class exists. Data as such facts like, the number of rooms in a house, the number of companies registered in a country, the number of children in a family, etc.

Example

In a survey of 40 families in a village, the number of children per family was recorded and the following data obtained.

3	1	3	2	1	5	6	6
2	0	0	3	4	2	1	2
1	3	1	5	3	3	2	4
2	2	3	0	2	1	4	5
4	2	3	4	1	2	5	4

Represent the data in the form of a discrete frequency distribution.

Solution:

Frequency distribution of the number of children

Number of Children	Tally Marks	Frequency
0	|||	3
1	𝍸 ||	7
2	𝍸 𝍸	10
3	𝍸 |||	8

4	𝍸 |	6
5	𝍸	4
6	||	2
	Total	40

Grouped frequency distribution

Whenever the range of values of the variable is large for example 0 to 100 or 15 to 200 and if the data is represented by discrete frequency distribution, the data will still remain unwieldy and need further processing for condensation and statistical analysis. Grouped frequency distribution is called continuous frequency distribution. In this form of distribution refers to groups of values. This becomes necessary in the case of some variables which can take any fractional value and in which case an exact measurement is not possible. Hence a discrete variable can be presented in the form of a continuous frequency distribution.

Weekly wages (Rs.)	Number of Employees
1500-2000	4
2000-2500	12
2500-3000	22
3000-3500	33
3500-4000	16
4000-4500	8
4500-5000	5
Total	100

To understand the contraction of the Grouped frequency distribution, the following technical terms need definition and its calculations.

- Class interval

The various groups into which the values of the variable are classified are known as class intervals or simply classes. For example, the symbol 25-35 represents a group or class which includes all the values from 25 to 35.

- Class limits

  - The two values (maximum and minimum) specifying the class intervals are called the class limits. The lowest value is called the lower limit and the height value the upper limit of the class.

  - Length or width of the class is defined as the difference between the upper and the lower limits of the class. That is Class mark or Midpoint of a class = (Lower Limit+Upper Limit)/2

Example

Consider a class denoted as 25 – 50

- The class 25-50 includes all the values in between 25 to 50
- The lower limit of the class is 25 and the upper limit 50
- The length of the class is given by:

  Length = Upper limit – Lower limit = 50 – 25 = 25.

- The class mark or mid value of the class is given by

  Mid value = (Lower Limit + Upper Limit) / 2

  = (25 + 50) / 2 = 37.5

Nominal:

Let's start with the easiest one to understand. Nominal scales are used for labelling variables, without any quantitative value. "Nominal" scales could simply be called "labels." Here are some examples, below. Notice that all of these scales are mutually exclusive (no overlap) and none of them has any numerical significance. A good way to remember all of this is that

"nominal" sounds a lot like "name" and nominal scales are kind of like "names" or labels.

What is your gender?	What is your hair colour?	What is your live?
M – Male	1- Brown	A - North of the equator
F - female	2- Black	B - South of the equator
	3- Blonde	C - Neither, In the international space station
	4- Gray	
	5- Other	

Ordinal:

Ordinal scales, it is the order of the values is what's important and significant, but the differences between each one is not really known. Take a look at the example below. In each case, we know that a = 4 is better than a = 3 or 2, but we don't know–and cannot quantify – how *much* better it is. For example, is the difference between "OK" and "Unhappy" the same as the difference between "Very Happy" and "Happy?" We can't say. Ordinal scales are typically measures of non-numeric concepts like satisfaction, happiness, discomfort, etc. "Ordinal" is easy to remember because is sounds like "order" and that's the key to remember with "ordinal scales"–it is the *order* that matters, but that's all you really get from these.

How do you feel today?	How satisfied are you with our service
1 – Very Unhappy	1- Very unsatisfied
2 – Unhappy	2- Somewhat Unsatisfied
3 – OK	3- Neutral
4 – Happy	4- Somewhat Satisfied
5 – Very Happy	5- Very Satisfied

Interval:

Interval scales are numeric scales in which we know not only the order, but also the exact differences between the values. The classic

example of an interval scale is Celsius temperature because the difference between each value is the same. For example, the difference between 60 and 50 degrees is a measurable 10 degrees, as is the difference between 80 and 70 degrees. Time is another good example of an interval scale in which the increments are known, consistent, and measurable. Interval scales are nice because the realm of statistical analysis on these data sets opens up.

For example, central tendency can be measured by mode, median, or mean; standard deviation can also be calculated. Like the others, you can remember the key points of an "interval scale" pretty easily. "Interval" itself means "space in between," which is the important thing to remember–interval scales not only tell us about order, but also about the value between each item. Here's the problem with interval scales: they don't have a "true zero." For example, there is no such thing as "no temperature." Without a true zero, it is impossible to compute ratios. With interval data, we can add and subtract, but cannot multiply or divide. Confused? Ok, consider this: 10 degrees + 10 degrees = 20 degrees. No problem there. 20 degrees is not twice as hot as 10 degrees, however, because there is no such thing as "no temperature" when it comes to the Celsius scale. I hope that makes sense. Bottom line, interval scales are great, but we cannot calculate ratios, which brings us to our last measurement scale...

Ratio:

Ratio scales are the ultimate nirvana when it comes to measurement scales because they tell us about the order, they tell us the exact value between units, and they also have an absolute zero–which allows for a wide range of both descriptive and inferential statistics to be applied. At the risk of repeating myself, everything above about interval data applies to ratio

scales + ratio scales have a clear definition of zero. Good examples of ratio variables include height and weight. Ratio scales provide a wealth of possibilities when it comes to statistical analysis. These variables can be meaningfully added, subtracted, multiplied, divided (ratios). Central tendency can be measured by mode, median, or mean; measures of dispersion, such as standard deviation and coefficient of variation can also be calculated from ratio scales.

This Device Provides Two Examples of Ratio Scales (height and weight)

1.10 DIAGRAMMATIC AND GRAPHICAL REPRESENTATION

One of the most convincing and appealing ways in which statistical results may be presented is through diagrams and graphs. Representation of statistical data by means of pictures, graphs and geometrical figures is called diagrammatic and graphical representations. The difference between the two is that in the case of diagrammatic representation the quantities are represented by diagrams and pictures and in case of graphical representation they are represented by points which are plotted on a graph paper.

Diagrammatic Representation:

Diagrammatic representation is used when the data relating to different times and places are given and they are independent of one another. One of the most convincing and appealing ways in which statistical results may be presented is through diagrams and graphs. Just one diagram is enough to represent a given data more effectively than thousand words. Moreover even a layman who has nothing to do with numbers can also

understands diagrams. Evidence of this can be found in newspapers, magazines, journals, advertisement, etc. An attempt is made in this chapter to illustrate some of the major types of diagrams and graphs frequently used in presenting statistical data. Diagram is a visual form for presentation of statistical data, highlighting their basic facts and relationship. If we draw diagrams on the basis of the data collected they will easily be understood and appreciated by all. It is readily intelligible and save a considerable amount of time and energy.

Significance of Diagrams and Graphs:

Diagrams and graphs are extremely useful because of the following reasons.

- They are attractive and impressive.
- They make data simple and intelligible.
- They make comparison possible
- They save time and labour.
- They have universal utility.
- They give more information.
- They have a great memorizing effect.

General rules for constructing diagrams:

The construction of diagrams is an art, which can be acquired through practice. However, observance of some general guidelines can help in making them more attractive and effective. The diagrammatic presentation of statistical facts will be advantageous provided the following rules are observed in drawing diagrams.

- A diagram should be neatly drawn and attractive.
- The measurements of geometrical figures used in diagram should be accurate and proportional.
- The size of the diagrams should match the size of the paper.

- Every diagram must have a suitable but short heading.
- The scale should be mentioned in the diagram.
- Diagrams should be neatly as well as accurately drawn with the help of drawing instruments.
- Index must be given for identification so that the reader can easily make out the meaning of the diagram.
- Footnote must be given at the bottom of the diagram.
- Economy in cost and energy should be exercised in drawing diagram.

Types of diagrams:

In practice, a very large variety of diagrams are in use and new ones are constantly being added. For the sake of convenience and simplicity, they may be divided under the following heads:

- One-dimensional diagrams
- Two-dimensional diagrams
- Three-dimensional diagrams
- Pictograms and Cartograms

One-dimensional diagrams

In such diagrams, only one-dimensional measurement, i.e height is used and the width is not considered. These diagrams are in the form of bar or line charts and can be classified as

- Line Diagram
- Simple Diagram
- Multiple Bar Diagram
- Sub-divided Bar Diagram
- Percentage Bar Diagram

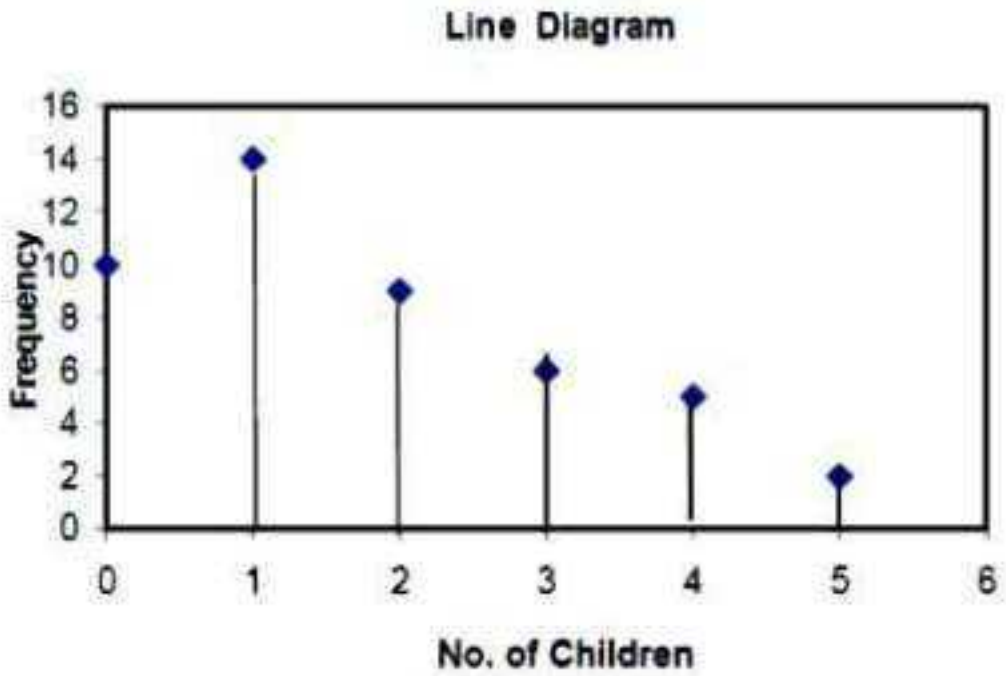
Line Diagram:

Line diagram is used in case where there are many items to be shown and there is not much of difference in their values. Such diagram is prepared by drawing a vertical line for each item according to the scale. The distance between lines is kept uniform. Line diagram makes comparison easy, but it is less attractive.

Example

Show the following data by a line chart.

No. of Children	0	1	2	3	4	5
Frequency	10	14	9	6	4	2



Two-dimensional Diagrams

In one-dimensional diagrams, only length 9 is taken into account. But in two-dimensional diagrams the area represents the data and so the length and breadth have both to be taken into account. Such diagrams are also called area diagrams or surface diagrams. The important types of area diagrams are:

- Rectangles

- Squares
- Pie-diagrams

Three-dimensional diagrams

Three-dimensional diagrams, also known as volume diagram, consist of cubes, cylinders, spheres, etc. In such diagrams three things, namely length, width and height have to be taken into account. Of all the figures, making of cubes is easy. Side of a cube is drawn in proportion to the cube root of the magnitude of data.

Pictograms and Cartograms

Pictograms are not abstract presentation such as lines or bars but really depict the kind of data we are dealing with. Pictures are attractive and easy to comprehend and as such this method is particularly useful in presenting statistics to the layman. When Pictograms are used, data are represented through a pictorial symbol that is carefully selected. Cartograms or statistical maps are used to give quantitative information as a geographical basis. They are used to represent spatial distributions. The quantities on the map can be shown in many ways such as through shades or colours or dots or placing pictogram in each geographical unit.

GRAPHICAL REPRESENTATION:

The graphical representation is used when we have to represent the data of a frequency distribution and a time series. A graph represents mathematical relationship between the two variables whereas a diagram does not. A graph is a visual form of presentation of statistical data. A graph is more attractive than a table of figure. Even a common man can understand the message of data from the graph. Comparisons can be made between two or more phenomena very easily with the help of a graph. Finally graphs are more obvious, precise and accurate than diagrams and are quite helpful to the statistician for the study of slopes, rates of changes and estimation, whenever possible. However here we shall discuss only some important types of graphs which are more popular and they are

- Histogram
- Frequency Polygon
- Frequency Curve
- Ogive
- Lorenz Curve

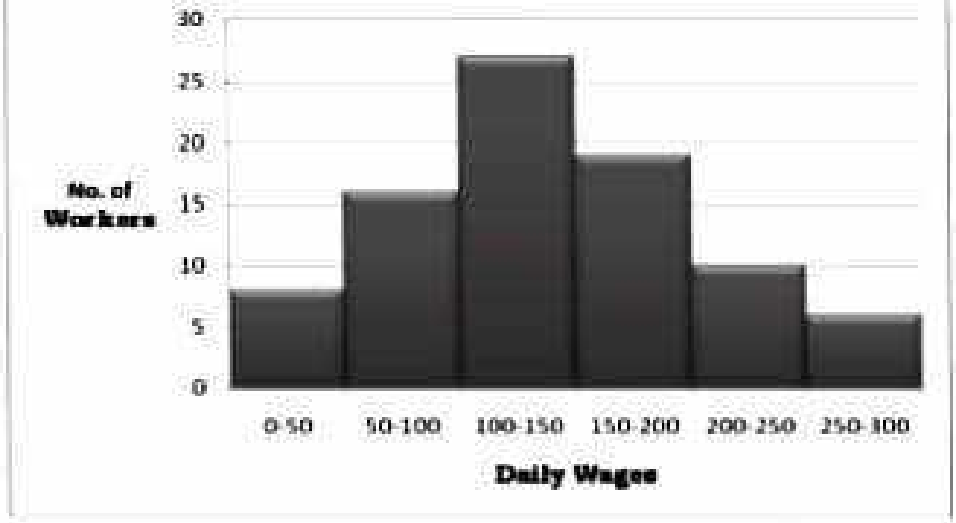
Histogram

A histogram consists of bars or rectangles which are erected over the class intervals, without giving gaps between bars and such that the areas of the bars are proportional to the frequencies of the class intervals. It is a bar chart or graph showing the frequency of occurrence of each value of the variable being analysed. In histogram, data are plotted as a series of rectangles. Class intervals are shown on the 'X-axis' and the frequencies on the 'Y-axis'. The height of each rectangle represents the frequency of the class interval. Each rectangle is formed with the other so as to give a continuous picture. Such a graph is also called staircase or block diagram. However, we cannot construct a histogram for distribution with open-end classes. It is also quite misleading if the distribution has unequal intervals and suitable adjustments in frequencies are not made.

Example

Draw a histogram for the following data.

Daily wages	Number of Workers
0-50	8
50-100	16
100-150	27
150-200	19
200-250	10
250-300	6



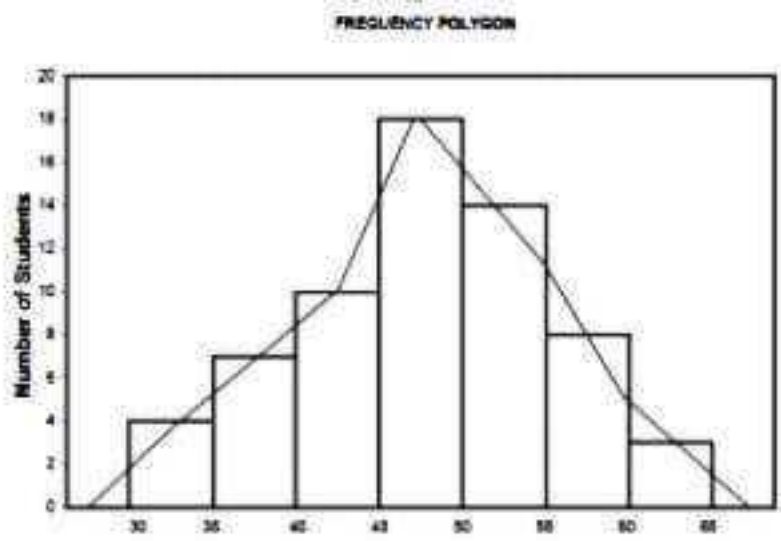
Frequency Polygon

Frequency Polygon is another device of graphic presentation of a frequency distribution. In case of discrete frequency distribution frequency polygon is obtained on plotting the frequencies on the vertical axis (y-axis) against the corresponding values of the variable on the horizontal axis (x-axis) and joining the points so obtained by straight lines. In case of grouped or continuous frequency distribution the construction of frequency polygon is consist in plotting the frequencies of different classes (along y-axis). The points so obtained are joined by straight lines to obtain the frequency polygon.

Example

Draw a Frequency polygon for the following data.

Weight (in Kg.)	30-35	35-40	40-45	45-50	50-55	55-60	60-65
No. of Students	4	7	10	18	14	8	3



Frequency Curve

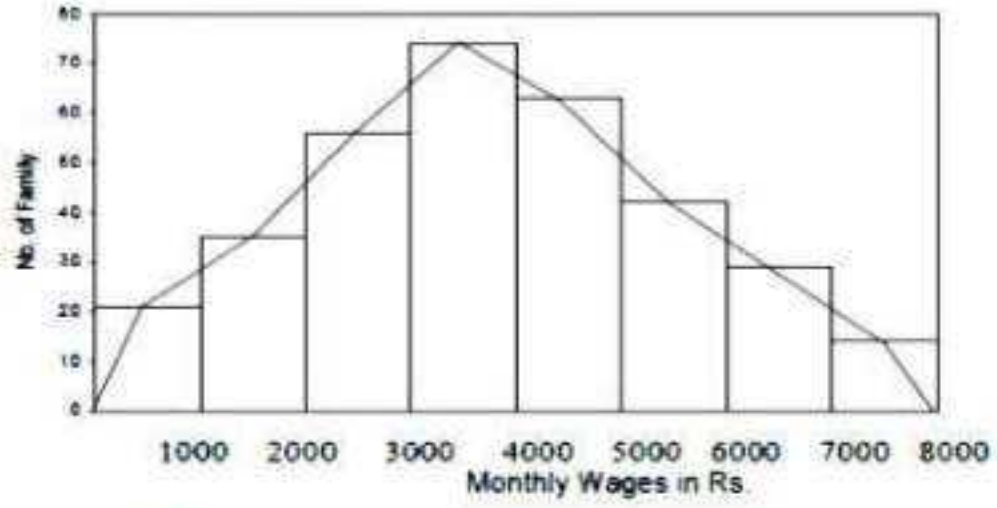
A frequency curve is a smooth free hand curve drawn through the vertices of frequency polygon. The object of smoothing of the frequency polygon is to eliminate as far as possible, the random or erratic changes that might be present n the data. The area enclosed shape is smooth one and not with sharp edges.

Example

Draw a frequency curve for the following data.

Monthly Wages (in Rs)	No. of family
0-1000	21
1000-2000	35
2000-3000	56
3000-4000	74
4000-5000	63
5000-6000	40
6000-7000	29
7000-8000	14

FREQUENCY CURVE



Ogives or cumulative frequency curves:

For a set of observations, we know how to construct a frequency distribution. In some cases we may require the number of observations less than a given value or more than a given value. This is obtained by a accumulating (adding) the frequencies upto (or above) the give value. This accumulated frequency is called cumulative frequency. These cumulative frequencies are then listed in a table is called cumulative frequency table. The curve table is obtained by plotting cumulative frequencies is called a cumulative frequency curve or an ogive. There are two methods of constructing ogive namely: a) The 'less than ogive' method, b) The 'more than ogive' method

In less than ogive method we start with the upper limits of the classes and go adding the frequencies. When these frequencies are plotted, we get a rising curve. In more than ogive method, we start with the lower limits of the classes and from the total frequencies we subtract the frequency of each class. When these frequencies are plotted we get a declining curve.

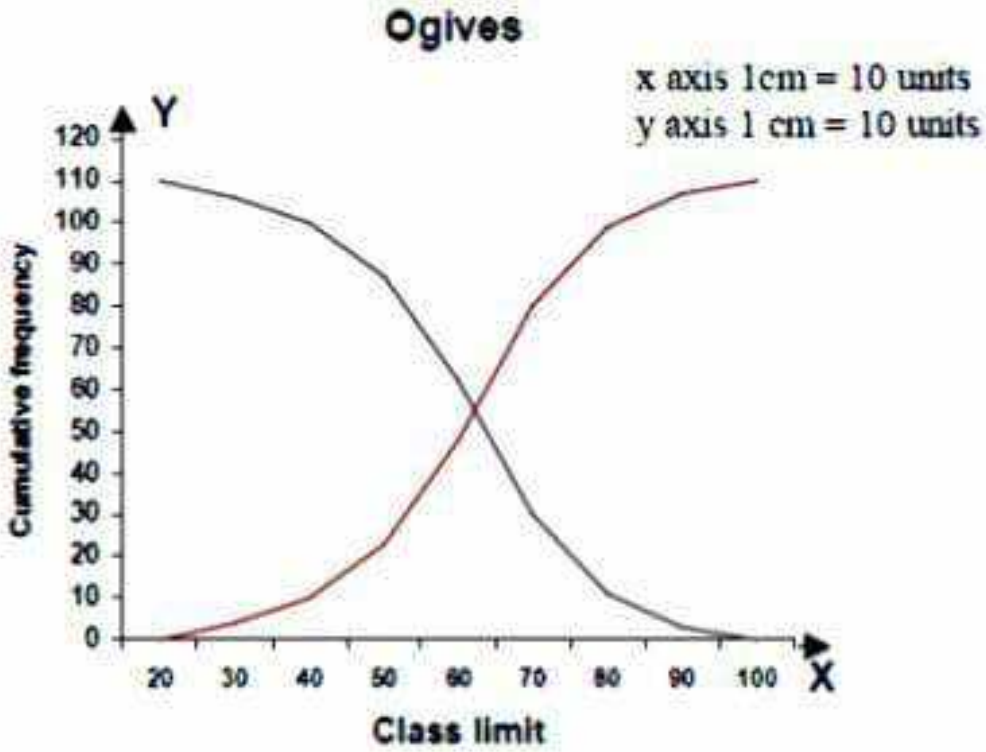
Example

Draw the O gives for the following data.

C.I	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
F	4	6	13	25	32	19	8	3

Solution

Class limit	Less than ogive	More than ogive
20	0	110
30	4	106
40	10	100
50	23	87
60	48	62
70	80	30
80	99	11
90	107	3
100	110	0



Unit II

Classification of data:

Classification is a process of grouping or sorting out the collected data according to certain common properties.

Data expressed in ascending or descending order is called an Array.

The differences between the largest and smallest item in the raw data. is called the range of the data.

Example:

Consider the following raw data which gives the marks of 100 Students

41	37	13	15	37	39	23	20
30	40	33	36	29	11	8	19
35	46	44	34	7	16	35	10
28	29	33	49	41	40	30	21
11	16	31	34	39	41	35	33
28	21	36	33	1	32	40	31
47	18	9	31	37	43	26	12
10	37	8	11	31	36	35	41
31	35	33	34	39	49	48	12
21	20	19	41	37	33	34	8

19	48
31	36
21	22
13	6
2	18
16	4
34	19
7	10
34	33
11	6

Raw data may be summarized by using tally marks. This results in a table which gives the number of time particular value occurs in the data. for example:

For the data given above using the following table:

Marks	No. of students Tally marks	Total
1	1	1
2	1	1
3	–	0
4	1	1
5	–	0
6	11	2
7	11	2
8	111	3
9	1	1
10	11	2
11	1111	4
12	11	2
13	11	2
14	–	0
15	–	0
16	111	3
17	1	1
18	11	2
19	1111	4
20	11	2

Marks	No. of student Tally marks	Total
21	III	3
22	I	1
23	I	1
24	II	2
25	–	0
26	II	2
27	–	0
28	II	2
29	III	3
30	II	2
31	~~HH~~ II	7
32	I	1
33	~~HH~~ II	7
34	IIII	4
35	~~HH~~	5
36	IIII	4
37	~~HH~~	5
38	–	0
39	I	1
40	IIII	4

Marks	No. of student Tally mark	Total
41	~~HH~~	5
42	–	0
43	I	1
44	I	1
45	–	0
46	I	1
47	I	1
48	II	2
49	II	2
Total	mark	100

Problems:

From the following data of the weekly wages of workers, employed in a certain factory, construct a frequency table with classes. 10 - 19.99, 20 - 29.99, 30 - 39.99 etc.

Wages in Rupees

10	100	90	30	99	25	70	32	76	19
68	76	31	39	89	40	66	27	109	42
93	53	97	43	29	92	28	95	36	105
67	55	47	108	37	86	46	102	44	68
47	81	77	48	50	87	41	88	59	80
52	85	56	61	58	72	69	118	82	78
69	54	71	60	63	73	65	79	64	61

Frequency distribution of the wages of 70 workers.

Wages (Rs)	Tally Bars	No. of Workers.
10 - 19.99	II	2
20 - 29.99	IIII	4
30 - 39.99	~~HH~~ I	6
40 - 49.99	~~HH~~ IIII	9
50 - 59.99	~~HH~~ III	8
60 - 69.99	~~HH~~ ~~HH~~ II	12
70 - 79.99	~~HH~~ IIII	9
80 - 89.99	~~HH~~ III	8
90 - 99.99	~~HH~~ I	6
100 - 109.99	IIII	4
110 - 119.99	II	2

8. Prepare a frequency table by taking the variable as the no. of. letter in each word from the passage given below selecting an appropriate class intervals.

"By Statistics we mean aggregates of facts affected to a Marksheet extent by multiplicity of causes numerically expressed enumerated or estimated according to reasonable standards of accuracy. collected in a Systematic manner for a Predetermined purpose and placed in relation to each other".

H.W

Class	No. of letters Tally Mark	Frequency
1 - 3	𝍸 𝍸 𝍸 II	17
4 - 6	𝍸 IIII	9
7 - 9	𝍸 IIII	9
10 - 12	𝍸 II	7
13 - 15	I	1
	Total	43

H.W:

The following are the weights in kilogram of group of 55 students.

42	74	40	60	82	105	41	61	75	83	63
53	100	76	84	50	67	65	71	77	56	95
68	69	104	80	79	79	54	73	59	81	100
66	49	77	90	84	76	42	64	69	70	86
72	50	79	52	103	96	51	86	78	94	71

Prepare a frequency table taking the magnitude of each class interval as 10 kilograms and first class interval as equal to 40 and less than 50.

Ans:

Weight (Kgs)	Tally bars	Frequency
40 and less than 50	𝍸	5
50 and less than 60	𝍸 III	8
60 and less than 70	𝍸 𝍸	10

70 and less than 80	||| ||| |||	15
80 and less than 90	||| |||	8
90 and less than 100	||||	4
100 and less than 110	||||	21
110 and less than 120		1
		55

2. Prepare a frequency table for the following figures related to weekly wages in (Rupees) of workers in a factory taking a class interval of 5.

70	68	74	76	110	106	107	86	74	75
71	76	106	72	102	94	83	74	88	92
73	74	105	73	101	92	86	76	87	93
86	76	103	83	100	96	71	79	85	96
89	80	102	86	90	102	76	80	86	95
96	81	101	74	95	109	75	82	78	94
95	89	106	96	97	108	76	83	77	86
92	70	103	94	96	106	72	84	75	88
100	73	96	93	92	107	73	86	79	83
110	102	96	92	93	100	76	88	82	84.

Wages (Rs)	Tally bars	Frequency
65 - 70	II	2
70 - 75	𝍸 𝍸 III	13
75 - 80	𝍸 𝍸 𝍸 I	16
80 - 85	𝍸 𝍸 II	12
85 - 90	𝍸 𝍸 II	12
90 - 95	𝍸 𝍸 II	12
95 - 100	𝍸 𝍸 I	11
100 - 105	𝍸 𝍸 I	11
105 - 110	𝍸 IIII	9
110 - 115	II	2
	total	100

Bar Diagram:-

Bar diagram is a popular form of diagrammatic representation. This diagram consists of a series of rectangular bars standing on a common base. The bars are all of equal width and equally spaced. The length of bars are proportional to their magnitude. The comparsion among the bars is based only on lengths. This type of diagram is called "One dimensional diagram".

Bar diagrams are two

types.

1. Vertical Bardiagram.
2. Horizontal Bardiagram.

The Bar diagrams can be classified as:

1. Simple bardiagram.
2. Multiple bardiagram.
3. Component (or) Subdivided bardiagram.
4. Percentage bardiagram.

(i) Simple Bardiagram:-

A simple bardiagram represents the magintude of a single variable like Sales, Production, Profit etc..

Example:1

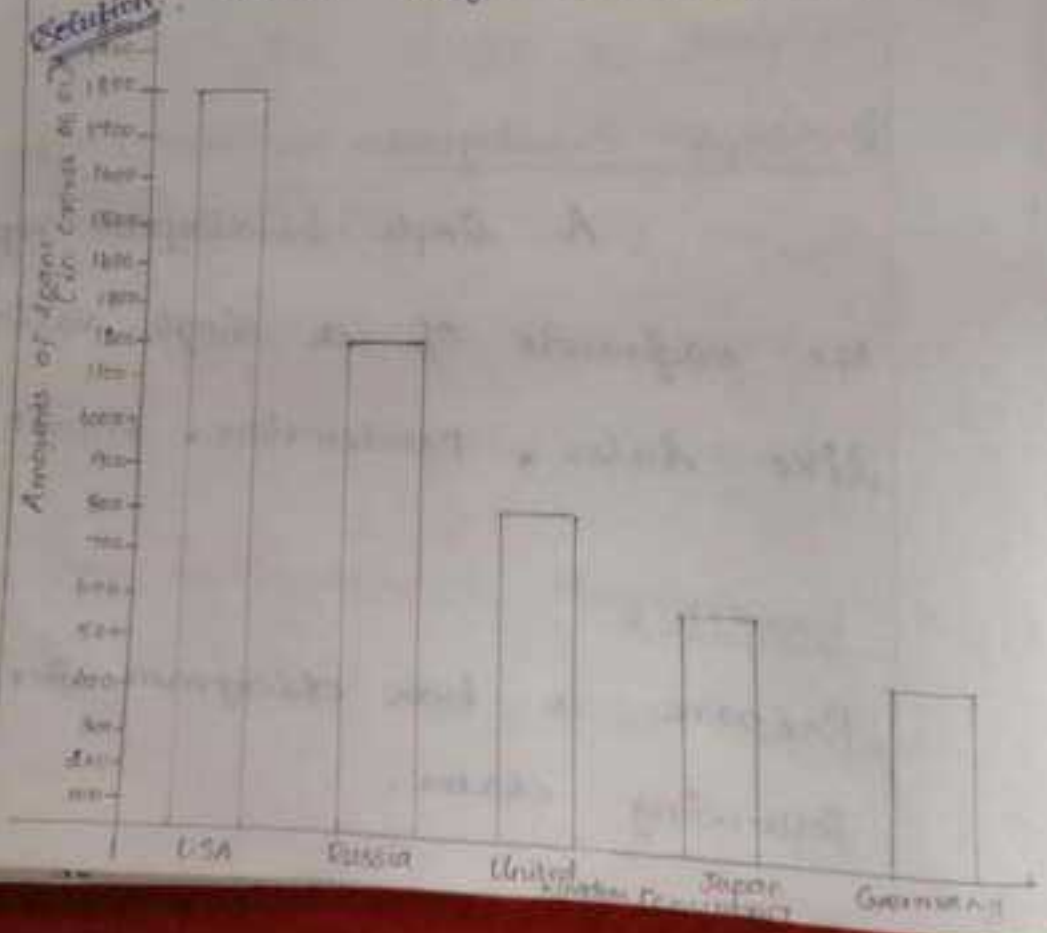
Prepare a bar diagram for the following data.

