Unit: 1-9

Course Code: 9309

Credit Hours: 3

AIOU

Introduction to Statistics for Economists

B.S. Economics (4 Years)

ALLAMA IQBAL OPEN UNIVERSITY
www.aiou.edu.pk

Study Guide

INTRODUCTION TO STATISTICS FOR ECONOMISTS

BS Economics (4 Year)

Code No. 9309 / ECO4004 Units 1 – 9

Credit Hours: 3

DEPARTMENT OF ECONOMICS
FACULTY OF SOCIAL SCIENCES AND HUMANITIES
ALLAMA IQBAL OPEN UNIVERSITY

1st Edition…………………………………….2023

Quantity………………………………………1000

Layout :………………………………………..Naeem Akhtar

Printing Coordinator:…………………………Dr. Saramd Iqbal

Printer…………………………………………AIOU, Islamabad

Publisher………………………………………. AIOU, Islamabad

Course Team

1.	Incharge	Dr. Fouzia Jamshaid
2.	Course Development Coordinator	Dr Muhammad Ilyas
3.	Writer	Dr. Zahid Iqbal
4.	Reviewer	Dr Muhammad Ilyas
5.	Editor	Mr. Fazal Karim

CONTENTS

1 – Introduction to the Course

If you invest in financial markets, you may want to predict the price of a stock in six months from now on the basis of company performance measures and other economic factors. As a university student, you may be interested in knowing the dependence of the mean starting salary of a college graduate, based on your GPA. These are just some examples that highlight how statistics are used in our modern society. To figure out the desired information for each example, you need data to analyze or knowledge of Statistics.

The purpose of this course is to introduce you to the subject of statistics as a science of data. There is data abound in this information age; how to extract useful knowledge and gain a sound understanding of complex data sets has been more of a challenge. In this course, we will focus on the fundamentals of statistics, which may be broadly described as the techniques to collect, clarify, summarize, organize, analyze, and interpret numerical information.

This course will begin with a brief overview of the discipline of statistics and will then quickly focus on descriptive statistics, introducing graphical methods of describing data. You will learn about combinatorial probability and random distributions, the latter of which serves as the foundation for statistical inference. On the side of inference, we will focus on both estimation and hypothesis testing issues. We will also examine the techniques to study the relationship between two or more variables; this is known as regression.

By the end of this course, you should gain a sound understanding of what statistics represent, how to use statistics to organize and display data, and how to draw valid inferences based on data by using appropriate statistical tools.

There are nine units in total. First five unit are devoted to Introduction to Statistics, presentation, central tendency and variability. In today's technologically advanced world, we have access to large volumes of data. The first step of data analysis is to accurately summarize all of this data, both graphically and numerically, so that we can understand what the data reveals. To be able to use and interpret the data correctly is essential to making informed decisions. For instance, when you see a survey of opinion about a certain TV program, you may be interested in the proportion of those people who indeed like the program. In these units, you will learn about descriptive statistics, which are used to summarize and display data. After completing each unit, you will know how to present your findings once you have collected data. For example, suppose you want to buy a new mobile phone

with a particular type of a camera. Suppose you are not sure about the prices of any of the phones with this feature, so you access a website that provides you with a sample data set of prices, given your desired features. Looking at all of the prices in a sample can sometimes be confusing. A better way to compare this data might be to look at the mean, median price and the variation of prices. The mean, median and variation are two ways out of several ways that you can describe data. You can also graph the data so that it is easier to see what the price distribution looks like.
Probabilities affect our everyday lives. In this unit, you will learn about probability and its properties, how probability behaves, and how to calculate and use it. You will study the fundamentals of probability and will work through examples that cover different types of probability questions. These basic probability concepts will provide a foundation for understanding more statistical concepts, for example, interpreting polling results. Though you may have already encountered concepts of probability, after this unit, you will be able to formally and precisely predict the likelihood of an event occurring given certain constraints.

Probability theory is a discipline that was created to deal with chance phenomena. For instance, before getting a surgery, a patient wants to know the chances that the surgery might fail; before taking medication, you want to know the chances that there will be side effects; before leaving your house, you want to know the chance that it will rain today. Probability is a measure of likelihood that takes on values between 0 and 1, inclusive, with 0 representing impossible events and 1 representing certainty. The chances of events occurring fall between these two values.

The skill of calculating probability allows us to make better decisions. Whether you are evaluating how likely it is to get more than 50% of the questions correct on a quiz if you guess randomly; predicting the chance that the next storm will arrive by the end of the week; or exploring the relationship between the number of hours students spend at the gym and their performance on an exam, an understanding of the fundamentals of probability is crucial.

We will also talk about random variables. A random variable describes the outcomes of a random experiment. A statistical distribution describes the numbers of times each possible outcome occurs in a sample. The values of a random variable can vary with each repetition of an experiment. Intuitively, a random variable, summarizing certain chance phenomenon, takes on values with certain probabilities. A random variable can be classified as being either discrete or continuous, depending on the values it assumes. Suppose you count the number of people who go to a coffee shop between 4 p.m. and 5 p.m. and the amount of waiting time that they spend in that hour. In this case, the number of people is an example of a discrete random variable and the amount of waiting time they spend is an example of a continuous random variable.

In unit 8, we will discuss situations in which the mean of a population, treated as a variable, depends on the value of another variable. One of the main reasons why we conduct such analyses is to understand how two variables are related to each other. The most common type of relationship is a linear relationship. For example, you may want to know what happens to one variable when you increase or decrease the other variable. You want to answer questions such as, "Does one variable increase as the other increases, or does the variable decrease?" For example, you may want to determine how the mean reaction time of rats depends on the amount of drug in bloodstream.

In unit 8 and 9, you will also learn to measure the degree of a relationship between two or more variables. Both correlation and regression are measures for comparing variables. Correlation quantifies the strength of a relationship between two variables and is a measure of existing data. On the other hand, regression is the study of the strength of a linear relationship between an independent and dependent variable and can be used to predict the value of the dependent variable when the value of the independent variable is known.

The Study Guide in your hand provides you the introduction of each Unit followed by the objectives of the Unit. In each Unit throughout the Study Guide, we have given self-assessment questions. They are meant to assist your comprehension after reading the Unit the useful reading list is also provided for each Unit.

This is basic Statistics of 3 credit hours course on Statistics for Economist-I, specially designed for BS Economics students learning through distance education system of the Allama Iqbal Open University. We hope that you will find this course useful and interesting one. Suggestions for the improvement of course as well as the Study Guide will be highly appreciated.

2 – Course Learning Outcomes

The desired result of all introductory statistics or basic statistics courses is to produce statistically educated students, which means that students should develop the ability to think statistically.

The following goals reflect major strands in the collective thinking expressed in the statistics education literature. They summarize what a student should know and understand at the conclusion of a first course in statistics. Achieving this knowledge will require learning some statistical techniques, but mastering specific techniques is not as important as understanding the statistical concepts and principles that underlie such techniques.

The main objectives of the course are to enable you:

1. Students should become critical consumers of statistically-based results reported in popular media, recognizing whether reported results reasonably follow from the study and analysis conducted.
2. Students should be able to recognize questions for which the investigative process in statistics would be useful and should be able to answer questions using the investigative process.
3. Students should be able to produce graphical displays and numerical summaries and interpret what graphs do and do not reveal.
4. Students should recognize and be able to explain the central role of variability in the field of statistics.
5. To have introduction of statistics as a field of knowledge and its scope and relevance to other disciplines of natural and social sciences.
6. To equipped and prepare students for advance courses in the field of statistics.
7. To achieve the capability of critical thinking about data and its sources; have idea about variables and their types and scale measures.
8. Be able to calculate and interpret descriptive statistics (able to classify, tabulate, describe and display data using software).

3- Structure of the Study Guide

The course "Introduction to the Statistics for Economists" a three credit hours course consists of nine units. A unit is a study of 12–16 hours of course work for two weeks. The course work of one unit will include study of compulsory reading materials and suggested books. You should make a timetable for studies to complete the work within the allocated time.

This study guide/course has been organized to enable you to acquire the skill of self-learning. For each unit an introduction is given, to help you to develop an objective analysis of the major and sub-themes, discussed in the prescribed reading materials. Besides this, learning outcomes of each unit are very specifically laid down to facilitate in developing logical analytical approach. Summary of main topics has also been included in the contents to understand the topics. We have given you a few self-assessments questions and activities which are not only meant to facilitate you in understanding the required reading materials, but also to provide you an opportunity to assess yourself. Recommended books and important links have been given to understand the main topics. Key terms have also been included in the study guide.

Every course has a study package including study guides, assignments and tutorial schedule uploaded by the University. For the books suggested at the end of each unit you can visit online resources, a nearby library/study center or the Central Library at main campus in AIOU.

Course Materials

The primary learning materials for this course are:

- Readings (e.g., study guides, recommended books, online links and scholarly articles)
- Lectures, (tutorial and workshops)
- Other resources.

All course materials are free to access and can be found through the links provided in each unit and sub-unit of the course. Pay close attention to the notes that accompany these course materials, as they will instruct you as to what specifically to read or watch at a given point in the course and help you to understand how these individual materials fit into the course. You can also access a list all the materials used in this course by clicking on resources mentioned in each unit.

Technical Requirements

This course is delivered online through Learning Management System (LMS). You will be required to have access to a computer or web-capable mobile device and have consistent access to the internet either to view or download the necessary course resources and to attempt any auto-graded course assessments and the final exam.

Methods of Instruction

Following are the methods for directing this guide and course also and then you will be able to understand the macroeconomics course through.

- Lecture online
- Mandatory workshops
- Workshop Quizzes
- Class discussion during workshops
- Individual, paired and small group exercises
- Use of library for research projects
- Use of videos lectures
- Use of the internet

Types of Assignments

- Students must complete assignments from the recommended books and other sources also.
- Students must be able to research and complete the assignments, which will include library, Internet and another media research.

Activities

In most units, different types of activities are mentioned for better understanding of the course. If you thoroughly study the materials and follow the links and videos, then you will be able to understand the course in the easiest way.

4- How to Use the Study Guide

Before attending a tutorial meeting, it is imperative to prepare yourself in the following manner to get maximum benefit of it. You are required to follow the following steps:

Step 1

Go through them.

1. Course Outlines
2. Course Introduction
3. Course Learning Outcomes
4. Structure of the Course
5. Assessment Methods
6. Recommended Books
7. Suggested Readings

Step 2

Read the whole unit and make notes of those points which you could not fully understand or wish to discuss with your course tutor.

Step 3

Go through the self-assessment questions at the end of each unit. If you find any difficulty in comprehension or locating relevant material, discuss it with your tutor.

Step 4

Study the compulsory recommended books at least for three hours in a week recommended in your study guide. AIOU Tries to read it with the help of a specific study guide for the course. You can raise questions on both during your tutorial meetings and workshops.

Step 5

First go through assignments, which are mandatory to solve/complete for this course. Highlight all the points you consider difficult to tackle, and then discuss in detail with your tutor. This exercise will keep you regular and ensure good results in the form of higher grades.

Assessment

For each three credit hours course, a student will be assessed as follow:

- Two Assignments (continuous assessment during semester).
- Final Examination (three-hours written examination will take place at the end of each semester)
- Mandatory participation in the workshop (as per AIOU policy)
- Workshop Quizzes
- Group discussion
- Presentation

Assignments

- Assignments are written exercises that are required to complete at home or place of work after having studied 9 units/study guides with the help of compulsory and suggested reading material within the scheduled study period. (See the assignments scheduled).
- For this course 02 assignments are uploaded on the AIOU portal along with allied material. You are advised to complete your assignments within the required time and upload it to your assigned tutor.
- This is compulsory course work, and its successful completion will make you eligible to take the final examination at the end of the semester.
- You will upload your assignments to your appointed tutor, whose name is notified to you for assessment and necessary guidance through concerned Regional Office of AIOU. You can also locate your tutor through AIOU website. Your tutor will return your online assignments after marking and providing necessary academic guidance and supervision.

Workshops

- The online mandatory workshops through (LMS) of Bachelor Studies BS Economics (4, Year) courses will be arranged during each semester or as-per AIOU policy. Attendance and course quizzes are compulsory in workshops. A student will not be declared pass until he/she attends the workshop satisfactorily and actively.
- The duration of a workshop for each 03-credit course will be as per AIOU policy.

Revision before the Final Examination

It is very important that you revise the course as systematically as you have been studying.

You may find the following suggestions helpful.

- Go through the course unit one by one, using your notes during tutorial meetings to remind you of the key concepts or theories. If you have not already made notes, do so now.
- Prepare a chronology with short notes on the topics/events/personalities included in all units.
- Go through your assignments and check your weak areas in each case.

- Test yourself on each of the main topics, write down the main points or go through all the notes.
- Make sure to attend the workshops and revise all the points that you find difficult to comprehend.
- Try to prepare various questions with your fellow-students during last few tutorial meetings. A group activity in this regard is helpful. Each student should be given a topic and revise his topics intensively, summarize it and revise in group, then all members raise queries and questions. This approach will make your studies interesting and provide you an opportunity to revise thoroughly.
- For the final exam paper, go through last semesters' papers. This can clarify questions and deciding how to frame an answer.
- Before your final exams, make sure that,
  - ➢ you get your roll-number slip
  - ➢ you know the exact location of the examination center
  - ➢ you know the date and time of the examination.

Note:

This study guide has been developed to guide the students about the course "Introduction to Statistics for Economists". In this context we want to make it clear that you are not bound to depend entirely upon the recommended books in the study guide. In case you are unable to find any recommended book, please free to consult any other book which covers the main contents of the course.

Moreover, you can get information regarding your Assignments, Workshop Schedule, Assignment Results, Tutors, and Final Examination from the AIOU website: www.aiou.edu.pk and through your LMS account. You are advised to regularly visit the university website to update yourself about the activities.

5 – Prescribed Readings

1. Bluman, A.G. (2004). Elementary Statistics. A Step by Step Approach. 5th Edition. McGraw-Hill Companies Incorporated. London.
2. Chaudhary, S.M. & KAmnal, S. (2017). Introduction to Statistical Theory Part-I. Eighth Edition. Ilmi Kitab Khana. Lahore.
3. Chaudhary, S.M. & KAmnal, S. (2017). Introduction to Statistical Theory Part-II. 8th Edition. Ilmi Kitab Khana. Lahore.
4. Daniel, W.W. (1995). Biostatistics: A foundation for Analysis in Health sciences. Sixth Edition. John Wiley and sons Incorporated. USA.
5. Harper, W.M. (1991). Statistics. Sixth Edition. Pitman Publishing, Longman Group, United Kingdom.
6. Hoel, P.G. (1976). Elementary Statistics. 4th Edition. John Wiley and Sons Incorporated, NewYork.
7. Kiani, G. H., & Akhtar, M. S. (2012). Basic statistics, Majeed Book Depot.
8. Khan, A. A., Mirza, S. H., Ahmad, M. I., Baig, I. & Yaqoob, M. (2011). Business Statistics, Qureshi Brothers Publishers.
9. Millar, R.L.: Intermediate Microeconomics, McGraw-Hill, Latest Edition.
10. Russel, R.R. and M. Wilkinson: Microeconomics: A Synthesis of modern and Neo-Classical Theory, John Wiley and Sons, New York, 1978.
11. Scherer, F.M.: Industrial Market Structure and Economics Performance.
12. Varian, H.R: Microeconomic Analysis, Norton W.W. Ince, New York, Latest Edition.

UNIT 01

INTRODUCTION

Written By: **Dr. Zahid Iqbal**
Reviewed By: **Dr. Muhammad Ilyas**

CONTENTS

Pages

Background

Statistics has been defined differently by different authors from time to time. One can find more than a hundred definitions in the literature of statistics.

The following are some important definitions of statistics.

1. Statistics is the branch of science which deals with the collection, classification and tabulation of numerical facts as the basis for explanations, description and comparison of phenomenon – Lovitt
2. The science which deals with the collection, analysis, and interpretation of numerical data - Corxton & Cowden
3. The science of statistics is the method of judging collective, natural or social phenomenon from the results obtained from the analysis or enumeration or collection of estimates -King
4. Statistics may be called the science of counting or science of averages or statistics is the science of the measurement of social organism, regarded as whole in all its manifestations – Bowley
5. Statistics is a science of estimates and probabilities -Boddington
6. Statistics is a branch of science, which provides tools (techniques) for decision making in the face of uncertainty (probability) - Wallis and Roberts

Objectives

After studying this unit, you will be able to;

1. Explain why knowledge of statistics is important.
2. Define statistics and provide an example of how statistics is applied.
3. Differentiate between descriptive and inferential statistics.
4. Classify variables as qualitative or quantitative, and discrete or continuous.

