SMA 2272: Statistics

Course Purpose: This course introduces the students to the essentials required to understand issues related to measurement and how to generate descriptive information and statistical analysis from these measurements.

Course Objectives: The aim of this course is to provide students with the necessary statistical background for analyzing data and drawing inference from the analyses.

Expected Outcomes: On completion of the unit the student will be able to:

1. Summarise the main features of a data set (exploratory data analysis).
    (a) Summarise a set of data using a table or frequency distribution, and display it graphically using a line plot, a box plot, a bar chart, histogram, stem and leaf plot, or other appropriate elementary device.
    (b) Describe the level/location of a set of data using the mean, median, mode, as appropriate.
    (c) Describe the spread/variability of a set of data using the standard deviation, range, interquartile range, as appropriate.
    (d) Explain what is meant by symmetry and skewness for the distribution of a set of data.
2. Explain the concepts of probability.
    (a) Explain what is meant by a set function, a sample space for an experiment, and an event.
    (b) Define probability as a set function on a collection of events, stating basic axioms.
    (c) Derive basic properties satisfied by the probability of occurrence of an event, and calculate probabilities of events in simple situations.
    (d) Derive the addition rule for the probability of the union of two events, and use the rule to calculate probabilities.
    (e) Define the conditional probability of one event given the occurrence of another event, and calculate such probabilities.
    (f) Derive Bayes' Theorem for events, and use the result to calculate probabilities.
    (g) Define independence for two events, and calculate probabilities in situations involving independence.
3. Apply various probability distribution functions and sampling techniques in a real life situation.
4. Demonstrate an understanding of statistical data analysis.
5. Investigate linear relationships between variables using correlation analysis and regression analysis.
    (a) Draw scatterplots for bivariate data and comment on them.
    (b) Define and calculate the correlation coefficient for bivariate data, explain its interpretation and perform statistical inference as appropriate.

(c) Explain what is meant by response and explanatory variables.

(d) State the usual simple regression model (with a single explanatory variable).

(e) Derive and calculate the least squares estimates of the slope and intercept parameters in a simple linear regression model.

(f) Calculate R^2 (coefficient of determination) and describe its use to measure the goodness of fit of a linear regression model.

(g) Use a fitted linear relationship to predict a mean response or an individual response with confidence limits.

(h) State the usual multiple linear regression model (with several explanatory variables).

Course Description: Classical and axiomatic approaches to probability. Compound and conditional probability, including Bayes' theorem. Concept of discrete random variable: expectation and variance. Data: sources, collection, classification and processing. Frequency distributions and graphical representation of data, including bar diagrams, histograms and stem-and-leaf diagrams. Measures of central tendency and dispersion. Skewness and kurtosis. Correlation. Fitting data to a best straight line.

Pre-Requisites: Mathematics for Science.

Reference Books:

1. Mathematical statistics. Freund, John E - 7th ed. - Prentice Hall International, 2004. 614 pages. ISBN: 978-0131246461.

2. Introduction to Probability Models. Sheldon M. Ross, 5th edition.

3. Introduction to the Theory of Probability with Statistical Application. Hogg and Graig.

4. Introduction to the Theory of Statistics. Mood and Graybill

5. Modern probability Theory and its Applications. Parzen

Teaching Methodology: The method of instruction will be lectures, interactive tutorials and any other presentations/demonstrations the lecturer will deem fit towards enhancing understanding of the concepts taught in class.

Lectures: 2 Hours per week; Tutorials: 2 hours per week.

Course Assessment: During the period of study, assessment will be conducted by CATs (Continuous Assessment Tests), regular assignments and a final Examination at end of the unit. The composition for continuous assessment shall be as follows: 10 % Assignments, 20 % Tests, and regular examination at end of semester 70 %.

Contents

Lecture 1: An Introduction to Statistics

"A knowledge of statistics is like a knowledge of foreign language of algebra; it may prove of use at any time under any circumstance" Bowley.

***Objective:** The aim of the present lesson is to enable the students to understand the meaning, definition, nature, importance and limitations of statistics.*

1 Introduction

Statistics is a discipline, which scientifically deals with data, and is often described as the science of data. In dealing with statistics as data, statistics has developed appropriate methods of collecting, presenting, summarizing, and analysing data, and thus consists of a body of these methods. Statistics has become an integralpartofourdailylives. Everyday, weareconfrontedwithsomeformofstatisticalinformationthrough newspapers, magazines and other forms of communication. Such statistical information has become highly influential in our lives. Indeed, the famous science fiction writer H.G. Wells had predicted nearly a century ago that

> *"Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write".*

Thus, the subject of statistics in itself, has gained considerable importance in affecting the processes of our thinking and decision making.

The subject of statistics is primarily concerned with making decisions about various properties of some population of interest such as stock market trends, unemployment rates in various sectors of industries, demographic shifts, interest rates, inflation rates over the years and so on. By way of examples, consider the following statistical statements.

- The crime rate has gone up by 15% of what it was last year.
- The average salary of a Professor at DeKUT is $50,000 per year.
- The rate of inflation in Kenya is expected to remain above 15% per year for the next 5 years.
- Less than 20% of all high school graduates enter colleges for higher education, and less than 40% of those who do enter colleges, actually graduate.
- Majority of Kenyans consider Japanese cars superior in quality than Chinese cars.

All the above statements represent statistical conclusions in some form.

Statistics is a scientific discipline concerned with collection, description, analysis, and interpretation of data obtained from observation or experiment.

1. **Collection of data:** Once an investigator has collected data through a survey, it is necessary to edit these data in order to correct any apparent inconsistencies, ambiguities, recording errors or for that matter any mistake that can enter into the actual computations. But even before the data has been

collected and edited, it is assumes that these can be suitably classified according to some common characteristic of the population sampled.

2. **Description of data:** The organized data can now be presented in the form of tables or diagrams or graphs. This presentation in an orderly manner facilitates the understanding as well as analysis of data.
3. **Analysis of data:** The basic purpose of data analysis is to make it useful for certain conclusions. This analysis may simply be a critical observation of data to draw some meaningful conclusions about it or it may involve highly complex and sophisticated mathematical techniques. Some simple statistical tools such as calculations of averages, dispersion of data around averages and percentages are commonly used to analyze data.

4. **Interpretation of data:** Interpretation means drawing conclusions from the data which form the basis of decision making. Correct interpretation requires a high degree of skill and experience and is necessary in order to draw valid conclusions.

1.1 Categories of Statistics

There are two major divisions of statistics such as **descriptive statistics** and **inferential statistics**.

The term **descriptive statistics** deals with collecting, summarizing, and simplifying data, which are otherwise quite unwieldy and voluminous. It seeks to achieve this in a manner that meaningful conclusions can be readily drawn from the data. Descriptive statistics may thus be seen as comprising methods of bringing out and highlighting the latent characteristics present in a set of numerical data. It not only facilitates an understanding of the data and systematic reporting thereof in a manner; and also makes them amenable to further discussion, analysis, and interpretations.

The first step in any scientific inquiry is to collect data relevant to the problem in hand. When the inquiry relates to physical and/or biological sciences, data collection is normally an integral part of the experiment itself. In fact, the very manner in which an experiment is designed, determines the kind of data it would require and/or generate. The problem of identifying the nature and the kind of the relevant data is thus automatically resolved as soon as the design of experiment is finalized. It is possible in the case of physical sciences. In the case of social sciences, where the required data are often collected through a questionnaire from a number of carefully selected respondents, the problem is not that simply resolved. For one thing, designing the questionnaire itself is a critical initial problem. For another, the number of respondents to be accessed for data collection and the criteria for selecting them has their own implications and importance for the quality of results obtained. Further, the data have been collected, these are assembled, organized, and presented in the form of appropriate tables to make them readable. Wherever needed, figures, diagrams, charts, and graphs are also used for better presentation of the data. A useful tabular and graphic presentation of data will require that the raw data be properly classified in accordance with the objectives of investigation and the relational analysis to be carried out.

A well thought-out and sharp data classification facilitates easy description of the hidden data characteristics by means of a variety of summary measures. These include measures of central tendency, dispersion, skewness, and kurtosis, which constitute the essential scope of descriptive statistics.

Descriptive statistics summarize population data numerically or graphically by deriving

- statistics pertaining to central tendency such as the mean, median, or mode
- statistics pertaining to dispersion around the central tendency such as the range or standard deviation
- statistics or graphs depicting the shape of a distribution

Inferential statistics, also known as inductive statistics, goes beyond describing a given problem situation by means of collecting, summarizing, and meaningfully presenting the related data. Instead, it consists of methods that are used for drawing inferences, or making broad generalizations, about a totality of observations on the basis of knowledge about a part of that totality. The totality of observations about which an inference may be drawn, or a generalization made, is called a population or a universe. The part of totality, which is observed for data collection and analysis to gain knowledge about the population, is called a sample. The desired information about a given population of our interest; may also be collected even by observing all the units comprising the population. This total coverage is called *census*. Getting the desired value for the population through census is not always feasible and practical for various reasons. A part from time and money considerations making the census operations prohibitive, observing each individual unit of the population with reference to any data characteristic may at times involve even destructive testing. In such cases, obviously, the only recourse available is to employ the partial or incomplete information gathered through a sample for the purpose. This is precisely what inferential statistics does. Thus, obtaining a particular value from the sample information and using it for drawing an inference about the entire population underlies the subject matter of inferential statistics. Consider a situation in which one is required to know the average body weight of all the college students in a given cosmopolitan city during a certain year. A quick and easy way to do this is to record the weight of only 500 students, from out of a total strength of, say, 10000, or an unknown total strength, take the average, and use this average based on incomplete weight data to represent the average body weight of all the college students. In a different situation, one may have to repeat this exercise for some future year and use the quick estimate of average body weight for a comparison. This may be needed, for example, to decide whether the weight of the college students has undergone a significant change over the years compared.

Inferential statistics helps to evaluate the risks involved in reaching inferences or generalizations about an unknown population on the basis of sample information. for example, an inspection of a sample of five battery cells drawn from a given lot may reveal that all the five cells are in perfectly good condition. This information may be used to conclude that the entire lot is good enough to buy or not.

Since this inference is based on the examination of a sample of limited number of cells, it is equally likely that all the cells in the lot are not in order. It is also possible that all the items that may be included in the sample are unsatisfactory. This may be used to conclude that the entire lot is of unsatisfactory quality, whereas the fact may indeed be otherwise. It may, thus, be noticed that there is always a risk of an inference about a population being incorrect when based on the knowledge of a limited sample. The rescue in such situations lies in evaluating such risks. For this, statistics provides the necessary methods. These centres on quantifying in probabilistic term the chances of decisions taken on the basis of sample information being incorrect. This requires an understanding of the what, why, and how of probability and probability distributions to equip ourselves with methods of drawing statistical inferences and estimating the degree of reliability of these inferences.

Inferential statistics allow one to infer population parameters based upon sample statistics and to model relationships within the data. The categories of inferential statistics are

- Estimation is the group of statistics which allow for the estimation about population values based upon sample data. The two types of statistics in this category are population parameter estimates and confidence intervals.

- Modeling allows us to develop mathematical equations which describe the interrelationships between two or more variables.

- Hypothesis testing allows us to test for whether a particular hypothesis we've developed is supported by a systematic analysis of the data.

1.2 Commonly used Terms in Statistics

1.2.1 Populations and Samples

In statistics, we are interested in obtaining information about a total collection of elements, which we will refer to as the **population**. A population is defined as the set of all individuals, items, or data of interest. This is the group about which scientists will generalize. A characteristic (usually numeric) that describes a population is referred to as a **population parameter**.

The population is often too large for us to examine each of its members. In such case, we try to learn about the population by choosing and then examining a sub group of its elements. The subgroup of a population is called a **sample**. A sample is defined as a set of selected individuals, items, or data taken from a population of interest. A characteristic (usually numeric) that describes a sample is referred to as a **sample statistic**. The total collection of all elements that we are interested in is called a **target population**.

A sample of k members of a population is said to be a **random sample**, sometimes called a **simple random sample**, if the members are chosen in such a way that all possible choices of the k members are equally likely.

Example 1.1. On the basis of the following example, we will identify the population, sample, population parameter, and sample statistic: Suppose you read an article in the local college newspaper citing that the average college student plays 2 hours of video games per week. To test whether this is true for your school, you randomly approach 20 fellow students and ask them how long (in hours) they play video games per week. You find that the average student, among those you asked, plays video games for 1 hour per week. Distinguish the population from the sample.

Answer. In this example, all college students at your school constitute the population of interest, and the 20 students you approached is the sample that was selected from this population of interest. Since it is purported that the average college student plays 2 hours of video games per week, this is the population parameter (2 hours). The average number of hours playing video games in the sample is the sample statistic (1 hour).

1.3 Characteristics of Statistics

In order for the quantitative and numerical data to be identified as statistics, it must possess certain identifiable characteristics. Some of these characteristics are described below:

1. **Statistics are aggregate of facts.** Single or isolated facts or figures cannot be called statistics as these cannot be compared or related to other figures within the same framework. For example, a single birth in a hospital is not statistics, as it has no significance for analysis purposes. However, when such information about many births in the same hospital or birth information for different hospitals is collected, then this information can be compared and analyzed and thus this data would constitute statistics.

2. **Statistics, generally are not the outcomes of a single cause, but are affected by multiple causes.** There are a number of forces working together that affect the facts and figures. For example, when we say that the crime rate in Kenya has increased by 15% over the last year, a number of factors might have affected this change. These factors may be: general level of economy such as state of economic recession, unemployment rate, extent of use of drugs, areas affected by crime, extent of legal effectiveness, social structure of the family in the area and so on. While these factors can be isolated by themselves, the effects of these factors cannot be isolated and measured individually. It is generally not possible to segregate and study the effect of each of these forces individually.
3. **Statistics are numerically expressed.** All statistics are stated in numerical figures which means that these are quantitative information only.

4. **Statistical data is collected in a systematic manner.** The procedure for collecting data should be predeterminedandwellplannedandsuchdatacollectionshouldbeundertakenbytrainedinvestigators. Haphazard collection of data can lead to erroneous conclusions.

5. **Statistics are collected for a predetermined purpose.** The purpose and objective of collecting pertinent data must be clearly defined, decided upon and determined prior to data collection. This would facilitate the collection of proper and relevant data. For example, data on the heights of students would be irrelevant if considered in connection with the ability to get admission in a college, but may be relevant when considering recruits to join the army.

6. **Statistics are enumerated or estimated according to reasonable standard of accuracy.** There are basically two ways of collecting data. one is the actual counting or measuring, which is the most accurate way. For example, the number of people attending a football game can be accurately determined by counting the number of tickets sold and redeemed at the gate.

   The second way of collecting data is by estimation and is used in situations where actual counting or measuring is not feasible or where it involves prohibitive cost. For example, the crowd at a campaign rally can be estimated by using visual observation or by taking samples of some segments of the crowd and then estimating the total number of people on that basis of these samples. Estimates, based on the samples cannot be as precise and accurate as actual counts or measurements, but these should be consistent with the degree of accuracy desired.

7. **Statistics must be placed in relation to each other.** The main objective of data collection is to facilitate a comparative or relative study of the desired characteristics of the data. The comparisons of facts and figures may be conducted regarding the same characteristics over a period of time from

a single source or it may be from various sources at any one given time. For example, prices of different items in a store as such would not be considered statistics. However, prices of one product in different stores constitute statistical data since these prices are comparable. Also, the changes in the price of a product in one store over a period of time would also be considered statistical data since these changes provide for comparison over a period of time.

1.4 Functions of Statistics

1. **It condenses and summarizes voluminous data into a few presentable, understandable and precise figures.** The raw data, as is usually available, is voluminous and haphazard. It is generally not possible to draw any conclusions from the raw data as collected. Hence it is necessary and desirable to express this data in few numerical values. For example, stock market prices of individual stocks and their trends are highly complex to comprehend, but a graph of prices trends gives us the overall picture at a glance.

2. **It facilitates classification and comparison of data.** Arrangement of data with respect to different characteristics facilitates comparison and interpretation. For example, data on age, height, gender, and family income of college students gives us a much better picture of students when the data is categorized relative to these characteristics.

3. **It helps in determining functional relationship between two or more phenomenon.** Statistical techniques such as correlational analysis assist in establishing the degree of association between two or more independent variables. For example, the *coefficient of correlation* between literacy and employment gives us the degree of association between extent of training and industrial productivity.

4. **It helps in predicting future trends (Forecasting).** Statistical methods are highly useful tools in analyzing the past data and predicting some future trends. For example, the sales for a particular product for the next year can be computed by knowing the sales for the same product over the previous years, the current market trends and the possible changes in the variable that affect the demand of the product.

5. **It helps the central management and the government in formulating policies.** Example, the recently conducted census, will be used as a source of information for planning by the government for the next 10 years until another census is conducted in 2019.

1.5 Scope of Statistics

Some of the important areas where the knowledge of statistics is usefully applied are as follows:

1. **Government.** Various departments of the government collect and interpret vast amount of data and information for efficient functioning and decision making.

2. **Economics.** Statistics are widely used in economics study and research. The subject of economics is mainly concerned with production and distribution of wealth as well as savings and investments. Some of the areas of economic interest in which statistical tools are used are as follows:

   (a) Statistical methods are extensively used in measuring and forecasting Gross National Product (GNP).

   (b) Economic stability is primarily judged by statistical studies of business cycles.

   (c) Statistical analyzes of population growth, unemployment figures, rural or urban population shifts and so on influence much of the economic policy making.

   (d) Econometric models which involve application of statistical methods and used for optimum utilization of resources available.

   (e) Financial statistics are necessary in the fields of money and banking including consumer savings and credit availability.

3. **Physical, Natural and Social Sciences.** In physical sciences, as an example, the science of meteorology uses statistics in analyzing the data gathered by satellites in predicting weather conditions.

4. **Statistics and Research.** There is hardly any advanced research going on without the use of statistics in one form or another. Statistics are used extensively in medical, pharmaceutical and agricultural research. The effectiveness of a new drug is determined by statistical experimentation and evaluation.

5. **Other Areas.** Statistics are commonly used by insurance companies, stock brokerage firms, banks, public utility companies and so on. Statistics are also immensely useful to politicians since they can predict their chance of winning through the use of sampling techniques in random selection of voters sampled and studying their attitude on issues and policies.

1.6 Limitations of Statistics

Statistics has a number of limitations, pertinent among them are as follows:

1. **It does not deal with individual values.** Statistics only deals with aggregate values. For example, the marks obtained by one student in a class does not carry any meaning in itself, unless it can be compared with a set standard or with other students in the same class or with his own marks obtained earlier.
2. **It cannot deal with qualitative characteristics.** Statistics is not applicable to qualitative characteristics such as honesty, kindness, goodness, colour, poverty, beauty, and so on, since these cannot be expressed in quantitative terms. The characteristics, however, can be statistically dealt with if some quantitative values can be assigned to these with logical criterion.

3. **Statistical conclusions are not universally true.** Since statistics is not an exact science, as is the case with natural sciences, the statistical conclusions are true only under certain assumptions.

4. **Statistical interpretation requires a high degree of skill and under standing of the subject.** In order to get meaningful results, it is necessary that the data be properly and professionally collected and critically interpreted. it requires extensive training to read and analyze statistics in its proper context.

5. **Statistics can be misused.** The famous statement that *'figures don't lie but the liars can figure'*, is a testimony to the misuse of statistics. Thus, inaccurate or incomplete figures, can be manipulated to get desirable references. Example, advertising slogans such as 4 out of 5 dentists recommend brand X tooth paste gives us the impression that 80% of all dentists recommended this brand. This may not be true since we don't know how big the sample is or whether the sample represents the entire population or not.

   Another example is the opinion polls after the news. We are normally given a percentage but not told the sample size of the total number of people who called to respond to the questions.

6. There are certain phenomena or concepts where statistics cannot be used. This is because these phenomena or concepts are not amenable to measurement. For example, beauty, intelligence, courage cannot be quantified. Statistics has no place in all such cases where quantification is not possible.

7. Statistics reveal the average behaviour, the normal or the general trend. An application of the 'average' concept if applied to an individual or a particular situation may lead to a wrong conclusion and sometimes may be disastrous. For example, one may be misguided when told that the average depth of a river from one bank to the other is four feet, when there may be some points in between where its depth is far more than four feet. On this understanding, one may enter those points having greater depth, which may be hazardous.

8. Since statistics are collected for a particular purpose, such data may not be relevant or useful in other situations or cases. For example, secondary data (i.e., data originally collected by someone else) may not be useful for the other person.

9. Statistics are not 100 per cent precise as is Mathematics or Accountancy. Those who use statistics should be aware of this limitation.

10. In statistical surveys, sampling is generally used as it is not physically possible to cover all the units or elements comprising the universe. The results may not be appropriate as far as the universe is concerned. Moreover, different surveys based on the same size of sample but different sample units may yield different results.

11. At times, association or relationship between two or more variables is studied in statistics, but such a relationship does not indicate cause and effect' relationship. It simply shows the similarity or dissimilarity in the movement of the two variables. In such cases, it is the user who has to interpret the results carefully, pointing out the type of relationship obtained.

12. A major limitation of statistics is that it does not reveal all pertaining to a certain phenomenon. There is some background information that statistics does not cover. Similarly, there are some other aspects related to the problem on hand, which are also not covered. The user of Statistics has to be well informed and should interpret Statistics keeping in mind all other aspects having relevance on the given problem.

Apart from the limitations of statistics mentioned above, there are misuses of it. Many people, knowingly or unknowingly, use statistical data in wrong manner. Let us see what the main misuses of statistics are so that the same could be avoided when one has to use statistical data. The misuse of Statistics may take several forms some of which are explained below.

(i) **Sources of data not given:** At times, the source of data is not given. In the absence of the source, the reader does not know how far the data are reliable. Further, if he wants to refer to the original source, he is unable to do so.

(ii) **Defective data:** Another misuse is that sometimes one gives defective data. This may be done knowingly in order to defend one's position or to prove a particular point. This apart, the definition used to denote a certain phenomenon may be defective. For example, in case of data relating to unemployed persons, the definition may include even those who are employed, though partially. The question here is how far it is justified to include partially employed persons amongst unemployed ones.

(iii) **Unrepresentative sample:** In statistics, several times one has to conduct a survey, which necessitates to choose a sample from the given population or universe. The sample may turn out to be unrepresentative of the universe. One may choose a sample just on the basis of convenience. He may collect the desired information from either his friends or nearby respondents in his neighbourhood even though such respondents do not constitute a representative sample.

(iv) **Inadequate sample:** Earlier, we have seen that a sample that is unrepresentative of the universe is a major misuse of statistics. This apart, at times one may conduct a survey based on an extremely inadequate sample. For example, in a city we may find that there are 100,000 households. When we have to conduct a household survey, we may take a sample of merely 100 households comprising only 0.1 per cent of the universe. A survey based on such a small sample may not yield right information.

(v) **Unfair Comparisons:** An important misuse of statistics is making unfair comparisons from the data collected. For instance, one may construct an index of production choosing the base year where the production was much less. Then he may compare the subsequent year's production from this low base. Such a comparison will undoubtedly give a rosy picture of the production though in reality

it is not so. Another source of unfair comparisons could be when one makes absolute comparisons instead of relative ones. An absolute comparison of two figures, say, of production or export, may show a good increase, but in relative terms it may turnout to be very negligible. Another example of unfair comparison is when the population in two cities is different, but a comparison of overall death rates and deaths by a particular disease is attempted. Such a comparison is wrong. Likewise, when data are not properly classified or when changes in the composition of population in the two years are not taken into consideration, comparisons of such data would be unfair as they would lead to misleading conclusions.

(vi) **Unwanted conclusions:** Another misuse of statistics may be on account of unwarranted conclusions. This may be as a result of making false assumptions. For example, while making projections of population in the next five years, one may assume a lower rate of growth though the past two years indicate otherwise. Sometimes one may not be sure about the changes in business environment in the near future. In such a case, one may use an assumption that may turn out to be wrong. Another source of unwarranted conclusion may be the use of wrong average. Suppose in a series there are extreme values, one is too high while the other is too low, such as 800 and 50. The use of an arithmetic average in such a case may give a wrong idea. Instead, harmonic mean would be proper in such a case.

(vii) **Confusion of correlation and causation:** In statistics, several times one has to examine the relationship between two variables. A close relationship between the two variables may not establish a cause-and-effect-relationship in the sense that one variable is the cause and the other is the effect. It should be taken as something that measures degree of association rather than try to find out causal relationship.

1.7 Self-Test Questions

1. What are the major limitations of Statistics? Explain with suitable examples.

2. Distinguish between descriptive Statistics and inferential Statistics.

Lecture 2: Data Types and Measures

2 Data Types and Measures

2.1 Introduction

Statistical data are the basic raw material of statistics. Data may relate to an activity of our interest, a phenomenon, or a problem situation under study. They derive as a result of the process of measuring, counting and/or observing. Statistical data, therefore, refer to those aspects of a problem situation that can be measured, quantified, counted, or classified. Any object subject phenomenon, or activity that generates data through this process is termed as a variable. In other words, a variable is one that shows a degree of variability when successive measurements are recorded. In statistics, data are classified into two broad categories: quantitative data and qualitative data. This classification is based on the kind of characteristics that are measured.

Quantitative data are those that can be quantified in definite units of measurement. These refer to characteristics whose successive measurements yield quantifiable observations. Depending on the nature of the variable observed for measurement, quantitative data can be further categorized as continuous and discrete data.

Obviously, a variable may be a continuous variable or a discrete variable.

(i) **Continuous data** represent the numerical values of a continuous variable. A continuous variable is the one that can assume any value between any two points on a line segment, thus representing an interval of values. The values are quite precise and close to each other, yet distinguishably different. All characteristics such as weight, length, height, thickness, velocity, temperature, tensile strength, etc., represent continuous variables. Thus, the data recorded on these and similar other characteristics are called continuous data. It may be noted that a continuous variable assumes the finest unit of measurement. Finest in the sense that it enables measurements to the maximum degree of precision.

(ii) **Discrete data** are the values assumed by a discrete variable. A discrete variable is the one whose outcomes are measured in fixed numbers. Such data are essentially count data. These are derived from a process of counting, such as the number of items possessing or not possessing a certain characteristic. The number of customers visiting a departmental store everyday, the incoming flights at an airport, and the defective items in a consignment received for sale, are all examples of discrete data.

Qualitative data refer to qualitative characteristics of a subject or an object. A characteristic is qualitative in nature when its observations are defined and noted in terms of the presence or absence of a certain attribute in discrete numbers. These data are further classified as nominal and rank data.

(i) **Nominal data** are the outcome of classification into two or more categories of items or units comprising a sample or a population according to some quality characteristic. Classification of students according to sex (as males and females), of workers according to skill (as skilled, semi-skilled, and unskilled), and of employees according to the level of education (as matriculates, undergraduates, and post-graduates), all result into nominal data. Given any such basis of

classification, it is always possible to assign each item to a particular class and make a summation of items belonging to each class. The count data so obtained are called nominal data.

(ii) **Ordinal/Rank data**, on the other hand, are the result of assigning ranks to specify order in terms of the integers $1, 2, 3, \cdots, n$. Ranks may be assigned according to the level of performance in a test. a contest, a competition, an interview, or a show. The candidates appearing in an interview, for example, may be assigned ranks in integers ranging from I to n, depending on their performance in the interview. Ranks so assigned can be viewed as the continuous values of a variable involving performance as the quality characteristic.

You cannot use statistics without data. Different statistical methods are appropriate for different types of data. Moreover, different statistical analyses require different representations of the same data. This means that we have to know something about how data are categorized, represented, manipulated and managed. Statistical analysis requires a significant amount of time preparing data for analysis.

Developing a good understanding of the kinds of data and data measurement is necessary because the kind of data you are analyzing essentially dictates the type of statistical analysis you perform. Data can be classified as either numerical (quantitative) or categorical (qualitative).

Data is either provided to you or you collect it yourself. In the latter case, it will be worth your while to think about how you enter (key in) the data. For example, counts are represented as nonnegative integers while measurements are real numbers. Furthermore, when it comes to analyzing and presenting data, the same method will display data differently based on their type.

Categorical data is data that can be sorted according to a category and each value is from a set of nonoverlapping values. Examples of categorical data would include eye color (green, brown, blue, etc.) and managerial level (supervisor, mid-level, executive).

- Categorical variables are typically measured on a **nominal scale**. Nominal level variables are those that can simply be grouped; there's no underlying numeric order to them and any ordering is arbitrary or artificial. Our examples above of eye color and managerial level are both measured at a nominal level. Other examples of categorical data that's measured on a nominal scale include type of industry, state of residence, marital status, and favorite food. Please note that you might have responses "dummy-coded" with numbers to represent a response such as 0s representing male and 1s representing female, but even though there are numbers it's still a nominal scale because the ordering is completely arbitrary; we could have had 0s representing female and 1s representing male.

Data involve the *values* of a variable and there are several types of variable:

Numerical data can be classified into two types: **discrete** and **continuous**. The distinction between discrete and continuous data is that discrete data can only take one of a set of particular values, whereas continuous data can take any value within a specified range (or the possible values are so close together that they can be considered to occupy a continuous range).

Discrete data arise from counting, *eg* numbers of actuaries, chemists, number of atoms, numbers of claims.

Continuous data arise from measuring, eg height, amount, age.

From a strictly mathematical point of view, the distinction we make here between continuous and discrete is not correct. For our purpose, the distinction is useful.

CATEGORICAL Factors are also called categories or enumerated types. Think of a factor as a set of category names. Factors are qualitative classification of objects. Categories do not imply order. A black cat is different from a brown cat. It is neither larger or smaller.

Attribute (or dichotomous) data have only two categories, eg yes/no, male/female, claim/no claim.
***Nominal data* have several unordered categories, eg type of policy, nature of claim.**

***Ordinal data* have several ordered categories, eg questionnaire responses such as "strongly in favour / ⋯ / strongly against".**

Sometimes we use the levels to indicate order, but not necessarily magnitude. For example, we can define the label of presidential candidates as implying order from the most popular (having the most number of voters) to the least popular. Ordinal data do not reveal this kind of information. For example, we generally agree that rabbis are faster than turtles. We rarely know by how much.

Examples: Here are some examples of categorical data: a division of a population into males and females, thenumberofdotsthatappearonthefaceofadie, headortailinflippingacoin, speciesandcolourofflowers.

Categorical data may be presented in graphs. However, the location of categories along the x or the y axes does not imply order.

2.2 Levels of Measurement

Many statistical tests will require that variables in a study be measured on a certain scale of measurement. In the early 1940s, Harvard psychologist S. S. Stevens coined the terms nominal, ordinal, interval, and ratio to classify the scales of measurement (Stevens, 1946). Scales of measurement refer to how the properties of numbers can change with different uses.

In all, scales of measurement are characterized by three properties: order, differences, and ratios. Each property can be described by answering the following questions:

1. *Order:* Does a larger number indicate a greater value than a smaller number?

2. *Differences:* Does subtracting two numbers represent some meaningful value?

3. *Ratio:* Does dividing (or taking the ratio of) two numbers represent some meaningful value?

1. **Nominal Level (in name only):** Nominal scales are measurements where a number is assigned to represent something or someone. Qualities with no ranking/ordering; no numerical or quantitative value. Data consists of names, labels and categories. a. Taos, Acoma, Zuni and Cochiti are names of four native American pueblos. b. Car colors for a certain model are: red, silver, blue and black. Examples of nominal variables include a person's race, gender, nationality, sexual orientation, hair and eye color, season of

birth, marital status, or other demographic or personal information. A researcher may code men as 1 and women as 2. These numbers are used to identify gender and nothing more.

Coding refers to the procedure of converting a nominal value to a numeric value.

2. **Ordinal Level:** An ordinal scale of measurement is one that conveys order alone. This scale indicates that some value is greater or less than another value. Examples of ordinal scales include finishing order in a competition, education level, and rankings. These scales only indicate that one value is greater or less than another, so differences between ranks do not have meaning.

Can be arranged in some order, but the differences between the data values are meaningless. a. Of 17 fishing reels rated: 6 were rated good quality, 4 were rated better quality, and 7 were rated best quality. b. Out of a high school class of 319, Walter ranked 4th, June ranked 12th, and Jim ranked 20 th.

3. **Interval Level:** Interval scales are measurements where the values have no true zero and the distance between each value is equidistant.

Equidistant scales are those values whose intervals are distributed in equal units. A true zero describes values where the value 0 truly indicates nothing. Values on an interval scale do not have a true zero.

Data values can be ranked and the differences between data values are meaningful. However, there is no intrinsic zero, or starting point, and the ratio of data values are meaningless. Note: Calendar dates and Celsius & Fahrenheit temperature readings have no meaningful zero and ratios are meaningless. a. The years in which democrats won presidential elections. b. Body temperature in degrees Celsius (or Fahrenheit) of trout swimming in the North River. c. Building A was built in 1284, Building B in 1492 and Building C in 5 bce.

4. **Ratio Level:** Ratio scales are similar to interval scales in that scores are distributed in equal units. Yet, unlike interval scales, a distribution of scores on a ratio scale has a true zero. This is an ideal scale in behavioral research because any mathematical operation can be performed on the values that are measured. Common examples of ratio scales include counts and measures of length, height, weight, and time. For scores on a ratio scale, order is informative. For example, a person who is 30 years old is older than another who is 20. Differences are also informative. For example, the difference between 70 and 60 seconds is the same as the difference between 30 and 20 seconds (the difference is 10 seconds). Ratios are also informative on this scale because a true zero is defined tO truly means nothing. Hence, it is meaningful to state that 60 pounds is twice as heavy as 30 pounds.

Similar to interval, except there is a true zero, or starting point, and the ratios of data values have meaning. a. Core temperature of stars measured in degrees Kelvin. b. Time elapsed between the deposit of a check and the clearance of that check. c. Length of trout in the North

Question 2.1. *Answer the following dating agency questionnaire and state what type of data is required in each question:*

(a) How old are you? (Give your age last birthday.) (b)

How tall are you? (State as accurately as you can.)

(c) What gender are you?

(d) What colour are your eyes?

(d) Do you smoke? 6. How would you rate your looks? (10 =Drop-dead gorgeous, 1= Seen better days)

(e) height of trees (continuous)

(f) concentration of a pollutant in the air in unit of parts per million (discrete)

(g) weight of an animal (continuous)

(h) number of dogs in a park (discrete)

(i) average number of birds per flock (continuous)

(j) density of animal population (continuous)

Question 2.2. *With the help of the tutor or otherwise collect the following data.*

1. DATA SET 1: Age in years of the current students in your class.
2. DATA SET 2: Home province of the current students in your class.
3. DATA SET 3: Number of siblings (including self) of the current students in your class.
4. DATA SET 4: Shoe size of the current students in your class.

Question 2.3. *Indicate whether the following variables are (a) Qualitative (Nominal or Ordinal) or (b) Quantitative (Discrete or Continuous)*

1. The classification of BCM certificate after the course
2. The number of Kenyan couples who were married last year
3. The speed of rallying cars
4. Marital status of University staff
5. The age of your siblings
6. The taste of Oranges
7. The number of subjects offered in school last year
8. The length of your last call
9. The number of airtime denominations in Kenya
10. The district from which BCM students come from
11. The hair styles of the BCM students

12. The number of course in your University

13. The number of employees at the Kenya National Bureau of Statistics

2.3 Data Collection

Statistical investigation is a comprehensive process and requires systematic collection of data about some group of people or objects, describing and organizing the data, analyzing this data with the help of different statistical methods, 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 conclusions and hence, utmost importance must be given to this process and every possible precaution should be taken to ensure accuracy while gathering and collecting data.

Statistical data, may be classified under two categories depending upon the sources utilized. These categories are:

Data sources could be seen as of two types, viz., secondary and primary. The two can be defined as under: (i) Secondary data: They already exist in some form: published or unpublished - in an identifiable secondary source. They are, generally, available from published source(s), though not necessarily in the form actually required. (ii) Primary data: Those data which do not already exist in any form, and thus have to be collected for the first time from the primary source(s). By their very nature, these data require fresh and first-time collection covering the whole population or a sample drawn from it.

2.3.1 Primary data

Primary data is data which is collected by the investigator him/herself for the purpose of a specific inquiry or study. Such data is original in character and is generated by surveys conducted by individuals or research institutions. e.g. if we wish to study the attitude of Nairobi residents about the introduction of a new transport system a survey would be conducted directly on the commuters through personal interviews or mailed questionnaires etc. Such data collected would be considered as primary data.

2.3.2 Secondary data

When an investigator uses the data which has already been collected by others, such data is called *secondary data.* This data is primary data for the agency that collected it and becomes secondary data for someone else who uses this data for his own purposes.

The secondary data can be obtained from journals, reports, government publications, publications of professional and research organizations etc. For example if a researcher desires to analyze the weather conditions of different regions, he can get the required information or data from the records of the metrology department, economic surveys data from the KNBS.

The following steps may be considered in the primary data collection process:

1. **Planning the study**

Before any procedures for data collection are undertaken, the purpose and scope of the study must be clearly specified. If any similar studies have been conducted prior to the current one, then the investigator may want to use some secondary data in his own study and may redefine his objectives on the basis of the previous studies conducted.

The scope of the study must take into consideration the field to be covered and the time period in which to conduct the study. The time span is very important because in certain areas, the conditions change very quickly and hence by the time the study is completed, it may become irrelevant.

2. **Methods of Collecting data**

   There are many methods used to collect or obtain data for statistical analysis. Three of the most popular methods are:

   (a). Direct Observation

   (b). Experiments

   (c). Surveys

   Surveys: A *survey* solicits information from people; e.g. Gallup polls; pre-election polls; marketing surveys. The *Response Rate* (i.e. the proportion of all people selected who complete the survey) is a key survey parameter. Surveys may be administered in a variety of ways, e.g.

   - Personal Interview,
   - Telephone Interview, and
   - Self-Administered Questionnaire.

3. **Sampling**

   Recall that statistical inference permits us to draw conclusions about a population based on a sample.

   Sampling (i.e. selecting a sub-set of a whole population) is often done for reasons of *cost* (it's less expensive to sample 1,000 television viewers than 100 million TV viewers) and *practicality* (e.g. performing a crash test on every automobile produced is impractical).

   In any case, the *sampled population* and the *target population* should be **similar** to one another.

   A *sampling plan* is just a method or procedure for specifying how a sample will be taken from a population. We will focus our attention on one of these methods; **Simple Random Sampling.**

   A *simple random sample* is a sample selected in such a way that every possible sample of the same size is equally likely to be chosen. Drawing three names from a hat containing all the names of the students in the class is an example of a simple random sample: any group of three names is as equally likely as picking any other group of three names.

   Two major types of error can arise when a sample of observations is taken from a population: *sampling error* and *non-sampling error*.

Sampling error refers to differences between the sample and the population that exist only because of the observations that happened to be selected for the sample. Non-sampling errors are more serious and are due to mistakes made in the acquisition of data or due to the sample observations being selected improperly.

4. **Data Organization**

   Data that describes or measure a single attribute, say height of a tree, are called **univariate**. They are composed of a set of observations of objects about which a single value is obtained. **Bivariate data** are represented in pairs. **Multivariate data** are composed of a set of observations or objects. Each observation contain a number of values that represent this object.

   Statistical analysis usually involves more than one data file. Often we use several files to store different data that relate to a single analysis. We then need to somehow relate data from different files. This requires careful consideration of how the data are to be organized. Once you commit the data to a particular organization it is difficult to change. The way the data are organized will then dictate how easy they are to prepare for different types of statistical analyses.

   Data are organized into tables and tables are related to each other. The tables, their relationship and other auxiliary information form a database.

5. **Data manipulation**

   The core of working with data is the ability to subset, merge, split, and perform other such data operations. Applying various operations to subsets of the data wholesale is as important.

2.4 Exercise Questions

1. What is the difference between primary data and secondary data?

2. Differentiate between the terms parameter and statistic. Which one will be the result of taking a sample?

3. What is the difference between a sample and a census? Why can it be advantageous to select a sample instead of carrying out a census?

4. Statistics is a means of collecting numerical facts or data. Differentiate between a sample survey and a census; hence explain the major rewards of the sample survey method over census method.

Lecture 3: Data Presentation

3 The Presentation of Data

This lecture deals with descriptive statistics, that is, the methodology for describing or summarising a set of data using tables, diagrams and numerical measures. Presenting the data in a descriptive form is usually the first stage in any statistical analysis, as it allows us to spot any patterns in the data.

In this lecture we will learn methods for presenting and describing sets of data in tables and graphics. Often, tables requires a good deal of data manipulations. Graphics is an important tool not only in presenting data but also in cleaning them and latter analyzing them.

3.1 The Tabulation of Data

A frequency distribution table is a list of values for a variable and the number of times they appear in the data.

Ungrouped Data

Suppose we have a collection of measurements given by numbers. Some may occur only once, while others may be repeated several times. If we write down the numbers as they appear, the processing of them is likely to be cumbersome. This is known as "**ungrouped (or raw) data**". When the data set contains only a relative small number of distinct or different values, it is convenient to represent it in an **ungrouped frequency distribution table** which present each distinct value along with its frequency of occurrence.

The data from a discrete distribution can be summarised using a frequency distribution, that is, by counting the number of 0's, 1's, 2's, etc. For example, the number of children in a sample of 80 families might be summarised as follows:

Number of children under 16, x	Number of families in sample, f
0	8
1	12
2	28
3	19
4	7
5	4
6	1
7	1
8 or more	0

Question 3.1. *How would you calculate how many children there were altogether?*

Steps of Construction of Ungrouped frequency distribution table:

1. Identify the smallest and the largest value in the data set.

2. Tally the number of times each value is appearing in the data.

3. Count the number of tallies of each quantity and record them as the frequency for the value.

Example 3.1. ***Construction of ungrouped frequency distribution table:*** The following data represents the number of days of sick leave taken by each of 50 worker of a given company over the last 6 weeks:

17	13	8	9	16	12	8	11	9	13
11	11	11	16	19	12	10	13	10	15
16	12	9	11	11	13	8	11	15	10
15	16	10	16	18	12	14	12	11	8
12	12	12	12	9	8	8	10	15	13

Solution. Since the data set contains only a relative small number of distinct, or different values, it is convenient to represent it in a frequency table below which presents each distinct value along with its frequency of occurrence.

Number of Leave Days	Frequency
8	6
9	4
10	5
11	8
12	9
13	5
14	1
15	4
16	5
17	1
18	1
19	1

Example 3.2. Consider the simple data below

16	14	15	13	14	16	15	15	14	12
17	16	13	16	15	14	18	13	15	17

Construct an ungrouped frequency distribution table.

Solution. The corresponding ungrouped frequency distribution table is

x	12	13	14	15	16	17	18
Tallies	/	///	////		////	//	/
Frequency	1	3	4	5	4	2	1

Note: While tallying is used for 5 counts and NOT /////.

Question 3.2. *The following data represent the number of days 30 undergraduate students abscond lectures in a semester in the university.*

5	13	9	12	7	4	8	6	6	10	7	11	10	8	15
8	6	9	12	10	7	11	10	8	12	9	7	10	7	8

Construct a frequency table for these data.

Ranked Data

A slightly more convenient method of tabulating a collection of data would be to arrange them in rank order, so making it easier to see how many times each number appears. This is known as "ranked data".

Example 3.3.

Two types of frequency distributions that are most often used are the *categorical frequency distribution* and the *grouped frequency distribution*. The procedures for constructing these distributions are shown now.

Categorical Frequency Distributions

The categorical frequency distribution is used for data that can be placed in specific categories, such as nominal- or ordinal-level data. For example, data such as political affiliation, religious affiliation, or major field of study would use categorical frequency distributions.

Grouped Frequency Distribution Tables

For about forty or more items in a set of numerical data, it usually most convenient to group them together into between 5 and 20 "**classes**" of values, each covering a specified range or "**class interval**". For example; if we are interested in the distribution of weights in kg of students as follows:

$$45 - 49, \qquad 50 - 54, \qquad 55 - 59, \qquad 60 - 64, \cdots, \quad 85 - 89.$$

Each item is counted every time it appears in order to obtain the "**class frequency**" and each class interval has the same "**class width**". Too few classes means that the data is over-summarised, while too many classes means that there is little advantage in summarising at all. When data is arranged this way we call it a **grouped data**. The resulting table is called a **grouped frequency distribution table.**

Here, we use the convention that the lower boundary of the class is included while the upper boundary is excluded. Each item in a particular class is considered to be approximately equal to the "**class midpoint**"; that is, the average of the two "**class boundaries**". A "grouped frequency distribution table" normally has columns which show the class intervals, class mid-points, class frequencies, and "cumulative frequencies", the last of these being a running total of the frequencies themselves. There may also be a column of "tallied frequencies", if the table is being constructed from the raw data without having first arranged the values in rank order.

Steps in the Construction of a Grouped Frequency Distribution

1. Select the number of classes k, using some professional judgement so as k falls between 5 and 20. One such good guideline is to pick k such that $2^k \geq n$, so that if the sample size, $n = 20$, $k = 5$ because $2^5 = 32 > n$ and if $n = 80$, $k = 7$ because $2^7 = 128 > n$. To be more specific, we can solve for k to get $k \geq \dfrac{\log n}{\log 2}$.

2. Find the largest and smallest values and compute the **working range** denoted by R.

R = Maximum Value − Desired *Lower Class Limit (LCL) of starting class.*

LCL of the starting class is normally the Minimum value in the data or any other value slightly less than the minimum value.

3. Identify the smallest unit of measurement (u) used in the data collection. The value of u can be inferred from the given data or the given starting value (usually tens (10), ones(1), oneth (0.1) and tenth (0.01) etc. For example

10	20	40	60	$u = 10$
15	12	11	52	$u = 1$
2.8	1.6	1.7	5.6	$u = 0.1$

Estimate the class interval (i) (sometimes denoted by c) as

i = Round up $\left(\frac{R}{k}\right)$ to the nearest u.

Note: You must Round Up, not Round Off. For $u = 1$, Round Up $(5.2) = 6$ not 5 and for $u = 0.1$ Round Up $(5.21) = 5.3 \text{ not } 5.2$. If $\frac{R}{k}$ is exact (no remainder when divided by u) add one to the number of classes.

4. The starting value used in calculation of R above is picked as the lower class limit (LCL) of the of the first class. Add the class interval i to this LCL successfully to get the rest of the lower class limits.

5. Find the Upper Class Limit (UCL) of the first class by subtracting u from the LCL of the second class. Then continue to add the class interval i to this UCL to find the rest of the upper limits.

6. If necessary, find the class boundaries (CB) for each class as follows. Lower Class Boundary LCB = $LCL - 0.5u$

Upper Class Boundary $UCB = UCL + 0.5u$

Upper Class Boundary UCB = UCL + 0.5u

7. Tally the number of observations falling in each class and find the frequencies.

Note: A value x falls into a class $LCL - UCL$ only if $LCB \leq x < UCB$. That is x can be equal to LCB but not UCB of that class.

8. Record the number of tallies in each category as the class frequencies.

9. Compute the cumulative frequencies to confirm that the last value of the column is equal to the sum of the frequencies.

10. Compute the midpoints of each class using the class boundaries.

Example 3.4. The Dean of the Faculty of Science wishes to determine the amount of studying BCM students do. He selects a random sample of 40 students and records the number of hours each student studies per week as follows

15.0	23.7	19.7	15.4	18.3	23.0	17.5	20.8	13.5	20.7
17.4	18.6	12.9	20.3	23.7	21.4	18.3	29.8	17.1	18.9
10.3	26.1	15.7	24.0	17.8	32.8	23.2	24.5	27.1	16.6
9.2	16.5	30.8	29.6	24.6	12.5	21.6	28.4	27.9	22.4

Organize the data into a grouped frequency distribution.

Solution. First, we select the number of classes

$$n = 40,\ 2^k \geq 40 \Rightarrow k = 6.$$

Identify the smallest and the largest values of the data. In this case the smallest value is 9.2 and the largest is 32.8. The range is given by

R = Maximum Value − Desired *Lower Class Limit (LCL) of starting class.*

$$Range(R) = 32.8 - 9.2 = 23.6$$

$$n = 40,\ k = 6,\ u = 0.1$$

Therefore the class interval (i) is given by

$$i = \text{Round Up}\left[\frac{R}{k}\right] = \text{Round Up}\left[\frac{23.6}{6}\right] = \text{Round Up } [3.933]$$

$$= 4.0 \text{ to the nearest } u$$

The LCL of the first class should be 9.2 (because it was used in the calculation of Range), adding 4 to 9.2 gives 13.2, adding 4 to 13.2 gives 17.2 and so on.

For the upper class limits, the smallest unit of measurement is 0.1 so $u = 0.1$, UCL of the first class = $13.2 - 0.1 = 13.1$. Adding 4 to 13.1 gives 17.1, adding 4 to 17.1 gives 21.1 and so on.

The frequency table is as given below:

Class-intervals	Class-boundaries	Tallies	Frequency (*f*)	*cf*
9.20 - 13.1	9.15 - 13.15	////	4	4
13.2 - 17.1	13.15 - 17.15		7	11
17.2 - 21.1	17.15 - 21.15		11	22
21.2 - 25.1	21.15 - 25.15		10	32
25.2 - 29.1	25.15 - 29.15	////	4	36
29.2 - 33.1	29.95 - 33.15	////	4	40

Note: The last column in the table contains cumulative frequencies (*cf*) which gives the total number of observations equal to or less than the UCB of a particular class. The cumulative frequency is obtained by successively adding the frequencies of values of the variable from the lowest to highest value or class.

Example 3.5. Suppose in the previous example we were to use 9.0 as the starting value of the first class, then the interval would become

$$\text{Range } R = 32.8 - 9.0 = 23.8$$

$$i = \text{Round up } \left[\frac{R}{k}\right] = \text{Round up } [3.967] = 4.0 \text{ to the nearest } u$$

The frequency table would be

Class	9.0 - 12.9	13.0 - 16.9	17.0 - 20.9	21.0 - 24.9	25.0 - 28.9	29.0 - 32.9
Tally	////				////	////
Frequency	4	6	12	10	4	4

Example 3.6. Consider the data below

2370	1970	1540	1830	1500	2300	1750
1740	1860	1290	2030	2370	2140	1830
1030	2610	1570	2400	1780	3280	2320
920	1650	3080	2960	2460	1250	2160

Organize the data into a grouped frequency distribution.

Solution. $n = 28$, we pick the smallest k such that $2^k > 28 \Rightarrow k = 5$, If 920 is the starting (minimum) value,

$$\text{Range } (R) = 3280 - 920 = 2360$$

Therefore the class interval (i) is given by

$$\frac{R}{k} = \frac{2360}{5} = 472$$

$$i = \text{Round UP } (472) = 480 \text{ to the nearest } u = 10$$

Note: When the value of $i = \frac{R}{k}$ has no remainder, we increase the number of classes by 1, thus $new k = k+1$. Our $k = 6$.

The corresponding frequency table is

Class	920-1370	1380-1830	1840-2290	2300-2750	2760-3210	3220-3670
Tally	////				//	/
Frequency	4	9	5	7	2	1

Example 3.7. Consider the figures recorded in the table below which gives the weight of Oranges measured to the nearest gram.