India's foreign debt as on 1-4-2000

Source of borrowing	Amounts of loan (in crores of Rs)
U.S.A.	1800
Russia.	1200
United Kingdom.	800
Japan.	600
Germany.	500.

India's foreign debt as on 1-4-2000.

Solution:



Multiple Bardiagram:-

Multiple Bardiagram is used for comparing two or more set of Statistical data. Bars are constructed side by side to represent the sets of values for comparison.

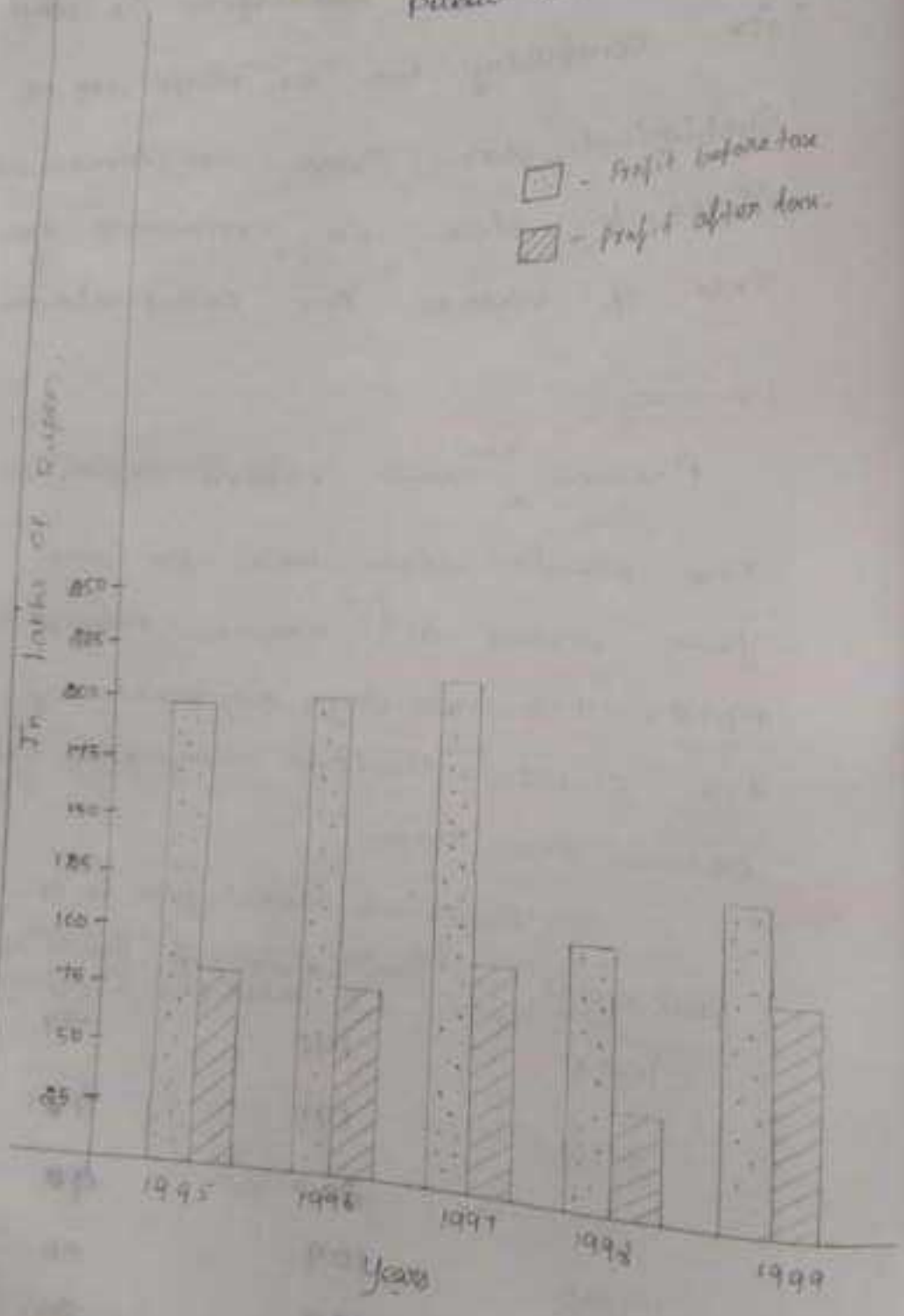
Example:

Present the profit before tax and the profit after tax for the year ended 31st March 1995, 1996, 1997, 1998 and 1999 respectively of the public limited company mentioned below the data.

Financial Highlights of the public limited company

Year ended in 31st March	Profit before tax (In lakhs of Rs)	Profit After tax (In Lakhs of Rs)
1995	190	79
1996	191	941
1997	200	90
1998	109	36
1999	127	89.

Financial Highlights of the public limited company.



H.W 1- Represent a following data by a simple bar diagram.

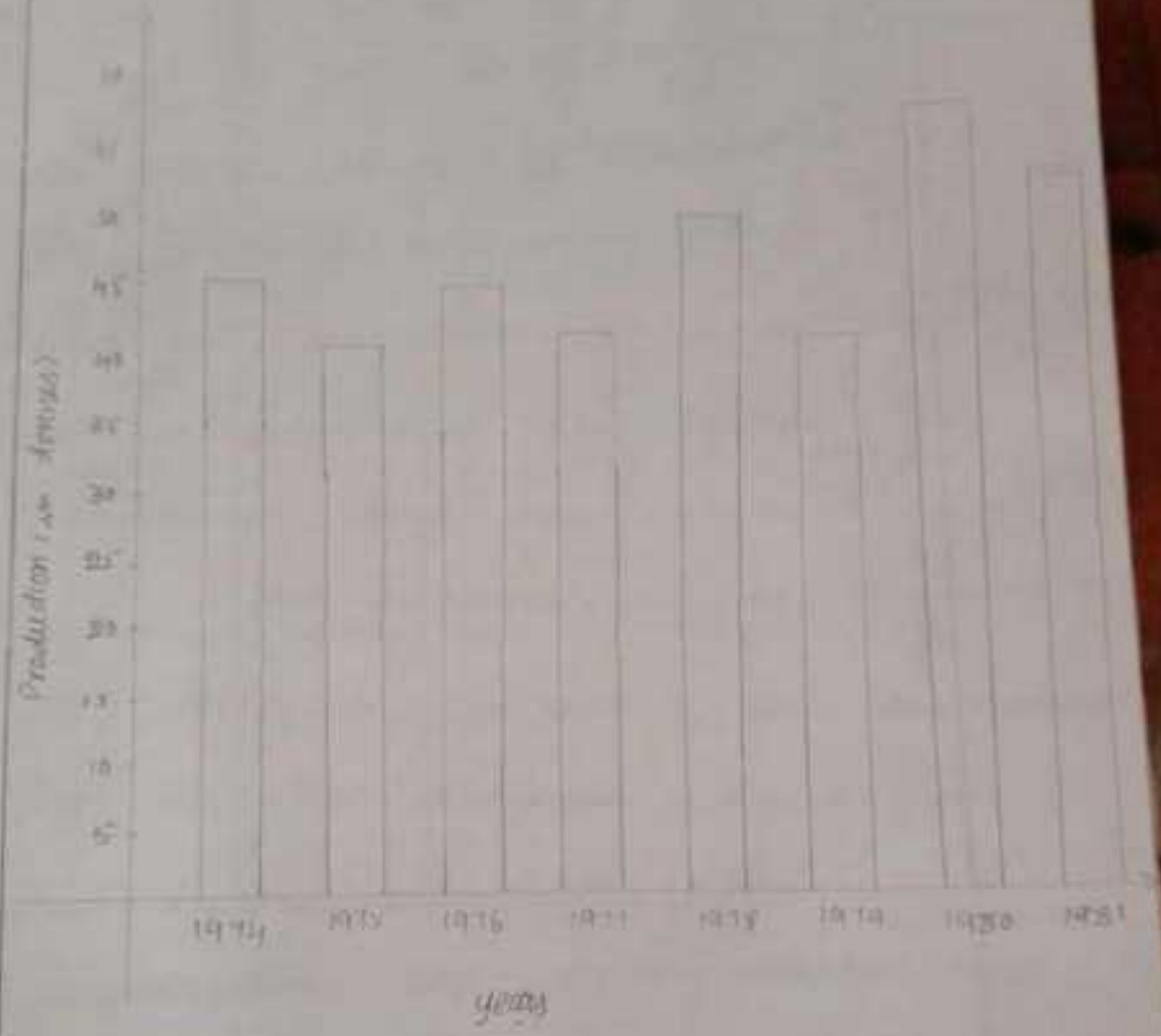
Year	production (in tonnes)
1974	45
1975	40
1976	44
1977	41
1978	49
1979	42
1980	55
1981	50.

2. Represent the following data by a suitable diagram showing the difference between proceeds and costs.

Proceeds and costs of the farm.
[In 1000 of rupees].

Year	Total proceeds	Total costs.
1950	22.0	19.5
1951	27.8	21.7
1952	28.3	30.0
1953	20.3	25.6
1954	32.7	26.1
1955	33.8	31.1

1. Solution:



2. Solution:

Proceeds and costs of a firm

(in thousands of Rs).

(In thousands Rs).

	1950	1951	1952	1953	1954	1955
Sales						
Total Profit						

(iii) Component or Subdivided Bardiagram:-

Comp.

Here each bar representing the total value is subdivided into its different component parts. This enables comparison between three different components and also between a component and the whole.

For example; where population of different countries are represented by a bar diagram, each bar can be subdivided into sub populations representing males and females.

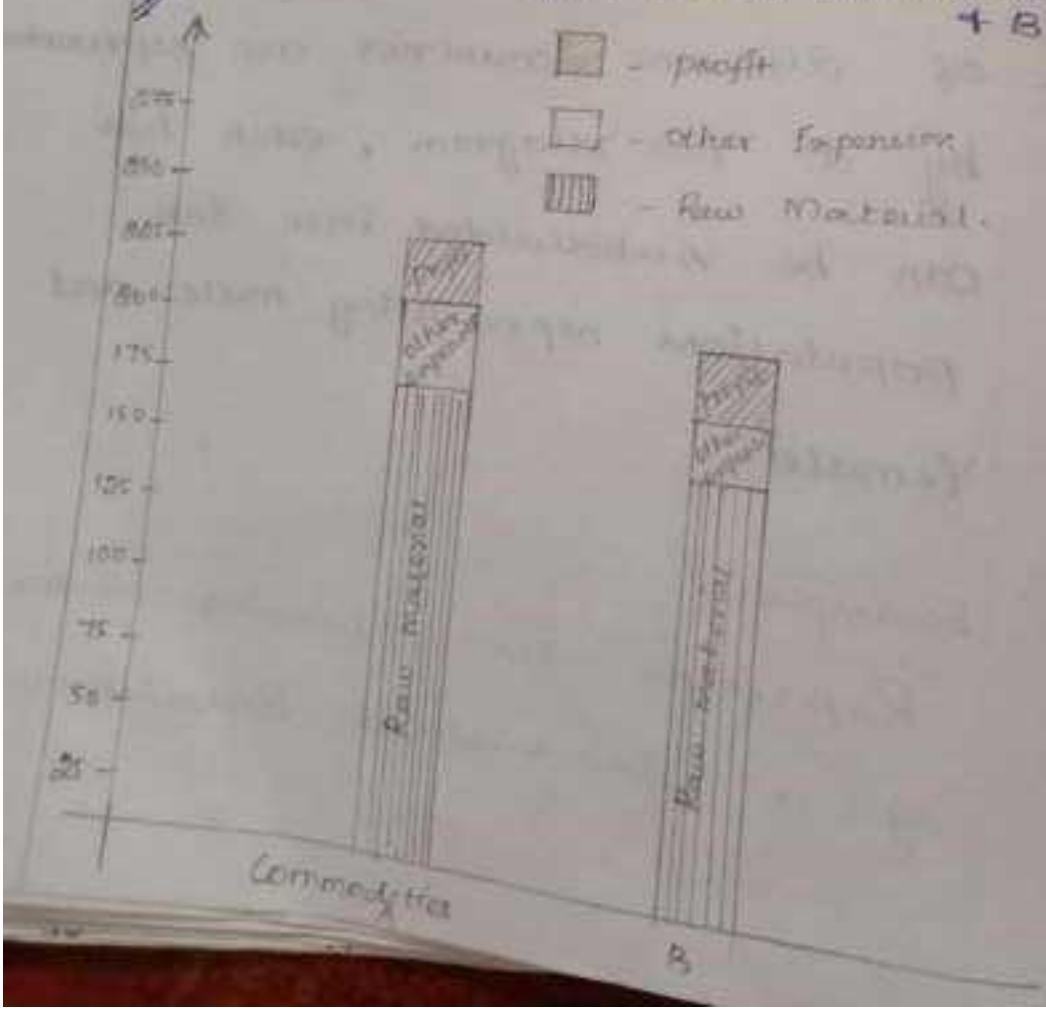
Example :-

Represent the following data by a subdivided Bardiagram.

Distribution of revenue in commodities A and B.

Particulars per unit	Commodities A (Rs)	B (Rs)
Profit per unit	3	2
Quantity Sold	75	100
Value of raw material	175	150
Other production expensions	30	25
Profit	80	25

Distribution of revenue in commodities A & B.



(iv) Percentage Bardiagram:

This is another form of Component Bardiagram. Here the components are not the actual values but percentage of the whole. The main difference between the component bardiagram and percentage bardiagram is that the bars in the component bar diagram are of different heights. Since the totals may be different. whereas in the percentage bardiagram the bars are of equal height Since each bar represents 100 Per cent.

Example:-

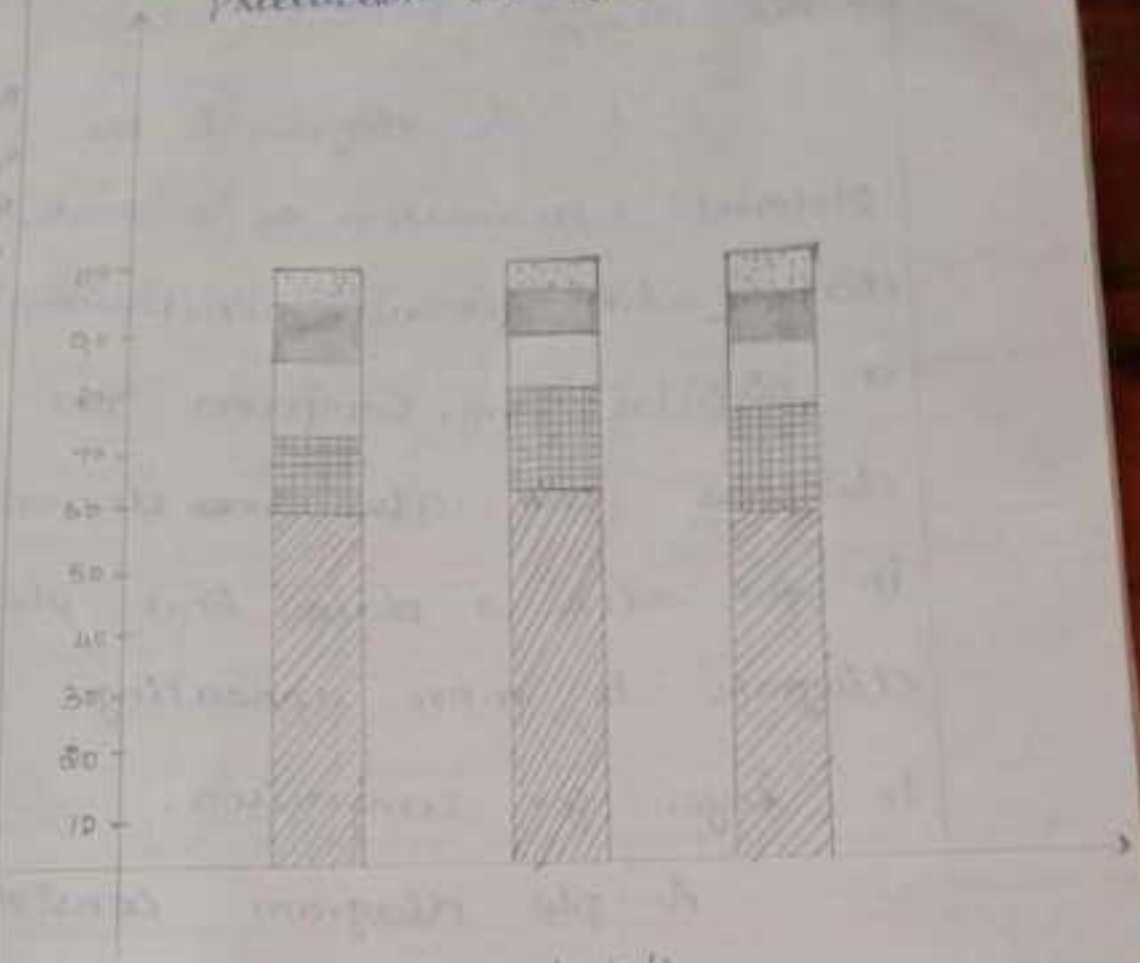
Represent the following data by means of percentage Subdivided bardiagrams.

Particulars	1979	1980	1981
Raw Materials	2,160	2600	2,700
Labour	540	700	810
Direct Expenses	360	200	360
Factory Expenses	360	300	360
Office Expenses	180	200	270
Total Cost	3600	4000	4500

Ans:

Particulars	Expressed in percentage yrs.		
	1979	1980	1981
Raw Materials	$\frac{2160}{36} = 60$	$\frac{2600}{40} = 65$	$\frac{2700}{45} = 60$
Labour	$\frac{540}{36} = 15$	$\frac{700}{40} = 17.5$	$\frac{810}{45} = 18$
Direct Expenses	$\frac{360}{36} = 10$	$\frac{200}{40} = 5$	$\frac{360}{45} = 8$
Factory Expenses	$\frac{360}{36} = 10$	$\frac{300}{40} = 7.5$	$\frac{360}{45} = 8$
Office Expenses	$\frac{180}{36} = 5$	$\frac{200}{40} = 5$	$\frac{270}{45} = 6$
Total cost	100	100	100

Production cost of Scooter in 3 years



- Raw materials
- Labour
- Direct expenses
- Factory Expenses
- office expenses.

04 Pie Diagram:

A pie diagram is the pictorial representation of a statistical data with several subdivision in a circular form. Component bar diagrams can also be drawn in for such a data. But pie diagram is more appealing to eyes for comparison.

A pie diagram consists of a circle subdivided into several sectors by radius. The area of the sectors is proportional to the values of the components.

Inorder to draw a pie diagram the different values of the components are expressed in

degrees taking the whole value as 360°.

Example:-

Draw a pie diagram of the following data relating to areas under different food crops

Food Crops	Rice	Wheat	Barley	Jowar	Bajra	Maize	other
Area (in acres)	8	8	4	2	2	5	11

Sol:-

We Expressed the given values in degrees in the following table

Food crops	Area	In degrees
Rice	8	$\frac{8}{40} \times 360 = 72$
Wheat	8	$\frac{8}{40} \times 360 = 72$
Barley	4	$\frac{4}{40} \times 360 = 36$
Jowar	2	$\frac{2}{40} \times 360 = 18$
Bajra	2	$\frac{2}{40} \times 360 = 18$
Maize	5	$\frac{5}{40} \times 360 = 45$
Others.	11	$\frac{11}{40} \times 360 = 99$
Total	40	360

2 Draw a circular diagram for the following data.

Type of commodity	Expenditure in rupees	
	Family A	Family B
Food	300	500
Rent	200	350
Clothes	125	250
Education	110	225
Miscellaneous	75	125
Savings	90	150

H.W 1. Draw a pie diagram.

Compartment No	1	2	3	4
Space limit (000 c ft)	180	160	140	180

2.

Blood Group	Frequency			
	Gypsies	Indians	Hungarians	Total
O	343	313	344	1000

3.

Males	Females	Girls	Boys	Total
2000	1,800	4,200	2000	10,000

Unit: 3

Measures of Central tendency

Measures of Central tendency are "Statistical constants which enables us to comprehent in a single effort the significance of the whole".

The following are the five measures of central tendency which are the common use.

(i) Arithmetic Mean (Mean)
(ii) Median
(iii) Mode.
(iv) Geometric Mean.
(v) Harmonic Mean.

i) Arithmetic Mean: (ungrouped data).

Arithmetic mean of n observations $x_1, x_2, \ldots, x_n$ is defined by.

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\Sigma x_i}{n}$$

Note: (grouped data)

Suppose $x_1, x_2, \ldots, x_n$ be the distinct values of a variable with corresponding frequencies $f_1, f_2 \ldots f_n$ then.

$$\bar{x} = \frac{\Sigma f_i x_i}{\Sigma f_i (= N)}, \quad i = 1, 2 \ldots n$$

Example:

Consider the 10 numbers 18, 15, 18, 16, 17, 18, 15, 19, 17, 17 find the mean.

Soln:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

$$\bar{x} = \frac{18+15+18+16+17+18+15+19+17+17}{10}$$

$$\bar{x} = \frac{170}{10}$$

$$\bar{x} = 17$$

Example:

The frequency distribution of the above data is.

x_i	15	16	17	18	19
f_i	2	1	3	3	1

Calculate the arithmetic mean.

Soln:

$$\bar{x} = \frac{\Sigma f_i x_i}{\Sigma f_i}$$

$$\bar{x} = \frac{(8 \times 15) + (1 \times 16) + (3 \times 17) + (3 \times 18) + (1 \times 19)}{8 + 1 + 3 + 3 + 1}$$

$$\bar{x} = \frac{30 + 16 + 51 + 54 + 19}{10}$$

$$\bar{x} = \frac{170}{10}$$

$$\bar{x} = 17 //$$

H.W

1. Calculate the A.M from the following frequency table.

Weight in kgs	50	48	46	44	42	40
No. of. persons	18	14	16	13	11	9

Sol:

$$\bar{x} = \frac{\Sigma f_i x_i}{\Sigma f_i}$$

$$\bar{x} = \frac{(12\times50)+(14\times48)+(16\times46)+(13\times44)+(11\times42)+(9\times40)}{12+14+16+13+11+9}$$

$$\bar{x} = \frac{600+672+736+572+462+360}{75}$$

$$\bar{x} = \frac{3402}{75}$$

$$\bar{x} = 45.36$$

Method : 3 (Step division Method) (Coded deviation)

If we take A as the new origin and take h units of the variate $x_i = 1$ unit of the new variate u_i then.

$$u_i = \frac{x_i - A}{h}$$

ie, $hu_i = x_i - A$

$$x_i = hu_i + A.$$

then the arithmetic mean for the variate x_i is calculated as follows.

$$\bar{x} = \frac{\Sigma f_i x_i}{\Sigma f_i (= N)}$$

$$\bar{x} = \frac{\Sigma f_i (h u_i + A)}{N}$$

$$\bar{x} = \frac{\Sigma f_i h u_i + \Sigma f_i A}{N}$$

$$= \frac{h \Sigma f_i u_i + A \Sigma f_i}{N}$$

$$= h \frac{\Sigma f_i u_i}{N} + A \frac{\Sigma f_i}{N}$$

$$= h \bar{u} + A \frac{N}{N}$$

$$\bar{x} = h \bar{u} + A$$

1. Calculate the average mark from the following data.

Marks	No. of Students
0 - 10	5
10 - 20	12
20 - 30	15
30 - 40	25
40 - 50	8
50 - 60	3
60 - 70	2

Sol:

Here A = 35

h = 10

Marks	f_i (no. of students)	mid x_i	$u_i = \frac{x_i - A}{h}$	$f_i u_i$
0-10	5	5	$\frac{5-35}{10} = -3$	-15
10-20	12	15	-2	-24
20-30	15	25	-1	-15
30-40	25	35	0	0
40-50	8	45	1	8
50-60	3	55	2	6
60-70	2	65	3	6
Total	70			-34

$$\bar{x} = h\bar{u} + A$$

$$= \frac{h \Sigma f_i u_i}{\Sigma f_i} + A$$

$$= 35 + 10 \times \frac{(-34)}{70}$$

$$= 35 + (-4.86)$$

$$\bar{x} = 30.14$$

2. The marks scored by 60 students in an examination in Statistics are given below from a frequency distribution with class intervals of 10 and calculate the arithmetic mean.

Marks	f_i
0-10	4
10-20	4
20-30	9
30-40	20
40-50	12
50-60	6
60-70	3
70-80	2

60

$A = 45/35$

$h = 10$

Marks	f_i	mid x_i	$u_i = \frac{x_i - A}{h}$	$f_i u_i$
0-10	4	5	-3	-12
10-20	4	15	-2	-8
20-30	9	25	-1	-9
30-40	20	35	0	0
40-50	12	45	1	12
50-60	6	55	2	12
60-70	3	65	3	9
70-80	2	75	4	8
Total	60			12

$$\bar{x} = h\bar{u} + A$$

$$= \frac{h \Sigma f_i u_i}{\Sigma f_i} + A \quad = 10 \times \left(\frac{12}{60}\right) + 35$$

$$= 35 + 10 \left(\frac{12}{60}\right)$$

$$= 35 + \frac{12}{6}$$

$$\bar{x} = 37$$

14-10 1. Calculate arithmetic mean for following data.

Marks	Students
0-10	33
10-20	53
20-30	108
30-40	221
40-50	153
50-60	322
60-70	439
70-80	586
80-90	495
90-100	50

2.

Temp, °C	No. of. days
-40 to -30	10
-30 to -20	28
-20 to -10	30
-10 to 0	42
0 to 10	65
10 to 20	180
20 to 30	10

1. Soal:

Masa	fi	midxi	$u_i = \frac{x_i - A}{h}$	$f_i u_i$
0 - 10	33	5	-4	-132
10 - 20	53	15	-3	-159
20-30	108	25	-2	-216
30-40	221	35	-1	-221
40-50	153	45	0	0
50-60	322	55	1	322
60-70	429	65	2	858
70-80	526	75	3	1578
80-90	495	85	4	1980
90-100	50	95	5	250
Total	2400			4280

$A = 45, \quad h = 10$

$\bar{x} = h\bar{u} + A$

$= \frac{h \Sigma f_i u_i}{\Sigma f_i} + A$

$= \frac{10\,(4280)}{2400} + 45$

$= 62.83$ //

2.

Temp°c	f_i	mid x_i	$u_i = \frac{x_i - A}{h}$	$f_i u_i$
-40 to -30	10	-35	-3	-30
-30 to -20	23	-25	-2	-46
-20 to -10	30	-15	-1	-30
-10 to 0	42	-5	0	0
0 to 10	65	5	1	65
10 to 20	150	15	2	360
20 to 30	10	25	3	30
Total	365			339

$A = -5$

$h = 10$

$\bar{x} = h\bar{u} + A$

$= h \frac{\Sigma f_i u_i}{\Sigma f_i} + A$

$= \frac{10(339)}{365} + (-5)$

$= 9.2876 + (-5)$

$\bar{x} = 4.288$

Combined mean:

$$\bar{X}_{12} = \frac{N_1\bar{x}_1 + N_2\bar{x}_2}{N_1 + N_2} \rightarrow \text{formula}$$

where N_1 is the number of items in the first groups N_2 is the number of items in the second group $\bar{x}_1$ is the arithmetic mean of the first group $\bar{x}_2$ is the arithmetic mean of the second group

Example:

The mean wags of 100 labours working in the factory running two shifts of 70 and 30, workers respectively is Rs. 84. The mean wage of 70 labour working in morning shift is 90. find the mean wags of 30 workers working in evening shift.

Solution:

Here, $\bar{X}_{12} = 84$

$N_1 = 70$

$N_2 = 30$

$\bar{X}_1 = 90$

$\bar{X}_2 = ?$

$$\bar{X}_{12} = \frac{N_1\bar{X}_1 + N_2\bar{X}_2}{N_1 + N_2} \quad -①$$

$$① \Rightarrow 84 = \frac{70 \times 90 + 30\bar{X}_2}{70 + 30}$$

$$84 = \frac{6300 + 30\bar{X}_2}{100}$$

$$8400 = 6300 + 30\bar{X}_2$$

$$8400 - 6300 = 30\bar{X}_2$$

$$30\bar{X}_2 = 8400 - 6300$$

$$\bar{X}_2 = \frac{8400 - 6300}{30}$$

$$\bar{X}_2 = 70.$$

2 There are three sections in B.Com. 1st year in a certain college, the number of student in each section and the average marks obtained by them in the paper of statistics in the semester exam as follows

Section	Average marks in Statistics	No. of. Students
A	75	50
B	60	60
C	50	50

$$\bar{X}_{123} = \frac{N_1\bar{X}_1 + N_2\bar{X}_2 + N_3\bar{X}_3}{N_1 + N_2 + N_3} \quad -①$$

Here, $N_1 = 50$

$N_2 = 60$

$N_3 = 50$

$\bar{X}_1 = 75$

$\bar{X}_2 = 60$

$\bar{X}_3 = 50$

$$\text{(1)} \Rightarrow \bar{X}_{123} = \frac{(50\times75)+(60\times60)+(50\times50)}{50+60+50}$$

$$= \frac{3750+3600+2500}{160}$$

$$\bar{X}_{123} = 61.56 \text{ //}.$$

---

3. The mean age of a combined group of men and women is 30 years. if the mean age of the group of men is 32 and that of the group of the women is 27 find the percentage of men and women in the group.

$$\bar{X}_{12} = \frac{N_1\bar{x}_1 + N_2\bar{X}_2}{N1+N_2}$$

Here, $\bar{x}_{12} = 30$

$\bar{x}_1 = 32$

$\bar{x}_2 = 27$

$N_1 = ?$, $N_2 = ?$

Let, the percentage of men to be N_1 the percentage of women to be N_2.

$\therefore N_1 + N_2 = 100$

$\Rightarrow N_2 = 100 - N$

$$① \Rightarrow \quad 30 = \frac{N_1(32) + (100 - N_1)27}{100}$$

$$3000 = 32N_1 + 2700 + 27N_1$$

$$300 = 5N_1$$

$$N_1 = \frac{300}{5}$$

$$N_1 = 60\%.$$

Hence, the percentage of men in the

group is 60 and the perend age of women in the group is 40.

H.W. 1. The mean mark in physics in 100 student of a class was 72. The mean marks of the boys was 75, while their number was 70. Find out mean marks of girls in the class.