1.1 Meaning of Statistics

All definitions clearly point out the four aspects of statistics. Statistics is the science which deals with methods of collecting, classifying, presenting and interpreting numerical data.

Statistics is the discipline concerned with the collection, organization, and interpretation of numerical data, especially as it relates to the analysis of population characteristics by inference from sampling. The discipline of statistics addresses all elements of analysis, from study planning to the final presentation of results. Statistics is more than a compilation of computations techniques; it is a means of learning from data; it is "the servant of all sciences" (Neyman, 1955).

Functions of Statistics:

Statistics has four major functions.

- ➢ Collection of Data.
- ➢ Presentation of Data.
- ➢ Analysis of Data and
- ➢ Interpretation of results.

Population:

By population we mean aggregate of units which are under investigation according to some pre - determined objective and are available in specified area at a specified time period.

Population is of two types.

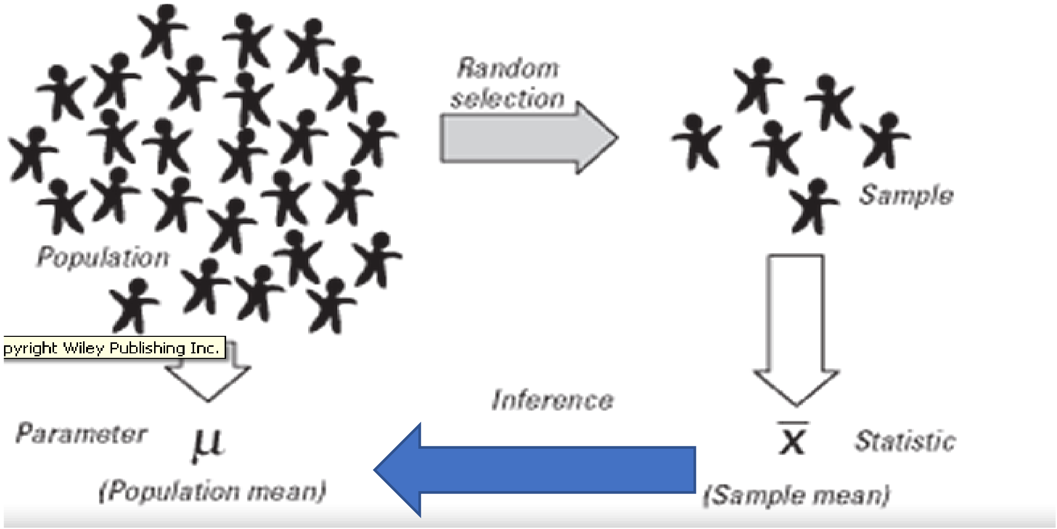
1) Finite population
2) Infinite population

Example:

1. All registered voter in Islamabad city
2. All students of Allma Iqbal Open University
3. All daily minimum temperatures in January for major Pakistani cities.

Sample:

A representative part of population which is under investigation is called a sample. Following figure illustrates the idea of population and sample



Parameter:

The numerical Characteristic of population is known as parameter.
Statistic:
The numerical Characteristic of sample is known as statistic

Properties of Representative Samples

Estimates calculated from sample data are often used to make inferences about populations. If a sample is representative of a population, then statistics calculated from sample data will be close to corresponding values from the population. Samples contain less information than full populations, so estimates from samples about population quantities always involve some uncertainty.

Random sampling, in which every potential sample of a given size has the same chance of being selected, is the best way to obtain a representative sample. However, it often impossible or impractical to obtain a random sample. Nevertheless, we often will make calculations for statistical inference as if a sample was selected at random, even when this is not the case. Thus, it is important to understand both how to conduct a random sample in practice and the properties of random samples.

Main divisions of Statistics:

Following are the main divisions of Statistics:

- **Descriptive statistics:** classification and diagrammatic representation of data.
- **Inferential Statistics:** to draw conclusion about population on the basis of sample drawn from it.
- **Data:** Any **measurement** of one or more characteristics recorded either from population or sample.

1.2 Importance of Statistics

There are three major functions in any business enterprise in which the statistical methods are useful. These are as follows:

(i) The planning of operations: This may relate to either special projects or to the recurring activities of a firm over a specified period.

(ii) The setting up of standards: This may relate to the size of employment, volume of sales, fixation of quality norms for the manufactured product, norms for the daily output and so forth.

(ii) The function of control: This involves comparison of actual production achieved against the norm or target set earlier. In case the production has fallen short of the target, it gives remedial measures so that such a deficiency does not occur again.

A worth noting point is that although these three functions-planning of operations, setting standards and control-are separate, but in practice they are very much interrelated.

Different authors have highlighted the importance of Statistics in business. For instance, Croxton and Cowden give numerous uses of Statistics in business such as project planning, budgetary planning and control, inventory planning and control, quality control, marketing, production and personnel administration. Within these also they have specified certain areas where Statistics is very relevant. Another author, Irwing W. Burr, dealing with the place of statistics in an industrial organization, specifies a number of areas where statistics is extremely useful. These are: customer wants and market research, development design and specification, purchasing, production, inspection, packaging and shipping, sales and complaints, inventory and maintenance, costs, management control, industrial engineering and research.

Statistical problems arising in the course of business operations are multitudinous.

As such, one may do no more than highlight some of the more important ones to emphasis the relevance of statistics to the business world. In the sphere of production, for example, statistics can be useful in various ways.

Statistical quality control methods are used to ensure the production of quality goods. Identifying and rejecting defective or substandard goods achieves this. The sale targets can be fixed on the basis of sale forecasts, which are done by using varying methods of forecasting. Analysis of sales affected against the targets set earlier would indicate the deficiency in achievement, which may be on account of several causes: (i) targets were too high and unrealistic (ii) salesmen's performance has been poor (iii) emergence of increase in competition (iv) poor quality of company's product, and so on. These factors can be further investigated.

Another sphere in business where statistical methods can be used is personnel management. Here, one is concerned with the fixation of wage rates, incentive norms and performance appraisal of individual employees. The concept of productivity is very relevant here. On the basis of measurement of productivity, the productivity bonus is awarded to the workers. Comparisons of wages and productivity are undertaken in order to ensure increases in industrial productivity. Statistical methods could also be used to ascertain the efficacy of a certain product, say, medicine. For example, a pharmaceutical company has developed a new medicine in the treatment of bronchial asthma. Before launching it on a commercial basis, it wants to ascertain the effectiveness of this medicine. It undertakes experimentation involving the formation of two comparable groups of asthma patients. One group is given this new medicine for a specified period and the other one is treated with the usual medicines. Records are maintained for the two groups for the specified period. This record is then analyzed to ascertain if there is any significant difference in the recovery of the two groups. If the difference is really significant statistically, the new medicine is commercially launched.

Application of Statistics

Statistics plays a vital role in every field of human activity. Statistics helps in determining the existing position of per capita income, unemployment, population growth rates, housing, schooling medical facilities, etc., in a country.

Now statistics holds a central position in almost every field, including industry, commerce, trade, physics, chemistry, economics, mathematics, biology, botany, psychology, astronomy, etc., so the application of statistics is very wide. Now we shall discuss some important fields in which statistics is commonly applied.

Business

Statistics plays an important role in business. A successful businessman must be very quick and accurate in decision making. He knows what his customers want; he should therefore know what to produce and sell and in what quantities.

Statistics helps businessmen to plan production according to the taste of the customers, and the quality of the products can also be checked more efficiently by using statistical methods. Thus, it can be seen that all business activities are based on statistical information. Businessmen can make correct decisions about the location of business, marketing of the products, financial resources, etc.

Economics

Economics largely depends upon statistics. National income accounts are multipurpose indicators for economists and administrators, and statistical methods are used to prepare these accounts. In economics research, statistical methods are used to collect and analyze the data and test hypotheses. The relationship between supply and demand is studied by statistical methods; imports and exports, inflation rates, and per capita income are problems which require a good knowledge of statistics.

Mathematics

Statistics plays a central role in almost all natural and social sciences. The methods used in natural sciences are the most reliable but conclusions drawn from them are only probable because they are based on incomplete evidence.
Statistics helps in describing these measurements more precisely. Statistics is a branch of applied mathematics. A large number of statistical methods like probability averages, dispersions, estimation, etc., is used in mathematics, and different techniques of pure mathematics like integration, differentiation and algebra are used in statistics.

Banking

Statistics plays an important role in banking. Banks make use of statistics for a number of purposes. They work on the principle that everyone who deposits their money with the banks does not withdraw it at the same time. The bank earns profits out of these deposits by lending it to others on interest. Bankers use statistical approaches based on probability to estimate the number of deposits and their claims for a certain day.

State Management (Administration)

Statistics is essential to a country. Different governmental policies are based on statistics. Statistical data are now widely used in making all administrative decisions. Suppose if the government wants to revise the pay scales of employees in view of an increase in the cost of living, and statistical methods will be used to determine the rise in the cost of living. The preparation of federal and provincial government budgets mainly depends upon statistics because it helps in estimating the expected expenditures and revenue from different sources. So statistics are the eyes of the administration of the state.

Accounting and Auditing

Accounting is impossible without exactness. But for decision making purposes, so much precision is not essential; the decision may be made on the basis of approximation, know as statistics. The correction of the values of current assets is made on the basis of the purchasing power of money or its current value.
In auditing, sampling techniques are commonly used. An auditor determines the sample size to be audited on the basis of error.

Natural and Social Sciences

Statistics plays a vital role in almost all the natural and social sciences. Statistical methods are commonly used for analyzing experiments results, and testing their significance in biology, physics, chemistry, mathematics, meteorology, research, chambers of commerce, sociology, business, public administration, communications and information technology, etc.

Astronomy

Astronomy is one of the oldest branches of statistical study; it deals with the measurement of distance, and sizes, masses and densities of heavenly bodies by means of observations. During these measurements errors are unavoidable, so the most probable measurements are found by using statistical methods.
Example: This distance of the moon from the earth is measured. Since history, astronomers have been using statistical methods like method of least squares to find the movements of stars and many mores.

1.3 Observation and Variable

Our reliance on statistics can be examined against the backdrop of empiricism and "the scientific method." **Empiricism** (from the Greek *empirikos* - experience) means "based on observation." **The scientific method** is not an actual method -- at least in the normal sense -- for there are no orderly rules of progress and no set procedures to follow. Nevertheless, it is based on a combination of empiricism and theory which uses several overlapping stages of reasoning. These stages of reasoning include:

Observation, in which the scientist observes what is happening, collects information, and studies facts relevant to the problem. In this stage, statistics suggests what can most advantageously be observed and how data might be collected.

Variable

To put it in very simple terms, a variable is an entity whose value varies. A variable is an essential component of any statistical data. It is a feature of a member of a given sample or population, which is unique, and can differ in quantity or quantity from another member of the same sample or population. Variables either are the primary quantities of interest or act as practical substitutes for the same. The importance of variables is that they help in operationalization of concepts for data collection. For example, if you want to do an experiment based on the severity of urticaria, one option would be to measure the severity using a scale to grade severity of itching. This becomes an operational variable. For a variable to be "good," it needs to have some properties such as good reliability and validity, low bias, feasibility/practicality, low cost, objectivity, clarity, and acceptance. Variables can be classified into various ways as discussed below.

Quantitative vs qualitative variable

A variable can collect either qualitative or quantitative data. A variable differing in quantity is called a quantitative variable (e.g., weight of a group of patients), whereas a variable differing in quality is called a qualitative variable (e.g., the Fitzpatrick skin type)

A simple test which can be used to differentiate between qualitative and quantitative variables is the subtraction test. If you can subtract the value of one variable from the other to get a meaningful result, then you are dealing with a quantitative variable (this of course will not apply to rating scales/ranks).

Quantitative variables can be either discrete or continuous Variable

Discrete variables are variables in which no values may be assumed between the two given values (e.g., number of Heads or the number of tails when coin is tossed or number appears when a dice is rolled).

Continuous variables, on the other hand, can take any value in between the two given values (e.g., height (between 5ft to 6ft) or weight (between 70kg and 71kg) it may takes any values). One way of differentiating between continuous and discrete variables is to use the “mid-way” test. If, for every pair of values of a variable, a value exactly mid-way between them is meaningful, the variable is continuous. For example, two values for the time taken for a weal to subside can be 10 and 13 min. The mid-way value would be 11.5 min which makes sense. However, for a number of weals, suppose you have a pair of values – 5 and 8 – the midway value would be 6.5 weals, which does not make sense.

Under the umbrella of qualitative variables, you can have nominal/categorical variables and ordinal variables

Nominal/categorical variables are, as the name suggests, variables which can be slotted into different categories (e.g., gender or type of psoriasis).
Ordinal variables or ranked variables are similar to categorical, but can be put into an order (e.g., a scale for severity of itching).

Dependent and independent variables

In the context of an experimental study, the dependent variable (also called outcome variable) is directly linked to the primary outcome of the study. For example, in a clinical trial on psoriasis, the PASI (psoriasis area severity index) would possibly be one dependent variable. The independent variable (sometime also called explanatory variable) is something which is not affected by the experiment itself but which can be manipulated to affect the dependent variable. Other terms sometimes used synonymously include blocking variable, covariate, or predictor variable. Confounding variables are extra variables, which can have an effect on the experiment. They are linked with dependent and independent variables and can cause spurious association. For example, in a clinical trial for a topical treatment in psoriasis, the concomitant use of moisturizers might be a confounding variable. A control variable is a variable that must be kept constant during the course of an experiment.

1.4 Collection of Data

Data sources could be seen as of two types, viz., secondary and primary. The two can be defined as under:

(i) **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.

(ii) **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.

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.

1.5 Summary

In a summarized manner, ‘Statistics’ means numerical information expressed in quantitative terms. As a matter of fact, data have no limits as to their reference, coverage and scope. At the macro level, these are data on gross national product and shares of agriculture, manufacturing and services in GDP (Gross Domestic Product). At the micro level, individual firms, how so ever small or large, produce extensive statistics on their operations. The annual reports of companies contain variety of data on sales, production, expenditure, inventories, capital employed and other activities. These data are often field data, collected by employing scientific survey techniques. Unless regularly updated, such data are the product of a one-time effort and have limited use beyond the situation that may have called for their collection. A student knows statistics more intimately as a subject of study like economics, mathematics, chemistry, physics and others. It 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.

1.6 SELF-ASSESSMENTS QUESTIONS

1. Define Statistics. Explain its types, and importance to trade, commerce and business.
2. “Statistics is all-pervading”. Elucidate this statement.
3. Write a note on the scope and limitations of Statistics.
4. What are the major limitations of Statistics? Explain with suitable examples.
5. Distinguish between descriptive Statistics and inferential Statistics.

1.7 SUGGESTED READINGS

Bluman, A.G. (2004). Elementary Statistics. A Step by Step Approach. 5th Edition. McGraw-Hill Companies Incorporated. London.

Chaudhary, S.M. & Kamal, S. (2017). Introduction to Statistical Theory Part-I. 8th Edition. Ilmi Kitab Khana. Lahore.

Chaudhary, S.M. & Kamal, S. (2017). Introduction to Statistical Theory Part-II. 8th Edition. Ilmi Kitab Khana. Lahore.

Daniel, W.W. (1995). Biostatistics: A foundation for Analysis in Health sciences. Sixth Edition. John Wiley and sons Incorporated. USA.

Harper, W.M. (1991). Statistics. Sixth Edition. Pitman Publishing, Longman Group, United Kingdom.

Hoel, P.G. (1976). Elementary Statistics. 4th Edition. John Wiley and Sons Incorporated, NewYork.

Kiani, G. H., & Akhtar, M. S. (2012). Basic statistics, Majeed Book Depot.

Khan, A. A., Mirza, S. H., Ahmad, M. I., Baig, I. & Yaqoob, M. (2011). Business Statistics, Qureshi Brothers Publishers.

