

MAR GREGORIOS COLLEGE OF ARTS & SCIENCE

Block No.8, College Road, Mogappair West, Chennai – 37

Affiliated to the University of Madras
Approved by the Government of Tamil Nadu
An ISO 9001:2015 Certified Institution



DEPARTMENT OF COMPUTER SCIENCE

SUBJECT NAME: ALLIED STATISTICS I

SUBJECT CODE: SBA0C

SEMESTER: II

PREPARED BY: PROF. D. SELVARAJ

Allied - Paper I - Applied Statistics I (SBAOC)

Note:

The emphasis is solely upon the applicational understanding and practice of statistical methods, with specific reference to problems in earth sciences.

UNIT - 1:

Nature and scope of statistical methods and their limitations and their applications in Geography - spatial data and statistical methods classification, tabulation and diagrammatic representation of various type of statistical data - frequency curves and Ogives - Graphical; determination of percentiles, quartiles and their uses, Lorenz curve.

UNIT - 2:

Measures of location - Arithmetic mean, median, mode, Geometric mean, Harmonic mean and their properties - merits and demerits.

UNIT - 3:

Measures of dispersion - Range, mean deviation, quartile deviation, standard deviation, coefficient of variation, skewness and kurtosis - and their properties.

UNIT - 4:

Probability of an event- Finitely additive probability space addition and multiplication theorems - Independence of events - conditional probability - Bayes' theorem - simple problems.

UNIT - 5:

Concepts of random variable - Distribution function - Mathematical Expectation - Moments of random variable - Moment generating function - simple problem.

Books for Study:

Wonnacott, R.J. & Wonnacott, T. H. (1985): Introductory Statistics. 4th edition John Wiley & Sons.

David Ebdon (1977): Statistics in Geography - A practical approach Basil Blackwell, Oxford.

Gregory. S. (1964): Statistical Methods and Geographer, Longman, London.

Books for Reference:

Snedecor, G.W., & Cochran, W.G.: Statistical Methods, Oxford and IBH.

Burr, I.W.: Applied Statistical Methods, Academic Press.

Aslam Mahmood and Moonis Raza, (1977): Statistical methods in Geographical studies. Rajesh publications, New Delhi.

Hammond.R. and Mc. Cullagh.P. (1974): Quantitative Techniques in Geography; An introduction, Clarendon Press, Oxford.

Science in Geography Series: (I-IV) Oxford University Press, London.

(I) Development of Geographical methods

(II) Data Collection.

(III) Data Description & Presentation

(IV) Data use and interpretation.

ALLIED STATISTICS I

UNIT I

NATURE, SCOPE AND LIMITATIONS OF STATISTICS

Introduction

The term “statistics” is used in two senses : first in plural sense meaning a collection of numerical facts or estimates—the figure themselves. It is in this sense that the public usually think of statistics, e.g., figures relating to population, profits of different units in an industry etc. Secondly, as a singular noun, the term ‘statistics’ denotes the various methods adopted for the collection, analysis and interpretation of the facts numerically represented. In singular sense, the term

‘statistics’ is better described as statistical methods. In our study of the subject, we shall be more concerned with the second meaning of the word ‘statistics’.

Definition

Statistics has been defined differently by different authors and each author has assigned new limits to the field which should be included in its scope. We can do

no better than give selected definitions of statistics by some authors and then come to the conclusion about the scope of the subject. A.L. Bowley defines, “Statistics

may be called the science of counting”. At another place he defines, “Statistics may be called the science of averages”. Both these definitions are narrow and throw light only on one aspect of Statistics. According to King, “The science of

statistics is the method of judging collective, natural or social, phenomenon from the results obtained from the analysis or enumeration or collection of estimates”.

Many a time counting is not possible and estimates are required to be made. Therefore, Boddington defines it as “the science of estimates and probabilities”.

But this definition also does not cover the entire scope of statistics. The statistical methods are methods for the collection, analysis and interpretation of numerical data and form a basis for the analysis and comparison of the observed phenomena.

In the words of Croxton&Cowden, “Statistics may be defined as the collection, presentation, analysis and interpretation of numerical data”. Horace Secrist has given an exhaustive definition of the term statistics in the plural sense. According to him: “By statistics we mean aggregates of facts affected to a marked extent by a multiplicity of causes numerically expressed, enumerated or estimated according to reasonable standards of accuracy collected in a systematic manner for a pre-determined purpose and placed in relation to each other”.

This definition makes it quite clear that as numerical statement of facts, ‘statistic’ should possess the following characteristics.

1. Statistics are aggregate of facts

A single age of 20 or 30 years is not statistics, a series of ages are. Similarly, a single figure relating to production, sales, birth, death etc., would not be statistics although aggregates of such figures would be statistics because of their comparability and relationship.

2. Statistics are affected to a marked extent by a multiplicity of causes A number of causes affect statistics in a particular field of enquiry, e.g., in

production statistics are affected by climate, soil, fertility, availability of raw materials and methods of quick transport.

3. Statistics are numerically expressed, enumerated or estimated. The subject of statistics is concerned essentially with facts expressed in numerical form—with their quantitative details but not qualitative descriptions. Therefore, facts indicated by terms such as 'good', 'poor' are not statistics unless a numerical equivalent, is assigned to each expression. Also this may either be enumerated or estimated, where actual enumeration is either not possible or is very difficult.

4. Statistics are enumerated or estimated according to reasonable standard of accuracy. Personal bias and prejudices of the enumeration should not enter into the counting or estimation of figures, otherwise conclusions from the figures would not be accurate. The figures should be counted or estimated according to reasonable

standards of accuracy. Absolute accuracy is neither necessary nor sometimes possible in social sciences. But whatever standard of accuracy is once

adopted, should be used throughout the process of collection or estimation.

5. Statistics should be collected in a systematic manner for a predetermined purpose. The statistical methods to be applied on the purpose of enquiry since figures are

always collected with some purpose. If there is no predetermined purpose, all the efforts in collecting the figures may prove to be wasteful. The purpose of a series

of ages of husbands and wives may be to find whether young husbands have young wives and the old husbands have old wives.

6. Statistics should be capable of being placed in relation to each other. The collected figure should be comparable and well-connected in the same department of inquiry. Ages of husbands are to be compared only with the corresponding ages of wives, and not with, say, heights of trees.

Functions of Statistics

The functions of statistics may be enumerated as follows :

(i) To present facts in a definite form : Without a statistical study our ideas are likely to be vague, indefinite and hazy, but figures help as to represent things in their true perspective. For example, the statement that some students out of 1,400 who had appeared, for a certain examination, were

declared successful would not give as much information as the one that 300 students out of 400 who took the examination were declared successful.

(ii) To simplify unwieldy and complex data : It is not easy to treat large numbers and hence they are simplified either by taking a few figures to serve as a representative sample or by taking average to give a bird's eye view of the large

masses. For example, complex data may be simplified by presenting them in the form of a table, graph or diagram, or representing it through an average etc.

(iii) To use it as a technique for making comparisons : The significance of certain figures can be better appreciated when they are compared with others of the same type. The comparison between two different groups is best represented by certain statistical methods, such as average, coefficients, rates, ratios, etc.

(iv) To enlarge individual experience : An individual's knowledge is limited to what he can observe and see; and that is a very small part of the social organism.

His knowledge is extended in various ways by studying certain conclusions and results, the basis of which are numerical investigations. For example, we all have a general impression that the cost of living has increased. But to know to what extent the increase has occurred, and how far the rise in prices has affected different income groups, it would be necessary to ascertain the rise in prices of articles consumed by them.

(v) To provide guidance in the formulation of policies : The purpose of statistics is to enable correct decisions, whether they are taken by a businessman or Government. In fact statistics is a great servant of business in management,

governance and development. Sampling methods are employed in industry in tackling the problem of standardisation of products. Big business houses maintain a

separate department for statistical intelligence, the work of which is to collect, compare and coordinate figures for formulating future policies of the firm regarding production and sales.

(vi) To enable measurement of the magnitude of a phenomenon : But for the development of the statistical science, it would not be possible to estimate the

population of a country or to know the quantity of wheat, rice and other agricultural commodities produced in the country during

Importance of Statistics
 These days statistical methods are applicable everywhere. There is no field of work in which statistical methods are not applied. According to A L. Bowley, ‘A

knowledge of statistics is like a knowledge of foreign languages or of Algebra, it may prove of use at any time under any circumstances”. The importance of the statistical science is increasing in almost all spheres of knowledge, e g., astronomy,

biology, meteorology, demography, economics and mathematics. Economic

planning without statistics is bound to be baseless. Statistics serve in administration, and facilitate the work of formulation of new policies. Financial institutions and investors utilise statistical data to summaries the past experience.

Statistics are also helpful to an auditor, when he uses sampling techniques or test checking to audit the accounts of his client.

LIMITATIONS OF STATISTICS

The scope of the science of statistic is restricted by certain limitations :

1. The use of statistics is limited numerical studies: Statistical methods cannot be applied to study the nature of all type of phenomena. Statistics deal with only such phenomena as are capable of being quantitatively measured and numerically expressed. For, example, the health, poverty and intelligence of a group of individuals, cannot be quantitatively measured, and thus are not suitable subjects for statistical study.
2. Statistical methods deal with population or aggregate of individuals rather than with individuals. When we say that the average height of an Indian is 1 metre 80 centimetres, it shows the height not of an individual but as found by the study of all individuals.
3. Statistical relies on estimates and approximations : Statistical laws are not exact laws like mathematical or chemical laws. They are derived by taking a majority of cases and are not true for every individual. Thus the statistical inferences are uncertain.

4. Statistical results might lead to fallacious conclusions by deliberate manipulation of figures and unscientific handling. This is so because statistical results are represented by figures, which are liable to be manipulated. Also the data placed in the hands of an expert may lead to fallacious results. The figures may be stated without their context or may be applied to a fact other than the one to which they really relate. An interesting example is a survey made some years ago which reported that 33% of all the University teachers. Whereas the University had only three girls student at that time and one of them married to a teacher.

Distrust of Statistics

Due to limitations of statistics an attitude of distrust towards it has been developed.

There are some people who place statistics in the category of lying and maintain that, “there are three degrees of comparison in lying-lies, dammed lies and statistics”. But this attitude is not correct. The person who is handling statistics may be a liar or inexperienced. But that would be the fault not of statistics but of the person handling them. The person using statistics should not take them at their

face value. He should check the result from an independent source. Also only experts should handle the statistics otherwise they may be misused. It may be noted that the distrust of statistics is due more to insufficiency of knowledge regarding the nature, limitations and uses of statistics than to any fundamental inadequacy in the science of statistics. Medicines are meant for curing people, but if they are unscientifically handle by quacks, they may prove fatal to the patient. In both the cases, the medicine is the same; but its usefulness or harmfulness depends

upon the man who handles it. We cannot blame medicine for such a result.

Similarly, if a child cuts his finger with a sharp knife, it is not a knife that is to be blamed, but the person who kept the knife at a place that the child could reach it.

These examples help us in emphasising that if statistical facts are misused by some people it would be wrong to blame the statistics as such. It is the people who are to

be blamed. In fact statistics are like clay which can be moulded in any way.

Collection of data

For studying a problem statistically first of all, the data relevant thereto must be collected. The numerical facts constitute the raw material of the statistical process.

The interpretation of the ultimate conclusion and the decisions depend upon the accuracy with which the data are collected. Unless the data are collected with

sufficient care and are as accurate as is necessary for the purposes of the inquiry, the result obtained cannot be expected to be valid or reliable. Before starting the collection of the data, it is necessary to know the sources from which the data are to be collected.

Primary and Secondary Sources

The original compiler of the data is the primary source. For example, the office of the Registrar General will be the primary source of the decennial population census

figures. A secondary source is the one that furnishes the data that were originally compiled students at John Hopkins University had married by someone else. If the population census figures issued by the office of the Registrar-General are published in the Indian year Book, this publication will be the secondary source of the population data. The source of data also are classified according to the character of the data yielded by them. Thus the data which are gathered from the primary source is known as primary data and the one gathered from the secondary source is known as secondary data. When an investigator is making use of figures which he has obtained by field enumeration, he is said to be using primary data and when he is making use of figures which he has obtained from some other source, he is said to be using secondary data.

Choice between Primary and Secondary Data

An investigator has to decide whether he will collect fresh (primary) data or he will compile data from the published sources. The former is reliable per se but the latter

can be relied upon only by examining the following factors:—

- (i) source from which they have been obtained;
- (ii) their true significance;

(iii) completeness and

(iv) method to collection.

In addition to the above factors, there are other factors to be considered while making choice between the primary or secondary data :

(i) Nature and scope of enquiry.

(ii) Availability of time and money.

(iii) Degree of accuracy required and

(iv) The status of the investigator i.e., individual, Pvt. Co., Govt. etc.

However, it may be pointed out that in certain investigations both primary and secondary data may have to be used, one may be supplement to the other.

Methods of Collection of Primary Data

The primary methods of collection of statistical information are the following :

1. Direct Personal Observation,
2. Indirect Personal Observation,
3. Schedules to be filled in by informants
4. Information from Correspondents, and
5. Questionnaires in charge of enumerators

The particular method that is decided to be adopted would depend upon the nature and availability of time, money and other facilities available to the investigation.

1. Direct Personal Observation

According to this method, the investigator obtains the data by personal observation. The method is adopted when the field of inquiry is small. Since the by someone else. If the population census figures issued by the office of the

Registrar-General are published in the Indian year Book, this publication will be the secondary source of the population data. The source of data also are classified according to the character of the data yielded by them. Thus the data which are gathered from the primary source is known as primary data and the one gathered from the secondary source is known as secondary data. When an investigator is making use of figures which he has obtained by field enumeration, he is said to be using primary data and when he is making use of figures which he has obtained from some other source, he is said to be using secondary data.

Choice between Primary and Secondary Data

An investigator has to decide whether he will collect fresh (primary) data or he will compile data from the published sources. The former is reliable per se but the latter can be relied upon only by examining the following factors :—

- (i) source from which they have been obtained;
- (ii) their true significance;
- (iii) completeness and
- (iv) method to collection.

In addition to the above factors, there are other factors to be considered while making choice between the primary or secondary data :

- (i) Nature and scope of enquiry.
- (ii) Availability of time and money.
- (iii) Degree of accuracy required and
- (iv) The status of the investigator i.e., individual, Pvt. Co., Govt. etc.

However, it may be pointed out that in certain investigations both primary and secondary data may have to be used, one may be supplement to the other.

Methods of Collection of Primary Data

The primary methods of collection of statistical information are the following :

1. Direct Personal Observation,
2. Indirect Personal Observation,
3. Schedules to be filled in by informants
4. Information from Correspondents, and
5. Questionnaires in charge of enumerators

The particular method that is decided to be adopted would depend upon the nature and availability of time, money and other facilities available to the investigation.

1. Direct Personal Observation

According to this method, the investigator obtains the data by personal observation. The method is adopted when the field of inquiry is small. Since the investigator is closely connected with the collection of data, it is bound to be more

accurate. Thus, for example, if an inquiry is to be conducted into the family budgets and living conditions of industrial labour, the investigator himself live in the industrial area as one of the industrial workers, mix with other residents and

make patient and careful personal observation regarding how they spend, work and live.

2. Indirect Personal Observation

According to this method, the investigator interviews several persons who are either directly or indirectly in possession of the information sought to be collected.

It may be distinguished from the first method in which information is collected directly from the persons who are involved in the inquiry. In the case of indirect personal observation, the persons from whom the information is being collected are known as witnesses or informants. However it

should be ascertained that the informants really pass the knowledge and they are not prejudiced in favour of or against a particular view point.

This method is adopted in the following situations:

1. Where the information to be collected is of a complete nature.
2. When investigation has to be made over a wide area.
3. Where the persons involved in the inquiry would be reluctant to part with the information. This method is generally adopted by enquiry committee or commissions appointed by government.
3. Schedules to be filled in by the informants

Under this method properly drawn up schedules or blank forms are distributed among the persons from whom the necessary figures are to be obtained. The informants would fill in the forms and return them to the officer in charge of

investigation. The Government of India issued slips for the special enumeration of scientific and technical personnel at the time of census. These slips are good examples of schedules to be filled in by the informants.

The merit of this method is its simplicity and lesser degree of trouble and pain for the investigator. Its greatest drawback is that the informants may not send back the schedules duly filled in.