2. The four parts of a destribution are as follows.

	Frequency	mean.
Part 1	50	61
Part 2	100	70
Part 3	120	80
Part 4	30	83

find the mean of the entire distribution

1. Solution:

Here, $\bar{X}_{12} = 72$

$\bar{X}_1 = 75$

$N_1 = 70$

$N_2 = 30$

$\bar{X}_2 = ?$

$$\bar{X}_{12} = \frac{N_1\bar{X}_1 + N_2\bar{X}_2}{N_1+N_2} \quad -①$$

$$① \Rightarrow 72 = \frac{70(75) + 30\bar{X}_2}{100}$$

$$7200 = 5250 + 30\bar{X}_2$$

$$30\bar{X}_2 = 7200 - 5250$$

$$\bar{X}_2 = \frac{1950}{30}$$

$$\bar{X}_2 = 65 //.$$

2. We know that,

$$\bar{x}_{1234} = \frac{N_1\bar{x}_1 + N_2\bar{x}_2 + N_3\bar{x}_3 + N_4\bar{x}_4}{N_1 + N_2 + N_3 + N_4}$$

$$= \frac{50 \times 61 + 100 \times 10 + 120 \times 80 + 30 \times 83}{50 + 100 + 120 + 30}$$

$$= \frac{3050 + 1000 + 9600 + 2490}{300}$$

$$= \frac{22140}{300}$$

$$\bar{x}_{1234} = 73.8 \text{ //.}$$

Weighted Mean:

Let $x_1, x_2, \ldots x_n$ be n number Suppose with each x_i there is associated with weight w_i then the weighted average

or weighted mean of $x_1, x_2 \ldots x_n$ is defined by,

$$\overline{x}_w = \frac{\Sigma w_i x_i}{\Sigma w_i}, \quad \text{where } i = 1, 2, \ldots n$$

Example Calculate the weighted mean of the following food articles from the table given below.

Articles of food	Quantity in kgs	Per kg (Price)
Rice	30	4.50
wheat	10	2.75
Sugar	5.5	6.25
Oil	3.5	16.50
flavour	4.5	4.00
ghee	1.5	40.00
onion	9	3.25

Solution:

Weighted mean,

$$\overline{x}_w = \frac{\Sigma w_i x_i}{\Sigma w_i}$$

Articles of food	Quantity in kgs	Price per kg.	$w_i x_i$
Rice	30	4.50	135.00
wheat	10	9.75	27.50
Sugar	5.5	6.25	34.38
oil	3.5	16.50	57.75
flavour	4.5	4.00	18.00
ghee	1.5	40.00	60.00
onion	9	3.25	29.25
Total	64		361.88

$$\overline{x}_w = \frac{\Sigma w_i x_i}{\Sigma w_i}$$

$$= \frac{361.88}{64}$$

$$\overline{x}_w = \text{Rs. } 5.65 \text{ A.}$$

1 An examination was held to decide about the award of a scholarship in the university of Delhi. The weight of various

Subject ware different. The marks obtained by two candidate of 100 in each Subject are given below.

Subject	weight (w_i)	Student A marks	Student B marks.
Tamil	4	70	80
English	3	90	75
Mathematics	2	50	60
Physics	1	56	45

If the candidate getting highest marks is to be awared the scholarship who should get it.

Solution:

Subject	weight	Student A's mark		Student B's Mark	
		x_i	$w_i x_i$	x_i	$w_i x_i$
Tamil	4	70	280	80	320
English	3	90	270	75	225
Mathe-matics	2	50	100	60	120
Physics	1	50	50	45	45
Total			700		710

Student A.

$$\bar{x}_w = \frac{\Sigma x_i w_i}{\Sigma w_i}$$

$$= \frac{700}{10}$$

$$= 70\%.$$

Student B,

$$\bar{x}_w = \frac{\Sigma x_i w_i}{\Sigma w_i}$$

$$= \frac{710}{10}$$

$$= 71\%.$$

The average is the highest in Student B.
Student B got the Scholarship.

---

1. Show that the arithmatic mean of the first n natural numbers is $\frac{1}{2}(n+1)$.

2. The weighted mean of the first n natural numbers whose weighted are equal to the corresponding numbers is equal to $\frac{1}{3} \times (2n+1)$

Solutions:

$\bar{x} = \frac{\sum x_i}{n}$ $\quad \bar{x} = \sum \frac{x_i}{n}$

$\bar{x} = \frac{1+2+3+\cdots+n}{n}$

Arithmetic mean of the first n natural number $= \frac{\sum x_i}{n}$

$= \frac{1+2+3+\cdots+n}{n}$

$= \frac{n(n+1)/2}{n}$

$= \frac{n(n+1)}{n \times 2}$

$\bar{x} = \frac{n+1}{2}$ //.

$= \frac{1}{2}(n+1)$

$1+2+\cdots+n = \frac{n(n+1)}{2}$

$1^2+2^2+3^2+\cdots+n^2 = \frac{n(n+1)(2n+1)}{6}$

2. The weighted arithmetic mean of the first n natural numbers $= \frac{\sum w_i x_i}{\sum w_i}$

$= \frac{1^2+2^2+3^2+\cdots+n^2}{1+2+3+\cdots+n}$

$= \frac{n(n+1)(2n+1)/6}{n(n+1)/2}$

$= \frac{n(n+1)(2n+1)}{6} \times \frac{2}{n(n+1)}$

$= \frac{2n+1}{3}$

$= \frac{n(n+1)(2n+1)}{6} \times \frac{2}{n(n+1)}$ $\frac{\frac{2n+1}{6}}{1/2}$

$= \frac{(2n+1)}{3}$ //. $= \frac{1}{3}(2n+1)$ $= \frac{2n+}{6}$

$= \frac{2n+1}{3}$

H.W

Find the weighted arithmetic mean for the following data.

Price per kg.	Quantity Sold in kg.
1.36	14
1.40	11
1.44	9
1.48	6
1.52	4
1.56	2

Solution:

Price per kg	Quantity in kg	$w_i x_i$
1.36	14	19.04
1.40	11	15.4
1.44	9	12.96
1.48	6	8.88
1.52	4	6.08
1.56	2	3.12
Total	46	65.48

$$\bar{x}_w = \frac{\Sigma w_i x_i}{\Sigma w_i}$$

$$= \frac{65.48}{46}$$

$$= 1.423 //.$$

Median:

Median of the frequency distribution is the value of the variate which devides the total frequency into two equal parts. In otherwords median is the value of the variate for which the cumulative frequency is $\frac{1}{2}N$ where N is the total frequency.

In this case of ungrouped data if n values of the variate are arranged in ascending or descending order of magnitude the median is the middle value if n is odd and is take as the arithmetic mean of two middle values if n is even.

Example:-

Consider the values 54, 81, 84, 71, 61, 57, 68, 54, 54, 67, 49. Find the Median.

Soln:-

Arranging these values in ascending order of magnitude we get, 49, 54, 54, 56, 57, 61, 67, 68, 71, 81, 84.

Since there are 11 items.

$\therefore$ median = 61 //

Example: 2

Find the median marks of 10 Students in Statistics test whose marks are given as 40, 90, 61, 68, 72, 43, 50, 84, 75, 33.

Soln:

Arranging in ascending order of magnitude we get,

33, 40, 43, 50, 61, 68, 70, 75, 84, 90.

$\frac{61+68}{2}$.

Here $n = 10$

hence median is the average of two middle values. 61 and 68

$\therefore$ Median $= \frac{61+68}{2} = \frac{129}{2}$

$= 64.5$ //

Note: In the case of the discrete frequency distribution we calculate the median as follows.

(i) Calculate $\frac{1}{2}N = \frac{1}{2}\Sigma f_i$ Σf_i

(ii) Find the cumulative frequency just greater than $\frac{N}{2}$. $\frac{N \Sigma f_i}{2}$

(iii) The corresponding value of the variate is the Median.

Example :1

Consider the following discrete frequency distribution.

x	f	less than cumulative frequency
1	5	5
2	9	14
3	18	32
4	12	44
5	9	53
6	7	60
Total	60	

Here $N = \Sigma f_i = 60$

$\therefore \frac{1}{2} N = \frac{1}{2} \times 60$

$\frac{1}{2} N = \frac{1}{2} \times 60 = 30$

$= 30$

30

The value of x for which cumulative frequency is just greater than 30 is given by $x = 3$.

$\therefore x = 3$ is Median of the frequency distribution.

Definition :-

For a grouped frequency distribution the median class is defined to be the class where the lesser than cumulative frequency is just greater than $\frac{N}{2}$.

Theorem:-

The Median of a grouped frequency distribution is given by

$$\text{Median} = l + \frac{\left(\frac{N}{2} - m\right) h}{f_k}$$

Where, l - is the lower boundary of the median class.

m - is the cumulative frequency above the median class.

f_k - is the frequency corresponding to the median class.

h - is the width of the class.

Example:-

Class	Frequency	less than Cumulative frequency
0 - 5	3	3
5 - 10	8	11
10 - 15	10	21
15 - 20	10	31
20 - 25	9	40
25 - 30	7	47
30 - 35	21	68
35 - 40	15	83
40 - 45	11	94
45 - 50	6	100

Soln:

Here $\Sigma f_i = N = 100$

$\therefore \frac{N}{2} = \frac{100}{2} = 50$

hence the Median Class is 30 - 35

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

here $l = 30$

$m = 47$

$f_k = 21$

$h = 5$

$$\text{Median} = 30 + \frac{(50 - 47)5}{21}$$

$$= 30 + \frac{15}{21}$$

$$= 30.714 //$$

H.W 1. Find the median of the height of 11 Student are given by 66, 65, 64, 70, 61, 60, 56, 63, 60, 67, 62.

2. Find the median of the following distribution.

class	frequency	less than cumulative frequency
10-15	8	8
15-20	15	23
20-25	39	62
25-30	47	109
30-35	52	161
35-40	41	202
40-45	28	230
45-50	16	246
50-55	4	250

Sohn:

Here $\Sigma f_i = N = 250$

$$\frac{N}{2} = \frac{250}{2} = 125$$

hence, the median class is 30-35.

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

here $l = 30$

$m = 109$

$f_k = 52$

$h = 5$

$$\text{Median} = 30 + \frac{(185 - 109)5}{52}$$

$$= 30 + \frac{80}{52}$$

$$= 31.538$$

$$= 32.04$$

---

1. Soln:

Arranging ascending order of magnitude we get,

56, 60, 60, 61, 62, 63, 64, 65, 66, 67, 70

Since, there are 11 items

Mode:

In a distribution the value of the variate which occurs most frequently and around which the other values of variates cluster density is called the mode or modal value of the distribution.

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

Example:-

1. Consider the discrete frequency distribution.

x	1	2	3	4	5	6	7	8	9
y	8	13	47	105	88	9	5	3	2

Sol^{n}:

Here, the maximum frequency is 105.

∴ The value corresponding to this maximum frequency is 4.

Hence, mode is 4.

Note::

Sometimes it may be difficult to find the mode of a discrete frequency distribution by inspection. For example, If the maximum frequency is repeated or if the maximum frequency occurs at the very begining or at the end of the distribution or if there are inequalities in the distribution one cannot find the mode value by inspection.

In such cases we determine the value of the mode by the method of grouping as follows.

2. Find the mode of the following frequency distribution.

Size of shoes	3	4	5	6	7	8	9	10
Persons wearing it	10	38	28	42	45	15	8	4

Soln:

By inspection it is difficult to say which is the modal value because though the highest frequency is 45. which is greater around 42.

Hence, we have the following grouping table.

Sizes of shoes	I	II	III	IV	V	VI
3	10					
		38		76		
4	28					
			66		108	
5	38					
		80				185
6	42					
7	45		87	102		
		60			68	
8	15					
						30
9	8		23			
		15				
10	7					

• In column 1 the original frequencies are written

• Column 2 is obtained by the combining the frequencies 2 by 2.

• If we leave the first frequency in column 1 and combining the rest of the frequencies 2 by 2 we get column 3.

• In column 4 we combine frequency in column 1, 3 by 2 starting from the first frequency.

• The combining frequencies in column 1, 3 by 3 after leaving the first frequency results in column 5.

• Column 6 is obtained by combining the frequencies of column 5 3 by 3 leaving the first two frequencies so find mode we form the following analysis table.

Column Number	Size of Shoes having maximum frequency
I	7
II	5, 6.
III	6, 7
IV	6, 7, 8
V	4, 5, 6
VI	5, 6, 7.

In the analysis table size 6 occurs maximum times (5 times). Hence the modal size of the shoes is 6.

∴ Hence mode is 6.

H.W. ① Find the mode of the following distribution.

Size	2	3	4	5	6	7	8	9	10	11	12
Frequency	3	8	10	12	16	14	10	8	17	5	4

Ans:

Size	I	II	III	IV	V	VI
2	3					
3	8	11		21		
4	10		18		30	
5	12	22				38
6	16		28	42		
7	14	30				
8	10		24		40	32
9	8	18		35		
10	17		25		30	
11	5	22				26
12	4		9			
13	1	5		10		

Column number	Size having maximum frequency
I	10
II	6, 7
III	5, 6
IV	5, 6, 7
V	6, 7, 8
VI	4, 5, 6

In this analysis table size 6 occurs (5 times) maximum times.
∴ Hence the Mode is 6.

In the case of grouped frequency distribution the mode is computed by the formula $\text{Mode} = l + \frac{(f - f_1)h}{2f - f_1 - f_2}$

where l is the lower boundary of the modal class (class having maximum frequency). f is the maximum frequency f_1 and f_2 are frequencies of the classes proceeding and following the modal class.

h is width of the class.

An alternate formula for finding the mode is given by

$$\text{Mode} = l + \frac{hf_2}{f_1 + f_2}.$$

Note!

A frequency distribution may have more than one mode

in which it is called Multi Modal distribution. If there is only one mode it is called Bimodal distribution

Note 2:-

There is an relationship between Mean, Median, Mode

Mean − Mode = 3 (Mean − Median)

i.e., Mode = 3 Median − 2 Mean

Problems:-

1. Calculate the mode for the frequency distribution.

Mark	No. of Students
0 − 9	6
10 − 19	29
20 − 29	87
30 − 39	181
40 − 49	247
50 − 59	263
60 − 69	133
70 − 79	43

80 - 89	9
90 - 99	2

Solu:

Marks	I	II	III	IV	V	VI
0 - 9	6			122	297	
10 - 19	29	35				
20 - 29	87		116			515
30 - 39	181	268				
40 - 49	247		428	691		
50 - 59	263	510			643	
60 - 69	133		396			439
70 - 79	43	176		185		
80 - 89	9		52		54	
90 - 99	2	11				

Analysis table:-

Columns	Classes having maximum frequency.
I	50-59
II	40-49, 50-59
III	30-39, 40-49
IV	30-39, 40-49, 50-59
V	40-49, 50-59, 60-69
VI	20-29, 30-39, 40-49

From the analysis table the modal class is 40-49.

Since the class 40-49 occurs maximum times.

The true class limit of the modal class is 39.5 - 49.5.

Hence $l = 39.5$

$f_1 = 181$

$f_2 = 263$

$h = 10$

$\therefore$ Mode $= l + \frac{h f_2}{f_1 + f_2}$.

$= 39.5 + \frac{10 \times 263}{181 + 263}$

$= 39.5 + \frac{2630}{444}$

$= 45.42$ //

2. Given that the mode of the following frequency distribution of 70 students is 58.75. Find the missing frequency f_1 and f_2.

Class	Frequency	less than c.f
52 - 55.	15	15
55 - 58	f_1	$15 + f_1$
58 - 61	25 (x)	$40 + f_1$
61 - 64	f_2	$40 + f_1 + f_2 = N$

Sol^n:

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

Since, $N = 70$

$\therefore 40 + f_1 + f_2 = 70$

$f_1 + f_2 = 70 - 40$

$f_1 + f_2 = 30 \longrightarrow (1)$

here, $l = 58$

$h = 3$

$f = 25$

given mode = 58.75

$$58.75 = 58 + \frac{3(85-f_1)}{2\times 85 - f_1 - f_2}$$

$$58.75 - 58 = \frac{3(85-f_1)}{50 - f_1 - f_2}$$

$$0.75 = \frac{75-3f_1}{50-f_1-f_2}$$

$$0.75(50-f_1-f_2) = 75-3f_1$$

$$37.5 - 0.75f_1 - 0.75f_2 = 75 - 3f_1$$

$$37.5 - 0.75f_1 - 0.75f_2 - 75 + 3f_1 = 0$$

$$-37.5 + 2.25f_1 - 0.75f_2 = 0$$

$$2.25f_1 - 0.75f_2 = 37.5 \quad \text{—(2)}$$

(1) $\Rightarrow$ $f_1 = 30 - f_2$

(2) $\Rightarrow$ $2.25 \times (30 - f_2) - 0.75f_2 = 37.5$

$$67.5 - 2.25f_2 - 0.75f_2 - 37.5 = 0$$

$$30 - 3f_2 = 0$$

$$-3f_2 = -30$$

$$f_2 = \frac{-30}{-3}$$

$$\boxed{f_2 = 10}$$

① $\Rightarrow$ $f_1 + 10 = 30$

$$f_1 = 30 - 10$$

$$\boxed{f_1 = 20}$$ //

Q.10. Find the mode

Class	Frequency
11 - 15	8
16 - 20	15
21 - 25	39
26 - 30	47
31 - 35	52
36 - 40	41
41 - 45	28

46-50	16
51-55	4

2. Calculate the mode

x	1-9	9-17	17-25	25-33	33-41	41-49	49-57
f	20	31	27	15	10	7	8

3. Find the mode.

Weight	No. of Students.
90 - 100	3
100 - 110	2
110 - 120	18
120 - 130	22
130 - 140	21
140 - 150	19
150 - 160	10
160 - 170	3

170 - 180	2

Answer:-

Soln:-

Hence we have the following analysis table.

Class	I	II	III	IV	V	VI.
11 - 15	8					
16 - 20	15	23		62		
21 - 25	39		54		101	
26 - 30	47	86				138
31 - 35	52		99	140		
36 - 40	41	93			121	
41 - 45	28		69			85
46 - 50	16	44		48		
51 - 55	4		20			

Columns	Classes having Maximum frqncy
I	31 - 35
II	31 - 35 , 36 - 40
III	26 - 30 , 31 - 35
IV	26 - 30, 31 - 35, 36 - 40
V	31 - 35, 36 - 40, 41 - 45
VI	21 - 25, 26 - 30, 31 - 35

From the analysis table the modal class is 31 - 35. Since the class 31 - 35 occurs maximum times

The true class limits of the modal class is 31 - 35.

$$\text{Mode} = l + \frac{hf_2}{f_1 + f_2}$$

here, $l = 30.5$

$f_1 = 47$

$f_2 = 41$

$h = 5$

$$\text{Mode} = 30.5 + \frac{5(41)}{47+41}$$

$$= 30.5 + \frac{205}{88}$$

$$= 32.829$$

$$= 32.83 \text{ //}$$

8. Ans:-

x	I	II	III	IV	V	VI
1-9	20					
9-17	31	51		78		
17-25	27		58		73	
25-33	15	42				52
33-41	10		25	32		
41-49	7	17			25	
49-57	8		15			

Analysis table:-

Columns	Maximum frequency
I	9-17
II	1-9, 9-17
III	9-17, 17-25
IV	1-9, 9-17, 17-25
V	9-17, 17-25, 25-33
VI	17-25, 25-33, 33-41

From the analysis table class is 9-17.

9-17 occurs maximum times.

The true class limit of the model class is 8.5 - 17.5

Hence $l = 8.5$

$f_1 = 20$

$f_2 = 27$

$h = 9$

$$\text{Mode} = l + \frac{hf_2}{f_1 + f_2}$$

$$= 8.5 + \frac{9 \times 27}{20 + 27}$$

$$= 8.5 + \frac{243}{47}$$

$$= 8.5 + 5.17$$

$$= 13.67 //$$

Unit : III

Measures of Dispersion.

Range :-

Range is the most simple and obvious measure of dispersion. It is the difference between the maximum and the minimum values of the variate.

Example :-

The following are the wages of 8 workers of a factory. Find the range wages in Rupees 1400, 1450, 1520, 1380, 1485, 1495, 1575, 1440.

Sol^n :-

Range = Maximum value − Minimum value.

Range = 1575 - 1280.

= 195

Standard Deviation :-

The Standard Deviation σ of the frequency distribution is defined by $\sigma = \left[\frac{\Sigma f_i (x_i - \bar{x})^2}{N}\right]^{1/2}$ where $N = \Sigma f_i$ and $\bar{x}$ is the arithmetic mean of the frequency distribution.

The Square of the Standard deviation of a frequency distribution is called the variance of the frequency distribution. Hence ~~variance~~ Variance $= \sigma^2$.

Note:

$$\sigma = \left[\frac{\Sigma f_i x_i^2}{N} - \left(\frac{\Sigma f_i x_i}{N}\right)^2\right]^{1/2}$$

Problems :-

Find (i) Range

(ii) Standard Deviation for the following marks of 10 Students.

20, 22, 27, 30, 40, 48, 45, 32, 31, 25

Soln:-

(i) Range = Maximum value - Minimum value

= 48 - 20

= 28,

(ii) Standard Deviation

$$\sigma = \left[\frac{\Sigma x_i^2}{n} - \left(\frac{\Sigma x_i}{n} \right)^2 \right]^{1/2} \quad \text{---- (1)}$$

$\Sigma x_i^2 = 11652$

$\Sigma x_i = 330$

$n = 10$

① ⇒ $\sigma = \left[\frac{11652}{10} - \left(\frac{330}{10}\right)^2\right]^{1/2}$

$= \sqrt{(76.2)}$

$= 8.73$ //

S.D

Find Standard Deviation

Mark	10	9	8	7	6	5	4	3	2	1
frequency	1	5	11	15	12	7	3	3	0	1

Quartile Deviation:- (Semi inter quatile range)

Consider the frequency distribution with total frequency N. The value of the variate for which the cumulative frequency is $N/4$ is called the first quartile or lower quartile. and it is denoted by Q_1.

The value of the variate for which the cumulative frequency is $3N/4$ is called the third quartile or

Upper quartile and it is denoted by Q_3

Median is the second quartile and it is denoted by Q_2.

Ungrouped data:-

Let $i = \left[\frac{1}{4}(n+1)\right]$ = the integral part of $\frac{1}{4}(n+1)$.

Let $q = \frac{1}{4}(n+1) - \left[\frac{1}{4}(n+1)\right]$

hence q is the fractional part.

$\therefore Q_i = x_i + q(x_{i+1} - x_i)$

$Q_3 = x_i + q(x_{i+1} - x_i)$

Where $i = \left[\frac{3}{4}(n+1)\right]$ and

$q = \frac{3}{4}(n+1) - \left[\frac{3}{4}(n+1)\right]$

Grouped data:-

$$Q_1 = l + \frac{(N/4 - m)h}{f_k} \quad \text{and}$$

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

where $l \rightarrow$ is the lower limit of the class in which the particular quartile lies.

$f_k \rightarrow$ is the frequency of this class.

$h \rightarrow$ is the width of the class

$m \rightarrow$ is the cumulative frequency of the preceding class.

Problems:-

1. Find the quartiles Q_1, Q_3 given by the following data.

66, 65, 64, 70, 61, 60, 56, 63, 60, 67, 62

Sol:

Here $n = 11$

56, 60, 60, 61, 62, 63, 64, 65, 66, 67, 70

First Quartile

$$Q_1 = x_i + q(x_{i+1} - x_i) \longrightarrow (1)$$

where $i = \left[\frac{1}{4}(n+1)\right]$ and

$$q = \frac{1}{4}(n+1) - \left[\frac{1}{4}(n+1)\right]$$

Arranging the given data in ascending order

56, 60, 60, 61, 62, 63, 64, 65, 66, 67, 70

$$\therefore \quad i = \left[\frac{1}{4}(11+1)\right]$$

$$= \left(\frac{12}{4}\right]$$

$$= 3$$

$$\therefore \quad q = \frac{1}{4}(11+1) - \left[\frac{1}{4}(11+1)\right]$$

$$= \frac{12}{4} - \frac{12}{4}$$

$$= 3 - 3$$

$$= 0$$

① $\Rightarrow$ $Q_1 = x_3 + 0$

$= 60.$

$\therefore Q_1 = 60$ //

Third Quartile:-

$$Q_3 = x_i + q\,(x_{i+1} - x_i) \quad \text{——②}$$

where $i = \left[\frac{3}{4}(n+1)\right]$ and

$$q = \frac{3}{4}(n+1) - \left[\frac{3}{4}(n+1)\right]$$

$$i = \left[\frac{3}{4}(11+1)\right]$$

$$= \left[\frac{3 \times 12}{4}\right]$$

$$= 9$$

$$q = \left\{\frac{3}{4}(11+1)\right\} - \left[\frac{3}{4}(11+1)\right]$$

$$= \frac{12 \times 3}{4} - \frac{12 \times 3}{4}$$

$$= \frac{36}{4} - \frac{36}{4}$$

$$= 0$$

(2) $\rightarrow$ $Q_3 = x_9 + 0$

$$= 66$$

Note:-

Variance of the first n natural number is (i) $\sigma^2 = \frac{1}{12}(n^2-1)$. (ii) $\sigma^2 = \frac{1}{n_1+n_2}[n_1\sigma_1^2 + n_2\sigma_2^2 + n_1 d_1^2 + n_2 d_2^2]$

$d_1 = \bar{x_1} - \bar{x}$ & $d_2 = \bar{x_2} - \bar{x}$.

Find the Quartiles Q_1 and Q_3 for the following data.

40, 90, 61, 68, 78, 43, 50, 84, 75, 33.

Soln:

$$n = 10$$

First Quartile:

$$Q_1 = x_i + q(x_{i+1} - x_i) \longrightarrow (1)$$

$$i = \left[\frac{1}{4}(n+1)\right] \text{ and } q = \frac{1}{4}(n+1) - \left[\frac{1}{4}(n+1)\right]$$

Arranging the given data in ascending order

38, 40, 43, 50, 61, 68, 78, 75, 84, 90.

$$i = \left[\frac{1}{4}(10+1)\right]$$

$$= \left[\frac{11}{4}\right]$$

$$= [2.75]$$

$$= 2$$

$$\therefore q = \frac{1}{4}(10+1) - 2$$

$$= 2.75 - 2$$

$$= 0.75$$

① $\Rightarrow x_2 + 0.75(x_3 - x_2)$

$\Rightarrow 40 + 0.75(43 - 40)$

$\Rightarrow 40 + 2.25$

$\Rightarrow 42.25$

Third Quartile :-

$$Q_3 = x_i + q(x_{i+1} - x_i) \longrightarrow ②$$

$$i = \left[\frac{3}{4}(10+1)\right]$$

$$= \left[\frac{3}{4}(11)\right]$$

$$= \left[\frac{33}{4}\right]$$

$$= 8.25$$

$\therefore q = \frac{3}{4}(10+1) - 8$

$= 8.25 - 8$

$= 0.25$

(2) $\Rightarrow x_8 + 0.25\,(x_9 - x_8)$

$\Rightarrow 75 + 0.25\,(84 - 75)$

$\Rightarrow 75 + 0.25$

$\Rightarrow 77.5$

2. Soln:

Marks x_i	frequency f_i	$f_i x_i$	$f_i x_i^2$
10	1	10	100
9	5	45	405
8	11	88	704
7	15	105	735
6	12	72	432
5	7	35	175
4	3	12	48
3	3	9	27
2	0	0	0
1	1	1	1
N =	58	377	2627

$$\sigma = \left[\frac{\Sigma f_i x_i^2}{N} - \left(\frac{\Sigma f_i x_i}{N} \right)^2 \right]^{1/2}$$

$$= \left[\frac{2627}{58} - \left(\frac{377}{58} \right)^2 \right]^{1/2}$$

$$= \left[45.29 - 42.85 \right]^{1/2}$$

$$= \sqrt{3.04}$$

$$= 1.74 \text{ //.}$$

3. The following table gives the monthly wages of workers in a factory. Compute

(i) Standard Deviation.

(ii) Quartile Deviation.

(iii) Mean (Arithmetic) $\bar{x}$.

Monthly Wages	No. of workers
125 - 175	2
175 - 225	22
225 - 275	19
275 - 325	14
325 - 375	3
375 - 425	4
425 - 475	6
475 - 525	1
525 - 575	1

Soln:

(i) S.D

$$\sigma = \left[\frac{\Sigma f_i (x_i - \bar{x})^2}{N}\right]^{1/2} \longrightarrow (1)$$

mid x_i	f_i	u_i	$f_i u_i$	$f_i u_i^2$	C.f
150	2	−4	−8	32	2
200	22	−3	−66	198	24
250	19	−2	−38	76	43
300	14	−1	−14	14	57
350	3	0	0	0	60
400	4	1	4	4	64
450	6	2	12	24	70
500	1	3	3	9	71
550	1	4	4	16	72
Total	72		−108	373	

(iii) Mean $\bar{x} = A + h\bar{u}$

$= 350 + 50 \times \left(-\frac{108}{72}\right)$

$= 278.47$

(i) Standard Deviation:-

$$\sigma = h\left[\frac{\sum f_i u_i^2}{N} - \left(\frac{\sum f_i u_i}{N}\right)^2\right]^{1/2}$$

$$= 50\left[\frac{373}{72} - \left(\frac{-103}{72}\right)^2\right]^{1/2}$$

$$= 50\left[5.180 - (1.430)^2\right]^{1/2}$$

$$= 50\left[5.180 - 2.046\right]^{1/2}$$

$$= 50 \sqrt{3.134}$$

$$= 88.515$$

$$\sigma = 88.52 //$$

(ii) Quartile Deviation:-

$$Q_1 = l + \frac{(N/4 - m)h}{f_k}$$

Here $l = 175$

$N/4 = 18$

$m = 2$

$h = 50$ $f_k = 22$.

$$Q_1 = 175 + \frac{(18-2)50}{22}$$

$$= 211.86 //$$

Third Quartile :-

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

here,

$l = 275$

$3N/4 = 54$

$m = 43$

$h = 50$

$f_k = 14$

$$Q_3 = 275 + \frac{(54-43)50}{14}$$

$$= 314.285$$

$$= 314.29 //$$

Quartile Deviation $= \frac{1}{2}(Q_3 - Q_1)$.

$= \frac{1}{2}(314.29 - 211.36)$

$= 51.465$

$= 51.47$

Problems:-

1. Find (i) Mean (ii) Standard Deviation (iii) Quartile Deviat

Class	Frequency
11 - 15	8
16 - 20	15
21 - 25	39
26 - 30	47
31 - 35	52
36 - 40	41
41 - 45	28
46 - 50	16
51 - 55	4

2) Calculate Quartile Deviation.