UNIT 02

PRESENTATION OF DATA

Written By: **Dr. Zahid Iqbal**
Reviewed By: **Dr. Muhammad Ilyas**

CONTENTS

Pages

Introduction

In Statistics, presentation of data is very important. In real life problems, we have to deal with lot of data. Tables, Graphs and charts are used to summarize the data and to give the data an attractive look. This chapter will explain that how large data is summarized and presented in understandable form by using different statistical tools.

The presentation of data is not as easy as people think. There is an art to taking data and creating a story out of it that fulfills the purpose of the presentation.

This refers to the organization of data into tables, graphs or charts, so that logical and statistical conclusions can be derived from the collected measurements. Data may be presented in (3 Methods): - Textual - Tabular and - Graphical.

Whenever we hear the word statistics, we think there will be some information, data, figures, charts, graphs, diagrams, values or some numeric. Isn’t it? It means statistics relates to some data or values or numeric. Before discussing the data lets step back to the origin of statistics.

Statistics has developed gradually during the last few centuries. Now it is no longer restricted to the study of human population or the byproduct of administrative activities of the state. In the present era of information technology, statistics is regarded as one of the most import tools for making decisions and its scope has acquired broad spectrum in almost every sphere of life.

One of the number of meanings and definition of statistics is “the science of systematic collection, presentation, analysis and interpretation of numerical data to draw conclusions and to make decisions on the basis of such analysis”. In this sense the word statistics is used in singular.

Now, then what is data? Before interpreting the data lets understand the concept of observation. Anything that can be measured or observed is called an observation and the numbers or measurements that are collected as a result of observations is called data. In other words, the facts and figures that are collected, analyzed and interpreted are called data. Data is considered to be useful information.

Objectives

After studying this unit, you will be able to;

1. Discovery and communication are the two objectives of data visualization.
2. To introduce the students about the types of data and its presentation.
3. To give introduction of basic graphs, charts and diagrams.
4. Interpret a frequency table of quantitative data.
5. Be able to make a histogram or frequency polygon.
6. Differentiate normal distribution, positively skewed distribution and negatively skewed distribution.

2.1 Classification

It is the process of arranging observations into different classes or categories according to some common characteristics. The best example of classification is the process of sorting letters in a Courier Office. The data may be classified or represented by one, two or more characteristics at a time. If the data is classified according to one characteristic, it is called one-way classification and if the data is classified according to two characteristics, it is called two-way classification. As in Courier office the letters are firstly classified as district-wise which is an example of one way classification and then they are classified in to tehsil-wise that is second classification. In this manner the third classification may be mohallah or town. That is an example of three way classification. When the data is classified according to many characteristics, it is called many-way classification.

Classification is the process of arranging data into various groups, classes and subclasses according to some common characteristics of separating them into different but related parts.

Main objectives of classification:

1. To make the data easy and precise
2. To facilitate comparison
3. Classified facts expose the cause-effect relationship.
4. To arrange the data in proper and systematic way.

Construction of Frequency Distribution Table:

In statistics, a frequency distribution is a tabulation of the values that one or more variables take in a sample. Each entry in the table contains the frequency or count of the occurrences of values within a particular group or interval, and in this way the table summarizes the distribution of values in the sample.

The following steps are used for construction of frequency table.

i. The number of classes is to be decided.

ii. The appropriate number of classes may be decided by Yule's formula, which is as follows:

iii. $Number\ of\ classes = 2.5 \times n^{\frac{1}{4}}$.

iv. Another formula for no of Classes is
Number of Classes =1 + 3.33 log n

v. Where $n =$ *is the total number of observations.*

vi. The class interval is to be determined. It is obtained by using the relationship

$$Class\ interval = \frac{maximum\ value - minimum\ value}{no\ of\ classes}$$

The classification of the data primarily depends upon the following four basis:

i. Geographical (Spatial)
ii. Chronological (Temporal)
iii. Qualitative
iv. Quantitative

Some characteristics of a good classification are:

- Classification should be unambiguous.
- Classification should be stable.
- Classification should not be rigid.

Activity:

Provide some examples of classification based on spatial, temporal, qualitative and quantitative.

2.2 Tabulation

The process of making tables or arranging the data into rows and columns is called tabulation.
The following are the parts of tables which are involved in the construction of table.

Parts of a Table:

Title

Prefatory Notes

Stub	Box Head
	Column Caption
Row Captions	Body of the table

Footnote

Source note

i) Title:
It is the heading at the top of the table. It should be brief and self-explanatory. It describes the contents of the table.

ii) **Column captions and Box-head:**
The headings for different columns are called column captions and this part of column captions is called box-head.

iii) **Row captions and Stub:**
The headings for different rows are called row captions and this part of row captions is called stub.

iv) **Body of table:**
The entries in different cells of columns and rows in a table are called body of the table.

v) **Prefatory notes:**
The prefatory note is given after the title of the table. It is used to explain the contents of the data.

vi) **Footnotes:**
The footnotes are given at the end of the table. It is used to explain the contents of the data.

vii) **Source note:**
Source notes are given at the end of the table, which indicate the compiling agency, publication, the data and page of distribution.

Frequency Distribution:

A frequency distribution is a compact form of data in a table which displays the categories of observations according to their magnitudes and frequencies such that the similar or identical numerical values are grouped together. The number of values falling in a particular category is called the frequency of that category. It is denoted by f.

Construction of Frequency Distribution

Steps for the construction of frequency distribution:

i. Calculate the range of the data, where
Range=R=Maximum value in the data-Minimum value in the data

ii. Calculate the number of classes by the following formula:
C = 1 + 3.33 log n

iii. Decide about the width of the class by the following:
$h = \frac{R}{C}$ (approximately)

Open-end classes:

By open-end classes in a frequency table, either the lower limit of the 1st class or the upper limit of the last class is not a fixed number.

Class limits:

Each class is described by two numbers (the smaller number in the class limit is lower class limit and the upper number in the class limit is called upper class limit). These numbers are called class limits.

Class interval:

The class interval is the difference between the upper-class boundary and the lower-class boundary of the same class (not the difference between the class limits).

Class frequency:

The number of observations falling in a class is class is called class frequency.

Class mark:

The class mark or the midpoint is the value which divides the class into two equal parts. It is obtained by adding the lower- and upper-class limits or class boundaries of a class and dividing the resulting total by 2.

Class boundaries:

A class boundary is located midway between the upper limit of a class and the lower limit of the next class. The upper-class boundary of a class coincides with the lower-class boundary of the next class.

Cumulative Frequency:

It is obtained simply by adding the preceding frequencies including the frequency of that class.

Relative Frequency:

It is obtained by dividing the frequency of a class by the total frequency. It is generally expressed as a percentage.

Percentage Frequency

It is obtained by dividing the number of observations (frequency) within each data point or grouping of data points by the total number of observations and then multiply by Hundred. The sum of all the percentages corresponding to each data is 100.

Example:

The marks of 30 students of BS class are as follows:
51, 57, 64, 66, 71, 56, 58, 67, 80, 82, 71, 72, 70, 64, 66, 43, 30, 33, 38, 40, 46, 49, 55, 59, 60, 66, 70, 88, 70, 72
Make a suitable frequency distribution. Also find class boundaries and cumulative frequency.

Solution:

To construct a frequency distribution, we proceed as follow:

a. Range = R = Maximum value – Minimum value
Here Maximum Value = 92 Minimum Value = 30
So Range = R = 92 – 30 = 62

b. No. of classes = C =1 + 3.3 log 30 C = 1 + 3.3 log 30 here n = 30
C = 1 + 3.3 (1.4771) C = 1 + 4.87443
C = 5.87443 C = 6 (approximately)

c. Class interval = h = R / C = 62 / 6 = 10 (approximately)

Frequency distribution of students-marks data is:

Class Limits	**Tally**	**f**	**Class boundaries**	**Cumulative frequency**	**Relative Frequency**	**Percentage Frequency**
30-39	III	3	29.5-39.5	3	3/30=0.100	0.100*100=10%
40-49	IIII	4	39.5-49.5	3+4=7	4/30=0.133	0.133*100=13.3%
50-59	IIII I	6	49.5-59.5	7+6=13	6/30=0.200	0.200*100=20%
60-69	IIII II	7	59.5-69.5	13+7=20	7/30=0.233	0.233*100=23.3%
70-79	IIII II	7	69.5-79.5	20+7=27	7/30=0.233	0.233*100=23.3%
80-89	III	3	79.5-89.5	27+3=30	3/30=0.100	0.100*100=10%
Total		30			0.999 ≅ 1	99.9%≅100%

2.3 Diagrams and Graphs

Diagrammatic Presentation of Data gives an immediate understanding of the real situation to be defined by data in comparison to the tabular presentation of data or textual representations. Diagrammatic presentation of data translates pretty effectively the highly complex ideas included in numbers into more concrete and quickly understandable form. Diagrams may be less certain but are much more efficient than tables in displaying the data. There are many kinds of diagrams in general use.

Suppose you are interested to compare the marks of your mates in a test. How can you make the comparison interesting? It can be done by the diagrammatic representations of data. You can use a bar diagram, histograms, pie-charts etc. for this.

How will you find out the number of students in the various categories of marks in a certain test? What can you say about the marks obtained by the maximum students? Also, how can you compare the marks of your classmates in five other tests? Is it possible for you to remember the marks of each student in all subjects? No! Also, you don’t have the time to compare the marks of every student. Merely noting down the marks and making comparisons is not interesting at all.

A diagram is a symbolic representation of information according to visualization technique. Diagrams have been used since ancient times but became more prevalent during the Enlightenment. Sometimes, the technique uses a three-dimensional visualization which is then projected onto a two-dimensional surface. The word graph is sometimes used as a synonym for diagram.

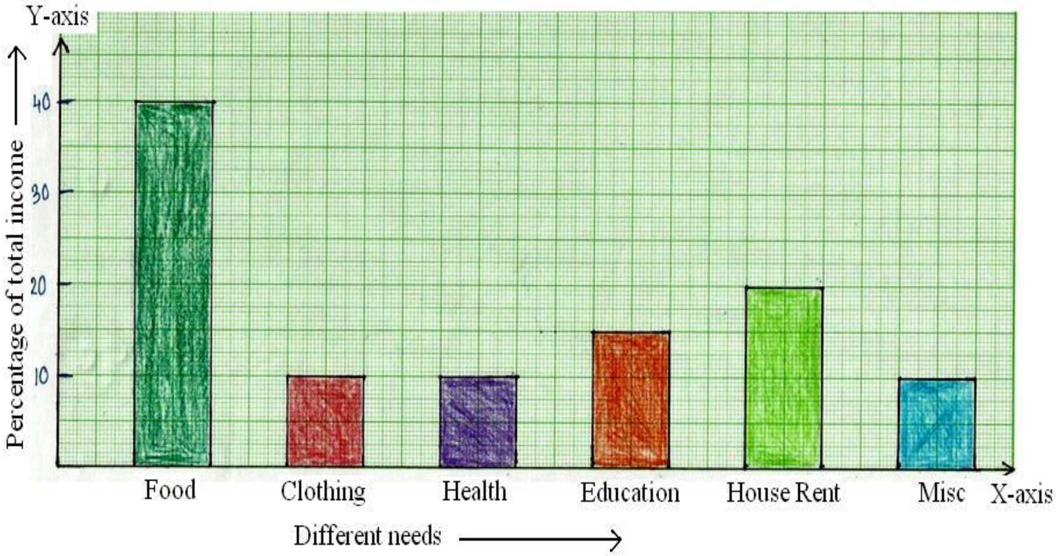
Simple Bar Diagram (Chart):

When the data consists of a single component and have not large variations, then a simple bar diagram is drawn. The first step in the construction is to arrange the data either in ascending or descending order if the data do not relate to time. Equi-spaced vertical or horizontal bards with moderate uniform width are then drawn. The length of bar is in proportion to the actual data.

The percentage of total income spent under various heads by a family is given below.

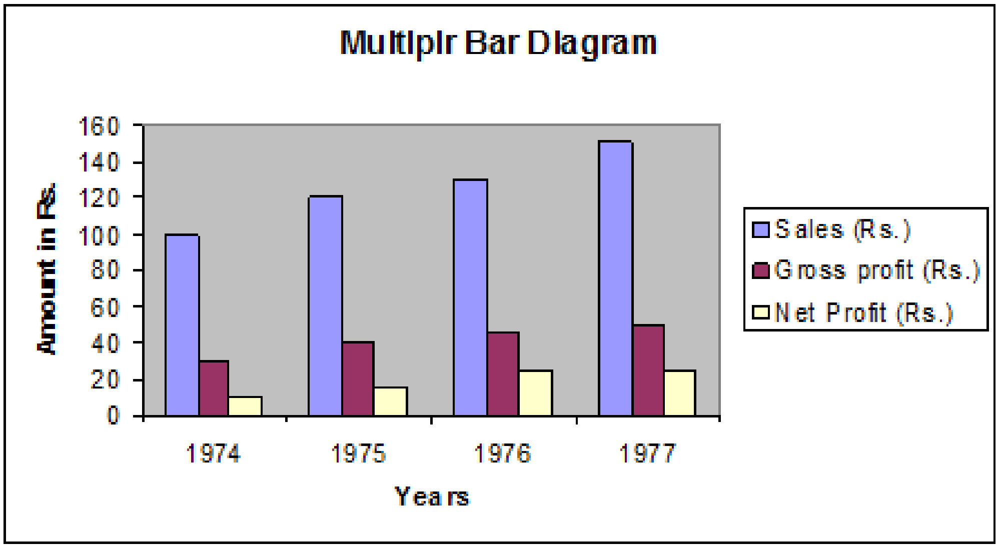
Different Heads	Food	Clothing	Health	Education	House Rent	Miscellaneous
% Age of Total Number	40%	10%	10%	15%	20%	5%

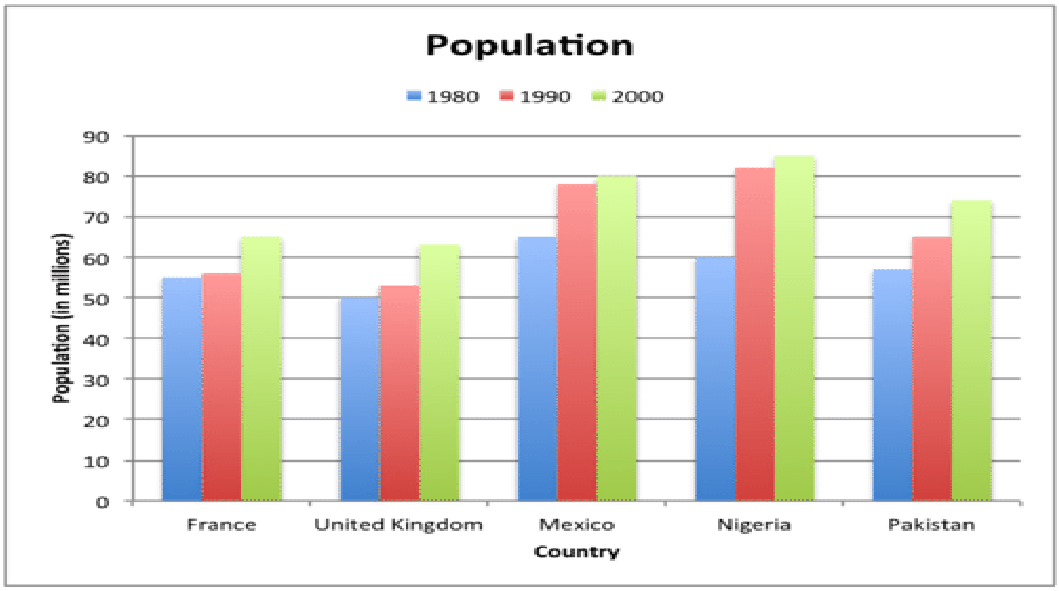
Represent the above data in the form of bar graph.



Multiple Bar Diagram

A multiple bar graph shows the relationship between different values of data. Each data value is represented by a column in the graph. In a multiple bar graph, multiple data points for each category of data are shown with the addition of columns.