4. Information from Correspondents

Under this method certain correspondents are appointed in different parts of the field of enquiry, who submit their reports to the Central Office in their own manner. For example, estimates of agricultural wages may be periodically furnished to the Government by village school teachers. The local correspondents being on the spot of the enquiry are capable of giving reliable information. But it is not always advisable to place much reliance on correspondents, who have often got their own personal prejudices. However, by this method, a rough and approximate estimate is obtained at a very low cost. This method is also adopted by various departments of the government in such cases where regular information is to be collected from a wide area.

Questionnaire in charge of Enumerations

A questionnaire is a list of questions directly or indirectly connected with the work of the enquiry. The answers to these questions would provide all the information

sought. The questionnaire is put in the charge of trained investigators whose duty is to go to all persons or selected persons connected with the enquiry. This method is usually adopted in case of large inquiries. The method of collecting data is relatively cheap. Also the information obtained is that of good quality.

The main drawback of this method is that the enumerator (i.e., investigator in charge of the questionnaire) may be a biased one and may not enter the answer given by the informant. Where there are many enumerators, they may interpret various terms in questionnaire according to their whims. To that extent the information supplied may be either inaccurate or inadequate or not comparable.

This drawback can be removed to a great extent by training the investigators before the enquiry begins. The meaning of different questions may be explained to them so that they do not interpret them according to their whims.

Drafting the Questionnaire

The success of questionnaire method of collecting information depends on the proper drafting of the questionnaire. It is a highly specialized job and requires great deal of skill and experience. However, the following general principle may be helpful in framing a questionnaire:

1. The number of the questions should be kept to the minimum fifteen to twenty five may be a fair number.
2. The questions must be arranged in a logical order so that a natural and spontaneous reply to each is induced.
3. The questions should be short, simple and easy to understand and they should convey one meaning.
4. As far as possible, quotation of a personal and pecuniary nature should not be asked.

5. As far as possible the questions should be such that they can be answered briefly in 'Yes' or 'No', or in terms of numbers, place, date, etc.

6. The questionnaire should provide necessary instructions to the Informants. For instance, if there is a question on weight. It should be specified as to whether weight is to be indicated in lbs or kilograms.

7. Questions should be objective type and capable of tabulation.

Specimen Questionnaire

We are giving below a specimen questionnaire of Expenditure Habits of Students residing in college Hostels.

Name of Student Class

State and District of origin Age

1. How much amount do you get from your father/guardian p.m. ?

2. Do you get some scholarship? If so, state the amount per month.

3. Is there any other source from which you get money regularly? (e.g. mother, brother or uncle).

4. How much do you spend monthly on the following items:

Rs.

College Tuition Fee

Hostel Food Expenses

Other hostel fees

Clothing

Entertainment

Smoking

Miscellaneous

Total

5. Do you smoke? If so what is the daily expenditure on it?

6. Any other item on which you spend money ?

Sources of Secondary Data

There are number of sources from which secondary data may be obtained. They may be classified as follow. :

1. Published sources, and

2. Unpublished sources.

1. Published Sources

The various sources of published data are :

1. Reports and official publications of-

(a) International bodies such as the International Monetary Fund, International Finance Corporation, and United Nations Organisation.

(b) Central and State Governments- such as the Report of the Patel Committee, etc.

2. Semi Official Publication. Various local bodies such as Municipal Corporation, and Districts Boards.

3. Private Publication of—

(a) Trade and professional bodies such as the Federation of India, Chamber of Commerce and Institute of Chartered Accountants of India.

(b) Financial and Economic Journals such as “Commerce”, ‘Capital’ etc.

(c) Annual Reports of Joint Stock Companies.

(d) Publication brought out by research agendas, research scholars, etc.

2. Unpublished Sources

There are various sources of unpublished data such as records maintained by various government and private offices, studies made by research institutions, scholars, etc., such source can also be used where necessary.

Census and Sampling Techniques of Collection of Data

There are two important techniques of Data collection,

- (i) Census enquiry implies complete enumeration of each unit of the universe,
- (ii) In a sample survey, only a

small part of the group, is considered, which is taken as representative. For example the population census in India implies the counting of each and every human being within the country. In practice sometimes it is not possible to examine every item in the population. Also many a time it is possible to obtain sufficiently accurate results by studying only a part of the “population”. For example, if the marks obtained in statistics by 10 students in an examination are

selected at random, say out of 100, then the average marks obtained by 10 students will be reasonably representative of the average marks obtained by all the 100 students. In such a case, the populations will be the marks of the entire group of 100 students and that of 10 students will be a sample.

3. COLLECTION OF DATA, CLASSIFICATION AND TABULATION

3.1 Introduction:

Everybody collects, interprets and uses information, much of it in a 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.

1. To describe the methods of collecting primary statistical information.
2. To consider the status involved in carrying out a survey.
3. To analyse the process involved in observation and interpreting.
4. To define and describe sampling.
5. To analyse the basis of sampling.
6. 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 possibleprecautions should be taken to ensure accuracy while collecting thedata.

3.2 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:

1. Time series data.
2. Spatial data
3. Spacio-temporal data.

3.2.1 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 1:

The following is the data for the three types of expenditures for a family for the four years 2001,2002,2003,2004.

Year	Food	Education	Others	Total
2001	3000	2000	3000	8000
2002	3500	3000	4000	10500
2003	4000	3500	5000	12500
2004	5000	5000	6000	16000

3.2.2 Spatial Data:

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

Example 2:

The population of the southern states of India in 1991.

State Population

Tamilnadu 5,56,38,318

Andhra Pradesh 6,63,04,854

Karnataka 4,48,17,398

Kerala 2,90,11,237

Pondicherry 7,89,416

3.2.3 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 3:

State Population

1981 1991

Tamil Nadu 4,82,97,456 5,56,38,318

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

3.3 Categories of data:

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

These categories are,

1. Primary data
2. Secondary data

3.3.1 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.

Example 4:

If a researcher is interested to know the impact of noon-meal scheme for the school children, he has to undertake a survey and collect data on the opinion of parents and children by asking relevant questions. Such a data collected for the purpose is called primary data.

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

1. Direct personal interviews.
2. Indirect Oral interviews.
3. Information from correspondents.
4. Mailed questionnaire method.
5. Schedules sent through enumerators.

1. Direct personal interviews:

The persons from whom informations are collected are known as informants. The investigator personally meets them and asks questions to gather the necessary informations. It is the suitable method for intensive rather than extensive field surveys. It suits best for intensive study of the limited field.

2. Indirect Oral Interviews:

Under this method the investigator contacts witnesses or neighbours or friends or some other third parties who are capable of supplying the necessary information. This method is preferred if the required information is on addition or cause of fire or theft or murder etc., If a fire has broken out a certain place, the persons living in

neighbourhood and witnesses are likely to give information on the cause of fire. In some cases, police interrogated third parties who are supposed to have knowledge of a theft or a murder and get some clues.

Enquiry committees appointed by governments generally adopt this method and get people's views and all possible details of facts relating to the enquiry. This method is suitable whenever direct sources do not exist or cannot be relied upon or would be unwilling to part with the information.

The validity of the results depends upon a few factors, such as the nature of the person whose evidence is being recorded, the ability of the interviewer to draw out information from the third parties by means of appropriate questions and cross examinations, and the number of persons interviewed. For the success of this method one person or one group alone should not be relied upon.

3. Information from correspondents:

The investigator appoints local agents or correspondents in different places and compiles the information sent by them.

Information to newspapers and some departments of Government come by this method. The advantage of this method is that it is cheap and appropriate for extensive investigations. But it may not ensure accurate results because the correspondents are likely to be negligent, prejudiced and biased. This method is adopted in those cases where information is to be collected periodically from a wide area for a long time.

4. Mailed questionnaire method:

Under this method a list of questions is prepared and is sent to all the informants by post. The list of questions is technically called questionnaire. A covering letter accompanying the questionnaire explains the purpose of the investigation and the importance of correct information and request the informants to fill in the blank spaces provided and to return the

form within a specified time. This method is appropriate in those cases where the informants are literates and are spread over a wide area.

5. Schedules sent through Enumerators:

Under this method enumerators or interviewers take the schedules, meet the informants and filling their replies. Often distinction is made between the schedule and a questionnaire. A schedule is filled by the interviewers in a face-to-face situation with the informant. A questionnaire is filled by the informant which he receives and returns by post. It is suitable for extensive surveys.

3.3.2 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:

1. Published sources, and
2. Unpublished sources.

1. Published Sources:

The various sources of published data are: Clinical and

other personal records, death certificates, published mortality statistics, census publications, etc. Examples include:

1. Official publications of Central Statistical Authority
2. Publication of Ministry of Health and Other Ministries
3. News Papers and Journals.
4. International Publications like Publications by WHO, World Bank UNICEF
5. Records of hospitals or any Health Institutions.

Note: A lot of secondary data is available in the internet. We can access it at any time for the further studies.

2. Unpublished Sources

All statistical material is not always published. There are various sources of unpublished data such as records maintained by various Government and private offices, studies made by research institutions, scholars, etc. Such sources can also be used where necessary

Precautions in the use of Secondary data

The following are some of the points that are to be considered in the use of secondary data

1. How the data has been collected and processed
2. The accuracy of the data
3. How far the data has been summarized
4. How comparable the data is with other tabulations
5. How to interpret the data, especially when figures collected for one purpose is used for another
 - Generally speaking, with secondary data, people have to compromise between what they want and what they are able to find.

3.4 Classification:

The collected data, also known as raw data or ungrouped data are always in an unorganised 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 assimilable 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., Delhi, Madurai, Bangalore, Mumbai etc.,

Objects of Classification:

The following are main objectives of classifying the data:

1. It condenses the mass of data in an easily assimilable form.
2. It eliminates unnecessary details.
3. It facilitates comparison and highlights the significant aspect of data.
4. It enables one to get a mental picture of the information and helps in drawing inferences.
5. 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

- a) Chronological classification
- b) Geographical classification
- c) Qualitative classification

d) Quantitative classification

e) Chronological classification:

In chronological classification the collected data are arranged according to the order of time expressed in years, months, weeks, etc.,

The data is generally classified in ascending order of time.

Example 5:

The estimates of birth rates in India during 1970 – 76 are

Year 1970 1971 1972 1973 1974 1975 1976

Birth Rate 36.8 36.9 36.6 34.6 34.5 35.2 34.2

b) Geographical classification:

In this type of classification the data are classified according to geographical region or place. For instance, the production of paddy in different states in Iraq, production of wheat in different countries etc.,

Example 6:

Country America China Denmark France Iraq

Yield of wheat in (kg/acre) 1925 893 225 439 862

c) Qualitative classification:

In this type of classification data are classified on the basis of same attributes or quality like sex, literacy, religion, employment etc.,

Such attributes cannot be measured along with a scale.

For example, if the population to be classified in respect to one attribute, say sex, then we can classify them into two namely that of males and females. Similarly, they can also be classified into 'married or 'single' on the basis of another attribute 'marital status'.

Thus when the classification is done with respect to one

attribute, which is dichotomous in nature, two classes are formed, one possessing the attribute and the other not possessing the attribute.

This type of classification is called simple or dichotomous classification.

A simple classification may be shown as under

Population
Female Male

The classification, where two or more attributes are considered and several classes are formed, is called a manifold classification. For example, if we classify population simultaneously with respect to two attributes, e.g. sex and marital status, then population are first classified with respect to 'sex' into 'males' and 'females'. Each of these classes may then be further classified into 'married' and 'single' on the basis of attribute 'employment' and as such Population are classified into four classes namely.

- (i) Male married
- (ii) Male single
- (iii) Female married
- (iv) Female single

Still the classification may be further extended by considering other attributes like marital status etc. This can be explained by the following chart

Population

Female Male

single married single married

d) Quantitative classification:

Quantitative classification refers to the classification of data

according to some characteristics that can be measured such as height, weight, etc.,

Diagrammatic representation of data – uses and limitations – simple, Multiple, Component and percentage bar diagrams – pie chart

Diagrams are various geometrical shape such as bars, circles etc. Diagrams are based on scale but are not confined to points or lines. They are more attractive and easier to understand than graphs.

Merits

- 1. Most of the people are attracted by diagrams.**
- 2. Technical Knowledge or education is not necessary.**
- 3. Time and effort required are less.**
- 4. Diagrams show the data in proper perspective.**
- 5. Diagrams leave a lasting impression.**
- 6. Language is not a barrier.**
- 7. Widely used tool.**

Demerits (or) limitations

- 1. Diagrams are approximations.**
- 2. Minute differences in values cannot be represented properly in diagrams.**
- 3. Large differences in values spoil the look of the diagram.**
- 4. Some of the diagrams can be drawn by experts only. eg. Pie chart.**
- 5. Different scales portray different pictures to laymen.**

Types of Diagrams

The important diagrams are

- 1. Simple Bar diagram.**
- 2. Multiple Bar diagram.**

3. Component Bar diagram.

4. Percentage Bar diagram.

5. Pie chart

6. Pictogram

7. Statistical maps or cartograms.

In all the diagrams and graphs, the groups or classes are represented on the x-axis and the volumes or frequencies are represented in the y-axis.

Simple Bar diagram

If the classification is based on attributes and if the attributes are to be compared with respect to a single character we use simple bar diagram.

Example

- 1. The area under different crops in a state.**
- 2. The food grain production of different years.**
- 3. The yield performance of different varieties of a crop.**
- 4. The effect of different treatments etc.**

Simple bar diagrams Consists of vertical bars of equal width. The heights of these bars are proportional to the volume or magnitude of the attribute. All bars stand on the same baseline. The bars are separated from each others by equal intervals. The bars may be coloured or marked.

Line Diagrams:

This kind of a diagram becomes suitable for representing data supplied chronologically in an ascending or descending order. Usually, it shows the behaviour of a variable over time. Successive values of a variable at different periods or places are plotted as separate points on a two dimensional plane and the locus of all those points joined together form a continuous line segment, called line diagram.

While tracing out such a diagram, the usual convention is to show the successive values of the variable under study along the vertical axis in an increasing order and the time dimension along the horizontal axis. It should carefully be noted that none of the two axes be too long or too short with respect to each other.

ADVERTISEMENTS:

This is very much necessary mainly to avoid unpredictable and wide fluctuations in the given values of the variable. The origin or the (0, 0) point at the left hand corner should clearly be mentioned so as to discard wrong impression on the process of drawing.

Two or more (but finite number of) line segments can also be drawn on the same quadrant when information on different variables over the same period or time are simultaneously represented using the same unit of measurement along the same axis. We can thus draw a number of line-diagrams for different data series on the same quadrant.

They can distinctly and attractively be displayed on a screen for presentation with various colourful lines. When the values of the variable under consideration change at a constant rate over the same successive time intervals, the diagram will take the shape of a straight line. Otherwise, it will represent various concave, convex or irregular curves when viewed from the origin.

Let us now represent a common line diagram below:

ADVERTISEMENTS:

Example:

Line diagrams showing total values of Exports and Imports during 1987-96 have been presented in Fig. 7.1. This figure has been drawn on the basis of data shown in Table 7.4.

Table 7.4 : Foreign Trade of India during 1987-96 (units in Rs. crores)

Year	Value of Export	Value of Import
1987—88	301	243
1988—89	295	226
1989—90	309	230
1990—91	260	184
1991—92	276	168
1992—93	184	85
1993—94	158	89
1994—95	166	160
1995—96	182	177

Two separate line diagrams showing fluctuations in the values of exports and imports of India during (1987—96) are shown below:

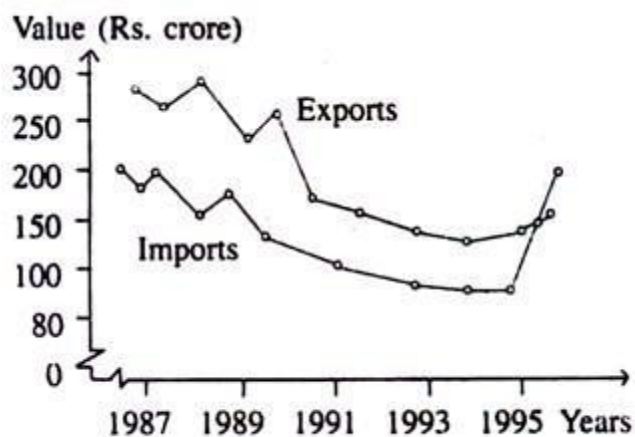


Fig. 7.1 : Line Diagrams

Bar Diagrams:

It is another well-known useful statistical weapon to represent raw data decently. This device is applied specially in a situation where the given data can be classified on the basis of a non-measurable criterion e.g., standards of college education in different states of India at the present time.