Organize the data into a group frequency distribution.

Frequency Distribution Tables

Thirdly, it is possible to save a little space by making a table in which each individual item of the ranked data is written down once only, but paired with the number of times it occurs. The data is then presented in the form of a "**frequency distribution table**". **Cumulative Frequency Tables**

Cumulative frequencies are obtained by accumulating the frequencies to give the total number of observations up to and including the value or group in question. For grouped data it is natural to relate the cumulative frequency to the upper boundaries of the groups.

Lecture 4: Graphical Displays

4 The Graphical Representations of data

4.1 Introduction

Visualization techniques are ways of creating and manipulating graphical representations of data. We use these representations in order to gain better insight and understanding of the problem we are studying - pictures can convey an overall message much better than a list of numbers. In this section we describe some graphical presentations of data.

Line or Dot Plots

Line plots are graphical representations of numerical data. A **line plot** is a number line with x's placed above specific numbers to show their frequency. By the **frequency** of a number we mean the number of occurrence of that number. Line plots are used to represent one group of data with fewer than 50 values.

Example 4.1. Suppose thirty people live in an apartment building. These are the following ages:

58	30	37	36	34	49	35	40	47	47
39	54	47	48	54	50	35	40	38	47
48	34	40	46	49	47	35	48	47	46

Make a line plot of the ages.

Line plots allow several features of the data to become more obvious. For example, outliers, clusters, and gaps are apparent.

- **Outliers** are data points whose values are significantly larger or smaller than other values, such as the ages of 30, and 58.
- **Clusters** are isolated groups of points, such as the ages of 46 through 50.
- **Gaps** are large spaces between points, such as 4 and 45.

Data from a frequency table can be represented graphically. Graphical representation of statistical data is used to bring to light the prominent features of the data at a glance and make visual comparisons of the data easier.

There are several different graphical displays for describing a frequency distribution data. Some of the commonly used displays are; Bar charts, Histograms, Frequency Polygons, and cumulative frequency curves or ogive, pie charts, and pictograph etc.

Stem and Leaf Plots

Another type of graph is the **stem-and-leaf plot**. It is closely related to the line plot except that the number line is usually vertical, and digits are used instead of x's. To illustrate the method, consider the following scores which twenty students got is a statistics test:

69	84	52	93	61	74	79	65	88	63
57	64	67	72	74	55	82	61	68	77

We divide each data value into two parts. The left group is called a stem and the remaining group of digits on the right is called a leaf. We display horizontal rows of leaves attached to a vertical column of stems. We can construct the following table

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

where the stems are the ten digits of the scores and the leaves are the one digits.

The disadvantage of the stem-and-leaf plots is that data must be grouped according to place value. What if one wants to use different groupings? In this case histograms, to be discussed below, are more suited.

If you are comparing two sets of data, you can use a **back-to-back stem-and-leaf plot** where the leaves are sets listed on either side of the stem as shown in the table below.

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

where the stems represents the tens digits of a science test scores and the leaves represent the ones digits.

Example 4.2. Suppose the members of your class scored the following percentages in a statistics test:

32	56	45	78	77	59	65	54	54	39
45	44	52	47	50	52	51	40	69	72
36	57	55	47	33	39	66	61	48	45
53	57	56	55	71	63	62	65	58	55

Construct a stem and leaf diagram.

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

Key: $6/2 = 62$.

Notice the stem and leaf display is visual representation of the data. It is easy to see that there are more marks in the fifties than in the seventies.

The Histogram

A "**histogram**" is a diagram which is directly related to a grouped frequency distribution table and consists of a collection of rectangles whose height represents the class frequency (to some suitable scale) and whose breadth represents the class width.

A histogram is a bar graph on which the bars are adjacent to each other with no space between them. In a histogram, each of the classes in the frequency distribution is represented by a vertical bar whose height is the class frequency of the interval. The horizontal endpoints of each vertical bar correspond to the class endpoints.

To construct a histogram, arrange the data in equal intervals. Represent the frequencies along the vertical axis and the scores along the horizontal axis. The true limits of any interval extend one half unit beyond the endpoints established for the interval and are represented in this manner on the horizontal axis. For example, the true limits of the interval 76-80 are 75.5 and 80.5. To get the proper perspective, the vertical axis should be approximately three-fourths as long as the horizontal axis.

Histograms are close relatives of bar plots. The main difference is that in histograms we are interested in the distribution of data. In other words, we wish to know if there is regularity in the number of observationsthatfallwithinacategory.

Thismeansthathowthedataarebinnedtakesonanadditionalimportance. For grouped data the height of each rectangle is the **relative frequency** (h) of a class given by

$$h = \frac{f}{i}$$

where f is the class frequency and i is the class interval. The value represented on the x-axis of a histogram is the middle point of the classes that determine the width of the bars (placed at the middle of the bar) or class boundaries (placed at the edges of the bars).

Remark: We will deal with equal class intervals' histograms only.

Example 4.3. Draw a Histogram for the following distribution giving the marks obtained by 60 students of a class in a college.

Marks:	20-24	25-29	30-34	35-39	40-44	45-49	50-54
Number of Students:	3	5	12	18	14	6	2

Solution. Here class intervals given are of inclusive type. The upper limit of a class is not equal to the lower limit of its following class, therefore class boundaries will have to be determined. After the adjustment, the distribution will be as below

Practice Problems

Question 4.1. Illustrate the following set of measurements on a histogram:

72	82	56	73	87	89	72	86	88	76
86	69	84	85	62	97	70	78	84	93
70	60	91	76	83	94	65	72	92	81
98	78	88	76	96	89	90	83	74	80

Question 4.2. The height of 100 maize plants was measured, to the nearest cm, one month after planting.

Height:	1-20	21-40	41-60	61 - 80
Number of Plants:	12	28	54	6

Construct the corresponding histogram.

Question 4.3. Construct a histogram for the following scores earned by a group of high school students on a Statistical Aptitude Examination.

Score	**Number of students**
400-449	20
450-499	35
500-549	50
550-599	50
600-649	40
650-699	20
700-749	10

Question 4.4. The weights of 40 football players are as follows:

210	181	192	164	170	186	205	194
178	161	175	195	172	188	196	182
206	188	165	202	178	163	190	198
187	198	174	172	183	208	185	162
203	172	196	184	185	176	197	184

(a) Construct a frequency distribution for the given data.

(b) Make a histogram for the given data.

Question 4.5. The following table shows some test scores from a statistics class.

65	91	85	76	85	87	79	93
82	75	100	70	88	78	83	59
87	69	89	54	74	89	83	80
94	67	77	92	82	70	94	84
96	98	46	70	90	96	88	72

(a) Construct a frequency distribution

(b) Construct the corresponding histogram.

Question 4.6. Suppose a sample of 38 female university students were asked their weights in pounds. This was actually done, with the following results:

130	108	135	120	97	110
130	112	123	117	170	124
120	133	87	130	160	128
110	135	115	127	102	130
89	135	87	135	115	110
105	130	115	100	125	120
120	120				

(a) Construct a frequency distribution

(b) Construct the corresponding histogram.

Question 4.7. The table below shows the response times of calls for police service measured in minutes.

34	10	4	3	9	18	4
3	14	8	15	19	24	9
36	5	7	13	17	22	27
3	6	11	16	21	26	31
32	38	40	30	47	53	14
6	12	18	23	28	33	
3	4	62	24	35	54	
15	6	13	19	3	4	
4	20	5	4	5	5	
10	25	7	7	42	44	

Construct a frequency distribution and the corresponding histogram.

Question 4.8. A nutritionist is interested in knowing the percent of calories from fat which Students intake on a daily basis. To study this, the nutritionist randomly selects 25 students and evaluates the percent of calories from fat consumed in atypical day. The results of the study are as follows:

24%	18%	33%	25%	30 %
42%	40%	33%	39%	40 %
45%	35%	45%	25%	27 %
23%	32%	33%	47%	23 %
27%	32%	30%	28%	36 %

Construct a frequency distribution and the corresponding histogram.

Frequency Polygon

It is another method of representing a frequency distribution of a graph. Frequency polygons are more suitable than histograms whenever two or more frequency distributions are to be compared.

Frequency polygon of a grouped or continuous frequency distribution is a straight line graph. The frequencies of the classes are plotted against the mid-values of the corresponding classes. The points so obtained are joined by straight lines (segments) to obtain the frequency polygon. It can also be obtained by connecting midpoints of the top of the rectangles in a Histogram. The gaps at both ends are extended to the next lower and the next upper class mark (imaginary classes with frequency zero). For grouped data, a straight line graph is drawn with class frequency plotted against class mark (midpoint).

Example 4.4. The following data show the number of accidents sustained by 313 drivers of a public utility company over a period of 5 years. Draw the frequency polygon.

No. of accidents:	0	1	2	3	4	5	6	7	8	9	10	11
No. of drivers:	80	44	68	41	25	20	13	7	5	4	3	2

Example 4.5. Draw a Histogram and Frequency polygon from the following distribution giving marks of 50 students in statistics.

Marks:	0-9	10-19	20 – 29	30 – 39	40 – 49	50 – 59	60 – 69	70 – 79	80-89
No.:	0	2	3	7	13	13	9	2	1

Note that, here we will first draw histogram and then the mid-points of the top of bars are joined by line segments to get the frequency polygon.

Remark: Note that frequency polygon can be drawn even without converting the given distribution into classes. The frequencies are plotted against the corresponding mid-points (given) and joined by line segment.

Frequency Curve

Frequency curve is similar like frequency polygon, only the difference is that the points are joined by a free hand curve instead of line segments as we join in frequency polygon.

Let us study the following examples to understand the concept of frequency curve.

Example 4.6. Draw a frequency curve for the following data:

Solution. Here we will take ages on horizontal axis and number of students on vertical axis. We will plot the given frequencies against mid-points of the given class interval and then join these points by free hand curve. Extremities (first and the last point plotted) are joined to the mid-points of the neighboring class intervals.

Cumulative Frequency Polygon

A line graph of cumulative frequency plotted against the upper class boundaries (UCBs) is called an **Ogive** or **cumulative frequency curve**. The Ogive curve is very useful in estimating the median and the measures of location as we will see later.

Example 4.7. The data below give the marks secured by 70 students at a certain examination:

(a) Draw the ogive curve.

(b) Use the ogive curve to estimate the percentage of students getting less than 45.

Solution. We will plot the cumulative frequencies on the vertical axis against the upper class boundaries of the corresponding class on the horizontal axis. We will then join the points by a smooth free hand to get an ogive.

To estimate the number of students getting marks less than 45, draw a perpendicular to the X-axis (representing marks) at $X = 45$, meeting the ogive at point P. From P draw a perpendicular PM on the Y-axis (representing number of students). Then, from the graph $OM = 26.8 \approx 27$ is the number of candidates getting score 45 or less. Hence, the percentage of students getting less than 45 marks is given by

$$\frac{27}{70} \times 100 = 38.57$$

(Here total students = 70).

Bar Graphs

Bar Graphs, similar to histograms, are often useful in conveying information about categorical data where the horizontal scale represents some non-numerical attribute. In a bar graph, the bars are non overlapping rectangles of equal width and they are equally spaced. The bars can be vertical or horizontal. The length of a bar represents the quantity we wish to compare.

Example 4.8. The areas of the various continents of the world (in millions of square miles) are as follows: 11.7 for Africa; 10.4 for Asia; 1.9 for Europe; 9.4 for North America; 3.3 Oceania; 6.9 South America; 7.9 Soviet Union. Draw a bar chart representing the above data and where the bars are horizontal.

A double bar graph is similar to a regular bar graph, but gives 2 pieces of information for each item on the vertical axis, rather than just 1.

Practice Problems

Question 4.9. Given are several gasoline vehicles and their fuel consumption averages.

Buick 27 mpg

BMW	28 mpg
Honda Civic	35 mpg
Geo	46 mpg
Neon	38 mpg
Land Rover	16 mpg

(a) Draw a bar graph to represent these data.

(b) Which model gets the least miles per gallon? the most?

Question 4.10. The following data gives the number of murder victims in the U.S in 1978 classified by the type of weapon used on them: Gun, 11,910; cutting/stabbing, 3,526; blunt object, 896; strangulation/beating, 1,422; arson, 255; all others 705. Construct a bar chart for this data. Use vertical bars.

Line Graphs

A **line graph** (or **time series plot**) is particularly appropriate for representing data that vary continuously. Aline graph typically shows the trend of a variable over time. To construct a time series plot, we put time on the horizontal scale and the variable being measured on the vertical scale and then we connect the points using line segments.

Example 4.9. The population (in millions) of the US for the years 1860-1950 is as follows: 31.4 in 1860 ; 39.8 in 1870; 502, in 1880; 62.9 in 1890;76.0 in 1900; 92.0 in 1910;105.7 in 1920; 122.8 in 1930; 131.7 in 1940; and 151.1 in 1950. Make a time plot showing this information.

Circle Graphs or Pie Chart

Another type of graph used to represent data is the circle graph. A **circle graph** or **pie chart**, consists of a circular region partitioned into disjoint sections, with each section representing a part or percentage of a whole. To construct a pie chart we first convert the distribution into a percentage distribution. Then, since a complete circle corresponds to 360 degrees, we obtain the central angles of the various sectors by multiplying the percentages by 3.6. Each sector is then labeled by the group it represents and indicate the corresponding percentages.

Steps in constructing a pie chart.

1. Add up the given quantities (Let S be the sum of the values)
2. For each quantity calculate angle represented as $\frac{X}{S} \times 360^\circ$
3. For each quantity calculate percentage represented as $\frac{X}{S} \times 100\ \%$
4. Draw a circle and divide it into sectors using the angles calculated in step 2.
5. Label each sector by the group represented and indicate the corresponding percentage.

We illustrate this method in the next example.

Example 4.10. Asurvey of1000 adultsuncoveredsomeinterestinghousekeeping secrets. Whenunexpected company comes, where do we hide the mess? The survey shows that 68% of the respondents toss their

mess in the closet, 23% shove things under the bed, 6% put things in the bath tub, and 3% put the mess in the freezer. Make a circle graph to display this information.

Solution. We first find the central angle corresponding to each case:

$$\begin{aligned} \textit{in closet} \quad & 68 \times 3.6 = 244.8 \\ \textit{under bed} \quad & 23 \times 3.6 = 82.8 \\ \textit{in bathtub} \quad & 6 \times 3.6 = 21.6 \\ \textit{in freezer} \quad & 3 \times 3.6 = 10.8 \end{aligned}$$

Note that

$$244.8 + 82.8 + 21.6 + 10.8 = 360.$$

The pie chart is given in Figure

Example 4.11. Asampleof250studentswereaskedtoindicatetheirfavouriteTVstationandtheirresponses were as follows; KBC - 52, CITIZEN - 28, KTN - 63, STV - 15 and NTV - 92 viewers. Draw a pie chart representing this information.

Station	No. of Viewers	Angle	Percent
KBC	52	74.88	20.80
Citizen	28	40.32	11.20
KTN	63	90.72	25.20
STV	15	21.60	6.00
NTV	93	132.48	36.80

The corresponding pie chart is given below:

Practice Problems

Question 4.11. The table below shows the ingredients used to make a sausage and mushroom pizza.

Ingredient	%
Sausage	7.5
Cheese	25
Crust	50
Tomato Sauce	12.5
Mushroom	5

Plot a pie chart for the data.

Question 4.12. A newly qualifies teacher was given the following information about the regional origins of the pupils in a class.

Region	No. of pupils
Central	12
Rift Valley	7
Coast	2
Western	3

Nyanza	6
TOTAL	30

Plot a pie chart representing the data.

Question 4.13. The following table represents a survey of people's favorite ice cream flavor.

Flavor	*Number of people*
Vanilla	21.0 %
Chocolate	33.0 %
Strawberry	12.0 %
Raspberry	4.0 %
Peach	7.0 %
Neopolitan	17.0 %
Others	6.0 %

Plot a pie chart to representing the data.

Question 4.14. In Kenya, approximately 45% of the population has blood type O; 40% type A; 11% type B; and 4% type AB. Illustrate this distribution of blood types with a pie chart.

Pictograph

Another type of chart which has been used widely is the **pictograph**. In a pictograph, a symbol or icon is used to represent a quantity of items. A pictograph needs a title to describe what is being presented and how the data are classified as well as the time period and the source of the data. It is also called **pictogram**. Example of a pictograph is given in Figure...

A disadvantage of a pictograph is that it is hard to quantify partial icons.

Practice Problems

Question 4.15. Make a pictograph to represent the data in the following table. Use ♣ to represent 10 glasses of lemonade.

Day	**Frequency**
Monday	15
Tuesday	20
Wednesday	30
Thursday	5
Friday	10

Scatter plots

A relationship between two sets of data is sometimes determined by using a **scatterplot**. Let's consider the question of whether studying longer for a test will lead to better scores. A collection of data is given below.

Study Hours	3	5	2	6	7	1	2	7	1	7
Score	80	90	75	80	90	50	65	85	40	100

Based on these data scatterplot has been prepared and is given in Figure... (Remember when making a scatterplot, do NOT connect the dots.)

The data displayed on the graph resembles a line rising from left to right. Since the slope of the line is positive, there is a positive correlation between the two sets of data. This means that according to this set of data, the longer I study, the better grade I will get on my exam score.

If the slope of the line had been negative (falling from left to right), a negative correlation would exist. Under a negative correlation, the longer I study, the worse grade I would get on my exam.

If the plot on the graph is scattered in such a way that it does not approximate a line (it does not appear to rise or fall), there is no correlation between the sets of data. No correlation means that the data just doesn't show if studying longer has any affect on my exam score. *(will de done later in correlation and Regression analysis.*

Lecture 5: Measures of Central Tendency

5 Measures of Central Tendency

5.1 Introduction

In the previous lecture, we saw how raw data is converted into frequency distributions and visual displays. We will now examine statistical methods for describing typical values in the data as well as the extent to which the data are spread out.

There is a tendency in almost every statistical data that most of the values concentrate at the centre which is referred as "central tendency". The typical values which measure the central tendency are called **measures of central tendency** or **measures of location**. Measures of central tendency are commonly known as "Averages". They are also known as **first order measures**. Averages always lie between the lowest and the highest observation.

The purpose for computing an average value for a set of observations is to obtain a single value which is representative of all the items and which the mind can grasp simply and quickly. The single value is the point or location around which the individual items cluster.

There are a number of different quantities, which can be used to estimate the central point of a sample. The various types of measures of central tendency for statistical distribution discussed in this lecture include; Arithmetic mean or simply the mean, Weighted arithmetic mean, Geometric mean, Harmonic mean, median and Quartiles, deciles, percentiles, and mode. Of these, arithmetic mean, geometric mean and harmonic mean are called mathematical averages; median and mode are called positional averages.

5.2 Arithmetic Mean

By far the most common measure for describing the location of a set of data is the mean. The mean (or average) of observations, as we know, is the sum of the values of all the observations divided by the total number of observations.

The **arithmetic mean** is defined as the sum of the data values divided by the number of observations. Also referred to as the *arithmetic average* or simply the *mean*.

5.2.1 For Simple or Ungrouped data

Arithmetic Mean is defined as the sum of all the observations divided by the total number of observations in the data and is denoted by $\bar{X}$, which is read as 'X- bar'.

In general, if $X_1, X_2, \cdots, X_n$, or X_i, $i = 1,2,\cdots,n$ are the n observations of variable x, then the arithmetic mean is defined by

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$
$$= \frac{1}{n}\sum_{i=1}^{n} X_i$$

Where

$\bar{X}$ = sample mean

X_i = the ith data value in the sample Σ = the sum of n = number of data value in the sample

Example 5.1. Find the arithmetic mean for the following data representing marks in six subjects at the university examination of a student. The marks are 74, 89, 93, 68, 85 and 76.

Solution. $n = 6$, and $\bar{x} = \sum_{i=1}^{n} X_i$.

$$\bar{X} = \frac{74 + 89 + 93 + 68 + 85 + 76}{6} = \frac{485}{6}$$
$$= 80.83$$

Example 5.2. Compute the sample mean for the values,

9,3,4,2,1,5,8,4,7, and 3.

Solution. $n = 10$, $X_1 = 9$, $X_3 = 4,\cdots, X_{10} = 3$,

$$\sum_{i=1}^{5} X_i = 9 + 3 + 4 + 2 + 1 = 19$$
$$\sum_{i=3}^{5} X_i = 4 + 2 + 1 = 7$$

Using this notation we can calculate the average as

$$\bar{x} = \frac{1}{10}\sum_{i=1}^{n} X_i = \frac{9+3+\cdots+7+3}{10} = \frac{46}{10} = 4.6$$

Example 5.3. A sample of five executive received the following amounts of bonus last year: 14,000, 15 ,000, 17,000, 16,000, and y. Find the value of y if the average bonus for these five executives is 15,400.

Solution. Since these values represent a sample size of 5, the sample mean is

$$\begin{aligned}
\bar{y} &= \frac{1}{n}\sum_{i=1}^{5} Y_i \\
&= \frac{14,000 + 15,000 + 17,000 + 16,000 + Y}{5} = 15,400 \\
&\Rightarrow 62,000 + Y = 15,400 \times 5 \\
Y &= 177,000 - 62,000 \\
&= 15,000
\end{aligned}$$

Example 5.4. Find the mean of the numbers 5, 2, 3 ,7, and 3.

Solution. The mean is given as

$$\bar{X} = \frac{5+2+3+7+3}{5} = \frac{20}{5} = 4$$

Question 5.1. Find arithmetic mean for the following data; 425 , 408 , 441 , 435 , 418.

5.2.2 For Grouped data (or) discrete data with frequencies

If $X_1, X_2, \cdots, X_n$ are the values of the variable X with corresponding frequencies, $f_1, f_2, \cdots, f_n$, then the arithmetic mean of X, where $\sum f_i = n$, is given by

$$\begin{aligned}
\bar{X} &= \frac{X_1 f_1 + X_2 f_2 + \cdots + X_n f_n}{f_1 + f_2 + \cdots + f_n} \\
&= \frac{\Sigma f_i X_i}{\Sigma f_i} = \frac{1}{n}\sum_{i=1}^{n} f_i X_i
\end{aligned}$$

For example, for the family size distribution data the mean number of children in the sample is from the frequency distribution:

$$\bar{X} = \frac{\sum f_i X_i}{\sum f_i} = \frac{186}{80} = 2.325$$

Example 5.5. Calculate arithmetic mean for the following data.

Age in years:	11	12	13	14	15	16	17
No. of Students:	7	10	16	12	8	11	5

Solution. The data is represented in the table below:

Age in years (X)	No. of Students (f)	fX
11	7	77
12	10	120
13	16	208
14	12	168
15	8	120
16	11	176
17	5	85
Totals:	$n = 69$	$\sum fX = 954$

$$\sum f = n = 69, \quad \sum fX = 954$$

$$\bar{X} = \frac{\sum fX}{n} = 13.83 \text{ years}$$

Question 5.2. Calculate the average bonus paid per member from the following data:

Bonus (in Ksh.):	40	50	60	70	80	90	100
No. of persons:	2	5	7	6	4	8	3

5.2.3 Grouped data with class intervals and frequencies

For grouped data i.e., when the data is represented as a frequency distribution table with class intervals, the mean is calculated using the above formula:

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

with x_i standing for the class mid-point (i.e., the mid value of the class interval).

For grouped data the mid-point of each group would normally be used in the frequency distribution to determine the mean.

Steps:

(i) Obtain the mid-point of each class interval

$$\text{Mid-point} = \frac{(\text{lower limit} + \text{upper limit})}{2}$$

(ii) Multiply these mid-points by the respective frequency of each class interval and obtain the total $\Sigma f_i x_i$.

(iii) Divide the total obtained by step (2) by the total frequency Σf_i.

Example 5.6. Find the arithmetic mean for the following data representing marks of 60 students.

Marks:	10-19	20-29	30-39	40-49	50-59	60-69	70-79
No. of Students:	8	15	13	10	7	4	3

Solution.

Marks	No. of Students f_i	Mid-points x_i	$f_i x_i$
10-19	8		
20-29	15		
30-39	13		
40-49	10		
40-59	7		
60-69	4		
70-79	3		
Totals:			$\sum f_i x_i =$

$$\bar{x} = \frac{\Sigma f_i x_i}{\Sigma f_i} = \frac{1}{60} =$$

Hence the average marks are .

Example 5.7. Given the following frequency distribution, calculate the mean.

Classes	Class boundaries	Midpoints (x_i)	Frequency (f_i)	$f_i x_i$
$13-17$	$12.5-17.5$	15	2	30
$18-22$	$17.5-22.5$	20	22	440
$23-27$	$22.5-27.5$	25	19	475
$28-32$	$27.5-32.5$	30	14	420
$33-37$	$32.5-37.5$	35	3	105
$38-42$	$37.5-42.5$	40	4	160
$43-47$	$42.5-47.5$	45	6	270
$48-52$	$47.5-52.5$	50	1	50
$53-57$	$52.5-57.5$	55	1	55
			$\Sigma f_i = 72$	$\Sigma f_i x_i = 2005$

The arithmetic mean is given by

$$\bar{x} = \frac{\Sigma f_i x_i}{\Sigma f_i} = \frac{2005}{72} = 27.85$$

Hence the mean weekly wage is 27.85.

This new method of finding the mean is known as the **Direct Method**.

Sometimes when the numerical values of x_i and f_i are large, finding the product of x_i and f_i becomes tedious and time consuming. So, for such situations, let us think of a method of reducing these calculations.

We can do nothing with the f_i's, but we can change each x_i to a smaller number so that our calculations become easy. How do we do this? What about subtracting a fixed number from each of these x_i's?

The first step is to choose one among the x_i's as the assumed mean, and denote it by 'a'. Also, to further reduce our calculation work, we may take 'a' to be that x_i which lies in the centre of $x_1, x_2, \cdots, x_n$.

The next step is to find the difference d_i between 'a' and each of the x_i's, that is, the **deviation** of A from each of the x_i's.

i.e.,

$$d_i = x_i - a$$

The third step is to find the product of d_i with the corresponding f_i, and take the sum of all the $f_i d_i$'s. So, the mean of the deviations, $\bar{d} = \dfrac{\sum f_i d_i}{\sum f_i}$.

Now, let us find the relation between $\bar{d}$ and $\bar{x}$. Since in obtaining d_i, we subtracted 'a' from each x_i, so, in order to get the mean $\bar{x}$, we need to add A to $\bar{d}$. This can be expressed mathematically as: Mean of deviations,

$$\begin{aligned}\bar{d} &= \frac{\sum f_i d_i}{\sum f_i} = \frac{\sum f_i (x_i - a)}{\sum f_i} \\ &= \frac{\sum f_i x_i}{\sum f_i} - a\frac{\sum f_i}{\sum f_i} \\ &= \bar{x} - a\end{aligned}$$

So,

$$\bar{x} = a + \bar{d}$$

If the classes are of equal width the work of calculating the mean is made easy by change of origin and scale. The assumed mean method gives the mean as:

$$\bar{x} = a + \frac{\sum f_i d_i}{\sum f_i}$$

This method is better illustrated by an example.

Example 5.8. Solve the above example on wages using the assumed mean method.

Solution. Let the assumed mean be $a = 35$.

(x_i)	(f_i)	$d_i = (x_i - a)$	$f_i d_i$
15	2	-20	-40
20	22	-15	-330
25	19	-10	-190
30	14	-5	-70
35	3	0	0
40	4	5	20
45	6	10	60
50	1	15	15
55	1	20	20
	$\sum f_i = 72$		$\sum f_i d_i = -515$

Using the formula;

$$\bar{x} = a + \frac{\sum f_i d_i}{\sum f_i} = 35 + \frac{-515}{72}$$
$$= 35 - 7.1528 = 27.85$$

The method discussed above is called the **Assumed Mean Method**.

Activity 5.1. From the Table ,... find the mean by taking each of x_i (i.e., ... and so on) as a. What do you observe? You will find that the mean determined in each case is the same, i.e., 27.85. (Why?)

So, we can conclude that the value of the mean obtained does not depend on the choice of 'a'. If after subtracting assumed mean a from each of the observations, the values are still large, one could divide the deviations with a constant value c so that $u_i = \frac{x_i - a}{c}$, where a is the assumed mean and c is the class size.

Observe that in Table ..., the values in Column 4 are all multiples of 15..... So, if we divide the values in the entire Column 4 by 15, we would get smaller numbers to multiply with f_i (Here, 15 is the class size of each class interval.)

Now, we calculate u_i as above and continue as before (i.e., find $f_i u_i$ and then $\Sigma f_i u_i$. Taking c = 5, let us construct Table...

X	f	$u = \frac{(X-35)}{5}$	fu
15	2	-4	-8
20	22	-3	-66
25	19	-2	-38
30	14	-1	-14
35	3	0	0
40	4	1	4
45	6	2	12
50	1	3	3
55	1	4	4
	$\Sigma f = 72$		$\Sigma fu = -103$

Let

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

Here, again let us find the relationship between $\bar{u}$ and $\bar{x}$.

We have,

$$u_i = \frac{x_i - a}{c}$$

Therefore,

$$\bar{u} = \frac{\sum f_i \left(\frac{x_i - a}{c}\right)}{\sum f_i} = \frac{1}{c}\left[\frac{\sum f_i x_i - a\sum f_i}{\sum f_i}\right]$$
$$= \frac{1}{c}\left[\frac{\sum f_i x_i}{\sum f_i} - a\frac{\sum f_i}{\sum f_i}\right] = \frac{1}{c}[\bar{x} - a]$$
$$c\bar{u} = \bar{x} + a$$

So;

$$\begin{aligned}\bar{u} &= a + c\bar{u} \\ &= a + c\left(\frac{\sum f_i u_i}{\sum f_i}\right)\end{aligned}$$

Now substituting the values of a, c, $\sum f_i u_i$ and $\sum f_i$ from Table..., we get

$$\bar{x} = a + c\left(\frac{\sum f_i u_i}{\sum f_i}\right) = 35 + 5\left(\frac{-103}{72}\right) = 27.85$$

So the mean obtained is 27.85.

The method discussed above is called the **Step-deviation method** or sometimes **coding method**.

We note that:

- the step-deviation method will be convenient to apply if all the d_i's have a common factor.
- The mean obtained by all the three methods is the same.
- The assumed mean method and step-deviation method are just simplified forms of the direct method.
- The formula $\bar{x} = a + c\bar{u}$ still holds if a and c are not as given above, but are any non-zero numbers such that $u_i = \dfrac{x_i - a}{c}$.

Let us apply these methods in another example.

Example 5.9.

Remark: The result obtained by all the three methods is the same. So the choice of method to be used depends on the numerical values of x_i and f_i. If x_i and f_i are sufficiently small, then the direct method is an appropriate choice. If x_i and f_i are numerically large numbers, then we can go for the assumed mean method or step-deviation method. If the class sizes are unequal, and x_i are large numerically, we can still apply the step-deviation method by taking c to be a suitable divisor of all the d_i's.

5.3 Weighted Arithmetic Mean

One of the limitations of the arithmetic mean discussed above is that it gives equal importance to all the items. But these are cases where the relative importance of the different items is not the same. In these cases, weights are assigned to different items according to their importance. The term 'weight' stands for the relative importance of the different items.

If $x_1, x_2, \cdots, x_n$ are the n values of the variable X with the corresponding weights $w_1, w_2, \cdots, w_n$, then the weighted mean is given by;

$$\bar{x}_w = \frac{w_1x_1 + w_2x_2 + \cdots + w_nx_n}{w_1 + w_2 + \cdots + w_n} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

Where $\bar{x}_w$ = Weighted Arithmetic Mean and Σw_i = Sum of the weights.

The weights are normally assigned as measures of the importance of a subject to the issue under consideration.

Example 5.10. Calculate the weighted mean for the following data.

X	28	25	20	32	40
w	3	6	4	5	8

Solution. The data is represented in the table below:

x_i	w_i	$w_i x_i$
28	3	84
25	6	150
20	4	80
32	5	160
40	8	320
Totals:	26	794

$$\bar{x}_w = \frac{\sum w_i x_i}{\sum w} = \frac{794}{26} = 30.54 \text{ units}$$

Example 5.11. Consider the table below with marks obtained by two students James (marks x) and Jane (marks y). The subjects are to be used in determining who joins an Engineering course whose requirements is a mean of 58% in the four subjects.

For both students, mean $\bar{X}$ = $225/4$ = 56.25 implying that they are both unqualified but how much of history is required in the course?

Subject	% marks (X)	% marks (Y)	Weight (w)	wX	wY
Mathematics	25	70	3.6	90.0	252.0
English	87	45	2.3	200.1	103.5
History	83	35	1.5	124.5	52.5
Physics	30	75	2.6	78.0	195.0
Totals	225	225	10	492.6	603.0

Solution. If the subjects are given weights depending on their usefulness to the programme (column 4),

the weighted means are;

For James

$$\bar{X}_w = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i} = \frac{492.6}{10} = 49.26$$

which clearly falls below the required mean. For Jane

$$\frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i} = \frac{603}{10} = 60.3$$

which clearly above the required mean. Jane therefore qualifies for the admission into the programme.

Example 5.12. A tycoon has three house girls who he pays Ksh.2,000 per month each, two watchmen who receives Ksh.2,500 per month each and some gardeners who he pays ksh3,500 each. If he pays out an average of Ksh.2,850 per month to these people. Find the number of gardeners?

Solution. The weighted mean,

$$\bar{X}_w = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i}$$

Let n be the number of gardeners, then,

$$\begin{aligned} 2,850 &= \frac{(3 \times 2,000) + (2 \times 2,500) + (n \times 3,500)}{3 + 2 + n} \\ \Rightarrow 8,550 + 5,700 + 2,850n &= 11,000 + 3,500n \\ 3,250 &= 650n \\ n &= \frac{3,250}{650} = 5 \end{aligned}$$

There are 5 gardeners.

Question 5.3. A candidate obtained the following marks in percentages in an examination. English 64 , Mathematics 93, Economics 72, Accountancy 85 and Statistics 79. The weights of these subjects are 2, 3 , 3, 4, 1 respectively. Find the candidate's weighted mean.

Question 5.4. A teacher has decided to use a weighted average in figuring final grades for his students. The midterm examination will count 40%, the final examination will count 50% and quizzes 10%. Compute the average mark obtained for a student who got 90 marks for midterm examination, 80 marks for final and 70 for quizzes.

5.3.1 Merits of Arithmetic Mean

(1) It is rigidly defined.

(2) It is easy to understand and easy to calculate.

(3) It is based on each and every observation of the series.

(4) It is capable for further mathematical.

(5) It is least affected by sampling fluctuations.

5.3.2 Demerits of Arithmetic Mean

(1) It is very much affected by extreme observations.

(2) It can not be used in case of open end classes.

(3) It can not be determined by inspection nor it can be located graphically.

(4) It can not be obtained if a single observation is missing.

(5) It is a value which may not be present in the data.

5.4 Geometric Mean

The Geometric Mean is a special type of average where we multiply the numbers together and then take a square root (for two numbers), cube root (for three numbers) etc. The geometric mean is technically defined as the nth root of the product of n numbers, i.e., for a set of numbers $x_1, x_2, \cdots, x_n$, the geometric mean is defined as

$$\left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} = \sqrt[n]{x_1 x_2 \cdots x_n}.$$

For instance, the geometric mean of two numbers, say 4 and 9, is just the square root of their product, that is, $\sqrt{4 \cdot 9} = 6$. As another example, the geometric mean of the three numbers 4, 1, and 1/32 is the cube root of their product (1/8), which is 1/2, that is, $\sqrt[3]{4 \cdot 1 \cdot 1/32} = 1/2$.

Example 5.13. Calculate the Geometric Mean of 10, 51.2 and 8 ?

Solution. First we multiply them:

$$10 \times 51.2 \times 8 = 4096$$

Then (as there are three numbers) take the cube root:

$$\sqrt[3]{4096} = 16.$$

Hence the Geometric Mean is 16.

Question 5.5. Calculate the Geometric Mean of 3 and 27

Question 5.6. The Geometric Mean of three numbers is 8. Two of the numbers are 4 and 32. What is the third number?

Question 5.7. Find the Geometric Mean of the values 10, 5, 15, 8, 12.

(**ForGroupedData**)Ifwehaveaseriesof$n$positivevalueswithrepeatedvaluessuchas $x_1, x_2, x_3, \cdots, x_n$ are repeated $f_1, f_2, f_3, \cdots, f_n$ times respectively then the Geometric mean denoted by G is given as:

$$\text{G.M of} X = \sqrt[n]{x_1^{f_1} \cdot x_2^{f_2} \cdots x_n^{f_n}}$$

where $n = f_1 + f_2 + f_3 + \cdots + f_n$.

Example 5.14. Find the Geometric Mean of the following Data

X	13	14	15	16	17
f	2	5	13	7	3

Solution. Here

$$\begin{aligned} x_1 = 13, x_2 = 14, x_3 = 15, x_4 = 16, x_5 = 17, \\ f_1 = 2, f_2 = 5, f_3 = 13, f_4 = 7, f_5 = 3 \\ n = \sum f = f_1 + f_2 + f_3 + f_4 + f_5 = 2 + 5 + 13 + 7 + 3 = 30 \end{aligned}$$

Using the formula of geometric mean for grouped data, geometric mean in this case will become:

$$\begin{aligned} \text{G.M of} X &= \sqrt[n]{x_1^{f_1} \cdot x_2^{f_2} \cdots x_n^{f_n}} \\ &= \sqrt[30]{(13)^2 \cdot (14)^5 \cdot (15)^{13} \cdot (16)^7 \cdot (17)^3} \\ &= \sqrt[30]{2.33292 \times 10^{35}} = \left. 2.33292 \times 10^{35} \right)^{\frac{1}{30}} \\ &= 15.0984 \approx 15.10 \end{aligned}$$

The method explained above for the calculation of geometric mean is useful when the numbers of values in given data are small in number and the facility of electronic calculator is available. When a set of data contains large number of values then we need an alternative way for computing geometric mean. The modified or alternative way of computing geometric mean is given as under:

For Ungrouped data	For Ungrouped data
G.M of X = Antilog$\left(\frac{\Sigma \log x}{n}\right)$	G.M of X = Antilog $\left(\frac{\Sigma f \log x}{\Sigma f}\right)$

The logarithm of the Geometric mean is the weighted average of different values of $\log x$, whose weights are the frequencies.

$$\log G = \frac{1}{n} \sum f_i \log x_i = \frac{1}{n} \left[f_1 \log x_1 + f_2 \log x_2 + \cdots + f_n \log x_n \right]$$

Example 5.15. Find the Geometric Mean of the values 10, 5, 15, 8, 12

x	log$x$
10	1.0000
5	0.6990
15	1.1761
8	0.9031
12	1.0792
Total	$\Sigma \log x$ = 4.8573

G.M of X = Antilog$\left(\frac{\Sigma \log x}{n}\right)$

$$= \text{Antilog}\left(\frac{4.8573}{5}\right)$$

$$= \text{Antilog}(0.9715)$$

$$= 9.36$$

Example 5.16. Calculate the Geometric mean of the given data.

X	15	20	25	30	35	40	45	50
f	2	22	29	24	7	8	6	2

Solution. The Geometric mean is given by

$$G = \left[x_1^{f_1} \cdot x_2^{f_2} \cdots x_n^{f_n}\right]^{\frac{1}{n}} \qquad \text{where} \quad n = \Sigma f$$

$$\log G = \frac{1}{n}\sum f_i \log x_i$$

$$= \frac{1}{n}\left[f_1 \log x_1 + f_2 \log x_2 + \cdots + f_n \log x_n\right]$$

$$= \frac{1}{100}\left[2\log 15 + 22\log 20 + 29\log 25 + 24\log 30 + 7\log 35 + 8\log 40 + 6\log 45 + 2\log 50\right]$$

$$= \frac{1}{100}\left[143.903\right]$$

$$= 1.43908$$

$$G = 27.48$$

Example 5.17. Find the Geometric Mean for the following distribution of students marks:

Marks	0-39	40-49	50-59	60-69	70-100
No. of Students	8	12	13	11	8

The geometric mean must be used when working with percentages, which are derived from values, while the standard arithmetic mean works with the values themselves. The geometric mean of a data set is less than the data set's arithmetic mean unless all members of the data set are equal, in which case the geometric and arithmetic means are equal.

Question 5.8. Calculate the Geometric Mean of 2, 4, 8 and 16

Question 5.9. What is the Geometric Mean of 1, 3, 9, 27 and 81 ?

Question 5.10. Calculate the Geometric Mean of the first eight natural numbers.

5.4.1 Properties of Geometric mean

(i). Geometric mean gives more weight to logarithms

(ii). The geometric mean is not much affected by fluctuations of sampling

(iii). It is not easy to interpret

(iv). It does not lend itself to algebraic manipulations

(v). It cannot be calculated when there is a negative or a zero value in the data

(vi). It is useful in getting (calculating) average growth rate e.g population growth rate.

5.5 Harmonic Mean

Harmonic mean is another measure of central tendency and also based on the same concept like arithmetic mean and geometric mean. Like arithmetic mean and geometric mean, harmonic mean is also useful for quantitative data. Harmonic mean is defined in following terms:

> *Harmonic mean is quotient of "number of the given values" and "sum of the reciprocals of the given values".*

Harmonic mean in mathematical terms is defined as follows:

For Ungrouped data	For Ungrouped data
H.M of; $X = \frac{n}{\Sigma\left(\frac{1}{x}\right)}$	H.M of; $X = \frac{\Sigma f}{\Sigma\left(\frac{f}{x}\right)}$

The Harmonic mean, H, of n non zero different values of the variable values each occurring with frequencies $f_1, f_2, \cdots, f_n$ is given by the formula.

$$H = \frac{\Sigma f_i}{\Sigma\left(\frac{f_i}{x_i}\right)}$$

$$\frac{1}{H} = \frac{\Sigma\left(\frac{f_i}{x_i}\right)}{\Sigma f_i}$$

Thus, the harmonic mean of the variable is the reciprocal of the arithmetic mean of their reciprocals.

Example 5.18. Calculate the harmonic mean of the numbers: 13.5, 14.5, 14.8, 15.2 and 16.1

Example 5.19. Given the following frequency distribution of first year students of a particular course. Calculate the Harmonic Mean.

Age (Years)	13	14	15	16	17
Number of students	1	4	12	6	2

Solution. The given distribution belongs to a grouped data and the variable involved is age of first year students. While the number of students represent frequencies.

Ages (Years) x	Number of Students f	$\frac{f}{x}$
13	1	
14	4	
15	12	
16	6	

17	2	

Example 5.20. Calculate the harmonic mean for the given below:

Marks	30 – 39	40 – 49	50 – 59	60 – 69	70 – 79	80 – 89	90 – 99
f	3	4	12	21	33	26	8

Example 5.21. Calculate the Harmonic mean of the data in the example above.

x	$\frac{1}{x}$	frequency	$f \cdot \frac{1}{x}$
15	0.066	2	0.13
20	0.050	22	1.10
25	0.040	29	1.16
30	0.033	24	0.80
35	0.029	7	0.20
40	0.025	8	0.20
45	0.022	6	0.13
50	0.020	2	0.04
		$\Sigma f = 100$	$\Sigma f_i/x_i = 3.76$

The harmonic mean is given by

$$H = \frac{\Sigma f_i}{\Sigma\left(\frac{f_i}{x_i}\right)}$$
$$= \frac{100}{3.76}$$
$$= 26.60$$

5.5.1 Properties of the Harmonic Mean

(i). Gives much importance to small values of the data.

(ii). It is not easy to interpret.

(iii). It does not lend itself to algebraic manipulation.

(iv). Harmonic mean does not exist if any of the values of the data is zero.

(v). it is suitable when observations are dealing with rates i.e., speed in Km/h.

5.6 Combined Mean

Consider two sets of data; Data set A: 78, 66, 43, 56, 76, 26, 57, 42 and Data set B: 65, 52, 42, 53, 53, with sample sizes $n_1 = 8$ and $n_2 = 5$ respectively.

Let $\bar{x}_1$ and $\bar{x}_2$ be the arithmetic means of data set 1 and data set 2 respectively, then

$$\bar{x}_1 = \frac{78 + 66 + 43 + 56 + 76 + 26 + 57 + 42}{8} = \frac{444}{8}$$
$$= 55.5$$

$$\bar{x}_2 = \frac{65 + 52 + 42 + 53 + 53}{5} = \frac{265}{5}$$
$$= 53$$

Next, is the mean of the combined sets of data the average of the two means? Let us check;

Let the combined mean be denoted by $\bar{x}$, then

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$
$$= \frac{78 + 66 + 43 + 56 + 76 + 26 + 57 + 42 + 65 + 52 + 42 + 53 + 53}{13}$$
$$= \frac{709}{13}$$
$$= 54.54$$

The average of the two means is given by

$$\bar{x} = \frac{\bar{x}_1 + \bar{x}_2}{2} = \frac{55.5 + 53.0}{2}$$
$$= 54.25$$

The two quantities are clearly not equal, but

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} = \frac{8 \times 55.5 + 5 \times 53.0}{8 + 5} = \frac{709}{13}$$
$$= 54.54$$

Therefore the mean of combined sets of data is NOT $(\bar{x}_1 + \bar{x}_2)/2$ however if $n_1 = n_2$ the case will hold.

In general if there are k data sets with means $\bar{x}_1, \bar{x}_1, \cdots, \bar{x}_k$ then the combined mean $\bar{x}$ for all of the data sets is

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + \cdots + n_k\bar{x}_k}{n_1 + n_2 + \cdots + n_k}$$

where n_i denotes the number of values in the data set $i = 1, 2, \cdots, k$.

If there are two groups containing n_1 and n_2 observations with means $\bar{X}_1$ and $\bar{X}_2$ respectively, then the combined arithmetic mean of two groups is given by

$$\bar{X}_c = \frac{n_1\bar{X}_1 + n_2\bar{X}_2}{n_1 + n_2}$$

The above formula can be generalized for more than two groups. If $n_1, n_2, \cdots, n_k$ are sizes of k groups with means $\bar{X}_1, \bar{X}_2 \cdots, \bar{X}_k$ respectively then the mean $\bar{X}_c$ of the combined group is given by

$$\bar{X}_c = \frac{n_1\bar{X}_1 + n_2\bar{X}_2 + \cdots + n_k\bar{X}_k}{n_1 + n_2 + \cdots + n_k}$$

Example 5.22. If average salaries of two groups of employees are Ksh.1500 and Ksh. 2200 and there are 80 and 70 employees in the two groups respectively. Find the mean of the combined group.

Solution. Given;

Group I	Group II
$n_1 = 80$	$n_2 = 70$
$\bar{X}_1 = 1500$	$\bar{X}_2 200$

$$\bar{X}_c = \frac{n_1\bar{X}_1 + n_2\bar{X}_2}{n_1 + n_2} = \frac{80(1500) + 70(2200)}{80 + 70}$$
$$= \frac{274,000}{150} = 1,826.67$$

The average monthly salary of the combined group of 150 employees is Ksh.1826.67

Question 5.11. The mean weight of a group of 50 workers is 58 kgs. The second group consists of 60 workers with average weight 62 kgs. and there are 90 workers in the third group with average weight 56 kgs. Find the average weight of the combined group.

Question 5.12. The average daily wages for 120 workers in a factory are Ksh.78. The average wage for 80 male workers out of them is Ksh.92 Find the average wage for the remaining female workers.

Question 5.13. There are three groups in a class of 200 Students. The first groups contains 80 Students with average marks 65, the second group consists of 70 Students with average marks 74. Find the average marks of the Students from the third group if the average for the entire class is 68.

Question 5.14. Average daily wage of 60 male workers in a firm is 120 and that of 40 females is 100. Find the mean wage of all the workers.

Question 5.15. The average wage of 100 male workers is 80 and that 50 female workers is 75. Find the mean wage of workers in the company.

5.7 Mode

Mode is the value which occurs the greatest number of times in the data. When each value occur the same numbers of times in the data, there is no mode. If two or more values occur the same numbers of time, then there are two or more modes and distribution is said to be multi-mode. If the data having only one mode the distribution is said to be uni-model and data having two modes, the distribution is said to be bi-model.

The mode is the most frequently occurring value in a set of observation. For example, given 2, 3, 4, 5, 4 , the mode is 4, because there are more fours than any other number. Data may have two modes. In this case we say the data are *bimodal*, and observations with more than two modes are referred to as *multi-modal*. Note that the mode does not have important mathematical properties for future use. Also, the mode is not a helpful measure of location, because there can be more than one mode or even no mode.

5.7.1 Mode from Ungrouped Data

Mode is calculated from ungrouped data by inspecting the given data. We pick out that value which occur the greatest numbers of times in the data.

Example 5.23. Find the mode of the values

$$5, 8, 9, 11, 9, 12, 9$$

Solution. The most frequently occurring value is 9. Therefore the mode is 9. This is a unimodal distribution.

Note: The mode may not be unique.

Example 5.24. Find the mode of the numbers.

$$5, 8, 9, 11, 9, 12, 9, 13, 15, 14, 14, 14$$

Solution. The modes are 9 and 14 each occurring three times. This is a bimodal distribution. **Example 5.25.** Find the mode of the values

$$5, 8, 9, 11, 12, 13, 15, 14$$

Solution. There is no mode.

5.7.2 Mode from Grouped Data

Whenfrequencydistributionwithequalclassintervalsizes, theclasswhichhasmaximumfrequencyiscalled model class. In a grouped frequency distribution the mode is calculated using an interpolation formula which is derived as follows:

Using similar triangles AND and CNB

$$\frac{L_1}{L_2} = \frac{AD}{BC} = \frac{d_1}{d_2}$$

d_1- difference between frequency density of the modal class and the class preceding the modal class.
d_2- difference between the frequency density of the modal class and that following the modal class.

When the class widths (sizes) are constant say c, frequencies are proportional to frequency densities so

$$\frac{L_1}{L_2} = \frac{\Delta_1}{\Delta_2} = \frac{d_1}{d_2} \qquad (1)$$

where Δ_1 is the difference between the frequency of the modal class and the frequency of the class preceding the modal class.

Δ_2 is the difference between the frequency of the modal class and the frequency of the class following the modal class.

But also, $L_1 + L_2 = c$

$$L_1 + \frac{\Delta_2 L_1}{\Delta_1} = c$$
$$\Delta_1 L_1 + \Delta_2 L_2 = c\Delta_1$$
$$(\Delta_1 + \Delta_2)L_1 = c\Delta_1$$

$$L_1 = \frac{c\Delta_1}{\Delta_1 + \Delta_2}$$

The mode is given by

$$\begin{aligned} M_0 &= Lcb + L_1 \\ &= Lcb + \frac{\Delta_1 c}{\Delta_1 + \Delta_2} \\ \text{Mode} &= lcb + \frac{(f_m - f_1)c}{(f_m - f_1) + (f_m - f_2)} \\ &= Lcb + \frac{(f_{m_0} - f_1)c}{2f_0 - f_1 - f_2} \end{aligned}$$

where;

lcb−lower class boundary of the modal class

f_{m0}−raw frequency of the modal class f_1− frequency of the class preceding the modal class. f_2− frequency of the class following the modal class.