Wages	No. of wages
below 35	14
35 - 37	160
38 - 40	95
41 - 43	24
over 43	7

1. Soln:

A = 33 h = 5

Class	mid x_i	f_i	u_i	$f_i u_i$	$f_i u_i^2$	C.F
10.5 - 15.5	13	8	-4	-32	128	8
15.5 - 20.5	18	15	-3	-45	135	23
20.5 - 25.5	23	39	-2	-78	156	62
25.5 - 30.5	28	47	-1	-47	47	109
30.5 - 35.5	33	52	0	0	0	161
35.5 - 40.5	38	41	1	41	41	202
40.5 - 45.5	43	28	2	56	112	230
45.5 - 50.5	48	6	3	18	144	236
50.5 - 55.5	53	4	4	16	64	250
Total		250		-41	827	

(i) Mean

$$\bar{x} = A + h\bar{u}$$

$$= 33 + 5\left(\frac{\Sigma f_i u_i}{\Sigma f_i}\right)$$

$$= 33 + 5\times\left(\frac{-41}{250}\right)$$

$$= 33 - 0.82$$

$$= 32.18$$

(ii) Standard Deviation.

$$\sigma = h\left[\frac{\Sigma f_i u_i^2}{N} - \left(\frac{\Sigma f_i u_i}{N}\right)^2\right]^{1/2}$$

$$= 5\left[\frac{827}{250} - \left(\frac{-41}{250}\right)^2\right]^{1/2}$$

$$= 5\left[3.308 - 0.026\right]^{1/2}$$

$$= 5\left[3.282\right]^{1/2}$$

$$= 5\sqrt{3.282}$$

$$= 5 \times 1.811 = 9.055$$

(iii) Quartile Deviation

$$Q_1 = l + \frac{(N/4 - m)h}{f_k}$$

here $l = 20.5$

$N/4 = 62.5$

$m = 23$

$h = 5$

$f_k = 39$

$$Q_1 = 20.5 + \frac{(62.5 - 23)5}{39}$$

$$= 20.5 + \frac{39.5 \times 5}{39}$$

$$= 25.56 //$$

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

$l = 35.5$ $\quad h = 5$

$3N/4 = 187.5$ $\quad f_k = 41$

$m = 161$

$$Q_3 = 35.5 + \frac{(187.5 - 161) \times 5}{41}$$

$$= 38.73 //$$

Quartile Deviation $= 1/2 (Q_3 - Q_1)$

$= 1/2 [38.73 - 35.56]$

$= 6.585$

$= 6.959 //$

Wages	No. of. wages.	Cumulative frequency
below 35	14	14
34.5 - 37.5	60	74
37.5 - 40.5	95	169
40.5 - 43.5	24	193
Over 43	7	200
Total	200.	

First Quartile:-

$$Q_1 = l + \frac{(N/4 - m)h}{f_k}$$

here, $l = 34.5$

$m = 14$

$N/4 = 50$

$h = 3$

$f_k = 60$

$$Q_1 = 34.5 + \frac{(50-14)3}{60}$$

$$= 36.3 //$$

Third Quartile:-

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

here, $l = 37.5$

$3N/4 = 150$

$m = 74$

$h = 3$

$f_k = 95$

$$Q_3 = 37.5 + \frac{(150-74)_2}{95}$$

$$= 39.9$$

Quartile Deviation $= \frac{1}{2}(Q_3 - Q_1)$.

$$= \frac{1}{2}(39.9 - 36.3)$$

$$Q = 1.8 //$$

Note:

(i) The Standard Deviation σ is independent of change of origin and is dependent on change of scale

(ii) $\sigma^2 = S^2 - d^2$, where $d = \bar{x} - A$

Mean deviation:

$$\text{Mean deviation} = \frac{\Sigma f_i |x_i - A|}{N}$$

Coefficient of Variance [C.V] of the frequency distribution is defined to be

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

Unit - II

Correlation

Consider a set of bivariate data $x_i, y_i \quad i = 1, 2, \ldots, n$ if there is a change in one variable corresponding to change in other variable we say that the variable are correlated. If the two variable deviate in the same direction the correlation is said to be direct or positive. If they always deviate in the opposite direction the correlation is said to be inverse or negative. If the change in one variable corresponse to the proposal to the other variable then the correlation is perfect.

Karl pearson's coefficiant of correlation:

Karl pearson's coefficiant of correlation between the variable x and y is difined by $r(x,y) = \frac{\Sigma(xi-\bar{x})(yi-\bar{y})}{n\sigma x\,\sigma y}$

where $\bar{x} = \frac{\Sigma xi}{n}$, $\bar{y} = \frac{\Sigma yi}{n}$

$\sigma x = \sqrt{\frac{\Sigma(xi-\bar{x})^2}{n}}$ & $\sigma y = \sqrt{\frac{\Sigma(yi-\bar{y})^2}{n}}$

co-variance between x & y is difined by

covariance $(x,y) = \frac{\Sigma(xi-\bar{x})(yi-\bar{y})}{n}$

Hence $r(x,y) = \frac{cov(x,y)}{\sigma x\,\sigma y}$

1) The Heights and weights case of 5 students are given below.

Height in cm (x)	160	161	162	163	164
Weight in kg (y)	50	53	54	56	57

Find the correlation between x & y

$$\bar{x} = \frac{\sum x_i}{n}$$

$$= \frac{160 + 161 + 162 + 163 + 164}{5}$$

$$= 162$$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{50 + 53 + 54 + 56 + 57}{5}$$

$$= 54$$

x	y	$x - \bar{x}$ $x - 162$	$y - \bar{y}$ $y - 54$	$(x - \bar{x})^2$	$(y - \bar{y})^2$	$(x - \bar{x})(y - \bar{y})$
160	50	-2	-4	4	16	8
161	53	-1	-1	1	1	1
162	54	0	0	0	0	0
163	56	1	2	4	4	2
164	57	2	3	1	9	6
		$\Sigma(x_i - \bar{x})$ $= 0$	$\Sigma(y_i - \bar{y})$ $= 0$	$\Sigma(x - \bar{x})^2$ $= 10$	$\Sigma(y - \bar{y})^2$ $= 30$	$\Sigma(x - \bar{x})(y - \bar{y})$ $= 17$

$$\sigma_x^2 = \frac{\Sigma(x_i - \bar{x})^2}{n}$$

$$= \frac{10}{5} = 2$$

$\sigma_x = \sqrt{3}$

$\sigma_y^2 = \frac{\Sigma(y_i - \bar{y})^2}{n}$

$= \frac{30}{5} = 6$

$\sigma_y = \sqrt{6}$

correlation between x & y is

$r_{(x,y)} = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{n \sigma_x \sigma_y}$

$= \frac{17}{5 \times \sqrt{3} \times \sqrt{6}} = \frac{17}{5 \times 2\sqrt{3}}$

$= \frac{17}{5 \times 3.464} = \frac{17}{17.320} = 0.98$

Theorem (1)

1) Proove that $r_{(x,y)} = \frac{n \Sigma x_i y_i - \Sigma x_i \Sigma y_i}{[n \Sigma x_i^2 - (\Sigma x_i)^2]^{1/2} [n \Sigma y_i^2 - (\Sigma y_i)^2]^{1/2}}$

Proof:

we have $r_{(x,y)} = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{n \sigma_x \sigma_y}$ ——①

$\Sigma (x_i - \bar{x})(y_i - \bar{y}) = \Sigma [x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x} \bar{y}]$

$= \Sigma x_i y_i - \Sigma x_i \bar{y} - \Sigma \bar{x} y_i + \Sigma \bar{x} \bar{y}$

$$= \Sigma x_i y_i - \bar{y}\Sigma x_i - \bar{x}\Sigma y_i + n\bar{x}\bar{y}$$

$$= \Sigma x_i y_i - \bar{y} n\bar{x} - \bar{x} n\bar{y} + n\bar{x}\bar{y}$$

$$= \Sigma x_i y_i - n\bar{x}\bar{y} - n\bar{x}\bar{y} + n\bar{x}\bar{y}$$

$\bar{x} = \frac{\Sigma x_i}{n}$, $\bar{y} = \frac{\Sigma y_i}{n}$

$$= \Sigma x_i y_i - n \frac{\Sigma x_i}{n} \cdot \frac{\Sigma y_i}{n}$$

$$= \Sigma x_i y_i - \frac{\Sigma x_i \Sigma y_i}{n}$$

$$= \frac{n\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{n} \longrightarrow (2)$$

$$\sigma_x^2 = \frac{\Sigma (x_i - \bar{x})^2}{n}$$

$$= \frac{\Sigma (x_i^2 - 2x_i\bar{x} + \bar{x}^2)}{n}$$

$$= \frac{\Sigma x_i^2 - \Sigma 2x_i\bar{x} + \Sigma \bar{x}^2}{n}$$

$$= \frac{\Sigma x_i^2}{n} - \frac{2\bar{x}\Sigma x_i}{n} + \frac{\Sigma \bar{x}^2}{n}$$

$$= \frac{\Sigma x_i^2}{n} - \frac{2\bar{x}\,\Sigma x_i}{n} + \frac{n\bar{x}^2}{n}$$

$$= \frac{\sum x_i^2}{n} - \frac{2\bar{x} n \bar{x}}{n} + \bar{x}^2$$

$$= \frac{\sum x_i^2}{n} - 2\bar{x}^2 + \bar{x}^2$$

$$= \frac{\sum x_i^2}{n} - \bar{x}^2$$

$$= \frac{\sum x_i^2}{n} - \left(\frac{\sum x_i}{n}\right)^2$$

$$= \frac{\sum x_i^2}{n} - \frac{(\sum x_i)^2}{n^2}$$

$$= \frac{n\sum x_i^2 - (\sum x_i)^2}{n^2}$$

$$\sigma_x = \left[\frac{n\sum x_i^2 - (\sum x_i)^2}{n}\right]^{1/2} \rightarrow (2)$$

III^{ly} $$\sigma_y = \frac{n\sum y_i^2 - (\sum y_i)^2}{n} \rightarrow (4)$$

Sub (2), (3) & (4) in (1)

$$r_{(x,y)} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n} \times \frac{n}{n\left[n\sum x_i^2 - (\sum x_i)^2\right]^{1/2} \left[n\sum y_i^2 - (\sum y_i)^2\right]^{1/2}}$$

$$r_{(x,y)} = \frac{n\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{\left[n\Sigma x_i^2 - (\Sigma x_i)^2\right]^{1/2}\left[n\Sigma y_i^2 - (\Sigma y_i)^2\right]^{1/2}}$$

Hence proved.

2) Theorem(2) Prove that the correlation coefficient is independent of the change of origin and scale.

A & B → origin

h & K → scale

Let $u_i = \frac{x_i - A}{h}$

$v_i = \frac{y_i - B}{K}$

To prove $r_{(x,y)} = r_{(u,v)}$

$$u_i = \frac{x_i - A}{h}$$

$$hu_i = x_i - A$$

$$x_i = hu_i + A$$

$$\frac{x_i}{n} = \frac{hu_i}{n} + \frac{A}{n}$$

$$\frac{\sum x_i}{n} = \frac{\sum hu_i}{n} + \frac{\sum A}{n}$$

$$\bar{x} = h\bar{u} + \frac{nA}{n}$$

$$\bar{x} = h\bar{u} + A$$

$$x_i - \bar{x} = hu_i + A - h\bar{u} - A$$

$$= hu_i - h\bar{u}$$

$$= h(u_i - \bar{u})$$

$$(x_i - \bar{x})^2 = h^2 (u_i - \bar{u})^2$$

$$\frac{(x_i - \bar{x})^2}{n} = \frac{h^2 (u_i - \bar{u})^2}{n}$$

$$\frac{\sum (x_i - \bar{x})^2}{n} = h^2 \frac{\sum (u_i - \bar{u})^2}{n}$$

$$\sigma_x^2 = h^2 \sigma_u^2$$

$$\sigma_x = h \sigma_u$$

Similarly $y_i - \bar{y} = k(v_i - \bar{v})$

$$\sigma_y = k \sigma_v$$

$$r_{(x,y)} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n \sigma_x \sigma_y}$$

$$= \frac{\sum k(u_i - \bar{u})\, k(v_i - \bar{v})}{n (k\sigma_u)(k\sigma_v)}$$

$$= \frac{\sum (u_i - \bar{u})(v_i - \bar{v})}{n \sigma_u \sigma_v}$$

$$r_{(x,y)} = r_{(u,v)}$$

Hence proved.

Theorem (3)

3) prove that $-1 \le r \le 1$

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n \sigma_x \sigma_y}$$

$$= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}} \cdot \sqrt{\frac{\sum (y_i - \bar{y})^2}{n}}}$$

$$= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n \frac{\sqrt{\sum (x_i - \bar{x})^2}}{\sqrt{n}} \cdot \frac{\sqrt{\sum (y_i - \bar{y})^2}}{\sqrt{n}}}$$

$$= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}}$$

(or)

$$\beta_{yx} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}}$$

Let $(x_i - \bar{x}) = a_i$, $y_i - \bar{y} = b_i$

$$r = \frac{\sum a_i b_i}{\sqrt{\sum a_i^2}\sqrt{\sum b_i^2}}$$

squaring,

$$r^2 = \frac{(\sum a_i b_i)^2}{\sum a_i^2 \sum b_i^2} \rightarrow (1)$$

By schwartz inequality we have,

$$(\sum a_i b_i)^2 \leq \sum a_i^2 \sum b_i^2$$

$$(1) \Rightarrow r^2 \leq \frac{\sum a_i^2 \sum b_i^2}{\sum a_i^2 \sum b_i^2}$$

$$r^2 \leq 1$$

(ie) $|r| \leq 1$

(ie) $-1 \leq r \leq 1$

Hence proved

Note:

(i) If $r = 1$ the correlation is perfect and positive

(ii) If $r = -1$ the correlation is perfect and negetive.

(iii) If $r = 0$ the variables are uncorrelated.

(iv) If the variables x & y are uncorrelated then $cov(x,y) = 0$

Theorem: (4)

Prove that $r(x,y) = \frac{\sigma x^2 + \sigma y^2 - (\sigma x - y)^2}{2 \sigma x \sigma y}$

$$\sigma_{x-y}^2 = \frac{\Sigma[(x_i - y_i) - (\bar{x} - \bar{y})]^2}{n}$$

$$\sigma_{x-y}^2 = \frac{\Sigma[(x_i - y_i) - (\bar{x} - \bar{y})]^2}{n}$$

$$= \frac{\Sigma[x_i - y_i - \bar{x} + \bar{y}]^2}{n}$$

$$= \frac{\Sigma [(x_i - \bar{x}) - (y_i - \bar{y})]^2}{n}$$

$$= \frac{\Sigma [(x_i - \bar{x})^2 - 2(x_i - \bar{x})(y_i - \bar{y}) + (y_i - \bar{y})^2}{n}$$

$$= \frac{\Sigma (x_i - \bar{x})^2}{n} - \frac{2\Sigma (x_i - \bar{x})(y_i - \bar{y})}{n} + \frac{\Sigma (y_i - \bar{y})^2}{n}$$

$$\rightarrow r_{xy} = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{n \sigma_x \sigma_y}$$

$$= \sigma_x^2 - \frac{2 r_{xy} n \sigma_x \sigma_y}{n} + \sigma_y^2$$

$$\sigma_{x-y}^2 = \sigma_x^2 - 2 r_{xy} \sigma_x \sigma_y + \sigma_y^2$$

$$2 r_{xy} \sigma_x \sigma_y = \sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2$$

$$r_{xy} = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2 \sigma_x \sigma_y}$$

∴ Hence proved

1) Ten Students obtained the following % of mark in the college internal test(x) and in the final university exam(Y) find the correlation co-efficient between the marks of two test

x	51	69	63	49	58	60	65	63	46	58
y	49	72	75	50	48	60	70	48	60	52

We have

$$r_{(x,y)} = r_{uv}$$

Let $u_i = x_i - 50$

$v_i = y_i - 48$

$$r_{uv} = \frac{n\Sigma u_i v_i - \Sigma u_i \Sigma v_i}{\left[n\Sigma u_i^2 - (\Sigma u_i)^2\right]^{1/2}\left[n\Sigma v_i^2 - (\Sigma v_i)^2\right]^{1/2}}$$

x_i	y_i	$u_i = x_i - 50$	$v_i = y_i - 48$	u_i^2	v_i^2	$u_i v_i$
51	49	1	1	1	1	1
63	72	13	24	169	576	312
63	75	13	27	169	729	351
49	50	−1	2	1	4	−2
50	48	0	0	0	0	0
60	60	10	12	100	144	120
65	70	15	22	225	484	330
63	48	13	0	169	0	0
46	60	−4	12	16	144	−48
50	56	0	8	0	64	0
		$\Sigma u_i = 60$	$\Sigma v_i = 108$	$\Sigma u_i^2 = 850$	$\Sigma v_i^2 = 2146$	$\Sigma u_i v_i = 1064$

$$r_{uv} = \frac{10 \times 1064 - 60 \times 108}{\left[10 \times 850 - (60^2)\right]^{1/2} \left[10 \times 2146 - 108^2\right]^{1/2}}$$

$$= \frac{10640 - 6480}{(8500 - 3600)^{1/2} \left[21460 - 11664\right)^{1/2}}$$

$$= \frac{4160}{(4900)^{1/2}(9796)^{1/2}}$$

$$= \frac{4160}{70 \times 98.97} = \frac{4160}{6927.9}$$

$$= 0.6$$

2) Find the correlation coefficient between two variables

X	300	350	400	450	500	550	600	650	700
Y	800	900	1000	1100	1200	1300	1400	1500	1600

We have

$$\gamma_{xy} = \gamma_{uv}$$

Let $u_i = \frac{x_i - 500}{50}$

$$v_i = \frac{y_i - 1200}{100}$$

$$\gamma_{uv} = \frac{n\Sigma u_i v_i - \Sigma u_i \Sigma v_i}{\left[n\Sigma u_i^2 - (\Sigma u_i)^2\right]^{1/2}\left[n\Sigma v_i^2 - (\Sigma v_i)^2\right]^{1/2}}$$

x_i	y_i	$u_i = \frac{x_i - 500}{50}$	$v_i = \frac{y_i - 1200}{100}$	u_i^2	v_i^2	$u_i v_i$
300	800	−4	−4	16	16	16
350	900	−3	−3	9	9	9
400	1000	−2	−2	4	4	4
450	1100	−1	−1	1	1	1
500	1200	0	0	0	0	0
550	1300	1	1	1	1	1
600	1400	2	2	4	4	4
650	1500	3	3	9	9	9
700	1600	4	4	16	16	16
		$\Sigma u_i = 0$	$\Sigma v_i = 0$	$\Sigma u_i^2 = 60$	$\Sigma v_i^2 = 60$	$\Sigma u_i v_i = 60$

$$r_{uv} = \frac{9 \times 60 - 0 \times 0}{[9 \times 60 - (0)^2]^{1/2} [9 \times 60 - (0)^2]^{1/2}}$$

$$r_{uv} = \frac{9 \times 60 - 0}{(9 \times 60 - 0)^{1/2} (9 \times 60 - 0)^{1/2}}$$

$$= \frac{9 \times 60}{(9 \times 60 - 3600)^{1/2} [9 \times 60 - 3600]^{1/2}}$$

$$= \frac{540}{(540)^{1/2} (540)^{1/2}}$$

$$= \frac{540}{(540 - 3600)^{1/2} (540 - 3600)^{1/2}}$$

$$= \frac{540}{(-3060)^{1/2} (-3060)^{1/2}}$$

$$= \frac{540}{-3060}$$

$$= \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

$$= \frac{S_{xy}}{S_{xy}} = 1$$

$r = 1$ then the correlation between perfect and positive.

3) A programmer while writing a program for correlation co-efficient between 2 variables x & y from 30 pairs of observations obtained the following result $\Sigma x = 300$, $\Sigma x^2 = 3110$, $\Sigma y = 210$, $\Sigma y^2 = 2000$, $\Sigma xy = 2100$ at the time of checking it was found that he had copied down 2 pairs (x_i, y_i) as $(10, 20)$ and $(12, 10)$ instead of the correct value $(10, 15)$ and $(20, 15)$ Obtain the correct value of the correlation co-efficient.

$\Sigma x = 300$, $\Sigma x^2 = 3718$, $\Sigma y = 210$, $\Sigma y^2 = 2000$

$\Sigma xy = 2100$

(x_i, y_i) as $(18, 20)$ & $(12, 10)$ → wrong values

$(10, 15)$ & $(20, 15)$ → correct values

corrected

$$\Sigma x = 300 - 18 - 12 + 10 + 20$$

$$= 300$$

$$\Sigma x^2 = 3718 - 18^2 - 12^2 + 10^2 + 20^2$$

$$= 3750$$

$$\Sigma y = 210 - 20 - 10 + 15 + 15$$

$$= 210$$

$$\Sigma y^2 = 2000 - 20^2 - 10^2 + 15^2 + 15^2$$

$$= 1950$$

$$\Sigma xy = 2100 - (18 \times 20) - (12 \times 10) + (10 \times 15) + (20 \times 15)$$

$$= 2070$$

The corrected values are $\Sigma x = 300$ & $\Sigma x^2 = 3750$

$\Sigma y = 210$, $\Sigma y^2 = 1950$, $\Sigma xy = 2070$

Here $n = 30$

$$r_{xy} = \frac{n\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{[n\Sigma x_i^2 - (\Sigma x_i)^2]^{1/2} [n\Sigma y_i^2 - (\Sigma y_i)^2]^{1/2}}$$

$$= \frac{30 \times 2070 - 300 \times 210}{[30 \times 3750 - (300)^2]^{1/2} [30 \times 1950 - (210)^2]^{1/2}}$$

$$= \frac{-900}{(22500)^{1/2} (19600)^{1/2}}$$

$$= \frac{-900}{150 \times 120}$$

$$= \frac{-900}{18000}$$

$$= -1/20$$

$$= -0.05$$

1) If x and y are two variable prove that the correlation co-efficient between $ax+b$ & $cy+d$ is $r_{ax+b, cy+d}$

$$r_{ax+b, cy+d} = \frac{ac}{|ac|} r_{xy} \text{ if } ac \neq 0$$

Let $u_i = ax_i + b$, $v_i = cy_i + d$

$$\bar{u} = \frac{\sum u_i}{n}$$

$$\bar{u} = \frac{\sum (ax_i + b)}{n}$$

$$= \frac{\sum ax_i}{n} + \frac{\sum b}{n}$$

$$= \frac{a\sum x_i}{n} + \frac{nb}{n}$$

$$= a\frac{\sum x_i}{n} + b$$

$$\bar{u} = a\bar{x} + b$$

$$\bar{v} = c\bar{y} + d$$

$$r_{uv} = \frac{\sum (u_i - \bar{u})(v_i - \bar{v})}{n\sigma_u \sigma_v}$$

$$\sigma_u^2 = \frac{\sum (u_i - \bar{u})^2}{n}$$

$$= \frac{\sum [(ax_i + b) - (a\bar{x} + b)]^2}{n}$$

$$= \frac{\sum [ax_i - a\bar{x}]^2}{n}$$

$$= \frac{\sum [a(x_i - \bar{x})]^2}{n}$$

$$= \frac{a^2 \Sigma (x_i - \bar{x})^2}{n}$$

$$= a^2 \sigma_x^2$$

III^{ly} $\sigma_v^2 = c^2 \cdot \sigma_y^2$

$$\sigma_u^2 \sigma_v^2 = a^2 \sigma_x^2 \cdot c^2 \cdot \sigma_y^2$$

$$(\sigma_u \cdot \sigma_v)^2 = (ac)^2 (\sigma_x \cdot \sigma_y)^2$$

Taking square root

$$\sigma_u \sigma_v = |ac| \sigma_x \sigma_y$$

$$\gamma_{uv} = \frac{\Sigma (u_i - \bar{u})(v_i - \bar{v})}{n \sigma_u \sigma_v}$$

$$= \frac{\Sigma [(ax_i + b - a\bar{x} - b)(cy_i + d - c\bar{y} - d)]}{n |ac| \sigma_x \sigma_y}$$

$$= \frac{\Sigma [(ax_i - a\bar{x})(cy_i - c\bar{y})]}{n |ac| \sigma_x \sigma_y}$$

$$= \frac{ac \Sigma (x_i - \bar{x})(y_i - \bar{y})}{n |ac| \sigma_x \sigma_y}$$

$r_{uv} = \frac{ac}{|ac|} r(x,y)$

Hence proved

2) If x, y and z are uncorrelated variables each having same standard deviation obtain the correlation co-efficient between $x+y$ & $y+z$

x, y & z are uncorrelated variables

x and y are uncorrelated $\Rightarrow cov(x,y) = 0$

(ie) $\frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n} = 0$

$\Sigma(x_i - \bar{x})(y_i - \bar{y}) = 0$

y and z are uncorrelated $\Rightarrow cov(y,z) = 0$

$\Sigma(y_i - \bar{y})(x_i - \bar{z}) = 0$

z and x are uncorrelated $\Rightarrow cov(z,x) = 0$

$\Sigma(x_i - \bar{z})(x_i - \bar{x}) = 0$

Also given $\sigma_x = \sigma_y = \sigma_z = \sigma$

To find correlation co-efficient between

$x+y$ & $y+z$

Let $u_i = x_i + y_i$, $v_i = y_i + z_i$

$\bar{u} = \bar{x} + \bar{y}$ $\quad \bar{v} = \bar{y} + \bar{z}$

$$r_{uv} = \frac{\Sigma (u_i - \bar{u})(v_i - \bar{v})}{n \sigma_u \sigma_v} \rightarrow (1)$$

$$\Sigma (u_i - \bar{u})(v_i - \bar{v}) = \Sigma \left\{ [(x_i + y_i) - (\bar{x} + \bar{y})][(y_i + z_i) - (\bar{y} + \bar{z})] \right\}$$

$$= \Sigma \left[(x_i + y_i - \bar{x} - \bar{y})[(y_i + z_i - \bar{y} - \bar{z})] \right]$$

$$= \Sigma \left[(x_i - \bar{x}) + (y_i - \bar{y}) \right] \left[(y_i - \bar{y}) + (z_i - \bar{z}) \right]$$

$$= \Sigma \left[(x_i - \bar{x})(y_i - \bar{y}) + (x_i - \bar{x})(z_i - \bar{z}) + (y_i - \bar{y}) (y_i - \bar{y}) + (y_i - \bar{y})(z_i - \bar{z}) \right]$$

$$= \Sigma (x_i - \bar{x})(y_i - \bar{y}) + \Sigma (x_i - \bar{x})(z_i - \bar{z}) + \Sigma (y_i - \bar{y})^2 + \Sigma (y_i - \bar{y})(z_i - \bar{z})$$

$$= 0 + 0 + \Sigma (y_i - \bar{y})^2 + 0$$

$= \Sigma (y_i - \bar{y})^2$ $\left[\sigma_y^2 = \frac{\Sigma (y_i - \bar{y})^2}{n} \right.$

$= n\sigma_y^2$ $\Sigma (y_i - \bar{y})^2 = \sigma_y^2 n$

$\Sigma (u_i - \bar{u})(u_i - \bar{u}) = n\sigma_y^2$

$= n\sigma^2$

$\sigma_u^2 = \frac{\Sigma (u_i - \bar{u})^2}{n}$

$= \frac{\Sigma [(x_i + y_i) - (\bar{x} + \bar{y})]^2}{n}$

$= \frac{\Sigma (x_i + y_i - \bar{x} - \bar{y})^2}{n}$

$= \frac{\Sigma [(x_i - \bar{x}) + (y_i - \bar{y})]^2}{n}$

$= \frac{\Sigma [(x_i - \bar{x})^2 + (y_i - \bar{y})^2 + 2(x_i - \bar{x})(y_i - \bar{y})]}{n}$

$= \frac{\Sigma (x_i - \bar{x})^2}{n} + \frac{\Sigma (y_i - \bar{y})^2}{n} + \frac{2\Sigma (x_i - \bar{x})(y_i - \bar{y})}{n}$

$\sigma_u^2 = \sigma_x^2 + \sigma_y^2$

$= \sigma^2 + \sigma^2 = 2\sigma^2$

$\sigma_u = \sqrt{2}\sigma$

III^dly $\sigma_v = \sqrt{2}\sigma$

$$\textcircled{1} \Rightarrow \gamma_{uv} = \frac{t\sigma^2}{t\sqrt{2}\sigma \cdot \sqrt{2}\sigma}$$

$$= \frac{1}{2}$$

1) Show that the variables $u = x\cos\alpha + y\sin\alpha$ and $v = y\cos\alpha - x\sin\alpha$ are uncorrelated if $\alpha = \frac{1}{2}\tan^{-1}\left(\frac{2\gamma_{xy}\sigma_x\sigma_y}{\sigma_x^2 - \sigma_y^2}\right)$

Let $u_i = x_i\cos\alpha + y_i\sin\alpha$

$v_i = y_i\cos\alpha - x_i\sin\alpha$

$\bar{u} = \frac{\Sigma(x_i\cos\alpha + y_i\sin\alpha)}{n}$

$= \bar{x}\cos\alpha + \bar{y}\sin\alpha$

Also $\bar{v} = \bar{y}\cos\alpha - \bar{x}\sin\alpha$

$u_i - \bar{u} = (x_i\cos\alpha + y_i\sin\alpha) - (\bar{x}\cos\alpha + \bar{y}\sin\alpha)$

$= (x_i - \bar{x})\cos\alpha + (y_i - \bar{y})\sin\alpha$

$v_i - \bar{v} = (y_i - \bar{y})\cos\alpha - (x_i - \bar{x})\sin\alpha$

u & v are uncorrelated

(i) $r_{uv} = 0$

(ii) $\frac{\Sigma (u_i - \bar{u})(v_i - \bar{v})}{n} = 0$

(iii) $\Sigma (u_i - \bar{u})(v_i - \bar{v}) = 0$

$$\Sigma \left[[(x_i - \bar{x})\cos\alpha + (y_i - \bar{y})\sin\alpha] [(y_i - \bar{y})\cos\alpha - (x_i - \bar{x})\sin\alpha] \right] = 0$$

$$\Sigma \left[(x_i - \bar{x})(y_i - \bar{y})\cos^2\alpha - (x_i - \bar{x})^2 \sin\alpha\cos\alpha + (y_i - \bar{y})^2 \sin\alpha\cos\alpha - (y_i - \bar{y})(x_i - \bar{x})\sin^2\alpha \right] = 0$$

$$\Sigma \left[(x_i - \bar{x})(y_i - \bar{y}) \right] (\cos^2\alpha - \sin^2\alpha) - \left[(x_i - \bar{x})^2 - (y_i - \bar{y})^2 \right] \sin\alpha\cos\alpha = 0$$

$$(\cos^2\alpha - \sin^2\alpha) \Sigma (x_i - \bar{x})(y_i - \bar{y}) - \sin\alpha\cos\alpha \Sigma \left[(x_i - \bar{x})^2 - (y_i - \bar{y})^2 \right] = 0$$

$$\cos 2\alpha \cdot n r_{xy} \sigma_x \sigma_y - \frac{2\sin\alpha\cos\alpha}{2} \left[\Sigma (x_i - \bar{x})^2 - \Sigma (y_i - \bar{y})^2 \right] = 0$$

$$\cos 2\alpha \cdot n r_{xy} \sigma_x \sigma_y - \frac{\sin 2\alpha}{2} \left[n\sigma_x^2 - n\sigma_y^2 \right] = 0$$

$$\cos 2\alpha \cdot n r_{xy} \cdot \sigma_x \sigma_y = \frac{1}{2} \sin 2\alpha \left[n(\sigma_x^2 - \sigma_y^2) \right]$$

$$2\cos 2\alpha \cdot n \cdot r_{xy} \sigma_x \sigma_y = \sin 2\alpha \, n (\sigma_x^2 - \sigma_y^2)$$

$$\frac{2 r_{xy} \sigma_x \sigma_y}{\sigma_x^2 - \sigma_y^2} = \frac{\sin 2\alpha}{\cos 2\alpha}$$

$$\frac{2 r_{xy} \sigma_x \sigma_y}{\sigma_x^2 - \sigma_y^2} = \tan 2\alpha$$

$$2\alpha = \tan^{-1}\left(\frac{2 r_{xy} \sigma_x \sigma_y}{\sigma_x^2 - \sigma_y^2}\right)$$

$$\alpha = \frac{1}{2}\tan^{-1}\left(\frac{2 r_{xy} \sigma_x \sigma_y}{\sigma_x^2 - \sigma_y^2}\right)$$

Hence proved.