This is very often called cross-section data. More precisely, a bar graph is formed as a collection of rectangles having the same width or breadth placed successively at equal distance. Practically, the height of each bar placed vertically represents the value of the variable on the identical class interval shown horizontally.

Usually, these bars are placed either vertically on the horizontal axis or horizontally on the vertical axis and they are thus known as vertical bar chart or horizontal bar chart. Conventionally vertical bar charts are formed with the time series data.

ADVERTISEMENTS:

Actually speaking, no formal rule as to how much space to be given in between the two bars is there. If necessary, no space in between two bars can be given. In some other cases, suitable and reasonable gaps in-between two bars may also be allowed.

Let us imprint simple and suitable examples of bar diagrams below:

(a) Simple Vertical Bar Diagram:

Volume of population in a number of states in India in 2001 is given below—represents the data with the aid of vertical bars.

Table 7.5 : Volume of Population in Five Different States in India in 2001

States	: Andhra	Bihar	Mahr.	U.P.	W.B.
Popn. (lakh)	: 434	564	503	889	444

ADVERTISEMENTS:

Fig. 7.2 Shows population of a number of 5 States in India in a particular year (2001):

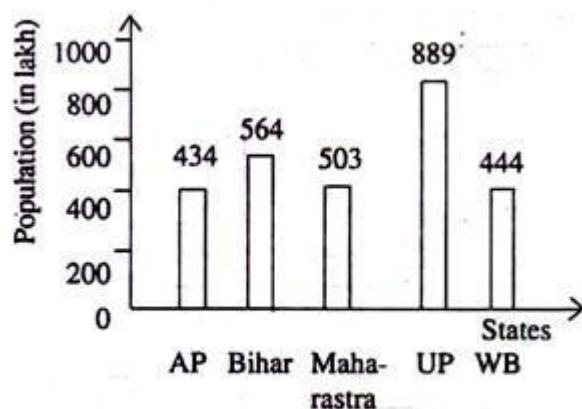


Fig. 7.2 : Vertical Bar Diagram

(b) Horizontal Bar Diagram:

Volume of production and profit of five different organisations operating under a particular industry with separate productive capacities are given below for the two successive years 2011 and 2012.

We represent the information through an ideal bar diagram. Here Fig. 7.3 is drawn below on the basis of Table 7.6. We have chosen this horizontal bar diagram to facilitate comparison of performances of 5 organisations for the years 2011 and 2012, respectively.

ADVERTISEMENTS:

Table 7.6 : Production and Profit of Five Different Organisations in 2011-12

Organi- sation	2011		2012	
	Production (Thousand)	Profit (Rs. Thousand)	Prodn. (Thou.)	Profit (Rs. Thou.)
A	12	11	14	11
B	06	5.5	06	06
C	25	26	27	30
D	02	1.8	1.5	1.2
E	07	64	07	07

Horizontal bars show production (in thousands) and profit (Rs. thousand) of five organisations of India in the financial year 2011-12.

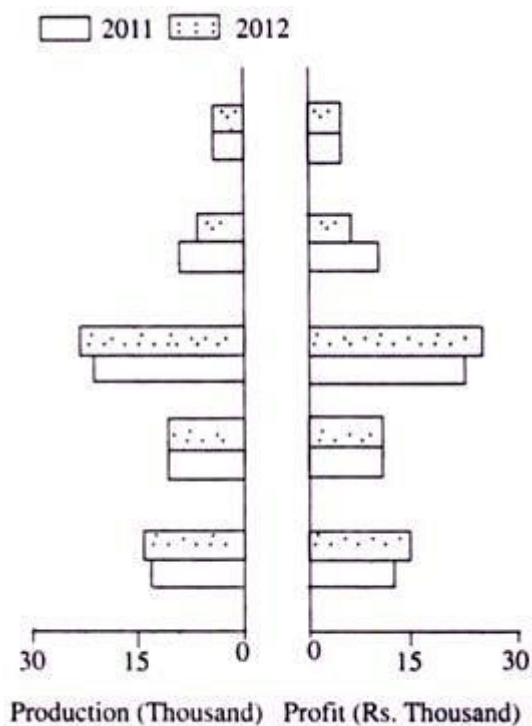


Fig. 7.3 : Horizontal Bar Diagram

(c) *Multiple or Component Bar Diagram*

These diagrams are used in a situation where two or more related categories are to be compared simultaneously.

Consider the following example:

Labour employment and their percentages in 2000 and 2010 in a factory is given below. Represent them in terms of multiple or component bar diagrams.

ADVERTISEMENTS:

Table 7.7 : Sex-wise Labour Employment in a Factory in 2000 and 2010

Labour Employ.	Numbers		Percentages	
	2000	2010	2000	2010
Male	350	300	35	15
Female	400	1200	40	60
Boys	150	100	15	05
Girls	100	400	10	20
Total	1000	2000	100	100

Component bar diagrams show number of labourers of different categories and their respective percentages for the years 2000 and 2010.

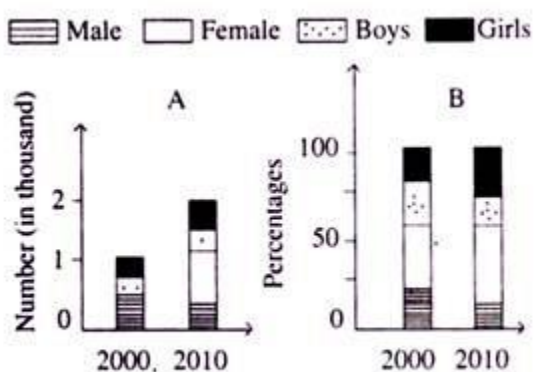


Fig. 7.4 : Component Bar Diagram

Pie Diagram:

It is another effective statistical device to represent quantitative data obtainable on many occasions simply and diagrammatically. When the various parts of the values of a variable possesses different properties then to express the inherent relationship among them and also with the aggregate value of the variable, pie diagram possibly is the best device.

Here, the aggregate value of the variable is expressed as the total area of a circle with a reasonable radius. The entire area in the circle is subdivided into a number of parts by several radii which are separately related to the total area of the circle and also maintain the same proportional relation with the angle at the centre.

For drawing it correctly, we convert the particular given values of the variable as a percentage of the total value of the variable. As the angle at the centre is 360° , it is supposed to express 100 p.c. value of the variable where 1 p.c. value of the variable is equivalent to an angle of 3.6° at the centre.

ADVERTISEMENTS:

We can thus easily convert the individual given values of the variable into the required angles at the centre. Then we draw a complete circle taking any standard radius and put the angles found from the numerical exercise separately at the centre. Each separate part in the circle signifies a

particular section of the data. Let us represent a simple pie diagram below constructed with the usual method prescribed and followed for its computation by converting the following information into that diagram.

Example:

Expenditure incurred by the Planning Commission of India on Education in the last 5-year economic plan.

Table 7.8(A): Educational Expenditure in the Last Five-year Economic Plan:

Table 7.8(A) : Educational Expenditure in the Last Five-year Economic Plan

Stages of education	Total expenditure (in Rs. crores)
I. Primary	209
II. Secondary	88
III. Higher Education	82
IV. Others	39
Total	418

Let us first convert the given data into respective percentages and then into the required angles to be shown at the centre in two more columns and represent them in the following way:

Table 7.8(B) : Calculation for Drawing Pie Diagram

Stages of education	Total expenditure (Rs. crores)	Percentage (%)	Angle at the centre (°)
I. Primary	209	50	180.0
II. Secondary	88	21	75.6
III. Higher Education	82	20	72.0
IV. Others	39	09	32.4
Total	418	100	360.0

Here, angle at the Centre = Percentage \times 3.6.

Pie diagram drawn below on the basis of Table 7.8 (B) shows expenditure on education at various stages in the last 5-year economic plan.

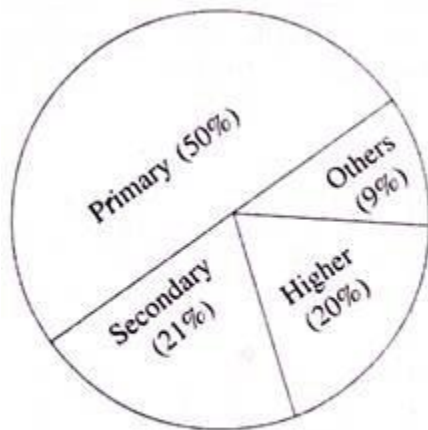


Fig. 7.5 : Pie Diagram

Cumulative-Frequency Curve

Group data are also represented by a curve called ogive or cumulative-frequency curve. As the name suggests, in this representation cumulative frequencies of different class intervals play an important role.

Method of Constructing an Ogive:

Step I: Prepare a frequency-distribution table with overlapping class intervals to make the distribution continuous.

Step II: Prepare the cumulative-frequency table for the distribution.

Step III: Plot points on the horizontal axis (class-interval axis) corresponding to the upper limits of the class intervals.

Step IV: Draw perpendiculars to the horizontal axis at the points representing upper limits of the class intervals. The length of the perpendiculars should represent the cumulative frequencies of

the corresponding class intervals. In other words, plot points with coordinates (a, c), where a = upper limit of a class interval and c = the corresponding cumulative frequency.

Step V: Join the points plotted in step IV freehand to get a smooth curve.

Example 1: The table given below shows the marks obtained by 80 students in science. Construct (i) less than ogive (ii) more than ogive.

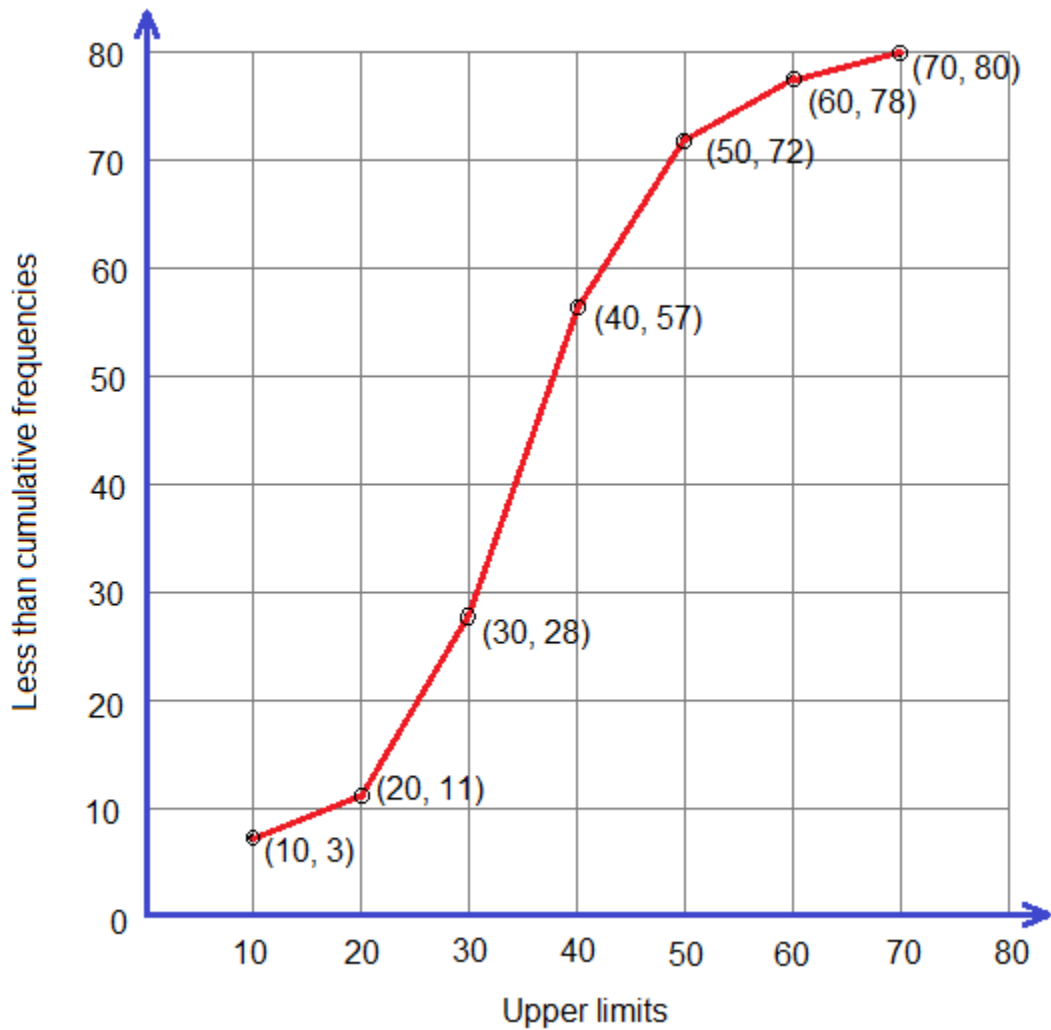
<i>Marks</i>	<i>0-10</i>	<i>10-20</i>	<i>20-30</i>	<i>30-40</i>	<i>40-50</i>	<i>50-60</i>	<i>60-70</i>
<i>No. of students</i>	<i>3</i>	<i>8</i>	<i>17</i>	<i>29</i>	<i>15</i>	<i>6</i>	<i>2</i>

Solution: Here,

Less than cumulative frequency table:

<i>Marks</i>	<i>No. of students</i>	<i>Upper limit</i>	<i>Less than c.f.</i>
<i>0-10</i>	<i>3</i>	<i>10</i>	<i>3 (less than 10)</i>
<i>10-20</i>	<i>8</i>	<i>20</i>	<i>11 (less than 20)</i>
<i>20-30</i>	<i>17</i>	<i>30</i>	<i>28 (less than 30)</i>
<i>30-40</i>	<i>29</i>	<i>40</i>	<i>57 (less than 40)</i>
<i>40-50</i>	<i>15</i>	<i>50</i>	<i>72 (less than 50)</i>
<i>50-60</i>	<i>6</i>	<i>60</i>	<i>78 (less than 60)</i>
<i>60-70</i>	<i>2</i>	<i>70</i>	<i>80 (less than 70)</i>

The less than ogive graph,



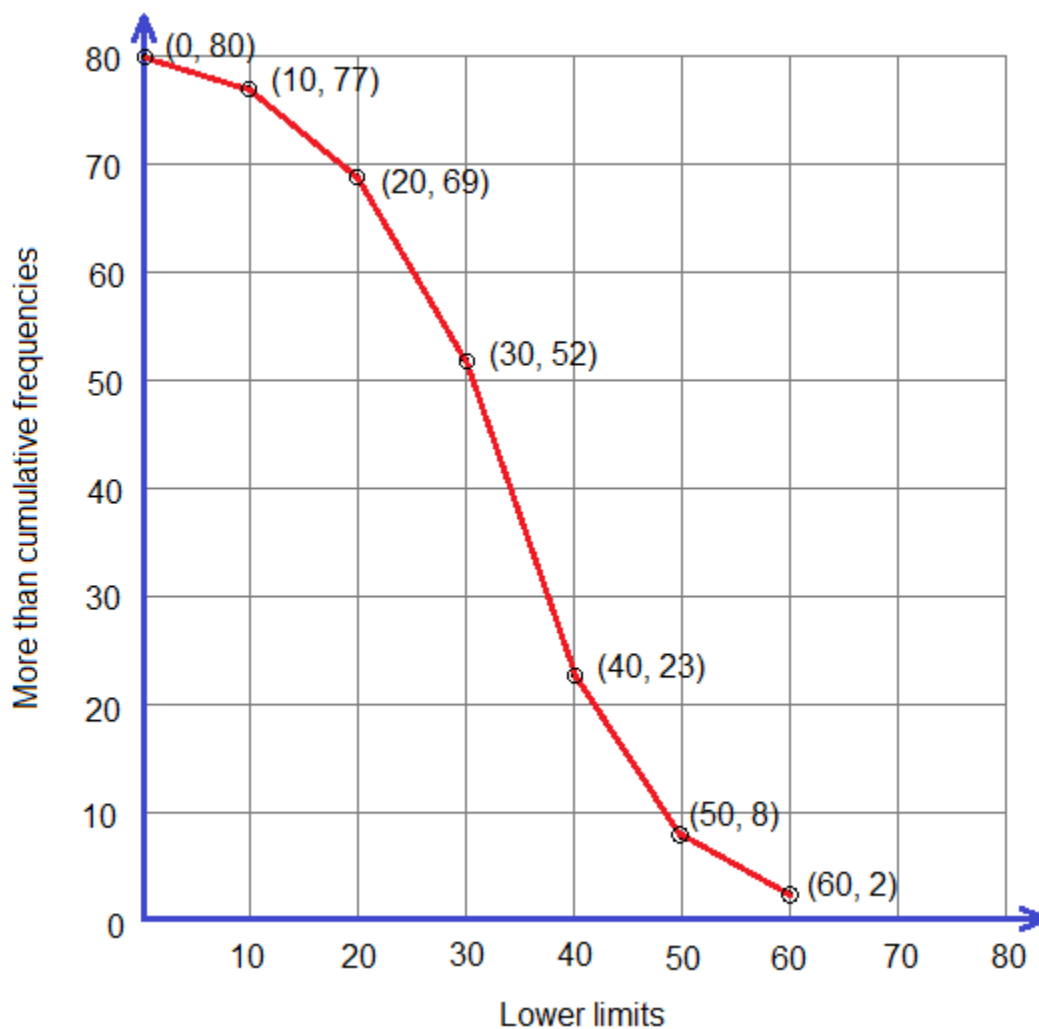
Step VI: Join the first point to the point representing the lower limit of the first class interval in continuation with the smooth curve in step V. (This part of the curve may be a dotted line segment.)