Using the same diagram the formula may be modified to cater for the case when the classes do not have constant width. In this case the mode is given by;

$$\begin{aligned} M_0 &= Lcb + L_1 \\ &= Lcb + \frac{d_1 \times c_{m_0}}{d_1 + d_2} \end{aligned}$$

where;

lcb−lower class boundary of the modal class

c_{m0}−Is the class size (width) of the modal class

d_1 and d_2 − as defined above.

Example 5.26. Calculate the mode for the following frequency distribution of marks obtained by 50 students in Statistics.

Marks	6 – 10	11 – 15	16 – 20	21 – 25	26 – 30	31 – 35	36 – 40	41 – 45	45 – 50
Frequency	5	6	15	10	5	4	2	2	1

Solution. The modal class is $16-20$. Thus, lcb = 15.5, $f_1 = 6$, $f_2 = 10$, $f_{m_0} = 15$ and $c = 5$. The mode is then given by

$$\begin{aligned} M_0 &= Lcb + \frac{(f_{m_0} - f_1) \times c}{2f_0 - f_1 - f_2} \\ &= 15.5 + \frac{(15-6)5}{30-6-10} \\ &= 15.5 + 3.214 \\ &= 18.71 \end{aligned}$$

Example 5.27. Calculate the mode for the following frequency distribution scores for 80 students.

Scores	Frequency	Class interval	Frequency density
$5-20$	8	16	0.50
$21-40$	12	20	0.60
$41-55$	18	15	1.20
$56-87$	40	32	1.25
$88-95$	2	8	0.25
	$\Sigma f = 80$		

The modal class is $56-87$, lcb = 55.5, $d_1 = (1.25-1.20) = 0.05$, $d_2 = (1.25-0.25) = 1.00$ and $c_{m_0} = 32$. The mode is given by;

$$\begin{aligned} M_0 &= Lcb + L_1 \\ &= Lcb + \frac{d_1 \times c_{m_0}}{d_1 + d_2} \\ &= 55.5 + \frac{0.05 \times 32}{0.05 + 1.00} \\ &= 55.5 + 1.52 \\ &= 57.02 \end{aligned}$$

5.8 Median

The median by definition refers to the middle value in the arrayed data. It means that when the data are arranged, median is the middle value if the number of values is odd and the mean of the two middle values, if the numbers of values is even. A value which divides the arrayed set of data in two equal parts is called median, the values greater than the median is equal to the values smaller than the median. The 50% observations lie below the value of the median and 50% observations lie above it. Median is called a positional average. It is denoted by $\tilde{X}$ read as X- tilde, and also sometimes denoted by M.

5.8.1 Median of Ungrouped data

There are two cases for calculating the median of ungrouped data. The number of observations in ungrouped data may be odd or even. The procedure for calculating the median of odd and even observations is different and is explained below.

The first step is to arrange the data in ascending or descending order of magnitude.

Number of observations (odd): If the observations n are odd in number, then:

$$\text{Median} = \text{value of} \left(\frac{n+1}{2}\right)^{th} \text{ term}$$

The median is the middle number if there is an odd number of observations.

Example 5.28. Find the median for the following set of observations

$$65, 38, 79, 85, 54, 47, 72$$

Solution. Arrange the values in ascending order:

$$38, 47, 54, 65, 72, 79, 85 ;$$

Here, n = 7 (odd number)

The middle observation is 65, therefore Median= 65 units.

Using the formula:

$$\text{Median} = \text{value of} \left(\frac{n+1}{2}\right)^{th} \text{ term}$$
$$= \text{value of} \left(\frac{7+1}{2}\right)^{th} \text{ term}$$
$$= 4^{th} \text{ term}$$
$$Median = 65 \text{ units}$$

Number of observations (Even): If the observations are Even in number, then median is the arithmetic mean of two central values. These two central values are computed as:

$$\text{First central value} = \frac{n}{2}$$
$$\text{Second central value} = \frac{n+2}{2}$$

Median = Arithmetic Mean of value of First central value and Second central value

i.e, adding the two middle values and divided by two, where n = number of observations. If there is an even number of data in the array, the median is the average of the two middle numbers.

Example 5.29. Find the median for the following set of data:

$$74, 66, 69, 68, 73, 70$$

Solution. First, we arrange the data in an ordered array:

$$66, 68, 69, 73, 70, 74$$

Since there is an even number of data, the average of the middle two numbers (i.e., 69 and 73) is the median $(142/2 = 71)$. Note that in general, location of the median is = $(n+1)/2$ where n = total number of items.

Generally, the median provides a better measure of location than the mean when there are some extremely large or small observations (i.e., when the data are skewed to the right or to the left).

Example 5.30. Find the median for the following data.

$$25, 98, 67, 18, 45, 83, 76, 35$$

Solution. Arrange the values in ascending order

$$18, 25, 35, 45, 67, 76, 83, 98$$

$n = 8$ (even number)

The pair $45, 67$ can be considered as the middle pair.

Example 5.31. Find the median for the following data sets.

(i) 3, 5, 2, 7, 8

(ii) 3, 5, 2, 7, 8, 11

Solution:

(i) First arrange the values in ascending order of

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

The median is 5

(ii) 2, 3, 5, 7, 8 , 11

$$\frac{5+7}{2} = \frac{12}{2} = 6$$

Note: The results of median will not be affected by arranging in ascending or descending order.

5.8.2 Median of Ungrouped data with frequencies

The median for ungrouped data with frequencies is obtained by finding the size of $\Sigma f_i/2$ th value. Here Σf_i is the sum of frequencies and can be even or odd.

Steps:

(1) Arrange the data in ascending or descending order of magnitude with respective frequencies.

(2) Find the cumulative frequency *(c.f)* less than type.

(3) Find $\Sigma f_i/2$, Σf_i = total frequency.

(4) Find the cumulative frequency *c.f* column that is either equal or greater than $\Sigma f_i/2$ and determine the value of the variable corresponding to it.

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

Example 5.32. Consider the following frequency distribution. Calculate the median.

X	1	2	3	4	5	6	7
f	2	10	15	20	25	18	10

Solution. First we compute the cumulative frequency of the data.

Observation X	Frequency f	Cumulative frequency (C.F)
1	2	2
2	10	12
3	15	27
4	20	47
5	25	72
6	18	90
7	10	100
Total	100	

$$\text{Median} = \text{size of } \left(\frac{n+1}{2}\right)^{\text{th}} \text{ term} = \frac{100+1}{2} = 50.0.$$

The median is 5 because 51st item lies corresponding to 5.

5.8.3 Median of Grouped Data

In case of a grouped frequency distribution median can be calculated with the help of either the interpolation formula or cumulative frequency distribution curve.

Location of Median using interpolation method

The median for grouped data, we find the cumulative frequencies and then calculated the median number $n/2$. The median lies in the group (class) which corresponds to the cumulative frequency in which $n/2$ lies. We use following formula to find the median.

$$\text{Median} = lcb + \frac{\left(\frac{n}{2} - cum_L\right)c}{f_m}$$

where

lcb = lower class boundary of the median class

n = Total of all the frequencies

f_m = Frequency of the median class

cum_L = Cumulative frequency up to the class preceding the median class.

c = Class size or width of the median class

Steps:

1. To locate the median class, divide the cumulative frequency n by 2, since the median is the $\frac{n}{2}$ the value of the variable when data is arranged in ascending order.
2. Up to the *lcb* of the interval containing the median we have cum_L items say.
3. If we assume that the values are evenly distributed in the interval containing the median, then to reach the median from the *lcb* we add

$$\frac{(\frac{n}{2} - cum_L)c}{f_{me}}$$

Example 5.33. Calculate median for the following data.

Group	20 – 24	25 – 29	30 – 34	35 – 39	40 – 44	45 – 49
Frequency	2	6	10	14	8	3

Solution. The class boundaries and cumulative frequency are given in the table below:

Group	Frequency f	Class boundary	Cumulative frequency
$20-24$	2	$19.5-24.5$	2
$25-29$	6	$24.5-29.5$	8
$30-34$	9	$29.5-34.5$	17
$35-39$	14	$34.5-39.5$	31
$40-44$	8	$39.5-44.5$	39
$45-49$	11	$44.5-49.5$	50

$$\begin{aligned}\text{Median} &= lcb + \frac{\left(\frac{n}{2} - cum_L\right)c}{f_m} \\ &= 34.5 + \frac{\left(\frac{50}{2} - 17\right)5}{14} = 34.5 + \frac{20}{7} \\ &= 38.257\end{aligned}$$

Graphical Location of Median

Median and other partition values can be located from the graph of the cumulative frequency curve (Ogive Polygon). For median, we calculate $n/2$. On Y -axis, we mark the height equal to $n/2$ and from this point we draw a straight line parallel to X-axis which intersects the polygon at the point m. from the point m, we draw a perpendicular which touches the X-axis at M. This point on X-axis is the median. Similarly, for the lower quartile we take height equal to $n/4$ on the Y -axis. From this we draw a line parallel to X-axis which the polygon at the point q. From this point we draw perpendicular on X-axis which touches it at the point Q_1 which is the first quartile. For upper quartile take the height on Y -axis equal to $3n/4$.

Now, that you have studied about all the three measures of central tendency, let us discuss which measure would be best suited for a particular requirement.

The mean is the most frequently used measure of central tendency because it takes into account all the observations, and lies between the extremes, i.e., the largest and the smallest observations of the entire data. It also enables us to compare two or more distributions. For example, by comparing the average (mean) results of students of different schools of a particular examination, we can conclude which school has a better performance.

However, extreme values in the data affect the mean. For example, the mean of classes having frequencies more or less the same is a good representative of the data. But, if one class has frequency, say 2, and the five others have frequency 20, 25, 20, 21, 18, then the mean will certainly not reflect the way the data behaves. So, in such cases, the mean is not a good representative of the data.

In problems where individual observations are not important, and we wish to find out a 'typical' observation, the median is more appropriate, e.g., finding the typical productivity rate of workers, average wage in a country, etc. These are situations where extreme values may be there. So, rather than the mean, we take the median as a better measure of central tendency.

In situations which require establishing the most frequent value or most popular item, the mode is the best choice, e.g., to find the most popular Television programme being watched, the consumer item in greatest demand, the colour of the vehicle used by most of the people, etc.

Remarks:

1. There is an empirical relationship between the three measures of central tendency:

$$3\text{Median} = \text{Mode} + 2\ \text{Mean}$$

2. The median of grouped data with unequal class sizes can also be calculated. However, it is not discussed here.

Question 5.16. If the median of the distribution given below is 28.5, find the values of x and y.

Class Interval	0 – 9	10 – 19	20 – 29	30 – 39	40 – 49	50 – 59
Frequency	5	x	20	15	y	5

6 Partition Values

If the values of the variate are arranged in ascending or descending order of magnitudes then we have seen above that median is that value of the variate which divides the total frequencies in two equal parts. Similarly the given series can be divided into four, ten and hundred equal parts. The values of the variate dividing into four equal parts are called **Quartile**, into ten equal parts are called **Decile** and into hundred equal parts are called **Percentile**.

6.1 Quartiles

The Quartiles are those values which divide the set of observations into four equal parts. There are three quartiles called, first quartile, second quartile and third quartile. The first quartile is also called lower quartile and is denoted by Q_1 is the value that lies between the smallest value and the median. The median is the second quartile Q_2. The third quartile is also called upper quartile and is denoted by Q_3 is the value that lie midway between the median and the largest value. The lower quartile Q_1 is a point which has 25 % observations less than it and 75% observations are above it. The upper quartile Q_3 is a point with 75 % observations below it and 25% observations above it.

6.1.1 Quartile for Individual Observations (Ungrouped Data)

$$Q_1 = \text{Value of } \left(\frac{n+1}{4}\right)^{\text{th}} \text{ item}$$

$$Q_2 = \text{Value of } 2\left(\frac{n+1}{4}\right)^{\text{th}} \text{ item} = \text{Value of } \left(\frac{n+1}{2}\right)^{\text{th}} \text{ item} = \textit{Median}$$

$$Q_3 = \text{Value of } 3\left(\frac{n+1}{4}\right)^{\text{th}} \text{ item}$$

Example 6.1. For the data we can arrange the values in ascending order and assign each value a rank (position) to get

Arranged values	10	11	12	13	14	17	18	19	23	25
Position	1	2	3	4	5	6	7	8	9	10

Solution. Then

$$\begin{aligned} Q_2 &= \frac{2}{4}(10+1)^{th}\ value = 5.5^{th}\ value \\ 5.5^{th}\ value &= \frac{1}{2}(5^{th}\ value + 6^{th}\ value) \\ &= 5^{th}\ value - \frac{1}{2}(5^{th}\ value) + \frac{1}{2}(6^{th}\ value) \\ &= 5^{th}\ value + \frac{1}{2}(6^{th}\ value - 5^{th}\ value) \\ &= 14 + 0.5(17-14) = 14 + 1.5 \\ &= 15.5 \end{aligned}$$

which is the second quartile or the median of the data set, this is the same value we got in example 3.1.1 above. The formula used here ie referred to as **linear interpolation** and works even in general cases to locate the actual value of the n- tile.

For example 5.7^{th} value

$$5.7^{th}\ value = 5^{th}\ value + 0.7(6^{th}\ value - 5^{th}\ value)$$

Generally, the interpolation formula for the $n.d^{th}$ value, (n, whole part and d, the decimal part) is;

$$n.d^{th}\ value = n^{th}\ value + 0.d \times [(n+1)^{th}\ value - n^{th}\ value]$$

Example 6.2. The wheat production (in Kg) of 20 acres is given as: 1120, 1240, 1320, 1040, 1080, 1200 , 1440, 1360, 1680, 1730, 1785, 1342, 1960, 1880, 1755, 1720, 1600, 1470, 1750, and 1885. Find the quartile deviation and coefficient of quartile deviation.

Solution. After arranging the observations in ascending order, we get 1040, 1080, 1120, 1200, 1240, 1320 , 1342, 1360, 1440, 1470, 1600, 1680, 1720, 1730, 1750, 1755, 1785, 1880, 1885, 1960.

$$Q_1 = \text{Value of } \left(\frac{n+1}{4}\right)^{\text{th}} \text{ item}$$
$$= \text{Value of } \left(\frac{20+1}{4}\right)^{\text{th}} \text{ item}$$
$$= \text{Value of } (5.25)^{\text{th}} \text{ item}$$
$$= 5th \text{ value} + 0.25\,(6th \text{ value} - 5th \text{ value}) = 1240 + 0.25(1320 - 1240)$$
$$Q_1 = 1240 + 20 = 1260$$

$$Q_3 = \text{Value of } \left(\frac{3(n+1)}{4}\right)^{\text{th}} \text{ item}$$
$$= \text{Value of } \left(\frac{3(20+1)}{4}\right)^{\text{th}} \text{ item}$$
$$= \text{Value of } (15.75)^{\text{th}} \text{ item}$$
$$= 15\text{th value} + 0.75\,(16\text{th value} - 15\text{th value}) = 1750 + 0.75(1755 - 1750)$$
$$Q_3 = 1750 + 3.75 = 1753.75$$

6.1.2 Quartile for a Frequency Distribution (Discrete Data)

$$Q_1 = \text{Value of } \left(\frac{n+1}{4}\right)^{\text{th}} \text{ item} \qquad \left(n = \sum f\right)$$
$$Q_2 = \text{Value of } 2\left(\frac{n+1}{4}\right)^{\text{th}} \text{ item} = \text{Value of } \left(\frac{n+1}{2}\right)^{\text{th}} \text{ item} = \textit{Median}$$
$$Q_3 = \text{Value of } 3\left(\frac{n+1}{4}\right)^{\text{th}} \text{ item}$$

Example 6.3. Calculate the quartile deviation and coefficient of quartile deviation from the data given below:

6.1.3 Quartile for Grouped Frequency Distribution

The three quartiles can be calculated for grouped data by the formula.

$$Q_i = lcb + \frac{\left(i \cdot \frac{n}{4} - cum_L\right) c}{f_{Q_i}} \quad \text{for } i = 1,2,3$$

where:

lcb = lower class boundary of the median class

N = Total of all the frequencies cum_L = Cumulative frequency up to the class preceding the median class.

f_{Qi} = Frequency of the i th quartile class

c = Class size or width of the ith class

When i = 2 we get;

$$\text{Median } = lcb + \frac{\left(\frac{n}{2} - cum_L\right) c}{f_m}$$

Where;

lcb = lower class boundary of the median class

n = Total of all the frequencies

f_m = Frequency of the median class

cum_L = Cumulative frequency up to the class preceding the median class.

c = Class size or width of the median class

Example 6.4. Calculate the three quartiles for the following frequency distribution of marks obtained by 50 students.

Marks	Frequency	Cumulative Frequency
$6-10$	5	5
$11-15$	6	11
$16-20$	15	26
$21-25$	10	36
$26-30$	5	41
$31-35$	4	45
$36-40$	2	47
$41-45$	2	49
$46-50$	1	50
	$\Sigma f = 50$	

$$\text{The} Q_i = lcb + \frac{\left(i \cdot \frac{n}{4} - cum_L\right) c}{f_{Q_i}} \qquad \text{for} \qquad i = 1, 2, 3$$

The first quartile Q_1 is the $\frac{50}{4})^{th} = 12.5^{th}$ observation from the smallest. This value is found in the $16-20$ class interval.

$$\begin{aligned} lcb = 15.5, \qquad f_{Q_1} = 15, \qquad cum_L = 11, \qquad c = 5 \\ &= 15.5 + \frac{(\frac{50}{4} - 11) \times 5}{15} \\ &= 15.5 + \frac{(12.5 - 11) \times 5}{15} \\ &= 16.0 \end{aligned}$$

The median is the $\frac{50}{2})^{th}$ or 25^{th} observation. This occurs in class $16-20$ so that

$$lcb = 15.5, \qquad f_{Q_2} = 15, \qquad cum_L = 11, \qquad c = 5$$

$$\begin{aligned} Q_2 &= 15.5 + \frac{(\frac{50}{4} - 11) \times 5}{15} \\ &= 15.5 + \frac{(25 - 11) \times 5}{15} \\ &= 15.5 + \frac{70}{15} \\ &= 20.17 \end{aligned}$$

The third quartile is $\frac{150}{4})^{th}$ or 37.5^{th} observation. This occurs in class $26-30$ so that

$$lcb = 25.5, \qquad f_{Q_3} = 5, \qquad cum_L = 36, \qquad c = 5$$

$$\begin{aligned} Q_3 &= 25.5 + \frac{(\frac{150}{4} - 36) \times 5}{15} \\ &= 25.5 + \frac{(25 - 11) \times 5}{15} \\ &= 15.5 + (37.5 - 36) \\ &= 27.0 \end{aligned}$$

6.2 Deciles

The deciles are the partition values which divides the set of observations into ten equal parts. There are nine deciles namely $D_1, D_2, D_3, \cdots, D_9$. The first decile is D_1 is a point which has 10% of the observations below it.

6.2.1 Deciles for Individual Observations (Ungrouped Data)

D_1 = Value of $\left(\frac{n+1}{10}\right)^{th}$ item $\left(n = \sum f\right)$

D_2 = Value of $2\left(\frac{n+1}{10}\right)^{th}$ item

...

D_9 = Value of $9\left(\frac{n+1}{10}\right)^{th}$ item

Example 6.5.

6.2.2 Decile for a Frequency Distribution (Discrete Data):

D_1 = Value of $\left(\frac{n+1}{10}\right)^{th}$ item

D_2 = Value of $2\left(\frac{n+1}{10}\right)^{th}$ item

...

$$D_9 = \text{Value of } 9\left(\frac{n+1}{10}\right)^{th} \text{ item}$$

6.2.3 Decile for Grouped Frequency Distribution

$$\text{The} D_i = lcb + \frac{c}{f_{D_i}}\left(\frac{in}{10} - cum_L\right) \quad \text{for} \quad i = 1,2,3,\cdots,9$$

6.3 Percentiles

The percentiles are the points which divide the set of observations into one hundred equal parts. These points are denoted by $P_1, P_2, P_3, \cdots, P_{99}$, and are called the first, second, third, ⋯, ninety ninth percentiles. The percentiles are calculated for very large number of observations like workers in factories and the population in counties or countries. The percentiles are usually calculated for grouped data. The first percentile denoted by P_1 is calculated as

$$P_1 = \text{Value of}\left(\frac{n+1}{100}\right)^{th} \text{ item}$$

We find the group in which the $\frac{n+1}{100})^{th}$ item lies and then P_1 is interpolated from the formula.

$$\text{The} P_i = lcb + \frac{c}{f_{P_i}}\left(\frac{in}{100} - cum_L\right) \quad \text{for} \quad i = 1,2,3,\cdots,99$$

6.4 Estimation of Measures of Location from Ogive Curves

The measures of location such as quartiles, deciles and percentile can also be estimated from accurately drawn cumulative frequency (Ogive) curves. The following example illustrates this.

Example 6.6. A factory producing rechargeable batteries might be interested in what percent of their batteries last up to 30 hours, what percent last more than 35 ours, etc. To obtain this information the data from the experiment on the time batteries lasted is to be represented in a cumulative frequency table and curve.

6.5 Measures of Location from Grouped Data

When observation have been grouped into classes, the problem of estimating a particular observation is slightly complicated because real values of the individual observations are not known. The class in which the observation lies can easily be identified using the cumulative frequency column. To determine the k^{th} observation when classes are arranged in ascending order, we first determine the class where it falls and assume that the observations are distributed uniformly or equally distributed throughout the class. After this, identify the corresponding;

1. Lower class boundary LCB of the class containing the value
2. Cumulative frequency of the previous class p_{cf}
3. The frequency f of the particular class
4. The class interval i of particular class

Then the k^{th} value which may be the $N-$tile is given by

$$k^{th}\ value = LCB + \frac{k - p_{cf}}{f} \times i$$

For example, consider the frequency table below

Class-interval	8 - 12	13 - 17	18 - 22	23 - 27	28 - 32	33 - 37
Frequency (*f*)	3	10	12	9	5	1
cf	3	13	25	34	39	40

15^{th} value lies in class $18-22$ because the previous cumulative frequency is $13 < 15$ and the following one is $34 > 15$. Then, $LCB = 17.5$, $k = 15$, $p_{cf} = 13$, $f = 12$ and $i = 22.5 - 17.5 = 5$.

$$15^{th}\text{ value } = 17.5 + \frac{15 - 13}{12} \times 5 = 17.5 + 0.83 = 18.33$$

Earlier, we found that the median is located at position

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

for ungrouped data. But with grouped data we normally drop the 1 in $(n + 1)$ to get

$$\frac{n}{2}$$

as the position of the median value. Here the median $= \frac{40}{2} = 20^{th}$ value.

$$20^{th}\ value\ = 17.5 + \frac{20 - 13}{12} \times 5 = 17.5 + 2.92 = 20.42$$

What is the 85th percentile?

$$P_{85} = 85\ \frac{n^{th}}{100}\ value\ = 34^{th}\ value = 22.5 + \frac{34 - 25}{9} \times 5 = 22.5 + 5 = 27.5$$

Example 6.7. The table below shows the frequency of weekly withdrawals of money from a certain Bank. Use the data to answer the following questions that follows.

Amount Withdrawn	**Frequency**
1,000 - 4,999	10
5,000 - 8,999	14
9,000 - 12,999	20
13,000 - 16,999	16
17,000 - 20,999	12
21,000 - 24,999	8

(a) Using the coding method calculate the mean and

(b) The standard deviation

(c) The position of the withdrawal sheet reading Ksh10,000.00 if the sheets are arranged in ascending order.

Solution. To choose the assumed mean, we note that the two middle classes are 9,000 - 12,999 and 13,000 - 16,999 with frequencies 20 and 16 respectively. We choose assumed mean $A = 10{,}999.5$ as the

midpoint of the one with a higher frequency. The class interval $c = 12{,}999.5 - 8.999.5 = 4{,}000$ which is a good choice of scaling constant, then

Class-interval	x	f	$x - A$	$d = \frac{(x-A)}{c}$	d^2	fd	fd^2	cf
1,000 - 4,999	2,999.5	10	-8,000	-2	4	-20	40	10
5,000 - 8,999	6,999.5	14	-4,000	-1	1	-14	14	24
9,000 - 12,999	10,999.5	20	0	0	0	0	0	44
13,000 - 16,999	14,999.5	16	4,000	1	1	16	16	60
17,000 - 20,999	18,999.5	12	8,000	2	4	24	48	72
21,000 - 24,999	22,999.5	8	12,000	3	9	24	72	80
		80				30	190	

(a). The mean of the amount withdrawn

$$\bar{x} = A + c\frac{\sum fd}{\sum f} = 10,999.5 + 4,000\,\frac{30}{80} = 12,499.5$$

(b). The standard deviation

$$s = c\sqrt{\frac{\sum fd^2}{\sum f} - \left(\frac{\sum fd}{\sum f}\right)^2} = 4,000\sqrt{\frac{190}{80} - \left(\frac{30}{80}\right)^2} = 5,979.13$$

(c). The position of a withdrawal sheet which reads Ksh10,000.

k^{th} value = $10{,}000 = 8{,}999.5 + 4{,}000(k - 24)/20 \Rightarrow 2{,}001 = 400k - 9{,}600$

$$\Rightarrow 11{,}601 = 400k, \Rightarrow k = 29$$

The sheet is approximately in the 29th position.

6.6 Boxplots

The summary information contained in the quartiles is highlighted in a graphic display called a **boxplot**. The center half of the data, extending from the first to the third quartile, is represented by a rectangle. The median is identified by a bar within this box. A line extends from the third quartile to the maximum and another line extends from the first quartile to the minimum. (For large data sets the lines may only extend to the 95th and 5th percentiles).

Note: Percentiles can be computed using the formula

$$p_i = Lcb + \frac{(\frac{in}{100} - cum_L) \times c}{f_{p_i}}$$

6.7 Properties of Measures of Central Tendency

From the Example .. we note that if m is a known measure of location such as the mean, the mode or any measures of location for a given set of data $x_1, x_2, \cdots, x_n$ then,

1. A constant value (k say) added or subtracted from the each of the values in the data set translates the measures by the same constant. That is, New m= Old m + k. For the data set 2, 3, 4, 5, and 6 with mean $\bar{x} = 4$. If we add $k = 20$ to every value to have 22, 23, 24, 25, and 26 then the mean $\bar{x} = 24$.

2. If each value in the set is multiplied with a constant k, the new measure of location is given by $k \times m$. For the data is 2, 3, 4, 5, and 6 each of the values is multiplied or divided by $k = 2$, we have for multiplication the new median $m = 2 \times 4 = 8$ and for division as $m = \frac{4}{2} = 2$.

Lecture 6: Measures of Dispersion

7 Measures of Dispersion

7.1 Introduction

The measures of central tendency help to locate the center of the distribution, but they do not reveal how the observations are spread out on either side of the center. Although the measures of central tendency provide useful information about the data, they depend largely on the extent to which the data are dispersed. The degree to which numerical data tend to spread about the average value is called the variation or dispersion of the data. The data sets may have common means, medians and modes and identical frequencies in the modal class, yet they may differ widely in their spread of values about the measures of central tendencies. Consider the four data sets and their measures of central tendency.

Data Sets	A	B	C	D
	8	8	8	4
	8	8	6	12
	8	6	7	8
	8	10	9	16
	8	8	10	0
Mean	8	8	8	8
Mode	8	8		
Median	8	8	8	8

From the given measures, it is not possible to differentiate the four sets in the absence of the raw data. Other measures are required to make the comparison. The measures of **dispersion** or **spread** which is the degree of scatter or variation of the variable about the central value can be considered. There are various measures of dispersion which include; Range, Inter-Quartile Range, Quartile deviations, Mean Absolute Deviation, Variance and Standard Deviation.

7.2 Range

The range is the simplest of all the measure of dispersion and is defined as the difference between the largest and smallest value of a given data set. Range is denoted by R and is given by:

$$\text{Range } (R) = x_{\max} - x_{\min}$$

where $x_{\max}$ is the maximum value and $x_{\min}$ is the smallest observation in the sample (data set).

If this new measure is included in the example above, we get

Data Set	A	B	C	D
Mean	8	8	8	8
Mode	8	8		
Median	8	8	8	8
Range	0	4	4	16

It is still not possible to differentiate between the data sets. The range is easy to calculate but as seen in this example, it is unsatisfactory as it involves only two values, regardless of what the other values are.

The major disadvantage of the range is that it does not include all of the observations. Only the two most extreme values are included and these two numbers may be untypical observations. For example, given that the ages for a sample of 8 students at CSC are: 24, 18, 22, 19, 25, 20, 23, and 21, the range for this data set is:

$$25 - 18 = 7.$$

Note: In case of grouped frequency distribution, range will be the difference between UCB of the highest class and LCB of the lowest class.

7.3 Inter-Quartile Range (IQR)

The Inter-Quartile Range (IQR) is defined as the difference between the upper and the lower quartile. It is given by

$$IQR = Q_3 - Q_1$$

It can also be defined in terms of the deciles and percentiles as follows:

$$IQR = D_{7.5} - D_{2.5} = P_{75} - P_{25}$$

where D and P are as defined above.

The IQR is more stable than the range as it makes use of 50% of the data. It is not affected by extreme values (abnormally small or abnormally large) since extreme values are already removed. Since it ignores the other 50% of the data, it cannot be regarded as a reliable measure.

7.4 Quartile Deviation

The quartile deviation also called semi-interquartile range (SIQR) or semi-quartile range or interquartile range is defined as the average of the IQR.

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

i.e., the difference between the third and first quartiles divided by 2.

7.5 Mean Absolute Deviation (MAD)

The word "deviation" refers to the variations of each observation from the mean and "absolute deviation" means the positive numerical value of the deviation (negative sign ignored). MAD is the average of the absolute deviations from the mean. Let $x_1, x_2, \cdots, x_n$ be n given observations then the MAD about the mean is given by

$$MAD = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}| \qquad \text{for } i = 1,2,\cdots,n$$

For ungrouped data $x_1, x_2, \cdots, x_n$ with corresponding frequencies $f_1, f_2, \cdots, f_n$, the mean absolute deviation is given by;

$$\text{MAD} = \frac{\sum_{i=1}^{n} f_i|x_i - \bar{x}|}{\sum f_i},$$

For grouped data set, mean absolute deviation is as above but x_i's are mid-points of the classes.

Example 7.1. Consider the following data set:

$$3,6,\ 9,\ 3,\ 10,\ 7,\ 12,\ 1,\ 13,\ 15,\ 6,\ 5$$

Find (a) mean, (b) IQR (c) Quartile deviation and (d) mean absolute deviation

Solution. Let

(a). Mean

$$\bar{x} = \frac{\sum x_i}{n} = \frac{3+6+9+\cdots+6+5}{12} = 7.5$$

(b). Arranging the observations in ascending order:

$$1,\ 3,\ \underline{3},\ 5,\ 6,\ 6,\ 7,\ 9,\ \underline{\underline{10}},\ 12\ ,13,\ 15$$

$$\begin{aligned}
IQR &= Q_3 - Q_1 \\
Q_1 &= \left(\frac{13}{4}\right)^{th} value \\
&= 3.25^{th}\, value = 3^{rd} + 0.25(4^{th} - 3^{rd}) \\
Q_1 &= 3 + 0.25(5-3) = 3 + 0.5 = 3.5
\end{aligned}$$

$$\begin{aligned}
Q_3 &= 3(3.25^{th}) = 9.75^{th}\, value \\
&= 9^{th} + 0.75(10^{th} - 9^{th}) \\
&= 100.75(12-10) \\
&= 10 + 1.5 \\
&= 11.5 \\
IQR &= 11.5 - 3.5 = 8
\end{aligned}$$

(c). The Quartile deviation is given by

$$\text{Quartile Deviation } = \frac{1}{2}(IQR)$$

(d).

$$MAD = \frac{|3-7.5| + |6-7.5| + \cdots + |5-7.5|}{12}$$
$$= \frac{43}{12} = 3.5833 \approx 3.58$$

7.6 Variance and Standard Deviation

The range and IQR only involves only two values while ignoring the negative sign in order to compute MAD is not the only option. We know that the square of a number is always positive. We can therefore attempt to use the average of squared deviation from the mean denoted by $s62$ and define it as

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

where $x_1, x_2, \cdots, x_n$ are the set of n observations. This is referred to as the **variance** of the data set some books denoted it by σ^2.

Note: The variance formula given above is a good estimate (unbiased) of the population variance only if the sample is large ($n \geq 30$). If this is not the case the formula because;

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

Throughout these notes, n is used instead of $n - 1$ for the variance calculation (Assume that sample is a large sample or the variance is not for inference).

Example 7.2. Consider the set of values: 3, 8, and 4 whose mean is 5. Now, $(3-5)+(8-5)+(4-5) = -2 + 3 - 1 = 0$. In other words, $\Sigma\,(x_i - \bar{x}) = 0$. This is generally the case for any data set.

For the three values

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$
$$= \frac{(3-5)^2 + (8-5)^2 + (4-5)^2}{3}$$
$$= \frac{4+9+1}{3}$$
$$= \frac{14}{3} = 4.67$$

Our interest was to get average deviation from the mean and squaring was only meant to remove the negative sign in a reasonable way. To reverse the squaring we find the square root of the variance to get standard deviation denoted by s. This approach of computing variance is tedious because the mean has

to be calculated first. it becomes even more tedious if the mean is a decimal number. An easier computational formula is the expanded form which we quote directly as

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

7.6.1 Calculations from a Frequency Distribution

Let $x_1, x_2, \cdots, x_n$ be a set of n observations with corresponding frequencies $f_1, f_2, \cdots, f_n$ then, the mean is given by

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

Note: If the data is grouped, obtain x from the midpoints as

$$x = \frac{LCL + UCL}{2}$$

. For such data MAD is given by

$$MAD = \frac{\sum f_i |x_i - \bar{x}|}{\sum f_i}$$

and the variance by,

$$s^2 = \frac{\sum f_i (x_i - \bar{x})^2}{\sum f_i} = \frac{\sum f_i x_i^2}{\sum f_i} - \left(\frac{\sum f_i x_i}{\sum f_i}\right)^2$$

Note: For grouped frequency distribution, the use of x from the midpoint values instead of actual observations leads to just estimates (not accurate answers).

7.7 Assumed Mean and Coding Method

If the observations are too large such that the manual computation of totals is tedious, we may take any number among the observations (preferably the most central one) and use it as a **working mean** or **assumed mean**. This assumed mean is denoted by A. The deviations of the observations from the assumed mean (A), denoted by $d = x - A$, are used in place of x for all the observations. When this is done, the following formulas are used; For the mean

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

and for the variance

$$s^2 = \frac{\sum fd^2}{\sum f} - \left(\frac{\sum fd}{\sum f}\right)^2$$

Example 7.3. The masses (x) in kilograms of 30 bridging students who arrived the first day was recorded as they reported for the course and the following results calculated $\Sigma x = 1530$ and $\Sigma x^2 = 80{,}604$.

(a). Find the mean and the standard deviation

(b). Find Find the mean and the standard deviation if afterwards the weighing machine was discovered to be under weighing them by 2 kg.

(c). On the second day two students weighing 48 kg and 56 kg were absent. find the mean and the standard deviation of weight of those who were present.

Solution. (a). The mean and standard deviation

$$\bar{x} = \frac{\sum x}{n} = \frac{1530}{30} = 51$$

$$\begin{aligned} s &= \sqrt{\frac{\sum x_i^2}{n} - \left(\frac{\sum x_i}{n}\right)^2} \\ &= \sqrt{\frac{80604}{30} - \left(\frac{1530}{30}\right)^2} \\ &= \sqrt{85.5} = 9.26 \end{aligned}$$

(b). The new mean = 51 + 2 = 53 kg and new standard deviation s = 9.26 (remains the same).

(c). New sum

$$\sum_{i=1}^{n} x_i = 1530 - (48 + 56) = 1426$$

New sum of squares

$$\sum_{i=1}^{n} x^2 = 80,604 - (48^2 + 56^2) = 75,164$$

New sample size

$$n = 30 - 2 = 28$$

Mean

$$\frac{1426}{28} = 50.93$$

New Standard deviation

$$s = \sqrt{\frac{75164}{28} - \left(\frac{1426}{28}\right)^2} = \sqrt{90.71} = 9.52$$

If after subtracting assumed mean A from each of the observations, the values are still too large, one could divide the deviations with a constant value c so that $d = (x - A)/c$. The value of c is usually the class interval for grouped data. This gives rise to the following computational formulas. This method is referred

$$\bar{x} = A + c\left(\frac{\sum fd}{\sum f}\right)$$

to as **coding method**. and for the variance

$$s^2 = c^2\left[\frac{\sum fd^2}{\sum f} - \left(\frac{\sum fd}{\sum f}\right)^2\right]$$

Example 7.4. The masses in grams of some fruits are given in the table below

363.7	346.4	377.4	341.7	359.8	361.2	385.7	363.5	354.2	375.3
372.2	364.3	373.3	379.4	351.4	369.5	385.5	365.5	385.5	369.5

(a). Starting with 340 group the data into classes of interval 10. From the obtained frequency table calculate

(b). Using a suitable assumed mean, calculate the mean and the standard deviation.

(c). Using coding method, calculate the mean and the standard deviation.

Solution. Since the middle class is $360 - 369$, we let the assumed mean (A) be its midpoints = 364.5.

Class	Tally	freq f	midpoints (x)	$d = x - A$	fd	fd^2
340 - 349	II	2	344.5	-20	-40	800
350 - 359	III	3	354.5	-10	-30	300
360 - 369		7	364.5	0	0	0
370 - 379		5	374.5	10	50	500
380 - 389	III	3	384.5	20	60	1200
Totals		20			40	2800

$$\sum f = 20, \sum fd = 40, \sum fd^2 = 2800$$

$$\begin{aligned} \text{Mean, } \bar{x} &= A + \frac{\sum fd}{\sum f} = 364.5 + \frac{40}{20} \\ &= 364.5 + 2 \\ &= 366.5 \end{aligned}$$

$$\begin{aligned} \text{Standard deviation, } s &= \sqrt{\frac{\sum fd^2}{\sum f} - \left(\frac{\sum fd}{\sum f}\right)^2} \\ &= \sqrt{\frac{2800}{20} - 2^2} \\ &= \sqrt{136} = 11.66 \end{aligned}$$

7.8 Standard Deviation

Sometimes called the **root mean squared deviation** of a set of n numbers $x_1, x_2, \cdots, x_n$ denoted by s and given by the formula

$$s = \sqrt{\frac{\Sigma_{i=1}^{n}(x_i - \bar{x})^2}{n}}, \text{ where } \bar{x} = \frac{1}{n}\Sigma_{i=1}^{n} x_i$$

If the values $x_1, x_2, \cdots, x_n$ occur with frequencies $f_1, f_2, \cdots f_n$ respectively, the formula becomes;

$$s = \sqrt{\frac{\Sigma_{i=1}^{n} f_i(x_i - \bar{x})^2}{\Sigma f_i}}, \qquad \text{where } \bar{x} = \frac{\Sigma f_i x_i}{\Sigma f_i}$$

Example 7.5. A quality control supervisor has taken a sample of 16 bolts from the output of a thread-cutting machine and tested their tensile strengths. The results, in tons of force required for breakage, are s follows:

2.20	**1.95**	**2.15**	**2.08**	**1.85**	**1.92**
2.23	**2.19**	**1.98**	**2.07**	**2.24**	**2.31**
1.96	**2.30**	**2.27**	**1.89**		

(a) Determine the mean, median, range, and SIQR.

(b) Calculate the mean absolute deviation.

(c) Calculate the standard deviation and variance.

If the grouped frequency distribution has equal widths then the change of origin and scale (the so called Assumed mean or Coding method) can simplify the calculations for the standard deviation to

Example 7.6. Calculate the standard deviation for the wages given in example 1 above;

7.9 Variance

It is defined to be the square of the standard deviation and is thus denoted by s^2. For computational purposes the two formulae for standard deviation are given as;

$$S = \sqrt{\frac{\Sigma x_i^2}{n} - \left(\frac{\Sigma x_i}{n}\right)^2}$$ for un-grouped data

$$S = \sqrt{\frac{\Sigma f_i x_i^2}{\Sigma f_i} - \left(\frac{\Sigma f_i x_i}{\Sigma f_i}\right)^2}$$ for grouped data

Example 7.7. Calculate the standard deviation of the wages in example 1 above;

x_i	x_i^2	f_i	$f_i x_i$	$f_i x_i^2$
15	225	2	30	450
20	400	22	440	8800
25	625	19	475	11,875
30	900	14	420	12,600
35	1225	3	105	3,675
40	1600	4	160	6,400
45	2025	6	270	12,150
50	2500	1	50	2,500
55	3025	1	55	3,025

$\Sigma f_i = 72$, $\Sigma f_i x_i = 2005$, $\Sigma f_i x^2_i = 61{,}475$

Hence the standard deviation;

$$= \sqrt{\frac{\Sigma f_i x_i^2}{\Sigma f_i} - \left(\frac{\Sigma f_i x_i}{\Sigma f_i}\right)^2}$$
$$= \sqrt{\frac{61,475}{72} - \left(\frac{2005}{72}\right)^2}$$
$$= \sqrt{853.85 - 775.47}$$
$$= 8.85$$

7.10 Properties of Measures of Dispersion

Consider the set of values; 103, 108, and 104. The mean is 105. The variance of these values is

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$
$$= \frac{(103-105)^2 + (108-105)^2 + (104-105)^2}{3}$$
$$= \frac{4+9+1}{3} = 14/3 = 4.67$$

This is the same answer we got earlier with 3, 8, and 4. Note that each of these values have been modified by adding 100.

Generally the dispersion of observations remains the same if a constant is added to each of the data items.

7.11 Combined Variance

If there are two sets of data consisting of n_1 and n_2 observations with s_1^2 and s_2^2 as their respective variances, then the variance of the combined set consisting of $n_1 + n_2$ observations is

$$S^2 = \left[\frac{n_1(s_1^2 + d_1^2) + n_2(s_2^2 + d_2^2)}{n_1 + n_2}\right]$$

where d_1 and d_2 are the differences of the means $\bar{x}_1$ and $\bar{x}_2$, from the combined mean $\bar{x}$ respectively.

Example 7.8. Find the combined standard deviation of two series A and B.

	Series A	Series B
Mean	50	40
Standard deviation	5	6
No. of items	100	150

Solution. Given $\bar{x}_1 = 50$ and $\bar{x}_2 = 40$, $s_1^2 = 25$ and $s_2^2 = 36$, $n_1 = 36$ and $n_2 = 150$.

$$\text{Combined mean } \bar{x} = \frac{100 \times 50 + 150 \times 40}{100 + 150} = 44.$$

$d_1 = \bar{x}_1 - \hat{x} = 50 - 44 = 6$ and $d_2 = \bar{x}_2 - \bar{x} = 40 - 44 = -4$.

$$\text{Combined variance } = \frac{100(25+36) + 150(36+16)}{100+150}$$
$$= 55.6$$

$$\text{Therefore, combined } SD = \sqrt{55.6} = 7.46$$

7.12 Relative measures of Dispersion

The measure of dispersion namely range, quartile deviations, inter-quartile deviation, Mean deviation, standard deviation, root mean square deviation (these have been discussed above) are said to be absolute measure of dispersion, since they are expressed in terms of units of observations (Cm., Km., Kg., Kes. etc.). We know that different units can not be compared; for example a centimeter can not be compared with kilograms. Therefore, the dispersion in different units can not be compared.

The absolute measures of dispersion discussed above do not facilitate comparison of two or more data sets in terms of their variability. If the units of measurement of two or more sets of data are same, comparison between such sets of data is possible directly in terms of absolute measures. But conditions of direct comparison are not met, the desired comparison can be made in terms of the *relative measures.*

Also the measures of dispersion depend on the measures of central tendency. Therefore, it is needed to define some measures which are independent of the units of measurement and can be adjusted for measures of central tendency. Such type of measures are called **relation measures of dispersion** or **coefficients of dispersion**. These relative measures are pure numbers and are usually expressed as percentages. They are useful to compare two series in different units and also to compare variations of two series having different magnitudes.

Some of the relative measures of dispersion (or coefficient of dispersion) which are commonly used include; Quartile coefficient of dispersion, Coefficient of mean dispersion and Coefficient of variation or coefficient of dispersion.

7.12.1 Coefficient of range

7.12.2 Quartile coefficient of deviation

Another relative measure in terms of quartile deviations is **Coefficient of quartile deviation** and is defined as

$$Q_r = \frac{Q_3 - Q_1}{Q_3 + Q_1} \times 100$$

7.12.3 Coefficient of mean deviation

7.12.4 Coefficient of Variation (C.V.)

For distributions having the same mean, the distribution with the largest standard deviation has the greatest variation. But when considering distributions with different means, decision makers can't compare the uncertainty in distribution only by comparing standard deviations. In this case, the coefficient of variation is used, i.e., the coefficients of variation for different distributions are compared, and the distribution with the largest coefficient of variation value has the greatest relative variation.

The *coefficient of variation* is the ratio between the *standard deviation* of a sample and its *mean*, i.e., it reflects the variation in a distribution relative to the mean:

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

The coefficient of variation is usually expressed in percentages:

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

It allows us to compare the dispersions of two different distributions if their means are positive. The coefficientofvariationforadistributioncanbecalculatedtocomparethevaluesobtainedwithanotherdistributio n. The greater dispersion corresponds to the value of the coefficient of greater variation.
For example, Mark teaches two sections of statistics. He gives each section a different test covering the same material. The mean score on the test for the day section is 27, with a standard deviation of 3.4. The mean score for the night section is 74 with a standard deviation of 8.0. Which section has the greatest variation or dispersion of scores?

Direct comparison of the two standard deviations shows that the night section has the greatest variation. But comparing the coefficient of variations show quite different results:

C.V.(day) = $(3.4/27) \times 100 = 12.6\%$ and C.V.(night) $= (\frac{8}{94}) \times 100 = 8.5\ \%$

Thus, based on the size of the coefficient of variation, Mark finds that the night section test results have a smaller variation relative to its mean than do the day section test results.

Example 7.9. A distribution is $\bar{x} = 140$ and $\sigma = 28.28$ and the other is $\bar{x} = 150$ and $\sigma = 24$. Which of the two has a greater dispersion?

$$C.V_1 = \frac{28.28}{140} \times 100 = 20.2\ \%$$
$$C.V_2 = \frac{24}{150} \times 100 = 16\ \%$$

The first distribution has a higher dispersion.

Question 7.1. If the mean of the values 14, y, 17, 16, and y is $y + 0.4$, find the value y.

Question 7.2. Find the (a). mean, (b). median, (c). 3rd Quartile, (d). 8th Decile (e). 85th Percentile, and (f). mode of the following data below.

11 12 3 14 7 18 9 23 5 18 4 10

(a) Add 10 to each value of the given data and repeat (a). to (f).

(b) Subtract 5 from each of the given values and repeat (a). to (f).

Question 7.3. For 3 sample data sets $n_1 = 10$, $\bar{x}_1 = 15.4$, $n_2 = 15$, $\bar{x}_2 = 6.2$ and $n_3 = 12$, $\bar{x}_3 = 3.8$. Find the combined mean.

Question 7.4. Find the variance and standard deviation of the following values

15 23 19 16 18 23 14 12

(a) Complete the column for the mid-interval values and the other columns for the table below.

Class	Freq f	Midpoints (x)	x^2	fx	fx^2
1 - 10	2	5.5	30.25	11	60.5
11 - 20	14				
21 - 30	22				
31 - 40	26				
41 - 50	16				
	$\sum f$			$\sum fx$	$\sum fx^2$

(b) Calculate an estimate of the mean mark and the standard deviation.

(c) Explain why the mean and standard deviation are estimates and not the exact value.

Question 7.5. The data in the table below represents the weight (in kg) of 50 sacks of potatoes leaving a farm shop.

10.4	11.2	9.3	11.3	10.0	9.9	8.7	9.2	10.6	10.7
10.0	10.5	9.6	10.8	11.3	10.2	9.4	11.6	8.8	10.6
9.3	8.5	10.3	8.9	11.0	10.6	10.9	9.6	10.1	12.8
11.3	10.4	10.0	9.7	10.2	10.0	9.5	10.3	10.6	10.0
9.6	8.2	11.5	9.5	10.6	8.1	9.9	10.4	9.7	10.2

(a) Organize the data into a grouped frequency distribution starting with class

(b) Estimate the sample mean from the grouped frequency table.

(c) Calculate the median of the data

(d) Find the modal class and the mode of the data.

(e) Calculate the sample standard deviation.

Lecture 7: Measures of Skewness and Kurtosis

8 Measures of Skewness and Kurtosis

8.1 Introduction

Measure of central tendency gives us an idea about the average of the given set of observations, while a measure of dispersion gives the idea on how the observations are scattered about a central value or among themselves. The frequency distributions differ in three ways: Average value, Variability or dispersion, and Shape. Since the first two, that is, average value and variability or dispersion have already been discussed in previous lectures, here our main spotlight will be on the shape of frequency distribution. Generally, there are two comparable characteristics called **skewness** and **kurtosis** that help us to understand a distribution.

8.2 Skewness: Meaning and Definition

Two distributions may have the same mean and standard deviation but may differ widely in their overall appearance as can be seen from the following:

In both these distributions the value of mean and standard deviation is the same ($\bar{X} = 15, \sigma = 5$). But it does not imply that the distributions are alike in nature. The distribution on the left-hand side is symmetrical one whereas the distribution on the right-hand side is asymmetrical or skewed. Measures of skewness help us to distinguish between different types of distributions.

Measures of skewness tell us the direction and the extent of skewness. In symmetrical distribution the mean, median and mode are identical. The more the mean moves away from the mode, the larger the asymmetry or skewness. A distribution is said to be 'skewed' when the mean and the median fall at different points in the distribution, and the balance (or centre of gravity) is shifted to one side or the other-to left or right.

The above definitions show that the term 'skewness' refers to "lack of symmetry" i.e., when a distribution is not symmetrical (or is asymmetrical) it is called a skewed distribution.

Lack of symmetry is called Skewness. If a distribution is not symmetrical then it is called skewed distribution. So, mean, median and mode are different in values and one tail becomes longer than other. The skewness may be positive or negative.

The concept of skewness will be clear from the following three diagrams showing a symmetrical distribution. a positively skewed distribution and a negatively skewed distribution.

1. **Symmetrical Distribution.** In a symmetrical distribution the values of mean, median and mode coincide. The spread of the frequencies is the same on both sides of the centre point of the curve.

2. **Asymmetrical Distribution.** A distribution, which is not symmetrical, is called a skewed distribution and such a distribution could either be positively skewed or negatively skewed.

3. **Positively Skewed Distribution.** In the positively skewed distribution the value of the mean is maximum and that of mode least-the median lies in between the two. In the positively skewed distribution the frequencies are spread out over a greater range of values on the high-value end of the curve (the right-hand side) than they are on the low-value end.
   If the frequency curve has longer tail to right the distribution is known as positively skewed distribution and Mean > Median > Mode.