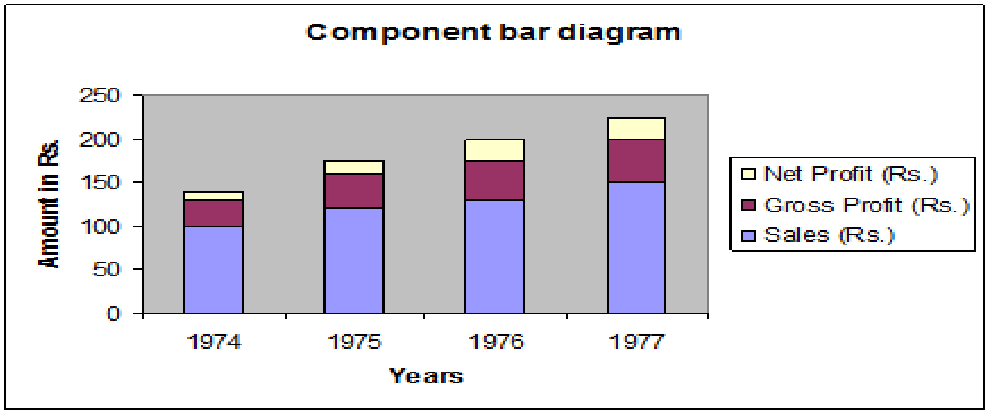


Subdivided Bar Diagram

This is also called Component bar diagram. Instead of placing the bars for each component side by side we may place these one on top of the other. This will result in a component bar diagram.

Example: Draw a component bar diagram for the following data

Year	Sales (Rs.)	Gross Profit (Rs.)	Net Profit (Rs.)
1974	100	30	10
1975	120	40	15
1976	130	45	25
1977	150	50	25



Pie Diagram

Pie diagram is a circular diagram where the whole circle represent a 'total' and the components of the total are represented by sectors of the pie diagram. Pie diagram is also called sector diagram. It is a popular diagram and is drawn when the components are to be shown for comparison. The total angle of the circle is 360^0 and the total quantity to be represented is taken equal to 360^0. The angles for each components are calculated and these angles are made in the circle to show different components.

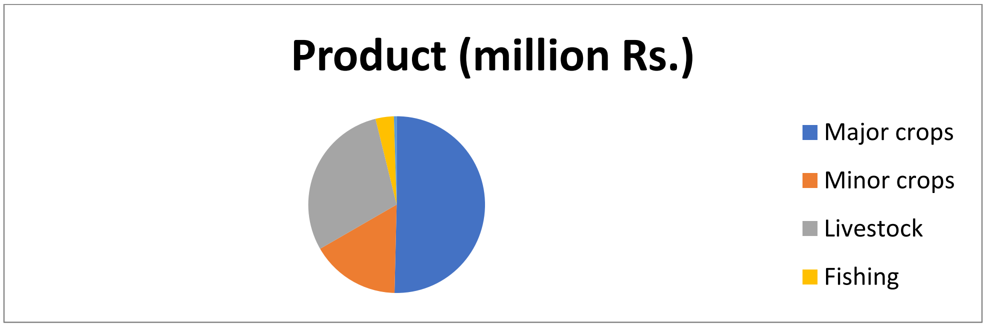
Example: The data on Agricultural Product at current factor cost for Pakistan for the year 1983-84 is given below. Make a pie diagram to represent the data.

Sub-sector	Product (million Rs.)
Major crops	46321
Minor crops	14971
Livestock	27096
Fishing	3082
Forestry	457

Source: Punjab Development Statistics, 1984

Solution: The necessary calculations to make the pie diagram are shown below and the diagram is shown.

Sub-sectors	Agriculture Product (million Rs.)	Angles of a sub-sectors
Major crops	46231	46231/91837 * 360 = 181.2
Minor crops	14971	14971/91837 *360 = 58
Livestock	27096	27096/91837 * 360 = 106.2
Fishing	3082	3082/91837 * 360 = 12.1
Forestry	457	457/91837 * 360 = 1.8
Total	91837	360



Graphs to describe **categorical variable** are bar diagram, pie diagram, pareto diagram and so on.
Graphs to describe **numerical variable** are histogram, ogive, stem and leaf plot.

2.4 SELF-ASSESSMENTS QUESTIONS

2.1. Construct a **frequency distribution table** for the following data
25 32 45 8 24 42 22 12 9 15 26 35 23
41 47 18 44 37 27 46 38 24 43 46 10 21 36
45 22 18.

2.2 Mercury contamination can be particularly high in certain types of fish. The mercury content (ppm) on the hair of 40 fishermen in a region thought to be particularly vulnerable are given below (From paper "Mercury content of commercially imported fish of the Seychelles, and hair mercury levels of a selected part of the population." *Environ. Research*, (1983), 305-312.)
13.26 32.43 18.10 58.23 64.00 68.20 35.35 33.92 23.94 18.28
22.05 39.14 31.43 18.51 21.03 5.50 6.96 5.19 28.66 26.29
13.89 25.87 9.84 26.88 16.81 38.65 19.23 21.82 31.58 30.13
42.42 16.51 21.16 32.97 9.84 10.64 29.56 40.69 12.86 13.80
Construct frequency distribution of the above data, also calculate the cumulative and percentage frequency distribution.

2.3 You are working for the Transport manager of a large chain of supermarkets which hires cars for the use of its staff. Your boss is interested in the weekly distances covered by these cars. Mileages recorded for a sample of hired vehicles from 'Fleet 1' during a given week yielded the following data:
138 164 150 132 144 125 149 157 161 150 168 126
138 186 163 146 158 140 109 136 148 152 144 145
145 109 154 165 135 156 146 183 105 108 135 153
140 135 142 128

a) Construct a frequency distribution.
b) Construct a pie chart
c) Construct steam and leaf plot.
d) Construct histogram and ogive curve.
e) Construct a Bar Diagram

SUGGESTED READINGS

Bluman, A.G. (2004). *Elementary Statistics. A Step by Step Approach*. 5th Edition. McGraw-Hill Companies Incorporated. London.

Chaudhary, S.M. & Kamal, S. (2017). *Introduction to Statistical Theory Part-I*. 8th Edition. Ilmi Kitab Khana. Lahore.

Chaudhary, S.M. & Kamal, S. (2017). *Introduction to Statistical Theory Part-II*. 8th Edition. Ilmi Kitab Khana. Lahore.

Daniel, W.W. (1995). *Biostatistics: A foundation for Analysis in Health Sciences*. Sixth Edition. John Wiley and sons Incorporated. USA.

Harper, W.M. (1991). *Statistics*. Sixth Edition. Pitman Publishing, Longman Group, United Kingdom.

Hoel, P.G. (1976). *Elementary Statistics*. 4th Edition. John Wiley and Sons Incorporated, New York.

Kiani, G. H., & Akhtar, M. S. (2012). *Basic statistics*, Majeed Book Depot.

Khan, A. A., Mirza, S. H., Ahmad, M. I., Baig, I. & Yaqoob, M. (2011). *Business Statistics*, Qureshi Brothers Publishers.

UNIT 03

MEASURE OF CENTRAL TENDENCY

Written By: **Dr. Zahid Iqbal**
Reviewed By: **Dr. Muhammad Ilyas**

CONTENTS

Pages

Introduction

Measure of Central tendency is a **single value** within the range of data which reflect the complete data set and **falls in the center** of the array**.** The purpose of measures of central tendency is to identify the location of the center of various distributions.

Objectives

After studying this unit, you will be able to;

1. Understand why the mean is the balancing point in a distribution of scores.
2. Understand the differences between statistics and parameters.
3. Understand the strengths and weaknesses of the mean, median and mode as measures of central tendency and when you might use one rather than the others.
4. Understand when you might a particular measure of central tendency to describe a set of data.
5. Understand why are there different formulas for calculating the median for an odd versus even number of scores for a variable.
6. Understand the purposes of measures of central tendency.
7. Calculate and interpret measures of central tendency (mode, median, mean) for a set of data.
8. Identify the mode from a frequency distribution table or figure.

3.1 Importance and Properties of Averages

- To present a brief picture of data- It helps in giving a brief description of the main feature of the entire data.
- Essential for comparison- It helps in reducing the data to a single value which is used for doing comparative studies.
- Helps in decision making- Most of the companies use measuring central tendency to plan and develop their businesses economy.
- Formulation of policies- Many governments rely on this medium while forming any policies

What are the good properties of good measures of central tendency?

i. It should be based on all observation of a set of values
ii. It should be rigorously defined
iii. It should be least affected by extreme values
iv. It should be easily computable
v. It should fluctuate least from sample to sample drawn from population.

3.2 Type of Averages

(a) Mean [Arithmetic Mean (AM), weighted Mean (WM), Geometric Mean (GM) and Hormonic Mean (HM)].
(b) Median
(c) Mode

3.3 Mean

The mean is the arithmetic average of all the observations in the data. It is also the balancing point of the data. The mean is found by adding up all of the observations and dividing by the total number of observations, either N or n depending upon whether you are dealing with the population or sample. The formula for the mean is

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} = \frac{\sum x}{n}$$

Where x_i are is the ith observation

Properties of the Arithmetic Mean
The arithmetic mean is a widely used measure of location. It has several important properties:

1. To compute a mean, the data must be measured at the interval or ratio level. Recall from Chapter 1 that ratio-level data include such data as ages, incomes, and weights, with the distance between numbers being constant.

2. All the values are included in computing the mean.

3. The mean is unique. That is, there is only one mean in a set of data.

4. The sum of the deviations of each value from the mean is zero. Expressed symbolically:

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

As an example, the mean of 3, 8, and 4 is 5. Then:

$$\Sigma(x - \bar{x})) = (3 - 5) + (8 - 5) + (4 - 5) = -2 + 3 - 1 = 0$$

Thus, we can consider the mean as a balance point for a set of data. To illustrate, we have a long board with the numbers 1, 2, 3, . . . , 9 evenly spaced on it. Suppose three bars of equal weight were placed on the board at numbers 3, 4, and 8, and the balance point was set at 5, the mean of the three numbers. We would find that the STATISTIC A characteristic of a sample.

EXAMPLE**:** Ufone is studying the number of monthly minutes used by clients in a particular cell phone rate plan. A random sample of 12 clients showed the following number of minutes used last month. 90, 77, 94, 89, 119, 112, 91, 110, 92, 100, 113, 83.

What is the arithmetic mean number of minutes used last month?
SOLUTION Using formula the sample mean is: Sample mean = Sum of all values in the sample Number of values in the sample Mean = Σx/ n = (90 + 77 + … + 83)/ 12 = 1,170 /12 = 97.5

The arithmetic mean number of minutes used last month by the sample of cell phone users is 97.5 minutes.

Weighted Mean

The weighted mean is a convenient way to compute the arithmetic mean when there are several observations of the same value. To explain, suppose the nearby Restaurant sold medium, large, and Biggie-sized soft drinks for Rs100, Rs 150, and

200, respectively. Of the last 10 drinks sold, 3 were medium, 4 were large, and 3 were Biggie-sized. To find the mean price of the last 10 drinks sold, we could use formula

Sample Mean = (100+100+100 + 150+150+150+150 + 200+200+200)/10
mean = 150

The mean selling price of the last 10 drinks is Rs. 150. An easier way to find the mean selling price is to determine the weighted mean. That is, we multiply each observation by the number of times it occurs.

We will refer to the weighted mean as $\overline{x_w}$. This is read “x bar sub w.”
$\overline{x_w}$ = = 3(100) + 4(150) + 3(200) /10 = 1500 /10 = 150

In this case, the weights are frequency counts. However, any measure of importance could be used as a weight. In general, the weighted mean of a set of numbers designated x_1, x_2, x_3, . . . , x_n with the corresponding weights w_1, w_2, w_3, . . . , w_n is computed by:

WEIGHTED MEAN= $\overline{x_w}$ = $$\frac{(w1x1 + w2x2 + w3x3 + ... + wnxn)}{(w1 + w2 + w3 + ... + wn\ )}$$

Illustration of above example

X_i	W_i	W_i X_i
100	3	3 x 100=300
150	4	4 x 150=600
200	3	3 x 200=600
Total	10	1500

$\overline{x_w}$ =1500/10 = 150

EXAMPLE**:** The Carter Construction Company pays its hourly employees Rs1650, Rs1900, or Rs. 2500 per hour. There are 26 hourly employees, 14 of whom are paid at the Rs.1650 rate, 10 at the Rs1900 rate, and 2 at the Rs.25.00 rate. What is the mean hourly rate paid the 26 employees?

SOLUTION To find the mean hourly rate, we multiply each of the hourly rates by the number of employees earning that rate. From formula

WEIGHTED MEAN= $\overline{x_w}$ = $\frac{(w1x1 + w2x2 + w3x3 + ... + wnxn)}{(w1 + w2 + w3 + ... + wn)}$

the mean hourly rate is $\overline{x_w}$ = {14(1650) + 10(1900) + 2(2500)}/(14 + 10 + 2)
= 47100/ 26 = Rs.1811.54

The weighted mean hourly wage is rounded to Rs. 1811.54.

a) Ungrouped Data

If the weights of 7 ear-heads of sorghum are 89, 94, 102, 107, 108, 115 and 126 g. find arithmetic mean.

$$A.M = \bar{x} = \frac{\sum x}{n} = \frac{89 + 94 + \cdots + 123}{7} = 105.86$$

b) Grouped Data

The following are the 405 soybean plant heights collected from a particular plot. Find the arithmetic mean of the plants height by **direct and indirect method.**

Plant height (cms)	8-12	13-17	18-22	23-27	28-32	33-37	38-42	43-47	48-52	53-57
No. of plants (f_i)	6	17	25	86	125	77	55	9	4	1

Solution:

1) Direct Method:

$$Arithmetic\ mean = \bar{x} = \frac{\sum f_i x_i}{\sum f_i}$$

$$where\ x = mid\ value\ of\ the\ corresponding\ classes$$
$$f = frequency$$

2) Indirect Method:

$$Arithmetic\ mean = \bar{x} = a + \frac{\sum f_i d_i}{N} X\ h$$

Where, d_i is the deviation$\left(d_i = \frac{x_i - a}{h}\right)$, a= assumed mean (central of X)=30 h=class interval = 5

Class interval	Frequency (f)	Mid value x_i	$f_i x_i$	$d_i = \frac{x_i - a}{h}$	$f_i d_i$
8-12	6	10	60	-4	-24
13-17	17	15	255	-3	-51
18-22	25	20	500	-2	-50
23-27	86	25	2150	-1	-86
28-32	125	30	3750	0	0
33-37	77	35	2695	1	77
38-42	55	40	2200	2	110
43-47	9	45	405	3	27
48-52	4	50	200	4	16
53-57	1	55	55	5	5
Total	405		$\sum f_i x_i = 12270$		$\sum f_i d_i = 24$

1) Direct Method:

$$Arithmetic\,mean = \bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{12270}{405} = 30.2963$$

2) Indirect Method:

a=assumed mean=30 (Which is in the Mid of X$_i$)
h=class interval=5 and N= Σf

$$Arithmetic\,mean = \bar{x} = a + \frac{\sum f_i d_i}{N} X\,h$$

$$A.M = \bar{x} = a + \frac{\sum f_i d_i}{N} X\,h = 30 + \left(\frac{24}{405}\right) X\,5$$

A.M = 30 + 120/405
A.M =30+0.2963
A.M = 30.2963

Geometric Mean

The geometric mean is useful in finding the average change of percentages, ratios, indexes, or growth rates over time. It has a wide application in business and economics because we are often interested in finding the percentage changes in sales, salaries, or economic figures, such as the gross domestic product, which compound or build on each other. The geometric mean of a set of n positive numbers is defined as the nth root of the product of n values. The formula for the geometric mean is written:

$$\text{Geometric Mean} = \text{GM} = \sqrt[n]{x_1\, x_2 \ldots\ldots x_n}$$

The geometric mean will always be less than or equal to (never more than) the arithmetic mean. Also, all the data values must be positive. As an example of the geometric mean, suppose you receive a 5% increase in salary this year and a 15% increase next year. The average annual percent increase is 9.886%, not 10.0%. Why is this so?

We begin by calculating the geometric mean. Recall, for example, that a 5% increase in salary is 105%. We will write it as 1.05.
$\text{GM} = \sqrt{}(1.05)(1.15) = 1.09886$

This can be verified by assuming that your monthly earning was Rs. 3,000 to start and you received two increases of 5% and 15%.