More than cumulative frequency table:

<i>Marks</i>	<i>No. of students</i>	<i>Upper limit</i>	<i>More than c.f.</i>
0-10	3	0	80 (more than 0)
10-20	8	10	$80-3 = 77$ (more than 10)
20-30	17	20	$77-8 = 69$ (more than 20)
30-40	29	30	$69-17 = 52$ (more than 30)
40-50	15	40	$52-29 = 23$ (more than 40)
50-60	6	50	$23-15 = 8$ (more than 50)
60-70	2	60	$8-6 = 2$ (more than 60)

The more than ogive graph,





Less Than And More Than Ogive

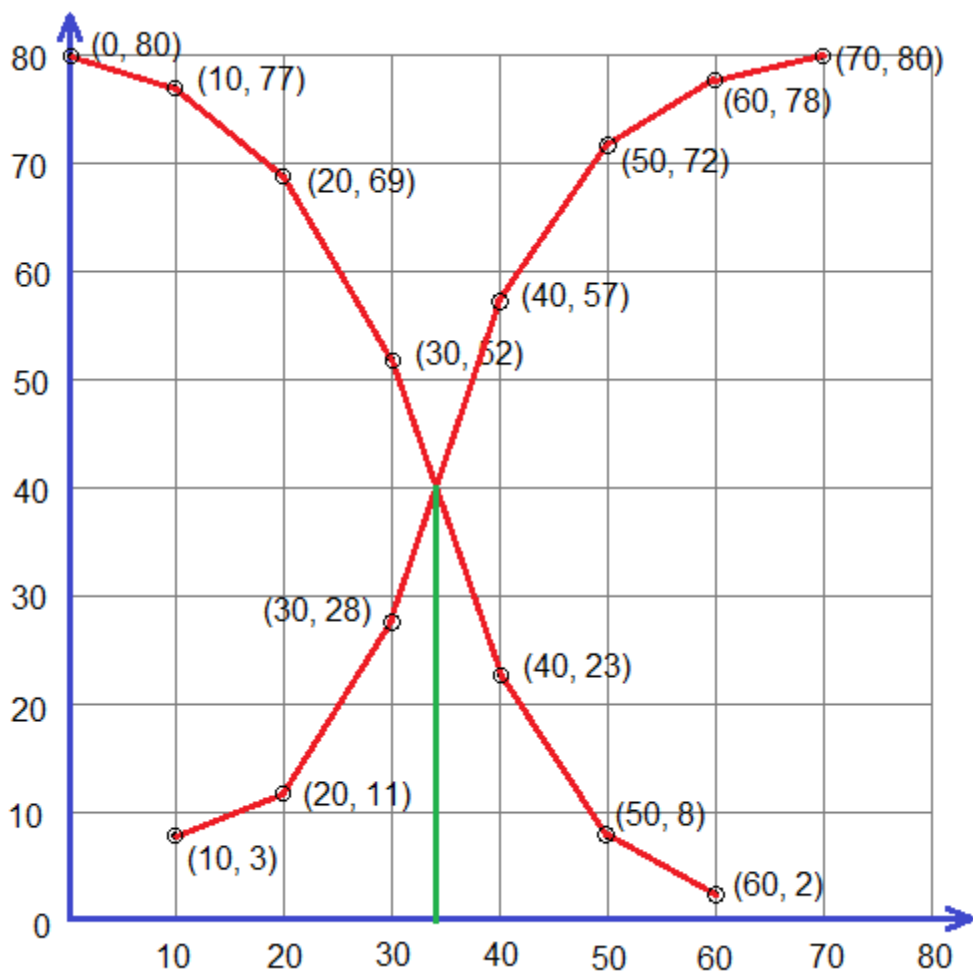
If we draw the both less than ogive and more than ogive of a distribution on the same graph paper, they intersect at a point. The foot of the perpendicular drawn from the point of intersection of two ogives to the x-axis gives the value of median of the distribution.

For example,

Marks	f	Upper limit	Less than c.f.	Lower limit	More than c.f.
0-10	3	10	3	0	80
10-20	8	20	11	10	77

20-30	17	30	28	20	69
30-40	29	40	57	30	52
40-50	15	50	72	40	23
50-60	6	60	78	50	8
60-70	2	70	80	60	2

Less than ogive and more than ogive combined graph,



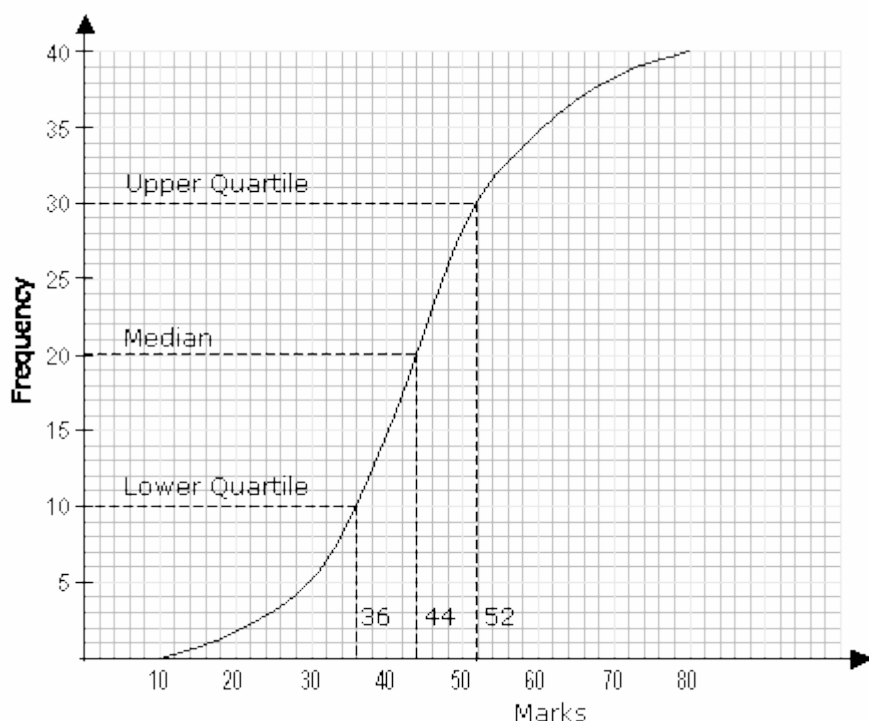
From the graph, the perpendicular drawn from the point of intersection of two ogives meets x -axis at 34.14 (approx.) units from the origin. So, the required median of the given distribution is 34.14.

Median, Quartiles And Percentiles (Grouped Data) in OGIVE

In this lesson, we will learn how to obtain the median, quartiles and percentiles from the cumulative frequency graph of the distribution (grouped data).

Example:

The following cumulative frequency graph shows the distribution of marks scored by a class of 40 students in a test.



Use the graph to estimate

- the median mark
- the upper quartile
- the lower quartile
- the interquartile range

Solution:

a) **Median** corresponds to the **50th percentile** i.e. 50% of the total frequency.

$$50\% \text{ of the total frequency} = \frac{50}{100} \times 40 = \frac{1}{2} \times 40 = 20$$

From the graph, 20 on the vertical axis corresponds to 44 on the horizontal axis. The median mark is 44.

b) The **upper quartile** corresponds to the **75th percentile** i.e. 75% of the total frequency.

$$75\% \text{ of the total frequency} = \frac{75}{100} \times 40 = \frac{3}{4} \times 40 = 30$$

From the graph, 30 on the vertical axis corresponds to 52 on the horizontal axis. The upper quartile is 52.

c) The **lower quartile** corresponds to the **25th percentile** i.e. 25% of the total frequency.

$$25\% \text{ of the total frequency} = \frac{25}{100} \times 40 = \frac{1}{4} \times 40 = 10$$

From the graph, 10 on the vertical axis corresponds to 36 on the horizontal axis. The lower quartile is 36.

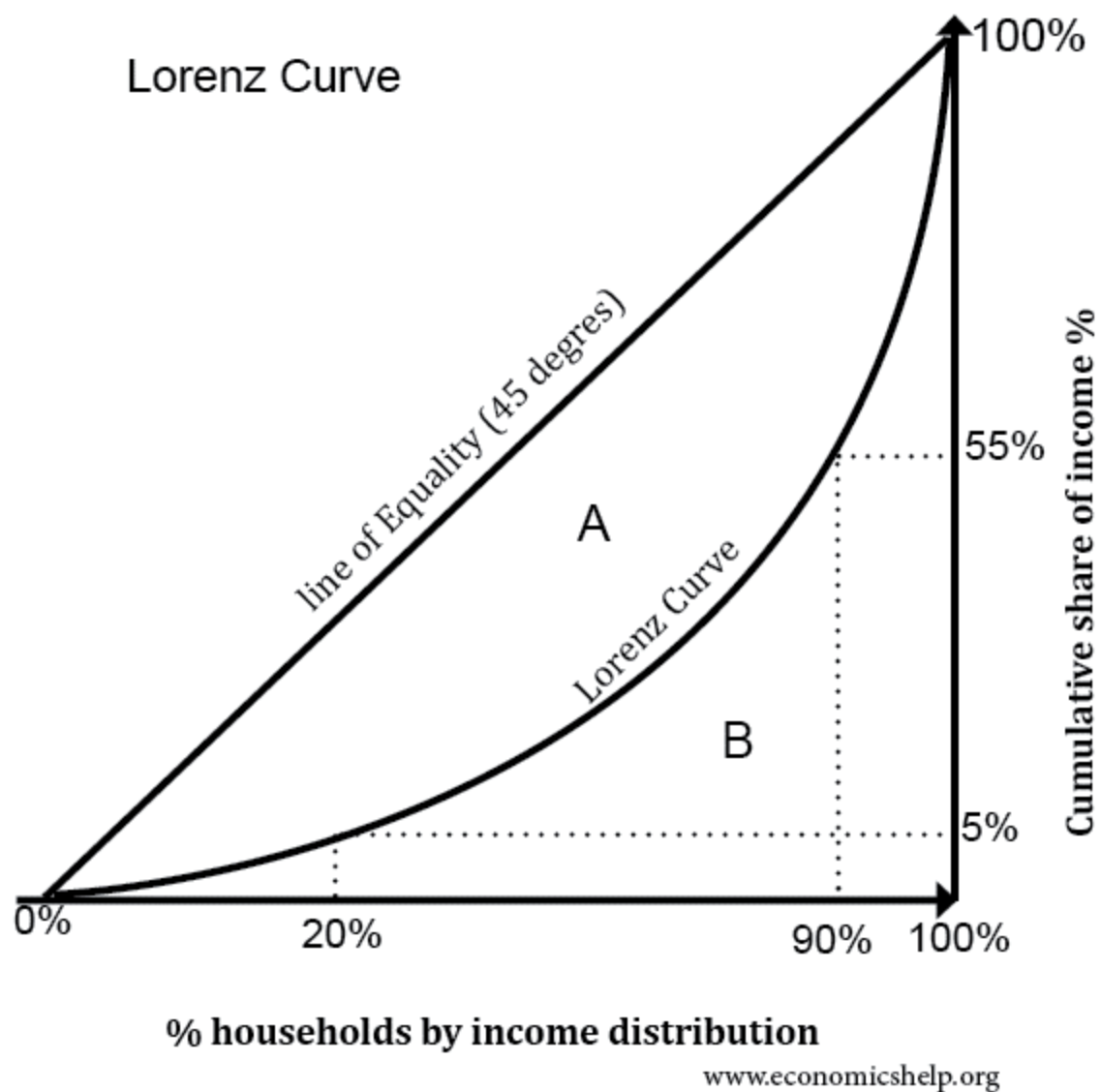
$$\begin{aligned} \text{d) The interquartile range} &= \text{upper quartile} - \text{lower quartile} \\ &= 52 - 36 = 16 \end{aligned}$$

Lorentz Curve

- **Definition:** The Lorenz curve is a way of showing the distribution of income (or wealth) within an economy. It was developed by Max O. Lorenz in 1905 for representing wealth distribution.
- The Lorenz curve shows the cumulative share of income from different sections of the population.
- If there was perfect equality – if everyone had the same salary – the poorest 20% of the population would gain 20% of the total income. The poorest 60% of the population would get 60% of the income.



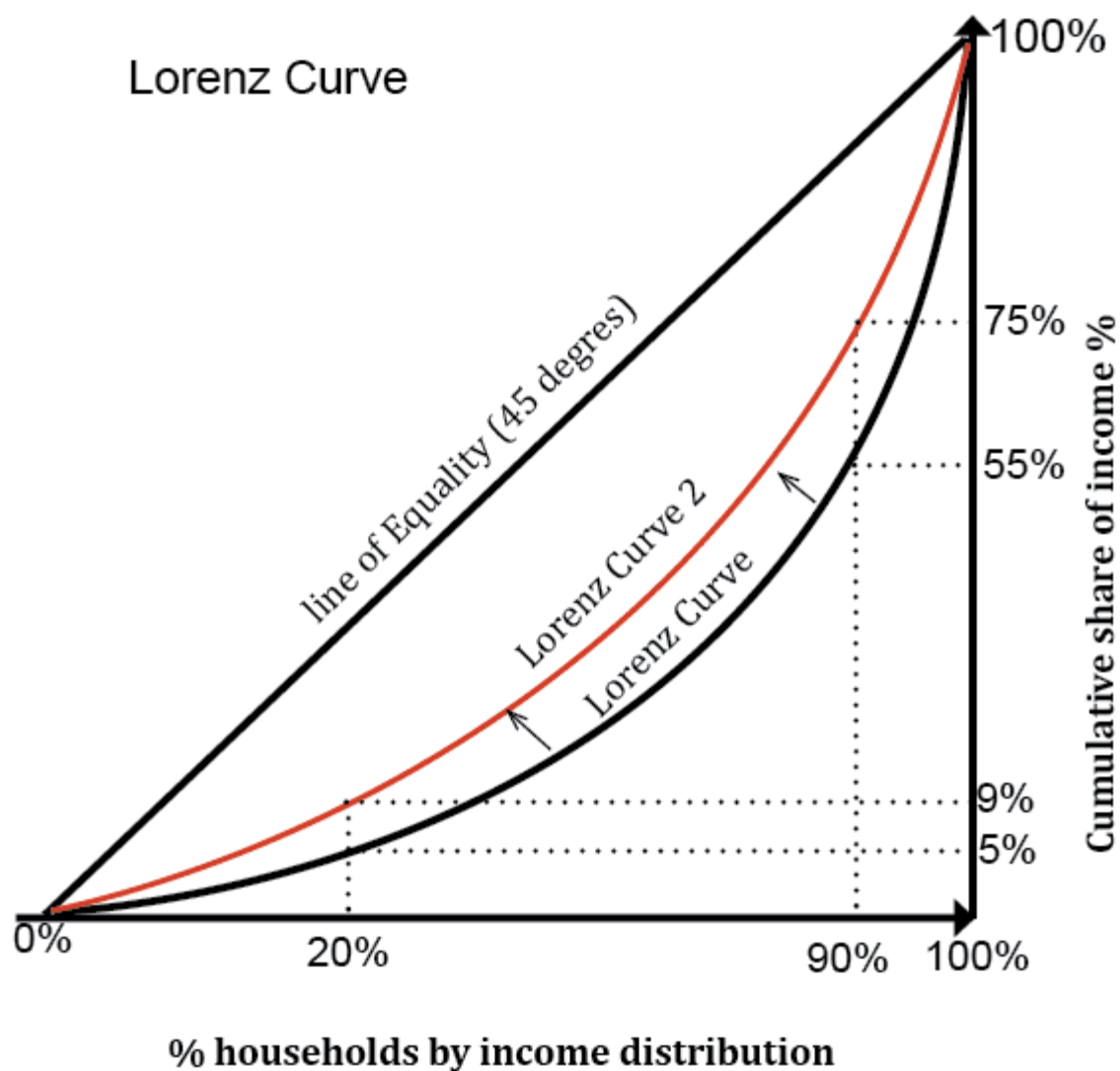
Diagram of Lorenz curve



In this Lorenz curve, the poorest 20% of households have 5% of the nation's total income.