4. **Negatively Skewed Distribution.** In a negatively skewed distribution the value of mode is maximum and that of mean least-the median lies in between the two. In the negatively skewed distribution the position is reversed, i.e. the excess tail is on the left-hand side.

   If the frequency curve has longer tail to left the distribution is known as negatively skewed distribution and Mean < Median < Mode.

It should be noted that in moderately symmetrical distributions the interval between the mean and the median is approximately one-third of the interval between the mean and the mode. It is this relationship, which provides a means of measuring the degree of skewness.

8.3 Test of Skewness

In order to ascertain whether a distribution is skewed or not the following tests may be applied. Skewness is present if:

1. The values of mean, median and mode do not coincide.

2. When the data are plotted on a graph they do not give the normal bellshaped form i.e. when cut along a vertical line through the centre the two halves are not equal.

3. The sum of the positive deviations from the median is not equal to the sum of the negative deviations.

4. Quartiles are not equidistant from the median.

5. Frequencies are not equally distributed at points of equal deviation from the mode.

8.4 Measures of Skewness

There are four measures of skewness, each divided into absolute and relative measures. The relative measure is known as the coefficient of skewness and is more frequently used than the absolute measure of skewness. Further, when a comparison between two or more distributions is involved, it is the relative measure of skewness, which is used. The measures of skewness are: (i) Karl Pearson's measure, (ii) Bowley's measure, (iii) Kelly's measure, and (iv) Moment's measure. These measures are discussed briefly below:

8.4.1 Karl Pearson's Measure

The formula for measuring skewness as given by Karl Pearson is as follows:

The formula for measuring skewness as given by Karl Pearson is as follows:

$$\text{Skewness} = \text{Mean} - \text{Mode}$$

$$\text{Coefficient of Skewness} = \frac{\text{Mean} - \text{Mode}}{\text{Standard deviation}}$$

In case the mode is indeterminate, the coefficient of skewness is:

$$\text{SK}_p = \frac{\text{Mean} - (3\ \text{Median} - 2\ \text{Mean})}{\text{Standard deviation}} = \frac{3\,(\text{Mean} - \text{median})}{\text{standard deviation}}$$

Now this formula is equal to the earlier one.

$$\frac{3\,(\text{Mean} - \text{median})}{\text{standard deviation}} = \frac{\text{Mean} - \text{Mode}}{\text{Standard deviation}}$$

Or 3 Mean − 3 Median = Mean − Mode

Or Mode = Mean − 3 Mean + 3 median.

Or Mode = 3 Mean − 2 Mean.

The direction of skewness is determined by ascertaining whether the mean is greater than the mode or less than the mode. If it is greater than the mode, then skewness is positive. But when the mean is less than the mode, it is negative. The difference between the mean and mode indicates the extent of departure from symmetry. It is measured in standard deviation units, which provide a measure independent of the unit of measurement. It may be recalled that this observation was made in the preceding chapter while discussing standard deviation. The value of coefficient of skewness is zero, when the distribution is symmetrical. Normally, this coefficient of skewness lies between ±1. If the mean is greater than the mode, then the coefficient of skewness will be positive, otherwise negative.

Example 9.1. Given the following data, calculate the Karl Pearson's coefficient of skewness.

$$\sum X = 452, \quad \sum X^2 = 24, \quad 270, \text{Mode} = 43.7, \quad \text{and } n = 10.$$

Solution. Pearson's coefficient of skewness is:

$$\text{Coefficient of Skewness} = \frac{\text{Mean} - \text{Mode}}{\text{Standard deviation}}$$

$$\bar{X} = \frac{\sum X}{n} = \frac{452}{10} = 45.2$$

$$\text{SD} = \sqrt{\frac{\sum X^2}{n} - \left(\frac{\sum X}{n}\right)^2} = \sqrt{\frac{24,270}{10} - \left(\frac{452}{10}\right)^2}$$

$$= \sqrt{2427 - (45.2)^2} = 19.59$$

Applying the values of mean, mode and standard deviation in the above formula,

$$\text{SK}_p = \frac{45.2 - 43.7}{19.59} = 0.08$$

This shows that there is a positive skewness though the extent of skewness is marginal.

Question 9.1. From the following data, calculate the measure of skewness using the mean, median and standard deviation:

X	$10-19$	$20-29$	$30-39$	$40-49$	$50-59$	$60-69$	$70-79$
f	18	30	40	55	38	20	16

8.4.2 Bowley's Measure

Bowley developed a measure of skewness, which is based on quartile values. The formula for measuring skewness is:

$$\text{Skewness} = Q_3 + Q_1 - 2M$$

Bowley's coefficient of skewness is:

$$\text{sk}_B = \frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1}$$

Where Q_3 and Q_1 are upper and lower quartiles and M is the median. The value of this skewness varies between ± 1. In the case of open-ended distribution as well as where extreme values are found in the series, this measure is particularly useful. In a symmetrical distribution, skewness is zero. This means that Q_3 and Q_1 are positioned equidistantly from Q_2 that is, the median. In symbols, $Q_3 - Q_2 = Q_2 - Q_1$. In contrast, when the distribution is skewed, then $Q_3 - Q_2$ will be different from $Q_2 - Q_1$. When $Q_3 - Q_2$ exceeds $Q_2 - Q_1$ then skewness is positive. As against this; when $Q_3 - Q_2$ is less than $Q_2 - Q_1$ then skewness is negative. Bowley's measure of skewness can be written as:

$$\text{Skewness} = (Q_3 - Q_2) - (Q_2 - Q_1) = (Q_3 - Q_2 - Q_2 + Q_1 = Q_3 + Q_1 - 2Q_2.$$

However, this is an absolute measure of skewness. As such, it cannot be used while comparing two distributions where the units of measurement are different. In view of this limitation, Bowley suggested a relative measure of skewness as given below:

$$\begin{aligned}\text{Relative skewness} &= \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)} \\ &= \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}.\end{aligned}$$

Example 9.2. For a distribution, Bowley's coefficient of skewness is -0.56, $Q_1 = 16.4$ and Median $= 24.2$. What is the coefficient of quartile deviation?

Solution. Bowley's coefficient of skewness is:

$$\text{sk}_B = \frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1}$$

Substituting the values in the above formula,

$$\begin{aligned} \mathrm{sk}_B &= \frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1} \\ \mathrm{sk}_B &= \frac{Q_3 + 16.4 - (2 \times 24.2)}{Q_3 - 16.4} \\ -0.56 &= \frac{Q_3 + 16.4 - 48.4}{Q_3 - 16.4} \\ Q_3 &= 26.4 \end{aligned}$$

Now, we have the values of both the upper and the lower quartiles.

$$\begin{aligned} \text{Coefficient of Quartile deviation} &= \frac{Q_3 - Q_1}{Q_3 + Q_1} \\ &= \frac{26.4 - 16.4}{26.4 + 16.4} = \frac{10}{42.8} \approx 0.234 \end{aligned}$$

.

8.4.3 Kelly's Measure

Kelly developed another measure of skewness, which is based on percentiles. The formula for measuring skewness is as follows:

$$\begin{aligned} \text{Coefficient of skewness} &= \frac{P_{90} - 2P_{50} + P_{10}}{P_{90} - P_{10}} \\ &= \frac{D_1 + D_9 - 2M}{D_9 - D_1} \end{aligned}$$

Where P and D stand for percentile and decile respectively. In order to calculate the coefficient of skewness by this formula, we have to ascertain the values of 10th, 50th and 90th percentiles. Somehow, this measure of skewness is seldom used. All the same, we give an example to show how it can be calculated.

8.5 Moment Coefficient of Skewness

The moment coefficient of skewness of a data set is skewness:; $g_1 = \frac{m_3}{m_2^{3/2}}$ where $m_3 = \sum(x - \bar{x})^3/n$ and $m_2 = \sum(x - \bar{x})^2/n$, $\bar{x}$ is the mean and n is the sample size, as usual. m_3 is called the third moment of the data set. m_2 is the variance, the square of the standard deviation.

8.6 Skewness of a distribution

Skewness is a measure of symmetry, or more precisely, the departure from symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point.

A frequency distribution is said to be **positively** or **right skewed** if the longer tail is towards the positive x direction, so that the peak of the frequency curve lies to the left of the center. If on the other hand, the peak of the curve lies to the right of the center so that the longer tail of the frequency curve is towards the left or negative x direction the distribution is said to be **negatively** or **left skewed**.

8.6.1 Measures of Skewness

The degree of skewness is measured by coefficiency of skewness.

(a). **The Pearsonian Measure of Skewness**

$$S \cdot K_p = \frac{\text{Mean} - \text{mode}}{\text{Standard deviation}}$$

There is an empirical relationship connecting the mean, the median and the mode.

$$\text{Mean} - \text{Mode} = 3\,(\text{Mean} - \text{Median})$$

The Pearsonian measure of Skewness is then given as

$$S \cdot K_p = \frac{3\,(\text{Mean} - \text{Median})}{\text{Standard deviation}}$$

(b). **The Quartile Coefficient of Skewness**

$$S \cdot K_Q = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)}$$

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

(c). **The Percentile Coefficient of Skewness** This is given by

$$S \cdot K_c = \frac{(P_{90} - P_{50}) - (P_{50} - P_{10})}{(P_{90} - P_{50}) + (P_{50} - P_{10})}$$

$$= \frac{(P_{90} + P_{10} - 2P_{50})}{(P_{90} - P_{10})}$$

(d). **The Moment Coefficient of Skewness**

This is given by

$$a_3 = \frac{m_3}{s^3} = \frac{\frac{1}{n}\Sigma f(x - \bar{x})^3}{\left[\frac{1}{n}\Sigma f(x - \bar{x})^2\right]^{\frac{3}{2}}}$$

Note: For a Normal distribution (perfectly symmetrical bell shaped curve) the measure of Skewness a_3 is zero. Any symmetric data should have a skewness near zero. Negative values for the skewness indicate data that are skewed left and positive values for the skewness indicate data that are skewed right. By skewed left, we mean that the left tail is long relative to the right tail. Similarly, skewed right means that the right tail is long relative to the left tail. Some measurements have a lower bound and are skewed right. For example, in reliability studies, failure times cannot be negative.

8.7 Kurtosis

Kurtosis is another measure of the shape of a frequency curve. It is a Greek word, which means bulginess. While skewness signifies the extent of asymmetry, kurtosis measures the degree of peakedness of a frequency distribution. Karl Pearson classified curves into three types on the basis of the shape of their peaks. These are mesokurtic, leptokurtic and platykurtic. These three types of curves are shown in figure below:

It will be seen from Fig. 3.2 that mesokurtic curve is neither too much flattened nor too much peaked. In fact, this is the frequency curve of a normal distribution. Leptokurtic curve is a more peaked than the normal curve. In contrast, platykurtic is a relatively flat curve. The coefficient of kurtosis as given by Karl Pearson is$\beta_2 = \mu_4/\mu_2^2$. In case of a normal distribution, that is, *mesokurtic curve*, the value of $\beta_2 = 3$. If $\mu_2 > 3$, the curve is called a *leptokurtic curve* and is more peaked than the normal curve. Again, when $\beta_2 < 3$, the curve is called a *platykurtic curve* and is less peaked than the normal curve. The measure of kurtosis is very helpful in the selection of an appropriate average. For example, for normal distribution, mean is most appropriate; for a leptokurtic distribution, median is most appropriate; and for platykurtic distribution, the quartile range is most appropriate.

8.8 Kurtosis of a distribution

Kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution. That is, data sets with high kurtosis tend to have a distinct peak near the mean, decline rather rapidly, and have heavy tails. Data sets with low kurtosis tend to have a flat top near the mean rather than a sharp peak. This is a measure of "peakedness" of a distribution.

8.8.1 Measures of Kurtosis

(a). **Percentile coefficient of kurtosis**

This is a measure of kurtosis based on both quartiles and percentiles and is given by:

$$k = \frac{\frac{1}{2}(Q_3 - Q_1)}{P_{90} - P_{10}}$$

For a normal (mesokurtic) distribution $k = 0.263$.

$k < 0.263$ platykurtic

$k > 0.263$ leptokurtic

$k \approx 0.263$ mesokurtic (medium peak)

(b). **Moment Measure of Kurtosis**

$$a_4 = \frac{m_4}{S_4} = \frac{M_4}{m_2^2}$$

For a normal (mesokurtic) distribution $a_4 = 3$. So that;

$a_4 < 3$ platykurtic

$a_4 > 3$ leptokurtic

$a_4 \approx 3$ mesokurtic (medium peak)

The kurtosis for a standard normal distribution is three. For this reason, some sources use the following definition of kurtosis (often referred to as "excess kurtosis"):

$$\text{Kurtosis} = \frac{\sum_{i=1}^{N}(Y_i - \bar{y})}{(N-1)s^4} - 3$$

This definition is used so that the standard normal distribution has a kurtosis of zero. In addition, with the second definition positive kurtosis indicates a "peaked" distribution and negative kurtosis indicates a "flat" distribution.

8.9 Practice Problems

1. The first four central moments of a distribution are 0, 2.5,0.7 and 8.75. Comment on the skewness and kurtosis of this distribution.

2. What do you mean by dispersion? What are the different measures of dispersion?

3. Why is the standard deviation the most widely used measure of dispersion? Explain.

4. Define skewness and Dispersion.

5. Define Kurtosis and Moments.

6. For a distribution, the first four moments about zero are 1, 7, 38 and 155 respectively. (i) Compute the moment coefficients of skewness and kurtosis. (ii) Is the distribution mesokurtic? Give reason.

7. The first four moments of a distribution about the value 4 are 1,4, 10 and 45. Obtain various characteristics of the distribution on the basis of the information given. Comment upon the nature of the distribution.

8. Define kurtosis. If $\beta_1 = 1$ and $\beta_2 = 4$ and variance = 9, find the values of β_3 and β_4 and comment upon the nature of the distribution.

Lecture 8: Correlations Analysis

10 Correlations Analysis

Objective: *The overall objective of this lesson is to give you an understanding of bivariate linear correlation, there by enabling you to understand the importance as well as the limitations of correlation analysis.* 7.1 Introduction

7.2 What is Correlation?

7.3 Correlation Analysis

7.3.1 Scatter Diagram

7.3.2 Correlation Graph

7.3.3 Pearson's Coefficient of Correlation

7.3.4 Spearman's Rank Correlation

7.3.5 Concurrent Deviation Method

7.4 Limitations of Correlation Analysis

7.5 Self-Assessment Questions

10.1 Introduction

Statistical methods of measures of central tendency, dispersion, skewness and kurtosis are helpful for the purpose of comparison and analysis of distributions involving only one variable i.e. univariate distributions. However, describing the relationship between two or more variables, is another important part of statistics.

There are situations where data appears as pairs of figures relating to two variables. A correlation problem considers the joint variation of two measurements neither of which is restricted by the experimenter. The regression problem discussed in this lecture considers the frequency distribution of one variable (called the dependent variable) when another (independent variable) is held fixed at each of several levels.

Examples of correlation problems are found in the study of the relationship between IQ and aggregate percentage of marks obtained by a person in the SSC examination, blood pressure and metabolism or the relation between height and weight of individuals. In these examples both variables are observed as they naturally occur, since neither variable is fixed at predetermined levels.

Examples of regression problems can be found in the study of the yields of crops grown with different amount of fertilizer, the length of life of certain animals exposed to different levels of radiation, and so on. In these problems the variation in one measurement is studied for particular levels of the other variable selected by the experimenter.

In many research situations, the key to decision making lies in understanding the relationships between two or more variables. For example, in an effort to predict the behavior of the bond market, a broker might find it useful to know whether the interest rate of bonds is related to the prime interest rate. While studying the effect of advertising on sales, an account executive may find it useful to know whether there is a strong relationship between advertising costs and sales volumes for a company.
The statistical methods of **correlation** (discussed in the present lesson) and **regression** (to be discussed in the next lesson) are helpful in knowing the relationship between two or more variables which may be related in same way, like interest rate of bonds and prime interest rate; advertising expenditure and sales; income and consumption; crop-yield and fertilizer used; height and weights; profits and sales; hours of study and student marks and so on.

In all these cases involving two or more variables, we may be interested in seeing:

- if there is any association between the variables;
- if there is an association, is it strong enough to be useful;
- if so, what form the relationship between the two variables takes;
- how we can make use of that relationship for predictive purposes, that is, forecasting; and
- how good such predictions will be.

When we collect data on two of such characteristics it is called bivariate data. It is generally denoted by (X,Y) where X and Y are the variables representing the values of the characteristics. Therefore, correlation analysis gives the idea about the nature and extent of relationship between two variables in the bivariate data.

10.2 Definition of Correlation

Correlation is a measure of association between two or more variables. When two or more variables very in sympathy so that movement in one tends to be accompanied by corresponding movements in the other variable(s), they are said to be correlated.

> *"The correlation between variables is a measure of the nature and degree of association between the variables"*.

As a measure of the degree of relatedness of two variables, correlation is widely used in exploratory research when the objective is to locate variables that might be related in some way to the variable of interest.

Correlation measures the degree of linear relation between the variables.*The existence of correlation between variables does not necessarily mean that one is the cause of the change in the other. It should noted that the correlation analysis merely helps in determining the degree of association between two variables, but it does not tell any thing about the cause and effect relationship.* While interpreting the correlation coefficient, it is necessary to see whether there is any cause and effect relationship between variables under study. If there is no such relationship, the observed is meaningless.

In correlation analysis, all variables are assumed to be random variables.

11 Types of correlation

Correlation can be classified in several ways. The important ways of classifying correlation are:

(i) Positive and negative,

(ii) Linear and non-linear (curvilinear) and

(iii) Simple, partial and multiple.

Positive and Negative Correlation: If both the variables move in the same direction, we say that there is a positive correlation, i.e., if one variable increases, the other variable also increases on an average or if one variable decreases, the other variable also decreases on an average. On the other hand, if the variables are varying in opposite direction, we say that it is a case of negative correlation; e.g., movements of demand and supply.

Linear and Non-linear (Curvilinear) Correlation: If the change in one variable is accompanied by change in another variable in a constant ratio, it is a case of linear correlation. Observe the following data:

X :	10	20	30	40	50
Y :	25	50	75	100	125

The ratio of change in the above example is the same. It is, thus, a case of linear correlation. If we plot these variables on graph paper, all the points will fall on the same straight line.

On the other hand, if the amount of change in one variable does not follow a constant ratio with the change in another variable, it is a case of non-linear or curvilinear correlation. If a couple of figures in either series X or series Y are changed, it would give a non-linear correlation.

Simple, Partial and Multiple Correlation: The distinction amongst these three types of correlation depends upon the number of variables involved in a study. If only two variables are involved in a study, then the correlation is said to be simple correlation. When three or more variables are involved in a study, then it is a problem of either partial or multiple correlation. In multiple correlation, three or more variables are studied simultaneously. But in partial correlation we consider only two variables influencing each other while the effect of other variable(s) is held constant.

Suppose we have a problem comprising three variables X, Y and Z. X is the number of hours studied, Y is I.Q. and Z is the number of marks obtained in the examination. In a multiple correlation, we will study the relationship between the marks obtained (Z) and the two variables, number of hours studied (X) and I.Q. (Y). In contrast, when we study the relationship between X and Z, keeping an average I.Q. (Y) as constant, it is said to be a study involving partial correlation.

In this lecture, we will study linear correlation between two variables.

Correlation does not necessarily mean causation

The correlation analysis, in discovering the nature and degree of relationship between variables, does not necessarily imply any cause and effect relationship between the variables. Two variables may be related

to each other but this does not mean that one variable causes the other. For example, we may find that logical reasoning and creativity are correlated, but that does not mean if we could increase peoples' logical reasoning ability, we would produce greater creativity. We need to conduct an actual experiment to unequivocally demonstrate a causal relationship. But if it is true that influencing someones' logical reasoning ability does influence their creativity, then the two variables must be correlated with each other. In other words, ***causation always implies correlation, however converse is not true.***

Correlation coefficient is merely a mathematical relationship and this has nothing to do with cause and effect relation. It only reveals co-variation between two variables. Even when there is no cause-and-effect relationship in bivariate series and one interprets the relationship as causal, such a correlation is called spurious or non-sense correlation.

11.1 Correlation Analysis

Correlation analysis is the process of finding how well (or badly) the line fits the observations, such that if all the observations lie exactly on the line of best fit, the correlation is considered to be 1 or unity. Correlation analysis measures the strength of the linear relationship between them. In particular, correlation analysis provides us with two important measures of this strength: (1) the coefficient of correlation and (2) the coefficient of determination.

Correlation analysis is a statistical technique used to indicate the nature and degree of relationship existing between one variable and the other(s). It is also used along with regression analysis to measure how well the regression line explains the variations of the dependent variable with the independent variable.

The commonly used methods for studying linear relationship between two variables involve both graphic and algebraic methods. Some of the widely used methods include:

1. Scatter Diagram
2. Correlation Graph
3. Pearson's Coefficient of Correlation
4. Spearman's Rank Correlation

12 Methods of studying correlation

Three important statistical tools used to measure correlation are scatter diagrams, Karl Pearson's coefficient of correlation and Spearman's rank correlation. A scatter diagram visually presents the nature of association without giving any specific numerical value. A numerical measure of linear relationship between two variables is given by Karl Pearson's coefficient of correlation. A relationship is said to be linear if it can be represented by a straight line. Spearman's coefficient of correlation measures the linear association between ranks assigned to individual items according to their attributes. Attributes are those variables which cannot be numerically measured such as intelligence of people, physical appearance, honesty, etc.

12.1 The Scatter diagram

The first step in correlation and regression analysis is to visualize the relationship between the variables. A scatter diagram is obtained by plotting the points (x_1,y_1), $(x_2,y_2),\cdots,(x_n,y_n)$ on a two-dimensional plane. If the points are scattered around a straight line, we may infer that there exist a *linear relationship* between the variables. If the points are clustered around a straight line with negative slope, then there exist *negative correlation* or the variables are inversely related (i.e, when x increases y decreases and vice versa). If the points are clustered around a straight line with positive slope, then there exist *positive correlation* or the variables are directly related (i.e, when x increases y also increases and vice versa).

A scatter diagram gives two very useful types of information. First, we can observe patterns between variables that indicate whether the variables are related. Secondly, if the variables are related we can get an idea of what kind of relationship (linear or non-linear) would describe the relationship.

Correlation examines the first question of determining whether an association exists between the two variables, and if it does, to what extent. Regression examines the second question of establishing an appropriate relation between the variables.

12.1.1 Scatter Diagram

Scatter diagram is one of the simplest methods of diagrammatic representation of a bivariate distribution. Under this method, both the variables are plotted on the graph paper by putting dots. The diagram so obtained is called "Scatter plot". By studying diagram, we can have rough idea about the nature and degree of relationship between two variables. The term scatter refers to the spreading of dots on the graph. We should keep the following points in mind while interpreting correlation:

- if the plotted points are very close to each other, it indicates high degree of correlation. If the plotted points are away from each other, it indicates low degree of correlation.
- if the points on the diagram reveal any trend (either upward or downward), the variables are said to be correlated and if no trend is revealed, the variables are uncorrelated.
- if there is an upward trend rising from lower left hand corner and going upward to the upper right hand corner, the correlation is positive since this reveals that the values of the two variables move in the same direction. If, on the other hand, the points depict a downward trend from the upper left hand corner to the lower right hand corner, the correlation is negative since in this case the values of the two variables move in the opposite directions.
- in particular, if all the points lie on a straight line starting from the left bottom and going up towards the right top, the correlation is perfect and positive, and if all the points like on a straight line starting from left top and coming down to right bottom, the correlation is perfect and negative.

The various diagrams of the scattered data in Figure 4-1 depict different forms of correlation.

Example 12.1. Given the following data on sales (in thousand units) and expenses (in thousand shillings) of a firm for 10 month:

Month:	J	F	M	A	M	J	J	A	S	O
Sales:	50	50	55	60	62	65	68	60	60	50
Expenses:	11	13	14	16	16	15	15	14	13	13

(a) Make a Scatter Diagram

(b) Do you think that there is a correlation between sales and expenses of the firm? Is it positive or negative? Is it high or low?

12.1.2 Correlation Graph

This method, also known as Correlogram is very simple. The data pertaining to two series are plotted on a graph sheet. We can find out the correlation by examining the direction and closeness of two curves. If both the curves drawn on the graph are moving in the same direction, it is a case of positive correlation. On the other hand, if both the curves are moving in opposite direction, correlation is said to be negative. If the graph does not show any definite pattern on account of erratic fluctuations in the curves, then it shows an absence of correlation.

12.1.3 Pearson's coefficient of correlation

A mathematical method for measuring the intensity or the magnitude of *linear relationship* between two variables was suggested by Karl Pearson (1867-1936), a great British Biometrician and Statistician and, it is by far the most widely used method in practice.

Karl Pearson's measure, known as Pearsonian correlation coefficient between two variables X and Y, usually denoted by $r(X,Y)$ or r_{xy} or simply r is a numerical measure of linear relationship between them and is defined as the ratio of the covariance between X and Y, to the product of the standard deviations of X and Y.

Mathematically;

$$r = \frac{\text{Cov}(X,Y)}{\sigma_x\, \sigma_y}$$

where, $(X_1,Y_1),(X_2,Y_2)\ \cdots(X_n,Y_n)$ are n pairs of observations of the variables X and Y in a bivariate distribution, Cov(X,Y), covariance between x and y, σ_x-standard deviation of x and σ_y-standard deviation of y.

Also

$$\text{Cov}\,(X,Y) = \frac{1}{n}\sum(X-\bar{X})(Y-\bar{Y}) = \frac{1}{n}\sum XY - \bar{X}\bar{Y} \quad (3)$$

$$\sigma_x = \sqrt{\frac{1}{n}\sum(X-\bar{X})^2} = \sqrt{\frac{1}{n}\sum X^2 - \bar{X}^2} \quad (4)$$

$$\sigma_y = \sqrt{\frac{1}{n}\sum(Y-\bar{Y})^2} = \sqrt{\frac{1}{n}\sum Y^2 - \bar{Y}^2}$$

(5) Thus by substituting Eqs. (5) in Eq. (2), we can write the Pearsonian correlation coefficient as

$$r_{xy} = \frac{\frac{1}{n}\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\frac{\sum(X-\bar{X})^2}{n}}\sqrt{\frac{\sum(Y-\bar{Y})^2}{n}}}$$
$$= \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})}\sqrt{\sum(Y-\bar{Y})}}$$
$$= \frac{\frac{\sum XY}{n} - \bar{X}\bar{Y}}{\sqrt{\left(\frac{\sum X^2}{n} - \bar{X}^2\right)\left(\frac{\sum Y^2}{n} - \bar{Y}^2\right)}}$$

or

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2]\,[n\Sigma y^2 - (\Sigma y)^2]}}$$

The Pearson's Correlation coefficient is also called as the product moment correlation coefficient.

Interpretation of Pearson's Correlation Coefficient

The sign of the correlation coefficient determines whether the correlation is positive or negative. The magnitude of the correlation coefficient determines the strength of the correlation.

Example 12.2. Calculate the Karl Pearson's correlation coefficient from the following:

$X:$	12	10	20	13	15
$Y:$	7	14	6	12	11

Solution. The table of calculations is given below, where $n = 5$;

X	Y	XY	X^2	Y^2
12	7	84	144	49
10	14	140	100	196
20	6	120	400	36
13	12	156	169	144
15	11	165	225	121
$\sum X = 70$	$\sum Y = 50$	$\sum XY = 665$	$\sum X^2 = 1038$	$\sum Y^2 = 546$

The Pearson's correlation coefficient *r* is given by;

$$r = \frac{\text{Cov}(X,Y)}{\sigma_x\,\sigma_y}$$

where;

$$\bar{X} = \frac{\sum X}{n} = \frac{70}{5} = 14, \qquad \bar{Y} = \frac{\sum Y}{n} = \frac{50}{5} = 10$$

$$\text{Cov}(X,Y) = \frac{\sum XY}{n} - \bar{X}\bar{Y} = \frac{665}{5} - 14 \times 10 = 133 - 14 = -7$$

$$\sigma_x = \sqrt{\frac{1}{n}\sum X^2 - \bar{X}^2} = \sqrt{\frac{1038}{5} - 14^2} = \sqrt{11.6} = 3.40$$

$$\sigma_y = \sqrt{\frac{1}{n}\sum Y^2 - \bar{Y}^2} = \sqrt{\frac{546}{5} - 10^2} = \sqrt{9.2} = 3.03$$

Substituting the values in the formula of r we get

$$r = \frac{\text{Cov}(X,Y)}{\sigma_x\, \sigma_y} = \frac{-7}{3.40 \times 3.03} = -0.68$$

Properties of Pearson's Correlation Coefficient

The following are important properties of Pearsonian correlation coefficient:

1. *Pearsonian correlation coefficient cannot exceed 1 numerically*. In other words it lies between ï£¡1 and +1.

   $$-1 \le r \le +1$$

   Remarks:

   (a) This property provides us a check on our calculations. If in any problem, the obtained value of r lies outside the limits ±1, this implies that there is some mistake in our calculations.

   (b) The sign of r indicate the nature of the correlation. Positive value of r indicates positive correlation, whereas negative value indicates negative correlation. r = 0 indicate absence of correlation.

2. *Pearsonian Correlation coefficient is independent of the change of origin and scale.*

   Mathematically, if given variables X and Y are transformed to new variables U and V by change of origin and scale, i. e.

   $$U = \frac{X - A}{h} \quad \text{and} \quad V = \frac{Y - B}{k}$$

   Where A, B, h and k are constants and $h > 0$, $k > 0$; then the correlation coefficient between X and Y is same as the correlation coefficient between U and V i.e.,

   $$r(X,Y) = r(U,V) \Rightarrow r_{xy} = r_{uv}$$

3. *Two independent variables are uncorrelated but the converse is not true.* If X and Y are independent variables then

   $$r_{xy} = 0$$

However, the converse of the theorem is not true i.e., uncorrelated variables need not necessarily be independent.

4. *Pearsonian coefficient of correlation is the geometric mean of the two regression coefficients, i.e.*

$$r_{xy} = \pm\sqrt{b_{xy} \cdot b_{yx}}$$

The signs of both the regression coefficients are the same, and so the value of r will also have the same sign.

5. *The square of Pearsonian correlation coefficient is known as the coefficient of determination.* Coefficient of determination, which measures the percentage variation in the dependent variable that is accounted for by the independent variable, is a much better and useful measure for interpreting the value of r.

12.1.4 Coefficient of Determination

Question 12.1. Calculate the coefficient of correlation between marks of students in the subjects of mathematics and statistics in a certain test conducted.

Maths(X):	28	25	32	16	20	15	19	17	40	30
Statistics(Y):	30	40	50	18	25	12	11	21	45	35

12.2 Spearman's Rank Correlation

Sometimes we come across statistical series in which the variables under consideration are not capable of quantitative measurement but can be arranged in serial order. This happens when we are dealing with qualitative characteristics (attributes) such as honesty, beauty, character, morality, etc., which cannot be measured quantitatively but can be arranged serially. In such situations Karl Pearsonï£¡s coefficient of correlation cannot be used as such. Charles Edward Spearman, a British Psychologist, developed a formula in 1904, which consists in obtaining the correlation coefficient between the ranks of N individuals in the two attributes under study.

The Pearsonian correlation coefficient between the ranks X and Y is called the rank correlation coefficient between the characteristics A and B for the group of individuals.

Spearman's rank correlation coefficient, usually denoted by r_s, R or ρ(Rho) is given by the equation

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)}. \quad (6)$$

Where d is the difference between the pair of ranks of the same individual in the two characteristics and n is the number of pairs.

Remarks: The value of r_s always lies between -1 and $+1$. The positive value of r_s indicates positive correlation (association) in the rank allocation. Whereas, the negative value of r_s indicates the negative correlation (association) in the rank allocation.

The calculation of rank correlation will be illustrated under three situations.

1. The ranks are given.

2. The ranks are not given. They have to be worked out from the data.

3. Ranks are repeated.

Case 1: When the ranks are given

Example 12.3. Data given below read the ranks assigned by two judges to 8 participants. Calculate the coefficient of rank correlation.

Participant	Ranks by Judge		Rank diff squared
	A	B	d^2
1	5	4	$(5-4)^2=1$
2	6	8	4
3	7	1	36
4	1	7	36
5	8	5	9
6	2	6	16
7	3	2	1
8	4	3	1
$n=8$			$\sum d^2=104$

Solution. Spearman's rank correlation coefficient is given by

$$r_s = 1 = \frac{6\sum d^2}{n(n^2-1)}$$

substituting the values from the table we get,

$$r_s = 1 = \frac{6\times 104}{8(8^2-1)} = -0.23$$

The value of correlation coefficient is -0.23. This indicates that there is negative association in rank allocation by the two judges A and B.

Example 12.4. Five persons are assessed by three judges in a beauty contest. We have to find out which pair of judges has the nearest approach to common perception of beauty.

Competitors

Judge	1	2	3	4	5
A	1	2	3	4	5
B	2	4	1	5	3
C	1	3	5	2	4

There are 3 pairs of judges necessitating calculation of rank correlation thrice. Formula (6) will be used.

The rank correlation between A and B is calculated as follows: 0.3 The rank correlation between A and C is calculated as follows: 0.5.

Similarly, the rank correlation between the rankings of judges B and C is 0.9. Thus, the perceptions of judges A and C are the closest. Judges B and C have very different tastes.

Spearman's rank correlation can also be used even if we are dealing with variables, which are measured quantitatively, i.e. when the actual data but not the ranks relating to two variables are given. In such a case we shall have to convert the data into ranks. The highest (or the smallest) observation is given the rank 1. The next highest (or the next lowest) observation is given rank 2 and so on. It is immaterial in which way (descending or ascending) the ranks are assigned. However, the same approach should be followed for all the variables under consideration.

Case 2: When the ranks are not given

Example 12.5. We are given the percentage of marks, secured by 5 students in Economics and Statistics. Then the ranking has to be worked out and the rank correlation is to be calculated.

Student	Marks in Statistics (X)	Marks in Economics (Y)
A	85	60
B	60	48
C	55	49
D	65	50
E	75	55

Student	Marks in Statistics (X)	Marks in Economics (Y)
A	1	1
B	4	5
C	5	4
D	3	3
E	2	2

Once the ranking is complete formula (6) is used to calculate rank correlation.

Example 12.6. Calculate the rank coefficient of correlation from the following data:

X :	75	88	95	70	60	80	81	50
Y :	120	134	150	115	110	140	142	100

Solution. The table below gives the calculations for coefficient of rank correlation.

X	Ranks R_X	Y	Ranks R_Y	$d = R_X - R_Y$	d^2
75	5	120	5	0	0
88	2	134	4	-2	4
95	1	150	1	0	0
70	6	115	6	0	0
60	7	110	7	0	0
80	4	140	3	1	1
81	3	142	2	1	1
50	8	100	8	0	0
					$\sum d^2 = 6$

$$\begin{aligned} r_s &= 1 - \frac{6\sum d^2}{n(n^2-1)} \\ &= 1 - \frac{6 \times 6}{8(8^2-1)} = 1 - \frac{36}{504} \\ &= 0.93 \end{aligned}$$

Hence, there is a high degree of positive correlation between X and Y.

Question 12.2. The data below gives the marks given by two examiners to a set of 10 students in a aptitude test. Calculate the Spearman's Rank correlation coefficent.

A	85	56	45	65	96	52	80	75	78	60
B	80	60	50	62	90	55	75	68	77	53

Case 3: When the ranks are repeated

If two or more data have the same value, then they are said to be "tied", and each of their ranks may be set equal to the mean of the ranks of the positions they occupy in the ordered data set.

In case of attributes if there is a tie i.e., if any two or more individuals are placed together in any classification w.r.t. an attribute or if in case of variable data there is more than one item with the same value in either or both the series then Spearman's correlation rank for calculating the rank correlation coefficient breaks down, since in this case the variables X [the ranks of individuals in characteristic A (1st series)] and Y [the ranks of individuals in characteristic B (2nd series)] do not take the values from 1 to n.

In this case common ranks are assigned to the repeated items. These common ranks are the arithmetic mean of the ranks, which these items would have got if they were different from each other and the next item will get the rank next to the rank used in computing the common rank. For example, suppose an item is repeated at rank 4. Then the common rank to be assigned to each item is $(4 + 5)/2$, i.e., 4.5 which is the average of 4 and 5, the ranks which these observations would have assumed if they were different. The next item will be assigned the rank 6. If an item is repeated thrice at rank 7, then the common rank to be assigned to each value will be $(7 + 8 + 9)/3$, i.e., 8 which is the arithmetic mean of 7, 8 and 9 viz., the ranks these observations would have got if they were different from each other. The next rank to be assigned will be 10.

If only a small proportion of the ranks are tied, this technique may be applied together with Eq.(6). If a large proportion of ranks are tied, it is advisable to apply an adjustment or a correction factor to Eq. (6) as explained below:

"In the Eq.(6) add the factor

$$\frac{m(m^2-1)}{12}$$

to $2\sum d$; where m is the number of times an item is repeated. This correction factor is to be added for each repeated value in both the series",i.e.,

$$r_s = 1 - \frac{6\left[\sum d^2 + \frac{m(m^2-1)}{12}\right]}{n(n^2-1)}$$

Example 12.7. The data given below scores assigned by two judges for 10 participants in the singing competition. Calculate the Spearman's Rank correlation coefficient.

Participant	A	B	R_A	R_B	d^2
1	28	35	9(8.5)	6	6.25
2	40	26	3	10 (9.5)	42.25
3	35	42	5 (4.5)	3	2.25
4	25	26	10	9 (9.5)	0.25
5	28	33	8 (8.5)	7	2.25
6	35	45	4 (4.5)	2	6.25
7	50	32	1	8	49
8	48	51	2	1	1
9	32	39	6	4	4
10	30	36	7	5	4
$n = 10$					$\sum d^2 = 117.5$

Explanation: In the column of A and B there is repetition of scores so while assigning the ranks we first assign the ranks by treating them as different values and then for repeated scores we assign the average rank.
For example, in column A the score 35 appears 2 times at number 4 and 5 in the order of ranking so we calculate the average rank as $(4 + 5)/2 = 4.5$. Hence the ranks assigned are 4.5 each. The other repeated scores can be ranked in the same manner.

Note: In this example, we can note that the ranks are in fraction e.g., 4.5, which is logically incorrect or meaningless. Therefore in the calculation of ρ we add a correction factor (C.F.) to $\sum d^2$ calculated as follows.

Value Repeated	Frequency m	$m(m^2-1)$
35	2	$2(2^2-1) = 6$
28	2	6
26	2	6
	Total	$\sum m(m^2-1) = 18$

Now,

$$C.F = \frac{\sum m(m^2-1)}{12} = \frac{18}{12} = 1.5$$

Therefore

$$\sum d^2 + \frac{\sum m(m^2-1)}{12} = 117.5 + 1.5 = 119$$

We use this value in the calculation of ρ. Now the Spearman's rank correlation coefficient is given by

$$r_s = 1 - \frac{\sum d^2 + \frac{\sum m(m^2-1)}{12}}{n(n^2-1)}$$

Substituting the values we get,

$$r_s = 1 - \frac{6 \times 119}{10(10^2-1)} = 1 - 0.72 = 0.28$$

Example 12.8. The values of X and Y are given as:

X	25	45	35	40	15	19	35	42
Y	55	60	30	35	40	42	36	48

In order to work out the rank correlation, the ranks of the values are worked out. Common ranks are given to the repeated items. The common rank is the mean of the ranks which those items would have assumed if they were slightly different from each other. The next item will be assigned the rank next to the rank already assumed. The formula of Spearman's rank correlation coefficient when the ranks are repeated is as follows

$$r_s = 1 - \frac{6\left[\sum d^2 + \frac{(m_1^3 - m_1)}{12} + \frac{(m_1^3 - m_1)}{12} + \cdots\right]}{n(n^2-1)} \qquad (7)$$

where $m_1, m_2, \cdots$, are the number of repetitions of ranks and $\frac{(m_1^3 - m_1)}{12}$ . . . their corresponding correction factors.

X has the value 35 both at the 4th and 5th rank. Hence both are given the average rank i.e.,

$$\frac{4+5}{2}\text{ th } = 4.5 \text{ th rank.}$$

X	Y	Rank of $X = R'$	Rank of $X = R''$	$d = R' - R''$	Deviation d^2
25	55	6	2	4	16
45	80	1	1	0	0
35	30	4.5	8	3.5	12.25
40	35	3	7	-4	16
15	40	8	5	3	9
19	42	7	4	3	9
35	36	4.5	6	-1.5	2.25
42	48	2	3	-1	1
					$\sum d^2 = 65.5$

The necessary condition thus is

$$\frac{m^3 - m}{12} = \frac{2^3 - 2}{12} = \frac{1}{2}$$

Using the equation

$$r_s = 1 - \frac{6\left[\sum d^2 + \frac{(m_1^3 - m_1)}{12} + \frac{(m_1^3 - m_1)}{12} + \cdots\right]}{n(n^2 - 1)}$$

Substituting the values of these expressions;

$$r_s = 1 - \frac{6(65.5 + 0.5)}{8(64 - 1)} = 1 - \frac{396}{504}$$
$$= 1 - 0.786 = 0.214$$

Thus there is positive rank correlation between X and Y . Both X and Y move in the same direction. However, the relationship cannot be described as strong.

Question 12.3. For a certain joint stock company, the prices of preference shares (X) and debentures (Y) are given below:

X:	73.2	85.8	78.9	75.8	77.2	81.2	83.8
Y:	97.8	99.2	98.8	98.3	98.3	96.7	97.1

Use the method of rank correlation to determine the relationship between preference prices and debenture prices.

Question 12.4. Using the data provided below, construct rankings and calculate the rank order correlation coefficient using Spearman's Rho. Determine the significance of the correlation. Show all work.

X= 82, 73, 95, 66, 84, 89, 51, 82, 75, 90, 60, 81, 34, 49, 82, 95, 49

Y= 76, 83, 89, 76, 79, 73, 52, 89, 77, 85, 48, 69, 51, 25, 74, 60, 50

Question 12.5. Using the data provided below, construct rankings for each variable and calculate the rank order correlation coefficient using Spearman's Rho. Determine if the correlation is statistically significant. Show all work.

Draw a conclusion related to the correlation.

Subject X Y

1 82 80

2 73 75

3 95 89

4 66 76

5 84 79

6 89 73

7 51 69

8 82 89

9 75 77

10 90 85

11 51 71

12 89 69

13 34 51

14 50 60

15 87 74

The process to calculate Spearman's ρ is quite simple using the R Environment.

```
library(Hmisc);
##No Ties x <- c(1,2,3,4,5,6,7);

y <-c(1,3,6,2,7,4,5);

rcorr(x,y,type="spearman");
##Ties

x2 <- c(1,2,3,4,5,6,7);
y2 <- c(1,3,6,2,7,4,6);
rcorr(x2,y2,type="spearman");
```

Remarks on Spearman's Rank Correlation Coefficient

1. We always have $\Sigma d = 0$, which provides a check for numerical calculations.
2. Since Spearman's rank correlation coefficient, r_s, is nothing but Karl Pearson's correlation coefficient, r, between the ranks, it can be interpreted in the same way as the Karl Pearson's correlation coefficient.
3. Karl Pearson's correlation coefficient assumes that the parent population from which sample observations are drawn is normal. If this assumption is violated then we need a measure, which is distribution free (or non-parametric). Spearman's ρ is such a distribution free measure, since no strict assumption are made about the from of the population from which sample observations are drawn.
4. Spearman's formula is easy to understand and apply as compared to Karl Pearson's formula. The values obtained by the two formulae, viz Pearsonian *r* and Spearman's *rho* are generally different.

The difference arises due to the fact that when ranking is used instead of full set of observations, there is always some loss of information. Unless many ties exist, the coefficient of rank correlation should be only slightly lower than the Pearsonian coefficient.

5. Spearman's formula is the only formula to be used for finding correlation coefficient if we are dealing with qualitative characteristics, which cannot be measured quantitatively but can be arranged serially. It can also be used where actual data are given. In case of extreme observations, Spearman's formula is preferred to Pearson's formula.
6. Spearman's formula has its limitations also. It is not practicable in the case of bivariate frequency distribution. For $n > 30$, this formula should not be used unless the ranks are given.

12.3 Limitations of Correlation Analysis

Correlation analysis is a statistical tool, which should be properly used so that correct results can be obtained. Sometimes, it is indiscriminately used by management, resulting in misleading conclusions. We give below some *errors* frequently made in the use of correlation analysis:

1. Correlation analysis cannot determine cause-and-effect relationship.One should not assume that a change in Y variable is caused by a change in X variable unless one is reasonably sure that one variable is the cause while the other is the effect.

2. Another mistake that occurs frequently is on account of misinterpretation of the coefficient of correlation. Suppose in one case $r = 0.7$, it will be wrong to interpret that correlation explains 70 percent of the total variation in Y . The error can be seen easily when we calculate the coefficient of determination. Here, the coefficient of determination r^2 will be 0.49. This means that only 49 percent of the total variation in Y is explained. Similarly, the coefficient of determination is misinterpreted if it is also used to indicate causal relationship, that is, the percentage of the change in one variable is due to the change in another variable.

3. Another mistake in the interpretation of the coefficient of correlation occurs when one concludes a positive or negative relationship even though the two variables are actually unrelated. For example, the age of students and their score in the examination have no relation with each other. The two variables may show similar movements but there does not seem to be a common link between them.

12.3.1 The Coefficient of Correlation

The coefficient of correlation denoted by r, with $(-1 \leq r \leq +1)$ is a number that indicates both the direction and the strength of the linear relationship between the dependent variable (y) and the independent variable (x):

- **Direction of the relationship** If r is positive, y and x are directly related - i.e.,when x increases, y will tend to increase. If r is negative, y and x are inversely related - i.e., when x increases, y will tend to decrease.

- **Strength of the relationship** The larger the absolute value of r, the stronger the linear relationship between y and x. If $r = -1$ or $r = +1$, the regression line will actually include all of the data points and the line will be a perfect fit.

Calculating the Coefficient of Correlation for a set of data involves combining the same terms that appear in the table above. The formula for r can be expressed as follows:

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

$r =$ coefficient of correlation, $-1 \leq r \leq +1$

$n =$ number of data points

Example 12.9. A production manager has compared the dexterity-test scores of ve assembly-line employees with their hourly productivity. The data are in the table below.

Employee	$x =$ score on Dexterity Test	$y =$ units produced in one hour
A	12	55
B	14	63
C	17	67
D	16	70
E	11	51

Solution. Referring to the dexterity-test example above, the coefficient of correlation between productivity (y) and dexterity-test score (x) can be computed as

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

$$= \frac{5\,(4362) - (70)(306)}{\sqrt{5(1006) - (70)^2} \cdot \sqrt{5(18,984) - (306)^2}}$$

$$= \frac{390}{11.40175 \times 33.6749}$$

$$= 0.9546$$

The coefficient of correlation ($r = 0.9546$) is positive, reflecting that productivity (y) is directly related to dexterity-test score (x). In other words, persons scoring higher on the dexterity test tend to record higher levels of productivity. This is also reflected in the positive slope of the regression line, $\hat{y} = 19.2 + 3.0x$.

12.3.2 The coefficient of Determination

Another measure of the strength of the relationship is the coefficient of determination, r^2. Its numerical value is the proportion of the variation in y that is explained by the regression line, $\hat{y} = b_0 + b_1x$. For the dexterity-test example, $r = 0.9546$, $r^2 = 0.911$, and 91.1% of the variation in productivity is explained by the dexterity test score.

Coefficient of correlation is a unit free measure of degree of linear relationship between two or more variables. It is denoted by r. The square of correlation coefficient i.e., r^2 is called the **coefficient of determination**.

Coefficient of determination measures the amount of variation in one variable that can be accounted for in terms of variation in the other(s). For instance if $r = 0.90$ then $r^2 = 0.81$ which implies that 81% of variation in one variable can be attributed to variation in the other. Correlation coefficient is just that and no more. That is the fact that a variable is correlated highly with the other does not imply that one variable is dependent on the other. No causation is implied.

The most common coefficient of correlation is the pearsonian coefficient of correlation given by;

$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2\Sigma(y-\bar{y})^2}}$$

which for computational convenience is also given as:

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2]\,[n\Sigma y^2 - (\Sigma y)^2]}}$$

Example 12.10. The following data refers to examination marks verses hours of study per week of a sample of 8 candidates that sat for statistics examinations in 2000:

Exam mark: (Y)	64	61	84	70	88	92	72	72
Hours of study: (X)	20	16	34	23	27	32	18	22

(a) Calculate the Pearson's product moment coefficient of correlation.

(b) Calculate the coefficient of determination between examination marks and hours of study.

x	y	xy	x^2	y^2
20	64	1280	400	4096
16	61	976	256	3721
34	84	2856	1156	7056
23	70	1610	529	4900
27	88	2376	729	7744
32	92	2944	1024	8464
18	72	1296	324	5184
22	71	1562	484	5041
192	602	14,900	4902	46,206

$$n = 8, \Sigma x = 192, \Sigma y = 602, \Sigma xy = 14{,}900, \Sigma x^2 = 4902, \Sigma y^2 = 46{,}206$$

Hence;

$$r = \frac{(8)(14,900) - (192)(602)}{\sqrt{[(8)(4902) - (192)^2]\,[(8)(46206) - (602)^2]}}$$
$$= \frac{361.6}{421.697663}$$
$$= 0.88$$

Therefore $r = 0.88$.

12.3.3 Coefficient of Determination

$r^2 = (0.88)62 = 0.77$ i.e., 77% of variation in examination marks can be explained in terms of variation in hours of study.

12.3.4 Properties of r

(i). r lies between -1 and 4, i.e., $-1 \leq r \leq 1$.

(ii). $r = -1$ means perfect negative linear correlation.

(iii). $r = +1$ means perfect positive linear correlation.

(iv). $r = 0$ means no linear correlation.

Question 12.6.

The ratings below are based on collisions claim experience and theft frequency for 12 makes of small, twodoor cars. higher numbers reflect higher claims and more frequent thefts, respectively.

Collision	Theft	Collision	Theft
103	103	106	97
97	113	139	425
105	81	110	82
115	68	96	81
127	90	84	59
104	79	105	167

(i) Determine the least-squares regression line for predicting the rate of collision claims on the basis of theft frequency rating.

(ii) Calculate and interpret the values of r and r^2.

(iii) If a new model were to have a theft rating of 110, what would be the predicted rating for collision claims?

Question 12.7. The following data describes fuel consumption and flying hours for turboprop general aviation aircraft from 1992 through 1997. Fuel consumption is in millions of gallons, flying times is in millions of hours.

Year

	1992	1993	1994	1995	1996	1997
Fuel consumed:	131.8	95.9	92.7	124.4	145.0	135.7
Flying time:	1.6	1.2	1.1	1.5	1.8	1.7

(i) Determine the least-squares regression line for predicting fuel consumption on the basis of flying time. (ii) Determine and interpret the coefficients of correlation and determination.