Raise 1 = Rs3,000(.05) = Rs. 150.00 ,
Raise 2 =Rs. 3,150(.15) = Rs. 472.50 ,
Total Rs. 622.50
Your total salary increase is Rs.622.50. This is equivalent to:
Rs. 3,000.00(.09886) = Rs. 296.59,
Rs. 3,296.58(.09886) =Rs. 325.91,
Total Rs. 622.50

Example: Compute the geometric mean of 2 and 8.

The formula of Geometric Mean is G.M= $\sqrt[n]{x_1 \times x_2 \times x_{3,} \ldots x_n}$

By putting the Values of X_1 and X_2

$$\text{GM} = \sqrt{2\times 8} = \sqrt{16} = 4$$

Example: Compute the Geometric mean of 2, 4, 8.

The formula of Geometric Mean is GM= $\sqrt[n]{x_1 \times x_2 \times x_{3,} \ldots x_n}$

By putting the Values of X_1 , X_2 and X_3

$$\text{GM} = \sqrt[3]{2\times 4 \times 8} = \sqrt[3]{2\times 4 \times 8} = \underline{\underline{4}}$$

Example: Calculate Geometric mean of the following data.

Solution:

x	**Log of** x
50	1.6990
72	1.8573
54	1.7324
82	1.9138
93	1.9685
	$\sum \log x = 9.1710$

GM= $\sqrt{50\times72\times54\times82\times93}$ =68.26

Or

$$\text{GM = Antilog } \frac{\sum \log x}{n} = \text{ Antilog } \frac{9.1710}{5} = \underline{\underline{Anti}}\log 1.8342 = 68.26$$

Example: Daily income of ten families are given below. Find out the Geometric Mean.

(Income Rs. 000) x	**log** x
85	1.9294
70	1.8451
15	1.1761
75	1.8751
500	2.6990
8	0.9031
45	1.6532
250	2.3979
40	1.6021
36	1.5563
	$\sum \log x = 17.6373$

$$\text{GM = Antilog of } \frac{\sum \log x}{n} = \text{Antilog } \frac{17.6373}{10} = 58.03$$

Example: For the grouped data given below obtain the geometric mean

X	10	100	1000	10000
F	2	3	2	3

Solution: By using the formula GM = Antilog $\frac{\sum f \log x}{n}$.

X	**F**	**Log x**	**f log x**
10 100 1000 10000	2 3 2 3	1 2 3 4	2 6 6 12
	n=$\sum f = 10$		$\Sigma f \log x = 26$

GM = Antilog $\dfrac{\Sigma f \log x}{n}$ = Antilog $\dfrac{26}{10}$ =398.1

Harmonic Mean

Example: Find the harmonic mean for the given data, 3, 5, 6, 6, 7, 10, 12.
Solution:

X	3	5	6	6	7	10	12	**Total**
1/X	0.3333	0.2000	0.1667	0.1667	0.1429	0.1000	0.0833	**1.2939**

The formula of Harmonic Mean is H.M= $\dfrac{n}{\sum\left(\frac{1}{x}\right)}$ = $\dfrac{7}{1.2939} = 5.8683$

Example: The monthly income of 10 families in a certain village are given below. Calculate the Harmonic Mean of monthly income.

Family	1	2	3	4	5	6	7	8	9	10
Income (in RS)	85	70	10	75	500	8	42	250	40	36

Solution: -

Family	**Income (x)**	**1/x**
1	85	0.01176
2	70	0.01426
3	10	0.1000
4	75	0.01333
5	500	0.0020
6	8	0.1250
7	42	0.0238
8	250	0.0040
9	40	0.0250
10	36	0.02778
n=10		$\sum$ (1/x) = 0.34693

Harmonic Mean = $\frac{n}{(1/x_1 + 1/x_2 + 1/x3\text{-----}1/xn)}$ OR $\frac{n}{\sum (1/x)}$

Harmonic Mean = $\frac{10}{0.34693}$ = $\underline{\underline{28.824}}$

Example: A truck company has 5 trucks to bring red soil from a pit of 5kms away from the brickyard.
The following table shows the time taken per load of all the 5 trucks.

Truck no	1	2	3	4	5
Minutes per hour	48	40	40	48	32

<u>Solution</u>: -

Truck no	**Minutes per hour**	**1/x**
1	48	0.0208
2	40	0.0250
3	40	0.0250
4	48	0.0208
5	32	0.0312
n = 5		$\sum$ x =0.1228

The formula for Harmonic Mean is HM= $\frac{n}{\sum \frac{1}{x}}$

Harmonic Mean = $\frac{n}{\sum (1/x)}$ = 5/(0.1228) = $\underline{\underline{40.716}}$

<u>Harmonic Mean of Grouped Data.</u>

Example: Calculate the harmonic Mean for the following data

Size of Items	6	7	8	9	10	11
Frequency	4	6	9	5	2	8

Solution: -

The formula of Harmonic Mean $= \dfrac{n}{\sum f\left(\dfrac{1}{x}\right)}$

X	**F**	**1/x**	**f (1/x)**
6	4	0.167	0.6668
7	6	0.143	0.8574
8	9	0.125	1.1250
9	5	0.111	0.5555
10	2	0.100	0.2000
11	8	0.090	0.7272
	$n = \sum f = 34$		$\sum f(1/x) = 4.1319$

$$\text{Harmonic Mean} = \frac{n}{\sum f(1/x)}$$

$$= \frac{34}{4.1319} = 8.23$$

3.4 Median

Median is the value of the variable that divides the ordered set of values into two equal halves. 50 percent values are to the left of the median and 50 percent are the right of the median.

Median for **odd number** of observations:

First, let’s examine these **five test** scores.

78 93 86 97 79

We need to put them **in order.**

78 79 86 93 97

The number in the middle is 86. Thus the Median is 86.

Median for **even number** of observation:

92 86 94 83 72 88

We need to put them **in order.**

72 83 86 88 92 94

Average of two middle is the median i.e. (86+ 88)/2 = 87

The median for this set is **87.**

Formula to calculate median:

Case-I: For **odd number** of observation

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

Case-II: For **even number** of observation:

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

Median for Grouped Data

The following are the 405 soybean plant heights collected from a particular plot. Find the Median of the plants height by**.**

The formula is, again,

$$\mathbf{Median = L + (\frac{n}{2} \quad - C) \times \frac{h}{f}}$$

Where:

L is the lower class boundary of the group containing the median

n is the total number of values and f is the frequency of the median group

C is the cumulative frequency of the groups before the median group

h is the Class Interval or the width

Example: Find the median, for the distribution of examination marks given below:

Marks	30 – 39	40- 49	50-59	60-69	70-79	80-89	90-99
No of students	08	87	190	304	211	85	20

Solution

Class Interval	Class Boundaries	Mid points (x)	Frequency (f)	Cumulative frequency (cf)
30 – 39	29.5 -39.5	34.5	08	08
40- 49	39.5- 49.5	44.5	87	95
50-59	49.5-59.5	54.5	190	285
60-69	59.5- 69.5	64.5	304	589
70-79	69.5 -79.5	74.5	211	800
80-89	79.5 -89.5	84.5	85	885
90-99	89.5 - 99.5	94.5	20	905
Total			905	

$n = \Sigma f = 905$ and $n/2 = \Sigma f / 2 = 905/2 = 452.5$th student which corresponds to marks in the class 60- 69 and class boundary 59.5 -69.5.

Therefore

Median = L+ (n/2 – C) x h/f
=59.5 + (452.5 – 285) x 10/304
Median = 59.5 + (167.5) x 10/304
Median = 59.5 + 1675/304
Median = 59.5 + 5.5098=65 Marks

3.5 Mode

Mode is that value of the variable which occurs **most frequently** in the series of observations of the variable.

A list of temperature for one week

Mon	Tues	Wed	Thurs	Fri	Sat	Sun
77	79	83	**77**	83	**77**	82

Here most frequently occurred number is 77.

Example: Find the Mode, for the distribution of examination marks given below:

Marks	30 – 39	40- 49	50-59	60-69	70-79	80-89	90-99
No of students	08	87	190	304	211	85	20

Solution

Class Interval	Class Boundaries	Mid points (x)	Frequency (f)	Cumulative frequency (cf)
30 – 39	29.5 -39.5	34.5	08	08
40- 49	39.5- 49.5	44.5	87	95
50-59	49.5-59.5	54.5	190	285
60-69	59.5- 69.5	64.5	304	589
70-79	69.5 -79.5	74.5	211	800
80-89	79.5 -89.5	84.5	85	885
90-99	89.5 - 99.5	94.5	20	905
Total			905	

Mode= L + $\frac{fm - f_1}{2fm - f_1 - f_2} \; X \; h$

Model class is that in which the frequency is highest i.e. frequency =304

Mode= 59.5 + $\frac{304-190}{2x304-190-211} \; X \; 10$

$$\text{Mode} = 59.5 + = \frac{114}{608-190-211} \text{ X } 10$$

$$\text{Mode} = 59.5 + \frac{114}{608-401} \text{ x } 10$$

$$\text{Mode} = 59.5 + \frac{114}{207} \text{ x } 10$$

Mode =59.5 + 5.507
Mode = 65.007
Mode = 65 Marks

3.6 Merits and Demerits of Averages

Mean

The arithmetic mean (or simply "mean") of a sample is the sum of the sampled values divided by the number of items in the sample.

Merits of Arithmetic Mean (AM)

1. Arithemetic Mean Rigidly Defined By Algebric Formula
2. It is easy to calculate and simple to understand
3. it based on all observations and it can be regarded as representative of the given data
4. It is capable of being treated mathematically and hence it is widely used in statistical analysis.
5. Arithmetic mean can be computed even if the detailed distribution is not known but some of the observation and number of the observation are known.
6. It is least affected by the fluctuation of sampling

Demerits of Arithmetic Mean

1. It can neither be determined by inspection or by graphical location
2. Arithmetic mean cannot be computed for qualitative data like data on intelligence honesty and smoking habit etc.
3. It is too much affected by extreme observations and hence it is not adequately represent data consisting of some extreme point
4. Arithmetic mean cannot be computed when class intervals have open ends

Median:

The median is that value of the series which divides the group into two equal parts, one part comprising all values greater than the median value and the other part comprising all the values smaller than the median value.

Merits of median

1. **Simplicity:-** It is very simple measure of the central tendency of the series. I the case of simple statistical series, just a glance at the data is enough to locate the median value.

2. **Free from the effect of extreme values: -** Unlike arithmetic mean, median value is not destroyed by the extreme values of the series.

4. **Certainty:** - Certainty is another merits is the median. Median values are always a certain specific value in the series.

5. **Real value:** - Median value is real value and is a better representative value of the series compared to arithmetic mean average, the value of which may not exist in the series at all.

6. **Graphic presentation:** - Besides algebraic approach, the median value can be estimated also through the graphic presentation of data.

6. **Possible even when data is incomplete:** - Median can be estimated even in the case of certain incomplete series. It is enough if one knows the number of items and the middle item of the series.

Demerits of median

1. **Lack of representative character:** - Median fails to be a representative measure in case of such series the different values of which are wide apart from each other. Also, median is of limited representative character as it is not based on all the items in the series.

2. **Unrealistic:-** When the median is located somewhere between the two middle values, it remains only an approximate measure, not a precise value.

3. **Lack of algebraic treatment: -**Arithmetic mean is capable of further algebraic treatment, but median is not. For example, multiplying the median with the number of items in the series will not give us the sum total of the values of the series.

However, median is quite a simple method finding an average of a series. It is quite a commonly used measure in the case of such series which are related to qualitative observation as and health of the student.

Mode:

The value of the variable which occurs most frequently in a distribution is called the mode.

Merits of mode:

1. **Simple and popular: -** Mode is very simple measure of central tendency. Sometimes, just at the series is enough to locate the model value. Because of its simplicity, it s a very popular measure of the central tendency.

2. **Less effect of marginal values:** - Compared top mean, mode is less affected by marginal values in the series. Mode is determined only by the value with highest frequencies.

3. **Graphic presentation:-** Mode can be located graphically, with the help of histogram.

4. **Best representative:** - Mode is that value which occurs most frequently in the series. Accordingly, mode is the best representative value of the series.

5. **No need of knowing all the items or frequencies:** - The calculation of mode does not require knowledge of all the items and frequencies of a distribution. In simple series, it is enough if one knows the items with highest frequencies in the distribution.

Demerits of mode:=

1. **Uncertain and vague:** - Mode is an uncertain and vague measure of the central tendency.

2. **Not capable of algebraic treatment:** - Unlike mean, mode is not capable of further algebraic treatment.

3. **Difficult: -** With frequencies of all items are identical, it is difficult to identify the modal value.

4. **Complex procedure of grouping:-** Calculation of mode involves cumbersome procedure of grouping the data. If the extent of grouping changes there will be a change in the model value.

5. **Ignores extreme marginal frequencies:-** It ignores extreme marginal frequencies. To that extent model value is not a representative value of all the items in a series. Besides, one can question the representative character of the model value as its calculation does not involve all items of the series.

3.7 SELF ASSESSMENT QUESTIONS

1. Consider the data below. This data represents the number of miles per gallon that 30 selected four-wheel drive sports utility vehicles obtained in city driving

12	17	16	14	16	18
16	18	17	16	17	15
15	16	16	15	16	19
10	14	15	11	15	15
19	13	16	18	16	20

   i) Calculate mean, median and mode of ungrouped data.
   ii) Construct the frequency distribution of the data.

2. A student recorded her scores on weekly math quizzes that were marked out of a possible 10 points. Her scores were as follows: 8, 5, 8, 5, 7, 6, 7, 7, 5, 7, 5, 5, 6, 6, 9, 8, 9, 7, 9, 9, 6, 8, 6, 6, 7. What is the Mean, Median and mode of her scores on the weekly math quizzes?

3. The following table of grouped data represents the weight (in pounds) of 100 computer towers. Calculate the mean, Median and Mode weight for a computer.

Weight (pounds)	Number of Computers
3 - 5	8
5 - 7	25
7 - 9	45
9 - 11	18
11 – 13	4

4. Calculate the Mean, Median and Mode from the frequency distribution for the weight of 120 students as given in the following Table;

Weights (Ibs)	110-119	120-129	130-139	140-149	150-159	160-169	170-179	180-189	190-199	200-209	210-219
f	1	4	17	28	25	18	13	6	5	2	1

SUGGESTED READINGS

Bluman, A.G. (2004). Elementary Statistics. A Step by Step Approach. 5th Edition. McGraw-Hill Companies Incorporated. London.

Chaudhary, S.M. & Kamal, S. (2017). Introduction to Statistical Theory Part-I. 8th Edition. Ilmi Kitab Khana. Lahore.

Chaudhary, S.M. & Kamal, S. (2017). Introduction to Statistical Theory Part-II. 8th Edition. Ilmi Kitab Khana. Lahore.

Daniel, W.W. (1995). Biostatistics: A foundation for Analysis in Health sciences. Sixth Edition. John Wiley and sons Incorporated. USA.

Harper, W.M. (1991). Statistics. Sixth Edition. Pitman Publishing, Longman Group, United Kingdom.

Hoel, P.G. (1976). Elementary Statistics. 4th Edition. John Wiley and Sons Incorporated, NewYork.

Kiani, G. H., & Akhtar, M. S. (2012). Basic statistics, Majeed Book Depot.

Khan, A. A., Mirza, S. H., Ahmad, M. I., Baig, I. & Yaqoob, M. (2011). Business Statistics, Qureshi Brothers Publishers.

UNIT 04

MEASURE OF DISPERSION

Written By: **Dr. Zahid Iqbal**
Reviewed By: **Dr. Muhammad Ilyas**

CONTENTS

Pages

Introduction

Dispersion means scattering of the observations among themselves or from a central value (Mean/ Median/ Mode) of data. We study the dispersion to have an idea about the variation. These measures give us an idea about the amount of dispersion in a set of observations. They give the answers in the same units as the units of the original observations.
There are two types of measures of dispersion.

1. **Absolute** measures of dispersion
2. **Relative** measures of dispersion

Difference between Absolute measures and Relative measures:

Absolute measures of Dispersion are expressed in same units in which original data is presented but these measures cannot be used to compare the variations between the two series. Relative measures are not expressed in units but it is a pure number. It is the ratios of absolute dispersion to an appropriate average such as co-efficient of Standard Deviation or Co-efficient of Mean Deviation.