2) Show that if x', y' are the deviation of the random variable X, Y from the respective mean. Then

(i) $r = 1 - \frac{1}{2N}\sum\left(\frac{x_i'}{\sigma_x} - \frac{y_i'}{\sigma_y}\right)^2$ and

(ii) $r = -1 + \frac{1}{2N}\sum\left(\frac{x_i'}{\sigma_x} + \frac{y_i'}{\sigma_y}\right)^2$

(iii) Deduce that $-1 \le r \le 1$

Sol:-

x' & y' are the deviation of the random variable x & y

$\therefore x' = x_i - \bar{x}$, $y_i = y_i - \bar{y}$

$$RHS = 1 - \frac{1}{2N} \sum \left(\frac{x'}{\sigma_x} - \frac{y'}{\sigma_y} \right)^2$$

$x' = x_i - \bar{x}$
$y = y_i - \bar{y}$

$$= 1 - \frac{1}{2N} \sum \left[\left(\frac{x'}{\sigma_x} \right)^2 - \frac{2 x' y'}{\sigma_x \sigma_y} + \left(\frac{y'}{\sigma_y} \right)^2 \right]$$

$$= 1 - \frac{1}{2N} \left[\frac{\sum (x')^2}{\sigma_x^2} - \frac{2 \sum x' y'}{\sigma_x \sigma_y} + \frac{\sum (y')^2}{\sigma_y^2} \right]$$

$$= 1 - \frac{1}{2N} \left[\frac{\sum (x_i - \bar{x})^2}{\sigma_x^2} - \frac{2 \sum (x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y} + \frac{\sum (y_i - \bar{y})^2}{\sigma_y^2} \right]$$

$$= 1 - \frac{1}{2N} \left[\frac{N \sigma_x^2}{\sigma_x^2} - \frac{2N r \sigma_x \sigma_y}{\sigma_x \sigma_y} + N \frac{\sigma_y^2}{\sigma_y^2} \right]$$

$$= 1 - \frac{1}{2N} \left[2N - 2Nr \right]$$

$= 1 - \frac{1}{2N} \cdot 2N(1-r)$

$= r - 1 + r$

$= r$

$= LHS$

(ii) $RHS \Rightarrow -1 + \frac{1}{2N} \sum \left[\frac{x_i'}{\sigma x} + \frac{y_i'}{\sigma y} \right]^2$

$= -1 + \frac{1}{2N} \sum \left[\frac{(x_i')^2}{\sigma x^2} + \frac{(y_i')^2}{\sigma y^2} + \frac{2 x_i' y_i'}{\sigma x \sigma y} \right]$ (iii)

$= -1 + \frac{1}{2N} \left\{ \right.$

$= -1 + \frac{1}{2N} \left[\frac{\sum (x_i')^2}{\sigma x^2} + \frac{\sum (y_i')^2}{\sigma y^2} + \frac{\sum 2 x_i' y_i'}{\sigma x \sigma y} \right]$

$= -1 + \frac{1}{2N} \left[\frac{\sum (x_i - \bar{x})^2}{\sigma x^2} + \frac{\sum (y_i - \bar{y})^2}{\sigma y^2} + \frac{2 \sum (x_i - \bar{x})(y_i - \bar{y})}{\sigma x \sigma y} \right]$

$= -1 + \frac{1}{2N} \left[\frac{N \sigma x^2}{\sigma x^2} + \frac{N \sigma y^2}{\sigma y^2} + 2N r \right]$

$= -1 + \frac{1}{2N} \left[N + N + 2Nr \right]$

$$= -1 + \frac{1}{2N}\left[2N + 2Nr\right]$$

$$= -1 + \frac{1}{2N} \cdot 2N\left[1 + r\right]$$

$$= -1 + 1 + r$$

$$= r$$

Hence = LHS

(iii) From (1)

$$1 - \frac{1}{2N}\sum\left(\frac{x_i'}{\sigma x} - \frac{y_i'}{\sigma y}\right)^2 \leq 1$$

$$r \leq 1 \rightarrow (a)$$

From (2)

$$-1 + \frac{1}{2N}\sum\left(\frac{x_i'}{\sigma x} + \frac{y_i'}{\sigma y}\right)^2 \geq -1$$

$$r \geq -1$$

$$\text{i.e., } -1 \leq r \rightarrow (b)$$

From (a) & (b) we get

$$-1 \leq r \leq 1$$

Hence proved

1) Let x, y be two variable with standard deviation σ_x & σ_y respectively

if $u = x + ky$, $v = x + \left(\frac{\sigma_x}{\sigma_y}\right) y$ & $\gamma_{uv} = 0$

then find the value of k.

Soln:

$u_i = x_i + ky_i$ $\qquad$ $v_i = x_i + \left(\frac{\sigma_x}{\sigma_y}\right) y_i$

$\bar{u} = \bar{x} + k\bar{y}$ $\qquad$ $\bar{v} = \bar{x} + \left(\frac{\sigma_x}{\sigma_y}\right) \bar{y}$

$u_i - \bar{u} = x_i + ky_i - \bar{x} - k\bar{y}$

$= (x_i - \bar{x}) + k(y_i - \bar{y})$

$v_i - \bar{v} = x_i + \left(\frac{\sigma_x}{\sigma_y}\right) y_i - \bar{x} - \left(\frac{\sigma_x}{\sigma_y}\right) \bar{y}$

$= (x_i - \bar{x}) + \left(\frac{\sigma_x}{\sigma_y}\right)(y_i - \bar{y})$

$\gamma_{uv} = 0$

$\text{cov}(u,v) = 0$

$\Sigma [(u_i - \bar{u})(v_i - \bar{v})] = 0$

$\Sigma \left[[(x_i - \bar{x}) + k(y_i - \bar{y})] \left[(x_i - \bar{x}) + \left(\frac{\sigma_x}{\sigma_y}\right)(y_i - \bar{y}) \right] \right] = 0$

$$\Sigma\left[(x_i-\bar{x})^2+\left(\frac{\sigma x}{\sigma y}\right)(x_i-\bar{x})(y_i-\bar{y})+k(y_i-\bar{y})(x_i-\bar{x})+k\left(\frac{\sigma x}{\sigma y}\right)(y_i-\bar{y})^2\right]=0$$

$$\left(\Sigma(x_i-\bar{x})^2+\left(\frac{\sigma x}{\sigma y}\right)\Sigma(x_i-\bar{x})(y_i-\bar{y})+k\Sigma(y_i-\bar{y})(x_i-\bar{x})+k\left(\frac{\sigma x}{\sigma y}\right)\Sigma(y_i-\bar{y})^2\right]=0$$

$$n\sigma x^2+\left(\frac{\sigma x}{\sigma y}\right)n\sigma x\sigma y\,r_{xy}+k\,n\sigma x\sigma y\,r_{xy}+k\left(\frac{\sigma x}{\sigma y}\right)n\sigma y^2=0$$

$$n\sigma x^2+n\sigma x^2 r_{xy}+k\,n\sigma x\sigma y\,r_{xy}+k\sigma x\sigma y=0$$

$$n\sigma x\left[\sigma x+\sigma x\,r_{xy}+k\sigma y\,r_{xy}+k\sigma y\right]=0$$

$$\sigma x\left[\sigma x+\sigma x\,r_{xy}+k\sigma y\,r_{xy}+k\sigma y\right]=0$$

$$\sigma x\left[\sigma x(1+r_{xy})+k\sigma y(1+r_{xy})\right]=0$$

$$\sigma x\,(1+r_{xy})*(\sigma x+k\sigma y)=0$$

$\sigma x=0$ (or) $1+r_{xy}=0$ (or) $\sigma x+k\sigma y=0$

$$\sigma x+k\sigma y=0$$

$$k\sigma y=-\sigma x$$

$$k=-\left(\frac{\sigma x}{\sigma y}\right)$$

If $r_{xy}\neq -1$, $\sigma x\neq 0$ we get $k=-\left(\frac{\sigma x}{\sigma y}\right)$

2nd internal.

Rank correlation:

Let (x_i, y_i) be the ranks of the i^{th} individual in the first & II ranking respectively in the coefficient of correlation between the rank (x_i, y_i) are called the rank correlation co efficient is denoted by ρ (rho)

Theorem:

P.T the rank correlation co.efficient ρ is $1 - \frac{6\Sigma(x-y)^2}{n(n^2-1)}$

consider the collection of n individuals Let x_i and y_i be the ranks of the i^{th} individuals

$$\bar{x} = \frac{\Sigma x_i}{n}$$

$$= \frac{1+2+\cdots+n}{n}$$

$$= \frac{n(n+1)}{2n}$$

$$= \frac{n+1}{2}$$

$$\sigma x^2 = \frac{\sum xi^2}{n} - \left(\frac{\sum xi}{n}\right)^2$$

$$\sum xi^2 = 1^2 + 2^2 + 3^2 + \cdots + n^2$$

$$= \frac{n(n+1)(2n+1)}{6}$$

$$\sigma x^2 = \frac{n(n+1)(2n+1)}{6n} - \left[\frac{n+1}{2}\right]^2$$

$$= \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4}$$

$$= \frac{2(n+1)(2n+1) - 3(n+1)^2}{12}$$

$$= \frac{(n+1)\left[2(2n+1) - 3(n+1)\right]}{12}$$

$$= \frac{(n+1)\left[4n+2-3n-3\right]}{12}$$

$$= (n+1)\frac{(n-1)}{12}$$

$$\sigma x^2 = \frac{n^2-1}{12}$$

$$\bar{x} = \bar{y} = \frac{n+1}{2}$$

$$\sigma x^2 = \sigma y^2 = \frac{n^2-1}{12}$$

Now $\Sigma (x-y)^2 = \Sigma (x-\bar{x}+\bar{y}-y)^2$ $[\because \bar{x}=\bar{y}]$

$= \Sigma [(x-\bar{x})-(y-\bar{y})]^2$

$= \Sigma [(x-\bar{x})^2 - 2(x-\bar{x})(y-\bar{y}) + (y-\bar{y})^2]$

$= \Sigma (x-\bar{x})^2 - 2\Sigma (x-\bar{x})(y-\bar{y}) + \Sigma (y-\bar{y})^2$

$= n\sigma x^2 - 2n\beta \sigma x \sigma y + n\sigma y^2$

$[\because \sigma x^2 = \sigma y^2 = \sigma^2]$

$= n\sigma^2 - 2n\beta\sigma^2 + n\sigma^2$

$= 2n\sigma^2 - 2n\beta\sigma^2$

$\Sigma (x-y)^2 = 2n\sigma^2 (1-\beta)$

$1-\beta = \frac{\Sigma (x-y)^2}{2n\sigma^2}$

$$\rho = 1 - \frac{\Sigma(x-y)^2}{2n\sigma^2}$$

$$= 1 - \frac{\Sigma(x-y)^2}{2n \cdot \frac{n^2-1}{12}}$$

$$= 1 - \frac{6\Sigma(x-y)^2}{n(n^2-1)}$$

This is called the spearman's ρ or formula for the rank correlation

1) Find the rank correlation coefficient between the height in cm and weight in kg of 6 soldiers in Indian Army.

Height (in cm)	165	167	166	170	169	172
Weight (in kg)	61	60	63.5	63	61.5	64

Height (in cm)	Rank for height (x)	weight (in kg)	Rank for weight (y)	$x-y$	$(x-y)^2$
165	6	61	5	1	1
167	4	60	6	-2	4
166	5	63.5	2	3	9
170	2	63	3	-1	1
169	3	61.5	4	-1	1
172	1	64	1	0	0

Here, $n = 6$

$$\rho = 1 - \frac{6\Sigma(x-y)^2}{n(n^2-1)}$$

$$= 1 - \frac{6\times 16}{6(36-1)}$$

$$= 1 - \frac{96}{6\times 35}$$

$$= 1 - \frac{96}{210}$$

$$\rho = 0.542 \text{ // Ans}$$

Note:

If two (or) more individuals get the same Rank in the ranking process we assign the common rank to the repeated values. This common rank is the average of the ranks, and the next item will get the rank next to the rank already assumed. As a result of this in the formula for the $\Sigma(x-y)^2$. where m is the number of times an item has repeated values. We add the factor $\frac{m(m^2-1)}{12}$ to This correlation factor added for each repeated rank of the variables (x, y).

1) From the following for data in marks obtaining to the ten B students in Physics and chemistry.

Calculate the rank correlation.

Physics	35	56	50	65	44	38	44	50	15	26
Chemistry	50	35	70	25	35	58	75	60	55	35

Physics	Rank(in Physics) (x)	Chemistry	Rank in chemistry (y)	x-y	$(x-y)^2$
35	8	50	6	2	4
56	2	35	8	-6	36
50	3.5	70	2	1.5	2.25
65	1	25	10	-9	81
44	5.5	35	8	-2.5	6.25
38	7	58	4	3	9
44	5.5	75	1	4.5	20.25
50	3.5	60	3	0.5	0.25
15	10	55	5	5	25
26	9	35	9	1	1

$\Sigma(x-y)^2 = 185$

$n = 10$

Here the marks 50 and 44 occur in x and 35 occurs thrice in y

Highest repeated twice

$$\Sigma(x-y)^2 = \Sigma(x-y)^2 + \frac{m(m^2-1)}{12}$$

Hence corrected $\Sigma(x-y)^2$ = actual

$$\Sigma(x-y)^2 = \text{actual } \Sigma(x-y)^2 + \frac{m(m^2-1)}{12}$$

$$\Sigma(x-y)^2 = 185 + \frac{2(2^2-1)}{12} + \frac{2(2^2-1)}{12} + \frac{3(3^2-1)}{12}$$

$$= 185 + \frac{2(3)}{12} + \frac{2(3)}{12} + \frac{3(8)}{12}$$

$$= 185 + \frac{6}{12} + \frac{6}{12} + \frac{24}{12}$$

$$= 185 + \frac{1}{2} + \frac{1}{2} + 2$$

$$= 185 + 3$$

$$= 188$$

$$\rho = 1 - \frac{6\Sigma(x-y)^2}{n(n^2-1)}$$

$$= 1 - \frac{6 \times 188}{10(100-1)}$$

$$= 1 - \frac{1128}{10(99)}$$

$$= 1 - \frac{1129}{990}$$

$$= \frac{990 - 1128}{990}$$

$$= -0.139 \text{ Ans.}$$

2) Calculate the rank correlation co-efficient for the following data

x	20	25	60	45	80	25	55	65	25	75
y	52	50	55	50	60	70	72	78	80	65

x	Rank x	y	Rank y	x − y	(x − y)²
20	10	52	8	2	4
25	8	50	9.5	−1.5	2.25
60	4	55	7	−3	9
45	6	50	9.5	−3.5	12.25

80	1	60	6	−5	25
25	8	70	4	4	16
55	5	72	3	2	4
65	3	78	2	1	1
25	8	80	1	7	49
75	2	63	5	−3	9

$\Sigma(x-y)^2 = 131.5$

$n = 10$

Here the ~~numbers~~ 25 occur thrice in x and 80 occur twice in y.

Hence corrected $\Sigma(x-y)^2$ = actual $\Sigma(x-y)^2 + \frac{m(m^2-1)}{12}$

$= 131.5 + \frac{3(3^2-1)}{12} + \frac{2(2^2-1)}{12}$

$= 131.5 + \frac{3(9-1)}{12} + \frac{2(4-1)}{12}$

$= 131.5 + \frac{3(8)}{12} + \frac{2(3)}{12}$

$= 131.5 + \frac{24}{12} + \frac{6}{12}$

$= 131.5 + 2 + \frac{1}{2}$

$= \frac{131.5 \times 2 + 2 \times 2 + 1}{2}$

$= 134$

$$\rho = 1 - \frac{6\Sigma(x-y)^2}{n(n^2-1)}$$

$$= 1 - \frac{6 \times 136}{10 \times (100-1)}$$

$$= 1 - \frac{804}{10 \times 99}$$

$$= 1 - \frac{804}{990}$$

$$= \frac{990 - 804}{990}$$

$$= 0.1878$$

x	
1	
2	
4	
3	
7	
6	
5	
8	

) Three Judges Asign the raks to 8 entries in a beauty contest

Judge Mr. X	1	2	4	3	7	6	5	8
Judge Mr. Y	3	2	1	5	4	7	6	8
Judge Mr. Z	1	2	5	4	5	7	8	6

Which pair of judges has a nearest approach to common taste in beauty.

We have to find

ρ_{xy}, ρ_{yz}, ρ_{zx}

x	y	z	x−y	$(x-y)^2$	y−z	$(y-z)^2$	z−x	$(z-x)^2$
1	3	1	−2	4	2	4	0	0
2	2	2	0	0	0	0	0	0
4	1	3	3	9	−2	4	−1	1
5	5	4	−2	4	1	1	1	1
7	4	5	3	9	−1	1	−2	4
6	7	7	−1	1	0	0	1	1
5	6	8	−1	1	−2	4	3	9
8	8	6	0	0	2	4	−2	4
				$\Sigma(x-y)^2$ = 28		$\Sigma(y-z)^2$ = 18		$\Sigma(z-x)^2$ = 20

$n = 8$

$$\rho_{xy} = 1 - \frac{6\Sigma(x-y)^2}{n(n^2-1)}$$

$$= 1 - \frac{6 \times 28}{8 \times (64-1)}$$

$$= 1 - \frac{6 \times 28}{8 \times 63}$$

$$= 1 - \frac{168}{504} = \frac{504 - 168}{504}$$

$$= 0.6666\ldots$$

$$\rho_{(yz)} = 1 - \frac{6\sum(y-z)^2}{n(n^2-1)}$$

$$= 1 - \frac{6(18)}{8(64-1)}$$

$$= 1 - \frac{6 \times 18}{8(63)}$$

$$= 1 - \frac{108}{504}$$

$$= \frac{504 - 108}{504}$$

$$= 0.7857$$

$$\rho_{(zx)} = 1 - \frac{6\sum(z-x)^2}{n(n^2-1)}$$

$$= 1 - \frac{6(20)}{8 \times 63}$$

$$= 1 - \frac{120}{504} = \frac{504 - 120}{504}$$

$$= 0.7619$$

Since $\rho_{yz} > \rho_{xy}$ and ρ_{zx}

Hence the judges Mr.y and Mr. z have nearest approach to common taste in beauty

2) The co efficient of Rank correlation of marks uptained by 10 students in Maths and physics pass found to be 0.8. It was latest discovered that the difference in ranks in two subjects uptained by one of the student was wrongly taken as 5 instead of 8. Find the correct co-efficient of rank correlation.

Soln:

Here $n = 10$

$\rho = 0.8$

$$0.8 = 1 - \frac{6\Sigma(x-y)^2}{10(10^2-1)}$$

$$0.8 = 1 - \frac{6\Sigma(x-y)^2}{990}$$

$$\frac{6\Sigma(x-y)^2}{990} = 1 - 0.8$$

$$\frac{6\Sigma(x-y)^2}{990} = 0.2$$

$$6\Sigma(x-y)^2 = 0.2 \times 990$$

$$6\Sigma(x-y)^2 = 198$$

$$\Sigma(x-y)^2 = \frac{198}{6}$$

$$\Sigma(x-y)^2 = 33$$

Corrected $\Sigma(x-y)^2 = 33 - 5^2 + 8^2$

$= 33 - 25 + 64$

$= 72$

$$\rho = 1 - \frac{6 \times 72}{10(10^2-1)}$$

$$= 1 - \frac{432}{990}$$

$$= \frac{990 - 432}{990}$$

$= 0.564$

5. Let $(x_1, x_2, \ldots x_n)$ be the ranks of n individuals according to the character of A and $y_1, y_2 \ldots y_n$ be the ranks of same individuals according to another character B. It is given that $x_i + y_i = 1+n$ for $i = (1, 2 \ldots n)$. Show that the values of the rank correlation coefficient ρ between the character A & B is (-1)

Given:

$$x_i + y_i = 1+n \longrightarrow (1)$$

Let d_i be the difference between the two ranks x_i & y_i for $i = 1, 2, \ldots n$

$$d_i = x_i - y_i \longrightarrow (2)$$

$$(1) - (2) \Rightarrow x_i + y_i - d_i = (1+n) - (x_i - y_i)$$

$$x_i + y_i - (x_i - y_i) = (1+n) - d_i$$

$$x_i + y_i - x_i + y_i = (1+n) - d_i$$

$$2y_i = (1+n) - d_i$$

$$d_i = (1+n) - 2y_i$$

Rank correlation $\rho = 1 - \frac{6\Sigma(x_i - y_i)^2}{n(n^2-1)} \rightarrow (3)$

$= 1 - \frac{6\Sigma d_i^2}{n(n^2-1)}$

$\Sigma d_i^2 = \Sigma [(1+n) - 2y_i]^2$

$= \Sigma [(1+n)^2 - 2(1+n)2y_i + (2y_i)^2]$

$= \Sigma [(1+n)^2 - 4(n+1)y_i + 4y_i^2]$

$= \Sigma (1+n)^2 - 4(n+1)\Sigma y_i + 4\Sigma y_i^2$

$= n(n+1)^2 - 4(n+1)\Sigma y_i + 4\Sigma y_i^2$

now $\Sigma y_i = 1+2+\cdots+n = \frac{n(n+1)}{2}$

$\Sigma y_i^2 = 1^2+2^2+\cdots+n^2 = \frac{n(n+1)(2n+1)}{6}$

$\Sigma d_i^2 = n(n+1)^2 - 4(n+1)\frac{(n+1)n}{2} + 4\frac{n(n+1)(2n+1)}{6}$

$= n(n+1)\left[(n+1) - 2(n+1) + 2\frac{(2n+1)}{3}\right]$

If the c
moves
explained
0.e. 96

$$= n(n+1)\left[\frac{3n+3-6n-6+4n+2}{3}\right]$$

$$= n(n+1)\left[\frac{n-1}{3}\right]$$

$$= \frac{n(n^2-1)}{3}$$

(3) $\Rightarrow$ $\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$

$$= 1 - \frac{6 \times \frac{n(n^2-1)}{3}}{n(n^2-1)}$$

$$= 1 - 2 \times \frac{n(n^2-1)}{n(n^2-1)}$$

$$= 1 - 2$$

$$\rho = -1$$

Hence proved.

The co-efficient of Rank correlation between marks in Statistics and Mathematics obtained by a certain group of student is -1. i.e. If the sum of the squares of the

difference in ranks is given to be 33. Find the number of students in a group.

$\rho = 0.8$

$\Sigma(x-y)^2 = 33$

$n = ?$

We have $\rho = 1 - \frac{6\Sigma(x-y)^2}{n(n^2-1)}$

$0.8 = 1 - \frac{6 \times 33}{n(n^2-1)}$

$0.8 = 1 - \frac{198}{n(n^2-1)}$

$\frac{198}{n(n^2-1)} = 1 - 0.8$

$\frac{198}{n(n^2-1)} = 0.2$

$n(n^2-1) = \frac{198}{0.2}$

$n(n^2-1) = 990$

$n(n^2-1) = 10(10^2-1)$

Here $n(n^2-1)$ is of the form $10(10^2-1)$

$$\boxed{n=10}$$

2) The co-efficient of rank correlation of the marks in ~~Statistics~~ optained by 10 students in physics and chemistry pass to be 0.5 it was latter discovered that the differenks in ranks the two subjects ~~was~~ optained by one of the students for strongly taken as 8 instead of 7. Find the correct co-efficient of rank correlation.

$$n=10 \quad , \quad \rho = 0.5$$

we have $\rho = 1 - \frac{6\Sigma(x-y)^2}{n(n^2-1)}$

$$0.5 = 1 - \frac{6\Sigma(x-y)^2}{10(10^2-1)}$$

$$\frac{6\Sigma(x-y)^2}{10(100-1)} = 1 - 0.5$$

$$\frac{6\Sigma(x-y)^2}{10 \times 99} = 0.5$$

$$\Sigma(x-y)^2 = \frac{0.5 \times 990}{6}$$

$$\Sigma(x-y)^2 = 82.5$$

corrected $\Sigma(x-y)^2 = 82.5 - 3^2 + 7^2$

$= 82.5 - 9 + 49$

$= 122.5$

correct rank correlation

$$\rho = 1 - \frac{6\Sigma(x-y)^2}{n(n^2-1)}$$

$$= 1 - \frac{6 \times 122.5}{990}$$

$= 1 - 0.742$

$= 0.258$ /.

3) Following are the marks obtained by 10 student in first semester in ancillary papers out of 75

Semester I (Ancillary I)	60	55	75	45	69	45	72	39	35	45

Semester II 70 58 73 49 60 49 60 55 60 CA
(Ancillary II)

Semester III 55 61 68 40 58 60 50 88 80 60
(Ancillary III)

~~Semester Rank Semester Rank Semester Rank~~
III

I	Rank x	II	Rank y	III	Rank z	x-y	y-z	x-z	$(x-y)^2$	$(y-z)^2$	$(x-z)^2$
70	4	70	2	55	6	2	-4	-2	4	16	4
55	5	58	6	61	2	-1	4	3	1	16	9
73	1	73	1	68	1	0	0	0	0	0	0
45	7	49	8.5	40	9	-1.5	-0.5	-2	2.25	0.25	4
64	3	60	4	58	5	-1	-1	-2	1	1	4
45	7	49	8.5	60	3.5	-1.5	5	3.5	2.25	25	12.25
72	2	60	4	60	7.5	-2	-3.5	5.5	4	12.25	30.25
49	9	55	7	38	10	2	-3	-1	4	9	1
55	10	60	4	50	7.5	6	-3.5	2.5	36	12.25	6.25
49	7	48	10	60	3.5	-3	6.5	3.5	9	42.25	12.25

$$\rho_{xy} = 1 - \frac{6\Sigma(x-y)^2}{n(n^2-1)}$$

$$= 1 - \frac{6 \times 63.5}{990}$$

$$= 1 - 0.385$$

$= 0.615$

$$\rho_{yz} = 1 - \frac{6\Sigma(y-z)^2}{n(n^2-1)}$$

$$= 1 - \frac{6 \times 134}{990}$$

$$= 1 - 0.818$$

$$= 0.182$$

$$\rho_{zx} = 1 - \frac{6\Sigma(z-x)^2}{n(n^2-1)}$$

$$= 1 - \frac{6 \times 83}{990}$$

$$= 1 - 0.503$$

$$= 0.497$$

4) A computer while calculating the correlation co-efficient between two variables x and y obtained the following constants $n = 25$, $\Sigma x = 125$, $\Sigma x^2 = 650$,

$\Sigma y = 100$, $\Sigma y^2 = 460$ & $\Sigma xy = 508$. It was later discovered at the time of checking that he had copied down 2 pairs of observations (x_i, y_i) as (6,14) & (8,6) instead of the correct values (8,12) & (6,8) Obtained the correct value of the correlation co-efficient between (x & y)

Soln:

$\Sigma x = 125$, $\Sigma x^2 = 650$

$\Sigma y = 100$, $\Sigma y^2 = 460$

$\Sigma xy = 508$

(x_i, y_i) as (6,14) & (8,6) - wrong

(8,12) & (6,8) - correct

corrected

$\Sigma x = 125 - 6 - 8 + 8 + 6$

$= 125$

$\Sigma x^2 = 650 - 6^2 - 8^2 + 8^2 + 6^2$

$= 650$

$\Sigma y = 100 - 14 - 6 + 12 + 8$

$= 100$

$\Sigma y^2 = 1460 - 14^2 - 6^2 + 12^2 + 8^2$

$= 460$

The corrected values are

$\Sigma x = 125$, $\Sigma x^2 = 650$, $\Sigma y = 100$, $\Sigma y^2 = 460$

$\Sigma xy = 520$

$\Sigma xy = 508 - (6 \times 8) - (8 \times 6) + (8 \times 12) + (6 \times 8)$

$= 520$

Here $n = 25$

$$r_{xy} = \frac{n\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{\left[n\Sigma x_i^2 - (\Sigma x_i)^2\right]^{1/2}\left[n\Sigma y_i^2 - (\Sigma y_i)^2\right]^{1/2}}$$

$$= \frac{25 \times 520 - 125 \times 100}{\left[(25 \times 650) - (125)^2\right]^{1/2}\left[(25 \times 460) - (100)^2\right]^{1/2}}$$

$$= \frac{13900}{(625)^{1/2}} \quad \frac{12500}{} \quad 2.475$$

$$= \frac{13000 - 12500}{[(16250)^{1/2} - (125)^{1/2}][(11500)^{1/2} - (100)^{1/2}]}$$

$$= \frac{13000 - 12500}{(625)^{1/2}\,(900)^{1/2}}$$

$$= \frac{500}{25 \times 30}$$

$$= \frac{500}{750}$$

$$\gamma_{xy} = 0.66$$

1) The following table shows how 10 students were ranked according to their achievements in the laboratory and lecture positions of biology course. Find the coefficients of rank correlation.

Laboratory	8	3	9	2	7	10	4	6	1	5
Lecture	9	5	10	1	8	7	3	4	2	6

Laboratory (x)	Rank	Lecture (y)	Rank	$x-y$	$(x-y)^2$
8	8	9		−1	1
3	8	5		−2	4
9	2	10		−1	1
2	9	1		1	1
7	4	8		−1	1
10	1	7		3	9
4	7	3		1	1
6	5	4		2	4
1	10	2		−1	1
5	6	6		−1	1

$\Sigma(x-y)^2 = 24$

$$\rho = 1 - \frac{6\Sigma(x-y)^2}{n(n^2-1)}$$

$$= 1 - \frac{6 \times 24}{10(10^2-1)}$$

$$= 1 - \frac{6 \times 24}{10(100-1)}$$

$$= 1 - \frac{6 \times 24}{10(99)}$$

$$= 1 - \frac{144}{990}$$

$$= \frac{990 - 144}{990}$$

$$= 0.855$$

2) 10 Students got the following % of marks in 2 subjects Economics & Statistics

Economics	78	65	36	98	25	75	82	90	62	35
Statistics	84	53	51	91	60	68	62	86	58	35

Calculate the rank correlation co-efficient

Economics	Rank in x	Statistics	Rank in y	x-y	$(x-y)^2$
78	4	84	3	1	1
65	6	53	8	-2	4
36	9	51	9	0	0

98	1	91	1	0	0
25	10	60	6	4	16
75	5	68	4	1	1
82	3	62	5	-2	4
90	2	86	2	0	0
62	7	58	7	0	0
39	8	47	10	-2	4

$\Sigma(x-y)^2 = 30$

$$\rho = 1 - \frac{6\Sigma(x-y)^2}{n(n^2-1)}$$

$$= 1 - \frac{6 \times 30}{10(10^2-1)}$$

$$= 1 - \frac{6 \times 30}{10(99)}$$

$$= 1 - \frac{6 \times 30}{990}$$

$$= 1 - \frac{180}{990}$$

$$= \frac{990 - 180}{990}$$

$$= 0.818$$

Regression

If there is a functional relationship between the variables x_i & y_i the points in the scatter diagram will cluster around some curve called the curve of regression. If a curve is a straight line it is called a line of regression between the two variables.