(iii) If there were 2.0 million flying hours during a given year, what would be the prediction for the amount of fuel consumed?

12.4 Practice Problems

The yield of a particular crop on a farm is thought to depend principally on the amount of rainfall in the growing season. The values of the yield Y, in tonnes per acre, and the rainfall X, in centimetres, for seven successive years are given in the table below.

Rainfall (cm)	12.3	13.7	14.5	11.2	13.2	14.1	12.0
Yield (tonnes per acre)	6.25	8.02	8.42	5.27	7.21	8.71	5.68

Calculate the linear correlation coefficient and interpret your result.

Lecture 9: Regression Analysis

13 Regression Analysis

Objectives: *The overall objective of this lesson is to give you an understanding of linear regression, there by enabling you to understand the importance and also the limitations of regression analysis*

13.1 Introduction

Regression Analysis is a very powerful tool in the field of statistical analysis in predicting the value of one variable, given the value of another variable, when those variables are related to each other.

Regression analysis is a statistical tool used in prediction of value of unknown variable from known variable.

Regression analysis is the mathematical process of using observations to find the line of best fit through the data in order to make estimates and predictions about the behaviour of the variables. This line of best fit may be linear (straight) or curvilinear to some mathematical formula. Regression analysis determines the nature of the linear relationship between two interval-or ratio-scale variables.

In this lecture we will focus only on **simple regression** – linear regression involving only two variables, dependent variable and an independent variable. Regression analysis for studying more than two variables at a time is known as **multiple regressions**.

13.2 Independent and Dependent variables

Simple regression involves only two variables; one variable is predicted by another variable. The variable to be predicted is called the **dependent variable**. The predictor is called the **independent variable**, or explanatory variable. For example, when we are trying to predict the demand for television sets on the basis of population growth, we are using the demand for television sets as the dependent variable and the population growth as the independent or predictor variable.

The decision, as to which variable is which sometimes, causes problems. If we are unsure, here are some points that might be of use:

- if we have control over one of the variables then that is the independent. For example, a manufacturer can decide how much to spend on advertising and expect his sales to be dependent upon how much he spends
- it there is any lapse of time between the two variables being measured, then the latter must depend upon the former, it cannot be the other way round
- if we want to predict the values of one variable from your knowledge of the other variable, the variable to be predicted must be dependent on the known one

13.3 Assumptions of Regression

- Existence of actual linear relationship.

- The regression analysis is used to estimate the values within the range for which it is valid.
- The relationship between the dependent and independent variables remains the same till the regression equation is calculated.

- The dependent variable takes any random value but the values of the independent variables are fixed.

- In regression, we have only one dependant variable in our estimating equation. However, we can use more than one independent variable.

13.4 Simple Regression Analysis

If two variables are significantly correlated, and if there is some theoretical basis for doing so, it is possible to predict values of one variable from the other. This observation leads to a very important concept known as 'Regression Analysis'.

Regression analysis, in general sense, means the estimation or prediction of the unknown value of one variable from the known value of the other variable. It is one of the most important statistical tools which is extensively used in almost all sciences ï£¡ Natural, Social and Physical. It is specially used in business and economics to study the relationship between two or more variables that are related causally and for the estimation of demand and supply graphs, cost functions, production and consumption functions and so on.

Regression analysis was explained by M. M. Blair as follows: "Regression analysis is a mathematical measure of the average relationship between two or more variables in terms of the original units of the data."

We are interested in the nature of relationship between two or more variables. It is a usual practice to observe the actual series of data. The observed series is then plotted on a diagram which is called a **scatter plot** or **scatter diagram**.

13.5 Regression Line

Regression line is the line which gives the best estimate of one variable from the value of any other given variable.

The regression line gives the average relationship between the two variables in mathematical form. The Regression would have the following properties:

$$\text{(a) } \sum(Y - Y_c) = 0 \text{ and}$$

$$\text{(b) } \sum(Y - Y_c)^2 = \text{minimum}$$

The task of bringing out linear relationship consists of developing methods of fitting a straight line, or a regression line as is often called, to the data on two variables. The line of Regression is the graphical or relationship representation of the best estimate of one variable for any given value of the other variable. The nomenclature of the line depends on the independent and dependent variables.

For two variables X and Y, there are always two lines of regression: (a) Regression line of X on Y: gives the best estimate for the value of X for any specific given values of Y. (b) Regression line of Y on X: gives the best estimate for the value of Y for any specific given values of X.

If X and Y are two variables of which relationship is to be indicated, a line that gives best estimate of Y for any value of X, it is called **Regression line of Y on X**. If the dependent variable changes to X, then best estimate of X by any value of Y is called **Regression line of X on Y**.

13.5.1 Regression line of Y on X

The variable Y depends on variable X. We can find the value of Y if known the value of X. This is called regression of Y on X. The regression equation is $(Y - \bar{Y}) = b_{yx}(X - \bar{X})$.

$$b_{yx} = \text{Regression coefficient of } Y \text{ on } X = \frac{\text{Cov}(X,Y)}{\text{Var}(X)}$$

where;

$$\text{Cov}(X,Y) = \frac{1}{n}\sum(X - \bar{X})(Y - \bar{Y}) = \frac{1}{n}\sum XY - \bar{X}\bar{Y}$$
$$\text{Var}(X) = \frac{1}{n}\sum(X - \bar{X})^2 = \frac{1}{n}\sum X^2 - \bar{X}^2$$

Used to find X.

Example 13.1. Obtain the two regression equations and hence find the value of X when $Y = 25$.

X:	8	10	12	15	20
Y:	15	20	30	40	45

Solution.

$$n = 5, \quad \sum X = 65, \quad \sum Y = 150, \quad \sum X^2 = 933, \quad \sum Y^2 = 5150, \quad \sum XY = 2180$$

Now the two regression equations are;

$$(X - \bar{X}) = b_{xy}(Y - \bar{Y}) \qquad \cdots\cdots X \quad \text{on } Y \qquad (i)$$
$$(Y - \bar{Y}) = b_{yx}(X - \bar{X}) \qquad \cdots\cdots Y \quad \text{on } X \qquad (ii)$$

where;

$$\bar{X} = \frac{1}{n}\sum X = \frac{65}{5}, \quad \text{and} \quad \bar{Y} = \frac{1}{n}\sum Y = \frac{150}{5} = 30$$

Also;

$$\text{Cov}(X,Y) = \frac{1}{n}\sum XY - \bar{X}\bar{Y} = \frac{2180}{5} - 13 \times 30 = 436 - 390 = 46$$
$$\text{Var}(X) = \frac{1}{n}\sum X^2 - \bar{X}^2 = \frac{933}{5} - 13^2 = 186.6 - 169 = 17.6$$
$$\text{Var}(Y) = \frac{1}{n}\sum Y^2 - \bar{Y}^2 = \frac{5150}{5} - 30^2 = 1030 - 900 = 130$$

Now we find,

$$b_{yx} = \text{Regression coefficient of } Y \text{ on } X = \frac{\text{Cov}(X,Y)}{\text{Var}(X)} = \frac{46}{17.6} = 2.61$$

and

$$b_{xy} = \text{Regression coefficient of } X \text{ on } Y = \frac{\text{Cov}(X,Y)}{\text{Var}(Y)} = \frac{46}{130} = 0.35$$

Now substituting the values of $\bar{X}, \bar{Y}, b_{xy}, b_{yx}$ in the regression equations we get,

$$(X - 13) = 0.35(Y - 30) \qquad \cdots\cdots X \quad \text{on} \quad Y \qquad (i)$$
$$(Y - 30) = 2.61(X - \bar{X}) \qquad \cdots\cdots Y \quad \text{on} \quad X \qquad (ii)$$

as the two regression equations.

Now to estimate X when $Y = 25$, we use the regression equation of X on Y, therefore

$$(X - 13) = 0.35(25 - 30) \Rightarrow X = 13 - 1.75 = 11.25.$$

Remark: From the above example, we can note some points about regression coefficients.

- Both the regression coefficients carry the same sign (+ or −).
- Both the regression coefficients cannot be greater than 1 in number, (e.g., -1.25 and -1.32) is not possible.
- Product of both the regression coefficients b_{xy} and b_{yx} must be < 1, i.e., $b_{xy} \times b_{yx} < 1$. Here $0.35 \times 2.61 = 0.91 < 1$. ***(Check this always!)***

13.5.2 Regression line of X on Y

The variable X depends on variable Y. We can find the value of X if known the value of Y. This is called regression of X on Y. The regression equation is $(X - \bar{X}) = b_{xy}(Y - \bar{Y})$.

$$b_{xy} = \text{Regression coefficient of } X \text{ on } Y = \frac{\text{Cov}(X,Y)}{\text{Var}(Y)}$$

where;

$$\text{Cov}(X,Y) = \frac{1}{n}\sum(X-\bar{X})(Y-\bar{Y}) = \frac{1}{n}\sum XY - \bar{X}\bar{Y}$$
$$\text{Var}(Y) = \frac{1}{n}\sum(Y-\bar{Y})^2 = \frac{1}{n}\sum Y^2 - \bar{Y}^2$$

Used to find Y.

13.6 Regression Equation /Line and Method of Least Squares

In regression analysis an attempt is made to determine a line (curve) which best fits the given pair of data. In case of linear relationship as in the figures (a) and (b) above a line with the equation of the form $y = a+bx$ where a and b are determined such that

$$S = \Sigma(y - a - bx)^2 \qquad \text{is a minimum.}$$

with the use of differential Calculus S is minimized for a and b which satisfy the Least Squares Normal Equations.

$$\Sigma y = na + b\Sigma x \tag{8}$$
$$\Sigma xy = a\Sigma x + b\Sigma x^2 \tag{9}$$

Solving for a and b simultaneously yields the formula

$$\hat{b} = \frac{n\Sigma xy - \Sigma x\Sigma y}{n\Sigma x^2 - (\Sigma x)^2}$$
$$\hat{a} = \frac{1}{n}\left[\Sigma y - \hat{b}\Sigma x\right]$$
$$= \bar{y} - \hat{b}\bar{x}$$

The coefficient b in the equation $y = a + bx$ is called the **regression coefficient** of y on x. The regression coefficient b measures the linear relationship between the two variables x and y.

In geometrical terms b is a rate of change of y with respect to x. i.e., the slope of the line $y = a + bx$. The regression line can be used for interpolation and extrapolation.

Example 13.2. The following data gives the observations on weekly income and expenditure for food for five households.

Weekly Income (x)	240	270	300	330	360
Expenditure on food (y)	200	220	240	245	250

(i). Plot the data on a scatter diagram.

(ii). Determine the least squares regression line of expenditure on weekly income.

(iii). Using the equation in (ii). estimate the expenditure on food for some one having a weekly income of 380.

Solution. We need to fit a line $y = a + bx$ where a and b are to be determined from the data using the least squares method.

Prepare the following table;

x	y	xy	$x2$
240	200	48,000	57,600
270	220	59,400	72,900
300	240	72,000	90,000
330	245	80,850	108,900
360	250	90,000	129,600
Σ1500	Σ1155	350,250	459,000

Using the formula;

$$\begin{aligned}\hat{b} &= \frac{n\Sigma xy - \Sigma x\Sigma y}{n\Sigma x^2 - (\Sigma x)^2} \\ &= \frac{(5)(350,250) - (1500)(1155)}{(5)(459,000) - (1500)^2} \\ &= 0.42\end{aligned}$$

$$\begin{aligned}\hat{a} &= \frac{1}{n}\left[\Sigma y - \hat{b}\Sigma x\right] \\ &= \frac{1}{5}\left[1155 - (0.42)(1500)\right] \\ &= 105\end{aligned}$$

Thus the least square regression line is

$$y = 105 + 0.42x$$

For $x = 380$,

$$y = 105 + (0.42)(380) = 264.6$$

For someone with a weekly wage of 380 he is expected to spend 264.6 on food.

Regression analysis involving two variables as discussed above is known as **simple regression**. If more than one independent variable are involved we talk of **multiple regression**. If in particular the regression is believed to be linear, i.e., of the form;

$$y = a + b_1x_1 + b_2x_2 + \cdots + b_nx_n$$

We can determine simultaneous equations resulting from the least squares normal equations for estimating $a, b_1, b_2 \cdots, b_n$ for two independent variables i.e., three variables the equation is given by $y = a + b_1x_1 + b_2x_2$ and the least squares normal equations are given by;

$$\Sigma y = an + b_1\Sigma x_i + b_2\Sigma x \quad (10)$$

$$\Sigma x_1 y = a\Sigma x_1 + b_1\Sigma x_1^2 + b_2\Sigma x_1 x_2 \quad (11)$$

$$\Sigma x_2 y = a\Sigma x_1 + b_1 x_1 x_2 + b_2\Sigma x_1^2 \quad (12)$$

13.6.1 The Simple Linear Regression Model

Model and Assumptions

The simple linear regression model is a linear equation having a *y*-intercept and a slope, with estimates of these population parameters based on sample data and determined by standard formulas. The model is described in terms of the population parameters as follows:

The simple linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Where

y_i = a value of the dependent variable, y

x_i = a value of the independent variable, x

β_0 = the y-intercept of the regression line

β_1 = the slope of the regression line

ϵ_i = random error, or residual.

According to this model, the y-intercept for the population of (x_i, y_i) pairs is β_0 and the slope is β_1. The ϵ_i term is the random error, or residual, for the ith observation or measurement - this residual is the difference between the actual value (y_i) and the expected value, $\mu_{y|x} = \beta_0 + \beta_1 x_i$, from the regression line. The y values may be scattered above and below the regression line, but the *expected* value of y for a given value of x will be given by the linear equation, $\mu_{y|x} = \beta_0 + \beta_1 x_i$.

Three assumptions underline the simple linear regression model:

1. For any given value of x, the y values are normally distributed with a mean that is on the regression line $\mu_{y|x} = \beta_0 + \beta_1 x_i$.

2. Regardless of the value of x, the standard deviation of the distribution of y values about the regression line is the same. The assumption of equal standard deviations about the regression line is called *homoscedasticity*.

3. The y values are statistically independent of each other. For example, if a given y value happens to exceed $\mu_{y|x} = \beta_0 + \beta_1 x_i$, this does not affect the probability that the next y value observed will also exceed $\mu_{y|x} = \beta_0 + \beta_1 x_i$.

Based on the sample data, the y-intercept and slope of the population regression line can be estimated. This result is the sample regression line:

Sample regression line;

$$\hat{y} = b_0 + b_1 x$$

where;

$\hat{y}$ = the estimated value of the dependent variable (y) for a given value of x.

b_0 = the y-intercept; this is the value of y where the line intersects the y-axis whenever $x = 0$

b_1 = the slope of the regression line.

x = a value for the independent variable.

The cap $(\hat{\cdot})$ over the y indicates that it is an estimate of the (unknown) "true" value of y. The equation is completely described by the y-intercept (b_0) and slope (b_1), which are sample estimates of their population counterparts, β_0 and β_1, respectively. An infinite number of possible equations can be fitted to a given scatter diagram, and each equation will have a unique combination of values for b_0 and b_1. However, only one equation will be the "best fit" as defined by the least squares criterion we are going to use.

13.6.2 The Least-Squares method

The Least-Squares method requires that the sum of the squared deviations between y values in the scatter diagram and y values predicted by the equation be minimized. In symbolic terms:

Least-squares criterion for determining the best-fit equation: The equation must be such that $\sum (y_i - \hat{y}_i)^2$ is minimized. Where y_i = the observed value of y for the given value of x. $\hat{y}_i$ = the predicted value of y for that x value, as determined from the regression equation.

13.6.3 Determining the Least-Squares Regression Line

Equations have been developed for proceeding from a set of data to the least-squares regression line. They are based on the methods of calculus and provide values for b_0 and b_1 such that the least-squares criterion is met. The least-squares regression equation or as simply the regression line.

Least-Squares regression line, $\hat{y} = b_0 + b_1 x$:

Slope: $$b_1 = \frac{(\Sigma x_i y_i) - n\bar{x}\bar{y}}{(\Sigma x_i^2) - n\bar{x}^2}$$

y – intercept: $b_0 = \bar{y} - b_1 \bar{x}$

where n = number of data points. With the slope determined, we take advantage of the fact that the least-squares regression equation passes through the point $(\bar{x},\bar{y})$. The equation for finding the y-intercept ($b_0 = \bar{y} - b_1\bar{x}$) is just a rearrangement of $\bar{y} = b_0 + b_1\bar{x}$.

Example:

A production manager has compared the dexterity-test scores of ve assembly-line employees with their hourly productivity. The data are in the table below.

Employee	x = score on Dexterity Test	y = units produced in one hour
A	12	55
B	14	63
C	17	67
D	16	70
E	11	51

Fit the best regression line to the data.

Solution:

The calculations necessary for determining the slope and y-intercept of the regression equation are shown below.

Data and preliminary calculations:

Employee	x = score on Dexterity Test	y = units produced in one hour	x_iy_i	x^2_i	y_i^2
A	12	55	660	144	3025
B	14	63	882	196	3969
C	17	67	1139	289	4489
D	16	70	1120	256	4900
E	11	51	561	121	2601
	$\sum x_i$	$\sum y_i$	$\sum x_iy_i$	$\sum x_i^2$	$\sum y_i^2$
	70	306	4362	1006	18984

$$\bar{x} = \frac{70}{5} = 14.0 \qquad \bar{y} = \frac{306}{5} = 61.2$$

Calculations for slope and y–intercept of Least-squares Regression line

$$\text{slope}, b_1 = \frac{\Sigma x_i y_i - n\bar{x}\bar{y}}{\Sigma x_i^2 - n\bar{x}^2} = \frac{4362 - 5(14.0)(61.2)}{1006 - 5(14.0)^2}$$

$$= \frac{78.0}{26.0} = 3.0$$

$$y - \text{intercept}, b_0 = \bar{y} - b_1\bar{x} = 61.2 - 3.0(14.0)$$

$$= 61.2 - 42.0$$

$$= 19.2$$

The least squares regression line is $\hat{y} = 19.2 + 3.0x$

Where

$\hat{y}$ = estimated units produced per

hour

x = score on manual dexterity test.

13.6.4 Point Estimates Using the Regression Line

Making point estimates based on the regression line is simply a matter of substituting a known or assumed value of x into the equation, then calculating the estimated value of y.

For example, if a job applicant were to score $x = 15$ on the manual dexterity test, we would predict this person would be capable of producing 64.2 units per hour on the assembly line. This is calculated as

$$\hat{y} = 19.2 + 3.0(15) = 64.2 \text{ units/hour},$$

estimated productivity.

Exercise:

1. For a sample of 8 employees, a personnel director has collected the following data on ownership of company stock versus years with the firm.

x = years	y = shares
6	300
12	408
14	560
6	252
9	288
13	650
15	630
9	522

(a) Determine the least-squares regression line and interpret the slope.

(b) For an employee who has been with the firm 10 years, what is the predicted number of shares of stock owned?

2. The following data represents x = boats sales and y = boat trailer sales from 1995 through 2000.

Year	Boat sales (Thousands)	Boat Trailer sales (Thousands)
1995	649	207
1996	619	194
1997	596	181
1998	576	174
1999	585	168
2000	574	159

(a) Determine the least squares regression line and interpret its slope.

(b) Estimate, for a year during which 500,000 boats are sold, the number of boat trailers that would be sold.

(c) What reasons might explain why the number of boat trailer sold per year is less than the number of boats sold per year?

Question 13.1. Scores made by students in a statistics class in the mid-term and final examination are given in the table below. Develop a regression equation which may be used to predict final examination scores from the mid-term score.

Student:	1	2	3	4	5	6	7	8	9	10
Mid-term:	98	66	100	96	88	45	76	60	74	82
Final:	90	74	98	88	80	62	78	74	86	80

We want to predict the final exam scores from the mid term scores.So, we let y- be the final exam scores and x be the mid-term exam scores.

Thus, the regression equation is given by $y = 40.7531 + 0.5127x$.

Lecture 11: Introduction to Probability

14 Probability

The word *probability* denotes the chance or likelihood of the occurrence of an event. The theory of probability deals with laws governing the chances of occurrence of phenomena, which are unpredictable in nature. To understand the concept of probability and learn the methods of calculating the probabilities, we should first understand some basic terms and concepts related to probability.

14.1 Definitions of Terms

Random Experiment: An operation which can produce some well defined outcomes is known as an *experiment.* e.g. tossing a coin is an experiment because if a coin is tossed either a head or tail will turn up. i.e. we have two possible outcomes. An experiment whose outcome cannot be determined in advance is called a *random experiment.*

Experiment: An experiment is any activity where we do not know for certain what will happen but we will observe what happens. For example:

- we will ask someone whether or not they have used certain products.
- we will observe the temperature at midday tomorrow.
- we will toss a coin and observe whether it shows 11heads" or "tails".

Outcome of an experiment: The possible result of the random experiment is called **outcome**. For example, when we toss a coin, there are two possible outcomes Heads (H) and Tail T; or when we throw a cubic die the possible outcomes of the number of dots on the uppermost face are 1, 2, 3, 4, 5, or 6.

Sample Space: The set of all possible outcomes in a random experiment is called a **sample space**. It is denoted by S. The outcomes listed in the sample space are called the sample points or sample elements. The sample space may be finite, countable infinite or infinite in nature. The number of sample points in the sample space may be denoted by $n(S)$.

Examples: Some examples of the experiments and their sample spaces are as follows:

- In throwing a die, the sample space $S = \{1,2,3,4,5,6\}$, $n(S) = 6$ and typical sample elements are 2 and 3.

- In tossing a fair coin the sample space $S = \{H,T\}$, $n(S) = 2$.
- When tossing three coins together, the sample space is

$$S = \{HHH,HHT,HTH,HTT,THH,THT,TTH,TTT\}$$

Sample size is $n(S) = 8$.

Events: The results or outcomes of experiments are called *events.* i.e. an event is a subset of the sample space. e.g

- For the sample space in tossing two coins, the subset $A = \{HH,HT,TH\}$ is the event that atleast one head occurs.
- The subset $B = \{TTT\}$ for the sample space of throwing three coins is the event of getting three tails. The number of sample points in an event E is denoted by $n(E)$.

Simple (Elementary) Events: An event consisting of single sample point is called a **simple event**. The six events in S of throwing a die are;

$$\{1\},\{2\},\{3\},\{4\},\{5\},\{6\}.$$

In a simultaneous toss of two coins, the event $\{HH\}$ of getting both heads is a simple event where $S = \{HH,HT,TH,TT\}$.

Null Event: It is the event containing no sample point in it. It is the impossible happening and is denoted by $\emptyset$, e.g in the experiment of throwing a cubic die, where the sample space is $S = \{1,2,3,4,5,6\}$.
We define the event

Event A: The number of dots appearing is a two digit number.

$A = \{\}$, i.e. $A = \emptyset$, and $n(A) = 0$.

Impossible Events: An impossible event is an event which will never occur i.e. the event that includes no elementary events at all. e.g. the event of getting three heads in tossing of two coins is an impossible event.

Sure Events: An event which is sure to occur is called a sure event. In throwing a die an event consisting of number lying between 1 and 6 is a sure event.

Equally likely Events: A number of events are said to be equally likely if any one of them cannot be expected to occur in preference to the other. e.g in tossing a fair coin, the two possible outcomes head and tail are equally likely i.e. we have no reason to accept that heads will appear more often than tails or vice versa.

Independent Events: If the occurrence of one event does not change the probability that another event will occur, we say that the events are **independent**.

Dependent Events: If the occurrence of one event does change the probability that another event will occur, we say that the events are **dependent**.

Exhaustive Events: Events are said to be exhaustive when they include all possible outcomes of a random experiment. In tossing a coin, exhaustive events are two, i.e., H or T, in rolling a die, there are six exhaustive cases, since any of the six numbers may appear on top.

Mutually Exclusive Events: Two or more events are mutually exclusive or disjoint events if the events cannot occur simultaneously, in other words, the occurrence of one of the events prevents the occurrence of others. If A and B are any two events defined on a sample space S and $A \cap B = \emptyset$ i.e. there is no common sample points between them, then the events A and B are said to be mutually exclusive.

e.g. in tossing the coin the appearance of head and tail are mutually exclusive events as they cannot occur simultaneously.

Mutually Exclusive and Exhaustive Events: The two events A and B are said to be mutually exclusive and exhaustive if they are disjoint and their union is S.

For example: when a cubic die is thrown, the sample space is

$$S = \{1,2,3,4,5,6\}$$

Now if the events A and B defined on S have the sample points as follows

$$A = \{1,2,3\}, \text{ and } B = \{4,5,6\}, \text{ then } A \cap B = \emptyset, \text{ and } A \cup B = S$$

Hence A and B are mutually exclusive and exhaustive.

14.2 Probability of an Event

There are three main ways in which we can measure probability. Different people argue in favour of the different views of probability and some will argue that each has its uses depending on the circumstances.

14.2.1 Classical definitions

If all possible outcomes are "equally likely" then we can adopt the *classical approach to measuring probability*. For example, if we tossed a fair coin, there are only two possible outcomes, a head or a tail both of which are equally likely and hence

$$P(\text{Head}) = \frac{1}{2} \quad \text{and} \quad P(\text{Tail}) = \frac{1}{2}.$$

The underlying idea behind this view of probability is symmetry. In this example, there is no reason to think that the outcome Head and the outcome Tail have different probabilities and so they should have the same probability. Since there are two outcomes and one of them must occur, both outcomes must have probability
$\frac{1}{2}$.

Another commonly used example is rolling a dice. There are six possible outcomes (1, 2, 3, 4, 5, 6) when a dice is rolled and each of them should have an equal chance of occurring. Hence the $P(1) = 1/6$, $P(2) = 1/6$, ⋯.

Other calculations can be made such as P(Even number) = $\frac{3}{6} = \frac{1}{3}$. This follows from the formula;

$$P(\text{Event}) = \frac{\text{Total number of outcomes in which event occurs}}{\text{Total number of possible outcomes}}$$

Note that this formula only works when all possible events are equally likely–not a practical assumption for most real life situations.

Every event associated with a random experiment is assigned a weight or measure of the chance of it occurring called its *probability*.

The probability of an occurrence of an event A in a sample S, written as $P(A)$ is defined by

$$\begin{aligned} P(A) &= \frac{\text{Number of distinct sample elements in A}}{\text{Number of distinct sample elements in S}} \\ &= \frac{n(A)}{n(S)} \end{aligned}$$

The probability of non occurrence of the event A is

$$\begin{aligned} P(A') &= \frac{\text{Number of unfavorable outcomes}}{\text{Total number of all possible outcomes}} \\ &= \frac{n(S) - n(A)}{n(S)} = 1 - \frac{n(A)}{n(S)} \\ &= 1 - P(A) \end{aligned}$$

Hence $P(A) + P(A') = 1$.

Example 14.1. Two unbiased dice are thrown. Find the probability that

(i). getting a sum of 6

(ii). the numbers shown are equal

(iii). the difference of the numbers shown is 1

(iv). the first die shows 6

(v). the total of numbers greater than 8

Solution. The two dice can be thrown in $6 \times 6 = 36$ ways.

+	1	2	3	4	5	6
1	1	2	3	4	5	6
2	2	4	6	8	10	12
3	3	6	9	12	15	18
4	4	8	12	16	20	24
5	5	10	15	20	25	30
6	6	12	18	24	30	12

Hence $n(S) = 36$ where S is the sample space.

(i). Let E_1, be the event of getting a sum of 6. So

$$E_1 = \{(1,5),(5,1),(2,4),(4,2),(3,3)\}$$

and $n(E_1) = 5$. Hence, the required probability

$$P(E_1) = \frac{n(E_1)}{n(S)} = \frac{5}{36}$$

(ii). Let E_2, be the event of getting equal numbers. So

$$E_2 = \{(1,1),(2,2),(3,3),(4,4),(5,5),(6,6)\}$$

and $n(E_2) = 6$. Hence, the required probability

$$P(E_2) = \frac{n(E_2)}{n(S)} = \frac{6}{36} = \frac{1}{6}$$

(iii). Let E_3, be the event of showing the difference of numbers is 1. So

$$E_3 = \{(1,2),(2,1),(3,2),(2,3),(4,3),(3,4),(4,5),(5,4),(5,6)(6,5)\}$$

and $n(E_3) = 10$. Hence, the required probability

$$P(E_3) = \frac{n(E_3)}{n(S)} = \frac{10}{36} = \frac{5}{18}$$

(iv). Let E_4, be the event of the first die showing 6. So

$$E_4 = \{(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\}$$

and $n(E_4) = 6$. Hence, the required probability

$$P(E_4) = \frac{n(E_4)}{n(S)} = \frac{6}{36} = \frac{1}{6}$$

(v). Let E_5, be the event of the total numbers greater than 8. So

$$E_5 = \{(3,6),(6,3),(4,5),(5,4),(5,5),(4,6),(6,4),(5,6),(6,5),(6,6)\}$$

and $n(E_5) = 10$. Hence, the required probability

$$P(E_5) = \frac{n(E_5)}{n(S)} = \frac{10}{36} = \frac{5}{18}$$

Example 14.2. Three unbiased coins are tossed.

(a). Write the sample space S

(b). Find the probability of

(i). all heads

(ii). at least 2 heads

(iii). at most 2 heads

Solution. Let S denote the sample space of tossing three coins. Therefore

$$S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$$

$n(S) = 8$.

(i). Let E_1, be the set of favourable events, then $E_1 = \{HHH\}$ and $n(E_1) = 1$. Hence, the required probability

$$P(E_1) = \frac{n(E_1)}{n(S)} = \frac{1}{8}$$

(ii). Let E_2, be the set of favourable events, then

$$E_2 = \{HHT, HTH, THH, HHH\}$$

and $n(E_2) = 4$. Hence, the required probability

$$P(E_2) = \frac{n(E_2)}{n(S)} = \frac{4}{8} = \frac{1}{2}$$

(iii). Let E_3, be the set of favourable events, then

$$E_3 = \{HHT, HTH, THH, THT, TTT\}$$

and $n(E_3) = 7$. Hence, the required probability

$$P(E_3) = \frac{n(E_3)}{n(S)} = \frac{7}{8}$$

Question 14.1. Two fair dice labeled 1 to 6 are rolled. Let A be the event that the product of the two numbers showing up is greater than 21 and let B be the event that the product is divisible by 6. Find $P(A)$, $P(B)$, $P(A \cap B)$ and $P(A \cup B)$.

Solution. One can use the following table as shown below;

×	1	2	3	4	5	6
1	1	2	3	4	5	6
2	2	4	6	8	10	12
3	3	6	9	12	15	18
4	4	8	12	16	20	24
5	5	10	15	20	25	30
6	6	12	18	24	30	36

14.2.2 Frequentist approach

When the outcome of an experiment are not equally likely, we can conduct experiments to give us some idea of how likely the different outcomes are. For example, suppose we were interested in measuring the probability of producing a defective item in a manufacturing process. This probability could be measured by monitoring the process over a reasonably long period of time and calculating the proportion of defective items. In a more simple case, if we did not believe that a coin was fair, we could toss the coin a large number of times and see how often we obtained a head. In both cases we perform the same experiment a large number of times and observe the outcome. This is the basis of the frequentist view. By conducting experiments the probability of an event can easily be estimated using the following formula;

$$P(\text{Event}) = \frac{\text{Total number of times an event occurs}}{\text{Total number of times experiment is done}}$$

The larger the experiment, the closer this probability is to the "true" probability. The frequentist view of probability regards probability as the long run relative frequency (or proportion).

14.2.3 Subjective/Bayesian approach

(Subjectivist) A subjective probability is an individual's degree of belief in the occurrence of an event. *This will not be covered. Its beyond the scope of this course.*

14.3 Laws of Probability

The probability of an event A of sample space S must satisfy the following axioms:

1. $0 \leq P(A) \leq 1$ for every event A of S.
2. $P(S) = 1$
3. $P(A \cup B) = P(A) + P(B)$ for any two events A and B on S for which $A \cap B = \varphi$.
4. If $A_1, A_2, \cdots$ is a sequence of mutually exclusive events, then

$$P(A_1 \cup A_2 \cup \cdots) = P(A_1) + P(A_2) + \cdots$$

5. Addition rule. If A and B are two events,

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

6. For three events A, B and C applying the above rule, twice, we obtain

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

Theorem 1: $P(\emptyset) = 0$

Proof.

$$A \cup \emptyset = A \quad \Rightarrow P(A \cup \emptyset) = P(A) \tag{13}$$

But $A \cap \emptyset = \emptyset$. Hence by rule 3, we have

$$P(A \cup \emptyset) = P(A) + P(\emptyset) \tag{14}$$

From (13) and (14) we have;

$$P(A) = P(A) + P(\emptyset) \Rightarrow P(\emptyset) = 0$$

Theorem 2: $P(A^c) = 1 - P(A)$
Proof.

$$A \cup A^c = S$$

also

$$A \cap A^c = \emptyset$$

$$P(A \cup A^c) = P(S) = 1 \qquad \text{by rule 2.}$$

$$\Rightarrow P(A) + P(A^c) = 1 \qquad \text{by rule 3.}$$

$$P(A^c) = 1 - P(A).$$

Theorem 3: If $A \subset B$, then $P(A) \leq P(B)$.

Proof. $A^c \cap B = B \backslash A$ *(backslash means subtract only elements common to both sets)*.

If $A \subset B$, then $B = A \cup (B \backslash A)$.

$$\Rightarrow P(B) = P(A) + P(B \backslash A) \qquad \text{by rule 3.}$$

$$P(B) \geq P(A) \quad \text{given that} \quad P(B \backslash A) \geq 0$$

Theorem 4: $P(A \backslash B) = P(A) - P(A \cap B)$.

$$(A \backslash B) \cup (A \cap B) = A \; P \; [(A \backslash B) \cup (A \cap B)] = P(A)$$

$$P(A \backslash B) + P(A \cap B) = P(A).$$

Theorem 5: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ From theorem 4 above;

$$P(A \backslash B) = P(A) - P(A \cap B).$$

Hence;

$$A \backslash (B \cup B) = A \cup B$$

$$P[(A \backslash B) \cup B] = P(A \cup B)$$

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

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

This is the theorem of **total probability** or the **generalized addition theorem**.

1. $P(\emptyset) = 0$
2. $P(A^c) = 1 - P(A)$
3. If $A \subset B$, then $P(A) \leq P(B)$
4. $P(A \backslash B) = P(A) - P(A \cap B)$
5. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Theorem 1: $P(\emptyset) = 0$

Theorem 2: $P(A^c) = 1 - P(A)$ **Theorem 3:** If $A \subset B$, then $P(A) \leq P(B)$.

Theorem 4: $P(A\backslash B) = P(A) - P(A \cap B)$.

Theorem 5: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

This is the theorem of **total probability** or the **generalized addition theorem**.

Example 14.3. Given events A and B such that $P(A) = \frac{1}{3}$, $P(B) = \frac{1}{4}$ and $P(A \cap B) = \frac{1}{6}$. Find

(i) $P(A^c)$

(ii) $P(B^c \cap A)$

(iii) $P(A \cup B)$

(iv) $P(A^c \cap B^c)$

(v) $P[(A \cup B)^c]$

Solution. From the theorems;

(i) $P(A^c)$

$$P(A^c) = 1 - P(A) = 1 - \frac{1}{3} \quad \text{i.e.,} \quad P(A^c) = \frac{2}{3}$$

(ii) $P(B^c \cap A)$

$$\begin{aligned} A &= A\backslash B \cup (A \cap B) \\ &= (A \cap B^c) \cup (A \cap B) \\ P(A) &= P\left[(A \cap B^c) \cup (a \cap B)\right] \\ \frac{1}{3} &= P(A \cap B^c) + \frac{1}{6} \end{aligned}$$

Thus,

$$P(A \cap B^c) = \frac{1}{3} - \frac{1}{6} = \frac{1}{6}$$

(iii) $P(A \cup B)$

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ \frac{1}{3} + \frac{1}{4} - \frac{1}{6} &= \frac{5}{12} \end{aligned}$$

(iv) $P(A^c \cap B^c)$

By Morgan's law on sets

$$\begin{aligned} (A^c \cap B^c) &= (A \cup B)^c \\ P(A^c \cap B^c) &= 1 - P(A \cup B) \\ &= 1 - \frac{5}{12} = \frac{7}{12}. \end{aligned}$$

(v) $P[(A \cup B)^c]$

$$P[(A \cup B)^c] = 1 - P(A \cup B)$$
$$= 1 - \frac{5}{12} = \frac{7}{12}$$

14.4 Law of Total Probability

Consider the set space, S, being divided into a Partition of n mutually exclusive events, E_i, where i = $1,2,3,\cdots,n$, then:

$$E_i \cup E_j = \emptyset, \quad \textbf{and}$$

$$E_1 \cup E_2 \cup E_3 \cup \cdots \cup E_n = S$$

For example, a partition of our dice sample space $S = \{1,2,3,4,5,6\}$ could be:

$$E_1 = \{1\}, E_2 = \{2,3\}, E_3 = \{4,5,6\}$$

There are many other probabilities, as long as the E_i's are mutually exclusive (i.e., there is no overlap between them) and together they make up the whole sample space (i.e., they are exhaustive) it is a partition.

And for any $A \subset S$:

$$A = (A \cap E_1) \cup (A \cap E_2) \cup (A \cap E_3) \cup \cdots \cup (A \cap E_n)$$

For example, if A is the event "roll an even number" we have $A = \{2,4,6\}$. This can be made up of all the intersections with our partition from above:

$$A \cap E_1 = \{\emptyset\} \; A \cap E_2 = \{2\}$$

$$A \cap E_3 = \{4,6\}$$

Hence

$$A = \{2,4,6\} = \emptyset \cup \{2\} \cup \{4,6\}$$

Therefore

$$P(A) = P(A \cap E_1) + P(A \cap E_2) + P(A \cap E_3) + \cdots + P(A \cap E_n)$$
$$= \sum_{j=1}^{n} P(A \cap E_j)$$

This result is known as the law of total probability.

14.5 Conditional Probability

Consider two events, A and B. We might wish to know the probability that event A occurred, given the occurrence of the event B. This is known as a conditional probability and is denoted thus:

$$P(A|B)$$

$P(A|B)$ is read as "the probability of event A occurring given that event B has already occurred" or "probability of A given B" for short. This is called a **conditional probability** as the probability depends (i.e., is conditional) on event B.

The conditional probability of A occurring given B can be expresses as:

$$P(A|B) = \frac{P(A \cap B)}{P(A \cap B) + P(A' \cap B)}$$

($A \cap B$) and $A^0 \cap B$) are mutually exclusive. Noting that ($A \cap B$) $\cup$ ($A^0 \cap B$) = B, this can be rearranged thus:

$$P(A|B) = \frac{P(A \cup B)}{P(B)}$$

This is the version of the formula that is commonly used in practice.
Example 14.4. Given $P(A) = 0.6$, $P(B) = 0.5$ and $P(A \cap B) = 0.4$. Find $P(A \cup B)$; $P(A|B)$; $P(B|A)$.
Solution.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
$$= 0.6 + 0.5 - 0.4$$
$$= 1.1 - 0.4 = 0.7$$

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.4}{0.5} = \frac{4}{5}$$

$$P(B|A) = \frac{P(A \cap B}{P(A)} = \frac{0.4}{0.6} = \frac{2}{3}$$

Example 14.5. For two events A, and B; $P(A) = \frac{2}{5}$, $P(B') = \frac{1}{3}$, $P(A \cup B) = \frac{5}{6}$*P*(only *A*); *P*(only one).

Solution.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
$$\frac{5}{6} = \frac{2}{5} +$$

14.5.1 The Multiplicative Rule

Rearranging the formula for conditional probabilities, we obtain:

$$P(A \cap B) = P(B) \cdot P(A|B).$$

Due to the symmetry, one can rewrite this as:

$$P(A \cap B) = P(A) \cdot P(B|A).$$

. Find $P(A \cap B)$; $P(A|B)$;
Events A and B are said to be independent if whether or not event B has occurred gives us no information on whether event A has occurred. This can be expressed algebraically as follows:

A and B are independent if *If*$P(A) = P(A|B) = P(A|B^0)$

Given that:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Then if A and B are independent

$$P(A \cap B) = P(A) \cdot P(B)$$

This is a special case of the multiplication rule when events A and B are independent.

Example 14.6. A friend tosses a coin three times. You accidentally notice that the first time the coin shows Head. What is the chance that the friend observes 2 Heads?

Solution. From the experiment, the sample space is given by

$$S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}, \quad n(S) = 8$$

When the first outcome is a Head, we have

$$\{HHH, HHT, HTT\ HTH\}$$

The two of them are;

$$\{HHT, HTH\}$$

Hence, the chance = $\frac{2}{4} = \frac{1}{2}$

Example 14.7. A pack of 52 cards has 13 spades. If two cards are drawn from the deck at random, what is the chance that both are spades?

Solution. Let $A-$ be the event that the first card is a spade and B- be the event that the second card is a spade.

$$P(A) = \frac{13}{52} = \frac{1}{4}.$$

Example 14.8. Let $P(A) = 0.8$ and $P(B) = 0.7$. Find $P(A \cap B)$ and $P(A \cup B)$ if A and B are;

(i). Independent events

(ii). Mutually exclusive events

Solution. (i). If the two events are independent, then

$$P(A \cap B) = P(A) \times P(B)$$
$$= 0.8 \times 0.7$$
$$= 0.56$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
$$= 0.8 + 0.7 - 0.56$$
$$= 0.94$$

(ii). If the two events are mutually exclusive events, then

$$P(A \cap B) = 0$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
$$= 0.8 + 0.7 - 0$$
$$= 1.5 \text{ This value is not a probability since } (0 \le P \le 1).$$

Hence events A and B are not mutually exclusive.

Example 14.9. Among 1000 applicants for admission to MBA course in a university, 600 were economic graduates and 400 were non-economic graduates; 30% of economics graduate applicants and 5% of noneconomic graduate applicants obtained admission. If an applicant selected at random is found to have been given admission, what is the probability that he/she is an economics graduate?

Solution. Let A be the event that the applicant selected at random is an economics graduate and B be the event that he/she is given admission.

$$n(S) = 1000, \quad n(A) = 600, \quad n(\bar{A}) = 400$$

Also

$$n(B) = \frac{600 \times 30}{100} + \frac{400 \times 5}{100} = 200$$

and

$$n(A \cap B) = \frac{600 \times 30}{100} = 180.$$

Thus, the required probability is given by

$$P(A|B) = \frac{n(A \cap B)}{n(B)} = \frac{180}{200} = 0.9.$$

Example 14.10. If A and B are two events such that $P(A) = 2/3$, $P(\bar{A} \cap B) = \frac{1}{6}$ and $P(A \cap B) = \frac{1}{3}$. Find $P(B)$, $P(A \cup B)$, $P(A|B)$, $P(B|A)$, $P(A^c \cup B)$, $P(A^c \cap B^c)$ and $P(B^c)$.

Also examine whether the events A and B are (a) Equally likely (b) Exhaustive (c) Mutually exclusive and (d) Independent.

Solution. The probabilities of various events are obtained as follows:

$$P(B) = P(\bar{A} \cap B) + P(A \cap B) = \frac{1}{6} + \frac{1}{3} = \frac{1}{2}.$$
$$P(A \cup B) = \frac{2}{3} + \frac{1}{2} - \frac{5}{6}$$
$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1}{3} \times \frac{2}{1} = \frac{2}{3}$$
$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{1}{3} \times \frac{3}{2} = \frac{1}{2}$$
$$P(\bar{A} \cup B) = P(\bar{A}) + P(B) - P(\bar{A} \cap B) = \frac{1}{3} + \frac{1}{2} - \frac{1}{6} = \frac{2}{3}.$$
$$P(\bar{A} \cap \bar{B}) = 1 - P(A \cup B) = 1 - \frac{5}{6} = \frac{1}{6}$$
$$P(\bar{B}) = 1 - P(B) = 1 - \frac{1}{2} = \frac{1}{2}.$$

(a) Since $P(A)$ 6= $P(B)$, A and B are not equally likely events.

(b) Since $P(A \cup B)$ 6= 1, A and B are not exhaustive events.

(c) Since $P(A \cap B)$ 6= 0, A and B are not mutually exclusive.

(d) Since $P(A)P(B) = P(A \cap B)$, A and B are independent events.

Example 14.11. Probability that an electric bulb will last for 150 days or more is 0.7 and that it will last at the most 160 days is 0.8. Find the probability that it will last between 150 and 160 days.

Solution. We are given, $P(A) = 0.7$ and $P(B) = 0.8$. Thus, $P(A \cup B) = 1$. We have to find $P(A \cap B)$.
This probability is given by

$$P(A \cap B) = P(A) + P(B) - P(A \cup B) = 0.7 + 0.8 - 1.0 = 0.5.$$

Question 14.2. Explain the meaning of conditional probability. State and prove the multiplication rule of probability of two events when

(a) they are not independent

(b) they are independent

Question 14.3. If $P(A) = \frac{1}{3}$, $P(B) = \frac{1}{2}$, $P(A|B) = \frac{1}{6}$, find $P(B|A)$ and $P(B|\bar{A})$.

Question 14.4. In a group of 80 students, 30 are taking mathematics, 20 are taking chemistry, and 10 are taking mathematics and chemistry. What is the probability that a randomly chosen student is taking either mathematics or chemistry?

14.5.2 Bayes' Theorem

Bayes' Theorem, named after the English mathematician Thomas Bayes (1702-1761), is an important formula that provides an alternative way of computing conditional probabilities.

Let A and B be two events. The probability that event B will occur, given that event A has occurred, is known-this is a "forward-looking" probability in the sense that event A occurred before event B. Suppose instead that you were asked to find the "backward-looking" probability that event A has occurred, given that event B has occurred. In other words, you are asked to find

$$P(A|B).$$

Bayes' Theorem gives a way to find this conditional probability by using the formula. The conditional probability that an event A has occurred, given that event B has occurred, is

$$P(A|B) = \frac{P(A) \cdot P(B|A)}{P(A) \cdot P(B|A) + P(A') \cdot P(B|A')}$$

Example 14.12. Two machines, A and B produce identical beads. Machine A has probability 0.1 of producing a defective bead each time, whereas machine B has probability 0.4 of producing a defective bead. Each machine produces one bead. One of these beads is selected at random, tested and found to be defective. What is the probability that it was produced by machine B?

Solution. The probabilities can be represented on a tree diagram. Let A be the event that the bead was produced by machine A, and B be the event that the bead was produced b machine B. Then,

$$\begin{aligned} P(D) &= P(A) \cdot P(D|A) + P(B) \cdot P(D|B) \\ &= (0.5 \times 0.1) + (0.5 \times 0.4) \\ &= 0.05 + 0.2 \\ &= 0.25 \end{aligned}$$

$$\begin{aligned} P(B|D) &= \frac{P(B \cap D)}{P(D)} = \frac{P(B) \cdot P(D|B)}{P(D)} \\ &= \frac{(0.5 \times 0.4)}{0.25} \\ &= 0.8 \end{aligned}$$

Example 14.13. Analysis of questionnaire completed by holiday makers showed that 0.75 classified their holiday as good at *Malindi.* The probability of hot weather in the resort is 0.6. If the probability of regarding the holiday as good given hot weather is 0.9, what is the probability that there was hot weather if a holiday maker considers his holiday good?

Solution. Let

H = Hot weather

G = Good holiday

Then

$$P(G) = 0.75, P(H) = 0.6, P(G|H) = 0.9$$

What we want to get is

$$P(H|G) = ?$$

Recall

$$P(G|H) = \frac{P(G \cap H)}{P(H)} \text{ and } P(H|G) = \frac{P(G \cap H)}{P(G)}$$

$$P(G|H) \cdot P(H) = P(H|G) \cdot P(G)$$

$$\begin{aligned} P(H|G) &= \frac{P(G|H) \cdot P(H)}{P(G)} \\ P(H|G) &= \frac{0.9 \times 0.6}{0.75} \\ &= 0.72 \end{aligned}$$

Example 14.14. Three machines A, B, and C produce respectively 60%, 30% and 10% of the total number of items of a factory. The percentages of defective output of these machines are respectively 2%, 3%, and 4%. An item is selected at random from the product and is found to be defective. Find the probability that the item was produced by machine C.

Solution. Let A_1, B_1, C_1 be the events that an item drawn at random is produced by machines A, B, C respectively and let X be the event that the product drawn is defective.

Then

$P(A_1)$ = probability that the product drawn was manufactured by A = 0.60

$P(B_1) = 0.30$

$P(C_1) = 0.10$

$P(X|A_1)$ = probability of drawing a defective item from the product manufactured by A = 0.02

$P(X|B_1) = 0.03$

$P(X|C_1) = 0.04$

We seek $P(C_1|X)$ = probability that an item is produced by machine C given that the item is defective.

By Bayes' Theorem

$$\begin{aligned} P(C_1|X) &= \frac{P(C_1) \cdot P(X|C_1)}{P(A_1)P(X|A_1) + P(B_1)P(X|B_1) + P(C_1)P(X|C_1)} \\ &= \frac{(0.10) \cdot (0.04)}{(0.60)(0.02) + (0.30)(0.03) + (0.10)(0.04)} \\ &= 0.08 \end{aligned}$$

Example 14.15. Suppose there are 3 urns containing 2 white and 3 black balls, 3 white and 2 black, and 4 white and 1 black balls respectively. There is equal probability of each urn being chosen. One ball is drawn from an urn chosen at random. What is the probability that a white ball is drawn?

Solution. Let A_i be the event that i^{th} urn is chosen $i = 1,2,3$ and $P(A_i) = \frac{1}{3}$ and B the event that a white ball is drawn.

Then

$$P(B) = P(A_1) \cdot P(B|A_1) + P(A_2) \cdot P(B|A_2) + P(A_3) \cdot P(B|A_3)$$

Here;

$$P(A_1) = P(A_2) = P(A_3) = \frac{1}{3} \text{ (each urn being equally likely to be chosen)}$$

$$P(B|A_1) = \frac{2}{5}, \qquad P(B|A_2) = \frac{3}{5}, \qquad P(B|A_3) = \frac{4}{5}$$

$$\begin{aligned} P(B) &= \frac{1}{3}\left[\frac{2}{5} + \frac{3}{5} + \frac{4}{5}\right] \\ &= \frac{3}{5} \end{aligned}$$

Bayes' Theorem can be generalized to include any number of mutually exclusive events whose union is the entire sample space. For instance, suppose in Figure C.16 that the events are mutually exclusive and that Then the conditional probability that the event has occurred, given that event has occurred, is

Reverend Thomas Bayes (1702-1761), introduced his theorem on probability which is concerned with a method for estimating the probability of causes which are responsible for the outcome of an observed effect.

Bayes' theorem makes use of conditional probability formula where the condition can be described in terms of the additional information which would result in the revised probability of the outcome of an event.