1. Absolute measures of dispersion

I. Range
II. Mean deviation.
III. Standard deviation and Variance
IV. Quartile deviation

2. **Relative** measures of dispersion

I. Coefficient of range
II. Coefficient of mean deviation
III. Co-efficient of variation
IV. Coefficient of quartile deviation.

Objectives

After studying this unit, you will be able to;

- Comparative Study: Measures of dispersion give a single value indicating the degree of consistency or uniformity of distribution. This single value helps us in making comparisons of various distributions.
- The smaller the magnitude (value) of dispersion, higher is the consistency or uniformity and vice-versa.
- Reliability of an Average: A small value of dispersion means low variation between observations and average. It means the average is a good representative of observation and very reliable.
- A higher value of dispersion means greater deviation among the observations. In this case, the average is not a good representative, and it cannot be considered reliable.
- Control the Variability: Different measures of dispersion provide us data of variability from different angles, and this knowledge can prove helpful in controlling the variation. Especially in the financial analysis of business and Medical, these measures of dispersion can prove very useful.

4.1 The Range

The range is the absolute difference between the highest and the smallest values in a set of data.

Range is defined as the difference between the maximum or largest and the minimum or smallest observation of the given data. If x_m denotes the maximum observation and x_0 denotes the minimum observation, then the range is defined as

Range= largest value - smallest value= $X_m – X_0$

Example:

Suppose we have the following data of weights in Ibs (Pounds)
126 68 130 129 139 119 115 128 100 186 84 99
The largest value among the data=X_m=186 lbs
The Smallest value among the data=X_0=68 lbs
Range= largest value - smallest value= $X_m – X_0$ =186 - 68=118 lbs
The coefficient of range can be calculated by using the following formulae

$$CR = \frac{x_m-x_0}{x_m+x_0},=\frac{186-68}{186+68}=\frac{118}{254}=0.465 \text{ or } 46.5\%$$

Example:

The heights (in centimeters) of second semester students of BS Statistics are measured nearest to whole number as 56, 71, 62, 65, 59, 67, 64, 68, 70, 63. Determine the range and coefficient range.
Solution: It is simple to find out that $x_0 = 56\ cm$ and $x_m = 71\ cm$,
therefore

$$R = x_m - x_m = 71 - 56 = 15\ cm$$

and

$$CR = \frac{x_m - x_0}{x_m + x_0} = \frac{71 - 56}{71 + 56} = \frac{15}{127} = 0.118\ \ or\ 11.8\%$$

Activity: Calculate Range and Coefficient of Range for the following information.
5 6 7 7 9 4 5

Activity: Calculate Range and Coefficient of Range for the following information.
0.30, 2.22, 0.71, 3.53, 2.15, 4.18, 0.16, 1.25, 2.46,
8.83, 1.51, 0.92, 2.49, 2.55, 2.35, 0.50, 2.17, 2.35,
0.08, 1.22, 0.31, 1.52, 0.69, 0.24, 0.80, 1.16, 2.98,

3.72 0.58, 6.57, 0.02, 3.93, 0.02, 1.96, 2.56, 2.61,
1.67, 0.23, 8.61, 4.84, 4.67, 4.63, 5.31, 1.11, 0.54,
1.95, 0.20, 0.57, 2.51, 1.98.

Range is based on two extreme observations. It gives no weight to the central values of the data. It is a poor measure of dispersion and does not give a good picture of the overall spread of the observations with respect to the center of the observations. Let us consider three groups of data which have the same range:

Group A: 30, 40, 40, 40, 40, 40, 50
Group B: 30, 30, 30, 40, 50, 50, 50
Group C: 30, 35, 40, 40, 40, 45, 50

In all the three groups the range is 50 – 30 = 20. In group A there is a concentration of observations in the center. In group B the observations are concentrated in the extreme corners, and in group C the observations are almost equally distributed in the interval from 30 to 50. The range fails to explain differences in the three groups of data. This defect in range cannot be removed even if we calculate the coefficient of the range, which is a relative measure of dispersion. If we calculate the range of a sample, we cannot draw any inferences about the range of the population.

4.2 The Mean Deviation

The mean deviation (MD) also called mean absolute deviation is defined as the mean of absolute deviations of the observations from some suitable average. Usually the mean deviation from mean or mean deviation from median is useful. The mean deviation from median is preferred in the sense that the sum of absolute deviations from median is minimum. Consider the calculation of mean deviation and coefficient of mean deviation (CMD) from ungrouped data set with values x_1, x_2, ... , x_n. The formulae for mean deviation from mean and the corresponding coefficient are

$$MD(mean) = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n} \quad \text{and} \quad CMD = \frac{MD(mean)}{mean} = \frac{MD(mean)}{\bar{x}}.$$

Similarly the formulae for mean deviation from median and the corresponding coefficient are

$$MD(median) = \frac{\sum_{i=1}^{n} |x_i - \tilde{x}|}{n} \quad \text{and} \quad CMD = \frac{MD(median)}{median} = \frac{MD(median)}{\tilde{x}}.$$

Now consider the calculation of mean deviation and coefficient of mean deviation for the grouped data in the form of following frequency distribution.

Example:

The weights (in kg) of second semester students of BS Statistics are measured nearest to one decimal point as 37.7, 40.3, 43.3, 44.5, 46.9, 47.6, 48.6, 51.5, 52.4, 53.8. Determine the mean deviation from mean and median and coefficient of mean deviation from mean and median.

Solution: First we compute the mean and median as

mean $= \bar{x} = \frac{\sum_{i=1}^{10} x_i}{10} = \frac{466.6}{10} = 46.66$ kg, and median $= \tilde{x} = \frac{46.9+47.6}{2} =$ 47.25 kg.

The next step is to find sum of the absolute deviations as

X	$\boldsymbol{x-\bar{x}}$	$\boldsymbol{x-\tilde{x}}$	$\boldsymbol{\lvert x-\bar{x}\rvert}$	$\boldsymbol{\lvert x-\tilde{x}\rvert}$
37.7	−8.96	−9.55	8.96	9.55
40.3	−6.36	−6.95	6.36	6.95
43.3	−3.36	−3.95	3.36	3.95
44.5	−2.16	−2.75	2.16	2.75
46.9	+0.24	−0.35	0.24	0.35
47.6	+0.94	+0.35	0.94	0.35
48.6	+1.94	+1.35	1.94	1.35
51.5	+4.84	+4.25	4.84	4.25
52.4	+5.74	+5.15	5.74	5.15
53.8	+7.15	+6.55	7.15	6.55
	Total		41.68	41.20

Now

$$MD(mean) = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n} = \frac{41.68}{10} = 4.17 \text{ kg}$$

and

$$CMD = \frac{MD(mean)}{\bar{x}} = \frac{4.17}{46.66} = 0.0894 = 8.94\ \%.$$

Similarly

$$MD(median) = \frac{\sum_{i=1}^{n} |x_i - \tilde{x}|}{n} = \frac{41.2}{10} = 4.12\ kg$$

and

$$CMD = \frac{MD(median)}{\tilde{x}} = \frac{4.12}{47.25} = 0.0872 = 8.72\%$$

Activity:

Calculate the mean deviation from mean and median and coefficient of mean deviation from mean and median from the following data.

6.28 6.42 5.52 6.09 5.71 6.18 5.80 6.10 6.09 6.06 6.11 5.95 6.25
6.10 6.02 6.16 5.61 5.97 5.92 5.89 6.11 5.56 5.70 5.63 6.13 5.94
6.17 6.14 5.80 5.97

4.3 The Variance and Standard Deviation

Standard deviation is the most commonly used measure of dispersion. It is a measure of spread of data about the mean. It is defined as the square root of sum of squared deviations of the observations from their mean divided by the number of observations. In other words, the standard deviation of observations $x_1,\ x_2, \dots ,$ $x_n\ is\ defined\ as$

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{1}{n-1}\left[\sum\nolimits_{i=1}^{n} x^2 - \frac{(\sum_{i=1}^{n} x)^2}{n}\right]}$$

And Variance of observations $x_1,\ x_2, \dots ,\ x_n$ is defined as

$$s^2 = Square\left[\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}\right] = Square\left[\sqrt{\frac{1}{n-1}\left[\sum\nolimits_{i=1}^{n} x^2 - \frac{(\sum_{i=1}^{n} x)^2}{n}\right]}\right]$$

The corresponding coefficient of standard deviation also called coefficient of variation (CV) is defined as

$$CV = \frac{s}{\bar{x}} \times 100.$$

The coefficient of variation is often used for comparing the consistency of two or more data sets beside for comparing the dispersion. For the grouped data the standard deviation is defined as

$$s = \sqrt{\frac{\sum_{i=1}^{k} f_i(x_i - \bar{x})^2}{\sum_{i=1}^{k} f_i}} = \sqrt{\frac{\sum_{i=1}^{k} f_i x_i^2}{\sum_{i=1}^{k} f_i} - \left(\frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i}\right)^2}.$$

Example: Compute standard deviation, Variance and Coefficient of variation for the following data. 56, 71, 62, 65, 59, 67, 64, 68, 70, 63

Solution: It is better to construct a table of calculation for such a question as shown below.

$\boldsymbol{x}$	$(\boldsymbol{x}-\bar{\boldsymbol{x}})$	$(\boldsymbol{x}-\bar{\boldsymbol{x}})^2$	$\boldsymbol{x}^2$	$\boldsymbol{x-64}$	$\boldsymbol{d}^2$
56	−8.5	72.25	3136	−8	64
59	−5.5	30.25	5041	−5	25
62	−2.5	6.25	3844	−2	4
63	−1.5	2.25	4225	−1	1
64	−0.5	0.25	3481	0	0
65	0.5	0.25	4489	1	1
67	2.5	6.25	4096	3	9
68	3.5	12.25	4624	4	16
70	5.5	30.25	4900	6	36
71	6.5	42.25	3969	7	49
645	**0**	**202.50**	**41805**	**5**	**205**

In order to Calculate the standard deviation, we first need the mean of the data which is computed as

$$\bar{x} = \frac{\sum_{i=1}^{10} x_i}{10} = \frac{645}{10} = 64.5 \text{ cm}$$

Now the standard deviation can be computed as

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{10}(x_i-64.5)^2}{9}} = \sqrt{\frac{202.5}{9}} = \sqrt{22.5} = 4.74 \text{ cm.}$$

And the variance is

s^2 = Square $\left[\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}\right]$ = $(4.74)^2$ = 22.5

Next we use the computing formula to compute the standard deviation as

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{1}{n-1}\left[\sum\nolimits_{i=1}^{n} x^2 - \frac{(\sum_{i=1}^{n} x)^2}{n}\right]}$$

$$s = \sqrt{\frac{1}{10-1}\left[\sum\nolimits_{i=1}^{n} 41805 - \frac{(645)^2}{10}\right]}$$

$$s = \sqrt{\frac{1}{9}\left[\sum\nolimits_{i=1}^{n} 41805 - \frac{416025}{10}\right]}$$

$$= \sqrt{(41805 - 41602.5)/9} = \sqrt{20.25/9} = \sqrt{22.5} = 4.74 \text{ cm.}$$

4.4 Coefficient of Variation (CV)

The most important and commonly used relative measure of dispersion is **Coefficient of variation (CV).** Coefficient of variation is the percentage ratio of standard deviation and the arithmetic mean. **It is usually expressed in percentage.** The formula for C.V. is

$$\boldsymbol{C.V} = \frac{standard\ deviation}{arithmatic\ mean} \times 100 = \frac{\sigma}{\bar{x}} \times 100$$

The coefficient of variation (CV) is the ratio of Standard deviation to the mean. The higher the coefficient of variation, the greater the level of dispersion around mean. It is generally expressed as a percentage. Without units, it allows for comparison between distributions of values whose scales of measurement are not comparable. When we are presented with estimated values, the CV relates the standard deviation of the estimate to the value of this estimates. The lower the value of the coefficient of variation, the more precise the estimate.

Example:

Below are the scores of two cricket players A & B in 10 innings. Calculate Coefficient of Variation for Player A and B and decide which player is more consistent?

Player A	204	68	150	30	70	95	60	76	24	19
Player B	99	190	130	94	80	89	69	85	65	40

Solution:

Coefficient of variation (CV) for player A is

$$cv_A = \frac{\sigma_A}{\bar{A}} \times 100 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots. (1)$$

Coefficient of variation (CV) for Player B is

$$cv_B = \frac{\sigma_B}{\bar{B}} \times 100 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots. (2)$$

Now

$$\bar{A} = \frac{204 + 68 + \cdots 19}{10} = 79.6$$

$$\sigma_A = \sqrt{\frac{\sum(A_i - \bar{A})^2}{n}} = \sqrt{\frac{(204 - 79.6)^2 + (68 - 79.6)^2 + \cdots (19 - 79.6)^2}{10}}$$

$$= 58.23$$

$$cv_A = \frac{\sigma_A}{\bar{A}} \times 100 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots. (1)$$

$$= \frac{58.23}{79.6} \times 100 = 73.15\%$$

Similarly

$$\bar{B} = \frac{99 + 190 + \cdots \ldots\ldots\ldots..40}{10} = 94.1$$

$$\sigma_B = \sqrt{\frac{\sum(B_i - \bar{B})^2}{n}} = \sqrt{\frac{(99-94.1)^2 + (190-94.1)^2 + \cdots (40-94.1)^2}{10}}$$

$$= 41.12$$

And

$$cv_B = \frac{\sigma_B}{\bar{B}} \times 100 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.(2)$$

$$= \frac{41.12}{94.1} \times 100 = 43.70\%$$

Coefficient of variation of A is greater than coefficient of variation of B and hence we conclude that player B is more consistent.

Activity: Calculate the variance, S.D and C.V from the following marks obtained by 9 students.45 32 37 46 39 36 41 48 36

Activity: Calculate Variance, Standard deviation and Coefficient of Variation using direct, shortcut and step deviation method for Continuous grouped data, the data are given below:

Income	35—39	40—44	45--49	50--54	55--59	60--64	65--69
Frequency	13	15	17	28	12	10	05

4.5 Moments

Beyond the measures of central tendency and dispersion explained earlier, there are measures that further describe the characteristics of a distribution. Moments are a set of statistical parameters to measure a distribution. Four moments are commonly used:

- 1st moment - Mean (describes central value)
- 2nd moment - Variance (describes dispersion)
- 3rd moment - Skewness (describes asymmetry)
- 4th moment - Kurtosis (describes peakedness)

The formula for calculating moments is as follows when data is ungroup:

1st moment = $\mu_1 = \sum(x - \bar{x})/\ n$
2nd moment = $\mu_2 = \sum(x - \bar{x})^2\ /n$

3rd moment = $\mu_3 = \sum(x - \bar{x})^3 / n$
4th moment = $\mu_4 = \sum(x - \bar{x})^4 / n$

The formula for calculating moments is as follows when data is group:

1st moment = $\mu_1 = \sum f(x - \bar{x}) / n$
2nd moment = $\mu_2 = \sum f(x - \bar{x})^2 / n$
3rd moment = $\mu_3 = \sum f(x - \bar{x})^3 / n$
4th moment = $\mu_4 = \sum f(x - \bar{x})^4 / n$

4.6 Skewness

The term 'skewness' refers to lack of symmetry or departure from symmetry, e.g., when a distribution is not symmetrical (or is asymmetrical) it is called a skewed distribution. The measures of skewness indicate the difference between the manner in which the observations are distributed in a particular distribution compared with a symmetrical (or normal) distribution. The concept of skewness gains importance from the fact that statistical theory is often based upon the assumption of the normal distribution. A measure of skewness is, therefore, necessary in order to guard against the consequence of this assumption. In a symmetrical distribution, the values of mean, median and mode are alike. If the value of mean is greater than the mode, skewness is said to be positive. In a positively skewed distribution, mean is greater than the mode and the median lies somewhere in between mean and mode. A positively skewed distribution contains some values that are much larger than most other observations. A distribution is positively skewed when the long tail is on the positive side of the peak. On the other hand, if the value of mode is greater than mean, skewness is said to be negative. The following diagrams could clarify the meaning of skewness.