If we fit a straight line by the principle of least squares to the points of the scatter diagram in such a way that the sum of the squares of the distance parallel to the y axis (x axis) from the points to the line to minimised we obtain a line of best fit for the data and it is called the regression line

y on x (x on y)

Theorem :- 1

The equation of the regression line of y on x is given by $y-\bar{y} = r\frac{\sigma_y}{\sigma_x}(x-\bar{x})$

Let $y = ax+b$ be the regression line of y on x

$$y_i = ax_i + b$$

$$y_i - ax_i - b = 0$$

$$(y_i - ax_i - b)^2 = 0$$

$$\Sigma(y_i - ax_i - b)^2 = 0$$

Let $S = \Sigma(y_i - ax_i - b)^2$

According to the principle of least Squares we have to determine the parameters a and b so that S is minimum

$$\frac{\partial S}{\partial a} = 0$$

$$\Rightarrow 2\Sigma(y_i - ax_i - b)(-x_i) = 0$$

$$\Rightarrow -2\Sigma(y_i - ax_i - b)(x_i) = 0$$

$$\Rightarrow \Sigma(x_i y_i - ax_i^2 - bx_i) = 0$$

$$\Rightarrow \Sigma x_i y_i - a\Sigma x_i^2 - b\Sigma x_i = 0$$

$$\Rightarrow a\Sigma x_i^2 + b\Sigma x_i = \Sigma x_i y_i \quad \text{--- (1)}$$

$$\frac{\partial S}{\partial b} = 0$$

$$\Rightarrow 2\Sigma(y_i - ax_i - b)(-1) = 0$$

$$\Rightarrow -2\Sigma(y_i - ax_i - b) = 0$$

$$\Rightarrow \Sigma(y_i - ax_i - b) = 0$$

$$\Rightarrow \Sigma y_i - a\Sigma x_i - nb = 0$$

$$\Rightarrow a\Sigma x_i + nb = \Sigma y_i \quad \text{--- (2)}$$

Equation (1) & (2) are called the normal equation

$\div$ing ② by n

$$② \Rightarrow a\frac{\Sigma x_i}{n} + \frac{nb}{n} = \frac{\Sigma y_i}{n}$$

$$a\bar{x} + b = \bar{y}$$

the regression of line passes through $(\bar{x}, \bar{y})$

Now shifting the origin to this point $(\bar{x}, \bar{y})$ by giving the transformation

$$X_i = x_i - \bar{x}, \quad Y_i = y_i - \bar{y}$$

$$X_i = x_i - \bar{x}$$

$$\Sigma X_i = \Sigma(x_i - \bar{x})$$

$$= \Sigma x_i - \Sigma \bar{x}$$

$$= n\bar{x} - n\bar{x}$$

$$= 0$$

$$\Sigma Y_i = 0$$

$\Rightarrow a\Sigma x_i + nb = \Sigma y_i$

$\Rightarrow a \times 0 + nb = 0$

$\Rightarrow nb = 0$

Hence $n \neq 0$, $b = 0$

Hence the line of regression becomes

$$Y = aX \rightarrow (3)$$

$\Rightarrow a\Sigma x_i^2 + b\Sigma x_i = \Sigma x_i y_i$

$\Rightarrow a\Sigma x_i^2 + 0\Sigma x_i = \Sigma x_i y_i$

$a\Sigma x_i^2 = \Sigma x_i y_i$

$$a = \frac{\Sigma x_i y_i}{\Sigma x_i^2}$$

$$a = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\Sigma (x_i - \bar{x})^2}$$

$$a = \frac{n \times \sigma_{xy}}{n \sigma_x^2}$$

$$a = r \frac{\sigma_y}{\sigma_x}$$

$\Rightarrow Y = aX$

$$y - \bar{y} = r \frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

Which is the regression line of y on x

Hence proved

Theorem :- ②

The equation of regression line of x on y is given by $(x-\bar{x}) = r\frac{\sigma_x}{\sigma_y}(y-\bar{y})$

Let $x = ay+b$ be the regression line of x on y

$$x_i = ay_i + b$$

$$x_i - ay_i - b = 0$$

$$\Sigma(x_i - ay_i - b)^2 = 0$$

$$\Sigma(x_i - ay_i - b)^2 = 0$$

Let $S = \Sigma(x_i - ay_i - b)^2$

According to the principle of least squares we have to determine the parameters a and b so that S is minimum

$$\frac{\partial S}{\partial a} = 0$$

$$\Rightarrow 2\Sigma(x_i - ay_i + b)(-y_i) = 0$$

$$\Rightarrow -2\Sigma(x_i - ay_i - b)y_i = 0$$

$$\Rightarrow \Sigma(x_i - ay_i - b)y_i = 0$$

$$\Rightarrow \Sigma x_i y_i - a\Sigma y_i^2 - b\Sigma y_i = 0$$

$$\Rightarrow a\Sigma y_i^2 + b\Sigma y_i = \Sigma x_i y_i \rightarrow (1)$$

$$\frac{\partial S}{\partial b} = 0$$

$$\Rightarrow 2\Sigma(x_i - ay_i - b)(-1) = 0$$

$$\Rightarrow -2\Sigma(x_i - ay_i - b) = 0$$

$$\Rightarrow \Sigma(x_i - ay_i - b) = 0$$

$$\Rightarrow \Sigma x_i - a\Sigma y_i - nb = 0 \rightarrow (2)$$

$$a\Sigma y_i + nb = \Sigma x_i \rightarrow (2)$$

Equation (1) & (2) are called the normal equation

÷ eqn ② by n

$$② \Rightarrow a\frac{\sum y_i}{n} + \frac{nb}{n} = \frac{\sum x_i}{n}$$

$$\Rightarrow a\bar{y} + b = \bar{x}$$

The regression of line passes through $(\bar{y}, \bar{x})$

Now shifting the origin to this point $(\bar{y}, \bar{x})$ by giving the transformation

$$X_i = x_i - \bar{x}, \quad Y_i = y_i - \bar{y}$$

$$X_i = x_i - \bar{x}$$

$$\sum X_i = \sum (x_i - \bar{x})$$

$$= \sum x_i - \sum \bar{x}$$

$$= n\bar{x} - n\bar{x}$$

$$= 0$$

III^{ly} $\sum Y_i = 0$

(2) $\Rightarrow a\Sigma y_i + nb = \Sigma x_i$

$\Rightarrow a \times 0 + nb = 0$

$\Rightarrow nb = 0$

Here $n \neq 0$, $b = 0$

Hence the line of regression becomes

$$x = ay \rightarrow (3)$$

(1) $\Rightarrow a\Sigma y_i^2 + b\Sigma y_i = \Sigma x_i y_i$

$\Rightarrow a\Sigma y_i^2 + 0\Sigma y_i = \Sigma x_i y_i$

$\Rightarrow a\Sigma y_i^2 = \Sigma x_i y_i$

$$a = \frac{\Sigma x_i y_i}{\Sigma y_i^2}$$

$$a = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\Sigma (y_i - \bar{y})^2}$$

$$a = \frac{r \sigma_x \sigma_y}{\sigma_y^2}$$

$$a = \frac{r \sigma_x}{\sigma_y}$$

(2) $x = aY$

$$x - \bar{x} = r \frac{\sigma_x}{\sigma_y} (y - \bar{y})$$

Hence proved

Note:

$\bar{x}, \bar{y}$ is the point of intersection of a & 2 regression line

The slope of the regression line of y on x is called the regression co-efficient of y on x an it is denoted by b_{yx}

Hence

$$b_{yx} = r \frac{\sigma_y}{\sigma_x}$$

IIIly

The regression co-efficient of x on y is given by

$$b_{xy} = r \frac{\sigma_x}{\sigma_y}$$

Theorem: 3 :-

correlation co-efficient is the geometric mean between the regression co-efficients

(ie) $r = \pm \sqrt{b_{yx} \cdot b_{xy}}$

Proof :-

We have $b_{yx} = r \frac{\sigma_y}{\sigma_x}$

$b_{xy} = r \frac{\sigma_x}{\sigma_y}$

$b_{yx} \cdot b_{xy} = r \frac{\sigma_y}{\sigma_x} \cdot r \frac{\sigma_x}{\sigma_y}$

$b_{yx} \cdot b_{xy} = r^2$

$r^2 = b_{yx} \cdot b_{xy}$

$r = \pm \sqrt{b_{yx} \cdot b_{xy}}$

Hence proved

The sign of the correlation co-efficient they same as the regression co-efficient

Theorem: 4

If one of the the regression co-efficient is greater than unity, the other is less than unity.

We have $b_{yx} = r\frac{\sigma_y}{\sigma_x}$ $b_{xy} = r\frac{\sigma_x}{\sigma_y}$

$$b_{yx} \cdot b_{xy} = r\frac{\sigma_y}{\sigma_x} \cdot r\frac{\sigma_x}{\sigma_y}$$

$$= r^2$$

$$= r^2 \leq 1$$

$$b_{yx} \cdot b_{xy} \leq 1$$

If $b_{yx} > 1$ then $b_{xy} < 1$

If $b_{xy} > 1$ then $b_{yx} < 1$

Theorem: 5)

Arithmetic mean of the regression co-efficient is greater than (or) equal to the correlation co-efficient.

Let b_{yx} & b_{xy} be the correlation co-efficient

P.T $\frac{b_{yx}+b_{xy}}{2} \geq r$

$b_{yx}+b_{xy} \geq 2r$

$r\frac{\sigma_y}{\sigma_x} + r\frac{\sigma_x}{\sigma_y} \geq 2r$

$r\left(\frac{\sigma_y}{\sigma_x} + \frac{\sigma_x}{\sigma_y}\right) \geq 2r$

$\frac{\sigma_y}{\sigma_x} + \frac{\sigma_x}{\sigma_y} \geq 2$

$\frac{\sigma_y^2 + \sigma_x^2}{\sigma_x \sigma_y} \geq 2$

$\sigma_y^2 + \sigma_x^2 \geq 2\sigma_x\sigma_y$

$\sigma_y^2 + \sigma_x^2 - 2\sigma_x\sigma_y \geq 0$

$(\sigma_x - \sigma_y)^2 \geq 0$

This condition is always true.

Theorem: 6

Regression co-efficient are indipenend of the change of origin but dependant on the change of the ~~Scale~~ Scale.

Let $u_i = \frac{x_i - A}{h}$, $v_i = \frac{y_i - B}{k}$

$hu_i = x_i - A$

$x_i = hu_i + A$

$\bar{x} = h\bar{u} + A$

$x_i - \bar{x} = hu_i + A - h\bar{u} - A$

$= hu_i - h\bar{u}$

$= h(u_i - \bar{u})$

$(x_i - \bar{x})^2 = h^2 (u_i - \bar{u})^2$

$\frac{(x_i - \bar{x})^2}{n} = \frac{h^2 (u_i - \bar{u})^2}{n}$

$\frac{\sum (x_i - \bar{x})^2}{n} = h^2 \frac{\sum (u_i - \bar{u})^2}{n}$

$kv_i = y_i - B$

$y_i = kv_i + B$

$\bar{y} = k\bar{v} + B$

$(y_i - \bar{y}) = k(v_i - \bar{v})$

$\sigma_x^2 = h^2 \sigma_u^2$

$\sigma_x = h \sigma_u$

Similarly $\sigma_y^2 = k^2 \sigma_v^2$

$\sigma_y = k \sigma_v$

$$r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n \sigma_x \sigma_y}$$

$$= \frac{\sum h(u_i - \bar{u}) * (v_i - \bar{v})}{n * h \sigma_u * \sigma_v}$$

$$= \frac{\sum (u_i - \bar{u})(v_i - \bar{v})}{n \sigma_u \sigma_v}$$

$$\therefore r_{xy} = r_{uv}$$

$$b_{xy} = r \frac{\sigma_x}{\sigma_y}$$

$$= r \frac{h(\sigma_u)}{k(\sigma_v)}$$

$$= \frac{h}{k} b_{uv}$$

$$b_{yx} = r \frac{\sigma_y}{\sigma_x}$$

$$= r \left(\frac{k}{h} \frac{\sigma_v}{\sigma_u}\right) = \frac{k}{h} b_{vu}$$

Hence the regression co-efficient are independent of origin A & B but dependent of the scale h & k

Hence proved

1) The following data relate to the marks of 10 students in the internal test and university Examination. for the maximum of 50 each

internal	25	28	30	32	35	36	38	39	42	
uni. marks	20	26	29	30	25	18	26	35	35	40

i) Obtain the two regression Equation & determine

(ii) The most likely internal mark for the university mark of 25

(iii) The most likely for the internal mark of 30

Let x be the internal mark & y the university mark

now $\bar{x} = \frac{25+28+30+32+35+36+38+39+42+45}{10}$

$= 35$

$\bar{y} = \frac{20+26+29+30+25+18+26+35+35+46}{10}$

$= 29$

x	y	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$	$(x_i-\bar{x})(y_i-\bar{y})$
25	20	-10	-9	100	81	90
28	26	-7	-3	49	9	21
30	29	-5	0	25	0	0
32	30	-3	1	9	1	-3
35	25	0	-4	0	16	0
36	18	1	-11	1	121	-11
38	26	3	-3	9	9	-9
39	35	4	6	16	36	24
42	35	7	6	49	36	42
45	46	10	17	100	289	170

$\Sigma(x_i-\bar{x})$ $\Sigma(y_i-\bar{y})$ $\Sigma(x_i-\bar{x})^2$ $\Sigma(y_i-\bar{y})^2$ $\Sigma(x_i-\bar{x})(y_i-\bar{y})$

$= 0$ $= 0$ $= 358$ $= 598$ $= 324$

$\Sigma(x_i-\bar{x})(y_i-\bar{y}) = 324$

$\Sigma(x_i-\bar{x})^2 = 358$

$\Sigma(y_i-\bar{y})^2 = 598$

$\sigma x^2 = \frac{\Sigma(x_i-\bar{x})^2}{n}$

$= \frac{358}{10}$

$= 35.8$

$\sigma x = 5.99$

$\sigma y^2 = \frac{\Sigma(y_i-\bar{y})^2}{n}$

$= \frac{598}{10}$

$= 5.98$

$\sigma y = 7.733$

$\gamma = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{n \sigma x \sigma y}$

$= \frac{324}{10 \times 5.98 \times 7.733}$

$= 0.7$ (approx)

Regression line of y on x

$$(y - \bar{y}) = r \frac{\sigma_y}{\sigma_x} (x - \bar{x})$$

$$y - 29 = 0.7 \times \frac{7.733}{5.98} (x - 35)$$

$$y - 29 = \frac{5.4131}{5.98} (x - 35)$$

$$y - 29 = 0.905 (x - 35$$

$$y - 29 = 0.905x - 31.675$$

$$y = 0.905x - 31.675 + 29$$

$$y = 0.905x - 2.675 \quad \text{--- (1)}$$

Regression line of x on y

$$x - \bar{x} = r \frac{\sigma_x}{\sigma_y} (y - \bar{y})$$

$$x - 35 = 0.7 \times \frac{5.98}{7.733} (y - 29)$$

$x - 35 = 0.5413(y - 29)$

$x - 35 = 0.5413y - 15.697$

$x = 0.5413y - 19.303$ → ④

8. when $y = 25$, $x = ?$

Subin ④

$x = 0.54 \times 25 - 19.34$

$= 32.84$

The most likely internal mark for u.m 25 is 32.84

when $x = 30$, $y = ?$

Subin ③

$y = 0.9 \times 30 - 2.5$

$y = 24.5$

The most likely u.marks for internal mark 35 is 24.5

Students obtain the following in the college internal test x and in the final univ. marks y

x	51	63	63	49	50	60	65	63	46	50
y	49	72	75	50	48	60	70	48	60	56

Estimate the univ. marks of a student who got 61 in the internal test

$xy \quad x-\bar{x} \quad y-\bar{y}$

$$\bar{x} = \frac{\sum x_i}{n}$$

$$= \frac{51+63+63+49+50+60+65+63+46+50}{10}$$

$$= 56$$

$$\bar{y} = \frac{\sum y_i}{n}$$

$$= \frac{49+72+75+50+48+60+70+48+60+56}{10}$$

$$= 58.8$$

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
51	49	−5	−9.6	25	96.04	49
62	77	7	+13.2	49	174.2	92.4
63	75	7	+11.2	49	262.4	113.4
49	50	−7	−13.8	49	−17.44	61.6
50	48	−6	−10.8	36	116.6	64.8
60	60	4	1.2	16	1.44	4.8
65	70	9	11.2	81	125.4	100.8
63	48	7	−10.8	49	116.6	−75.6
46	60	−10	1.2	100	1.44	−12
50	56	−6	−2.8	36	7.84	16.8
				$\Sigma(x_i-\bar{x})^2$	$\Sigma(y_i-\bar{y})^2$	$\Sigma(x_i-\bar{x})(y_i-\bar{y})$
				$=490$	$=979.6$	$=416$

$$\Sigma(x_i-\bar{x})(y_i-\bar{y}) = 416$$

$$\Sigma(x_i-\bar{x})^2 = 490$$

$$\Sigma(y_i-\bar{y})^2 = 979.6$$

$$\sigma_x^2 = \frac{\Sigma(x_i-\bar{x})^2}{n}$$

$$= \frac{490}{10}$$

$$= 49$$

$\sigma x = 7$

$$\sigma y^2 = \frac{\Sigma (y_i - \bar{y})^2}{n}$$

$$= \frac{971.6}{10}$$

$$= 97.96$$

$$\sigma y = 9.9$$

$$r = \frac{\Sigma (x - \bar{x})(y - \bar{y})}{n \sigma x \sigma y}$$

$$= \frac{416}{10 \times 9.9 \times 7}$$

$$= \frac{416}{693}$$

$$= 0.60$$

The regression line of y on x is

$$y - \bar{y} = r \cdot \frac{\sigma y}{\sigma x} (x - \bar{x})$$

$$y - 58.8 = 0.60 \times \frac{9.90}{7} (x - 56)$$

$$y - 58.8 = 0.85 (x - 56)$$

$y = 0.85x$

(i.e.) $y - 58.1 = 0.85x - 47.6$

$y = 0.85x - 47.6 + 58.8$

$y = 0.85x + 11.2$

when $x = 61$, $y = ?$

$y = 0.85x + 11.2$

$= 0.85 \times 61 + 11.2$

$= 51.85 + 11.2$

$\boxed{y = 63.05}$

d) Out of the two lines of Regression $x + 2y - 5 = 0$ & $2x + 3y - 8 = 0$ which one is the regression line of x on y

b_{xy}

b_{yx} $y^2 = b_{xy}\ b_{yx}$

The given two regression lines are

$x + 2y - 5 = 0$, $2x + 3y - 8 = 0$

$x = -2y + 5$ | $3y = -2x + 8$

$x = -2y + 5$ | $y = \frac{-2}{3}x + \frac{8}{3}$

Suppose the regression line of x on y

$x = -2y + 5$

Here $b_{xy} = -2$

The regression line of y on x

$y = -\frac{2}{3}x + \frac{8}{3}$

Here $b_{yx} = -\frac{2}{3}$

we have $r^2 = b_{xy} \cdot b_{yx}$

$= -2 \cdot -\frac{2}{3}$

$= \frac{4}{3}$

$r^2 = 1.33 > 1$

This is not possible.

∴ Our assumption is wrong

Hence regression line of x on y is

$2x + 3y - 8 = 0$

1) The 2 variables x and y for the regression line $3x + 2y - 26 = 0$ & $6x + y - 31 = 0$ find (i) The mean values of x & y

(ii) Prove that the correlation co-efficient between x & y

(iii) The variens of y if the varriens of x is 25.

(i) Since the two lines passes through $(\bar{x}, \bar{y})$

$3\bar{x} + 2\bar{y} = 26 \rightarrow (1)$

$6\bar{x} + \bar{y} = 31 \rightarrow (2)$

$(1) \times 2 \Rightarrow 6\bar{x} + 4\bar{y} = 52$

$(2) \Rightarrow 6\bar{x} + \bar{y} = 31$

$3\bar{y} = 21$

$\bar{y} = 7$

Sub $y = 7$ in ①

$3\bar{x} + 2 \times 7 = 26$

$3\bar{x} = 26 - 14$

$3\bar{x} = 12$

$\bar{x} = 4$

(ii) Suppose $3x + 2y - 26 = 0$ is the regression line of x on y

$3x = -2y + 26$

$x = \frac{-2}{3}y + \frac{26}{3}$

$b_{xy} = -2/3$

$6x + y - 31 = 0$ is the Regression line of y on x

~~$6x = -y + 31$~~ $y = -6x + 31$

$b_{yx} = -6$

$r^2 = b_{yx} \cdot b_{xy}$

$= -6 \times -2/3$

$= 4 > 1$

∴ our assumption is wrong

$3x + 2y - 26 = 0$ is the regression line of y on x

$2y = -3x + 26$

$y = \frac{-3}{2}x + \frac{26}{2}$

$\boxed{b_{yx} = -3/2}$

$6x + y - 31 = 0$ is the regression line of x on y

$6x = -y + 31$

$x = \frac{-y}{6} + \frac{31}{6}$

$\boxed{b_{xy} = -1/6}$

$r^2 = b_{xy} \cdot b_{yx}$

$= -1/6 \times -3/2$

$= 1/4$

$r = \pm 0.5$

$r = -0.5$

(iii) Variance of $x = 25$

$\sigma x^2 = 25$

$\sigma x = 5$

To find σy

we have,

$b_{xy} = r \cdot \frac{\sigma x}{\sigma y}$

$\frac{-1}{6} = -0.5 \cdot \frac{5}{\sigma y}$

$+\frac{\sigma y}{6} = +2.5$

$\sigma y = 2.5 \times 6$

$\sigma y = 15$

$\sigma y^2 = 225$

2) If $x = 4y + 5$ & $y = kx + 4$ are the regression line of x on y & y on x respectively (i) Show that $0 \leq k \leq 1/4$ (ii) If $k = 1/8$ find the means of the two variables x & y and the correlation co-efficient between them

(i) The regression line of x on y is

$$x = 4y + 5$$

$$\boxed{b_{xy} = 4}$$

Regression line of y on x is

$$y = kx + 4$$

$$\boxed{b_{yx} = k}$$

Now, $r^2 = b_{xy} \cdot b_{yx}$

$$= 4k$$

$$r^2 = 4k$$

we have,

$0 \leq r^2 \leq 1$

$0 \leq 4k \leq 1$

$0 \leq k \leq 1/4$

. Hence proved

(ii) If $k = 1/8$

$r^2 = 4k$

$= 4 \times 1/8$

$r^2 = 1/2$

$r = \pm \sqrt{\frac{1}{2}}$

$r = \pm 0.707$

$r = +0.707$ [$\because$ by r & hence y are positive]

~~[illegible]~~

$x = 4y + 5$

~~[illegible]~~

$x - 4y - 5 = 0 \rightarrow (i)$

$y = kx + 4$

$y = 1/8 x + 4$

$y = \frac{x + 32}{8}$

$8y = x + 4$

$x - 8y + 32 = 0 \rightarrow (b)$

The two regression lines passes through $(\bar{x}, \bar{y})$

$\bar{x} - 4\bar{y} - 5 = 0 \rightarrow (a)$

$\bar{x} - 8\bar{y} + 32 = 0 \rightarrow (b)$

$4\bar{y} - 37 = 0$

$4\bar{y} = 37$

$\bar{y} = 37/4$

$\boxed{\bar{y} = 9.25}$

Sub $\bar{y} = 9.25$ in (a)

$\bar{x} - 4 \times 9.25 - 5 = 0$

$\bar{x} - 37 - 5 = 0$

$\bar{x} - 42 = 0$

$\boxed{\bar{x} = 42}$

9) The variable x & y are connected by the equation $ax+by+c=0$. Show that $r_{xy} = -1$ (or) 1 according a & b are of the same sign or of opposite sign.

Writing $ax+by+c=0$ in the form

$$ax = -by - c$$

$$x = \frac{-b}{a}y - c/a$$

$$\boxed{b_{xy} = -b/a}$$

writing $ax+by+c=0$ in the form

$$by = -ax - c$$

$$y = -\frac{a}{b}x - c/b$$

$$\boxed{b_{yx} = -a/b}$$

Now, $r^2 = b_{xy} \cdot b_{yx}$

$$= \frac{-b}{a} \times \frac{-a}{b}$$

$$= 1$$

$$r^2 = 1$$

$$r = \pm 1$$

Suppose a & b are of same sign then

$r^2 = 1$

Hence $r = -1$ [$\because$ b_{xy} & b_{yx} are negative]

Suppose a & b are of opposite sign

then $r^2 = 1$

Hence $r = 1$ [$\because$ b_{xy} & b_{yx} are positive]

4) The following table shows the ages x & blood pressure y are given of

12 women

(i) find the correlation co-efficient between x & y

(ii) determine the regression equation of y on x

(iii) estimate the blood pressure of a women whose age is 45

Age (x)	56	42	72	36	63	47	55	49	38	42	68	60
Blood pressure (y)	147	125	160	118	149	128	150	145	115	140	152	155

Q) The equations of two regression lines obtained in a correlation analysis are $4x - 5y + 33 = 0$ & $20x - 9y - 107 = 0$. If the variance $y = 16$. find

(i) The mean values of x & y

(ii) The correlation coefficient between x & y

(iii) Standard deviation of x.

Sol) (i) Since the two lines passes through $\bar{x}, \bar{y}$

$$4\bar{x} - 5\bar{y} = -33 \rightarrow (1)$$

$$20\bar{x} - 9\bar{y} = 107 \rightarrow (2)$$

$(1) \times 5 \Rightarrow 20\bar{x} - 25\bar{y} = -165$

$(2) \Rightarrow 20\bar{x} - 9\bar{y} = 107$

$$-16\bar{y} = -272$$

$16\bar{y} = 272$

$\bar{y} = \frac{272}{16}$

$\boxed{\bar{y} = 17}$

Sub $\bar{y} = 17$ in (1)

$4\bar{x} - 5(17) = -33$

$4\bar{x} - 85 = -33$

$4\bar{x} = -33 + 85$

$4\bar{x} = 52$

$\bar{x} = 52/4$

$\boxed{\bar{x} = 13}$

(ii) Suppose $4x - 5y + 33 = 0$ is the regression line y on x

$4x = 5y - 33$

$x = \frac{5y}{4} - \frac{33}{4}$

$\boxed{b_{xy} = 5/4}$

Suppose $20\bar{x} - 9\bar{y} = 107$ is the regression line of y on x

$$-9\bar{y} = -20\bar{x} + 107$$

$$\bar{y} = \frac{+20}{+9}\bar{x} - \frac{107}{9}$$

$$\bar{y} = \frac{20}{9}\bar{x} - \frac{107}{9}$$

$$\boxed{b_{yx} = 20/9}$$

Now, $r^2 = b_{xy} \cdot b_{yx}$

$$= 5/x \times \frac{20}{9}$$

$$r^2 = 25/9 \qquad r^2 = 2.78 > 1$$

~~$x \leq 5/3$~~ Our assumption is wrong

$4x - 5y + 33 = 0$ is ~~$x \leq 5/3$~~ the regression line of y on x

$$-5y = -4x - 33$$

$$5y = 4x + 33$$

$$y = 4/5\,x + 33/5$$

$$\boxed{b_{yx} = 4/5}$$

$20x - 9y - 107 = 0$ is the regression line of x on y

$20x = 9y + 107$

$x = 9/20\,y + 107/20$

$\boxed{b_{xy} = 9/20}$

$r^2 = b_{yx} \cdot b_{xy}$

$= 4/5 \cdot 9/20$

$r^2 = 9/25 = 0.36$

$r = \pm 0.6$

$r = 0.6$

(iii) Given variance of $y = 16$

$\sigma_y^2 = 16$

$\sigma_y = 4$

To find the variance of x

$\sigma_x^2 = ?$

$b_{yx} = r \frac{\sigma_y}{\sigma_x}$

$4/5 = 0.6 \times 4/\sigma_x$

$4/5 = \frac{2.4}{\sigma_x}$

$4\sigma_x = 2.4 \times 5$

$4\sigma_x = 12$

$\sigma_x = 12/4$

$\sigma_x = 3$

$\sigma_x^2 = 9$

Standard deviation of x $\boxed{\sigma_x = 3}$

x or y

deviation

= 3

(a) (i) Let x be the Age
and y be the blood pressure

Now $\bar{x} = \frac{\sum x_i}{n}$

$= \frac{56+42+72+36+63+47+55+49+38+42+68+60}{12}$

$= \frac{628}{12}$

$\bar{x} = 52.33$

$\bar{y} = \frac{\sum y_i}{n}$

$= \frac{120+125+160+118+149+128+150+145+115+140+152+158}{12}$

$= \frac{1684}{12} = 140.33$

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
56	147	3.67	6.67	24.48	13.47	44.49
48	125	-10.33	-15.33	158.36	106.71	235.01
72	160	19.67	19.67	386.91	386.91	386.91
36	118	-16.33	-22.33	364.65	266.67	498.65
63	149	10.67	8.67	92.51	113.85	75.63
47	128	-5.33	-12.33	65.72	28.41	152.03
55	150	2.67	9.67	25.82	7.13	93.51
49	145	-3.33	4.67	-15.55	11.09	21.81
38	115	-14.33	-25.33	362.98	205.35	641.61
42	140	-10.33	-0.33	3.41	106.71	0.11
68	152	15.67	11.67	182.87	245.55	136.19
60	152	7.67	14.67	112.52	58.83	215.21
		$\Sigma(x-\bar{x})(y-\bar{y})$		$\Sigma(x-\bar{x})^2$	$\Sigma(y-\bar{y})^2$	
		$=1769.68$		$=1550.68$	$=2800.68$	

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

$$= \frac{1550.68}{12}$$

$$\sigma x^2 = 129.2$$

$$\sigma x = 11.37$$

$$\sigma y^2 = \frac{\Sigma(y-\bar{y})^2}{n}$$

$$= \frac{2800.68}{12}$$

$$= 16.44$$

$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{n\sigma x \sigma y}$$

$$= \frac{1764.68}{12 \times 11.37 \times 16.44}$$

$$= \frac{1764.68}{1970.19}$$

$$r = 0.90$$

(ii) The regression line of y on x

$$y - \bar{y} = r\frac{\sigma y}{\sigma x}(x - \bar{x})$$

$$y - 140.33 = 0.9\left(\frac{16.44}{11.37}\right)(x - 52.33)$$

$$y - 140.33 = 0.9(1.27)(x - 52.33)$$

$$y - 140.33 = 1.146x - 59.66$$

$$y = 1.146x - 59.66 + 140.33$$

$$y = 1.14x + 80.67$$

(iii) when $x = 45$, $y = ?$

$$y = 1.146(45) + 80.67$$

$$= 51.3 + 80.67$$

$$= 131.97$$

Theorem :

The angle between the two regression line is given by $\theta = \tan^{-1}\left[\left(\frac{r^2-1}{r}\right)\left(\frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2}\right)\right]$

Regression line of x on y is

$$x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y})$$

$$x - \bar{x} = r\frac{\sigma_x}{\sigma_y}y - r\frac{\sigma_x}{\sigma_y}\bar{y}$$

$$r\frac{\sigma_x}{\sigma_y} \cdot y = x - \bar{x} + r\frac{\sigma_x}{\sigma_y}\bar{y}$$

$$y = \frac{\sigma_y}{r\sigma_x}\left[x - \bar{x} + r\frac{\sigma_x}{\sigma_y}\bar{y}\right]$$

$$y = \frac{\sigma_y}{r\sigma_x} x - \frac{\sigma_y}{r\sigma_x}\left(\bar{x} - r\frac{\sigma_x}{\sigma_y}\bar{y}\right)$$

$$\boxed{m_2 = \frac{\sigma_y}{r\sigma_x}}$$

Regression line of y on x

$$y - \bar{y} = (x - \bar{x})\, r\frac{\sigma_y}{\sigma_x}$$

$$y = r\frac{\sigma_y}{\sigma_x} x - r\frac{\sigma_y}{\sigma_x}\bar{x} + \bar{y}$$

$$\boxed{m_1 = r\frac{\sigma_y}{\sigma_x}}$$

Let θ be the obtuse angle between two regression lines

$$\tan\theta = \frac{m_1 - m_2}{1 + m_1 m_2}$$

$$= \frac{r\frac{\sigma_y}{\sigma_x} - \frac{\sigma_y}{r\sigma_x}}{1 + r\frac{\sigma_y}{\sigma_x} \times \frac{\sigma_y}{r\sigma_x}}$$

$$= \frac{r\,\sigma_y/\sigma_x - \frac{\sigma_y}{r\sigma_x}}{1 + \frac{\sigma_y^2}{\sigma_x^2}}$$

$$= \frac{y^2 \delta y - \delta y}{2\delta x}$$

$$\frac{\delta x^2 + \delta y^2}{\delta x^2}$$

$$= y^2 \frac{\delta y}{\delta x} - \frac{\delta y}{\delta x}$$

$$= \frac{y^2 \delta y - \delta y}{2\delta x}$$

$$\frac{\delta x^2 + \delta y^2}{\delta x^2}$$

$$= \frac{\delta y (y^2 - 1)}{2\delta x}$$

$$\frac{\delta x^2 + \delta y^2}{\delta x^2}$$

$$= \frac{\delta y}{2\delta x} (y^2 - 1) \times \frac{\delta x^2}{\delta x^2 + \delta y^2}$$

$$= \frac{\delta y (y^2 - 1) \delta x}{2(\delta x^2 + \delta y^2)}$$

$$= \frac{\sigma x \sigma y (r^2 - 1)}{r(\sigma x^2 + \sigma y^2)}$$

$$\tan\theta = \left(\frac{r^2-1}{r}\right)\left(\frac{\sigma x \sigma y}{\sigma x^2 + \sigma y^2}\right)$$

$$\theta = \tan^{-1}\left[\left(\frac{r^2-1}{r}\right)\left(\frac{\sigma x \sigma y}{\sigma x^2 + \sigma y^2}\right)\right]$$

Hence Proved

Note: 1

The accute angle between the regression line is given by $\theta = \tan^{-1}\left[\left(\frac{1-r^2}{r}\right)\left(\frac{\sigma x \sigma y}{\sigma x + \sigma y^2}\right)\right]$

Note: 2

If $r = 0$ the $\theta = \tan^{-1}(\infty)$

$= \pi/2$

Thus If the two variables are uncorrelated then the lines of regression are perpendicular to each other.