The poorest 90% of the population holds 55% of the total income. That means the richest 10% of income earners gain 45% of total income.

Shift in the Lorenz Curve



www.economicshelp.org

In this example, there has been a reduction in inequality – the Lorenz curve has moved closer to the line of equality.

- The poorest 20% of the population now gain 9% of total income
- The richest 10% of the population used to gain 45% of total income but now only get 25% of total income

UNIT II

Arithmetic Mean

(a) To find A.M. for Raw data

For a raw data, the arithmetic mean of a series of numbers is sum of all observations divided by the number of observations in the series. Thus if x_1, x_2, \dots, x_n represent the values of n observations, then arithmetic mean (A.M.) for n observations is: (direct method)

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

There are two methods for computing the A.M :

- (i) Direct method
- (ii) Short cut method.

Example 5.1

The following data represent the number of books issued in a school library on selected from 7 different days 7, 9, 12, 15, 5, 4, 11 find the mean number of books.

Solution:

$$\begin{aligned} \bar{x} &= \frac{1}{n} \sum_{i=1}^n x_i \\ \bar{x} &= \frac{7+9+12+15+5+4+11}{7} \\ &= \frac{63}{7} = 9 \end{aligned}$$

Hence the mean of the number of books is 9

Short-cut Method to find A.M.

Under this method an assumed mean or an arbitrary value (denoted by A) is used as the basis of calculation of deviations (d_i) from individual values. That is if $d_i = x_i - A$

Then

$$\bar{x} = A + \frac{\sum_{i=1}^n d_i}{n}$$

Example 5.2

A student's marks in 5 subjects are 75, 68, 80, 92, 56. Find the average of his marks.

Solution:

Let us take the assumed mean, $A = 68$



x_i	$d_i = x_i - 68$
75	7
68	0
80	12
56	-12
92	24
Total	31

$$\begin{aligned}\bar{x} &= A + \frac{\sum_{i=1}^n d_i}{n} \\ &= 68 + \frac{31}{5} \\ &= 68 + 6.2 = 74.2\end{aligned}$$

The arithmetic mean of average marks is 74.2

(b) To find A.M. for Discrete Grouped data

If x_1, x_2, \dots, x_n are discrete values with the corresponding frequencies f_1, f_2, \dots, f_n . Then the mean for discrete grouped data is defined as (direct method)

$$\bar{x} = \frac{\sum_{i=1}^n f_i x_i}{N}$$

In the short cut method the formula is modified as

$$\bar{x} = A + \frac{\sum_{i=1}^n f_i d_i}{N} \quad \text{where } d_i = x_i - A$$

Example 5.3

A proof reads through 73 pages manuscript. The number of mistakes found on each of the pages are summarized in the table below. Determine the mean number of mistakes found per page.

No of mistakes	1	2	3	4	5	6	7
No of pages	5	9	12	17	14	10	6

Solution:

(i) Direct Method

x_i	f_i	$f_i x_i$
1	5	5
2	9	18
3	12	36
4	17	68
5	14	70
6	10	60
7	6	42
Total	N=73	299

$$\begin{aligned}\bar{x} &= \frac{\sum_{i=1}^n f_i x_i}{N} \\ &= \frac{299}{73} \\ &= 4.09\end{aligned}$$

The mean number of mistakes is 4.09

(ii) Short-cut Method

x_i	f_i	$d_i=x_i-A$	$f_i d_i$
1	5	-3	-15
2	9	-2	-18
3	12	-1	-12
4	17	0	0
5	14	1	14
6	10	2	20
7	6	3	18
	$\Sigma f_i=73$		$\Sigma f_i d_i=7$

$$\begin{aligned}\bar{x} &= A + \frac{\sum_{i=1}^n f_i d_i}{N} \\ &= 4 + \frac{7}{73} \\ &= 4.09\end{aligned}$$

The mean number of mistakes = 4.09

(c) Mean for Continuous Grouped data:

For the computation of A.M for the continuous grouped data, we can use direct method or short cut method.

Direct Method:

The formula is

$$\bar{x} = \frac{\sum_{i=1}^n f_i x_i}{N}, \quad x_i \text{ is the midpoint of the class interval}$$

Short cut method

$$\bar{x} = A + \frac{\sum_{i=1}^n f_i d_i}{N} \times C$$

$$d = \frac{x_i - A}{c}$$

where A - any arbitrary value
 c - width of the class interval

x_i is the midpoint of the class interval.

Example 5.4

The following the distribution of persons according to different income groups

Income (in ` 1000)	0 – 8	8 – 16	16 – 24	24 – 32	32 – 40	40 – 48
No of persons	8	7	16	24	15	7

Find the average income of the persons.

Solution :

Direct Method:

Class	f_i	x_i	$f_i x_i$
0-8	8	4	32
8 - 16	7	12	84
16-24	16	20	320
24-32	24	28	672
32-40	15	36	540
40-48	7	44	308
Total	N =77		1956

$$\begin{aligned}\bar{x} &= \frac{\sum_{i=1}^n f_i x_i}{N} \\ &= \frac{1956}{77} \\ &= 25.40\end{aligned}$$

Short cut method:



Class	f_i	x_i	$d_i = (x_i - A)/c$	$f_i d_i$
0 – 8	8	4	-3	-24
8 – 16	7	12	-2	-14
16 – 24	16	20	-1	-16
24 – 32	24	28 A	0	0
32 – 40	15	36	1	15
40 – 48	7	44	2	14
Total	N= 77			-25

$$\begin{aligned}\bar{x} &= A + \frac{\sum_{i=1}^n f_i d_i}{N} \times C \\ &= 28 + \frac{-25}{77} \times 8 = 25.40\end{aligned}$$

Merits

- It is easy to compute and has a unique value.
- It is based on all the observations.
- It is well defined.
- It is least affected by sampling fluctuations.
- It can be used for further statistical analysis.

Limitations

- The mean is unduly affected by the extreme items (outliers).
- It cannot be determined for the qualitative data such as beauty, honesty etc.
- It cannot be located by observations on the graphic method.

When to use?

Arithmetic mean is a best representative of the data if the data set is homogeneous. On the other hand if the data set is heterogeneous the result may be misleading and may not represent the data.

Weighted Arithmetic Mean

The arithmetic mean, as discussed earlier, gives equal importance (or weights) to each observation in the data set. However, there are situations in which values of individual observations in the data set are not of equal importance. Under these circumstances, we may attach, a weight, as an indicator of their importance to each observation value.

Definition

Let x_1, x_2, \dots, x_n be the set of n values having weights w_1, w_2, \dots, w_n respectively, then the weighted mean is,

$$\bar{x}_w = \frac{w_1 x_1 + w_2 x_2 + \dots + w_n x_n}{w_1 + w_2 + \dots + w_n} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$$

Uses of weighted arithmetic mean

Weighted arithmetic mean is used in:

- The construction of index numbers.
- Comparison of results of two or more groups where number of items in the groups differs.
- Computation of standardized death and birth rates.

Example 5.5

The weights assigned to different components in an examination or Component Weightage
Marks scored

Component	Weightage	Marks scored
Theory	4	60
Practical	3	80
Assignment	1	90
Project	2	75
	10	

Calculate the weighted average score of the student who scored marks as given in the table

Solution:

Component	Marks scored (x_i)	Weightage (w_i)	$w_i x_i$
Theory	60	4	240
Practical	80	3	240
Assignment	90	1	90
Project	75	2	150
Total		10	720

$$\begin{aligned}
 \text{Weighted average, } \bar{x} &= \frac{\sum w_i x_i}{\sum w_i} \\
 &= 720/10 \\
 &= 72
 \end{aligned}$$

Combined Mean:

Let \bar{x}_1 and \bar{x}_2 are the arithmetic mean of two groups (having the same unit of measurement of a variable), based on n_1 and n_2 observations respectively. Then the combined mean can be calculated using

$$\text{Combined Mean} = \bar{x}_{12} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$$

Remark : The above result can be extended to any number of groups.

Example 5.6

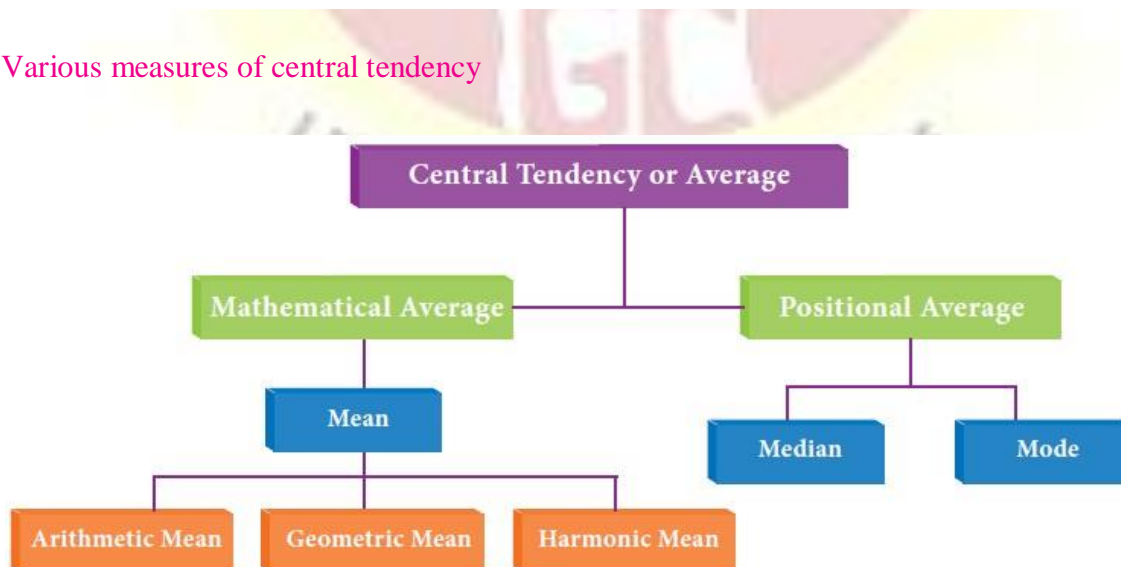
A class consists of 4 boys and 3 girls. The average marks obtained by the boys and girls are 20 and 30 respectively. Find the class average.

Solution:

$$n_1 = 4, \bar{x}_1 = 20, n_2 = 3, \bar{x}_2 = 30$$

$$\begin{aligned} \text{Combined Mean} &= \bar{x}_{12} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2} \\ &= \left[\frac{4 \times 20 + 3 \times 30}{4 + 3} \right] \\ &= \left[\frac{80 + 90}{7} = \frac{170}{7} \right] = 24.3 \end{aligned}$$

Various measures of central tendency



Median

The **median** is the middle number in a sorted, ascending or descending, list of numbers and can be more descriptive of that data set than the average. ... If there is an odd amount of numbers, the **median** value is the number that is in the middle, with the same amount of numbers below and above.

Question 1:

Find out the median.

S. No.	1	2	3	4	5	6	7	8	9
Marks Obtained	10	12	14	17	18	20	21	30	32

ANSWER:

10, 12, 14, 17, 18, 20, 21, 30, 32

$N = 9$

Median = $(N+1) \div 2$ th item
 Median = $(9+1) \div 2 = 5$ th item
 Median = 5th item

Thus, Median is given by the size of the 5th item. Therefore, Median of the data so given is 18.

Question 2:

Find the value of the median from the following data: 15, 35, 48, 46, 50, 43, 55, 49.

ANSWER:

First of all, we need to arrange the data in an ascending order. Thus, the data is presented below as:

15, 35, 43, 46, 48, 49, 50, 55

$N = 8$

Since the number of items in the series is even, thus the following formula is used to calculate the median.

$$\begin{aligned} \text{Median} &= \text{Size of } (N/2)\text{th item} + \text{Size of } (N/2+1)\text{th item} \\ &= \text{Size of } (82)\text{th item} + \text{Size of } (82+1)\text{th item} \\ &= \text{Size of 4th item} + \text{Size of 5th item} \\ &= 46 + 482 = 47 \end{aligned}$$

Thus, Median is 47.

Thus, Median is 47.

Question 3:

Calculate the value of median: 25, 20, 15, 45, 18, 7, 10, 64, 38, 12.

ANSWER:

First of all, we need to arrange the data in an ascending order. Thus, the data is presented below as:

7, 10, 12, 15, 18, 20, 25, 38, 45, 64

$N=10$

Since the number of items in the series is even, thus the following formula is used to calculate the median.

$$\begin{aligned} \text{Median} &= \text{Size of } (N/2)\text{th item} + \text{Size of } (N/2+1)\text{th item} \\ &= \text{Size of } (10/2)\text{th item} + \text{Size of } (10/2+1)\text{th item} \\ &= \text{Size of 5th item} + \text{Size of 6th item} \\ &= 18 + 20 = 19 \end{aligned}$$

Thus, Median is 19.

Question 7:

Find out the median size from the following:

Size	10–20	20–30	30–40	40–50
Frequency	42	25	58	40

ANSWER:

Size	Frequency (<i>f</i>)	Cumulative Frequency (<i>c.f.</i>)
10 – 20	42	42
20 – 30	25	67
30 – 40	58 (<i>f</i>)	125
40 – 50	40	165
	$N = \sum f = 165$	$N = \sum f = 165$

Median class is given by the size of $(\frac{N}{2})^{\text{th}}$ item, i.e. $(\frac{165}{2})^{\text{th}}$ item, which is 82.5^{th} item.

This corresponds to the class interval of 30 – 40, so this is the median class.
 $\text{Median} = l_1 + \frac{N/2 - c.f.}{f} \times i$, $\text{Median} = 30 + \frac{82.5 - 67}{58} \times 10$ or, $\text{Median} = 30 + 2.67 = 32.67$ Thus, $\text{Median} = 32.67$

MODE

The mode is the value that appears most often in a set of data values. If X is a discrete random variable, the mode is the value x at which the probability mass function takes its maximum value. In other words, it is the value that is most likely to be sample.

Mode is the most frequent number — that is, the number that occurs the highest number of times.

Index	New York City	Los Angeles
1	\$ 1.00	\$ 1.00
2	\$ 2.00	\$ 2.00
3	\$ 3.00	\$ 3.00
4	\$ 3.00	\$ 4.00
5	\$ 5.00	\$ 5.00
6	\$ 6.00	\$ 6.00
7	\$ 7.00	\$ 7.00
8	\$ 8.00	\$ 8.00
9	\$ 9.00	\$ 9.00
10	\$ 11.00	\$ 10.00
11	\$ 66.00	

Finding **Mode** of pizza prices in NY and LA

For the data set of NY, you can see \$3.00 appears twice and it has the most appearance.
Then **Mode of pizza prices in NY is \$3.00**

Case 1: Ungrouped Data

For ungrouped data, we just need to identify the observation which occurs maximum times.

Mode = Observation with maximum frequency

For example in the data: 6, 8, 9, 3, 4, 6, 7, 6, 3 the value 6 appears the most number of times.

Thus, mode = 6.

An easy way to remember mode is: Most Often Data Entered.

Note: A data may have no mode, 1 mode or more than 1 mode.

Depending upon the number of modes the data has, it can be called unimodal, bimodal, trimodal or multimodal.

The example discussed above has only 1 mode, so it is unimodal.

Case 2: Grouped Data

When the data is continuous, the mode can be found using the following steps:

Step 1: Find modal class i.e. the class with maximum frequency.