In a negatively skewed distribution, mode is greater than the mean and the median lies in between mean and mode. The mean is pulled towards the low-valued item (that is, to the left). A negatively skewed distribution contains some values that are much smaller than most observations. A distribution is negatively skewed when the long tail is on the negative side of the peak.

Generally, If Mean > Mode, the skewness is positive.
If Mean < Mode, the skewness is negative.
If Mean = Mode, the skewness is zero.

Skewness is measured in the following ways:

Karl Pearson's Coefficient of Skewness = $(Mean{-}Mode)/Standard\ Deviation$
or

Karl Pearson's Coefficient of Skewness = $3(Mean-Median)/Standard\ Deviation$

Moment based measure of skewness = $\beta_1 = \mu_3^2/\mu_2^3$
Pearson's coefficient of skewness = $\gamma 1 = \sqrt{\beta_1} = \mu_3/\mu_2^{3/2}$ is the most appropriate.

4.7 Kurtosis

Kurtosis refers to the degree of peakedness of a frequency curve. It tells how tall and sharp the central peak is, relative to a standard bell curve of a distribution. Kurtosis can be described in the following ways:

- Platykurtic– When the kurtosis < 0, the frequencies throughout the curve are closer to be equal (i.e., the curve is more flat and wide)

- Leptokurtic– When the kurtosis > 0, there are high frequencies in only a small part of the curve (i.e, the curve is more peaked)

- Mesokurtic- When the kurtosis = 0 To show the peakedness of a distribution, Kurtosis is measured in the following ways:

Moment based Measure of kurtosis = $\beta_2 = \mu_4/ \mu_2^2$
Coefficient of kurtosis = $\gamma_2 = \beta_2 – 3$

Example: Calculate first four moments about mean for ungrouped data for the following set of examination marks: 32, 36,36, 37, 39, 41, 45, 46, 48

X	32	36	36	37	39	41	45	46	48	ΣX=360
X - $\bar{X}$	-8	-4	-4	-3	-1	1	5	6	8	$\Sigma(X - \bar{X}) = 0$
$(X - \bar{X})^2$	64	16	16	9	1	1	25	36	64	$\Sigma(X - \bar{X})^2 = 232$
$(X - \bar{X})^3$	-512	-64	-64	-27	-1	1	125	216	512	$\Sigma(X - \bar{X})^3$=186
$(X - \bar{X})^4$	4096	256	256	81	-1	1	625	1296	4096	$\Sigma(X - \bar{X})^4$=10708

1st moment = $\mu_1 = \sum(x - \overline{x})/ n = 0$ Marks
2nd moment = $\mu_2 = \sum(x - \overline{x})^2/n = 232/10 = 23.2$ (Marks)^2
3rd moment = $\mu_3 = \sum(x - \overline{x})^3/ n = 186/10 = 18.6$ (Marks)^3
4th moment = $\mu_4 = \sum(x - \overline{x})^4/n = 10708/10 = 1070.8$ (Marks)^4

Skewness

Moment based measure of Skewness = $\gamma 1 = \sqrt{}\beta_1 = \mu_3/\mu_2^{3/2} = 186/\sqrt{12487168} = 186/3533.72$ Moment based measure of Skewness = $\gamma 1 = 0.0526$, The Data is very close to symmetry

Kurtosis

Moment based Measure of kurtosis = $\beta_2 = \mu_4/ \mu_2^2 = 1070.8/(23.2)^2 = 1070.8/538.24 = 1.989$

• Leptokurtic– When the kurtosis > 0, there are high frequencies in only a small part of the curve (i.e, the curve is more peaked)

Activity:

Calculate skewness and kurtosis for grouped data (using a continuous grouped case formula). The following distribution relates to the number of assistants in 50 retail establishments, the data are given below:

No of Assistant	0	1	2	3	4	5	6	7	8
Frequency	3	4	6	7	10	6	5	5	3

4.8 SELF ASSESSMENT QUESTIONS

Q1. The following data is of Batsman Score in a series

30, 91, 0, 64, 42, 80, 30,

Calculate variance, standard deviation, Co-efficient of Variation, Skewness and Kurtosis

Q2. The following table gives the frequency distribution of the amounts of telephone bills for April 2013 for a sample of 50 students**.**

Amount of telephone bills	Number of students
40-70	9
70-100	11
100-130	16
130-160	10
160-190	4

Calculate variance, standard deviation, Co efficient of Variation, Skewness and Kutosis

Q3. The production of jute goods in different days of first and second of the year are shown below

Class interval of production	2-2.5	2.5-3.0	3.0-3.5	3.5-4.0	4.0-4.5
No. of days in the first half of the year	12	48	70	35	15
No. of days in the second half of the year	5	38	80	50	7

In which part of the year the production level is homogeneous?

Q4. Terrier and SFP are two stocks traded on the New York Stock Exchange. For the past seven weeks Friday closing price (dollars per share) was recorded:

Terrier	32	35	34	36	31	39	41
SFP	51	55	56	52	55	52	57

1. Compute the range, standard deviation, variance , Coefficient of Variation, Skewness and Kurtosis for Terrier.
2. Compute the range, sample standard deviation, and sample variance, Coefficient of Variation, Skewness and Kurtosis for SFP.

SUGGESTED READINGS

Bluman, A.G. (2004). Elementary Statistics. A Step by Step Approach. 5th Edition. McGraw-Hill Companies Incorporated. London.

Chaudhary, S.M. & Kamal, S. (2017). Introduction to Statistical Theory Part-I. 8th Edition. Ilmi Kitab Khana. Lahore.

Chaudhary, S.M. & Kamal, S. (2017). Introduction to Statistical Theory Part-II. 8th Edition. Ilmi Kitab Khana. Lahore.

Daniel, W.W. (1995). Biostatistics: A foundation for Analysis in Health sciences. Sixth Edition. John Wiley and sons Incorporated. USA.

Harper, W.M. (1991). Statistics. Sixth Edition. Pitman Publishing, Longman Group, United Kingdom.

Hoel, P.G. (1976). Elementary Statistics. 4th Edition. John Wiley and Sons Incorporated, NewYork.

Kiani, G. H., & Akhtar, M. S. (2012). Basic statistics, Majeed Book Depot.

Khan, A. A., Mirza, S. H., Ahmad, M. I., Baig, I. & Yaqoob, M. (2011). Business Statistics, Qureshi Brothers Publishers.

UNIT 05

RANDOM VARIABLE AND PROBABILITY

Written By: **Dr. Zahid Iqbal**
Reviewed By: **Dr. Muhammad Ilyas**

CONTENTS

Pages

Introduction

Chance is what makes life worth living – if everything was known in advance, imagine the disappointment! If decision-makers had perfect information about the future as well as the present and the past, there would be no need to consider the concepts of probability. However, it is usually the case that uncertainty cannot be eliminated and hence its presence should be recognized and used in the process of decision- making. Information about uncertainty is often available to the decision-maker in the form of probabilities. This chapter introduces the fundamental concepts of probability. In other subjects (e.g. Management Science Methods) you may make full use of probabilities in decision trees and highlight ways in which such information can be used. Our treatment of probability in this module is quite superficial. The concepts of probability are simple but applying them in some circumstances can be very difficult! As a preliminary we consider the basic ideas concerning sets.

Objectives

After studying this unit, you will be able to understand the ideas of randomness and variability, and the way in which these link to probability theory to allow the systematic and logical collection of statistical techniques of great practical importance in many applied areas.

5.1 Random Experiment

An experiment is any well-defined, repeatable procedure, usually involving one or more chance events. One repetition of the procedure is called a trial. When a trial is conducted, it results in some outcome. (Note that, in the usual case where the experiment involves randomness, different trials can result in different outcomes.) A random variable is a measurable (numeric) quantity associated with the outcome of an experiment. An event is a statement about the outcome of the experiment that is either true or false.

This topic is returned to, and made more substantial use of, in the Statistics, economics and Management Mathematics courses.

• **Sample Space, S**. For a given experiment the sample space, S, is the set of all possible outcomes. • **Event, E.** This is a subset of S. If an event E occurs, the outcome of the experiment is contained in E.

Example When tossing a coin we might have the following sets/events: S = { H, T }

E = { H } or E = { T } (Note: H is the event a head appears, T a tail)

Example When throwing a die: S = {1,2,3,4,5,6} E = {3,4} F = {4,5,6 }

Example . Suppose you arrive at a railway station at a random time. There is a train once an hour. The random experiment is to observe the number of (rounded up) minutes that you wait before a train leaves.

The elementary outcomes here are the integers (whole numbers) 1 to 60, and the sample space is {1,2,3…60}. The event that 'you wait less than 10 minutes' is the subset {1,2,3,4,5,6,7,8,9}.

Example: We can discuss the experiment of drawing 5 cards at random from a deck of 52 playing cards. On a given trial, let's say the selected cards may be the four aces (spades, clubs, diamonds, and hearts) and the king of spades. This is the outcome of the trial. A different trial would probably result in different cards being selected, and hence a different outcome. Let's let **A** = the number of aces drawn. Then A is a random variable. For this particular trial, the value of **A** is 4. If the cards selected in the trial had been the 2, 3, 4, 5 and 6 of clubs, the value of **A** would have been 0.

5.2 Random Variable

1. A random variable is a variable which take a specific values with specific probabilities.

   It can be thought as a variable whose values depends on outcome of an uncertain event.

2. We usually use the capital alphabet to denote the random variables e.g. W, X,Y or Z etc.

Example: Let *X* be the outcome of the roll of a die. Then *X* is a random variable. Its possible values are 1, 2, 3, 4, 5, and 6; each of these possible values has probability 1⁄6.

The word "**random**" in the term "**random variable**" does *not* necessarily imply that the outcome is completely random in the sense that all values are equally likely. Some values may be more likely than others; "**random**" simply means that the value is uncertain.

When you think of a random variable, immediately ask yourself

- What are the possible values?
- What are their probabilities?

Example: Let *Y* be the *sum* of two dice rolls.

- Possible values: $\{2, 3, 4, 5, 6, 7, 8, 9, 10, 11\ , 12\}$.

Their probabilities: 2 has probability 1/36, 3 has probability 2/36, 4 has probability 3/36, etc. (The important point here is not the probabilities themselves, but rather the fact that such a probability can be assigned to each possible value.)

The probabilities assigned to the possible values of a random variable are its **distribution**. A distribution completely describes a random variable.
A random variable is called **discrete** if it has count ably many possible values; otherwise,

it is called **continuous.** For example, if the possible values are any of these:

- $\{1, 2, 3, \dots\ , \}$
- $\{\dots\ , -2, -1, 0, 1, 2, \dots\}$
- $\{0, 2, 4, 6, \dots\}$
- $\{0, 0.5, 1.0, 1.5, 2.0, \dots\}$

any finite set then the random variable is discrete.

If the possible values are any of these:

- all numbers between 0 and ∞
- all numbers between $-\infty$ and ∞
- all numbers between 0 and 1

then the random variable is continuous.

Sometimes, we approximate a discrete random variable with a continuous one if the possible values are very close together; e.g., stock prices are often treated as continuous random variables.

The following quantities would typically be modeled as discrete random variables:

- The number of defects in a batch of 20 items.
- The number of people preferring one brand over another in a market research study.
- The credit rating of a debt issue at some date in the future.

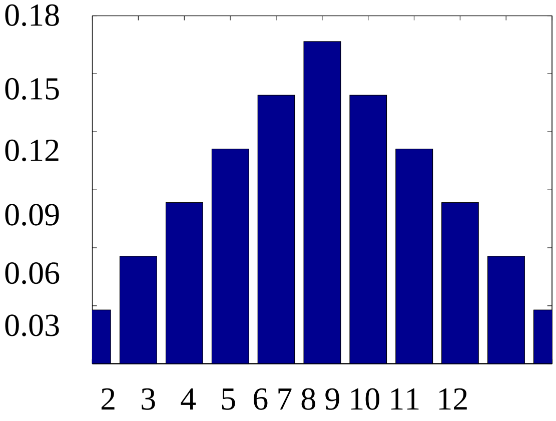
The following would typically be modeled as continuous random variables:

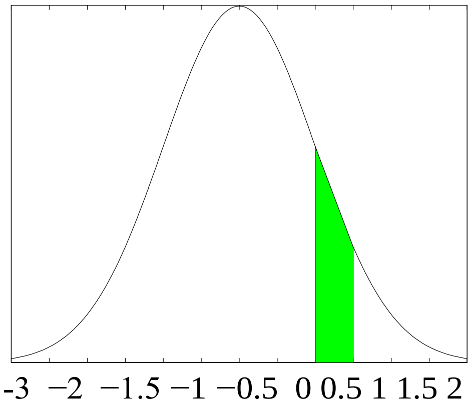
- The yield on a 10-year Treasury bond three years from today.
- The proportion of defects in a batch of 10,000 items.
- The time between breakdowns of a machine.

5.3 Discrete Distribution

The rule that assigns specific probabilities to specific values for a discrete random variable is called its **probability mass function** or **pmf or probability density function** or **pdf**. If X is a discrete random variable then we denote its pmf by P_X. For any value x, $P(X = x)$ is the probability of the event that $X = x$; i.e., $P(X = x)$ = probability that the value of X is x.

Example: If X is the outcome of the roll of a die, then $P(X = 1) = P(X = 2) = \cdots = P(X = 6) = 1/6$, and $P(X = x) = 0$ for all other values of x.





In Figure above the Left panel shows the probability mass function or pdf for the sum of two dice; the possible values are 2 through 12 and the heights of the bars give their probabilities. The bar heights sum to 1. Right panel shows a probability density for a continuous random variable. The probability $P\ (1 \leq X \leq\ \ 1.5)$ is given by the shaded area under the curve between 1 and 1.5. The total area under the curve is 1. The probability of any particular value, e.g., $P\ (X = 1)$ is zero because there is no area under a single point.

We always use capital letters for random variables. Lower-case letters like x and y stand for possible values (i.e., numbers) and are not random.

A pmf is graphed by drawing a vertical line of height $P\ (X = x)$ at each possible value x. It is similar to a histogram, except that the height of the line (or bar) gives the *theoretical probability* rather than the *observed frequency*.

5.4 Continuous Distribution

1. The distribution of a continuous random variable cannot be specified through a probability mass function because if X is continuous, then $P\ (X\ = x) = 0$ for all x; i.e., the probability of any particular value is zero. Instead, we must look at probabilities of *ranges* of values.

2. The probabilities of ranges of values of a continuous random variable are determined by a
density function. The density of X is denoted by $f(X)$.

The area under a density is always.

i. The probability that X falls between two points a and b is the area under $f(X)$ between the points a and b.
ii. The familiar bell-shaped curve is an example of a density.

3. The **cumulative distribution function** or **cdf** gives the probability that a random variable X takes values less than or equal to a given value x. Specifically, the cdf of X,
denoted by $F(X)$, is given by

$$F_X\ (x) = \ P\ (X \leq x).$$

So, $F_X\ (x)$ is the area under the density f_X to the left of x.

4. For a continuous random variable, $P\ (X = x) = 0$; consequently, $P\ (X \leq x) = \ P$

(X < x). For a discrete random variable, the two probabilities are not in general equal.

5. The probability that X falls between two points a and b is given by the difference between the cdf values at these points:

$$P\ (a < X \leq b) = F_X\ (b) - F_X\ (a).$$

Since $F_X(b)$ is the area under f_X to the left of b and since $F_X(a)$ is the area under f_X to the left of a, their difference is the area under f_X between the two points.

5.5 Expectations of Random Variables

1. The **expected value** of a random variable is denoted by $E[X]$. The expected value can be thought of as the “average” value attained by the random variable; in fact, the expected value of a random variable is also called its **mean**, in which case we use the notation $\mu_X = E(X)$. (μ is the Greek letter mu).

2. The formula for the expected value of a discrete random variable is this:

$$E[X] = \Sigma\ xP\ (X = x).\ \textit{for}\ \text{all possible}\ x$$

In words, the expected value is the sum, over all possible values x, of x times its probability
$P\ (X = x)$.

3. Example: The expected value of the roll of a die is

$$1(1/_6) \pm 2(1/_6) + 3(1/_6) + \cdots + 6\_\ (1/_6) = 21/6 = 3.5.$$

Notice that the expected value is not one of the possible outcomes:
you can’t roll a 3.5. However, if you average the outcomes of a large number of rolls, the result approaches 3.5.