The result is as follows:

Let $E_1, E_2, E_3, \cdots, E_n$ be a partition of S and define event $A \subset S$. Using the conditional probability:

$$P(E_i|A) = \frac{P(E_i \cap A)}{P(A)} \qquad (15)$$

also; the relationship:

$$P(E_i \cap A) = P(A \cap E_i) = P(E_i)P(A|E_i)$$

i.e., using $P(E_i \cap A) \equiv P(A \cap E_i)$ and then the Multiplication rule for $P(A \cap E_i)$ and the law of total probability

$$P(A) = \sum_{j=1}^{n} P(A \cap E_j)$$

then, by substituting for $P(E_i \cap A)$ and $P(A)$ in Eqn (15), the result is:

$$P(E_j|A) = \frac{P(E_i)P(A|E_i)}{\sum_{j=1}^{n} P(E_j)P(A|E_j)}, \qquad i = 1, 2, 3, \cdots, n$$

Essentially Baye's formula allows us to "turnaround" conditional probabilities, i.e., calculate $P(E_i|A)$ given only information about $P(A|E_i)$.

The value $P(E_j)$ are known as **prior probabilities**, the event A is some event which is known to have occurred and the conditional probability $P(E_i|A)$ is known as the **posterior probability**.

Bayes' theorem is frequently used in the analysis of decisions using decision trees where information is given in the form of conditional probabilities and the reverse of these probabilities must be found.

The general form of Bayes' Rule is

$$P(A|B) = \frac{P(A) \times P(B|A)}{P(B)}$$

Example 14.16. In a test, an examinee (student) either guesses or copies or knows the answer to multiple choice question with four choices, only one answer being correct. The probability that he makes a guess is $\frac{1}{3}$and the probability that he copies the answer is$\frac{1}{6}$. The probability that his answer is correct given that

he copies it is$\frac{1}{8}$. Find the probability that he knew the answer to the question given that he correctly answers it.

Solution. Let A be the event of answering by guess work, B be the event of answering by copying, C be the event of answering by knowing and D be the event of answering correctly.

$$P(A) = \frac{1}{3}; \qquad P(B) = \frac{1}{6}$$

As the question is answered by either guessing or copying or knowing

$$P(C) = 1 - P(A) - P(B) = 1 - \frac{1}{3} - \frac{1}{6} = \frac{1}{2}$$

$P(D|A)$ = probability of answering correctly by guessing

$= \frac{1}{4}$ (there are 4 choices and only one of them is correct)

$P(D|B)$ = probability of answering correctly by copying

$= \frac{1}{8}$ (given)

$P(D|C)$ = probability of answering correctly by knowing

= 1 (if he knows, he certainly answers correctly)

We have to find the probability that he knew when he answered correctly, i.e. $P(C|D)$

By Baye's theorem,

$$\begin{aligned} P(C|D) &= \frac{P(C) \cdot P(D|C)}{P(A)P(D|A) + P(B)P(D|B) + P(C)P(D|C)} \\ &= \frac{(0.5) \cdot (1)}{(1/3)(1/4) + (1/6)(1/8) + (1/2)(1)} \\ &= \frac{24}{29} \end{aligned}$$

14.6 Tree Diagrams

Many probability problems can be simplified by using a device called a **probability tree**. This is especially (but not exclusively) true in situations where actions are taken or decisions are made sequentially. Probability trees provide a systematic method of generating the elements of a suitable sample space and determining their probabilities. A tree diagram is a convenient representation of probabilities.

Tree diagrams can be a useful device for keeping track of conditional probabilities when using multiplication and partition rules. The idea is to draw a tree where each path represents a sequence of events. On any given branch of the tree we write the conditional probability of that event given all the

events on branches leading to it. The probability at any node of the tree is obtained by multiplying the probabilities on the branches leading to the node, and equals the probability of the intersection of the events leading to it.

Constructing the Tree for a Sequential Process: Probability trees may be shown growing from left to right or from top to bottom. The root of the tree corresponds to the starting point of the process. Line segments called branches connect the root to nodes representing the different outcomes that are possible at the first stage of the process. Each of those stage 1 nodes is connected to nodes representing the possible outcomes at the next stage, and the process continues until all stages are completed.

Consider and example where you have three (3) red balls, two (2) green balls and two (1) white ball in an urn. One ball is chosen randomly from the urn. If a red or white ball is chosen, a fair coin is flipped once. If the ball is green, the coin is flipped twice. We can construct a tree to enumerate the outcomes (the elements of the sample space) for this random process In the tree, the letters R, G and W represent the colors red, green and white, and the letters H and T represent heads and tails, respectively

14.6.1 Using the Probability Tree Diagram to Calculate Probabilities

Probabilities of Individual Outcomes: To determine the probability of an outcome, multiply the probabilities along its path. (This is a consequence of the Multiplicative Law of Probability.) For example, the probability of drawing a red ball followed by a tail is $(3/6)(1/2) = 1/4$, and the probability of drawing a green ball followed by two heads is $(2/6)(1/2)(1/2) = 1/12$.

Probabilities of Compound Events: Since different paths in a probability tree represent mutually exclusive events, we can add their probabilities without worrying about any overlap to find the probability of a compound event. (This is a consequence of the Additive Law of Probability.)

14.6.2 Drawing the Tree Diagram under Different Schemes

1. Drawing without Replacement

Example 14.17. Suppose that a bag contain 5 beads, 2 of them red and 3 blue. We randomly draw two beads, one after the other without replacing the first bead. Let R be the event that a red bead is picked and B be the event that a blue bead is picked. The tree diagram will have two stages (trials) each with two branches having changing probabilities as shown below:

2. Drawing with Replacement Suppose that there is replacement of the first bead before the second one is picked. The tree diagram will be similar to the first one but the probabilities will not change as the tree grows as shown below:

3. Drawing with Conditional Replacement Suppose that the replacement of the first bead is done only if it is blue, then the tree diagram will be similar as before but the probabilities will change only when a red bead is picked as shown below.

Question 14.5. A tourist decides between two plays, called "Good" (G) and "Bad" (B). The probability of the tourist choosing Good is $P(G) = 10\%$. A tourist choosing Good likes it (L) with 70% probability ($P(L \mid G) = .7$) while a tourist choosing Bad dislikes it with 80% probability ($P(D \mid B) = 0.8$).

a. Draw a probability decision tree diagram to illustrate the choices.

b. Calculate $P(L)$, the probability that the tourist liked the play he or she saw.

c. If the tourist liked the play he or she chose, what is the probability that he or she chose Good?

The use of tree diagrams may become tedious when the tree grows beyond four stages. We can make use of the binomial formula which will be discussed in the next two lectures. **Exercise Problems**

1. The probability that the judge selected to try a criminal case will arrive at the appropriate verdict is 0.95. That is, given a guilty defendant on trial, the probability is 0.95 that the judge will find him guilty and conversely, given an innocent man on trial, the probability is 0.95 that the judge will find him innocent. Suppose that the local police is quite diligent in its duties and that 99% of the people brought before the court are actually guilty. Find:

   (a) the probability that a defendant is found innocent.

   (b) the defendant is innocent, given that the judge finds him guilty.

   (c) the defendant is guilty, given that the judge finds him guilty.

2. A firm has four plants scattered around the city producing the same item. Plant A produces 30% of total production, plant B produces 25%, plant C produces 35%, and plant D produces 10%. The firm has a single warehouse in the city for storing finished products from all the plants. From the past performance records on the proportion of defectives, it has been found that 5%, 10%,15%, and 20 % of the items produced at A,B, C, and D respectively are defectives. before the shipment of an item to a dealer, one unit is selected at random and found to be defective. What is the probability that it was produced by plant C?

3. Three girls, Aileen, Barbara, and Cathy, pack biscuits in a factory. From the batch allotted to them Aileen packs 55%, Barbara 30%, and Cathy 15%. The probability that Aileen breaks some biscuits in a packet is 0.7, and the respective probabilities for Barbara and Cathy are 0.2 and 0.1. What is the probability that a packet with broken biscuits found by the checker was packed by Aileen?

4. When a person needs a mini-cab, it is hired from one of three firms, X, Y, and Z. Of the hirings 40% are from X, 50% are from Y and 10% are from Z. For cabs hired from X, 9% arrive late, the corresponding percentages for cabs hired from firms Y and Z belong 6% and 20% respectively. Calculate the probability that the next cab hired,

   (a) Will be from X and will not arrive late.

   (b) will arrive late.

5. In a bolt factory machines A, B, and C manufacture respectively 25%, 30%, and 45% of the bolts. The respective percentage defective bolts for machines A, B, and C are 2,1, and 0.5. A bolt is drawn at random from the production of this factory and is found to be defective. What is the probability that it was manufactured by machine C?

6. In the town of Corruptaville in the country of Burania 30% of the drivers are learners, 50% of the drivers are licensed but incompetent and bribed the examiner to make them pass. The remaining

20 % of drivers are licensed, competent and did not bribe the examiner. 10% of the learners, 80% of the incompetent drivers and 1% of the competent drivers drive carelessly. The probability that any careless driver has an accident is 70% and the probability that any careful driver has an accident is 20%. The police have been called to check up on an accident involving a driver and a pedestrian. What is the probability, correct to three significant figures, that

(a) the driver is incompetent?

(b) the driver is careful?

7. In a study made to find the relationship between IQ of a person and his academic achievements the following results were obtained.

   (a) proportion of graduates is 0.7
   (b) proportion of persons with IQ over 115 among graduates is 0.3.
   (c) proportion of persons with IQ above 115 among non-graduates is 0.2 given that a person has IQ above 115.

   Find the probability that the person is not a graduate.

8. A factory manufacturing memory card chips has three machines A, B, and C in operation. Machine A produces 50% of the cards and is known to have a rate of 3% defectives. Machine B produces 30 % of the memory cards with the rate of 4% dfective and machine C produces 20% of the cards with the rate of 5% defectives.

   An item selected at random from the memory cards manufactured by the company was found to be defective, what is the probability that it's defective? and what machine produced the card?

9. A box contains 12 light bulbs of which 5 are defective. All the bulbs look alike and have equal probability of being chosen. Three bulbs are picked up at random. What is the probability that at least 2 are defective ?

10. If you take a bus to work in the morning there is a 20% chance you'll arrive late. When you go by bicycle there is a 10% chance you'll be late. 70% of the time you go by bike, and 30% by bus. Given that you arrive late, what is the probability you took the bus?

11. At a police spot check, 10% of cars stopped have defective headlights and a faulty muffler. 15 % have defective headlights and a muffler which is satisfactory. If a car which is stopped has defective headlights, what is the probability that the muffler is also faulty?

12. In a large population, people are one of 3 genetic types A, B and C: 30% are type A, 60% type B and 10% type C. The probability a person carries another gene making them susceptible for a disease is .05 for A, .04 for B and .02 for C. If ten unrelated persons are selected, what is the probability at least one is susceptible for the disease?

13. Let A and B be events defined on the same sample space, with $P(A) = 0.3$, $P(B) = 0.4$ and $P(A|B) = 0.5$. Given that event B does not occur, what is the probability of event A?

14. Events A and B are independent with $P(A) = .3$ and $P(B) = .2$. Find $P(A \cup B)$.

15. Students A, B and C each independently answer a question on a test. The probability of getting the correct answer is .9 for A, .7 for B and .4 for C. If 2 of them get the correct answer, what is the probability C was the one with the wrong answer?

16. E and F are two events such $P(E) = 0.60$, $P(EorF) = 0.90$ and $P(EandF) = 0.50$. Find $P(F)$.

17. The probability that a randomly chosen adult resident of Kisumu city owns a boat is 0.16. The probability that a randomly chosen adult rents an apartment is 0.30. The probability that the adult owns a boat given he/she rents an apartment is 0.20. Find the probability that a randomly chosen adult rents an apartment and owns a boat

18. Assume that the probability is 95% that the jury selected to try a criminal case will arrive at the correct verdict whether innocent or guilty. Further, suppose that the local police force is quite diligent in performing its function, that 99% of the people brought to trial are in fact guilty. Given that a jury finds a defendant innocent, what is the probability that he is in fact innocent? Draw a tree diagram.

19. Medical researchers know that the probability of getting lung cancer if a person smokes is 0.34. The probability that a nonsmoker get lung cancer is 0.03. It is also known that 11% of the population smokes. What is the probability that a person with lung cancer was a smoker?
20. A sandwich is made with only one type of bread, one type of meat, and one type of cheese. There are 3 types of bread: white, wheat, or rye; 2 types of meat: turkey or roast beef; and 2 types of cheese: American or Swiss. Draw a tree diagram to show the number of sandwich choices.

21. A drawer contains 4 red socks, 3 white socks, and 3 blue socks. Without looking, you select a sock at random, replace it, and select a second sock at random. What is the probability that the first sock is blue and the second sock is red?

22. Two urns each contain green balls and blue balls. Urn I contains 4 green balls and 6 blue balls, and Urn II contains 6 green balls and 2 blue balls. A ball is drawn at random from each urn. What is the probability that both balls are blue?

23. A bag contains 6 purple marbles and 7 white marbles. Two marbles are drawn at random. One marble is drawn and not replaced. Then a second marble is drawn. What is the probability that the first marble is white and the second one is purple?

24. **Radar detection.** If an aircraft is present in a certain area, a radar correctly registers its presence with probability 0.99. If it is not present, the radar falsely registers an aircraft presence with probability 0.10. We assume that an aircraft is present with probability 0.05. What is the probability of false alarm (a false indication of aircraft presence), and the probability of missed detection (nothing registers, even though an aircraft is present)?

25. A company has installed a new computer system and some employees are having difficulty logging on to the system. They have been given training and the problems which arose during training were recorded and their probabilities calculated as follows:

- An employee has a probability of 0.9 of logging on successfully on the first attempt.
- If the employee logs in successfully then the employee will also be successful on each later attempt with probability 0.9.
- If the employee tries to log in and is not successful then the employee loses confidence and the probability of a successful log-in on later occasions drops to 0.5.

Use a tree diagram to find the following probabilities:

(a) An employee successfully logs on in each of the first three attempts.

(b) An employee fails in the first attempt but is successful in the next two attempts.

(c) An employee logs on successfully only once in three attempts.

(d) An employee does not manage to log on successfully in three attempts.

Lecture 12: Random Variables, Expected Value and Variance

15 Random Variables, Expected Value and Variance

15.1 Random Variables

A **random variable** is a function that associates a unique numerical value with every outcome of an experiment. The value of the random variable will vary from trial to trial of an experiment.

Suppose S is the sample space associated with the random experiment E, then to every sample point in S we can assign a real number denoted by a variable X called a **random variable** on S. For example; when we toss a fair coin three times the sample space is

$$S = \{\text{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}\}$$

If we define a variable X as the number of heads observed when a fair coin is tossed three times, then X takes values 0, 1, 2, 3, where

$$\{X = 0\} \Leftrightarrow \{TTT\},\ \{X = 1\} \Leftrightarrow \{HTT, THT, TTH\};$$
$$\{X = 2\} \Leftrightarrow \{HHT, HTH, THH\},\ \{X = 3\} \Leftrightarrow \{HHH\}.$$

Hence to each sample points in S we have assigned a real number, which uniquely determines the sample point. The variable X is called the random variable defined on the sample space S.

We can also find the probabilities of values 0, 1, 2, 3 of the random variable X as follows

$$P(\{X = 0\}) = P(\{TTT\}) = 1/8,$$
$$P(\{X = 1\}) = P(\{HTT, THT, TTH\}) = 3/8,\ P(\{X = 2\}) = P(\{HHT, HTH, THH\}) = 3/8,\ P(\{X = 3\}) = P(\{HHH\}) = 1/8.$$

We can now express these probabilities in the form of a table;

Value of X	0	1	2	3	Total
Probability $P(X)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$	1

This is called the **probability distribution** of random variable X. A probability distribution for a discrete random variable is a formula, table or graph that provides the probability associated with each value of the random variable.

In general probability distribution of X satisfies the following conditions;

(i) all $P(X)$ are positive, i.e., $0 \leq P(X) \leq 1$, (ii) $\mathrm{P}_{\text{all } x} P(X) = 1$.

Note: Here, the uppercase X is used for the random variable and lowercase x is used to denote (represent) a realization of X. Probabilities can be easily obtained from the probability distribution table as follows:

Probability of getting two or more heads

$$P(X > 1) = P(X = 2) = P(X = 3) = \frac{3}{8} + \frac{1}{8} = \frac{1}{2}$$

Probability of getting at least one head

$$P(X > 0) = 1 - P(X = 0) = 1 - \frac{1}{8} = \frac{7}{8}$$

A **random variable** X defined on the sample space S may be finite or infinite, at the same time it may take only countable values (without decimal) such variables are called **discrete random variables**. A discrete random variable is one which may take on only a countable number of distinct values such as $0, 1, 2, 3, 4, \cdots$. Examples of discrete random variables include the number of students in a class, the members teams in a football tournament, number of public holidays in a year, number of guests in attendance at a party, etc. On the other hand some variables like height, weigh, income do take any possible value on a given range and are called the **continuous random variables**.

15.2 Mathematical Expectation

Mathematical expectation refers to the mean or expected value of a random variable X whose distribution is known. The expected value, denoted by $\mathrm{E}(X)$, is a weighted average of realizations x of X where the weights are the corresponding probabilities. If $P(x)$ is the probability of various outcomes of X, then the mean of X or the expected value of X is given by

$$\mathrm{E}(X) = x_1 \cdot P(x_1) + x_2 \cdot P(x_2) + x_3 \cdot P(x_3) + \cdots + x_n \cdot P(x_n)$$

$$= \sum_{\text{all } x} XP(X)$$

In the example above (Coin tossed 3 times)

$$\mathrm{E}(X) = \sum_{\text{all } x} x \cdot p(x)$$

$$= 0 \times \frac{1}{8} + 1 \times \frac{3}{8} + 2 \times \frac{3}{8} + 3 \times \frac{1}{8}$$

$$= \frac{12}{8}$$

$$= 1.5$$

Example 15.1. A discrete random variable X has the following probability distribution.

X	-2	-1	0	1	2
$P(X)$:	k	0.2	$2k$	$2k$	0.1

Find k and also find the expected value of the random variable X.

Solution. *Since X is a random variable with given $P(X)$, it must satisfy the conditions of a probability distribution.*

X

$$P(X) = 1 \Rightarrow 5k + 0.3 = 1 \Rightarrow k = 0.7/5 = 0.14.$$

Now we can calculate the expected value by the formula $E(X) = {}^{\mathrm{P}}XP(X)$.

X	$P(x)$	$xP(x)$
-2	*0.14*	*-0.28*
-1	*0.2*	*-0.2*
0	*0.28*	*0*
1	*0.28*	*0.28*
2	*0.1*	*0.2*
Total	*1*	*0*

Example 15.2. A random variable follows the probability distribution given below;

X	0	1	2	3	4
$P(X)$	0.12	0.23	k	0.20	0.10

Obtain the value of k, and hence compute the expected value of X.
$k = 0.35$, $E(X) = 1.93$ *and* $Var(X) = 0.35$

Example 15.3. A coin is such that the tail is thrice as likely as the head. A game is played such that you earn 5 points for a head and lose 2 points for a tail after every toss. Let X be the total score after 4 consecutive tosses. Find the probability distribution of X and the expected number of points.

Solution. Let H be the event of observing a head and let X be the points earned, then $P(H) = 0.25$, and $P(T) + 0.75$, $n = 4$ and using the binomial formula, we have

$$P(y \text{ tails in 4 tosses} = {}^{4}C_y(0.75)^y(0.25)^{4-y})$$

$$\begin{aligned}
P(y = 0) &= {}^{4}C_0(0.75)^0(0.25)^4 = 0.0039 \\
P(y = 1) &= {}^{4}C_1(0.75)^1(0.25)^3 = 0.0469 \\
P(y = 2) &= {}^{4}C_2(0.75)^2(0.25)^2 = 0.2109 \\
P(y = 3) &= {}^{4}C_3(0.75)^3(0.25)^1 = 0.4219 \\
P(y = 4) &= {}^{4}C_4(0.75)^0(0.25)^4 = 0.3164
\end{aligned}$$

Outcome	No tail	1 tail	2 tail	3 tail	All tail
x	20	$15 - 2 = 13$	$10 - 4 = 6$	$5 - 6 = -1$	$0 - 8 = -8$
$P(X = x)$	0.0039	0.0469	0.2109	0.4219	0.3164

$$\mathrm{E}(X) = 20(0.0039) + 13(0.0469) + 6(0.2109) + -1(0.4219) + -8(0.3164) = -1$$

Example 15.4. A game is played as follows; throw a fair four sided dice and score eight times the number that faces down unless it is a four, you are given a second chance in which you score only four times that faces down. Let X be a random variable denoting the score for each player. Represent this information on a tree diagram showing the value of X and the corresponding probability, hence or otherwise find the mean of the scores;

x	4	8	16	12	24
$P(x)$	$\frac{1}{16}$	$\frac{5}{16}$	$\frac{5}{16}$	$\frac{1}{16}$	$\frac{1}{4}$

$$\begin{aligned}\mathrm{E}(X) &= \sum_{\text{all } x} x \cdot p(x) \\ &= 4 \times \frac{1}{16} + 8 \times \frac{5}{16} + 16 \times \frac{5}{16} + 12 \times \frac{1}{16} + 24 \times \frac{1}{4} \\ &= \frac{232}{16} \\ &= 14.5\end{aligned}$$

Question 15.1. A discrete random variable X takes the following values with the corresponding probabilities;

Question 15.1. A discrete random variable X takes the following values with the corresponding probabilities;

x	-3	-1	0	1	2	3
$P(x)$	0.1	0.2	0.1	0.2	0.15	0.25

Compute the following probabilities; (a). $P(X = -1)$ (b). $P(X = -2)$
(c). $P(X \leq 0)$ (d). $P(X$ is negative) (e). $\mathrm{E}(X)$

Question 15.2. A random variable X has the following probability distribution;

x	10	20	30	40
$P(X = x)$	a	$2a$	$4a$	$3a$

Find the value of a hence the expected value of X.

Question 15.3. The table below gives the probability distribution function of a random variable X.

x	0	1	2	3
$P(X = x)$	p	$2q$	$p+q$	q

Given that the mean of X is 1.375, find the values of p and q.

Question 15.4. A game is played as follows; throw a fair four sided die and score four times the number that faces down unless its a four. If its a four, you toss a fair coin whose sides are assigned 0 for tail and 2 for head and score 5 times the coins score. Let X be a random variable denoting the score for each player, representing this information on a tree diagram showing the value of X and the corresponding probability hence, find;

(a). the probability that you will score more than 8.

(b). the expected value of X.

(c). the probability that you will score more than the expected value.

15.3 Variance

The variance of a probability distribution of a discrete random variable provides a numerical measure of the spread and is given by the sum of the products of the squared deviations between the mean and all individual values of the random variable, taken one at a time and their respective probabilities. Thus variance is given by the formula:

$$\begin{aligned}\text{Var}(X) &= E(X - E(X))^2 \\ &= E(X - E(X))^2 \\ &= \sum_{i=1}^{n}(x_i - E(X))^2 \cdot P(X_i) \\ &= \sum_{i=1}^{n}(x_i - \mu)^2 \cdot P(X_i) \qquad i = 1, 2, \cdots, n\end{aligned}$$

Variance is denoted by $\sigma^2 = \text{Var}(X) = E(X - E(X))^2$ and the standard deviation is the square root of the variance.

From our example above of the three fair coins being tossed once, we can calculate the value of the variance, as follows, knowing that the mean of the distribution is 1.5.

No. of Heads (X)	$P(X)$	μ	$X - \mu$	$(X - \mu)^2$	$(X - \mu)^2 P(X)$
0	1/8	1.5	−1.5	2.25	0.28
1	3/8	1.5	−0.5	0.25	0.09
2	3/8	1.5	0.5	0.25	0.09
3	1/8	1.5	1.5	2.25	0.28
			X	2	

$$(X - \mu)\ P(X) = 0.74$$

Hence variance which is denoted by σ^2 is given as

$$\begin{aligned}\sigma^2 &= \sum_{i=1}^{n}(X - \mu)^2 P(X) \\ &= 0.74\end{aligned}$$

and standard deviation

$$\sigma = \sqrt{\sigma^2} = \sqrt{0.74} = 0.86$$

Question 15.5. Determine the mean, variance, and standard deviation of the following discrete probability distribution.

x	0	1	2	3	4
$P(x)$	0.10	0.30	q	0.20	0.10

Question 15.6. A random variable X has the following probability distribution:

X:	-2	-1	0	1	2	3
$P(x)$:	0.1	k	0.2	$2k$	0.3	k

Find the value of k. Find the expected value and variance of X.

Question 15.7. A random variable X has the following probability distribution:

X:	0	1	2	3	4	5
$P(x)$:	0.1	0.1	0.2	k	0.2	0.1

Find the value of k. Find the expected value and variance of X.

Question 15.7. A random variable X has the following probability distribution:

X:	0	1	2	3	4	5
$P(x)$:	0.1	0.1	0.2	k	0.2	0.1

Find the value of k. Find the expected value and variance of X.

Question 15.8. An unbiased coin is tossed four times. Find the expected value and variance of the random variable defined as number of Heads.

15.4 Discrete Probability Distribution

A discrete probability distribution, is the listing of all possible outcomes of an experiment together with their probabilities. This concept can be illustrated with the following example:

Example 15.5. A fair coin is tossed two times. The following is the list of all possible outcomes of this experiment and their respective probabilities.

Outcomes	Probability
TT	1 / 4
TH	1 / 4
HT	1 / 4
HH	1 / 4

Then the probability distribution of the number of heads obtained in these two tosses of the coin is given as follows:

Number of heads (X)	Probability $P(X = x)$
0	1/4
1	$1/4 + 1/4 = 1/2$
2	1/4
	1

Note: All probabilities must add up to 1.

A discrete probability distribution takes on discrete values that can be counted and it can only assume values only from a distinct predetermined set. The commonly used discrete probability distributions are: Binomial and Poisson distributions.

Conditions for a function to be a Probability Distribution Function

1. The probability that a random variable assumes a value X_i is always between 0 and 1. That is

$$0 \leq P(x_i) \leq 1$$

2. The sum of all probabilities $P(X_i)$ is equal to one.

$$\sum_{i=1}^{n} P(X_i) = 1$$

Example 15.6. The number of telephone calls received in an office between 9.00 A.M - 10.00 A.M has the probability distribution as shown in the table below:

The Probability distribution of the number of telephone calls.

No. of calls	**Probability** $P(X)$
0	0.05
1	0.20
2	0.25
3	0.20
4	0.10
5	0.15
6	0.05

(a). Verify that it is a probability function.

(b). Find the probability that there will be 3 or more calls.

(c). Find the probability that there will be even number of calls.

Solution. Clearly,

(a).

$$(i).\ 0 \leq P(x_i) \leq 1$$

$$(ii).\ \sum_{i=1}^{n} P(X_i) = 0.05 + 0.20 + 0.25 + 0.2 + .010 + 0.15 + 0.05 = 1$$

(b).

$$P(X \geq 3) = P(X = 4) + P(X = 5) + P(X = 6)$$
$$= 0.20 + 0.10 + 0.15 + 0.05$$
$$= 0.50$$

(c).

$$P(X = 0 \text{ or } 2 \text{ or } 4 \text{ or } 6) = P(X = 0) + P(X = 2) + P(X = 4) + P(X = 6)$$
$$= 0.05 + 0.25 + 0.10 + 0.05$$
$$= 0.40$$

15.5 The Mean or Expected Value of discrete probability distribution

The mean of the probability distribution is also known as the *Expected value*. Let X be a discrete random variable with the expected probability distribution $P(X)$. Then the expected value denoted as $E(X)$ is given by:

$$\mathrm{E}(X) = \sum_{i=1}^{n} x_i P(X_i) \quad i = 1,2,3\cdots,n.$$

Each value of the random variable is multiplied by the probability of occurrence of this value and then all these products are summed up.

It is also common in statistical literature to refer to the mean as Mathematical Expectation or the Expected value of the random variable X.

Example 15.7. Assume that we have three fair coins and we toss them simultaneously. The possible number of heads that can appear as s result of the random experiment are given in the following table:

Outcomes	No. of Heads	Probability
TTT	0	1 / 8
HTT	1	1 / 8
TTH	1	1 / 8
THT	1	1 / 8
THH	2	1 / 8
HHT	2	1 / 8
HTH	2	1 / 8
HHH	3	1 / 8

Thetablecanbesummarizedastothenumberofheadsoccurringintheentireexperimentandtheirrespective probabilities as follows:

Number of Heads (X)	$P(X)$	$X \cdot P(X)$
0	1/8	0
1	3/8	3/8
2	3/8	6/8
3	1/8	3/8
	1.0	12/8

The expected value (mean) for the number of heads in this experiment is

$$\begin{aligned} \mathsf{E}(X) &= \sum_{i=1}^{n} x_i P(X_i) \qquad i = 1, 2, \cdots, n \\ &= 0 + \frac{3}{8} + \frac{6}{8} + \frac{3}{8} \\ &= \frac{12}{8} = 1.5 \end{aligned}$$

This means that on an average, 1.5 heads can be expected to appear as a result of every random experiment of tossing three fair coins at any one time.

Example 15.8. In the telephone calls problem above find the mean of the telephone calls between 9 -10 am

X	$P(X)$	$X\,P(X)$
0	0.05	0
1	0.20	0.2
2	0.25	0.5
3	0.20	0.6
4	0.10	0.4
5	0.15	0.75
6	0.05	0.30
	1.00	2.75

$$\mu = \sum_{i=0}^{6} x_i P(X_i) = 2.75$$

Example 15.9. Suppose the hourly earnings X of a self employed landscaper gardener are given by the following probability function.

Hourly Earning X:	0	6	12	16
$P(X)$:	0.3	0.2	0.3	0.2

Find the gardener's Mean.

Solution. The Mean is given as:

$$\begin{aligned} \mu &= 0(0.3) + 6(0.2) + 12(0.3) + 16(0.2) \\ &= 0 + 1.2 + 3.6 + 3.2 \\ &= 8.0 \end{aligned}$$

16 Past Examination Papers

DEDAN KIMATHI UNIVERSITY OF TECHNOLOGY

University Examinations 2015/2016

EXAMINATION FOR THE FIRST YEAR SECOND SEMESTER DEGREE OF BACHELOR OF SCIENCE IN **COMPUTER SCIENCE**, BACHELOR OF SCIENCE IN **ACTUARIAL SCIENCE**, BACHELOR OF SCIENCE IN **INDUSTRIAL CHEMISTRY**, BACHELOR OF SCIENCE IN **LEATHER TECHNOLOGY** AND THIRD YEAR FIRST SEMESTER BACHELOR OF SCIENCE IN **INFORMATION TECHNOLOGY**.

SMA 2103-STA 2100: Probability and Statistics I

DATE: Wednesday, 20th April 2016 TIME: 11am -1.00 pm

Instructions: Answer QUESTION ONE and any other TWO QUESTIONS.

QUESTION ONE (30 marks) (COMPULSORY)

(a) Giving an example in each case differentiate the following terms as used in statistics:-

(i) Discrete and Continous data. **[2 marks]**

(ii) Nominal and Ordinal data. **[2 marks]**

(b) Applicants for an assembly job are required to take a test of manual dexterity. The times in seconds taken to complete the task for 19 applicants were as follows:-

63 ,229,165,77,49,74,67,59,66,102,81,72,59,74,61,82,48,70, 86

For these data find

(i) the median. **[1 mark]**

(ii) the upper and lower quartiles. **[2 marks]**

An outlier is defined as a value outside the range $\bar{x} \pm 2s$ where s is the standard deviation and $\bar{x}$ the mean of the data.

(iii) Identify any outliers in the data. **[3 marks]**

(iv) Illustrate the data by a box-plot. Outliers if any should be denoted by * and should not be included in the whiskers. **[2 marks]**

(c) The following table gives the one-way commuting distance (in nearest kms) of 30 working women in an Insurance company.

13	47	10	3	16	7
25	8	21	19	12	45
1	8	4	6	2	14
13	7	34	13	41	28
50	14	26	10	24	36

(i) Make a stem-leaf display of the data. **[2 marks]**

(ii) Taking class intervals of the form 1-5, 6-10, 11-15 and so on construct a frequency distribution and use it this the histogram. Comment on the skewness of the distribution. **[5 marks]**

(d) In an entrance examination in Mathematics and Statistics, of the 120 students appeared for the examination, 65 passed in Mathematics, 75 passed in Statistics and 35 passed in both the tests. A student is selected at random. What is the probability that the student has

(i) failed in both the tests. **[3 marks]**

(ii) passed in Mathematics given that the student has passed in at least one test. **[2 marks]** (e) Given a set of data $x_1,x_2,\ldots,x_n$ occurring with frequencies $f_1,f_2,\ldots,f_n$ respectively.

Define

(i) The r^{th} moment,m'_r about a point a. **[1 mark]**

(ii) The r^{th} moment, m_r about the mean . **[1 mark]**

(iii) Show that$m_4 = m'_4 - 4m'_1m'_3 + 6{m'_1}^2m'_2 - 3{m'_1}^4$ **[4 marks]**

QUESTION TWO (20 Marks)

(a) In an experiment, a bottle of milk of milk was brought from a cooler into a room whose temperature is $25^{\circ}C$. Its temperature $y^{\circ}C$ was recorded at time t minutes after it was brought in for 11 different values of t. The results are summarized as follows

$$\Sigma\, t = 44\ \Sigma\, t^2 = 180.4\ \Sigma\, ty = 824.5\ \Sigma\, y = 205$$

(i) regression lines y on t in the form $y = a + bt$. **[5 marks]**

(ii) Explain the practical significance of a and b **[2 marks]**

(iii) Use your equation to estimate the value of y when t=4.5 and t=20.
Comment with a reason on the reliability of each of the two estimates. **[4 marks]**

(c) The table below shows the marks obtained by six students in two examinations

student	A	B	C	D	E	F
English	38	62	56	42	59	48
Maths	64	84	84	60	73	89

(i) Calculate the Spearman's rank correlation coefficient and comment on the value. **[6 marks]**

(ii) The maths papers were remarked and one of the students awarded five more marks. Given that the other marks and the rank correlation coefficient were unchanged, state with reason which student received the extra marks. **[2 marks]**

(iii) Under what conditions would you expect the Spearman's rank correlation coefficient to be equal to the product-moment correlation coefficient. **[1 mark]**

QUESTION THREE (20 Marks)

(a) (i) Explain in words the meaning of the following symbol $P(A|B)$ where *A* and *B* are two events. **[1 mark]**

(ii) State the relationship between *A* and *B* if $P(A|B) = 0$ and $P(A|B) = P(A)$. **[2 marks]**

(b) When a car owner needs his car to be serviced he calls one of the three garages *A*, *B* or *C*. Of all his calls 30% to garage *A*, 10% to garage *B* and 60% to garage *C*.

The percentage of occasions when the garage called can take the car on that particular day are 20 % for *A*, 6% for *B* and 9% for *C*.

(i) Draw a tree diagram to show this information. **[2 marks]**

(ii) Find the probability that the garage called will *not* be able to service the call on that particular day. **[2 marks]**

(iii) Given that the car owner calls a garage and the car is serviced the same day find the probability that he called garage *C*. **[3 marks]**

(c) A bag contains five identical balls two of which are green while the others are red. The balls are successively drawn without replacement until both green balls are obtained.

Let *X* denote the number of draws required to obtain both green balls. Obtain the

(i) probability distribution of *X*(*show your working*). **[5 marks]**

(ii) mean and variance of *X*. **[5 marks]**

QUESTION FOUR (20 Marks)

(a) A discrete random variable X takes only the values $0,1,2,3,4,5$. The probability distribution of X is as follows

$$P(X=0)=P(X=1)=P(X=2)=p$$
$$P(X=3)=P(X=4)=P(X=5)=q$$
$$P(X\geq 2)=3P(X<2)$$

Determine the

(i) values of p and q. **[4 marks]**

(ii) mode of the distribution. **[1 mark]**

(iii) probability that the sum of any two independent observations from the distribution is 7. **[3 marks]**

(b) For a given data set the regression lines y on x and x on y are $y+0.219x = 20.8$ and $0.785y+x = 16.2$ respectively. Find

(i) the product moment correlation coefficient. **[4 marks]**

(ii) $\bar{x}$ and $\bar{y}$. **[2 marks]**

(c) From two samples x and y, the following statistics were obtained.

$$\sum_{i=1}^{9} x_i = 39 \quad \sum_{i=1}^{9} x_i^2 = 237 \quad \sum_{i=1}^{7} y_i = 27 \quad \sum_{i=1}^{7} y_i^2 = 131$$

(i) Determine the mean and the variance of the combined(pooled) sample. **[4 marks]**

(ii) Suppose 5 is added to each of the x_i and y_i. Find the new pooled mean and variance. **[2 marks]**

QUESTION FIVE (20 Marks)

The table below shows the frequency distribution of the masse 52 students in a college. Measurements were recorded to the nearest *kg*.

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

Estimate the

(a) (i) mode. **[2 marks]**

(ii) value of *x*, if 20% of the students were heavier than *x*. **[3 marks]**

(b) Compute the first four moments about the point 57 using the coding method and hence investigate the skewness and peakedness of this distribution. **[15 marks]**

DEDAN KIMATHI UNIVERSITY OF TECHNOLOGY

University Examinations 2015/2016

First Year Second Semester Examination for the Degree of Bachelor of Science in **Actuarial Science**, Bachelor of Science in **Computer Science**, Bachelor of Science in **Industrial Chemistry** and Bachelor of Science in **Leather Technology**

SMA 2103/STA 2100 Probability and Statistics I

DATE: 18TH MAY 2015 **TIME:** 11.00 AM - 1.00 PM

Instructions: Answer QUESTION ONE and any other TWO QUESTIONS.

QUESTION ONE (30 MARKS) (COMPULSORY)

(a) Giving an example in each case differentiate between the following terms as used in statistics:-

(i) Qualitative and quantitative data. **[2 marks]**

(ii) Nominal and Ordinal data. **[2 marks]**

(iii) Census and sample survey. **[2 marks]**

(b) Applicants for an assembly job are required to take a test of manual dexterity. The times in seconds taken to complete the task for 19 applicants were as follows:-

63 ,229,165,77,49,74,67,59,66,102,81,72,59,74,61,82,48,70, 86

For these data find

(i) the median. **[1 mark]**

(ii) the upper and lower quartiles. **[2 marks]**

An outlier is defined as a value less than $q_1 - 1.5(q_3 - q_1)$ or greater than $q_3 + 1.5(q_2 - q_1)$.

(iii) Identify any outliers in the data. **[3 marks]**

(iv) Illustrate the data by a box-plot. Outliers if any should be denoted by * and should not be included in the whiskers. **[3 marks]**

(c) A bag contains five identical balls each bearing one of the numbers 1,2,3,4 and 5. A ball is picked at random from the bag its number noted and then replaced. This was done 50 times and the following results obtained.

Number	1	2	3	4	5
Frequency	x	11	y	8	9

If the mean of is 2.7 find the standard deviation of the values. **[5 marks]**

(d) A company manufacturers T.V. sets. The probability that a set from this company fails during first month of its use is 0.02. Of those that do not fail during first month, the probability of failure in the next five months is 0.01. Of those that do not fail during the first six months, the probability of failure by the end of the first year is 0.001. The company replaces, free of charge, any set that fails during its warranty period. If 2,000 sets are sold, how many will have to be replaced if the warranty period is

(i) six months. **[2 marks]**

(ii) one year. **[2 marks]**

(e) Given a set of data $x_1,x_2,\ldots,x_n$ occurring with frequencies $f_1,f_2,\ldots,f_n$ respectively. Define

(i) The r^{th} moment,m'_r about a point a. **[1 mark]**

(ii) The r^{th} moment, m_r about the mean . **[1 mark]**

(iii) Show that$m_4 = m'_4 - 4m'_1m'_3 + 6m'^2_1m'_2 - 3m'^4_1$ **[4 marks]**

QUESTION TWO (20 Marks)

(a) Gross mean weekly earnings (y in Ksh. per week) for a sample of male clerical workers of varying ages (x, in complete years) in a large company are as follows:

Earnings, y	215	259	348	387	534	660	726	$\sum y = 3129$	$\sum y^2 = 1632011$
Age, x	18	20	23	28	35	45	55	$\sum x = 224$	$\sum x^2 = 8312$

You are also given that $\sum xy = 116,210$.

(i) Plot a scatter diagram of these data and comment on their suitability for simple linear regression analysis. **[3 marks]**

(ii) Obtain the coefficient of correlation between X and Y. **[4 marks]**

(iii) Write down the models for

(a) Simple linear regression of y on x.

(b) Simple linear regression of x on y.

Define your notation and explain clearly which model is better suited to fit the variables and data as defined in the table above. **[5 marks]**

(iv) (a) Fit the simple linear regression model of y on x to the data above, find the equation of the fitted regression line, draw this line on your scatter diagram, and use the equation to estimate the mean weekly earnings at age 50.

[7 marks]

(b) Your line manager asks you to use your model to estimate the mean weekly earnings at age 70. How would you answer him? **[1 mark]**

QUESTION THREE (20 Marks)

(a) A computer program generates random questions in arithmetic that have to be answered within a fixed time. The probability of answering the first question correctly is 0.8. Whenever a question is answered correctly, the next question generated is more difficult and the probability of a correct answer being given is reduced by 0.1. Whenever a question is answered wrongly, the next question is of the same standard and the probability of answering it correctly is not changed.

(i) Draw a tree diagram to show this information for the first two generated questions.

[2 marks]

(ii) Find the probability that the second question was answered wrongly.

[2 marks]

(iii) By extending the tree diagram find the probability that the second question is answered correctly given that the second question is answered correctly.

[4 marks]

(b) Two events A and B are independent. Given that $P(A) = 0.4$ and $P(A \cup B) = 0.7$ find $P(B)$. **[2 marks]**

(c) Two events C and D are such that $P(D|C) = 0.2$ and $P(C|D) = 0.25$. Given that $P(C \cup D) = 0.2$ find the probability that both events occur. **[4 marks]**

(d) A discrete random variable *X* takes the values 0,1,2 and 3 only.

Given that $P(X \leq 2) = 0.9$, $P(X \leq 1) = 0.5$ and $E(X) = 1.4$, find the 2^{nd} moment about the origin of *X*. **[5 marks]**

QUESTION FOUR (20 Marks)

(a) A random variable X has the following probability distribution.

$$P(X=x)=\begin{cases} kx^2, & x=1,2,3 \\ k(7-x)^2, & x=4,5,6 \\ 0, & \text{otherwise} \end{cases}$$

(i) Obtain the value of *k*.

Hence rewrite the probability distribution. **[3 marks]**

(ii) Obtain the mean and standard deviation of X. **[6 marks]**

(iii) State the mode of X. **[1 mark]**

(iv) Determine the mean and variance of a variable $Y = 2X + 2$. **[2 marks]**

(b) *A, B, C* and *D* toss a fair coin in turn, starting with A and the first to throw a head wins. The game can continue indefinitely until a head is thrown. However *D*, objects as the others have their first turn before him.

Compare the probability that *A* wins with the probability that *D* wins. **[8 marks]**

QUESTION FIVE (20 Marks)

(a) A doctor inquired from 10 of his patients the number of years they had smoked. For each patient he gave a grade between 0 and 100 of the extent of lung damage. The following table shows the results

No of years	15	22	25	28	31	33	36	39	42	48
Grade	30	50	55	30	57	35	60	72	70	75

Calculate the Spearman's rank correlation coefficient between the number of years of smoking and the extent of lung damage.

Comment on the figure your obtain. **[6 marks]**

(b) A sample of 51 people were asked to record the distance they had travelled by car in a given week. The distances to the nearest kilometer are shown below

67	76	85	42	93	48	93	46	52	72
77	53	41	48	86	78	56	80	70	70
66	62	54	85	60	58	43	58	74	44
52	74	52	82	78	47	66	50	67	87
78	86	94	63	72	63	44	47	57	68
81									

(i) Construct a suitable stem and leaf diagram to represent these data.

Comment on the shape of the distribution. **[4 marks]**

(ii) Starting with the interval 40-49, construct a frequency distribution table for the data. **[2 marks]**

(ii) Investigate the skewness and peakedness of this distribution. **[8 marks]**

DEDAN KIMATHI UNIVERSITY OF TECHNOLOGY

University Examinations 2015/2016

FIRST YEAR SECOND SEMESTER EXAMINATION FOR THE DEGREE OF BACHELOR OF **COMMERCE**, BACHELOR OF **BUSINESS ADMINISTRATION** AND BACHELOR OF **BUSINESS INFORMATION TECHNOLOGY**

HBC 2121: Introduction to Business Statistics

DATE: 13 TH APRIL 2015 **TIME:** 11.00 AM - 1.00 PM

Instructions: Answer QUESTION ONE and any other TWO QUESTIONS.

QUESTION ONE (30 MARKS) (COMPULSORY)

(a) Most undergraduate Business students will not go on to become actual practitioners of statistical research and analysis. Considering this fact, why should you bother to become familiar with business statistics concepts? ***(3 marks)***

(b) Distinguishbetweenqualitativeandquantitativevariables. Determinewhichofthefollowingvariables are qualitative and which are quantitative. Classify the quantitative variables further into discrete quantitative and continuous quantitative variables.

(i). Number of family members.

(ii). Your gender

(iii). The time you spend watching the English Premier league.

(iv). Your latest hairstyle. ***(4 marks)***

(c) The following data consist of the weights, in pounds, of 35 adults:

176, 154, 161, 125, 138, 142, 108, 115, 187, 158, 168, 162 135, 120, 134, 190, 195, 117, 142, 133 , 138, 151, 150, 168 172, 115, 148, 112, 123, 137, 186, 171, 166, 166, 179

(i) Organize the data in a table, using $100 - 119$ as the smallest interval.***(3 marks)***

(ii) Construct a frequency histogram based on the grouped data. ***(4 marks)***

(iii) Calculate the mean and the standard deviation of the data. ***(5 marks)***

(iv) In what interval is the median for these grouped data? Calculate the median of the data. ***(3 marks)***

(d) Given two events A and B such that $P(A|B) = 0.8$, $P(A) = 0.5$, and $P(B) = 0.25$.
Determine

i). $P(A \cap B)$. ***(1 mark)***

ii). $P(A \cup B)$. ***(1 mark)***

iii). Whether the two events A and B are mutually exclusive, independent or collectively exhaustive events.

(6 marks)

QUESTION TWO (20 Marks)

(a) (i) State and briefly explain giving an example the four components of a time series.

(4 marks)

(ii) The data below gives sales in millions for a certain company in Kenya from 2007 to 2015:

Year:	2007	2008	2009	2010	2011	2012	2013	2014	2015
Sales:	23	15	17	22	25	29	25	30	29

Obtain smoothed values using 4-point moving averages.

(6 marks)

(c) (i) IftheFisher'spriceindexis109.91andthePaasche'spriceindexis110.6, calculate the Laspeyre's index number. ***(2 marks)***

(ii) Define an index number. Explain two areas where index numbers are applied.

(2 marks)

(d) The table below gives the monthly cost in Ksh of some living necessities in a certain town for two time periods. Each necessity has been given a weight as a measure of its importance to living.

Item	**Cost**	**Cost**	**Weight**
	2005	2007	w
Food	1500	2000	5.5
Security	1300	1100	1.3
Transport	2000	2400	2.2
Rent	2500	3100	4.2
Health	1900	1800	4.1
Others	5500	5400	2.7

Taking 2007 as the base period, calculate the cost of living index interpret it.

(6 marks)

QUESTION THREE (20 Marks)

(a) In a study of the daily production of a company for over 50 days, the following data was obtained:

65	76	36	48	49	48	84	55	79	51
43	21	78	35	37	61	40	45	68	33
88	45	50	53	60	34	56	67	57	42
59	62	62	65	76	55	76	61	70	73
35	41	60	74	52	82	63	58	32	26

(i) Starting with a class $20 - 29$, group this data into a frequency distribution, and plot its ogive.***(6 marks)***

(ii) Using a suitable assumed mean, calculate the mean and standard deviation of this data. ***(4 marks)***

(b) Consider the following data whose mean is 9.

3	5	1	13	6	10	8
11	12	17	23	X	0	

Calculate

(i). the value of X. ***(2 marks)***

(ii). the median and third decile of the data. ***(4 marks)***

(iii). the coefficient of variation ***(3 marks)***

(iv). the mode of the data ***(1 mark)***

QUESTION FOUR (20 Marks)

(a) Differentiate between correlation and regression analysis. ***(2 marks)***

(b) The following data have been collected relating to returns which would have been earned from an investment of an equal sum of money in the shares of E.T. Plc. and a group of market shares.

Year:	1	2	3	4	5	6	7	8	9	10
E.T. Plc shares (Y):	7.8	11.0	15.2	23.1	29.7	37.4	44.6	52.8	60.2	63.9
Mkt shares (X):	11.1	12.3	18.5	25.4	28.7	33.8	37.7	39.6	44.7	45.5

Calculate the least squares regression of Y on X. ***(6 marks)***

(c) Recent unit prices in hundreds of shillings of various fruits and vegetables (items) in Nairobi and Nyeri were as follows:

Item:	A	B	C	D	E	F	G
Nairobi (X):	14	16	16	9	8	28	35
Nakuru (Y):	9	18	20	15	6	26	38

(i). Calculate the product moment coefficient of correlation. ***(6 marks)***

(ii). Calculate the coefficient of determination and interpret the value. ***(2 marks)***

QUESTION FIVE (20 Marks)

(a) Differentiate between the following terms as used in Hypothesis testing

(i) Null hypothesis and alternative hypothesis

(ii) Type I and Type II error ***(4 marks)***

(a) Find the probabilities that a random variable having a Standard Normal distribution will take on a value.

(i) greater than 1.72. ***(2 marks)***

(ii) less than -0.88 ***(2 marks)***

(iii) between 1.30 and 1.75 ***(2 marks)***

(b) Three machines A, B, and C produce respectively 60%, 30% and 10% of the total number of items of a factory. The percentages of defective output of these machines are respectively 2%, 3%, and 4 %. An item is selected at random from the product and is found to be defective. Find the probability that the item was produced by machine C.

(4 marks)

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

(6 marks)

DEDAN KIMATHI UNIVERSITY OF TECHNOLOGY

University Examinations 2013/2014

EXAMINATION FOR THE FIRST YEAR DEGREE OF BACHELOR OF SCIENCE IN **COMPUTER SCIENCE**, BACHELOR OF SCIENCE IN **ACTUARIAL SCIENCE** AND FOR THE THIRD YEAR IN THE DEGREE OF BACHELOR OF SCIENCE IN **INFORMATION TECHNOLOGY**

SMA 2103-STA 2100: Probability and Statistics I

DATE: 14 TH April 2014 TIME: 11.00 AM - 1.00 PM

Instructions: Answer QUESTION ONE and any other TWO QUESTIONS.

QUESTION ONE (30 marks) (COMPULSORY)

(a) Giving an example in each case differentiate the following terms as used in statistics:-

(i) Continuous and discrete data. **[2 marks]**

(ii) Nominal and Ordinal data. **[2 marks]**

(iii) Population and sample. **[2 marks]**

(b) The following data are the temperatures of effluent at discharge from a sewage treatment facility on consecutive days

43 47 51 48 52 50 46 49 45 52 46 51 44 49 46 51 49 45 44 50 48 50 49 50 Construct a box plot of the data and use comment on the skewness of this distribution. **[5 marks]**

(c) The mode of the following incomplete distribution of weights of 160 students is 56.