Step 2: Find mode using the following formula:

$$\text{Mode} = l + \left[\frac{f_m - f_1}{2f_m - f_1 - f_2} \right] \times h$$

where, l = lower limit of modal class,

f_m = frequency of modal class,

f_1 = frequency of class preceding modal class,

f_2 = frequency of class succeeding modal class,

h = class width

Consider the following example to understand the formula.

Example 1

Find the mode of the given data:

Marks Obtained	0-20	20-40	40-60	60-80	80-100
Number of students	5	10	12	6	3

Solution

The highest frequency = 12, so the modal class is 40-60.

l = lower limit of modal class = 40

f_m = frequency of modal class = 12

f_1 = frequency of class preceding modal class = 10

f_2 = frequency of class succeeding modal class = 6

h = class width = 20

Using the mode formula,

$$\text{Mode} = l + \left[\frac{f_m - f_1}{2f_m - f_1 - f_2} \right] \times h = 40 + \left[\frac{12 - 10}{2 \times 12 - 10 - 6} \right] \times 20 = 40 + \left[\frac{2}{8} \right] \times 20 = 45$$

\therefore Mode = 45

UNIT III

Measures of Dispersion

Range :

The **Range** is the difference between the lowest and highest values. **Example:** In {4, 6, 9, 3, 7} the lowest value is 3, and the highest is 9. So the **range** is $9 - 3 = 6$.

How to Find a Range

Example question 1: What is the range for the following set of numbers? 10, 99, 87, 45, 67, 43, 45, 33, 21, 7, 65, 98?

Step 1: Sort the numbers in order, from smallest to largest:
7, 10, 21, 33, 43, 45, 45, 65, 67, 87, 98, 99

Step 2: Subtract the smallest number in the set from the largest number in the set:

$$99 \quad - \quad 7 \quad = \quad 92$$

The range is 92

That's it!

Example question 2: What is the range of these [integers](#)?
14, -12, 7, 0, -5, -8, 17, -11, 19

Step 1: Sort the numbers in order, from smallest to largest:
-12, -11, -8, -5, 0, 7, 14, 17, 19

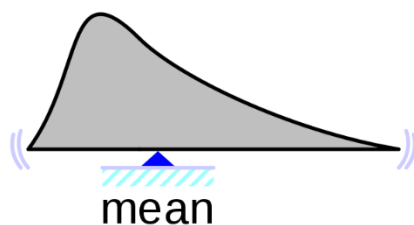
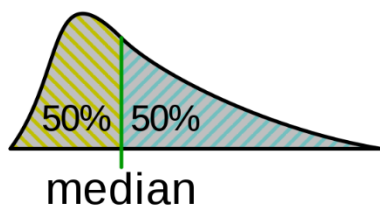
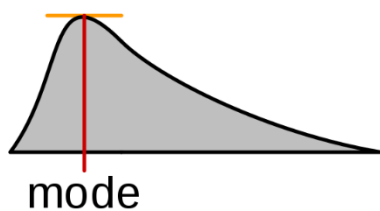
Step 2: Subtract the smallest number in the set from the largest number in the set:
 $19 - (-12) = 19 + 12 = 31$

The range is 31.

Mean Deviation

Mean Deviation Formula

The mean deviation is the mean of the absolute deviations of the observations or values from a suitable average. This suitable average may be the mean, [median or mode](#). We also know it as the mean absolute deviation. We shall now learn more about some important formulas, for example, the mean deviation formula for an individual series or a continuous series, etc.



1] Individual Series

$$\text{M.D.} = \frac{\sum |X - \overline{X}|}{N}$$

Where,

\sum	Summation
X	Observations or values
\overline{X}	Mean
N	Number of observations

2] Discrete Series

$$\text{M.D.} = \frac{\sum f |X - \overline{X}|}{\sum f}$$

Where,

\sum	Summation
X	Observations or values
\overline{X}	Mean
f	frequency of observations

3] Continuous Series

$$\text{M.D.} = \frac{\sum f |X - \overline{X}|}{\sum f}$$

Where,

\sum	Summation
X	Mid-value of the class
\overline{X}	Mean
f	frequency of observations

Mean deviation from Median

1] Individual Series

$$\text{M.D.} = \frac{\sum |X - M|}{N}$$

Where,

\sum	Summation
X	Observations or values
M	Median
N	Number of observations

2] Discrete Series

$$\text{M.D.} = \frac{\sum f |X - M|}{\sum f}$$

Where,

\sum	Summation
X	Observations or values
M	Median
f	frequency of observations

3] Continuous Series

$$\text{M.D.} = \frac{\sum f | X - \overline{X} |}{\sum f}$$

Where,

\sum	Summation
X	Mid-value of the class
M	Median
f	frequency of observations

Mean deviation from Mode

1] Individual Series

$$\text{M.D.} = \frac{\sum | X - \text{Mode} |}{N}$$

Where,

\sum	Summation
X	Observations or values
M	Mode
N	Number of observations

2] Discrete Series

$$M.D. = \frac{\sum f | X - Mode |}{\sum f}$$

Where,

\sum	Summation
X	Observations or values
Mode	Mode
f	frequency of observations

3] Continuous Series

$$M.D. = \frac{\sum f | X - Mode |}{\sum f}$$

Where,

\sum	Summation
X	Mid-value of the class
Mode	Mode
f	frequency of observations

Steps to Calculate the Mean Deviation:

1. Calculate the mean, median or mode of the series.
2. Calculate the deviations from the Mean, median or mode and ignore the minus signs.
3. Multiply the deviations with the frequency. This step is necessary only in the discrete and continuous series.
4. Sum up all the deviations.
5. Apply the formula.

The formula for the Co-efficient of Mean Deviation

- Co-efficient of Mean Deviation from Mean = $\frac{\text{M.D.}}{\overline{X}} XM.D.$
- Co-efficient of Mean Deviation from Median = $\frac{\text{M.D.}}{M} MM.D.$
- The Co-efficient of Mean Deviation from Mode = $\frac{\text{M.D.}}{\text{Mode}} ModeM.D.$

Solved Examples

Q.1. Calculate the mean deviation from the median and the co-efficient of mean deviation from the following data:

Marks of the students: 86, 25, 87, 65, 58, 45, 12, 71, 35.

Solution: Arrange the data in ascending order: 12, 25, 35, 45, 58, 65, 71, 86, 87.

Median = Value of the $\frac{N+1}{2}^{\text{th}}$ term $2N+1^{\text{th}}$ term

= Value of the $\frac{9+1}{2}^{\text{th}}$ term = 58 $29+1^{\text{th}}$ term = 58

Calculation of mean deviation:

X	$\left X - M \right $	$ X-M $
12	46	
25	33	
35	23	
45	13	
58	0	
65	7	
71	13	
86	28	
87	29	
N = 9	$\sum \left X - M \right = 460$	$\sum X-M =460$

$$M.D. = \frac{\sum \left| X - M \right|}{N} = \frac{460}{9} = 51.11$$

$$= \frac{460}{9} = 51.11$$

$$= 51.11$$

$$\text{Co-efficient of Mean Deviation from Median} = \frac{M.D.}{M} = \frac{51.11}{58} = 0.881$$

$$= \frac{51.11}{58} = 0.881$$

$$= 0.881$$

Quartile Deviation

What is Quartile Deviation?



Quartile deviation is based on the difference between the first quartile and the third quartile in the frequency distribution and the difference is also known as the interquartile range, the difference divided by two is known as quartile deviation or semi interquartile range.


When one takes half of the difference or variance between the 3rd quartile and the 1st quartile of a simple distribution or frequency distribution is the quartile deviation.


Formula

A Quartile Deviation (Q.D.) formula is used in statistics to measure spread or, in other words, to measure dispersion. This can also be called a Semi Inter-Quartile Range.

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

Quartile Deviation Formula = $\frac{Q_3 - Q_1}{2}$




- The formula includes Q3 and Q1 in the calculation, which is the top 25% and lower 25% of data respectively, and when the difference is taken between these two and when this number is halved, it gives measures of spread or dispersion.
- So, to calculate Quartile deviation, you need to first find out Q1, then the second step is to find Q3 and then make a difference of both, and the final step is to divide by 2.
- This is one of the best methods of dispersion for open-ended data.

Examples

You can download this Quartile Deviation Formula Excel Template here – [Quartile Deviation Formula Excel Template](#)

Example #1

Consider a data set of following numbers: 22, 12, 14, 7, 18, 16, 11, 15, 12. You are required to calculate the Quartile Deviation.

Solution:

First, we need to arrange data in ascending order to find Q3 and Q1 and avoid any duplicates.

7, 11, 12, 13, 14, 15, 16, 18, 22

Calculation of Q1 can be done as follows,

$$Q1 = \frac{1}{4} (9 + 1)$$

$$= \frac{1}{4} (10)$$

Q1=2.5 Term

Calculation of Q3 can be done as follows,

$$Q3 = \frac{3}{4} (9 + 1)$$

$$= \frac{3}{4} (10)$$

Q3= 7.5 Term

Calculation of quartile deviation can be done as follows,

- Q1 is an average of 2nd, which is 11 and adds the difference between 3rd & 4th and 0.5, which is $(12-11)*0.5 = 11.50$.

- Q3 is the 7th term and product of 0.5, and the difference between the 8th and 7th term, which is $(18-16)*0.5$, and the result is $16 + 1 = 17$.

$$\mathbf{Q.D. = Q3 - Q1 / 2}$$

Using the quartile deviation formula, we have $(17-11.50) / 2$

$$=5.5/2$$

$$\mathbf{Q.D.=2.75.}$$

Example #2

Harry Ltd. is a textile manufacturer and is working upon a reward structure. The management is in discussion to start a new initiative, but they first want to know how much their production spread is.

The management has collected its average daily production data for the last 10 days per (average) employee.

155, 169, 188, 150, 177, 145, 140, 190, 175, 156.

Use the Quartile Deviation formula to help management find dispersion.

Solution:

The number of observations here is 10, and our first step would be to arrange data in ascending order.

140, 145, 150, 155, 156, 169, 175, 177, 188, 190

Calculation of Q1 can be done as follows,

$$Q1 = \frac{1}{4} (n+1)\text{th term}$$

$$= \frac{1}{4} (10+1)$$

$$= \frac{1}{4} (11)$$

$$\mathbf{Q1 = 2.75^{\text{th}} \text{ Term}}$$

Calculation of Q3 can be done as follows,

$$Q3 = \frac{3}{4} (n+1)\text{th term}$$

$$= \frac{3}{4} (11)$$

$$\mathbf{Q3 = 8.25 \text{ Term}}$$

Calculation of quartile deviation can be done as follows,

- 2nd term is 145 and now adding to this $0.75 * (150 - 145)$ which is 3.75, and the result is 148.75
- 8th term is 177 and now adding to this $0.25 * (188 - 177)$ which is 2.75, and the result is 179.75

$$\text{Q.D.} = \text{Q3} - \text{Q1} / 2$$

Using the quartile deviation formula, we have $(179.75 - 148.75) / 2$
 $= 31 / 2$

$$\text{Q.D.} = 15.50.$$

Standard Deviation :

In statistics, the standard deviation is a measure of the amount of variation or dispersion of a set of values. A low standard deviation indicates that the values tend to be close to the mean of the set, while a high standard deviation indicates that the values are spread out over a wider range.

This is the formula for Standard Deviation:

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2}$$

The Formula Explained

First, let us have some example values to work on:



Example: Sam has 20 Rose Bushes.

The number of flowers on each bush is

9, 2, 5, 4, 12, 7, 8, 11, 9, 3, 7, 4, 12, 5, 4, 10, 9, 6, 9, 4

Work out the Standard Deviation.

Step 1. Work out the mean

In the formula above μ (the greek letter "mu") is the **mean** of all our values ...

Example: 9, 2, 5, 4, 12, 7, 8, 11, 9, 3, 7, 4, 12, 5, 4, 10, 9, 6, 9, 4

The mean is:

2020So:

$$\mu = 7$$

Step 2. Then for each number: subtract the Mean and square the result

This is the part of the formula that says:

$$(x_i - \mu)^2$$

So what is x_i ? They are the individual x values 9, 2, 5, 4, 12, 7, etc...

In other words $x_1 = 9$, $x_2 = 2$, $x_3 = 5$, etc.

So it says "for each value, subtract the mean and square the result", like this

Example (continued):

$$(9 - 7)^2 = (2)^2 = 4$$

$$(2 - 7)^2 = (-5)^2 = 25$$

$$(5 - 7)^2 = (-2)^2 = 4$$

$$(4 - 7)^2 = (-3)^2 = 9$$

$$(12 - 7)^2 = (5)^2 = \mathbf{25}$$

$$(7 - 7)^2 = (0)^2 = \mathbf{0}$$

$$(8 - 7)^2 = (1)^2 = \mathbf{1}$$

... etc ...

And we get these results:

Step 3. Then work out the mean of those squared differences.

To work out the mean, **add up all the values** then **divide by how many**.

First add up all the values from the previous step.

But how do we say "add them all up" in mathematics? We use "Sigma": Σ

The handy **Sigma Notation** says to sum up as many terms as we want:

Sigma Notation

We want to add up all the values from 1 to N, where N=20 in our case because there are 20 values:

Example (continued):

$$\sum_{i=1}^N (x_i - \mu)^2$$

Which means: Sum all values from $(x_1-7)^2$ to $(x_N-7)^2$

We already calculated $(x_1-7)^2=4$ etc. in the previous step, so just sum them up:

$$= 4+25+4+9+25+0+1+16+4+16+0+9+25+4+9+9+4+1+4+9 = \mathbf{178}$$

But that isn't the mean yet, we need to **divide by how many**, which is done by **multiplying by $1/N$** (the same as dividing by N):

Example (continued):

$$\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$$

Mean of squared differences = $(1/20) \times 178 = 8.9$

(Note: this value is called the "Variance")

Step 4. Take the square root of that:

Example (concluded):

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2}$$

$$\sigma = \sqrt{(8.9)} = 2.983...$$

Coefficient of variation

The **coefficient of variation (CV)** is a statistical measure of the relative dispersion of data points in a data series around the mean. In finance, the **coefficient of variation** allows investors to determine how much volatility, or risk, is assumed in comparison to the amount of return expected from investments.

Comparison of two data in terms of measures of central tendencies and dispersions in some cases will not be meaningful, because the variables in the data may not have same units of measurement.

For example consider the two data

	Weight	Price
Mean	8 kg	₹ 85
Standard deviation	1.5 kg	₹ 21.60

Here we cannot compare the standard deviations 1. 5kg and ₹21.60. For comparing two or more data for corresponding changes the relative measure of standard deviation, called “**Coefficient of variation**” is used.

Coefficient of variation of a data is obtained by dividing the standard deviation by the arithmetic mean. It is usually expressed in terms of percentage. This concept is suggested by one of the most prominent **Statistician Karl Pearson**.

Thus, coefficient of variation of first data (C.V1) = $\sigma_1/x_1 \times 100\%$

$$= \frac{\sigma_1}{\bar{x}_1} \times 100\%$$

and coefficient of variation of second data (C.V2) = $\sigma_2/x_2 \times 100\%$

$$= \frac{\sigma_2}{\bar{x}_2} \times 100\%$$

The data with lesser coefficient of variation is more consistent or stable than the other data. Consider the two data

A	500	900	800	900	700	400
B	300	540	480	540	420	240

Then, we get

	Mean	Standard deviation
A	700	191.5
B	420	114.9

If we compare the mean and standard deviation of the two data, we think that the two datas are entirely different. But mean and standard deviation of B are 60% of that of A. Because of the smaller mean the smaller standard deviation led to the misinterpretation.

To compare the dispersion of two data, coefficient of variation = $\sigma/x \times 100\%$

$$= \frac{\sigma}{\bar{x}} \times 100\%$$

The coefficient of variation of A = $191.5/700 \times 100\% = 27.4\%$

The coefficient of variation of $B = 114.9/420 \times 100\% = 27.4\%$

Thus the two data have equal coefficient of variation. Since the data have equal coefficient of variation values, we can conclude that one data depends on the other. But the data values of B are exactly 60% of the corresponding data values of A . So they are very much related. Thus, we get a confusing situation.

To get clear picture of the given data, we can find their coefficient of variation. This is why we need coefficient of variation.

Example 8.15

The mean of a data is 25.6 and its coefficient of variation is 18.75. Find the standard deviation.