Note: 3.

If $r = \pm 1$ then $\theta = \tan^{-1}(0)$

$\theta = 0$ (or) π

∴ The two regression lines are parallel

The two lines have the common point $(\bar{x}, \bar{y})$ Then the two line must be co-incident ∴ If there is a perfect correlation (Positive or Negative) between the 2 Variables then the two lines of regression co-inside

1) If θ is a accute angle between the two regression line show that $\sin\theta \leq 1 - r^2$

we have if θ is a accute angle between the two regression line

then $\theta = \tan^{-1}\left[\left(\frac{1-r^2}{r}\right)\cdot\left(\frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2}\right)\right]$

We assume that $\sigma_x^2 + \sigma_y^2 \geq 2\sigma_x\sigma_y$ → (1)

Suppose if it is not true

$$\sigma_x^2 + \sigma_y^2 < 2\sigma_x\sigma_y$$

$$\sigma_x^2 + \sigma_y^2 - 2\sigma_x\sigma_y < 0$$

$$(\sigma_x - \sigma_y)^2 < 0$$

So this is impossible $[\because (\sigma_x - \sigma_y)^2 > 0]$

Our assumption is corong

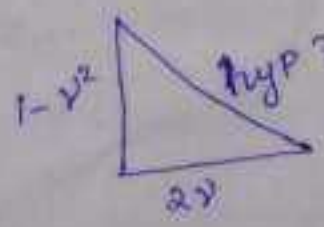
$$\sigma_x^2 + \sigma_y^2 \geq 2\sigma_x\sigma_y$$

$$\frac{1}{2} \geq \frac{\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2}$$

$$\frac{\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2} \leq \frac{1}{2}$$

$$\tan\theta \leq \frac{1-r^2}{r} \cdot \frac{1}{2}$$

$$\tan\theta \leq \frac{1-r^2}{2r}$$



$$(hyp)^2 = (1-x^2)^2 + (2x)^2$$

$$hyp = \sqrt{(1-x^2)^2 + 4x^2}$$

$$= \sqrt{1+x^4-2x^2+4x^2}$$

$$= \sqrt{1+x^4+2x^2}$$

$$= \sqrt{(x^2+1)^2}$$

$$= x^2+1$$

we have $\sin\theta \leq \frac{1-x^2}{x^2+1}$

$$\sin\theta \leq 1-x^2$$

Unit II

Moments, Skewness, Kurtosis

Moments :-

definition :-

The r^{th} moment about any point A, denoted by μ_r' of a frequency distribution (f_i / x_i) is defined by $\mu_r' = \frac{\sum f_i (x_i - A)^r}{N}$

When $A = 0$, we get

$\mu_r' = \frac{\sum f_i x_i^r}{N}$

Which is the r^{th} moment about the origin

The r^{th} moment about the arithmetic mean $\bar{x}$ of a frequency distribution is given by

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

μ_r is also called the r^{th} central moment.

Note :-

The first moment about origin coincides with the Arithmetic mean of the frequency distribution. and μ_2 is nothing but the varians of the frequency distribution.

Note-2

$$\mu_1' = \frac{\sum f_i (x_i - \bar{x})}{N}$$

$$= \frac{\sum f_i x_i}{N} - \frac{\sum f_i \bar{x}}{N}$$

$$= \bar{x} - \frac{N\bar{x}}{N}$$

$$= \bar{x} - \bar{x}$$

$$= 0$$

Note-3:

$$\mu_1' = \frac{\sum f_i (x_i - A)}{N}$$

$$= \frac{\sum f_i x_i}{N} - \frac{A \sum f_i}{N}$$

$$= \bar{x} - \frac{AN}{N}$$

$$\mu_1' = \bar{x} - A$$

$$\boxed{\bar{x} = \mu_1' + A}$$

Relation between μ_r and μ_r'

Theorem : 4.1

$$\mu_r = \mu_r' - {}^rC_1 \mu_{r-1}' \cdot \mu_1' + {}^rC_2 \mu_{r-2}' (\mu_1')^2 + \ldots\ldots\ldots + (-1)^{r-1} \cdot (r-1)(\mu_1')^r$$

We have,

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

$$= \frac{\sum f_i (x_i - A + A - \bar{x})^r}{N}$$

$$= \frac{\sum f_i [x_i - A - (\bar{x} - A)]^r}{N}$$

$$= \frac{\sum f_i [(x_i - A) - d]^r}{N}$$

where $d = \bar{x} - A$

$= \mu_1'$

$$= \frac{\sum f_i}{N} \left[(x_i - A)^r + {}^rC_1 (x_i - A)^{r-1} \cdot (-d) + {}^rC_2 (x_i - A)^{r-2} \cdot (-d)^2 + \ldots\ldots + {}^rC_{r-1} (x_i - A)(-d)^{r-1} + {}^rC_r (-d)^r \right]$$

$$= \frac{\sum f_i}{N}\left[(x_i - A)^r - {}^rC_1 (x_i - A)^{r-1} d + {}^rC_2 (x_i - A)^{r-2} \cdot d^2 + \ldots\ldots + {}^rC_{r-1} (x_i - A)(-1)^{r-1} \cdot d^{r-1} + (-1)^r \cdot d^r\right]$$

$$= \frac{\sum f_i (x_i - A)^r}{N} - {}^rC_1 \cdot d \frac{\sum f_i (x_i - A)^{r-1}}{N} + {}^rC_2 \cdot d^2 \frac{\sum f_i (x_i - A)^{r-2}}{N} + \ldots\ldots + {}^rC_{r-1} (-1)^{r-1} d^{r-1} \frac{\sum f_i (x_i - A)}{N} + (-1)^r d^r \frac{\sum f_i}{N}$$

$$= \mu_r' - {}^rC_1 \mu_{r-1}' \cdot \mu_1' + {}^rC_2 \mu_{r-2}' \cdot (\mu_1')^2 + \ldots\ldots + {}^rC_{r-1} (-1)^{r-1} \cdot (\mu_1')^{r-1} \cdot \mu_1' + (-1)^r \cdot (\mu_1')^r$$

$$= \mu_r' - {}^rC_1 \mu_{r-1}' \cdot \mu_1' + {}^rC_2 \mu_{r-2}' \cdot (\mu_1')^2 + \ldots + {}^rC_{r-1} (-1)^{r-1} \cdot (\mu_1')^{r-1} \cdot \mu_1' + (-1)^{r-1} \cdot (-1)(\mu_1')^r$$

$$= \mu_r' - {}^rC_1 \mu_{r-1}' \cdot \mu_1' + {}^rC_2 \mu_{r-2}' (\mu_1')^2 + \ldots\ldots + (-1)^{r-1} \cdot \mu_1^r + (-1)^{r-1} (-1) (\mu_1')^r$$

$$= \mu_r' - {}^rC_1 \mu_{r-1}' \cdot \mu_1' + {}^rC_2 \mu_{r-2}' (\mu_1')^2 + \ldots\ldots + (-1)^{r-1} (\mu_1')^r (r-1)$$

Note:

Put $r = 1, 2, 3, 4$ we have

$$\mu_1 = \mu_1' - {}^1C_1 \cdot \mu_{1-1}' \cdot \mu_1'$$

$$= \mu_1' - \mu_0' \cdot \mu_1'$$

$$= \mu_1' - \mu_1'$$

$$= 0$$

$$\mu_2 = \mu_2' - 2C_1 \mu_{2-1}' \cdot \mu_1' + 2C_2 \mu_{2-2}' (\mu_1')^2$$

$$= \mu_2' - 2\mu_1' \cdot \mu_1' + \mu_0' \cdot (\mu_1')^2$$

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

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

$$\mu_3 = \mu_3' - 3C_1 \mu_{3-1}' \cdot \mu_1' + 3C_2 \mu_{3-2}' \cdot (\mu_1')^2 - 3C_3 \mu_{3-3}' \cdot (\mu_1')^3$$

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

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

$$\mu_4 = \mu_4' - 4C_1 \mu_{4-1}' \cdot \mu_1' + 4C_2 \mu_{4-2}' (\mu_1')^2 - 4C_3 \mu_{4-3}' (\mu_1')^3 + 4C_4 \mu_{4-4}' (\mu_1')^4$$

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

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

Theorem : 4.2

$$\mu_r' = \mu_r + {}^rC_1\mu_{r-1}(\mu_1') + {}^rC_2\mu_{r-2}(\mu_1')^2 + \ldots + (\mu_1')^r$$

We have, $\mu_r' = \dfrac{\sum f_i(x_i - A)^r}{N}$

$$= \frac{\sum f_i\,(x_i - \bar{x} + \bar{x} - A)^r}{N}$$

$$= \frac{\sum f_i\,[(x_i - \bar{x}) + d]^r}{N} \quad \text{Where } d = \bar{x} - A = \mu_1'$$

$$= \frac{\sum f_i}{N}\left[(x_i - \bar{x})^r + {}^rC_1(x_i - \bar{x})^{r-1}d + {}^rC_2(x_i - \bar{x})^{r-2}d^2 + \ldots + {}^rC_{r-1}(x_i - \bar{x})\cdot d^{r-1} + d^r\right]$$

$$= \frac{\sum f_i}{N} \cdot \left[(x_i - \bar{x})^r + {}^rC_1 (x_i - \bar{x})^{r-1} \cdot \mu_1' + {}^rC_2 (x_i - \bar{x})^{r-2} (\mu_1')^2 \right.$$

$$\left. + \ldots\ldots + {}^rC_{r-1} (x_i - \bar{x}) (\mu_1')^{r-1} + \mu_1'^r \right]$$

$$= \frac{\sum f_i (x_i - \bar{x})^r}{N} + {}^rC_1 \frac{\sum f_i (x_i - \bar{x})^{r-1}}{N} \cdot \mu_1' + {}^rC_2 \quad (= 0)$$

$$\frac{\sum f_i (x_i - \bar{x})^{r-2}}{N} \cdot (\mu_1')^2 + \ldots\ldots + {}^rC_{r-1} \frac{\sum f_i (x_i - \bar{x})}{N} (\mu_1')^{r-1}$$

$$+ \frac{\sum f_i}{N} (\mu_1')^r$$

$$\therefore \frac{\sum f_i}{N} = \mu_r + {}^rC_1 \mu_{r-1} \cdot \mu_1' + {}^rC_2 \mu_{r-2} (\mu_1')^2$$

$$+ \ldots\ldots + (\mu_1')^r$$

Put $r = 2, 3, 4$ we get

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

$\mu_1 = \frac{\sum f_i (x_i - \bar{x})}{N} = 0$

$r = 2$

$$\mu_2' = \mu_2 + {}^2C_1 \mu_{2-1} (\mu_1') + {}^2C_2 \mu_{2-2} (\mu_1')^2$$

$$= \mu_2 + 2\mu_1 \mu_1' + 1 \times \mu_0 (\mu_1')^2$$

$$= \mu_2 + 2\mu_1^0 \mu_1' + (\mu_1')^2$$

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

$r = 3$

$$\mu_3' = \mu_3 + 3C_1 \mu_{3-1} (\mu_1') + 3C_2 \mu_{3-2} (\mu_1')^2 + 3C_3 \mu_{3-3} (\mu_1')^3$$

$$= \mu_3 + 3\mu_2 (\mu_1')^1 + 3\mu_1 (\mu_1')^2 + 1 \times \mu_0 (\mu_1')^3$$

$$= \mu_3 + 3\mu_2 (\mu_1') + 3\mu_1' (\mu_1')^2 + (\mu_1')^3$$

$$= \mu_3 + 3\mu_2 (\mu_1') + (\mu_1')^3$$

$r = 4$

$$\mu_4' = \mu_4 + 4C_1 \mu_{4-1} (\mu_1') + 4C_2 \mu_{4-2} (\mu_1')^2 + 4C_3 \mu_{4-3} (\mu_1')^3 + 4C_4 \mu_{4-4} (\mu_1')^4$$

$$= \mu_4 + 4\mu_3 \mu_1' + 6\mu_2 (\mu_1')^2 + 4\mu_1' (\mu_1')^3 + 1 \times \mu_0 (\mu_1')^4$$

$$= \mu_4 + 4\mu_3 \mu_1' + 6\mu_2 (\mu_1')^2 + (\mu_1')^4$$

Karl Pearson's β and γ coefficients:

$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$ and $\beta_2 = \frac{\mu_4}{\mu_2^2}$

$\gamma_1 = \sqrt{\beta_1}$ and $\gamma_2 = \beta - 3$

If $\beta_1 = 0$ then the frequency distribution is symmetric.

If $\beta_1 > 0$ then the frequency distribution has positive skewness

If $\beta_1 < 0$ then the frequency distribution has negative skewness.

Mean - Mode and Mean - Median may be taken as measure

of skewness

$\frac{\text{Mean} - \text{Mode}}{\sigma}$ and $\frac{3(\text{Mean} - \text{Median})}{\sigma}$ are called karl personi coefficient of skewness.

$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$ $\beta_2 = \frac{\mu_4}{\mu_2^2}$

Kurtosis:

Kurtosis is the degrees of peakedness of a distribution related to a Normal distribution

For a normal curve,

If $\beta_2 = 3$ or $\gamma_2 = 0$ Then it is Mesokurtic.

If $\beta_2 < 3$ or $\gamma_2 < 0$ then it is platykurtic

If $\beta_2 > 3$ or $\gamma_2 > 0$ then it is lepto kurtic

Problems:

1. Calculate the first four central moments for the following data to find β_1 and β_2 and discuss the nature of the distribution:

Datas:

x	0	1	2	3	4	5	6
f	5	15	17	25	19	14	5

$$\bar{x} = \frac{\sum f_i x_i}{N \Rightarrow \sum f_i}$$

$$= \frac{0\times5 + 1\times15 + 2\times17 + 3\times25 + 4\times19 + 5\times14 + 6\times5}{5+15+17+25+19+14+5}$$

$$= \frac{0+15+34+75+76+70+30}{100}$$

$$= \frac{300}{100} = 3$$

x	f_i	$x - \bar{x}$	$f_i(x_i - \bar{x})$	$f_i(x_i - \bar{x})^2$	$f_i(x_i - \bar{x})^3$	$f_i(x_i - \bar{x})^4$
0	5	−3	−15	45	−135	405
1	15	−2	−30	60	−120	240
2	17	−1	−17	17	−17	17
3	25	0	0	0	0	0
4	19	1	19	19	19	19
5	14	2	28	56	112	224
6	5	3	15	45	135	405
	$\sum f_i = 100$		$\sum f_i(x_i - \bar{x}) = 0$	$\sum f_i(x_i - \bar{x})^2 = 242$	$\sum f_i(x_i - \bar{x})^3 = -6$	$\sum f_i(x_i - \bar{x})^4 = 1310$

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

$r = 1, 2, 3, 4$

$$\mu_1 = \frac{\sum f_i (x_i - \bar{x})}{N} = 0$$

$$\mu_2 = \frac{\sum f_i (x_i - \bar{x})^2}{N} = \frac{242}{100} = 2.42$$

$$\mu_3 = \frac{\sum f_i (x_i - \bar{x})^3}{N} = \frac{-6}{100} = -0.06$$

$$\mu_4 = \frac{\sum f_i (x_i - \bar{x})^4}{N} = \frac{1310}{100} = 13.1$$

Hence The first four central moment are

$\mu_1 = 0$, $\mu_2 = 2.42$, $\mu_3 = -0.06$, $\mu_4 = 13.1$

$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(-0.06)^2}{(2.42)^3}$

$= 0.0003$

$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{13.1}{(2.42)^2} = 2.237$

Here $\beta_1 > 0$ then the distribution has possitive skewness.

$\beta_2 < 3$ then it is platykurtic.

2. Calculate the first four central moments for the following data to find β_1 and β_2 and discuss the nature of the distribution.

x	f
0	1
1	8
2	2
3	5
4	7
5	5
6	2
7	8

Data:

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

$$\bar{x} = \frac{\sum f_i x_i}{N}$$

$$= \frac{0\times1 + 1\times8 + 2\times28 + 3\times56 + 4\times70 + 5\times56 + 6\times28 + 7\times8 + 8\times1}{1+8+28+56+70+56+28+8+1}$$

$$= \frac{0+8+56+168+280+280+168+56+8}{256}$$

$$= \frac{1024}{256} = 4$$

x	f	$x-\bar{x}$	$f_i(x_i-\bar{x})$	$f_i(x_i-\bar{x})^2$	$f_i(x_i-\bar{x})^3$	$f_i(x_i-\bar{x})^4$
0	1	-4	-4	16	-64	256
1	8	-3	-24	72	-216	648
2	28	-2	-56	112	-224	448
3	56	-1	-56	56	-56	56
4	70	0	0	0	0	0
5	56	1	56	56	56	56
6	28	2	56	112	224	448
7	8	3	24	72	216	648

8	1	4	4	16	64	256
$\sum f_i$ = 256		$\sum f_i(x_i-\bar{x})$ = 0	$\sum f_i(x_i-\bar{x})^2$ = 512	$\sum f_i(x_i-\bar{x})^3$ = 0		$\sum f_i(x_i-\bar{x})^4$ = 2816

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

$r = 1, 2, 3, 4$

$$\mu_1 = \frac{\sum f_i (x_i - \bar{x})}{N}$$

$$= 0$$

$$\mu_2 = \frac{\sum f_i (x_i - \bar{x})^2}{N} = \frac{512}{256} = 2$$

$$\mu_3 = \frac{\sum f_i (x_i - \bar{x})^3}{N} = 0$$

$$\mu_4 = \frac{\sum f_i (x_i - \bar{x})^4}{N} = \frac{2816}{256} = 11$$

The first four central moment are

$\mu_1 = 0$, $\mu_2 = 2$, $\mu_3 = 0$, $\mu_4 = 11$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{0}{4} = 0$$

$$\beta_1 = 0$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{11}{4} = 2.75$$

Here $\beta_1 = 0$ then the frequency distribution is symmetric

$\beta_2 < 3$ then it is Platykurtic.

3. Calculate the values of β_1 and β_2 for the distribution given in the following table.

Marks	0-9	10-19	20-29	30-39	40-49
Frequency	11	20	16	37	17

Mid value = 24.5 = A

marks x_i	f_i	$x_i - A$	$f_i(x_i - A)$	$f_i(x_i - A)^2$	$f_i(x_i - A)^3$	$f_i(x_i - A)^4$
$\frac{L+U}{2}$						
4.5	11	-20	-220	4400	-88000	1760000
14.5	20	-10	-200	2000	-20000	200000
24.5	16	0	0	0	0	0
34.5	36	10	360	3600	36000	360000
44.5	17	20	340	6800	1,36000	2720000
	$\Sigma f_i = 100$		$\Sigma f_i(x_i - A) = 280$	$\Sigma f_i(x_i - A)^2 = 16800$	$\Sigma f_i(x_i - A)^3 = 64000$	$\Sigma f_i(x_i - A)^4 = 2542000$

$$\mu_3' = \frac{\Sigma f_i (x_i - A)^3}{N}$$

$$\mu_1' = \frac{\Sigma f_i (x_i - A)}{N}$$

$$= \frac{280}{100} = 2.8$$

$$\mu_2' = \frac{\Sigma f_i (x_i - A)^2}{N}$$

$$= \frac{16800}{100} = 168$$

$$\mu_3' = \frac{\sum f_i (x_i - A)^3}{N}$$

$$= \frac{64000}{100} = 640$$

$$\mu_4' = \frac{\sum f_i (x_i - A)^4}{N}$$

$$= \frac{25,92,000}{100}$$

$$= 25920$$

$\mu_1 = 0$

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

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

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

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

$= 168 - (2.8)^2$

$= 168 - 7.84$

$= 160.16$

$\mu_3 = \mu_3' - 3\mu_3' \cdot \mu_1' + 2(\mu_1')^3$

$= 640 - 3 \times 168 \times 2.8 + 2 \times (2.8)^3$

$= -729.296$

$\mu_4 = \mu_4' - 4\mu_3' \cdot \mu_1' + 6\mu_3' \cdot (\mu_1')^2 - 3(\mu_1')^4$

$= 25920 - 4 \times 640 \times 2.8 + 6 \times 168 \times (2.8)^2 - 3 \times (2.8)^4$

$= 26470.3232$

The first four central moment are

$\mu_1 = 0$, $\mu_2 = 160.16$, $\mu_3 = -727.296$,

$\mu_4 = 26470.3232$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(-727.296)^2}{(160.16)^3}$$

$$= 0.129 > 0$$

Here $\beta_1 > 0$ it has positive skewness

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{26470.3232}{(160.16)^2}$$

$$= 1.0322 < 3$$

Here $\beta_2 < 3$ it is platykurtic.

A. The first four moments of the distribution about $x = 2$ are 1, 2.5, 5.5, and 16.

(i) Calculate the four moment about the mean

(ii) about zero

Given $A = 2$

$\mu_1' = 1$, $\mu_2' = 2.5$, $\mu_3' = 5.5$, $\mu_4' = 16$

(i) To find the moments about the mean

$\mu_1 = 0$

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

$= 2.5 - 1$

$= 1.5$

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

$= 5.5 - 3 \times 2.5 \times 1 + 2 \times 1$

$= 5.5 - 7.5 + 2$

$= 0$

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

$= 16 - 4 \times 5.5 \times 1 + 6 \times 2.5 \times 1 - 3 \times 1$

$= 16 - 22 + 15 - 3 = 6$

(ii) To find the moments about zero

$$\mu_r' = \frac{\sum f_i x_i^r}{N}$$

$$\bar{x} = A + \mu_1'$$

$$= 2 + 1$$

$$= 3$$

$$\mu_1' = \frac{\sum f_i x_i}{N} = \bar{x}$$

$$= 3$$

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

$$= 1.5 + 3^2$$

$$= 1.5 + 9$$

$$= 10.5$$

$$\mu_3' = \mu_3 + 3\mu_2\mu_1' + (\mu_1')^3$$

$$= 0 + 3 \times 1.5 \times 3 + 3^3$$

$$= 40.5$$

$$\mu_4' = \mu_4 + 4\mu_3\mu_1' + 6\mu_2(\mu_1')^2 + (\mu_1')^4$$

$$= 6 + 4 \times 0 \times 3 + 6 \times 1.5 \times 9 + 3^4$$

$$= 6 + 6 \times 1.5 \times 9 + 81$$

$$= 168$$

5) The first four moments of the distribution about $x = A$ are $-1.5, 17, -30, 108$ find the first four moment

(i) about mean

(ii) about the origin.

(iii) also calculate β_1 and β_2

Given $A = 4$

$\mu_1' = -1.5, \mu_2' = 17, \mu_3' = -30, \mu_4' = 108$

(i) To find the moments about the mean

$$\mu_1 = 0$$

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

$$= 17 - (-1.5)^2$$

$$= 14.75$$

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

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

$$= -30 + 76.5 - 6.75$$

$$= 39.75$$

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

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

$$= 142.3125$$

(ii) To find the moments about zero

$$\mu_2' = \frac{\sum f_i x_i^2}{N}$$

$$\bar{x} = A + \mu_1'$$

$$= 4 - 1.5$$

$$= 2.5$$

$$\mu_1' = \frac{\sum f_i x_i}{N} = \bar{x}$$

$$= 2.5$$

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

$$= 14.75 + (2.5)^2$$

$= 21$

$\mu_3' = \mu_3 + 3\mu_2\mu_1' + (\mu_1')^3$

$= 39.75 + 3 \times 14.75 \times 2.5 + (2.5)^3$

$= 166$

$\mu_4' = \mu_4 + 4\mu_3\mu_1' + 6\mu_2(\mu_1')^2 + (\mu_1')^4$

$= 142.3125 + 4 \times 39.75 \times 2.5 +$

$6 \times 14.75 \times (2.5)^2 + (2.5)^3$

$= 1132$

(iii) $\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(39.75)^2}{(14.75)^3}$

$= 0.492237 > 0$

Here $\beta_1 > 0$ the the frequency distribution has positive skewness.

$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{142.3125}{(14.75)^2}$

$= 0.654122 < 3$

$\beta_2 < 3$ then it is platy kurtic

6) The first three moments of the distribution about $x=3$ are 2, 10, and 30

(i) Calculate the three moments above mean

(ii) about zero.

Let $A = 3$

$\mu_1' = 2$, $\mu_2' = 10$, $\mu_3' = 30$

(i) To find three moments about the mean

$\mu_1 = 0$

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

$= 10 - (4)$

$= 6$

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

$= 30 - 3 \times 10 \times 2 + 2 \times 8$

$= 30 - 60 + 16$

$= -14$

(ii) about the origin

$$\mu_1' = \frac{\sum f_i x_i}{N}$$

$$= 5$$

$\bar{x} = A + \mu_1'$

$= 3 + 2$

$= 5$

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

$= 6 + 25$

$= 31$

(8)

$$\mu_3' = \mu_3 + 3\mu_2\mu_1' + (\mu_1')^3$$

$= -14 + 3 \times 6 \times 5 + 125$

$= -14 + 90 + 125$

$= 201$

7) the first three moment about the origin are given by $\mu_1' = \frac{1}{2}(n+1)$, $\mu_2' = \frac{1}{6}(n+1)(2n+1)$, $\mu_3' = \frac{1}{4}n(n+1)^2$ examine the skewness of the distribution.