Weights (kgs)	30-40	40-50	50-60	60-70	70-80	80-90
No. of Students	20	36	x	y	15	5

Assuming that the weights are linearly distributed in each group estimate the

(i) values of x and y . **[5 marks]**

(ii) median weight . **[3 marks]**

(iii) mean and standard deviation of the distribution, using the coding method and an assumed mean of 55kgs . **[6 marks]**

(iv) Are values obtained in (ii) and (iii) above the actual values of those quantities? Give a reason for your answer **[2 marks]**

(d) Among 1,000 applicants for admission to M.A. economics course in a University, 600 were economics graduates and 400 were non-economics graduates; 30% of economics graduate applicants and 5% of non-economics graduate applicants obtained admission. If an applicant selected at random is found to have been given admission, what is the probability that he or she is an economics graduate? **[3 marks]**

QUESTION TWO (20 marks) (Optional)

(a) The following table presents sample data relating the number of study hours spent by students outside of class during a three-week period for a course in statistics and their scores in an examination given at the end of that period.

Sampled student	1	2	3	4	5	6	7	8
Study hours (x)	20	61	34	23	27	23	18	22
Examination grade (y)	64	61	84	70	88	92	72	77

(i) Plot a scatter diagram. **[3 marks]**

(ii) Using the calculator or otherwise find the regression line y on x.
Comment on the relationship between the marks scored and the number of hours of study. **[3 marks]**

(iii) Give the product moment correlation coefficient. **[1 mark]**

(iv) Draw the regression line y on x on the scatter diagram. **[1 mark]**

(v) If a student studied for 30 hours what is the expected mark? **[1 mark]**

(vi) If a student studied for 15 hours what is the expected mark?
Is this estimate reliable? Give a reason for your answer. **[3 marks]**

(b) Two variables X and Y are such that the regression equation y on x is $4x - 5y + 33 = 0$ and that of x on y is $20x - 9y = 107$. Obtain the

(i) Mean values of X and Y . **[2 marks]**

(ii) Correlation coefficient between X and Y . What is its sign? why? **[5 marks]**

(iii) Comment on the correlation between X and Y **[1 mark]**

QUESTION THREE (20 marks) (Optional)

(a) Three computer viruses arrived as an e-mail attachment. Virus A damages the system with probability 0.4. Independently of it, virus B damages the system with probability 0.5. Independently

of A and B, virus C damages the system with probability 0.2. Use a tree diagram to obtain the possible outcomes and hence determine the probability that the system gets damaged? **[5 marks]**

(b) A computer assembling company receives 24% of parts from supplier X, 36% of parts from supplier Y, and the remaining 40% of parts from supplier Z. Five percent of parts supplied by X, ten percent of parts supplied by Y, and six percent of parts supplied by Z are defective. If an assembled computer has a defective part in it, what is the probability that this part was received from supplier Z?**[4 marks]**

(c) A computer system has two components. Define the following events

A: first component is good

B: second component is good

Given that $P(A) = \frac{4}{5}, P(B|A) = \frac{17}{20}$ and $P(B|\bar{A}) = \frac{3}{4}$

Determine the probability that

(i) The second component is good. **[3 marks]**

(ii)) At least one of the components is good. **[3 marks]**

(iii) For the two events A and B are they independent or are they mutually exclusive? Verify your answers. **[5 marks]**

QUESTION FOUR (20 marks) (Optional)

(a) A random variable X has the following probability distribution.

x	1	2	3	4
$P(X = x)$	c	c^2	c^2+c	$3c^2 + 2c$

(i) Obtain the value of *c*.

Hence rewrite the probability distribution. **[4 marks]**

(ii) Obtain the mean and variance of X. **[5 marks]**

(iii) State the mode of X. **[1 mark]**

(iv) Find $P(X \geq E(X))$. **[2 marks]**

(b) The heights and the corresponding weights of a group of 9 randomly selected students were measured and the following results obtained.

Height (m) X	1.6	1.7	1.6	1.4	1.6	1.7	1.4	1.3	1.2
Weight (kg) Y	70	75	65	55	60	76	55	50	49

(i) Determine the Spearman's rank correlation coefficient.

Comment on the relationship between the heights and the weights of the students **[6 marks]**

(ii) For each value of X a constant k was added and each value of Y a constant k was subtracted. State giving a reason whether the value of the rank correlation coefficient increased, decreased or did not change. **[2 marks]**

QUESTION FIVE (20 marks) (Optional)

(a) Given a set of data $x_1, x_2, \ldots, x_n$ occurring with frequencies $f_1, f_2, \ldots, f_n$ respectively. Define

(i) The r^{th} moment, m'_r about a point a. **[1 mark]**

(ii) The r^{th} moment, m_r about the mean . **[1 mark]**

(iii) Show that $m_4 = m'_4 - 4m'_1m'_3 + 6{m'_1}^2m'_2 - 3{m'_1}^4$ **[4 marks]**

(b) The distribution of marks obtained by students in an exam is as follows

Marks	10 - 19	20 - 29	30 - 39	40 - 49	50 - 59	60 - 69
frequency	4	6	9	5	4	2

(i) Draw a histogram for this distribution. **[3 marks]**

(ii) Using an assumed mean of 34.5, estimate the first four moments about the mean. **[8 marks]**

(ii) Investigate the skewness and peakedness of distribution of the marks. **[4 marks]**

DEDAN KIMATHI UNIVERSITY OF TECHNOLOGY

University Examinations 2013/2014

FIRST YEAR SPECIAL/SUPPLEMENTARY EXAMINATION FOR THE DEGREE OF **BACHELOR OF COMMERCE**

HBC 2121 Introduction to Business Statistics

DATE: TH MARCH 2013 TIME: 2.00 PM-4.00 PM

Instructions: Answer QUESTION ONE and any other TWO QUESTIONS.

QUESTION ONE (30 marks) (COMPULSORY)

(a). Define the following terms as used in statistics.

(i) Nominal measurements

(ii) Ordinal measurements

(iii) Independent events

(iv) Coefficient of variation

(v) Mutually exclusive events **[5 marks]**

(b). The marks obtained by 30 students in a mathematics test marked out of 20 are as shown below

Marks	8	9	10	11	12	14	15	17	18	20
No. of Students	2	3	4	4	5	3	3	3	2	1

Compute

(i) The mode

(ii) Median

(iii) Mean

(iv) Standard deviation **[7 marks]**

(c). The first of the two groups has 100 items with mean 45 and variance 49. If the combined group has 250 items with mean 51 and variance 130, find the mean and standard deviation of the second group. **[5 marks]**

(d). The following data refers to Examination marks verses hours of study per week of a sample of eight candidates that sat for Business statistics examination in 2010.

Exam Mark (Y)	64	61	84	70	88	92	72	72
Hours of Study (X)	20	16	34	23	27	32	18	22

Calculate

(i) the Pearson's product moment coefficient of correlation. **[5 marks]**

(ii) the coefficient of determination between examination marks and hours of study. **[2 marks]**

(e). If 50% of the families subscribe to the morning newspaper, 65% of the families subscribe to the afternoon newspaper and 85% of the families subscribe to at least one of the two newspapers, what proportion of the families subscribe to both newspapers? **[3 marks]** (f).

(i) Under what conditions does $P(A|B) = P(A)$? **[1 marks]**

(ii) What is the addition rule of probability and for what type of events is it valid? **[2 marks]**

QUESTION TWO (20 marks)

(a). What is an index number?

Explain two areas where index numbers are applied. **[2 marks]**

(b). If the Fisher's price index is 109.91 and the Paasche's price index is 110.6, calculate the Laspeyre's index number. **[2 marks]**

(c). (i) Define the term "time series" and giving an example. **[2 marks]**

(ii) What is the aim of time series analysis? **[2 marks]**

(iii) By giving the relevant examples, briefly explain the components of a time series. **[6 marks]**

(d). The number of new stereo systems sold by an electrical store each quarter for four years is shown below.

Year	Quarters	Sales
2000	1	130
	2	105
	3	85
	4	120
2001	1	150
	2	115
	3	95
	4	140
2002	1	165
	2	120
	3	100
	4	140
2003	1	180
	2	140
	3	110
	4	160

Calculate a four point moving average trend. **[6 marks]**

QUESTION THREE (20 marks)

(a). The following table gives the marks obtained by first year students in a Marketing examination.

Marks	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
No. of Students	3	9	14	24	18	16	9	4

Calculate

(i) the mean

(ii) the Standard deviation

(iii) the median, and

(iv) the mode **[10 marks]**

(b). (i) State Bayes Theorem. **[2 marks]**

(ii) KY accounting firm has noticed that of the companies it audits, 85% show no inventory shortages, 10% show small inventory shortages and 5% show large inventory shortages. KY firm has devised a new accounting test for which it believes the following probabilities hold:

P(company will pass test | no shortages) = 0.90

P(company will pass test | small shortages) = 0.50

P(company will pass test | large shortages) = 0.20

Determine the probability if a company being audited fails this test has large or small inventory shortages. **[8 marks]**

QUESTION FOUR (20 marks)

(a). Statistics is a means of collection of numerical facts on data. What are the major advantages of sampling method over the census? **[4 marks]**

(b). The random variable X has a probability distribution shown in the table below

X	0	10	20	30
$P(X = x)$	0.1	p	0.45	q

Given that the mean of X is 16.5, find the value of p and q. **[4 marks]**

(c). The following data shows two groups of casual workers, their number and average wages they are paid.

	Group A	Group B	Combined Groups
Number of Workers	X	100	250
Mean wage	600	Y	6800
Standard Deviation	8	9	Z

(i) Calculate the missing entries; X, Y , and Z. **[8 marks]**

(ii) Calculate the coefficient of variation for Groups A and B. **[3 marks]**

(iii) From (ii) above, which group has a greater variability in income? **[1 marks]**

QUESTION FIVE (20 marks)

In a study of the daily production of a company for 50 days, the following data was obtained:

65	76	36	48	49	48	84	55	79	51
43	21	78	35	37	61	40	45	68	33
88	45	50	53	60	34	56	67	57	42
59	62	62	65	76	55	76	61	70	73
35	41	60	74	52	82	63	58	32	26

(i) Starting with a class $20 - 29$, group the data into a frequency distribution, represent the data in a histogram and plot its ogive. **[8 marks]**

(ii) State the modal class. **[1 marks]**

(iii) Using a suitable assumed mean, calculate the mean hence or otherwise calculate, the mode, the median and the standard deviation of the data above. **[11 marks]**

KUCT

Town Campus

Introduction to Business Statistics WKD
class ASSIGNMENT I: 60 marks

Instructions: Answer *all* Questions.

1. The ratings below are based on collisions claim experience and theft frequency for 12 makes of small, two-door cars. higher numbers reflect higher claims and more frequent thefts, respectively.

Collision	Theft	Collision	Theft
103	103	106	97
97	113	139	425
105	81	110	82
115	68	96	81
127	90	84	59
104	79	105	167

   (a) Plot the data on a scatter diagram

   (b) Determine the least-squares regression line for predicting the rate of collision claims on the basis of theft frequency rating.

   (c) Calculate and interpret the values of r and r^2.

   (d) If a new model were to have a theft rating of 110, what would be the predicted rating for collision claims?

2. A firm has four plants scattered around the city producing the same item. Plant A produces 30% of total production, plant B produces 25%, plant C produces 35%, and plant D produces 10%. The firm has a single warehouse in the city for storing finished products from all the plants. From the past performance records on the proportion of defectives, it has been found that 5%, 10%,15%, and 20 % of the items produced at A,B, C, and D respectively are defectives. before the shipment of an item to a dealer, one unit is selected at random and found to be defective. What is the probability that it was produced by plant C?

3. The masses in grams of some fruits are given in the table below

363.7	346.4	377.4	341.7	359.8	361.2	385.7	363.5	354.2	375.3
372.2	364.3	373.3	379.4	351.4	369.5	385.5	365.5	385.5	369.5

(a). Starting with 340 group the data into classes of interval 10 and draw an Ogive. From the obtained frequency table calculate

(b). Using a suitable assumed mean, calculate the mean and the standard deviation.

(c). Using coding method, calculate the mean and the standard deviation.

(d). The coefficient of variation.

(e). Draw a histogram and determine the mode.

4. The table below gives the monthly costs of some living necessities in two towns A and B. Each necessity has been given a weight as a measure of its importance to basic living. Christine has just been enrolled in a college near the two towns and is contemplating residing in one of the two towns.

Item	Cost (town A)	Cost (town B)	Weight
Food	1800	1700	2.0
Clothes	1000	1050	1.8
Shelter	8800	8300	2.2
Transport & comm	1500	1800	1.4
Security	1300	1200	1.6
Miscellaneous	2300	4300	1.0

Taking A as the base town, calculate the cost of living index and advice Christine accordingly.

5. The table below gives the probability distribution of a discrete random variable X given that $P(X < 13) = 0.75$. Find the value of k and q, hence calculate $E(X)$.

x	4	8	12	15	20
$P(X = x)$	k	0.25	0.3	q	0.1

KIMATHI UNIVERSITY COLLEGE OF TECHNOLOGY University Examinations 2012/2013

EXAMINATION FOR THE DEGREE OF BACHELOR OF SCIENCE IN **INFORMATION TECHNOLOGY** AND BACHELOR OF SCIENCE IN **ACTUARIAL SCIENCE**

SMA 2103/STA 2100: Probability and Statistics I

DATE: 23RD April 2012 TIME: 2 HOURS

Instructions: Answer QUESTION ONE and any other TWO QUESTIONS.

QUESTION ONE (30 marks) (COMPULSORY)

(a) The masses of 90 eggs to the nearest gram were recorded as follows Assuming that the are linearly

masses (g)	50-59	60-64	65-69	70-79
no. of eggs	18	20	x	y

distributed in each group and that 60% of the eggs have masses below 66.5g ,estimate the

(i) values of x and y . **[4 marks]**

(ii) mass of the most of the eggs . **[3 marks]**

(iii) mean and standard deviation of the distribution,using the coding method and an assumed mean of 67g . **[6 marks]**

(iv) Why are the values obtained in (ii)and (iii)above estimates of the actual values. **[1 mark]**

(b) An inspector working for a manufacturing company has a 99% chance of correctly identifying defective items and a 0.5% chance of incorrectly classifying a good item as defective. The company has evidence that its line produces 0.9% of nonconforming items.With the aid of a tree diagram,

i. What is the probability that an item picked at random for inspection is defective? **[2 marks]**

ii. If an item selected at random is classified as defective, what is the probability that it is indeed good? **[5 marks]**

iii. The marks obtained by eight students in maths and programming are as shown above. Calculate to 4 d.p. the Spearman's rank correlation coefficient and

maths	67	24	85	51	39	97	81	70
programming	70	59	71	38	55	62	80	76

comment on your results. **[6 marks]**

iv. Find the interquartile range of the following set of data

4,6,18,25,9,16,22,5,20,8 **[3 marks]**

QUESTION TWO (20 marks)

An old film is treated so as to improve its contrast. Preliminary tests on nine samples from a segment of the film produced the following results.

where x is the amount of chemical applied and y is the contrast index which is measured from 0 (no contrast)to 100 (max. contrast)

(a) Plot a scatter diagram. **[3 marks]**

sample	1	2	3	4	5	6	7	8	9
x	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
y	49	60	66	62	72	64	89	90	96

(b) It is discovered that one of the samples of the film was damaged and gave incorrect results. State which sample this could be. **[1 mark]**

(c) Ignoring the sample that was damaged calculate to 2d.p. the product moment correlation coefficient. **[8 marks]**

(d) State with reason whether it is sensible to conclude that x and y are linearly related. **[2 marks]**

(e) Obtain the regression line y on x. **[3 marks]**

(f) Use the regression line to estimate the contrast index corresponding to the damaged piece of film. **[2 marks]**

(g) State with reason whether it is sensible to estimate the contrast index when the amount of the chemical applied is zero. **[1 mark]**

QUESTION THREE (20 marks)

(a) Define the following terms as used in probability .

(i) Possibility space. **[1 mark]**

(ii) Exhaustive events **[1 mark]**

(iii) Independent events **[1 mark]**

(b) Two events A and B are such that $P(A) = \frac{8}{15}$, $P(B) = \frac{1}{3}$ and $P(A|B) = \frac{1}{5}$. Calculate the probability that:-

(i) Both events occur. **[2 marks]**

(ii) Only one of the events occurs. **[2 marks]**

(iii) None of the events occur. **[2 marks]**

(c) A group of 140 college students study maths. Each student takes only algebra, or statistics only or both algebra and statistics. If a student takes Statistics the probability that he takes Algebra is $\frac{1}{3}$while the probability that takes he statistics given that he takes algebra is $\frac{1}{5}$. Find the number of students taking

(i) Both algebra and statistics . **[5 marks]**

(ii) Only algebra. **[2 marks]**

(d) In large number of pens, the probability that a pen is defective is $\frac{1}{40}$. A student bought two such pens. Use a tree diagram to show the possible outcomes and hence compute the probability that at least one of the pens is defective. **[4 marks]**

QUESTION FOUR (20 marks)

(a) A game consists of tossing three unbiased coins simultaneously. The total score is calculated by awarding three points for each head and one point for each tail. Let the random variable X represents the total score

(i) Write down the probability distribution table for X. **[3 marks]**

(ii) Obtain the mean and variance of X. **[5 marks]**

(iii) State the mode of X. **[1 mark]**

(b) A discrete random variable X takes only the values $0,1,2,3,4,5$. The probability distribution of X is as follows

$$P(X=0)=P(X=1)=P(X=2)=p$$
$$P(X=3)=P(X=4)=P(X=5)=q$$
$$P(X\geq 2)=3P(X<2)$$

Determine

(i) The values of p and q. **[5 marks]**

(ii) The probability that the sum of any two independent observations from the distribution is 7. **[3 marks]**

(c) A discrete random variable X has p.m.f. $f(x)$.

Show that $Var(kX) = k^2 Var(X)$ **[3 marks]**

QUESTION FIVE (20 marks)

(a) Given a set of data $x_1, x_2, \ldots, x_n$ occurring with frequencies $f_1, f_2, \ldots, f_n$ respectively ,define

(i) The r^{th} moment about a point a. **[1 mark]**

(ii) The r^{th} moment about the origin . **[1 mark]**

(b) Let m_r' be the r^{th} moment about a point a and m_r be the r^{th} moment about the mean. Show that $m_4 = m_4 - 4m_1'm_3 + 6[m_1]^2m_1' - 3[m_1]^4$ **[5 marks]**

(c) The marks obtained in a test by a class of 120 students were as follows

marks	30-39	40-49	50-59	60-69	70-79	80-89	90-99
number of students	9	32	43	21	11	3	1

By taking moments about the point 54.5, investigate the skewness and peakedness of distribution of the marks. **[13 marks]**

KIMATHI UNIVERSITY COLLEGE OF TECHNOLOGY

University Examinations 2012/2013

FIRST YEAR SECOND SEMESTER EXAMINATION FOR THE DEGREE OF **BACHELOR OF COMMERCE**

HBC 2121 Introduction to Business Statistics

DATE: 23RD APRIL 2012 TIME: 11.00 AM- 1.00 PM

Instructions: Answer QUESTION ONE and any other TWO QUESTIONS.

QUESTION ONE (30 marks) (COMPULSORY)

(a). Most undergraduate business students will not go on to become actual practitioners of statistical research and analysis. Considering this fact, why should such individuals bother to become familiar with business statistics? **[2 marks]**

(b). Given two events A and B such that $P(A|B) = 0.4$, $P(A) = 0.06$, $P(B) = 0.10$. Find

(i) $P(A \cap B)$ and $P(A \cup B)$ **[2 marks]**

(ii) Determine whether the events A and B are mutually exclusive, independent or collectively exhaustive events. **[3 marks]**

(c). In a game organized by Prof. Makoha, the score for the game is a random variable X which takes a set of values $\{0,1,2,3,4\}$ according to the probability distribution below:

x	0	1	2	3	4
$P(x)$	0.1	a	b	0.2	0.1

Given the expectation (mean) of x is 1.8 determine the values of a and b and hence the variance of x. **[6 marks]**

(d). Given below is an incomplete frequency distribution of masses (in kg) of 100 students in a college. The classes are are of equal width.

Class Interval	Class boundary	Class midpoints	Frequency	Cumulative Frequency
20 - 29	19.5 - 29.5		3	
			8	
			?	
			18	
			25	
70 - 79			17	
			9	
			5	

(i). Fill the columns of the frequency distribution table. **[3 marks]**

(ii). Calculate the second quartile of the masses of students. **[3 marks]**

(iii). Find the modal class and calculate the mode. **[3 marks]**

(iv). Calculate the mean and standard deviation the data. **[5 marks]**

(e) Distinguish between the moment coefficient of skewness and moment coefficient of kurtosis. **[2 marks]**

QUESTION TWO (20 marks)

(a). Two data sets are such that data set A collected on a random variable X has $\sum_{i=1}^{30} x = 240$ and $\sum_{i=1}^{30} x^2 = 2520$ and data set B of ten observations collected on a random variable Y has a mean of 5 and variance of 6, find;

(i). Find the mean and the variance of data set A **[3 marks]**

(ii). The mean and variance of combined data sets **[5 marks]**

(iii). Write down the answers to (i) and (ii) if each of original values is multiplied by 2 **[2 marks]**

(b). A Battery manufacturer was interested in predicting the annual mantainance cost of the battery manufacturing machines based upon the age of the machine. A sample of ten machines revealed the following ages and mantainance costs during the previous year.

Age(years)	9	4	2	8	4	5	1	3	6	8
cost	40	12	8	27	15	17	5	10	25	31

(i) Define regression analysis. **[1 mark]**

(ii) Fit the simple linear regression model to the data using the least squares method. **[7 marks]**

(iii) Interpret the meaning of the slope in the linear regression model in part i). **[1 mark]**

(iv) Predict the mantainance cost of a machine that is 10 years old. **[1 mark]**

QUESTION THREE (20 marks)

(a). The following table gives the weights in kilograms of a certain product from some farmers. Using an assumed mean of $A = 655$;

Weight	310-400	410-500	510-600	610-700	710-800	810-900
Frequency	8	14	18	20	11	9

(i). Construct the coding method table. **[4 marks]**

(ii). Find the mean weight using the coding method **[2 marks]**

(iii). Find the standard deviation using the coding method **[4 marks]**

(b). An aircraft emergency locator transmitter (ELT) is a device designed to transmit a signal in the case of a crash. The Ultimate Manufacturing company makes 80% of the ELTs, the Bryant Company makes 15% of them, and the Charter air Company makes the other 5%. The ELTs made by Ultimate have a 4% rate of defects, the Bryant ELTs have a 6% rate of defects, and the Charter air ELTs have a 9% rate of defects (which helps to explain why Charter air has the lowest market share).

An ELT is randomly selected from the general population of all ELTs then tested and is found to be defective

(i) Find the probability that the ELT is defective. **[4 marks]**

(ii) If a randomly selected ELT is defective, find the probability that it was made by the Ultimate manufacturing company.**[3 marks]**

(iii) If a randomly selected ELT is defective, find the probability that it was made by the Charterair manufacturing company.**[3 marks]**

QUESTION FOUR(20 marks)

(a). The table below gives a probability distribution of a discrete random variable X. Given that $P(X < 130) = 0.62$, find the value of a and b hence calculate the standard deviation of (X).

x	40	80	120	150	200
$P(X = x)$	a	0.22	0.25	b	0.15

[6 marks]

(b). The following is information on the rail distance to destination and transportation times for ten shipments by a spare parts supplier data.

Customer:	A	B	C	D	E	F	G	H	I	J
Distance (X):	270	290	350	480	490	730	780	850	920	1010
Time(Days) (Y):	5	7	6	11	8	11	12	8	15	12

(i) Fit a regression line to the data. **[6 marks]**

(ii) Compare the regression line above with that obtained by regressing transportation time on the rail distance. **[2 marks]**

(iv) Calculate the Pearson's the coefficient of correlation. **[4 marks]**

(v) Calculate the coefficient of determination. **[2 marks]**

QUESTION FIVE (20 marks)

(a). The following table shows the sample prices and corresponding weights of various food items in Kenya for the year 2011 which were computed with respect to the 2009 as the base year.

Products	Weight (W)	Price Index (I)
Milk	16	148
Sugar	16	40
Wheat flour	k	129
Maize flour (5 kg)	26	111
Margarine (1 kg)	18	200
Rice (2 kg)	3	194

(i). If the price of maize flour was 300 in the year 2011. Find its price in 2009. **[4 marks]**

(ii). The combined composite index shows that the cost of items has increased by 28%. Find the value of k. **[6 marks]**

(b). Suppose that you have real time series data as

56,59,45,25,36,58,79

(ii). Compute a 4-point centered moving average. **[4 marks]**

(c). Distinguish between the following terms as used in Hypothesis testing

(i) Type I error and Type II error

(ii) Null hypothesis and Alternative hypothesis

(iii) Small sample test and large sample tests; give an example of the test.

[6 marks]

KIMATHI UNIVERSITY COLLEGE OF TECHNOLOGY

University Examinations 2011/2012

FIRST YEAR SECOND SEMESTER EXAMINATION FOR THE DEGREE OF **BACHELOR OF COMMERCE**

HBC 2121 Introduction to Business Statistics

DATE: 5 TH DECEMBER 2011 TIME: 2.00 PM-4.00 PM

Instructions: Answer QUESTION ONE and any other TWO QUESTIONS.

QUESTION ONE (30 marks) (COMPULSORY)

(a). Define the following terms as used in statistics.

(i) Nominal measurements

(ii) Ordinal measurements

(iii) Independent events

(iv) Coefficient of variation

(v) Mutually exclusive events **[5 marks]**

(b). The marks obtained by 30 students in a mathematics test marked out of 20 are as shown below

Marks	8	9	10	11	12	14	15	17	18	20
No. of Students	2	3	4	4	5	3	3	3	2	1

Compute

(i) The mode

(ii) Median

(iii) Mean

(iv) Standard deviation **[7 marks]**

(c). The first of the two groups has 100 items with mean 45 and variance 49. If the combined group has 250 items with mean 51 and variance 130, find the mean and standard deviation of the second group. **[5 marks]**

(d). The following data refers to Examination marks verses hours of study per week of a sample of eight candidates that sat for Business statistics examination in 2010.

Exam Mark (Y)	64	61	84	70	88	92	72	72
Hours of Study (X)	20	16	34	23	27	32	18	22

Calculate:

(i) the Pearson's product moment coefficient of correlation. **[5 marks]**

(ii) the coefficient of determination between examination marks and hours of study. **[2 marks]**

(e). If 50% of the families subscribe to the morning newspaper, 65% of the families subscribe to the afternoon newspaper and 85% of the families subscribe to at least one of the two newspapers, what proportion of the families subscribe to both newspapers? **[3 marks]**

(f). (i) Under what conditions does $P(A|B) = P(A)$? **[1 marks]**

(ii) What is the addition rule of probability and for what type of events is it valid? **[2 marks]**

QUESTION TWO (20 marks)

(a). What is an index number?

Explain two areas where index numbers are applied. **[2 marks]**

(b). If the Fisher's price index is 109.91 and the Paasche's price index is 110.6, calculate the Laspeyre's index number. **[2 marks]**

(c). (i) Define the term "time series" and giving an example. **[2 marks]**

(ii) What is the aim of time series analysis? **[2 marks]**

(iii) By giving the relevant examples, briefly explain the components of a time series. **[6 marks]**

(d). The number of new stereo systems sold by an electrical store each quarter for four years is shown below.

Year	Quarters	Sales
2000	1	130
	2	105
	3	85
	4	120
2001	1	150
	2	115
	3	95
	4	140
2002	1	165
	2	120
	3	100
	4	140
2003	1	180
	2	140
	3	110
	4	160

Calculate a four point moving average trend. **[6 marks]**

QUESTION THREE (20 marks)

(a). The following table gives the marks obtained by first year students in a Marketing examination.

Marks	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
No. of Students	3	9	14	24	18	16	9	4

Calculate

(i) the mean

(ii) the Standard deviation

(iii) the median, and

(iv) the mode **[10 marks]**

(b). (i) State Bayes Theorem. **[2 marks]**

(ii) KY accounting firm has noticed that of the companies it audits, 85% show no inventory shortages, 10% show small inventory shortages and 5% show large inventory shortages. KY firm has devised a new accounting test for which it believes the following probabilities hold:

$$P(\text{company will pass test} \mid \text{no shortages}) = 0.90$$
$$P(\text{company will pass test} \mid \text{small shortages}) = 0.50$$
$$P(\text{company will pass test} \mid \text{large shortages}) = 0.20$$

Determine the probability if a company being audited fails this test has large or small inventory shortages.**[8 marks]**

QUESTION FOUR (20 marks)

(a). Statistics is a means of collection of numerical facts on data. What are the major advantages of sampling method over the census? **[4 marks]**

(b). The random variable X has a probability distribution shown in the table below

X	0	10	20	30
$P(X = x)$	0.1	p	0.45	q

Given that the mean of X is 16.5, find the value of p and q. **[4 marks]**

(c). The following data shows two groups of casual workers, their number and average wages they are paid.

	Group A	Group B	Combined Groups
Number of Workers	X	100	250
Mean wage	600	Y	6800
Standard Deviation	8	9	Z

(i) Calculate the missing entries; X, Y , and Z. **[8 marks]**

(ii) Calculate the coefficient of variation for Groups A and B. **[3 marks]**

(iii) From (ii) above, which group has a greater variability in income? **[1 marks]**

QUESTION FIVE (20 marks)

In a study of the daily production of a company for 50 days, the following data was obtained:

65	76	36	48	49	48	84	55	79	51
43	21	78	35	37	61	40	45	68	33
88	45	50	53	60	34	56	67	57	42
59	62	62	65	76	55	76	61	70	73
35	41	60	74	52	82	63	58	32	26

(i) Starting with a class $20 - 29$, group the data into a frequency distribution, represent the data in a histogram and plot its ogive. **[8 marks]**

(ii) State the modal class. **[1 marks]**

(iii) Using a suitable assumed mean, calculate the mean hence or otherwise calculate, the mode, the median and the standard deviation of the data above. **[11 marks]**

KIMATHI UNIVERSITY COLLEGE OF TECHNOLOGY

University Examinations 2011/2012

SECOND YEAR FIRST SEMESTER EXAMINATION FOR THE DIPLOMA IN **PURCHASING AND SUPPLIES**

HPS/BCA 1303 BUSINESS STATISTICS

DATE: TH AUGUST 2011 TIME:2 HOURS

Instructions: Answer QUESTION ONE and any other TWO QUESTIONS.

QUESTION ONE (30 marks) (COMPULSORY)

QUESTION ONE (30 marks) (COMPULSORY)

(a). Giving at least one example in each case, distinguish between:-

(i). Qualitative and quantitative variables

(ii) Independent and mutually exclusive events **[4 marks]**

(b). State four reasons why statisticians would prefer to use sample data instead of population data. **[4 marks]**

(c). Find the mean, the mean absolute deviation and the 35th percentile of the following data.

$4, 6, 20, 3, 16, 13, 8, 23, 10, 25, 15$ **[7 marks]**

(d). A fair coin is tossed and a fair die is thrown at the same time. Find the probability of getting a head and a three (3) at the same time. **[2 marks]**

(e). The table below gives the probability distribution of a discrete random variable X given that $P(X < 13) = 0.75$. Find the value of k and q, hence calculate $E(X)$.

x	4	8	12	15	20
$P(X = x)$	k	0.25	0.3	q	0.1

[5 marks]

(f). State the four main component movements of a time series. **[4 marks]**

(g). For a certain data set of 20 values, $\sum x = 154$ and $\sum x^2 = 2045$. Find the mean and standard deviation of the data set after dropping a value of 19 from the data set. **[4 marks]**

QUESTION TWO (20 marks) (Optional)

(a). Twenty staff members in a construction company were surveyed to find out what their weekly wages were in euros. The results were as follows:

49.05	73.10	72.75	58.65	63.00	42.75	53.25	60.00	61.35	49.80
54.90	63.75	51.75	66.00	59.25	51.30	53.55	49.20	38.10	58.95

(i). State whether the data is discrete or continuous. **[1 marks]**

(ii). Determine the appropriate class interval and present this data in a frequency distribution table starting with the value of 38.00. **[5 marks]**

(iii). Draw a properly labeled histogram and a frequency polygon on the same graph for the data above. **[4 marks]**

(b). The table below shows the distribution of the weights of 100 students in a university.

Weight in kgs.	Number of students
60 - 62	5
63 - 65	18
66 - 68	42
69 - 71	27
72 - 74	8

(i). Calculate the mean weight. **[2 marks]**

(ii). Calculate the standard deviation of this data. **[3 marks]**

(iii). Calculate the median weight and the inter-quartile range. **[3 marks]**

(iv). Find the position of the person who weighs 70.6 kgs. **[2 marks]**

QUESTION THREE (20 marks) (Optional)

(a). In a certain company 65% of the workers can speak English, 75% can speak Kiswahili, while 15 % can neither speak English nor Kiswahili. An employee is randomly picked from this group. Find the probability that the person speaks;

(i). Kiswahili but not English **[4 marks]**

(ii). Both languages **[4 marks]**

(b). Marion can neither take a course in computers or in chemistry. If she takes the computer cause, then she will score an A with probability 0.5, if she takes the chemistry course, then she will score an A grade with probability$\frac{1}{3}$. Marion decides to base her decision on the flip of a coin.

(i). What is the probability that she will score an A in chemistry? **[6 marks]**

(ii). Given that Marion scores an A, what is the probability that she took the computer course. **[6 marks]**

QUESTION FOUR (20 marks) (Optional)

(a). The table below gives the monthly costs of some living necessities in two towns A and B. Each necessity has been given a weight as a measure of its importance to basic living. Christine has just been enrolled in a college near the two towns and is contemplating residing in one of the two towns.

Item	Cost (town A)	Cost (town B)	Weight
Food	1800	1700	2.0
Clothes	1000	1050	1.8
Shelter	8800	8300	2.2
Transport & comm	1500	1800	1.4
Security	1300	1200	1.6
Miscellaneous	2300	4300	1.0

Taking A as the base town, calculate the cost of living index and advice Christine accordingly. **[10 marks]**

(b). The data below gives sales in millions for a certain company in Kenya from 2000-2008.

Year	2000	2001	2002	2003	2004	2005	2006	2007	2008
Sales	2.3	1.5	1.7	2.2	2.5	2.9	2.5	3.0	2.9

Obtain smoothened values using 4-point moving averages. **[8 marks]**

(c). Two data sets are such that the first has 150 items with a mean of 55. If the combined set has 250 items with a mean of 51, calculate the mean of the second set. **[2 marks]**

QUESTION FIVE (20 marks) (Optional)

(b). If $P(A) = 0.6$, $P(A|B) = 0.4$,

(i). $P(A \cap B)$, **[2 marks]**

(ii). $P(A \cup B)$, **[2 marks]**

(iii). $P(B|A)$ state whether A and B independent events **[4 marks]**

(c). Find the number of classes and the class intervals for data with the following characteristic.

Maximum value 467, minimum value 323, sample size 50. **[2 marks]**

(b). A random variable X has probability distribution shown in the table below;

x	10	15	20	24
$P(x)$	a	0.24	$2a$	0.28

Find:

(i). the value of a. **[3 marks]**

(ii). the expected value of X **[3 marks]**

(c). A bag contains some rotten eggs and 40 good ones, it the probability of randomly picking a rotten egg from it is 1/5, find the number of rotten eggs. **[4 marks]**

KENYA METHODIST UNIVERSITY

END OF THIRD TRIMESTER 2009 EXAMINATIONS

FACULTY : BUSINESS AND MANAGEMENT STUDIES

DEPARTMENT : BUSINESS ADMINISTRATION

COURSE CODE : BUSS 113

COURSE TITLE : BUSINESS STATISTICS I

TIME : 2 HOURS

INSTRUCTIONS:

Answer Question ONE (COMPULSORY) and any other TWO Questions.

QUESTION ONE (30 marks) (COMPULSORY)

(a). Most undergraduate business students will not go on to become actual practitioners of statistical research and analysis. Considering this fact, why should such individuals bother to become familiar with business statistics? **[3 marks]**

(b). In a study of the daily production of a company for over 50 days, the following data was obtained:

65	76	36	48	49	48	84	55	79	51
43	21	78	35	37	61	40	45	68	33
88	45	50	53	60	34	56	67	57	42
59	62	62	65	76	55	76	61	70	73
35	41	60	74	52	82	63	58	32	26

(i). Starting with a class $20 - 29$, group this data into a frequency distribution, and plot its ogive.**[6 marks]**

(ii). Using a suitable assumed mean, calculate the mean and standard deviation of this data. **[5 marks]**

(c). Consider the following data whose mean is 9.

3	5	1	13	6	10	8
11	12	17	23	X	0	

Determine

(i). the value of X **[2 marks]**

(ii). the third decile of the data **[3 marks]**

(iii). the 65th percentile of the data **[2 marks]**

(iv). the mode of the data **[1 marks]**

(d). Given two events A and B such that $P(A|B) = 0.8$, $P(A) = 0.5$, and $P(B) = 0.25$.

Determine

i). $P(A \cap B)$. **[1 marks]**

ii). $P(A \cup B)$. **[1 marks]**

iii). Whether the two events A and B are mutually exclusive, independent or collectively exhaustive events. **[6 marks]**

QUESTION TWO (20 marks) (Optional)

(a). Explain the meaning of the following terms

i). Sample Space **[1 marks]**

ii). Mutually exclusive events **[1 marks]**

iii). Independent events **[1 marks]**

iv). Exhaustive events **[1 marks]**

(b). The weather in Nyeri over 18 months was observed and the mean temperatures in degrees Celsius reported as shown below:

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2007	27	25	22	20	17	15	13	16	19	18	20	23
2008	24	23	20	20	18	12						

(i). Is this data periodic? **[2 marks]**

(ii). Calculate the 6-month moving averages **[5 marks]**

(iii). On the same graph, plot the raw data and its moving averages. **[3 marks]**

(c). A sample of 40 electric batteries gives a mean life span of 600 hrs with a standard deviation of 20 hours. Another sample of 50 electric batteries gives a mean lifespan of 520 hours with a standard deviation of 30 hours. If these two samples were combined and used in a given project simultaneously, determine the combined new mean for the larger sample and hence determine the combined or pulled standard deviation. **[6 marks]**

QUESTION THREE (20 marks) (Optional)

(a). The following data have been collected relating to returns which would have been earned from an investment of an equal sum of money in the shares of E.T. Plc. and a group of market shares.

Year:	1	2	3	4	5	6	7	8	9	10
E.T. Plc shares (Y):	7.8	11.0	15.2	23.1	29.7	37.4	44.6	52.8	60.2	63.9
Mkt shares (X):	11.1	12.3	18.5	25.4	28.7	33.8	37.7	39.6	44.7	45.5

Calculate the least squares regression of Y on X. **[12 marks]**

(b). Recent unit prices in hundreds of shillings of various fruits and vegetables (items) in Nairobi and Nakuru were as follows:

Item:	A	B	C	D	E	F	G
Nairobi (X):	14	16	16	9	8	28	35
Nakuru (Y):	9	18	20	15	6	26	38

$$\Sigma X = 126,\ \Sigma Y = 132,\ \Sigma XY = 2975,\ \Sigma X^2 = 2862,\ \Sigma Y^2 = 3186$$

(i). Calculate the product moment coefficient of correlation. **[6 marks]**

(ii). Calculate the coefficient of determination. **[2 marks]**

QUESTION FOUR (20 marks) (Optional)

(a). The weights in grams of some 25 computer parts are given in the following frequency distribution table below

Weights	160.0–169.9	170.0–179.9	180.0–189.9	190.0–199.9	200.0–209.9
No. of parts	3	X	9	6	3

Calculate

(i). The missing frequency and hence find the mode. **[4 marks]**

(ii). Using the assumed mean A = 184.95. Find the mean and the standard deviation of the weights using the Coding method. **[6 marks]**

(iii). Calculate the third quartile of the weight. **[4 marks]**

(b). At a supermarket 60% of the customers pay using the credit card. Find the probability that in a randomly selected set of 10 customers.

(i). Exactly 2 pay by credit card. **[2 marks]**

(ii). None pays by credit card. **[1 marks]**

(iii). Determine the mean and standard deviation of the customers paying using credit cards. **[3 marks]**

QUESTION FIVE (20 marks) (Optional)

(a). Find the probabilities that a random variable having a Standard Normal distribution will take on a value.

(i). greater than 1.72 **[1 marks]**

(ii). less than -0.88 **[1 marks]**

(iii). between 1.30 and 1.75 **[2 marks]**

(b). Three machines A, B, and C produce respectively 60%, 30% and 10% of the total number of items of a factory. The percentages of defective output of these machines are respectively 2%, 3%, and 4 %.

An item is selected at random from the product and is found to be defective. Find the probability that the item was produced by machine C. **[8 marks]**

(c). A discrete random variable has the following probability distribution:

X	0	1	2	3
$P(X = x)$	p	$2q$	$p + q$	q

If $E(X) = 1.375$, find

(i). The value of p and q. **[3 marks]**

(ii). The standard deviation of X. **[4 marks]**

(iii). The probability that X exceeds the mean. **[1 marks]**

- The following data was observed and it is required to establish if there exists a relationship between the two.

X	15	24	25	30	35	40	45	65	70	75
Y	60	45	50	35	42	46	28	20	22	15

Compute the product moment coefficient of correlation (r).

Differentiate the following terms as used in statistics giving an example for each:

(i). Sample and population **[2 marks]**

(ii). Discrete and continuous random variables **[2 marks]**

KENYA METHODIST UNIVERSITY

END OF THIRD TRIMESTER 2009 EXAMINATIONS

FACULTY : BUSINESS AND MANAGEMENT STUDIES

DEPARTMENT : BUSINESS ADMINISTRATION

COURSE CODE : DPBA 018

COURSE TITLE : INTRODUCTION TO STATISTICS

TIME: $1\frac{1}{2}$**HOURS**

INSTRUCTIONS:

Answer Question ONE (COMPULSORY) and any other TWO Questions.

QUESTION ONE (30 marks) (COMPULSORY)

(a). Giving at least one example in each case, distinguish between:-

(i). Qualitative and quantitative variables

(ii) Independent and mutually exclusive events **[4 marks]**

(b). State four reasons why statisticians would prefer to use sample data instead of population data. **[4 marks]**

(c). Find the mean, the mean absolute deviation and the 35th percentile of the following data.

4, 6, 20, 3, 16, 13, 8, 23, 10, 25, 15

[7 marks]

(d). A fair coin is tossed and a fair die is thrown at the same time. Find the probability of getting a head and a three (3) at the same time. **[2 marks]**

(e). The table below gives the probability distribution of a discrete random variable X given that $P(X < 13) = 0.75$. Find the value of k and q, hence calculate $E(X)$.

x	4	8	12	15	20
$P(X = x)$	k	0.25	0.3	q	0.1

[5 marks]

(f). State the four main component movements of a time series. **[4 marks]**

(g). For a certain data set of 20 values, $\sum x = 154$ and $\sum x^2 = 2045$. Find the mean and standard deviation of the data set after dropping a value of 19 from the data set. **[4 marks]**

QUESTION TWO (20 marks) (Optional)

(a). Twenty staff members in a construction company were surveyed to find out what their weekly wages were in euros. The results were as follows:

49.05	73.10	72.75	58.65	63.00	42.75	53.25	60.00	61.35	49.80
54.90	63.75	51.75	66.00	59.25	51.30	53.55	49.20	38.10	58.95

(i). State whether the data is discrete or continuous. **[1 marks]**

(ii). Determine the appropriate class interval and present this data in a frequency distribution table starting with the value of 38.00. **[5 marks]**

(iii). Draw a properly labeled histogram and a frequency polygon on the same graph for the data above. **[4 marks]**

(b). The table below shows the distribution of the weights of 100 students in a university.

Weight in kgs.	Number of students
60 - 62	5
63 - 65	18
66 - 68	42
69 - 71	27
72 - 74	8

(i). Calculate the mean weight. **[2 marks]**

(ii). Calculate the standard deviation of this data. **[3 marks]**

(iii). Calculate the median weight and the inter-quartile range. **[3 marks]**

(iv). Find the position of the person who weighs 70.6 kgs. **[2 marks]**

QUESTION THREE (20 marks) (Optional)

(a). In a certain company 65% of the workers can speak English, 75% can speak Kiswahili, while 15 % can neither speak English nor Kiswahili. An employee is randomly picked from this group. Find the probability that the person speaks;

(i). Kiswahili but not English **[4 marks]**

(ii). Both languages **[4 marks]**

(b). Marion can neither take a course in computers or in chemistry. If she takes the computer cause, then she will score an A with probability 0.5, if she takes the chemistry course, then she will score an A grade with probability $\frac{1}{3}$. Marion decides to base her decision on the flip of a coin.

(i). What is the probability that she will score an A in chemistry? **[6 marks]**

(ii). Given that Marion scores an A, what is the probability that she took the computer course. **[6 marks]**

QUESTION FOUR (20 marks) (Optional)

(a). The table below gives the monthly costs of some living necessities in two towns A and B. Each necessity has been given a weight as a measure of its importance to basic living. Christine has just been enrolled in a college near the two towns and is contemplating residing in one of the two towns.

Item	Cost (town A)	Cost (town B)	Weight
Food	1800	1700	2.0
Clothes	1000	1050	1.8
Shelter	8800	8300	2.2
Transport & comm	1500	1800	1.4
Security	1300	1200	1.6
Miscellaneous	2300	4300	1.0

Taking A as the base town, calculate the cost of living index and advice Christine accordingly. **[10 marks]**

(b). The data below gives sales in millions for a certain company in Kenya from 2000-2008.

Year	2000	2001	2002	2003	2004	2005	2006	2007	2008
Sales	2.3	1.5	1.7	2.2	2.5	2.9	2.5	3.0	2.9

Obtain smoothened values using 4-point moving averages. **[8 marks]**

(c). Two data sets are such that the first has 150 items with a mean of 55. If the combined set has 250 items with a mean of 51, calculate the mean of the second set. **[2 marks]**

QUESTION FIVE (20 marks) (Optional)

(b). If $P(A) = 0.6$, $P(A|B) = 0.4$,

(i). $P(A \cap B)$, **[2 marks]**

(ii). $P(A \cup B)$, **[2 marks]**

(iii). $P(B|A)$ state whether A and B independent events **[4 marks]**

(c). Find the number of classes and the class intervals for data with the following characteristic.

Maximum value 467, minimum value 323, sample size 50. **[2 marks]**

(b). A random variable X has probability distribution shown in the table below;

x	10	15	20	24
$P(x)$	a	0.24	$2a$	0.28

Find

(i). the value of a. **[3 marks]**

(ii). the expected value of X **[3 marks]**

(c). A bag contains some rotten eggs and 40 good ones, it the probability of randomly picking a rotten egg from it is 1/5, find the number of rotten eggs. **[4 marks]**

16.0.1 Bernoulli Distribution

The Bernoulli distribution is an example of a discrete probability distribution. It is an appropriate tool in the analysis of proportions and rates.

A Bernoulli trial is a random experiment in which there are only two possible outcomes - *success* and *failure*.

- Tossing a coin and considering heads as success and tails as failure.
- Checking items from a production line: success = not defective, failure = defective.
- Phoning a call centre: success = operator free; failure = no operator free. A Bernoulli random

variable X takes the values 0 and 1 and

$$P(X = 1) = p$$
$$P(X = 0) = 1 - p$$

If an experiment has two possible outcomes, 'success' and 'failure' and their probabilities are respectively p and $1 - p$, then the number of successes, 0 or 1 has a Bernoulli distribution.

Definition 16.1. A random variable X has a Bernoulli distribution and it is referred to as a bernoulli random variable if and only if its probability distribution is given by

$$f(x;p) = p^x(1 - p)^{1-x} \quad for \qquad x = 0,1$$

In connection with the Bernoulli distribution, a success may be getting heads with a balanced coin, it may be catching pneumonia, it may be passing (or failing) an examination and it may be losing a race.

Note: Bernoulli distribution is a special case of the Binomial distribution. The mean and variance of the Bernoulli distribution are given as

$$\mathrm{E}(x) = p \qquad \text{and} \qquad \mathrm{Var}(x) = \sigma^2 = p(1 - p)$$

Question 16.1. Show that $\mathrm{E}(X) = p$ and $\mathrm{Var}(X) = p(1-p)$ for a random variable X which has a bernoulli distribution.

$$\begin{aligned}\mathrm{E}(X) = \sum_{allx} xP(x) &= \sum_{x=0}^{1} p^x(1-p)^{1-x} \\ &= 0 \cdot p^0(1-p)^1 + 1 \cdot p(1-p)^{1-1} \\ &= 0 + p(1-p)^0 = p\end{aligned}$$

$$\begin{aligned}\text{Var}(X) = \sum_{allx}(x-\mu)P(x) &= E(X^2) - E(X)^2 \\ &= 0^2(1-p) + 1^2(p-p^2) \\ &= p - p^2 = p(1-p).\end{aligned}$$

The moment generating function of a Bernoulli distribution is given by

$$\begin{aligned}M_X(t) = \text{E}(e^{tx}) &= \sum_{x=0}^{1} \cdot p^x(1-x)^{1-x} \\ &= e^0 \cdot p^0(1-p)^1 + e^t \cdot p(1-p)^0 \\ &= (1-p) + e^t p \\ &= 1 - p + e^t p \\ &= P(e^t - 1) + 1\end{aligned}$$

Example 16.1. A carton contain 4 good eggs and 6 bad eggs. If an egg is selected at random, then the random variable

$$X = \begin{cases} 0 \text{ if the egg is bad if the egg is good} \\ 1 \end{cases}$$

$$P(\text{good egg}) = \frac{4}{10} = \frac{2}{5}$$

$$P(X = 1) = (2/5)^1 \cdot (1 - 2/5)^0 = 2/5$$

$$E(X) = 2/5, Var(X) = p(1-p) = 2/5(1 - 2/5) = 6/25.$$

BERNOULLI TRIALS: Many experiments consist of a sequence of trials, where

(i) each trial results in a "success" or a "failure,"

(ii) the trials are **independent**, and

(iii) the probability of "success," denoted by p, $0 < p < 1$, is the **same** on every trial.

If the above conditions are satisfied then X is said to follow a binomial distribution. If a Bernoulli experiment is performed repeatedly, we obtain a sequence of **bernoulli trials**. In a finite sequence of bernoulli trials, we are usually interested in the number of 'success'. Since there are only two possible outcomes at each trial a sample of "n" bernoulli trials contains 2^n possible outcomes. As before $X = 0, 1$ for each trial. The probability of any particular sequence of outcomes in 'n' trials is obtained as the product of the probabilities of the n outcomes resulting at each trial.