Solution

Mean $\bar{x} = 25.6$, Coefficient of variation, C.V. = 18.75

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

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

$$18.75 = \frac{\sigma}{25.6} \times 100 ; \quad \sigma = 4.8$$

Example 8.16

The following table gives the values of mean and variance of heights and weights of the 10th standard students of a school.

	Height	Weight
Mean	155 cm	46.50 kg ²
Variance	72.25 cm ²	28.09 kg ²

Which is more varying than the other?

Solution

For comparing two data, first we have to find their coefficient of variations

Mean $\bar{x}_1 = 155\text{cm}$, variance $\sigma_1^2 = 72.25 \text{ cm}^2$

Therefore standard deviation $\sigma_1 = 8.5$

Coefficient of variation

$$C.V_1 = \frac{\sigma_1}{\bar{x}_1} \times 100\%$$

$$C.V_1 = \frac{8.5}{155} \times 100\% = 5.48\% \quad (\text{for heights})$$

Mean $\bar{x}_2 = 46.50 \text{ kg}$, Variance $\sigma_2^2 = 28.09 \text{ kg}^2$

Standard deviation $\sigma_2 = 5.3 \text{ kg}$

Coefficient of variation

$$C.V_2 = \frac{\sigma_2}{\bar{x}_2} \times 100\%$$

$$C.V_2 = \frac{5.3}{46.50} \times 100\%$$

= 11.40% (for weights)

$C.V_1 = 5.48\%$ and $C.V_2 = 11.40\%$

Since $C.V_2 > C.V_1$, the weight of the students is more varying than the height.

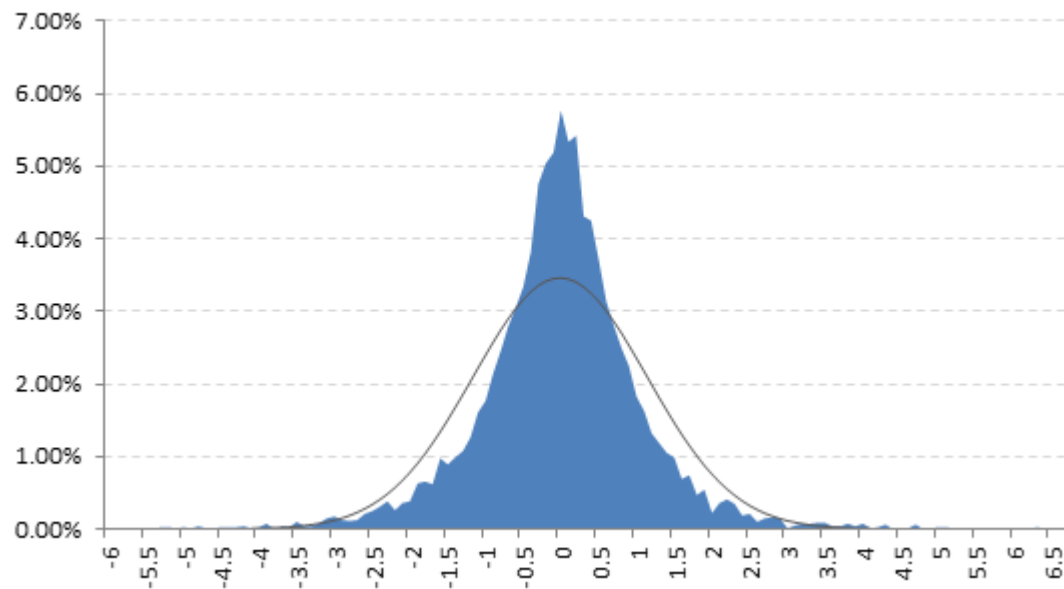
Coefficient of kurtosis

The **coefficient of kurtosis** is used to measure the peakness or flatness of a curve. It is based on the moments of the distribution. This **coefficient** is one of the measures of **kurtosis**

The **coefficient of kurtosis**, or simply kurtosis, **measures the peakedness of a distribution.**

High kurtosis means that values close to the mean are relatively more frequent and extreme values (very far from the mean) are also relatively more frequent. The values in between are relatively less frequent. If you plot a frequency histogram or another chart showing frequency of such distribution, it would have a sharp peak in the middle and fat tails.

Frequency vs. Normal Distribution



Conversely, low coefficient of kurtosis means that a distribution is less peaked and has thinner tails.

Coefficient of Kurtosis Formula

Coefficient of Kurtosis for a Population

$$K = n \frac{\sum_{i=1}^n (X_i - X_{avg})^4}{(\sum_{i=1}^n (X_i - X_{avg})^2)^2}$$

Coefficient of Kurtosis for a Sample

$$K = \frac{n(n+1)(n-1)}{(n-2)(n-3)} \frac{\sum_{i=1}^n (X_i - X_{avg})^4}{(\sum_{i=1}^n (X_i - X_{avg})^2)^2}$$

You can see a more detailed explanation of the formulas and their underlying logic [here](#)

Excess Kurtosis

Kurtosis is often measured and quoted in the form which is kurtosis relative to normal distribution. The coefficient of kurtosis for normal distribution is 3, therefore **excess kurtosis equals coefficient of kurtosis less 3**.

UNIT - IV:

Probability of an event

In an experiment, the **probability of an event** is the likelihood of that **event** occurring. **Probability** is a value between (and including) zero and one. If $P(E)$ represents the **probability of an event E**, then: $P(E) = 0$ if and only if E is an impossible **event**..

Probability Formulas

The probability formula is used to compute the probability of an event to occur. To recall, the likelihood of an event happening is called probability. When a random experiment is entertained, one of the first questions that come in our mind is: What is the probability that a certain event occurs? A probability is a chance of prediction. When we assume that, let's say, x be the chances of happening an event then at the same time $(1-x)$ are the chances for "not happening" of an event.

Similarly, if the probability of an event occurring is "a" and an independent probability is "b", then the probability of both the event occurring is "ab". We can use the formula to find the chances of an event happening.

Formula to Calculate Probability

The formula of the probability of an event is:

$$P(A) = \frac{\text{Number of Favourable Outcome}}{\text{Total Number of Favourable Outcomes}}$$

Probability Formula

Or,

$$P(A) = n(A)/n(S)$$

Where,

- $P(A)$ is the probability of an event “A”
- $n(A)$ is the number of favourable outcomes
- $n(S)$ is the total number of events in the sample space

Note: Here, the favourable outcome means the outcome of interest.

Sometimes, students get confused about the word “favourable outcome” with “desirable outcome”. In some of the requirements, losing in a certain test or occurrence of an undesirable outcome can be a favourable event for the experiments run.

Basic Probability Formulas

Let A and B are two events. The probability formulas are listed below:

All Probability Formulas List in Maths	
Probability Range	$0 \leq P(A) \leq 1$
Rule of Addition	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
Rule of Complementary Events	$P(A') + P(A) = 1$

All Probability Formulas List in Maths	
Disjoint Events	$P(A \cap B) = 0$
Independent Events	$P(A \cap B) = P(A) \cdot P(B)$
Conditional Probability	$P(A B) = P(A \cap B) / P(B)$
Bayes Formula	$P(A B) = P(B A) \cdot P(A) / P(B)$

Example Questions Using Probability Formulas

Example 1: What is the probability that a card taken from a standard deck, is an Ace?

Solution:

Total number of cards a standard pack contains = 52

Number of Ace cards in a deck of cards = 4

So, the number of favourable outcomes = 4

Now, by looking at the formula,

Probability of selecting an ace from a deck is,

$P(\text{Ace}) = (\text{Number of favourable outcomes}) / (\text{Total number of favourable outcomes})$

$P(\text{Ace}) = 4/52$

$= 1/13$

So we can say that the probability of getting an ace is 1/13.

Example 2: Calculate the probability of getting an odd number if a dice is rolled.

Solution:

Sample space (S) = {1, 2, 3, 4, 5, 6}

$n(S) = 6$

Let “E” be the event of getting an odd number, $E = \{1, 3, 5\}$

$n(E) = 3$

So, the Probability of getting an odd number is:

$P(E) = (\text{Number of outcomes favorable})/(\text{Total number of outcomes})$

$= n(E)/n(S)$

$= 3/6$

$= 1/2$

Addition Theorem of Probability

(i) If A and B are any two events then

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

(ii) If A, B and C are any three events then

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(A \cap C) + P(A \cap B \cap C)$$

Proof

(i) Let A and B be any two events of a random experiment with sample space S .

From the Venn diagram, we have the events only A , $A \cap B$ and only B are mutually exclusive and their union is $A \cup B$

Therefore, $P(A \cup B) = P[(\text{only } A) \cup (A \cap B) \cup (\text{only } B)]$

$= P(\text{only } A) + P(A \cap B) + P(\text{only } B)$

$$= [P(A) - P(A \cap B)] + P(A \cap B) + [P(B) - P(A \cap B)]$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

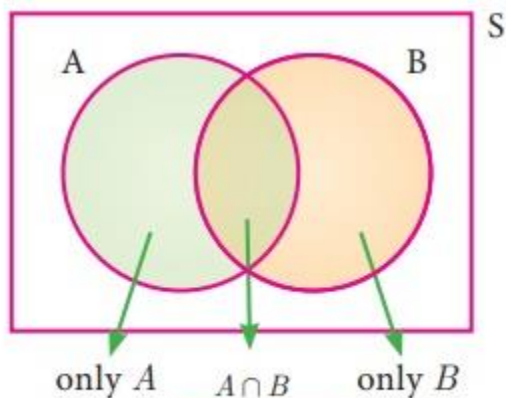


Fig. 8.10

(ii) Let A, B, C are any three events of a random experiment with sample space S .

Let $D = B \cup C$

$$P(A \cup B \cup C) = P(A \cup D)$$

$$= P(A) + P(D) - P(A \cap D)$$

$$= P(A) + P(B \cup C) - P[A \cap (B \cup C)]$$

$$= P(A) + P(B) + P(C) - P(B \cap C) - P[(A \cap B) \cup (A \cap C)]$$

$$= P(A) + P(B) + P(C) - P(B \cap C) - P(A \cap B) - P(A \cap C) + P[(A \cap B) \cap (A \cap C)]$$

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(C \cap A) + P(A \cap B \cap C)$$

Example 8.27

If $P(A) = 0.37$, $P(B) = 0.42$, $P(A \cap B) = 0.09$ then find $P(A \cup B)$.

Solution

$$P(A) = 0.37, P(B) = 0.42, P(A \cap B) = 0.09$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(A \cup B) = 0.37 + 0.42 - 0.09 = 0.7$$

Example 8.28

What is the probability of drawing either a king or a queen in a single draw from a well shuffled pack of 52 cards?

Solution

Total number of cards = 52

Number of king cards = 4

Probability of drawing a king card = $4/52$

Number of queen cards = 4

Probability of drawing a queen card = $4/52$

Both the events of drawing a king and a queen are mutually exclusive

$$\Rightarrow P(A \cup B) = P(A) + P(B)$$

Therefore, probability of drawing either a king or a queen = $4/52 + 4/52 = 2/13$

Example 8.29

Two dice are rolled together. Find the probability of getting a doublet or sum of faces as 4.

Solution

When two dice are rolled together, there will be $6 \times 6 = 36$ outcomes. Let S be the sample space.

Then $n(S) = 36$

Let A be the event of getting a doublet and B be the event of getting face sum 4.

Then $A = \{(1,1),(2,2),(3,3),(4,4),(5,5),(6,6)\}$

$B = \{(1,3),(2,2),(3,1)\}$

Therefore, $A \cap B = \{(2,2)\}$

Then, $n(A) = 6$, $n(B) = 3$, $n(A \cap B) = 1$.

$$P(A) = \frac{n(A)}{n(S)} = \frac{6}{36}$$

$$P(B) = \frac{n(B)}{n(S)} = \frac{3}{36}$$

$$P(A \cap B) = \frac{n(A \cap B)}{n(S)} = \frac{1}{36}$$

Therefore, $P(\text{getting a doublet or a total of 4}) = P(A \cup B)$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$= \frac{6}{36} + \frac{3}{36} - \frac{1}{36} = \frac{8}{36} = \frac{2}{9}$$

Hence, the required probability is $\frac{2}{9}$.

Example 8.30

If A and B are two events such that $P(A) = \frac{1}{4}$, $P(B) = \frac{1}{2}$ and $P(A \text{ and } B) = \frac{1}{8}$, find

(i) $P(A \text{ or } B)$ (ii) $P(\text{not } A \text{ and not } B)$.

Solution

$$(i) P(A \text{ or } B) = P(A \cup B)$$

$$= P(A) + P(B) - P(A \cap B)$$

$$P(A \text{ or } B) = \frac{1}{4} + \frac{1}{2} - \frac{1}{8} = \frac{5}{8}$$

$$(ii) P(\text{not } A \text{ and not } B) = P(\bar{A} \cap \bar{B})$$

$$= P(\overline{A \cup B})$$

$$= 1 - P(A \cup B)$$





$$P(\text{not } A \text{ and not } B) = 1 - 5/8 = 3/8$$

Example 8.31

A card is drawn from a pack of 52 cards. Find the probability of getting a king or a heart or a red card.

Solution

Total number of cards = 52; $n(S) = 52$

Suits of playing cards	Spade 	Heart 	Clavor 	Diamond 
Cards of each suit	A	A	A	A
	2	2	2	2
	3	3	3	3
	4	4	4	4
	5	5	5	5
	6	6	6	6
	7	7	7	7
	8	8	8	8
	9	9	9	9
	10	10	10	10
	J	J	J	J
	Q	Q	Q	Q
	K	K	K	K
Set of playing cards in each suit	13	13	13	13

Let A be the event of getting a king card. $n(A) = 4$

$$P(A) = \frac{n(A)}{n(S)} = \frac{4}{52}$$

Let B be the event of getting a heart card. $n(B) = 13$

$$P(B) = \frac{n(B)}{n(S)} = \frac{13}{52}$$

Let C be the event of getting a red card. $n(C) = 26$

$$P(C) = \frac{n(C)}{n(S)} = \frac{26}{52}$$

$$P(A \cap B) = P(\text{getting heart king}) = 1/52$$

$$P(B \cap C) = P(\text{getting red and heart}) = 13/52$$

$$P(A \cap C) = P(\text{getting red king}) = 2/52$$

$$P(A \cap B \cap C) = P(\text{getting heart, king which is red}) = 1/52$$

Therefore, required probability is

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(C \cap A) + P(A \cap B \cap C)$$

$$= 4/52 + 13/52 + 26/52 - 1/52 - 13/52 - 2/52 + 1/52 = 28/52$$

$$= \frac{4}{52} + \frac{13}{52} + \frac{26}{52} - \frac{1}{52} - \frac{13}{52} - \frac{2}{52} + \frac{1}{52} = \frac{28}{52} = \frac{7}{13}$$

$$= 7/13$$

Example 8.32

In a class of 50 students, 28 opted for NCC, 30 opted for NSS and 18 opted both NCC and NSS. One of the students is selected at random. Find the probability that

- (i) The student opted for NCC but not NSS.
- (ii) The student opted for NSS but not NCC.
- (iii) The student opted for exactly one of them.

Solution

Total number of students $n(S) = 50$.