$\mu_1' = \frac{1}{2}(n+1)$, $\mu_2' = \frac{1}{6}(n+1)(2n+1)$

$\mu_3' = \frac{1}{4}n(n+1)^2$

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

$= \frac{1}{6}(n+1)(2n+1) - \left[\frac{1}{2}(n+1)\right]^2$

$= \frac{1}{6}(n+1)(2n+1) - \frac{1}{4}(n+1)^2$

$= \frac{1}{2}(n+1)\left[\frac{1}{3}(2n+1) - \frac{1}{2}(n+1)\right]$

$= \frac{1}{2}(n+1)\left[\frac{2(2n+1) - 3(n+1)}{6}\right]$

$= \frac{1}{12}(n+1)\left[4n+2-3n-3\right]$

$= \frac{1}{12}(n+1)\left[n-1\right]$

$= \frac{1}{12}(n^2-1)$

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

$= \frac{1}{4}n(n+1)^2 - 3 \times \frac{1}{6}(n+1)(2n+1)\frac{1}{2}(n+1) + 2\left[\frac{1}{2}(n+1)\right]^3$

$= \frac{1}{4}n(n+1)^2 - \frac{1}{24}(n+1)^2(2n+1) + \frac{1}{8}(n+1)$

$= \frac{1}{4}(n+1)^2\left[n - \frac{1}{6}(2n+1) + (n+1)\right]$

$= \frac{1}{4}(n+1)^2\left[n - \frac{2n}{3} - \frac{1}{3} + n + 1\right]$

$= \frac{1}{4}(n+1)^2(0)$

$\mu_4 =$

$= 0$

$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{0}{\frac{1}{12}(n^2-1)} = 0$

Here $\beta_1 = 0$ then the distribution is symmetric

8) For a frequency distribution $\frac{fi}{xi}$ show that $\beta_2 \geq 1$

$\beta_2 \geq 1$

T.P $\beta_2 \geq 1$

(ie) T.P $\frac{\mu_4}{\mu_2^2} \geq 1$ $\frac{\mu_4}{\mu_2^2} \geq 1$

$\mu_4 \geq \mu_2^2$ $\mu_4 \geq \mu_2^2$

$$\mu_4 - \mu_2^2 \geq 0$$

now, $\mu_4 - \mu_2^2$

$$\frac{\sum f_i(x_i - \bar{x})^4}{N} - \left[\frac{\sum f_i(x_i - \bar{x})^2}{N}\right]^2$$

$$\frac{\sum f_i\left[(x_i - \bar{x})^2\right]^2}{N} - \left[\frac{\sum f_i (x_i - \bar{x})^2}{N}\right]^2$$

$$= \frac{\sum f_i z_i^2}{N} - \left[\frac{\sum f_i z_i}{N}\right]^2$$

where $(x_i - \bar{x})^2 = z_i$

$$= \sigma_{z_i}^2$$

$$\sigma_x = \sqrt{\frac{\sum fx^2}{N} - \left(\frac{\sum fx}{N}\right)^2}$$

≥ 0

$\therefore \mu_4 - \mu_2^2 \geq 0$

Hence $\beta_2 \geq 1$

9) Calculate the first four moments about $x = 4$ and Hence find the moments about the main of the following distribution also find

x	0	1	2	3	4	5	6	7	8	9	10
f	5	10	30	70	140	200	140	70	30	10	5

10) The first four moments of a distribution about $x = 4$ are 1, 4, 10, 45 respectively calculate the moments about the mean

$A = 4$

10) $\mu_1' = 1$, $\mu_2' = 4$, $\mu_3' = 10$, $\mu_4' = 45$

(b) To find the four moments about the mean:

$\mu_1 = 0$

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

$= 4 - (1)^2$

$= 4 - 1$

$= 3$

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

$$= 10 - 3\times 4\times 1 + 2\times 1^3$$

$$= 10 - 12 + 2$$

$$= 0$$

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

$$= 45 - 4\times 10\times 1 + 6\times 4\times 1 - 3\times 1$$

$$= 45 - 40 + 24 - 3$$

$$= 26$$

$A = 4$

9)

x	f	x − A	$f_i(x_i - A)$	$f_i(x_i - A)^2$	$f_i(x_i - A)^3$	$f_i(x_i - A)^4$
0	5	−4	−20	80	−320	1280
1	10	−3	−30	90	−270	810
2	30	−2	−60	120	−240	480
3	70	−1	−70	70	−70	70
4	140	0	0	0	0	0
5	200	1	200	200	200	200
6	140	2	280	560	1120	2240
7	70	3	210	630	1890	5670
8	30	4	120	480	1920	7680
9	10	5	50	250	1250	6250
10	5	6	30	180	1080	6480

$N = 710$

$$\mu_r' = \frac{\sum f_i (x_i - A)^r}{N}$$

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

$$\mu_1' = \frac{710}{710}$$

$$\boxed{\mu_1' = 1}$$

$$\mu_2' = \frac{2660}{710} = 3.75$$

$$\boxed{\mu_2' = 3.75}.$$

$$\mu_3' = \frac{6560}{710}$$

$$\boxed{\mu_3' = 9.24}.$$

$$\mu_4' = \frac{31160}{710}$$

$$\boxed{\mu_4' = 43.89}.$$

$$\bar{x} = \frac{\sum f_i x_i}{N}$$

$$N = \sum f_i$$

$\beta_1 = 0$ Symmetric

$\beta_1 > 0$ positive

$\beta_1 < 0$ Negative

$\beta_2 = 0$ mesokurtic

$\beta_2 < 0$ platy

$\beta_2 > 0$ lepto

$\mu_1 = 0$

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

$= 3.75 - (1)^2$

$\boxed{\mu_2 = 2.75}$

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

$= 9.24 - 3(3.75)(1) + 2(1)^3$

$= 9.25 - 11.25 + 2$

$\boxed{\mu_3 = -0.01}$

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

$= 43.89 - 4(9.74)(1) + 6(3.75)(1) - 3(1)$

$= 43.89 - 36.96 + 22.5 - 3$

$\boxed{\mu_4 = 26.43}$

$\beta_1 = \frac{M_3^2}{M_2^3}$

$= \frac{(-0.01)^2}{(2.75)^3}$

$= \frac{0.0001}{20.796875}$

$= 0.000004 > 0$ It has positive skewness.

$\beta_2 = \frac{M_4}{M_2^2}$

$= \frac{26.43}{(2.75)^2}$

$= \frac{26.43}{20.7961^2}$

$= 3.4993$ Lepto Kurtic

Curve fitting

Let x_i $i = 1, 2, \ldots, n$ be the values of independent variable and y_i $i = 1, 2 \ldots n$ be the corresponding values of dependent variable.

If the points (x_i, y_i) $i = 1, 2 \ldots n$ are plotted on a graph paper and be uptied a diagram called scatter diagram.

If there is a functional realation ship between x_i and y_i the points of the scatter diagram will be found to be consentrated round a sertern curve.

The process of finding such the functional realation ship between the variables is called curve fitting

Example:

The lines of regression can be got by fitting a linear curve to a given bi variate distribution.

Principle of least squares:

Let (x_i, y_i) $i = 1, 2, \ldots n$ be the observed set of values of the variable (x, y) Let $y = f(x)$ be a functional relationship between the variables (x, y) Then $d_i = y_i - f(x_i)$ which is the difference between the observed values of y and the value of y determine by the functional relation is called the residuals. The principle of least squares states that the parameters involved in $f(x)$ should be chosen in such a way that

$\sum d_i^2$ is minimum

Fitting a straight line:

Consider the fitting of a straight line $y = ax + b$ to the values (x_i, y_i) where $i = 1, 2, \ldots n$ the residual d_i is given by

$$d_i = y_i - f(x_i)$$

$$d_i^2 = [y_i - (ax_i + b)]^2$$

$$d_i^2 = [y_i - ax_i - b]^2$$

$$\sum d_i^2 = \sum [y_i - ax_i - b]^2 = R(\text{say})$$

according to the priciple of least square we have to determine the parameters a, b, so that R is minimum.

$$\frac{\partial R}{\partial a} = 0 \Rightarrow \frac{\partial}{\partial a}\left[\Sigma (y_i - ax_i - b)\right]^2 = 0$$

$$\Rightarrow 2\Sigma (y_i - ax_i - b)(-x_i) = 0$$

$$\Rightarrow -2\Sigma (y_i - ax_i - b)(x_i) = 0$$

$$\Rightarrow \Sigma (y_i x_i - ax_i^2 - bx_i) = 0$$

$$\Rightarrow \Sigma x_i y_i - a\Sigma x_i^2 - b\Sigma x_i = 0$$

$$\Rightarrow \Sigma x_i y_i = a\Sigma x_i^2 + b\Sigma x_i \longrightarrow (1)$$

$$\frac{\partial R}{\partial b} = 0 \Rightarrow \frac{\partial}{\partial b}\left[\Sigma (y_i - ax_i - b)\right]^2 = 0$$

$$\Rightarrow 2\Sigma (y_i - ax_i - b)(-1) = 0$$

$$\Rightarrow -2\Sigma (y_i - ax_i - b) = 0$$

$$\Rightarrow \Sigma (y_i - ax_i - b) = 0$$

$$\Rightarrow \Sigma y_i - a\Sigma x_i - \Sigma b = 0$$

$$\Rightarrow \Sigma y_i - a\Sigma x_i - nb = 0$$

$$\Rightarrow \Sigma y_i = a\Sigma x_i + nb \longrightarrow (2)$$

equation ① and ② are called normal equation. from these equations we have find a and b.

Fitting a second degree parabola :-

consider the fitting of the second degree parabola $y = ax^2 + bx + c$ to the values (x_i, y_i) $i = 1, \ldots, n$. The residual d_i given by $d_i = y_i - (ax_i^2 + bx_i + c)$

$$d_i = y_i - (ax_i^2 + bx_i + c)$$

$$d_i^2 = (y_i - ax_i^2 - bx_i - c)^2$$

$$\sum d_i^2 = \sum (y_i - ax_i^2 - bx_i - c)^2$$

According to the principle of least square we have to determine the parameters a, b, c so that R is minimum

$$\frac{\partial R}{\partial a} = 0 \rightarrow \frac{\partial}{\partial a}\left[\sum (y_i - ax_i^2 - bx_i - c)^2\right] = 0$$

$$\Rightarrow 2\sum (y_i - ax_i^2 - bx_i - c)(-x_i^2) = 0$$

$$\Rightarrow -2\sum (y_i - ax_i^2 - bx_i - c)(x_i^2) = 0$$

$\Rightarrow \sum (y_i - ax_i^2 - bx_i - c)(x_i^2) = 0$

$\Rightarrow \sum x^2 y_i - a\sum x_i^4 - b\sum x_i^3 - c\sum x_i^2 = 0$

$\Rightarrow \sum x_i^2 y_i = a\sum x_i^4 + b\sum x_i^3 + c\sum x_i^2 \rightarrow (1)$

$\frac{\partial R}{\partial b} = 0 \Rightarrow \frac{\partial}{\partial b}\left[\sum (y_i - ax_i^2 - bx_i - c)^2\right] = 0$

$\Rightarrow 2\sum (y_i - ax_i^2 - bx_i - c)(-x_i) = 0$

$\Rightarrow -2\sum (y_i - ax_i^2 - bx_i - c)(x_i) = 0$

$\Rightarrow \sum (y_i - ax_i^2 - bx_i - c)(x_i) = 0$

$\Rightarrow \sum x_i y_i - a\sum x_i^3 - b\sum x_i^2 - c\sum x_i = 0$

$\Rightarrow \sum x_i y_i = a\sum x_i^3 + b\sum x_i^2 + c\sum x_i = 0 \rightarrow (2)$

$\frac{\partial R}{\partial c} = 0 \Rightarrow \frac{\partial}{\partial c}\left[\sum (y_i - ax_i^2 - bx_i - c)^2\right] = 0 \rightarrow (3)$

$\Rightarrow 2\sum (y_i - ax_i^2 - bx_i - c)(-1) = 0$

$\Rightarrow \sum (y_i - ax_i^2 - bx_i - c) = 0$

$\Rightarrow \sum y_i - a\sum x_i^2 - b\sum x_i - nc = 0$

$\Rightarrow \sum y_i = a\sum x_i^2 - b\sum x_i - nc = 0$

$\Rightarrow \Sigma y_i = a\Sigma x_i^2 + b\Sigma x_i + nc \rightarrow ③$

equation ①, ②, ③ called normal equation from these equation from a, b, c.

1) fit a straight line to the following data.

x	0	1	2	3	4
y	2.1	3.5	5.4	7.3	8.2

Soln:-

Let us fit a straight line to the given data.

$y = ax + b \rightarrow ①$

We have to determine parameters a, b by using normal equations.

$\Sigma x_i y_i = a\Sigma x_i^2 + b\Sigma x_i \rightarrow ②$

$\Sigma y_i = a\Sigma x_i + nb \rightarrow ③$

x_i	y_i	$x_i y_i$	x_i^2
0	2.1	0	0
1	3.5	3.5	1
2	5.4	10.8	4
3	7.3	21.9	9
4	8.2	32.8	16
Σx_i	Σy_i	$\Sigma x_i y_i$	$\Sigma x_i^2 = 30$
$= 10$	26.5	$= 69$	

Sub these values in (1) & (2)

$$30a + 10b = 69 \longrightarrow (4)$$

$$10a + 5b = 26.5 \longrightarrow (5)$$

$(4) \Rightarrow 30a + 10b = 69$

$(5) \times 3 \Rightarrow 30a + 15b = 79.5 \longrightarrow (6)$

$(4) - (6) \Rightarrow -5b = -10.5$

$$b = \frac{+10.5}{+5}$$

$$b = +2.1$$

$$b = 2.1$$

$b = 2.1$ sub (4)

$30a + 10 \times 2.1 = 69$

$30a + 21 = 69$

$30a = 69 - 21$

$30a = 48$

$a = \frac{48}{30}$

$a = 1.6$

The straight line fitted for the given data is $y = 1.6x + 2.1$

2) fit a straight line $y = a + bx$ to the following data.

x	0	1	2	3	4
y	1	1.8	3.3	4.5	6.3

Let us fit a straigth line to the given data

$y = bx + a \rightarrow (1)$

We have to determine parameters a, b by using normal equations.

$\Sigma x_i y_i = a\Sigma x_i^2 + b\Sigma x_i \rightarrow (2)$

$\Sigma y_i = a\Sigma x_i + nb \rightarrow (3)$

x_i	y_i	$x_i y_i$	x_i^2
0	1	0	0
1	1.8	1.8	1
2	3.3	6.6	4
3	4.5	13.5	9
4	6.3	25.2	16
$\Sigma x_i = 10$	$\Sigma y_i = 16.9$	$\Sigma x_i y_i = 47.1$	$\Sigma x_i^2 = 30$

$30a + 10b = 47.1 \rightarrow (4)$

$10a + 5b = 16.9 \rightarrow (5)$

$(5) \Rightarrow 30a + 10b = 47.1$

$(5) \times 2 \rightarrow 30a + 10b = 33.8$

$10a = 13.3$

$a = \frac{13.3}{10}$

$\boxed{a = 1.33}$

$a = 1.33$ sub (4)

$30 \times 1.33 + 10b = 47.1$

$39.9 + 10b = 47.1$

$10b = 47.1 - 39.9$

$10b = 7.2$

$b = 7.2/10$

$b = 0.72$

The straight line fitted for the given data is

$y = 0.72x + 1.33$

3) fit a straight line to the following data and estimat the value of y corresponding to $x = 6$

x	0	5	10	15	20	25
y	12	15	17	22	24	30

Let us fit a straight line to the given data is

$$y = ax + b \longrightarrow (1)$$

we have to determine the parameters a and b by using the normal equations

$$\sum x_i y_i = a\sum x_i^2 + b\sum x_i \longrightarrow (2)$$

$$\sum y_i = a\sum x_i + nb \longrightarrow (3)$$

x_i	y_i	$x_i y_i$	x_i^2
0	12	0	0
5	15	75	25
10	17	170	100
15	22	330	225

20	24	480	400
25	30	750	625
$\sum x_i$ = 75	$\sum y_i$ = 120	$\sum x_i y_i$ = 1805	$\sum x_i^2$ 1375

Sub the values ① & ②

$$1375a + 75b = 1805 \rightarrow ④$$

$$75a + 6b = 120 \rightarrow ⑤$$

④ × 6 ⇒ $8250a + 450b = 10830 \rightarrow ⑦$

⑤ × 75 ⇒ $5625a + 450b = 9000 \rightarrow ⑧$

$$2625a = 1830$$

$$a = \frac{1830}{2625}$$

$$a = 0.6971$$

Sub $a = 0.6971$ ④

$$75 \times 0.6971 + 6b = 120$$

$$52.2825 + 6b = 120$$

$$6b = 120 - 52.2825$$

$$= 67.7175$$

$$b = \frac{67.7175}{6}$$

$$b = 11.2865$$

$$y = 0.6971x + 11.2865$$

when $x = 6$

$$y = 15.4691$$

1) fit a second degree parabola by taking xi as a indipendent variable

x	0	1	2	3	4
y	1	5	10	22	38

Soln:-

Let the second degree parabola to be fitted to the given data is

$$y = ax^2 + bx + c \longrightarrow (1)$$

we have to determine the parameters a, b, c by using the normal equations.

xi
0
1
2
3
4

$\Sigma xi = 10$

$\sum x_i^2 y_i = a\sum x_i^4 + b\sum x_i^3 + c\sum x_i^2 \rightarrow (1)$

$\sum x_i y_i = a\sum x_i^3 + b\sum x_i^2 + c\sum x_i \rightarrow (2)$

$\sum y_i = a\sum x_i^2 + b\sum x_i + nc \rightarrow (3)$

x_i	y_i	$x_i^2 y_i$	$x_i y_i$	x_i^2	x_i^3	x_i^4
0	1	0	0	0	0	0
1	5	5	5	1	1	1
2	10	40	20	4	8	16
3	22	198	66	9	27	81
4	38	608	152	16	64	256
$\sum x_i$ = 10	$\sum y_i$ = 76	$\sum x_i^2 y_i$ = 851	$\sum x_i y_i$ = 243	$\sum x_i^2$ = 30	$\sum x_i^3$ = 100	$\sum x_i^4$ = 354

sub the values (1), (2), and (3)

$354a + 100b + 30c = 851 \rightarrow (4)$

$100a + 30b + 10c = 243 \rightarrow (5)$

$30a + 10b + 5c = 76 \rightarrow (6)$

(5) $\Rightarrow 100a + 30b + 10c = 243 \rightarrow (7)$

(6) $\times 2 \Rightarrow 60a + 20b + 10c = 152 \rightarrow (8)$

(7) − (8): $40a + 10b = 91 \rightarrow (9)$

$354a + 100b + 30c = 851 \rightarrow (7)$

$(7) \times 3 \rightarrow 300a + 90b + 30c = 729 \rightarrow (8)$

$(7) - (8)$

$54a + 10b = 122 \rightarrow (9)$

$40a + 10b = 91 \rightarrow (9)$

$54a + 10b = 122 \rightarrow (10)$

$(10) - (9)$

$-14a = -31$

$a = \frac{-31}{-14}$

$a = 2.214$

$a = 2.071$ sub (6)

$40 \times 2.214 + 10 \times b = 91$

$88.56 + 10b = 91$

$10b = 91 - 88.56$

$10b = 2.44$

$b = \frac{2.44}{10}$

$b = 0.244$

$b = \frac{10.16}{10}$

$b = 1.016$

a, b value sub in (6) (8)

$30 \times 2.214 + 10 \times 1.016 + 5c = 75$

$72.29 + 5c = 75$

$5c = 75 - 72.29$

$5c = 2.71$

$$c = \frac{2.71}{5}$$

$$c = 0.542$$

The Straight line fitted for the given data is

$$y = 2.071x^2 + 1.016x + 0.542$$

a, b value sub in (8)

$$30 \times 2.214 + 10 \times 0.244 + 5c = 76$$

$$66.42 + 2.44 + 5c = 76$$

$$68.86 + 5c = 76$$

$$5c = 76 - 68.86$$

$$5c = 7.14$$

$$c = \frac{7.14}{5}$$

$$\boxed{c = 1.428}$$

the straight line fitted for the given data is

$$y = 2.214x^2 + 0.244x + 1.428.$$

1) Fit a straight line to the following data regarding x as the independing variable.

Years x	1911	1921	1931	1941	1951
Production y	10	12	8	10	12

Soln:

Let the straight line fitted to the given data

$$y = ax + b$$

$$u = \frac{x - A}{c}$$

Put $u = \frac{x - 1931}{10}$, $v = y - 10$

The straight line fitted to the given data is $v = au + b \rightarrow (1)$

we have to determine the parameters a and b by using normal equations

$$\Sigma u_i v_i = a\Sigma u_i^2 + b\Sigma u_i \rightarrow (2)$$

$$\Sigma v_i = a\Sigma u_i + nb \rightarrow (3)$$

x_i	y_i	$u_i = \frac{x_i - 1931}{10}$	$v_i = y_i - 10$	u_i^2	$u_i v_i$
1911	10	-2	0	4	0
1921	12	-1	2	1	-2
1931	8	0	-2	0	0
1941	10	1	0	1	0
1951	14	2	4	4	8
		$\sum u_i = 0$	$\sum v_i = 4$	$\sum u_i^2 = 10$	$\sum u_i v_i = 6$

Sub these values in ② & ③

$$10a + b(0) = 6$$

$$a = \frac{6}{10} = 0.6$$

$$a(0) + 5(b) = 4$$

$$5b = 4$$

$$b = 4/5 = 0.8$$

① $\Rightarrow$ $v = 0.6u + 0.8$

$$y - 10 = 0.6\left(\frac{x - 1931}{10}\right) + 0.8$$

$$y - 10 = 0.06\,(x - 1931) + 0.8$$

$$y - 10 = 0.06x - 115.86 + 0.8$$

$$y = 0.06x - 115.86 + 0.8 + 10$$

$$y = 0.06x - 105.06$$

2) Fit the curve $y = bx^a$ to the following data

x	1	2	3	4	5	6
y	1200	900	600	200	110	50

The given curve is $y = bx^a$

Taking log

$$\log y = \log bx^a$$

$$\log y = \log b + \log x^a$$

$$\log y = \log b + a \log x$$

$$\log y = a \log x + \log b$$

If is of the form $Y = AX + B \rightarrow (1)$

where $Y = \log y$, $A = a$, $X = \log x$,

$B = \log b$

we have to determine parameters A and B by using normal equations

$$\Sigma x_i y_i = A \Sigma x_i^2 + B \Sigma x_i \rightarrow (2)$$

$$\Sigma y_i = A \Sigma x_i + nB \rightarrow (3)$$

x_i	y_i	$X_i = \log x$	$Y_i = \log y$	x_i^2	$x_i Y_i$
			3.0791	0	0
1	1200	0			
			2.9542	0.0906	0.8894
2	900	0.3010			
			2.7781	0.2296	1.3254
3	600	0.4771			
			2.3010	0.3624	1.3852
4	200	0.6020			
			2.0413	0.4884	1.4266
5	110	0.6989			
			1.6989	0.6054	1.3219
6	50	0.7781			
		2.8571	14.8526	1.7214	6.3485
		2.86	14.85	1.77	6.35

$1.77A + 2.86B = 6.35 \rightarrow (4)$

$2.86A + 6B = 14.85 \rightarrow (5)$

$(4) \times 6 \Rightarrow \quad 10.62A + 17.16B = 38.10$

$(5) \times 2.86 \Rightarrow \quad 8.18A + 17.16B = 42.47$

$2.44A = -4.4$

$A = \frac{-4.4}{2.44}$

$= -1.8032$

$= -1.80$

Sub $A = -1.80$ in (5)

$2.86(-1.80) + 6B = 14.85$

$6B = 14.85 + 5.148$

$6B = 19.998$

$B = \frac{19.998}{6}$

$B = 3.33$

$Y = AX + B$

$Y = -1.80X + 3.33$

$A = a = -1.80$

$B = \log b = 3.333$

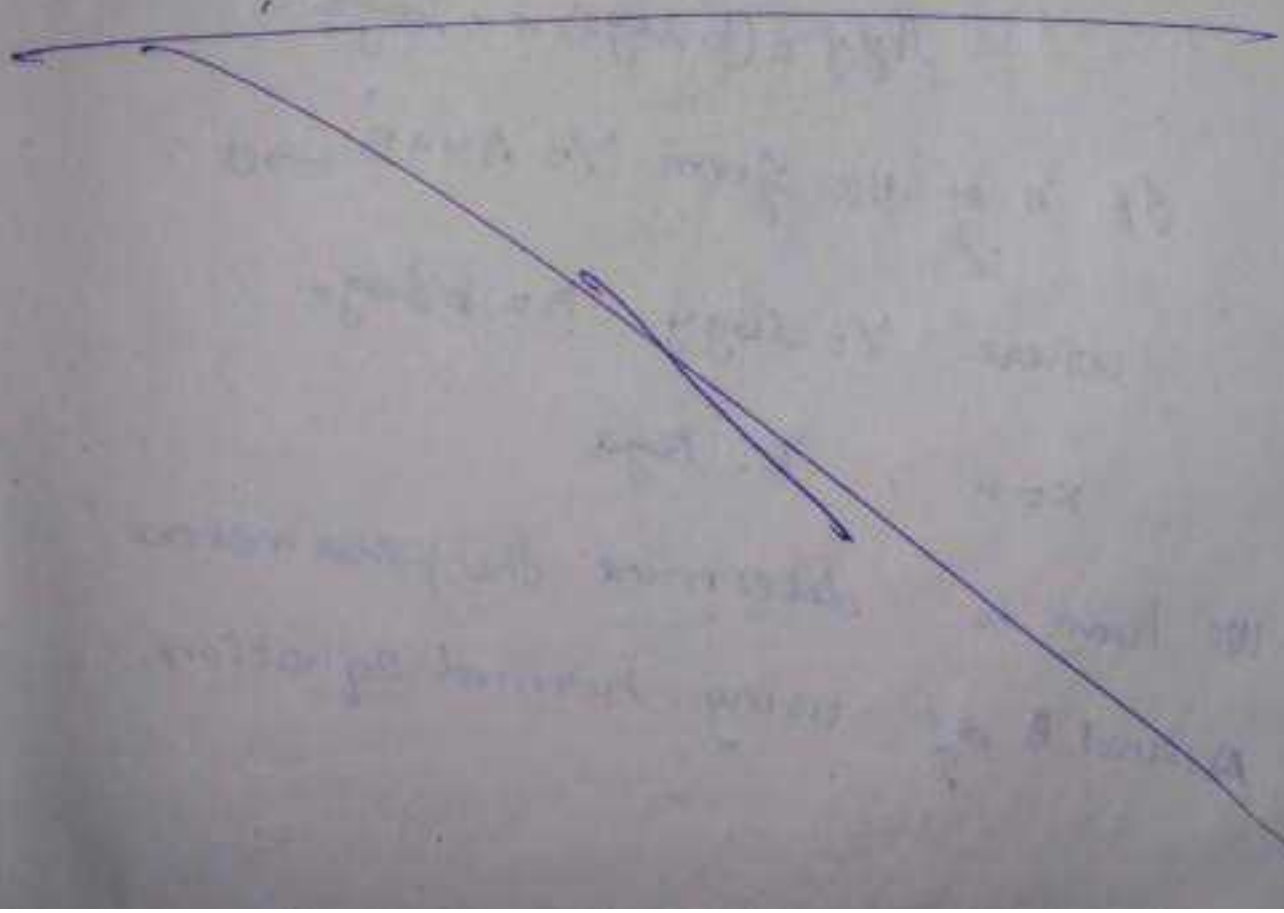
$b = \text{anti}\log(3.333)$

$= 2152.78$

The given curve is

$$y = bx^{a}$$

$$y = (2152.78)x^{-1.80}$$ Ans.



3) Explain the method of fitting the curve good fit $y = a e^{bx}$ $(a > 0)$

Taking log

$$\log y = \log a e^{bx}$$

$$\log y = \log a + \log e^{bx}$$

$$\log y = \log a + bx \log e$$

$$\log y = (b \log e) x + \log a$$

It is of the form $Y = Ax + B \rightarrow (1)$

where $Y = \log y$ $A = b \log e$

$x = x$, $B = \log a$

We have to determine the parameters A and B by using normal equation.

$$\Sigma x_i y_i = A\Sigma x_i^2 + B\Sigma x_i \rightarrow (2)$$

$$\Sigma y_i = A\Sigma x_i + nB \rightarrow (3)$$

From the two normal equations we get the values of A & B and a & b can be obtained from $B = \log a$

$$a = \text{antilog } B$$

$$A = b \log e$$

$$b = \frac{A}{\log e}$$

4) Explain the method of fitting the curve

$y = ka^{bx}$

Taking log,

$$\log y = \log ka^{bx}$$

$$\log y = \log k + \log a^{bx}$$

$$\log y = \log k + bx \log a$$

$$\log y = b \log a + \log k$$

$$\log y = (b \log a)x + \log k$$

It is from $Y = Ax + B \rightarrow (1)$

where $Y = \log y$, $A = b \log a$,

$$B = \log k$$

we have to determine the parameters A and B by using normal equations

$$\sum x_i Y_i = A \sum x_i^2 + B \sum x_i \rightarrow (2)$$

$$\sum y_i = A \sum x_i + nB \rightarrow (3)$$

a and b

From the two normal Equations we get the values of A & B and a & b can be obtained from

$$B = \log k$$

$$A = b \log a$$

$$b = \frac{A}{\log a}$$

$$b \log a = A$$

$$\log a = \frac{A}{b}$$

$$a = \text{anti log}\,(A/b)$$

Q. Fit a curve of the form $y = ab^x$ to the following data

Years (x)	1951	1952	1953	1954	1955	1956	1957
Production in tons (y)	201	263	314	395	427	504	612

Soln:

The given curve is $y = ab^x$

Taking log,

$$\log y = \log a + \log b^x$$

$$\log y = \log a + x \log b$$

$$\log y = x \log b + \log a$$

This is of the form $Y = AX + B \rightarrow (1)$

where $Y = \log y$, $A = \log b$, $B = \log a$

Put $X = x - 1954$

We have to determine the parameters A and B by using normal equations.

$$\Sigma x_i Y_i = A \Sigma x_i^2 + B \Sigma x_i \rightarrow (2)$$

$$\Sigma Y_i = A \Sigma x_i + nB \rightarrow (3)$$

x	y	$x_i = x - 1954$	$Y_i = \log y$	x_i^2	$x_i Y_i$
1951	201	-3	2.303	9	-6.909
1952	263	-2	2.419	4	-4.838
1953	314	-1	2.496	1	-2.496
1954	395	0	2.596	0	0
1955	427	1	2.63	1	2.63
1956	504	2	2.702	4	5.404
1956	612	3	2.786	9	8.358
		$\Sigma x_i = 0$	$\Sigma Y = 17.932$	$\Sigma x_i^2 = 28$	$\Sigma x_i Y_i = 2.149$

Sub these values in ③ & ④

$28A + 0 = 2.149$

$0 + 7B = 17.932$

$B = \frac{17.932}{7}$

$\boxed{B = 2.561}$

$28A = 2.149$

$A = \frac{2.149}{28}$

$A = 0.076$

$A = \log b$

$\log b = A$

$b = \text{antilog}(A)$

$= \text{antilog}(0.076)$

$b = 1.191$

$B = \log a$

$a = \text{antilog}(B)$

$= \text{antilog}(2.561)$

$= 363.9$

The required curve

$$y = a b^x$$

$$y = (363.97)(1.1911)^{x-1954}$$

2) Fit the exponential curve $y = ae^{bx}$ to the following data

x	0	2	4
y	5.02	10	31.62

The given curve is $y = ae^{bx}$

Taking log,

$$\log y = \log a + \log e^{bx}$$

$$= \log a + bx \log e$$

$$= \log a + (b \log e)x$$

$$\log y = (b \log e)x + \log a$$

This is of the form $Y = Ax + B \rightarrow (1)$

where $Y = \log y$, $A = b \log e$,

$B = \log a$

Now we have to determine the parameters A and B by using normal equations.

$$\Sigma xy = A\Sigma x^2 + B\Sigma x \rightarrow (3)$$

$$\Sigma y = A\Sigma x + nB \rightarrow (4)$$

x	y	$Y_i = \log y$	x_i^2	$x_i Y_i$
0	5.02	0.69	0	0
2	10	1	4	2
4	31.62	1.49	16	5.96
6		$\Sigma Y_i = 3.19$	$\Sigma x_i^2 = 20$	$\Sigma x_i Y_i = 7.96$

Sub these values in (3) & (4)

$$7.96 = 20A + 6B \rightarrow (5)$$

$$3.19 = 6A + 3B \rightarrow (6)$$

(5) $\Rightarrow$ $20A + 6B = 7.96$

(6) $\times 2 \Rightarrow$ $12A + 6B = 6.38$

$$8A = 1.58$$

$$A = 0.1975$$

$$A = 0.20$$

Sub in (6)

$$3.19 = 0.20 \times 6 + 3B$$

$$3.19 = 1.2 + 3B$$

$$3B = 1.99$$

$$B = 0.66$$

$a = \text{antilog}(B)$

$= \text{antilog}(0.66)$

$= 4.57$

$b = \frac{A}{\log e}$

$= \frac{0.20}{0.43}$

$= 0.465$

$y = 4.57\, e^{0.465x}$