Definition 16.2. A random variable X has a binomial distribution and it is referred to as a binomial random variable if and only if its probability distribution is given by

$$f(x;n,\theta)=\binom{n}{x}p^x(1-p)^{n-x} \qquad \text{for} \qquad x=0,1,2,\cdots,n$$

where
n - number of trials
p - probability of success
$1-p$ - probability of failure
x - is the random variable (the number of successes in n trials).
$\binom{n}{x}=\frac{n!}{x!(n-x)!}$

Thus, the number of successes in n trials is a random variable having a binomial distribution with parameters n and p. The name '**binomial distribution**' is derived from the fact that the values of $b(X;n,p)$ for $X = 0,1,2,\cdots,n$ are successive terms of the binomial expansion $[(1-p)+p]^n$; this shows that the sum of the probabilities equal 1, and it should.

Example 16.2. Each of the following situations represent **binomial experiments**. (Are you satisfied with the Bernoulli assumptions in each instance?)

(a) Suppose we flip a fair coin 10 times and let Y denote the number of tails in 10 flips. Here, $Y \sim b(n = 10;p = 0.5)$.

(b) In an agricultural experiment, forty percent of all plots respond to a certain treatment. I have four plots of land to be treated. If Y is the number of plots that respond to the treatment, then $Y \sim b(n = 4;p = 0.4)$.

(c) In rural Kenya, the prevalence rate for HIV is estimated to be around 8 percent. Let Y denote the number of HIV infected in a sample of 740 individuals. Here, $Y \sim b(n = 740;p = 0.08)$.

(d) It is known that screws produced by a certain company do not meet specifications (i.e., are defective) with probability 0.001. Let Y denote the number of defectives in a package of 40. Then, $Y \sim b(n = 40;p = 0.001)$.

(e) Toss a fair coin 100 times and let X be the number of heads. Then $X \sim B(100,0.5)$.

(f) A certain kind of lizard lays 8 eggs, each of which will hatch independently with probability 0.7. Let X denote the number of eggs which hatch. Then $X \sim B(8,0.7)$.

(g) What is the probability that at least 9 out of a group of 10 people who have been infected by a serious disease will survive, if the survival probability for the disease is 70 %?

Mean and Variance of Binomial Random Variables:

The probability function for a binomial random variable is

$$b(x;n,p)=\binom{n}{x}p^x(1-p)^{n-x}, \qquad x=0,1,2,\ldots$$

This is the probability of having x successes in a series of n independent trials when the probability of success in any one of the trials is p. If X is a random variable with this probability distribution,

$$\begin{aligned} E(X) &= \sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^{n-x} \\ &= \sum_{x=0}^{n} x \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \\ &= \sum_{x=1}^{n} \frac{n!}{(x-1)!(n-x)!} p^x (1-p)^{n-x} \end{aligned}$$

since the $x = 0$ term vanishes. Let $y = x - 1$ and $m = n - 1$. Subbing $x = y + 1$ and $n = m + 1$ into the last sum (and using the fact that the limits $x = 1$ and $x = n$ correspond to $y = 0$ and $y = m$, respectively)

$$\begin{aligned} E(X) &= \sum_{y=0}^{m} \frac{(m+1)!}{y!(m-y)!} p^{y+1} (1-p)^m \\ &= (m+1)p \sum_{y=0}^{m} \frac{m!}{y!(m-y)!} p^y (1-p)^m \\ &= np \sum_{y=0}^{m} \frac{m!}{y!(m-y)!} p^y (1-p)^m \end{aligned}$$

The binomial theorem says that Setting $a = p$ and $b = 1 - p$

$$(a+b)^m = \sum_{y=0}^{m} \frac{m!}{y!(m-y)!} a^y b^m$$

$$\sum_{y=0}^{m} \frac{m!}{y!(m-y)!} p^y (1-p)^m = \sum_{y=0}^{m} \frac{m!}{y!(m-y)!} a^y b^m = (a+b)^m = (p+1-p)^m = 1$$

so that

$E(X) = np$

Let us make use of the fact that $E(X^2) = E[X(X-1)]+E(X)$ and first evaluate $E[(X(X-1)]$. Similarly, but this time using $y = x - 2$ and $m = n - 2$.

$$
\begin{aligned}
E(X(X-1)) &= \sum_{x=0}^{n} x(x-1)\binom{n}{x}p^x(1-p)^{n-x} \\
&= \sum_{x=0}^{n} x(x-1)\frac{n!}{x!(n-x)!}p^x(1-p)^{n-x} \\
&= \sum_{x=2}^{n} \frac{n!}{(x-2)!(n-x)!}p^x(1-p)^{n-x} \\
&= n(n-1)p^2\sum_{x=2}^{n} \frac{(n-2)!}{(x-2)!(n-x)!}p^{x-2}(1-p)^{n-x} \\
&= n(n-1)p^2\sum_{y=0}^{m} \frac{m!}{y!(m-y)!}p^y(1-p)^{m-y} \\
&= n(n-1)p^2(p+(1-p))^m \\
&= n(n-1)p^2
\end{aligned}
$$

So the variance of X is

$$
\begin{aligned}
E(X^2) - (E(X))^2 &= E(X(X-1)) + E(X) - E(X)^2 \\
&= n(n-1)p^2 + np - (np)^2 = \\
& np(1-p).
\end{aligned}
$$

MGF FOR THE BINOMIAL DISTRIBUTION: Suppose that $Y \sim b(n;p)$. Then the mgf of Y is given by

$$P(X = 1) = p$$
$$P(X = 0) = 1 - p$$

$$
\begin{aligned}
M_X(t) = E(e^{tY}) &= \sum_{y=0}^{n} e^{ty}\binom{n}{y}p^y(1-p)^{n-y} \\
&= \sum_{y=0}^{n} \binom{n}{y} \left(pe^t\right)^y (1-p)^{n-y} = (q+pe^t)^n,
\end{aligned}
$$

where, $q = 1 - p$. The last step follows from noting that

$$\sum_{y=0}^{n} \binom{n}{y} \left(pe^t\right)^y (1-p)^{n-y},$$

is the binomial expansion of $(q + pe^t)^n$.

Question 16.2. Show that the mean and the variance of the binomial distribution are

$$\mu = n\,p \text{ and } \quad \sigma^2 = n\,p(1-p)$$

Hence, obtain the moment generating function of the binomial distribution.

Theorem 16.1.

$$b(x;n,\theta) = b(n - x,n,1 - \theta)$$

16.0.2 Binomial Distribution

The Binomial distribution is useful for problems in which we are concerned with determining the number of times an event is likely to occur or not occur during a given number of trials and consequently the probability of the event occurring or not occurring.

Binomial distribution is a theoretical probability distribution which was given by James Bernoulli. This distribution is applicable to situations with the following characteristics:

1. An experiment consists of a finite number n of repeated trials.
2. Each trial has only two possible, mutually exclusive, outcomes which are termed as a 'success' or a 'failure', or "good" or "bad", "Head" or "Tail" etc.
3. The probability of a success, denoted by p, is known and remains constant from trial to trial. The probability of a failure, denoted by q, is equal to $1 - p$, such that $p + q = 1$.
4. Different trials are independent, i.e., outcome of any trial or sequence of trials has no effect on the outcome of the subsequent trials.

   The sequence of trials under the above assumptions is also termed as *Bernoulli Trials*.

If the above conditions are satisfied then X is said to follow a Binomial distribution.

Definition 16.3. A random variable X has a Binomial distribution and it is referred to as a Binomial random variable if and only if its probability distribution is given by

$$f(x;n,p) = \binom{n}{x} p^x (1-p)^{n-x} \qquad for \qquad x = 0, 1, 2, \cdots, n$$

where
n - number of trials
p - probability of success
$1 - p$ - probability of failure
x - is the random variable (the number of successes in n trials).

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

Thus, the number of successes in n trials is a random variable having a Binomial distribution with parameters n and p. The name 'Binomial distribution' is derived from the fact that the values of $b(X;n,p)$ for $x = 0,1,2,\cdots,n$ are successive terms of the Binomial expansion $[(1 - p) + p]^n$; This shows that the sum of the probabilities equal 1, and it should.

This distribution is known as the binomial distribution with index n and probability p. We write this as $X \sim Bin(n,p)$.

Mean and Variance

If X is a random variable with a binomial $Bin(n,p)$ distribution then its mean and variance are

$$E(X) = np, \qquad \text{Var}(X) = np(1-p) = npq$$

For example, if $X \sim Bin(4,1/6)$ then

$$E(X) = np = 4 \times \frac{1}{6} = \frac{2}{3} = 0.667$$

and

$$\text{Var}(X) = np(1-p) = 4 \times \frac{1}{6} \times \frac{5}{6} = \frac{5}{9}$$

Also

$$SD(X) = \sqrt{\text{Var}(X)} = \sqrt{\frac{5}{9}} = 0.7454$$

Example 16.3. Find the probability of getting five heads and seven tails in 12 flips of a balanced coin.

Solution. Substituting $x = 5$, $n = 12$, and $\theta = 0.5$ into the formula for the binomial distribution.

$$f(5, 12, 0.5) = \binom{12}{5} (0.5)^5 (1-0.5)^{12-5} = 792\,(0.5)^{12} = 0.19$$

Example 16.4. Find the probability that seven of ten persons will recover from a tropical disease if we can assume independence and the probability is 0.80 that any one of them will recover from the disease.

Solution. Substituting $x = 7$, $n = 10$, and $\theta = 0.80$ into the formula for the binomial distribution.

$$f(7, 10, 0.80) = \binom{10}{7} (0.80)^7 (1-0.80)^{10-7} = 120\,(0.80)^7\,(0.20)^3 \approx 0.2$$

Example 16.5. At a supermarket 60% of the customers pay using the credit card. Find the probability that in a randomly selected set of 10 customers.

1. Exactly 2 pay by credit card.
2. None pays by credit card.
3. Less than 2 pay by credit card.
4. At most 3 pay by credit card.
5. More than seven pay by credit card.

Example 16.6. Five independent trials of an experiment are carried out, the probability of a successful outcome is p and failure is q. Write out the probability distribution distribution of X where x is the number of successful outcomes in five trials. Comment on your answer.

Example 16.7. The random variable X is defined binomially $B(7,0.2)$. Find to 3 d.p.

(a). $P(X = 3)$ (b). $P(1 < X \leq 4)$ (c). $P(X > 1)$

(d). $P(X \leq 3)$ (e). $P(X \geq 3)$

Example 16.8. A risky operation used for patients with no hope for survival has a survival rate of 80 %. Find the probability that exactly 4 of the next 5 patients operated on will survive. *Ans: 0.4096*

Example 16.9. A quiz, has 6 multiple choice questions, each with 3 alternatives. Find the probability of getting five or more correct. *Ans: 0.0178*

Example 16.10. A box contains a large number of pens. The probability that a pen is faulty is 0.1. How many pens would you need to select to be more than 95% certain of picking at least one faulty pen.

Example 16.11. Show that $\mathrm{E}(X) = np$ and $\mathrm{Var}(X) = npq$ of $X_n \sim B(n,p)$.

Example 16.12. The probability that it will be a fine day is 0.4.

1. Find the $\mathrm{E}(X)$ of fine days in a week.

2. Find the standard deviation in a week.

Example 16.13. A biased coin is tossed four times and the number of heads noted. The experiment was repeated 500 times in all. results are summarized as below;

Number of heads:	0	1	2	3	4
Frequency:	12	50	151	200	87

1. From the data, estimate the probability of obtaining a head when the coin is tossed.

2. Using binomial distribution in the same mean, calculate theoretical frequencies of $0,1,2,3,4$ heads.

Example 16.14. The probability of a student t being awarded a distinction in mathematics is 0.05. In a randomly selected group of 50 students. What is the most likely number of students awarded distinction.

16.0.3 The Poisson Distribution

The Poisson distribution is a very important discrete probability distribution which arises in many different contexts. We can think of a Poisson distribution as what becomes of a binomial distribution if we keep the mean fixed but let n become very large and p become very small, i.e. a large number of trials with a small probability of success in each. In general, it is used to model data which are counts of (random) events in a certain area or time interval, without a known fixed upper limit.

For example, consider the number of calls made in a 1 minute interval to an Internet service provider (ISP). The ISP has thousands of subscribers, but each one will call with a very small probability. If the ISP knows that on average 5 calls will be made in the interval, the actual number of calls will be a Poisson random variable, with mean 5.

If X is a random variable with a Poisson distribution with parameter λ(Greek lower case *lambda*) then the probability that $X = r$ is

$$f(X = r) = \frac{\lambda^r e^{-\lambda}}{r!} \quad r = 0,1,2,\cdots$$

We write $X \sim Po(\lambda)$. The parameter λ has a very simple interpretation as the rate at which events occur. The distribution has mean and variance

$$E(X) = \lambda, \qquad \text{Var}(X) = \lambda.$$

Conditions for a Poisson model.

(i) Events occur singly and at random in a given interval of time or space.

(ii) λ, the mean number of occurrences is known and is finite.

(iii). The variable X, is the number of occurrence in the given interval.

A random variable X has a Poisson distribution if the above conditions are satisfied.

The distribution is useful in describing the number of events that will occur in a specific period of time or a specific area or volume. For example the number of accidents per month at a busy intersection (junction) has a Poisson distribution.

Example 16.15. On average the school photocopier breaks down 8 times during the school week (MondayFriday). Assuming that the number of breakdowns can be modelled by Poisson distribution. find out the probability that it breaks down;

1. Five times in a week
2. Once on a Monday
3. 8 times in a fortnight

Solution. Let $\lambda = 8$

$$f(x, \lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$$

$$(a). \qquad P(X = 5) = \frac{e^{-8}8^5}{5!} =$$

$$(b). \qquad P(X = 1) = \frac{e^{-8}8}{1!} =$$

$$(c). \qquad \lambda = 8 \Rightarrow \lambda = 16 \quad \text{in two weeks}$$

$$P(X = 8) = \frac{e^{-16}16^8}{8!}$$

Example 16.16. The average number of trucks arriving on any one day at a truck depot in a certain city is known to be 2. What is the probability that on a given day fewer than nine trucks will arrive at the depot?

Solution. Let X be the number of trucks arriving on a given day. $\lambda = 12$

$$P(X < 9) = \sum_{x=0}^{8} P(x; 12) = 0.1550$$

Example 16.17. A maximum security prison reports that the number of escape attempts by prisoners per month has nearly a poisson distribution with mean equal to 1.5. Find

1. The probability of exactly 3 escape attempts during that month
2. The probability of at least one escape attempts during the next month.

Solution.

$$1. \qquad P(X=3) = f(3) = \frac{1.5^3 e^{-1.5}}{3!} = 0.1255$$

$$2. \qquad P(X \geq 1) = 1 - f(0) = 1 - \frac{1.5^0 e^{-1.5}}{0!} = 0.7769$$

1. $X \sim P(\lambda)$ with standard deviation 1.5. Find the $P(X \geq 3)$.
2. Show that if $X \sim P(\lambda)$, $M(t) = e^{\lambda(e_t - 1)}$.
3. A random variable X has a poisson distribution with mean equal to 2. Find
   (a) its pdf
   (b) its variance
   (c) $P(X \geq 1)$
   (d) $P(X = 3)$

16.0.4 Poisson Approximation to Binomial Distribution

When n is large (i.e. $n > 50$) and p is small (i.e. $p < 0.1$ then $X \sim B(n,p)$ can be approximated using a Poisson distribution. i.e. $\sim P(np)$.

Example 16.18. Eggs are packed in boxes of 500 on average 0.7 percent are found to be broken when unpacked. Find the probability that in a box of 500 eggs

1. Exactly 3 are broken
2. At least 2 are broken

Solution. Let X be the number of eggs broken in a box of 500 eggs.

$$X \sim B(500, 0.007)$$

$$\mathrm{E}(X) = np = 500(0.007) = \lambda$$

$$(a). \qquad P(X = 3) = \frac{e^{-\lambda}\lambda^x}{x!} =$$

$$\begin{aligned}(b). \qquad P(X \geq 2) &= P(X = 2) + P(X = 3) + \cdots \\ &= 1 - [P(X = 0) + P(X = 1)] \\ &= ???\end{aligned}$$

17 Random Variables, Expected Value and Variance

17.1 Random Variables

A **random variable** is a function that associates a unique numerical value with every outcome of an experiment. The value of the random variable will vary from trial to trial of an experiment.

Suppose S is the sample space associated with the random experiment E, then to every sample point in S we can assign a real number denoted by a variable X called a **random variable** on S. For example; when we toss a fair coin three times the sample space is

S = {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}

If we define a variable X as the number of heads observed when a fair coin is tossed three times, then X takes values 0, 1, 2, 3, where

$\{X = 0\} \Leftrightarrow \{TTT\}$, $\{X = 1\} \Leftrightarrow \{HTT, THT, TTH\}$;

$\{X = 2\} \Leftrightarrow \{HHT, HTH, THH\}$, $\{X = 3\} \Leftrightarrow \{HHH\}$.

Hence to each sample points in S we have assigned a real number, which uniquely determines the sample point. The variable X is called the random variable defined on the sample space S.

We can also find the probabilities of values 0, 1, 2, 3 of the random variable X as follows

$$\begin{aligned}&P(\{X = 0\}) = P(\{TTT\}) = 1/8, \\ &P(\{X = 1\}) = P(\{HTT, THT, TTH\}) = 3/8, \\ &P(\{X = 2\}) = P(\{HHT, HTH, THH\}) = 3/8, \\ &P(\{X = 3\}) = P(\{HHH\}) = 1/8.\end{aligned}$$

We can now express these probabilities in the form of a table;

Value of X	0	1	2	3	Total
Probability $P(X)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$	1

This is called the **probability distribution** of random variable X. A probability distribution for a discrete random variable is a formula, table or graph that provides the probability associated with each value of the random variable.

In general probability distribution of X satisfies the following conditions;

(i) all $P(X)$ are positive, i.e., $0 \leq P(X) \leq 1$,

(ii) $\sum_{\text{all } x} P(X) = 1$.

Note: Here, the uppercase X is used for the random variable and lowercase x is used to denote (represent) a realization of X. Probabilities can be easily obtained from the probability distribution table as follows: Probability of getting two or more heads

$$P(X > 1) = P(X = 2) = P(X = 3) = \frac{3}{8} + \frac{1}{8} = \frac{1}{2}$$

Probability of getting at least one head

$$P(X > 0) = 1 - P(X = 0) = 1 - \frac{1}{8} = \frac{7}{8}$$

A **random variable** X defined on the sample space S may be finite or infinite, at the same time it may take only countable values (without decimal) such variables are called **discrete random variables**. A discrete random variable is one which may take on only a countable number of distinct values such as $0, 1, 2, 3, 4, \cdots$. Examples of discrete random variables include the number of students in a class, the members teams in a football tournament, number of public holidays in a year, number of guests in attendance at a party, etc. On the other hand some variables like height, weigh, income do take any possible value on a given range and are called the **continuous random variables**.

17.2 Mathematical Expectation

Mathematical expectation refers to the mean or expected value of a random variable X whose distribution is known. The expected value, denoted by $\mathrm{E}(X)$, is a weighted average of realizations x of X where the weights are the corresponding probabilities. If $P(x)$ is the probability of various outcomes of X, then the mean of X or the expected value of X is given by

$$\begin{aligned}\mathrm{E}(X) &= x_1 \cdot P(x_1) + x_2 \cdot P(x_2) + x_3 \cdot P(x_3) + \cdots + x_n \cdot P(x_n) \\ &= \sum_{\text{all } x} XP(X)\end{aligned}$$

In the example above (Coin tossed 3 times)

$$\begin{aligned} \mathrm{E}(X) &= \sum_{\text{all } x} x \cdot p(x) \\ &= 0 \times \frac{1}{8} + 1 \times \frac{3}{8} + 2 \times \frac{3}{8} + 3 \times \frac{1}{8} \\ &= \frac{12}{8} \\ &= 1.5 \end{aligned}$$

Example 17.1. A discrete random variable X has the following probability distribution.

X	-2	-1	0	1	2
$P(X)$:	k	0.2	$2k$	$2k$	0.1

Find k and also find the expected value of the random variable X.

Solution. *Since X is a random variable with given $P(X)$, it must satisfy the conditions of a probability distribution.*

$$\sum P(X) = 1 \Rightarrow 5k + 0.3 = 1 \Rightarrow k = 0.7/5 = 0.14.$$

Now we can calculate the expected value by the formula $E(X) = \sum XP(X)$.

X	$P(x)$	$xP(x)$
-2	0.14	-0.28
-1	0.2	-0.2
0	0.28	0
1	0.28	0.28
2	0.1	0.2
Total	1	0

Example 17.2. A random variable follows the probability distribution given below;

X	0	1	2	3	4
$P(X)$	0.12	0.23	k	0.20	0.10

Obtain the value of k, and hence compute the expected value of X.
$k = 0.35$, $E(X) = 1.93$ *and* $Var(X) = 0.35$

Example 17.3. A coin is such that the tail is thrice as likely as the head. A game is played such that you earn 5 points for a head and lose 2 points for a tail after every toss. Let X be the total score after 4 consecutive tosses. Find the probability distribution of X and the expected number of points.

Solution. Let H be the event of observing a head and let X be the points earned, then $P(H) = 0.25$, and $P(T) + 0.75$, $n = 4$ and using the binomial formula, we have

$$P(y \text{ tails in 4 tosses} = {}^4C_y(0.75)^y(0.25)^{4-y})$$

$$\begin{aligned}
P(y = 0) &= {}^4C_0(0.75)^0(0.25)^4 = 0.0039 \\
P(y = 1) &= {}^4C_1(0.75)^1(0.25)^3 = 0.0469 \\
P(y = 2) &= {}^4C_2(0.75)^2(0.25)^2 = 0.2109 \\
P(y = 3) &= {}^4C_3(0.75)^3(0.25)^1 = 0.4219 \\
P(y = 4) &= {}^4C_4(0.75)^0(0.25)^4 = 0.3164
\end{aligned}$$

Outcome	No tail	1 tail	2 tail	3 tail	All tail
x	20	$15 - 2 = 13$	$10 - 4 = 6$	$5 - 6 = -1$	$0 - 8 = -8$
$P(X = x)$	0.0039	0.0469	0.2109	0.4219	0.3164

$$\mathrm{E}(X) = 20(0.0039) + 13(0.0469) + 6(0.2109) + -1(0.4219) + -8(0.3164) = -1$$

Example 17.4. A game is played as follows; throw a fair four sided dice and score eight times the number that faces down unless it is a four, you are given a second chance in which you score only four times that faces down. Let X be a random variable denoting the score for each player. Represent this information on a tree diagram showing the value of X and the corresponding probability, hence or otherwise find the mean of the scores;

x	4	8	16	12	24
$P(x)$	$\frac{1}{16}$	$\frac{5}{16}$	$\frac{5}{16}$	$\frac{1}{16}$	$\frac{1}{4}$

$$\begin{aligned}
\mathrm{E}(X) &= \sum_{\text{all } x} x \cdot p(x) \\
&= 4 \times \frac{1}{16} + 8 \times \frac{5}{16} + 16 \times \frac{5}{16} + 12 \times \frac{1}{16} + 24 \times \frac{1}{4} \\
&= \frac{232}{16} \\
&= 14.5
\end{aligned}$$

Question 17.1. A discrete random variable X takes the following values with the corresponding probabilities;

x	-3	-1	0	1	2	3
$P(x)$	0.1	0.2	0.1	0.2	0.15	0.25

Compute the following probabilities; (a). $P(X = -1)$ (b). $P(X = -2)$
(c). $P(X \leq 0)$ (d). $P(X$ is negative) (e). $\mathrm{E}(X)$

Question 17.2. A random variable X has the following probability distribution;

x	10	20	30	40
$P(X = x)$	a	$2a$	$4a$	$3a$

Find the value of a hence the expected value of X.

Question 17.3. The table below gives the probability distribution function of a random variable X.

x	0	1	2	3
$P(X = x)$	p	$2q$	$p+q$	q

Given that the mean of X is 1.375, find the values of p and q.

Question 17.4. A game is played as follows; throw a fair four sided die and score four times the number that faces down unless its a four. If its a four, you toss a fair coin whose sides are assigned 0 for tail and 2 for head and score 5 times the coins score. Let X be a random variable denoting the score for each player, representing this information on a tree diagram showing the value of X and the corresponding probability hence, find;

(a). the probability that you will score more than 8.

(b). the expected value of X.

(c). the probability that you will score more than the expected value.

17.3 Variance

The variance of a probability distribution of a discrete random variable provides a numerical measure of the spread and is given by the sum of the products of the squared deviations between the mean and all individual values of the random variable, taken one at a time and their respective probabilities. Thus variance is given by the formula:

$$\begin{aligned}\mathrm{Var}(X) &= E(X - E(X))^2 \\ &= E(X - E(X))^2 \\ &= \sum_{i=1}^{n}(x_i - E(X))^2 \cdot P(X_i) \\ &= \sum_{i=1}^{n}(x_i - \mu)^2 \cdot P(X_i) \qquad i = 1, 2, \cdots, n\end{aligned}$$

Variance is denoted by $\sigma^2 = \mathrm{Var}(X) = E(X - E(X))^2$ and the standard deviation is the square root of the variance.

From our example above of the three fair coins being tossed once, we can calculate the value of the variance, as follows, knowing that the mean of the distribution is 1.5.

No. of Heads (X)	$P(X)$	μ	$X-\mu$	$(X-\mu)^2$	$(X-\mu)^2P(X)$
0	1/8	1.5	-1.5	2.25	0.28
1	3/8	1.5	-0.5	0.25	0.09
2	3/8	1.5	0.5	0.25	0.09
3	1/8	1.5	1.5	2.25	0.28

$$\sum (X-\mu)^2 P(X) = 0.74$$

Hence variance which is denoted by σ^2 is given as

$$\begin{aligned}\sigma^2 &= \sum_{i=1}^{n} (X-\mu)^2 P(X) \\ &= 0.74\end{aligned}$$

and standard deviation

$$\sigma = \sqrt{\sigma^2} = \sqrt{0.74} = 0.86$$

Question 17.5. Determine the mean, variance, and standard deviation of the following discrete probability distribution.

x	0	1	2	3	4
$P(x)$	0.10	0.30	q	0.20	0.10

Question 17.6. A random variable X has the following probability distribution:

X:	-2	-1	0	1	2	3
$P(x)$:	0.1	k	0.2	$2k$	0.3	k

Find the value of k. Find the expected value and variance of X.

Question 17.7. A random variable X has the following probability distribution:

X:	0	1	2	3	4	5
$P(x)$:	0.1	0.1	0.2	k	0.2	0.1

Find the value of k. Find the expected value and variance of X.

Question 17.8. An unbiased coin is tossed four times. Find the expected value and variance of the random variable defined as number of Heads.

17.4 Discrete Probability Distribution

A discrete probability distribution, is the listing of all possible outcomes of an experiment together with their probabilities. This concept can be illustrated with the following example:

Example 17.5. A fair coin is tossed two times. The following is the list of all possible outcomes of this experiment and their respective probabilities.

Outcomes	Probability
TT	1 / 4
TH	1 / 4
HT	1 / 4
HH	1 / 4

Then the probability distribution of the number of heads obtained in these two tosses of the coin is given as follows:

Number of heads (X)	Probability $P(X = x)$
0	1/4
1	$1/4 + 1/4 = 1/2$
2	1/4
	1

Note: All probabilities must add up to 1.

A discrete probability distribution takes on discrete values that can be counted and it can only assume values only from a distinct predetermined set. The commonly used discrete probability distributions are: Binomial and Poisson distributions.

Conditions for a function to be a Probability Distribution Function

1. The probability that a random variable assumes a value X_i is always between 0 and 1. That is

$$0 \leq P(x_i) \leq 1$$

2. The sum of all probabilities $P(X_i)$ is equal to one.

$$\sum_{i=1}^{n} P(X_i) = 1$$

Example 17.6. The number of telephone calls received in an office between 9.00 A.M - 10.00 A.M has the probability distribution as shown in the table below:

The Probability distribution of the number of telephone calls.

No. of calls	**Probability** $P(X)$
0	0.05
1	0.20
2	0.25
3	0.20
4	0.10
5	0.15
6	0.05

(a). Verify that it is a probability function.

(b). Find the probability that there will be 3 or more calls.

(c). Find the probability that there will be even number of calls.

Solution. Clearly,

(a).

$$(i).\ 0 \leq P(x_i) \leq 1$$

$$(ii).\ \sum_{i=1}^{n} P(X_i) = 0.05 + 0.20 + 0.25 + 0.2 + .010 + 0.15 + 0.05 = 1$$

(b).

$$\begin{aligned} P(X \geq 3) &= P(X = 4) + P(X = 5) + P(X = 6) \\ &= 0.20 + 0.10 + 0.15 + 0.05 \\ &= 0.50 \end{aligned}$$

(c).

$$\begin{aligned} P(X = 0 \text{ or } 2 \text{ or } 4 \text{ or } 6) &= P(X = 0) + P(X = 2) + P(X = 4) + P(X = 6) \\ &= 0.05 + 0.25 + 0.10 + 0.05 \\ &= 0.40 \end{aligned}$$

17.5 The Mean or Expected Value of discrete probability distribution

The mean of the probability distribution is also known as the *Expected value*. Let X be a discrete random variable with the expected probability distribution $P(X)$. Then the expected value denoted as $E(X)$ is given by:

$$\mathrm{E}(X) = \sum_{i=1}^{n} x_i P(X_i) \quad i = 1,2,3\cdots,n.$$

Each value of the random variable is multiplied by the probability of occurrence of this value and then all these products are summed up.

It is also common in statistical literature to refer to the mean as Mathematical Expectation or the Expected value of the random variable X.

Example 17.7. Assume that we have three fair coins and we toss them simultaneously. The possible number of heads that can appear as s result of the random experiment are given in the following table:

Outcomes	No. of Heads	Probability
TTT	0	1 / 8
HTT	1	1 / 8
TTH	1	1 / 8
THT	1	1 / 8
THH	2	1 / 8
HHT	2	1 / 8
HTH	2	1 / 8
HHH	3	1 / 8

Thetablecanbesummarizedastothenumberofheadsoccurringintheentireexperimentandtheirrespective probabilities as follows:

Number of Heads (X)	$P(X)$	$X \cdot P(X)$
0	1/8	0
1	3/8	3/8
2	3/8	6/8
3	1/8	3/8
	1.0	12/8

The expected value (mean) for the number of heads in this experiment is

$$\begin{aligned} \mathsf{E}(X) &= \sum_{i=1}^{n} x_i P(X_i) \qquad i = 1, 2, \cdots, n \\ &= 0 + \frac{3}{8} + \frac{6}{8} + \frac{3}{8} \\ &= \frac{12}{8} = 1.5 \end{aligned}$$

This means that on an average, 1.5 heads can be expected to appear as a result of every random experiment of tossing three fair coins at any one time.

Example 17.8. In the telephone calls problem above find the mean of the telephone calls between 9 -10 am

X	$P(X)$	$X\,P(X)$
0	0.05	0
1	0.20	0.2
2	0.25	0.5
3	0.20	0.6
4	0.10	0.4
5	0.15	0.75
6	0.05	0.30
	1.00	2.75

$$\mu = \sum_{i=0}^{6} x_i P(X_i) = 2.75$$

Example 17.9. Suppose the hourly earnings X of a self employed landscaper gardener are given by the following probability function.

Hourly Earning X:	0	6	12	16
$P(X)$:	0.3	0.2	0.3	0.2

Find the gardener's Mean.

Solution. The Mean is given as:

$$\begin{aligned} \mu &= 0(0.3) + 6(0.2) + 12(0.3) + 16(0.2) \\ &= 0 + 1.2 + 3.6 + 3.2 \\ &= 8.0 \end{aligned}$$

17.5.1 Bernoulli Distribution

The Bernoulli distribution is an example of a discrete probability distribution. It is an appropriate tool in the analysis of proportions and rates.

A Bernoulli trial is a random experiment in which there are only two possible outcomes - *success* and *failure*.

- Tossing a coin and considering heads as success and tails as failure.
- Checking items from a production line: success = not defective, failure = defective.

- Phoning a call centre: success = operator free; failure = no operator free. A Bernoulli random

variable X takes the values 0 and 1 and

$$P(X = 1) = p$$
$$P(X = 0) = 1 - p$$

If an experiment has two possible outcomes, 'success' and 'failure' and their probabilities are respectively p and $1 - p$, then the number of successes, 0 or 1 has a Bernoulli distribution.

Definition 17.1. A random variable X has a Bernoulli distribution and it is referred to as a bernoulli random variable if and only if its probability distribution is given by

$$f(x;p) = p^x(1 - p)^{1-x} \quad for \qquad x = 0,1$$

In connection with the Bernoulli distribution, a success may be getting heads with a balanced coin, it may be catching pneumonia, it may be passing (or failing) an examination and it may be losing a race.

Note: Bernoulli distribution is a special case of the Binomial distribution. The mean and variance of the Bernoulli distribution are given as

$$\mathrm{E}(x) = p \qquad \text{and} \qquad \mathrm{Var}(x) = \sigma^2 = p(1 - p)$$

Question 17.9. Show that $\mathrm{E}(X) = p$ and $\mathrm{Var}(X) = p(1-p)$ for a random variable X which has a bernoulli distribution.

$$\begin{aligned}\mathrm{E}(X) = \sum_{allx} xP(x) &= \sum_{x=0}^{1} p^x(1-p)^{1-x} \\ &= 0 \cdot p^0(1-p)^1 + 1 \cdot p(1-p)^{1-1} \\ &= 0 + p(1-p)^0 = p\end{aligned}$$

$$\begin{aligned}\mathrm{Var}(X) = \sum_{allx}(x-\mu)P(x) &= E(X^2) - E(X)^2 \\ &= 0^2(1-p) + 1^2(p-p^2) \\ &= p - p^2 = p(1-p).\end{aligned}$$

The moment generating function of a Bernoulli distribution is given by

$$\begin{aligned}M_X(t) = \mathrm{E}(e^{tx}) &= \sum_{x=0}^{1} \cdot p^x(1-x)^{1-x} \\ &= e^0 \cdot p^0(1-p)^1 + e^t \cdot p(1-p)^0 \\ &= (1-p) + e^t p \\ &= 1 - p + e^t p \\ &= P(e^t - 1) + 1\end{aligned}$$

Example 17.10. A carton contain 4 good eggs and 6 bad eggs. If an egg is selected at random, then the random variable

$$X = \begin{cases} 0 & \text{if the egg is bad} \\ 1 & \text{if the egg is good} \end{cases}$$

$$P(\text{good egg}) = \frac{4}{10} = \frac{2}{5}$$
$$P(X = 1) = (2/5)^1 \cdot (1 - 2/5)^0 = 2/5$$
$$E(X) = 2/5, \quad Var(X) = p(1-p) = 2/5(1 - 2/5) = 6/25.$$

BERNOULLI TRIALS: Many experiments consist of a sequence of trials, where

(i) each trial results in a "success" or a "failure,"

(ii) the trials are **independent**, and

(iii) the probability of "success," denoted by p, $0 < p < 1$, is the **same** on every trial.

If the above conditions are satisfied then X is said to follow a binomial distribution. If a Bernoulli experiment is performed repeatedly, we obtain a sequence of **bernoulli trials**. In a finite sequence of bernoulli trials, we are usually interested in the number of 'success'. Since there are only two possible outcomes at each trial a sample of "n" bernoulli trials contains 2^n possible outcomes. As before $X = 0, 1$ for each trial. The probability of any particular sequence of outcomes in 'n' trials is obtained as the product of the probabilities of the n outcomes resulting at each trial.

Definition 17.2. A random variable X has a binomial distribution and it is referred to as a binomial random variable if and only if its probability distribution is given by

$$f(x; n, \theta) = \binom{n}{x} p^x (1-p)^{n-x} \qquad \text{for} \qquad x = 0,1,2,\cdots,n$$

where
n - number of trials
p - probability of success
$1 - p$ - probability of failure
x - is the random variable (the number of successes in n trials).
$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

Thus, the number of successes in n trials is a random variable having a binomial distribution with parameters n and p. The name '**binomial distribution**' is derived from the fact that the values of $b(X;n,p)$ for $X = 0,1,2,\cdots,n$ are successive terms of the binomial expansion $[(1 - p) + p]^n$; this shows that the sum of the probabilities equal 1, and it should.

Example 17.11. Each of the following situations represent **binomial experiments**. (Are you satisfied with the Bernoulli assumptions in each instance?)

(a) Suppose we flip a fair coin 10 times and let Y denote the number of tails in 10 flips. Here, $Y \sim b(n = 10; p = 0.5)$.

(b) In an agricultural experiment, forty percent of all plots respond to a certain treatment. I have four plots of land to be treated. If Y is the number of plots that respond to the treatment, then $Y \sim b(n = 4; p = 0.4)$.

(c) In rural Kenya, the prevalence rate for HIV is estimated to be around 8 percent. Let Y denote the number of HIV infected in a sample of 740 individuals. Here, $Y \sim b(n = 740; p = 0.08)$.

(d) It is known that screws produced by a certain company do not meet specifications (i.e., are defective) with probability 0.001. Let Y denote the number of defectives in a package of 40. Then, $Y \sim b(n = 40; p = 0.001)$.

(e) Toss a fair coin 100 times and let X be the number of heads. Then $X \sim B(100, 0.5)$.

(f) A certain kind of lizard lays 8 eggs, each of which will hatch independently with probability 0.7. Let X denote the number of eggs which hatch. Then $X \sim B(8, 0.7)$.

(g) What is the probability that at least 9 out of a group of 10 people who have been infected by a serious disease will survive, if the survival probability for the disease is 70 %?

Mean and Variance of Binomial Random Variables:

The probability function for a binomial random variable is

$$b(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, 2, \ldots$$

This is the probability of having x successes in a series of n independent trials when the probability of success in any one of the trials is p. If X is a random variable with this probability distribution,

$$\begin{aligned} E(X) &= \sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^{n-x} \\ &= \sum_{x=0}^{n} x \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \\ &= \sum_{x=1}^{n} \frac{n!}{(x-1)!(n-x)!} p^x (1-p)^{n-x} \end{aligned}$$

since the $x = 0$ term vanishes. Let $y = x - 1$ and $m = n - 1$. Subbing $x = y + 1$ and $n = m + 1$ into the last sum (and using the fact that the limits $x = 1$ and $x = n$ correspond to $y = 0$ and $y = m$, respectively)

$$\begin{aligned} E(X) &= \sum_{y=0}^{m} \frac{(m+1)!}{y!(m-y)!} p^{y+1} (1-p)^m \\ &= (m+1)p \sum_{y=0}^{m} \frac{m!}{y!(m-y)!} p^y (1-p)^m \\ &= np \sum_{y=0}^{m} \frac{m!}{y!(m-y)!} p^y (1-p)^m \end{aligned}$$

The binomial theorem says that

$$(a+b)^m = \sum_{y=0}^{m} \frac{m!}{y!(m-y)!} a^y b^m$$

Setting $a = p$ and $b = 1 - p$

$$\sum_{y=0}^{m} \frac{m!}{y!(m-y)!} p^y (1-p)^m = \sum_{y=0}^{m} \frac{m!}{y!(m-y)!} a^y b^m = (a+b)^m = (p+1-p)^m = 1$$

so that

$$E(X) = np$$

Let us make use of the fact that $E(X^2) = E[X(X-1)] + E(X)$ and first evaluate $E[(X(X-1)]$. Similarly, but this time using $y = x - 2$ and $m = n - 2$.

$$\begin{aligned}
E(X(X-1)) &= \sum_{x=0}^{n} x(x-1) \binom{n}{x} p^x (1-p)^{n-x} \\
&= \sum_{x=0}^{n} x(x-1) \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \\
&= \sum_{x=2}^{n} \frac{n!}{(x-2)!(n-x)!} p^x (1-p)^{n-x} \\
&= n(n-1)p^2 \sum_{x=2}^{n} \frac{(n-2)!}{(x-2)!(n-x)!} p^{x-2} (1-p)^{n-x} \\
&= n(n-1)p^2 \sum_{y=0}^{m} \frac{m!}{y!(m-y)!} p^y (1-p)^{m-y} \\
&= n(n-1)p^2 (p + (1-p))^m \\
&= n(n-1)p^2
\end{aligned}$$

So the variance of X is

$$\begin{aligned}
E(X^2) - (E(X))^2 &= E(X(X-1)) + E(X) - E(X)^2 \\
&= n(n-1)p^2 + np - (np)^2 \\
&= np(1-p).
\end{aligned}$$

MGF FOR THE BINOMIAL DISTRIBUTION: Suppose that $Y \sim b(n;p)$. Then the mgf of Y is given by

$$P(X = 1) = p$$
$$P(X = 0) = 1 - p$$

$$M_X(t) = E(e^{tY}) = \sum_{y=0}^{n} e^{ty} \binom{n}{y} p^y (1-p)^{n-y}$$
$$= \sum_{y=0}^{n} \binom{n}{y} \left(pe^t\right)^y (1-p)^{n-y} = (q + pe^t)^n,$$

where, $q = 1 - p$. The last step follows from noting that

$$\sum_{y=0}^{n} \binom{n}{y} \left(pe^t\right)^y (1-p)^{n-y},$$

is the binomial expansion of $(q + pe^t)^n$.

Question 17.10. Show that the mean and the variance of the binomial distribution are

$$\mu = n\,p \qquad \text{and} \qquad \sigma^2 = n\,p(1-p)$$

Hence, obtain the moment generating function of the binomial distribution.

Theorem 17.1.

$$b(x; n, \theta) = b(n - x, n, 1 - \theta)$$

17.5.2 Binomial Distribution

The Binomial distribution is useful for problems in which we are concerned with determining the number of times an event is likely to occur or not occur during a given number of trials and consequently the probability of the event occurring or not occurring.

Binomial distribution is a theoretical probability distribution which was given by James Bernoulli. This distribution is applicable to situations with the following characteristics:

1. An experiment consists of a finite number n of repeated trials.

2. Each trial has only two possible, mutually exclusive, outcomes which are termed as a 'success' or a 'failure', or "good" or "bad", "Head" or "Tail" etc.

3. The probability of a success, denoted by p, is known and remains constant from trial to trial. The probability of a failure, denoted by q, is equal to $1 - p$, such that $p + q = 1$.

4. Different trials are independent, i.e., outcome of any trial or sequence of trials has no effect on the outcome of the subsequent trials.

   The sequence of trials under the above assumptions is also termed as *Bernoulli Trials*.

If the above conditions are satisfied then X is said to follow a Binomial distribution.

Definition 17.3. A random variable X has a Binomial distribution and it is referred to as a Binomial random variable if and only if its probability distribution is given by

$$f(x;n,p) = \binom{n}{x} p^x (1-p)^{n-x} \qquad for \qquad x = 0, 1, 2, \cdots, n$$

where
n - number of trials
p - probability of success
$1 - p$ - probability of failure
x - is the random variable (the number of successes in n trials).

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

Thus, the number of successes in n trials is a random variable having a Binomial distribution with parameters n and p. The name 'Binomial distribution' is derived from the fact that the values of $b(X;n,p)$ for $x = 0,1,2,\cdots,n$ are successive terms of the Binomial expansion $[(1 - p) + p]^n$; This shows that the sum of the probabilities equal 1, and it should.

This distribution is known as the binomial distribution with index n and probability p. We write this as $X \sim Bin(n,p)$.

Mean and Variance

If X is a random variable with a binomial $Bin(n,p)$ distribution then its mean and variance are

$$E(X) = np, \qquad \text{Var}(X) = np(1-p) = npq$$

For example, if $X \sim Bin(4, 1/6)$ then

$$E(X) = np = 4 \times \frac{1}{6} = \frac{2}{3} = 0.667$$

and

$$\text{Var}(X) = np(1-p) = 4 \times \frac{1}{6} \times \frac{5}{6} = \frac{5}{9}$$

Also

$$SD(X) = \sqrt{\text{Var}(X)} = \sqrt{\frac{5}{9}} = 0.7454$$

Example 17.12. Find the probability of getting five heads and seven tails in 12 flips of a balanced coin.

Solution. Substituting $x = 5$, $n = 12$, and $\theta = 0.5$ into the formula for the binomial distribution.

$$f(5, 12, 0.5) = \binom{12}{5} (0.5)^5 (1-0.5)^{12-5} = 792\,(0.5)^{12} = 0.19$$

Example 17.13. Find the probability that seven of ten persons will recover from a tropical disease if we can assume independence and the probability is 0.80 that any one of them will recover from the disease.

Solution. Substituting $x = 7$, $n = 10$, and $\theta = 0.80$ into the formula for the binomial distribution.

$$f(7, 10, 0.80) = \binom{10}{7} (0.80)^7 (1-0.80)^{10-7} = 120\,(0.80)^7\,(0.20)^3 \approx 0.2$$

Example 17.14. At a supermarket 60% of the customers pay using the credit card. Find the probability that in a randomly selected set of 10 customers.

1. Exactly 2 pay by credit card.
2. None pays by credit card.
3. Less than 2 pay by credit card.
4. At most 3 pay by credit card.
5. More than seven pay by credit card.

Example 17.15. Five independent trials of an experiment are carried out, the probability of a successful outcome is p and failure is q. Write out the probability distribution distribution of X where x is the number of successful outcomes in five trials. Comment on your answer.

Example 17.16. The random variable X is defined binomially $B(7,0.2)$. Find to 3 d.p.

(a). $P(X = 3)$ (b). $P(1 < X \leq 4)$ (c). $P(X > 1)$
(d). $P(X \leq 3)$ (e). $P(X \geq 3)$

Example 17.17. A risky operation used for patients with no hope for survival has a survival rate of 80 %. Find the probability that exactly 4 of the next 5 patients operated on will survive. *Ans: 0.4096*

Example 17.18. A quiz, has 6 multiple choice questions, each with 3 alternatives. Find the probability of getting five or more correct. *Ans: 0.0178*

Example 17.19. A box contains a large number of pens. The probability that a pen is faulty is 0.1. How many pens would you need to select to be more than 95% certain of picking at least one faulty pen.

Example 17.20. Show that E(X) = np and Var(X) = npq of $X_n \sim B(n,p)$.

Example 17.21. The probability that it will be a fine day is 0.4.

1. Find the E(X) of fine days in a week.

2. Find the standard deviation in a week.

Example 17.22. A biased coin is tossed four times and the number of heads noted. The experiment was repeated 500 times in all. results are summarized as below;

Number of heads:	0	1	2	3	4
Frequency:	12	50	151	200	87

1. From the data, estimate the probability of obtaining a head when the coin is tossed.

2. Using binomial distribution in the same mean, calculate theoretical frequencies of $0,1,2,3,4$ heads.

Example 17.23. The probability of a student t being awarded a distinction in mathematics is 0.05. In a randomly selected group of 50 students. What is the most likely number of students awarded distinction.

17.5.3 The Poisson Distribution

The Poisson distribution is a very important discrete probability distribution which arises in many different contexts. We can think of a Poisson distribution as what becomes of a binomial distribution if we keep the mean fixed but let n become very large and p become very small, i.e. a large number of trials with a small probability of success in each. In general, it is used to model data which are counts of (random) events in a certain area or time interval, without a known fixed upper limit.

For example, consider the number of calls made in a 1 minute interval to an Internet service provider (ISP). The ISP has thousands of subscribers, but each one will call with a very small probability. If the ISP knows that on average 5 calls will be made in the interval, the actual number of calls will be a Poisson random variable, with mean 5.

If X is a random variable with a Poisson distribution with parameter λ(Greek lower case *lambda*) then the probability that X = r is

$$f(X = r) = \frac{\lambda^r e^{-\lambda}}{r!} \quad r = 0,1,2,\cdots$$

We write $X \sim Po(\lambda)$. The parameter λ has a very simple interpretation as the rate at which events occur. The distribution has mean and variance

$$E(X) = \lambda, \qquad \text{Var}(X) = \lambda.$$

Conditions for a Poisson model.

(i) Events occur singly and at random in a given interval of time or space.

(ii) λ, the mean number of occurrences is known and is finite.

(iii). The variable X, is the number of occurrence in the given interval.

A random variable X has a Poisson distribution if the above conditions are satisfied.

The distribution is useful in describing the number of events that will occur in a specific period of time or a specific area or volume. For example the number of accidents per month at a busy intersection (junction) has a Poisson distribution.

Example 17.24. On average the school photocopier breaks down 8 times during the school week (MondayFriday). Assuming that the number of breakdowns can be modelled by Poisson distribution. find out the probability that it breaks down;

1. Five times in a week
2. Once on a Monday
3. 8 times in a fortnight

Solution. Let $\lambda = 8$

$$f(x,\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$$

$$(a). \quad P(X=5) = \frac{e^{-8}8^5}{5!} =$$

$$(b). \quad P(X=1) = \frac{e^{-8}8}{1!} =$$

$$(c). \quad \lambda = 8 \Rightarrow \lambda = 16 \quad \text{in two weeks}$$

$$P(X=8) = \frac{e^{-16}16^8}{8!}$$

Example 17.25. The average number of trucks arriving on any one day at a truck depot in a certain city is known to be 2. What is the probability that on a given day fewer than nine trucks will arrive at the depot?

Solution. Let X be the number of trucks arriving on a given day. $\lambda = 12$

$$P(X<9) = \sum_{x=0}^{8} P(x;12) = 0.1550$$

Example 17.26. A maximum security prison reports that the number of escape attempts by prisoners per month has nearly a poisson distribution with mean equal to 1.5. Find

1. The probability of exactly 3 escape attempts during that month

2. The probability of at least one escape attempts during the next month.

Solution.

$$1. \quad P(X=3) = f(3) = \frac{1.5^3 e^{-1.5}}{3!} = 0.1255$$

$$2. \quad P(X \geq 1) = 1 - f(0) = 1 - \frac{1.5^0 e^{-1.5}}{0!} = 0.7769$$

1. $X \sim P(\lambda)$ with standard deviation 1.5. Find the $P(X \geq 3)$.

2. Show that if $X \sim P(\lambda)$, $M(t) = e^{\lambda(e_t - 1)}$.

3. A random variable X has a poisson distribution with mean equal to 2. Find

   (a) its pdf

   (b) its variance

   (c) $P(X \geq 1)$

   (d) $P(X = 3)$

17.5.4 Poisson Approximation to Binomial Distribution

When n is large (i.e. $n > 50$) and p is small (i.e. $p < 0.1$ then $X \sim B(n,p)$ can be approximated using a Poisson distribution. i.e. $\sim P(np)$.

Example 17.27. Eggs are packed in boxes of 500 on average 0.7 percent are found to be broken when unpacked. Find the probability that in a box of 500 eggs

1. Exactly 3 are broken

2. At least 2 are broken

Solution. Let X be the number of eggs broken in a box of 500 eggs.

$$X \sim B(500, 0.007)$$

$$\mathrm{E}(X) = np = 500(0.007) = \lambda$$

$$(a). \quad P(X=3) = \frac{e^{-\lambda}\lambda^x}{x!} =$$

$$(b). \quad P(X \geq 2) = P(X=2) + P(X=3) + \cdots$$
$$= 1 - [P(X=0) + P(X=1)]$$
$$= ???$$