Let A and B be the events of students opted for NCC and NSS respectively.

$$n(A) = 28, n(B) = 30, n(A \cap B) = 18$$

$$P(A) = \frac{n(A)}{n(S)} = \frac{28}{50}$$

$$P(B) = \frac{n(B)}{n(S)} = \frac{30}{50}$$

$$P(A \cap B) = \frac{n(A \cap B)}{n(S)} = \frac{18}{50}$$

- (i) Probability of the students opted for NCC but not NSS

$$P(A \cap \bar{B}) = P(A) - P(A \cap B) = \frac{28}{50} - \frac{18}{50} = \frac{10}{50} = \frac{1}{5}$$

- (ii) Probability of the students opted for NSS but not NCC.

$$P(\bar{A} \cap B) = P(B) - P(A \cap B) = \frac{30}{50} - \frac{18}{50} = \frac{12}{50} = \frac{6}{25}$$

(iii) Probability of the students opted for exactly one of them

$$\begin{aligned}
 &= P[(A \cap \bar{B}) \cup (\bar{A} \cap B)] \\
 &= P(A \cap \bar{B}) + P(\bar{A} \cap B) = \frac{1}{5} + \frac{6}{25} = \frac{11}{25}
 \end{aligned}$$

(Note that $(A \cap \bar{B})$, $(\bar{A} \cap B)$ are mutually exclusive events)

Multiplication Theorem of Probability :

- If A and B are two events associated with a random experiment, then $P(A \cap B) = P(A) \cdot P(B/A)$, if $P(A) \neq 0$ or $P(A \cap B) = P(B) \cdot P(A/B)$, if $P(B) \neq 0$.
- Extension of multiplication theorem:
If A_1, A_2, \dots, A_n are n events related to a random experiment, then $P(A_1 \cap A_2 \cap A_3 \cap \dots \cap A_n) = P(A_1) P(A_2/A_1) P(A_3/A_1 \cap A_2) \dots P(A_n/A_1 \cap A_2 \cap \dots \cap A_{n-1})$, where $P(A_i/A_1 \cap A_2 \cap \dots \cap A_{i-1})$, represents the conditional probability of the event A_i , given that the events A_1, A_2, \dots, A_{i-1} have already happened.
- Multiplication theorems for independent events:
If A and B are independent events associated with a random experiment, then $P(A \cap B) = P(A) \cdot P(B)$ i.e., the probability of simultaneous occurrence of two independent events is equal to the product of their probabilities. By multiplication theorem, we have $P(A \cap B) = P(A) \cdot P(B/A)$. Since A and B are independent events, therefore $P(B/A) = P(B)$. Hence, $P(A \cap B) = P(A) \cdot P(B)$.
- Extension of multiplication theorem for independent events:
If A_1, A_2, \dots, A_n are independent events associated with a random experiment, then $P(A_1 \cap A_2 \cap A_3 \cap \dots \cap A_n) = P(A_1) P(A_2) \dots P(A_n)$.
By multiplication theorem, we have $P(A_1 \cap A_2 \cap A_3 \cap \dots \cap A_n) = P(A_1) P(A_2/A_1) P(A_3/A_1 \cap A_2) \dots P(A_n/A_1 \cap A_2 \cap \dots \cap A_{n-1})$
Since $A_1, A_2, \dots, A_{n-1}, A_n$ are independent events, therefore $P(A_2/A_1) = P(A_2)$, $P(A_3/A_1 \cap A_2) = P(A_3)$, \dots , $P(A_n/A_1 \cap A_2 \cap \dots \cap A_{n-1}) = P(A_n)$
Hence, $P(A_1 \cap A_2 \cap A_3 \cap \dots \cap A_n) = P(A_1) P(A_2) \dots P(A_n)$.

Probability of at least one of the n independent events:

If p_1, p_2, \dots, p_n be the probabilities of happening of n independent events A_1, A_2, \dots, A_n respectively, then

(i) Probability of happening none of them

$$\begin{aligned} &= P(\bar{A}_1 \cap \bar{A}_2 \cap \bar{A}_3 \dots \cap \bar{A}_n) = P(\bar{A}_1) \cdot P(\bar{A}_2) \cdot P(\bar{A}_3) \dots P(\bar{A}_n). \\ &= (1 - p_1)(1 - p_2)(1 - p_3) \dots (1 - p_n) \end{aligned}$$

(ii) Probability of happening at least one of them

$$\begin{aligned} &= P(A_1 \cup A_2 \cup A_3 \dots \cup A_n) = 1 - P(\bar{A}_1)P(\bar{A}_2)P(\bar{A}_3) \dots P(\bar{A}_n). \\ &= 1 - (1 - p_1)(1 - p_2)(1 - p_3) \dots (1 - p_n) \end{aligned}$$

(iii) Probability of happening of first event and not happening of the remaining = $P(A_1)P(\bar{A}_2)P(\bar{A}_3) \dots P(\bar{A}_n)$

$$= p_1(1 - p_2)(1 - p_3) \dots (1 - p_n)$$

Theorem: If A and B are two independent events, then the **probability** that both will occur is equal to the product of their individual **probabilities**.

1. Solved

Problems:

Conditional Probability

In die and coin problems, unless stated otherwise, it is assumed coins and dice are fair and repeated trials are independent.

Problem

You purchase a certain product. The manual states that the lifetime TT of the product, defined as the amount of time (in years) the product works properly until it breaks down, satisfies

$$P(T \geq t) = e^{-t/5}, \text{ for all } t \geq 0. P(T > t) = e^{-t/5}, \text{ for all } t > 0.$$

For example, the probability that the product lasts more than (or equal to) 22 years is $P(T \geq 2) = e^{-25} = 0.6703$. I purchase the product and use it for two years without any problems. What is the probability that it breaks down in the third year?

You purchase a certain product. The manual states that the lifetime TT of the product, defined as the amount of time (in years) the product works properly until it breaks down, satisfies

$$P(T \geq t) = e^{-t/5}, \text{ for all } t \geq 0.$$

Let A be the event that a purchased product breaks down in the third year. Also, let B be the event that a purchased product does not break down in the first two years. We are interested in $P(A|B)$. We have

$$\begin{aligned} P(B) &= P(T \geq 2) = e^{-25} \\ &= e^{-25}. \end{aligned}$$

We also have

$$\begin{aligned} P(A) &= P(2 \leq T < 3) \\ &= P(T < 3) - P(T < 2) = e^{-35} - e^{-25} \\ &= e^{-35} - e^{-25}. \end{aligned}$$

Finally, since $A \subset B$, we have $A \cap B = A$. Therefore,

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} = \frac{P(A)}{P(B)} \\ &= \frac{e^{-35} - e^{-25}}{e^{-25}} \\ &= e^{-10} - 1. \end{aligned}$$

$$=0.1813$$

You toss a fair coin three times:

- What is the probability of three heads, HHHHHH?
- What is the probability that you observe exactly one heads?
- Given that you have observed *at least* one heads, what is the probability that you observe at least two heads?

Solution

- We assume that the coin tosses are independent.
 - $P(\text{HHH})=P(\text{H})\cdot P(\text{H})\cdot P(\text{H})=0.5^3=1/8$
 - To find the probability of exactly one heads, we can write

$$\begin{aligned} P(\text{One heads}) &= P(\text{HTT} \cup \text{THT} \cup \text{TTH}) = P(\text{HTT}) + P(\text{THT}) + P(\text{TTH}) \\ &= P(\text{HTT}) + P(\text{THT}) + P(\text{TTH}) \\ &= 1/8 + 1/8 + 1/8 = 3/8 = 0.375 \end{aligned}$$

-
- Given that you have observed *at least* one heads, what is the probability that you observe at least two heads? Let A_1 be the event that you observe at least one heads, and A_2 be the event that you observe at least two heads. Then

$$A_1 = S - \{\text{TTT}\}, \text{ and } P(A_1) = 7/8; A_2 = \{\text{HHT}, \text{HTH}, \text{THH}, \text{HHH}\}, \text{ and } P(A_2) = 4/8.$$

$$P(A_2 | A_1) = \frac{P(A_2 \cap A_1)}{P(A_1)} = \frac{3/8}{7/8} = 3/7.$$

Thus, we can write

$$P(A_2|A_1)P(A_1) = P(A_2 \cap A_1)P(A_1) = P(A_2 \cap A_1)P(A_1)$$

$$= P(A_2)P(A_1) = P(A_2)P(A_1)$$

$$= 0.5 \cdot 0.5 = 0.25$$

Bayes Theorem and problems

Solution

- We assume that the coin tosses are independent.

a. $P(HHH) = P(H) \cdot P(H) \cdot P(H) = 0.5^3 = \frac{1}{8}$

- b. To find the probability of exactly one heads, we can write

$$\begin{aligned} P(\text{One heads}) &= P(\text{HTT} \cup \text{THT} \cup \text{TTH}) = P(\text{HTT}) + P(\text{THT}) + P(\text{TTH}) \\ &= \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{3}{8} \end{aligned}$$

- c.

- d. Given that you have observed *at least* one heads, what is the probability that you observe at least two heads? Let A_1 be the event that you observe at least one heads, and A_2 be the event that you observe at least two heads. Then

$$A_1 = S - \{TTT\}, \text{ and } P(A_1) = \frac{7}{8};$$

$$A_2 = \{HHT, HTH, THH, HHH\}, \text{ and } P(A_2) = \frac{4}{8}.$$

$H, HHH\}$, and $P(A_2)=48$.

Thus, we can write

$$\begin{aligned} P(A_2|A_1)P(A_2|A_1) &= P(A_2 \cap A_1)P(A_1) = P(A_2 \cap A_1)P(A_1) \\ &= P(A_2)P(A_1) = P(A_2)P(A_1) \\ &= 48.87 = 47 = 48.87 = 47. \end{aligned}$$

Bayes Theorem and Problems

Bayes' theorem is a formula that describes how to update the probabilities of hypotheses when given evidence. It follows simply from the axioms of conditional probability, but can be used to powerfully reason about a wide range of problems involving belief updates.

Given a hypothesis HH and evidence EE , Bayes' theorem states that the relationship between the probability of the hypothesis before getting the evidence $P(H)P(H)$ and the probability of the hypothesis after getting the evidence $P(H \mid E)P(H|E)$ is

$$P(H \mid E) = \frac{P(E \mid H)}{P(E)} P(H). P(H|E) = P(E)P(E|H)P(H).$$

Many modern techniques rely on Bayes' theorem. For instance, spam filters use Bayesian updating to determine whether an email is real or spam, given the words in the email. Additionally, many specific techniques in statistics, such as calculating or , are best described in terms of how they contribute to updating hypotheses using Bayes' theorem.

Bayes' Theorem Example #1

You might be interested in finding out a patient's probability of having liver disease if they are an alcoholic. "Being an alcoholic" is the **test** (kind of like a litmus test) for liver disease.

- **A** could mean the event "Patient has liver disease." Past data tells you that 10% of patients entering your clinic have liver disease. $P(A) = 0.10$.
- **B** could mean the litmus test that "Patient is an alcoholic." Five percent of the clinic's patients are alcoholics. $P(B) = 0.05$.

- You might also know that among those patients diagnosed with liver disease, 7% are alcoholics. This is your $B|A$: the probability that a patient is alcoholic, given that they have liver disease, is 7%.

Bayes' theorem tells you:
 $P(A|B) = (0.07 * 0.1) / 0.05 = 0.14$

In other words, if the patient is an alcoholic, their chances of having liver disease is 0.14 (14%). This is a large increase from the 10% suggested by past data. But it's still unlikely that any particular patient has liver disease.

More Bayes' Theorem Examples

Bayes' Theorem Problems Example #2

Another way to look at the theorem is to say that one event follows another. Above I said "tests" and "events", but it's also legitimate to think of it as the "first event" that leads to the "second event." There's no one right way to do this: use the terminology that makes most sense to you.

In a particular pain clinic, 10% of patients are prescribed narcotic pain killers. Overall, five percent of the clinic's patients are addicted to narcotics (including pain killers and illegal substances). Out of all the people prescribed pain pills, 8% are addicts. *If a patient is an addict, what is the probability that they will be prescribed pain pills?*

Step 1: Figure out what your event "A" is from the question. That information is in the italicized part of this particular question. The event that happens first (A) is being prescribed pain pills. That's given as 10%.

Step 2: Figure out what your event "B" is from the question. That information is also in the italicized part of this particular question. Event B is being an addict. That's given as 5%.

Step 3: Figure out what the probability of event B (Step 2) given event A (Step 1). In other words, find what $(B|A)$ is. We want to know "Given that people are prescribed pain pills, what's the probability they are an addict?" That is given in the question as 8%, or .8.

Step 4: Insert your answers from Steps 1, 2 and 3 into the formula and solve.

$$P(A|B) = P(B|A) * P(A) / P(B) = (0.08 * 0.1) / 0.05 = 0.16$$

The probability of an addict being prescribed pain pills is 0.16 (16%).

Example #3: the Medical Test

A slightly more complicated example involves a medical test (in this case, a genetic test):

There are **several forms of Bayes' Theorem** out there, and they are all equivalent (they are just written in slightly different ways). In this next equation, "X" is used in place of "B." In addition, you'll see some changes in the denominator. The proof of why we can rearrange the equation like this is beyond the scope of this article (otherwise it would be 5,000 words instead of 2,000!). However, if you come across a question involving medical tests, you'll likely be using this alternative formula to find the answer:

$$\Pr(A|X) = \frac{\Pr(X|A) \Pr(A)}{\Pr(X|A) \Pr(A) + \Pr(X|\sim A) \Pr(\sim A)}$$

UNIT V

Concepts of random variable

A **random variable** is a **variable** whose value is unknown or a function that assigns values to each of an experiment's outcomes. ... **Random variables** are often used in econometric or regression analysis to determine statistical relationships among one another.

Discrete and Continuous Random Variables:

A **variable** is a quantity whose value changes.

A **discrete variable** is a variable whose value is obtained by counting.

Examples: number of students present

number of red marbles in a jar

number of heads when flipping three coins

students' grade level

A **continuous variable** is a variable whose value is obtained by measuring.

Examples: height of students in class

weight of students in class

time it takes to get to school

distance traveled between classes

A **random variable** is a variable whose value is a numerical outcome of a random phenomenon.

- A random variable is denoted with a capital letter
- The probability distribution of a random variable X tells what the possible values of X are and how probabilities are assigned to those values
- A random variable can be discrete or continuous

A **discrete random variable** X has a countable number of possible values.

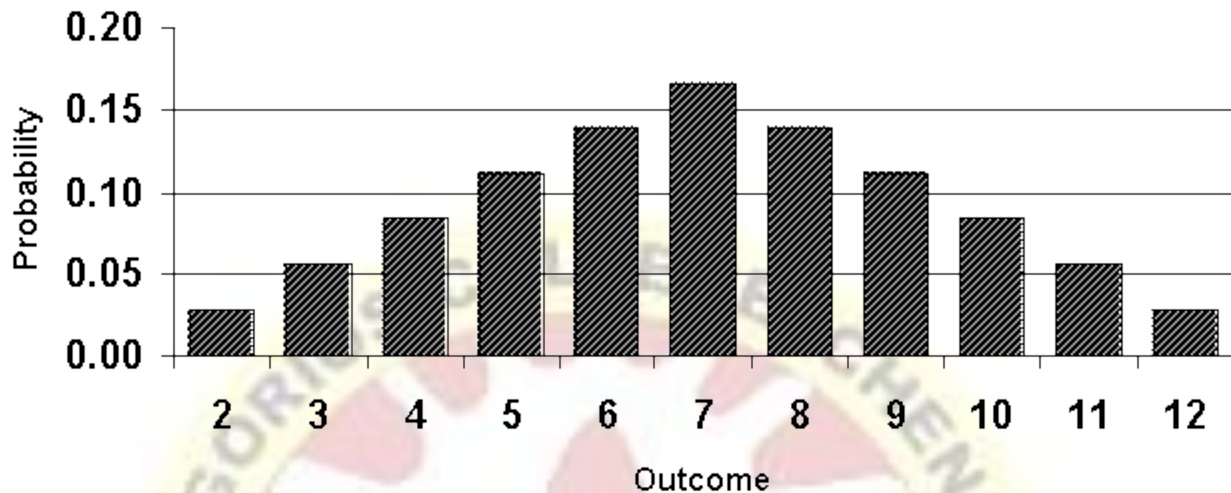
Example: Let X represent the sum of two dice.

Then the probability distribution of X is as follows:

X	2	3	4	5	6	7	8	9	10	11	12
$P(X)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

To graph the probability distribution of a discrete random variable, construct a **probability histogram**.

Probability Distribution of X



A **continuous random variable** X takes all values in a given interval of numbers.

T

- The probability distribution of a continuous random variable is shown by a **density curve**.
- The probability that X is between an interval of numbers is the area under the density curve between the interval endpoints
- The probability that a **continuous random variable** X is exactly equal to a number is zero

Means and Variances of Random Variables:

The mean of a discrete random variable, X , is its weighted average. Each value of X is weighted by its probability.

To find the mean of X , multiply each value of X by its probability, then add all the products.

$$\begin{aligned}\mu_X &= x_1 p_1 + x_2 p_2 + \cdots + x_k p_k \\ &= \sum x_i p_i\end{aligned}$$

The mean of a random variable X is called the **expected value** of X.

A **continuous random variable** is a **random variable** where the data can take infinitely many values. For example, a **random variable** measuring the time taken for something to be done is **continuous** since there are an infinite number of possible times that can be taken.

Moment Generating Function

The moment-generating function (mgf) of the (distribution of the) random variable Y is the function m_Y of a real parameter t defined by

$$m_Y(t) = E[et^Y],$$

for all $t \in \mathbb{R}$ for which the expectation $E[et^Y]$ is well defined.