4. We also define the expected value for a function of a random variable. If g is a function (for example, $g(x) = x^2$), then the expected value of $g(X)$ is $E\ [g(X)] = \Sigma\ g(x)P\ (X = x)$.all possible value of x.

For example, $E[X^2] = \Sigma\ x^2P\ (X = x)$, all possible x

In general, $E[g(X)]$ is not the same as $g(E[X])$. In particular, $E[X^2]$ is not the same as $(E[X])^2$.

5. The expected value of a continuous random variable cannot be expressed as a

sum; instead it is an integral involving the density. (If you don’t know what that means, don’t worry; we

won’t be calculating any integrals.).

6. The **variance** of a random variable X is denoted by either $Var[X]$ or $\sigma^2 x$. (σ is the Greek letter sigma.) The variance is defined by

$$\sigma^2_X = E[(X - \mu_X)^2];$$

this is the expected value of the squared difference between X and its mean. For a discrete distribution, we can write the variance as

$$\sigma^2_X = \Sigma(x - \mu_X)^2 P(X = x).$$

7. An alternative expression for the variance (valid for both discrete and continuous random variables) is

$$\sigma^2_X = E[(X^2)] - [\mu_X]^2.$$

$$\sigma^2_X = E[(X^2)] - [E(X)]^2.$$

This is the difference between the expected value of X^2 and the square of the mean of X.

8. The standard deviation of a random variable is the square-root of its variance and is denoted by σ_X. Generally speaking, the greater the standard deviation, the more spread-out the possible values of the random variable.

9. In fact, there is a Chebyshev rule for random variables: if $m > 1$, then the probability that X falls within m standard deviations of its mean is at least $1 - (1/m^2)$; that is,

$$P(\mu_x - m\sigma_X \leq X \leq \mu_X + m\sigma_X) \geq 1 - (1/m^2).$$

10. Find the variance and standard deviation for the roll of one die. Solution: We use the formula $Var[X] = E[X^2] (E[X])$. We found previously that $E[X] = 3.5$, so now we need to find $E[X^2]$. This is given by

Σ

Thus, $E[X^2] =$ $\Sigma x^2 P_X(x) = 1^2(\tfrac{}{6}) + 2^2(\tfrac{}{6}) + \cdots + 6^2(\tfrac{}{6}) = 15.167$. σ

$${}^2 x = Var[X] = E[X^2] - (E[X])^2 = 15.167 - (3.5)^2 = 2.917$$

and $\sigma = \sqrt{2.917} = 1.708$.

5.6 Linear Transformations of Random Variables

1. If X is a random variable and if a and b are any constants, then $a + bX$

is a **linear transformation** of X. It scales X by b and shifts it by a. A linear transformation of X is another random variable; we often denote it by Z.

Example: Suppose you have investments in Japan. The value of your investment (in yen) one month from today is a random variable X. Suppose you can convert yen to dollars at the rate of b dollars per yen after paying a commission of a dollars. What is the value of your investment, in dollars, one month from today?

Activity: Your salary is Rs. a per year. You earn a bonus of b dollars for every Rs. of sales you bring in. If X is what you sell, how much do you make?

Example: It takes you exactly 16 minutes to walk to the train station. The train ride takes X hours, where X is a random variable. How long is your trip, in minutes?

If $Z = a + bX$, then $\quad E[Z] = E[a + bX] = a + bE[X] \quad = a + b\mu_X$

and $\quad \sigma_Z^2 \;= Var\,[a + bX] = \quad {}_X \;\; b^2\sigma^{\,2}.$

2. Thus, the expected value of a linear transformation of X is just the linear transformation of the expected value of X. Previously, we said that $E[g(X)]$ and $g(E[X])$ are generally different. The *only* case in which they are the same is when g is a linear transformation: $g(x)=a + bx$.

3. Notice that the variance of $a + bX$ does not depend on a. This is appropriate: the variance is a measure of spread; adding a does not change the spread, it merely shifts the distribution to the left or to the right.

5.7 Jointly Distributed Random Variables

1. So far, we have only considered individual random variables. Now we turn to properties of several random variables considered at the same time. The outcomes of these different random variables may be related.

Examples

(a) Think of the price of each stock in the Pakistan exchange as a random variable; the movements of these variables are related.

(b) You may be interested in the probability that a randomly selected shopper buys prepared frozen meals. In designing a promotional campaign you might be even more interested in the probability that that same shopper also buys instant coffee and reads a certain magazine.

(c) The number of defects produced by a machine in an hour is a random variable. The number of hours the machine operator has gone without a break is another random variable. You might well be interested in probabilities involving these two random variables together.

2. The probabilities associated with multiple random variables are determined by their **joint distribution.** As with individual random variables, we distinguish discrete and continuous cases.

3. In the discrete case, the distribution is determined by a **joint probability mass function (Probability density Function, pdf)**.

For example, if X and Y are random variables, there joint pmf or pdf is $P_{X,Y}$ $(x, y) = P\ (X = x,\ Y = y) =$ probability that $X = x$ and $Y = y$.

For several random variables $X_1, \ldots, X_n$, we denote the joint pmf by $P(x_1, \ldots, X_n)$

4. It is often convenient to represent a joint pmf through a table. For example, consider a department with a high rate of turnover among employees. Suppose all employees are found to leave within 2-4 years and that all employees hired into this department have 1-3 years of previous work experience. The following table summarizes the joint probabilities of work experience (columns) and years stayed (rows):

	1	2	3
2	.03	.05	.22
3	.05	.06	.15
4	.14	.15	.15

Thus, the proportion of employees that had 1 year prior experience and stayed for 2 years is 0.03. If we let Y = years stayed and X = years' experience, we can express this as

$$P_{X,Y}\ (1, 2) = P\ (X = 1, Y = 2) = 0.03.$$

The table above determines all values of $P_{X,Y}\ (x, y)$.

5. What proportion of employees stay 4 years? What proportion are hired with just 1 year of experience?
These are questions about **marginal probabilities**; i.e., probabilities involving just one of the random variables. A marginal probability for one random variable

is found by adding up over all values of the other random variable; e.g.,

$$P\ (X = x) = \Sigma P\ (X = x,\ Y = y),$$

where the sum ranges over all possible y values. In the table, the marginal probabilities correspond to the column-sums and row-sums. So, the answers to the two questions just posed are 0.44 and 0.22 (the last row-sum and the first column-sum).

6. From a joint distribution we also obtain **conditional distributions**. The conditional distribution of X given $Y = y$ is

$$P_{X|Y}\ (x|y) = P\ (X = x|Y = y) = \ P\ (X = x,\ Y = y)\ /P\ (Y = y)$$

To find a conditional distribution from a table, divide the corresponding row or column by the row-sum or column-sum.

Example: What is the distribution of years stayed among employees with 1 year of experience? Since we are conditioning on 1 year of experience, we only need to consider the first column. Its sum is 0.22. The conditional probabilities are the entries of that column divided by 0.22.

$$P_{Y\ |X}\ (2|1) = 3/22,\ P_{Y\ |X}\ (3|1) = 5/22,\ P_{Y\ |X}\ (4|1) = 14/22.$$

Notice that these conditional probabilities sum to one (as they should), though the original column entries do not. Find the conditional distribution of prior experience among employees that stayed 4 years.

7. A joint distribution determines marginal distributions but the marginal distributions do not determine the joint distribution! (The row-sums and column-sums do not determine the table entries.)

8. Two discrete random variables X and Y are **independent** if their joint distribution is the product of their marginal distributions: $P(X = x, Y = y) = P(X = x)P(Y = y)$ for all x, y. Another way to express this is to say that $P(X = x\ Y = y) = P(X = x)$ for all x and y.

5.8 Covariance and Correlation

1. According to the table above, do employees hired with more years of experience tend to stay more years? This type of relationship between random variables is measured by **covariance** and **correlation**. The covariance between two random variables is

$$Cov[X, Y] = E[(X - \mu_X)(Y - \mu_Y)] = E[XY] - \mu_X\mu_Y .$$

If X tends to be large when Y is large, the covariance will be positive.

2. If two random variables are independent, their covariance is zero. However, the opposite is not (quite) true: two random variables can have zero covariance without being independent.

3. The **correlation coefficient** of X and Y is $\rho_{XY} = \frac{Cov[X,Y]}{\sigma X \sigma Y}$ $Corr[X, Y]$ is the ratio of the covariance to the product of the standard deviations of X and Y. (ρ is the Greek letter rho.)

4. The correlation coefficient has the following properties:

- It is always between -1 and 1.

A positive ρ_{XY} implies that X tends to be large when Y is large and vice-versa. A negative ρ_{XY} implies that X tends to be large when Y is small and vice-versa.

- Correlation measures the strength of **linear dependence** between two random variables. If
$Y = a + bX$ and b $f= 0$, the ρ_{XY} | = 1; its sign positive or negative if b is positive or negative. Conversely, if $|\rho_{XY}| = 1$ then $Y = a + bX$ for some values of ***a and b.***

- Independent random variables have zero correlation.

5. If $Y = X^2$, then the value of X completely determines the value of Y ; however, the correlation is not 1 because the relationship is not linear.

6. Find the covariance and correlation between years of experience and years stayed in the table above.

7. For any random variables X and Y , we have $E[X + Y] = E[X] + E[Y]$, regardless of whether or not X and Y are independent. More generally,

$$E[X_1 + X_2 + \cdots + X_n] = E[X_1] + E[X_2] + \cdots + E[X_n].$$

The variance is a bit more complicated: $Var[X + Y] = Var[X] + Var[Y] + 2Cov[X, Y]$.

More generally, $Var[aX + bY] = a^2Var[X] + b^2Var[Y] + 2abCov[X, Y]$.

In particular (with $a = 1$ and $b = -1$) $Var[X - Y] = Var[X] + Var[Y] - 2Cov[X, Y]$.

If X, Y are independent, then their covariance is zero and $Var[X + Y] = Var[X] + Var[Y]$.

For more than two random variables, we have

$Var[X_1 + \cdots + X_n] = Var[X_1] + \cdots + Var[X_n] + 2Cov[X_1, X_2] + ... + 2Cov[X_1, X_n]$
$+ \cdots + 2Cov[X_{n-1}, X_n]$;

there is a covariance term for each pair of variables. If the variables are independent, then this simplifies to

$$Var[X_1 + \cdots + X_n] = Var[X_1] + \cdots + Var[X_n].$$

If, in addition, $X_1, ... , X_n$ all have variance σ^2, then $V\,ar[X_1 + \cdots + X_n] = (\sigma^2 + \cdots + \sigma^2) = n\sigma^2$
and thus Standard Deviation $[X_1 + \cdots + X_n] = \sqrt{n}\ \sigma$.

Example. A population of interest has four members: Ali, Gulzar, Ibrar and Zeenat. A random experiment selects a sample of size two from the population without replacement. The sample space is:

S = {(Ali, Gulzar), (Ali, Ibrar), (Ali, Zeenat), (Gulzar, Ibrar), (Gulzar, Zeenat), (Ibrar, Zeenat)}.

The event that 'the sample includes Ibrar' is the subset: {(Ali, Ibrar), (Gulzar, Ibrar}, (Ibrar, Zeenat)}. This example shows that the elementary outcomes can themselves be sets.

5.9 Some Rules and Symbols

- **Union.** We write E ∪ F to mean the union of E and F. This set consisting of outcomes that belong to at least one of E or F. » is equivalent to 'either or both' in English. If you look at above example again, throwing a die, you will see that E ∪ F ={3,4,5,6}.

- **Intersection – (E ∩ F or E.F)** We write E ∩ F to mean the intersection of E and F. This set consisting of outcomes belonging to E and F. ∩ is equivalent to and in English. Returning to Example, E ∩ F = {4}.

• **Complement.** We write the complement of E as E^c . It indicates all the elements of a set not in event E. Looking at Example again, throwing a die, you can see that E^c is = {1,2,5,6}.

Discrete Random Variable: A numerical r.v. that takes on a countable number of values (there are gaps in the range of possible values).

Examples:

1. Number of phone calls received in a day by a company
2. Number of heads in 5 tosses of a coin

Continuous Random Variable : A numerical r.v. that takes on an uncountable number of values (possible values lie in an unbroken interval).

Examples: 1. Length of nails produced at a factory
2. Time in 100-meter dash for runners

If X is a random variable, the set of outcomes on which X takes a particular value (or range of values) is a subset of the sample space, which is to say, it is an event.

Thus, if we have a probability distribution on the sample space, we may therefore ask about quantities like

(i) $P(X = n)$, the probability that X takes the value n, or

(ii) $P(X \geq 5)$, the probability that the value of X is at least 5, or

(iii) $P(2 < X < 4)$, the probability that the value of X is strictly between 2 and 4.

A common way to tabulate all of this information is to make a list or table of all the possible values of X along with their corresponding probabilities. The associated function is called the probability density function of X:

Definition: If X is a random variable on the sample space S, then the function p(X) such that $P(X \in E)$ for any set of numbers E is called the probability density function (pdf) of X.

Explicitly, the value of p(a) on a real number a is the probability that the random variable X takes the value a.

For discrete random variables with a small number of outcomes, we usually describe the probability density function using a table of values. In certain situations, we can find a convenient formula for the values of the probability density function on arbitrary events, but in many other cases, the best we can do is simply to tabulate all the different values.

Example: If two standard 6-sided dice are rolled, find the probability distribution for the random variable X giving the sum of the outcomes. Then calculate (i) P(X=7), (ii) P(4< X<9), and (iii) P(X≤ 6).

To find the probability distribution for X, we identify all of the possible values for X and then tabulate the respective outcomes in which each value occurs.

We can see that the possible values for X are 2, 3, 4, ... , 12, and that they occur as follows:

Value (X)	Outcomes	Probabilities
2	(1, 1)	1/36
3	(1, 2), (2, 1)	2/36
4	(1, 3), (2, 2), (3, 1)	3/36
5	(1, 4), (2, 3), (3, 2), (4, 1)	4/36
6	(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)	5/36
7	1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)	6/36
8	(2, 6), (3, 5), (4, 4), (5, 3), (6, 2)	5/36
9	(3, 6), (4, 5), (5, 4), (6, 3)	4/36
10	(4, 6), (5, 5), (6, 4)	3/36
11	(5, 6), (6, 5)	2/36
12	(6, 6)	1/36

Then we have P(X = 7) = 6/ 36 = 1 /6 , P(4 < X < 9) = 4/ 36 + 5/ 36 + 6/ 36 + 5/ 36 = 5/ 9 , and
P(X ≤ 6) = 1 /36 + 2/ 36 + 3/ 36 + 4/ 36 + 5/ 36 = 15/36 =5/ 12

Example: If a fair coin is flipped 4 times, find the probability distributions for the random variable X giving the number of total heads obtained, and for the random variable Y giving the longest run of consecutive tails obtained. Then calculate (i) P(X = 2), (ii) P(X ≥ 3), (iii) P(1 < X < 4), (iv) P(Y = 1), (v) P(Y ≤ 3), and (vi) P(X = Y = 2).

For X, we obtain the following distribution:

Value (X)	Outcomes	Probability
0	(T T T T)	1/16
1	(T T T H), (T T HT), (T HT T), (HT T T)	1/4
2	(T T HH), (T HT H), (T HHT), (HT T H), (HT HT), (HHT T)	3/8
3	(T HHH), (HT HH), (HHT H), (HHHT)	1/4
4	(HHHH)	1/16

For Y , we obtain the following distribution: Value Outcomes Probability

Value (Y)	Outcomes	Probability
0	(HHHH)	1/16
1	(THTH), (THHT),(THHH),(HTHT),(HTHH), (HHT H), (HHHT)	7/16
2	(TTHT), (TTHH), (THTT), (HTT H), (HHTT)	5/16
3	(TTTH),(HTTT)	1/8
4	(TTTT)	1/16

We can then quickly compute
$P(X = 2) = 3 /8$, $P(X \geq 3) = 1 /4 + 1/16 = 5/16$, $P(1 < X < 4) = 3/8 + 1/4 = 5/8$,
$P(Y = 1) = 7/16$
and $P(Y \leq 3) = 1/16 + 7/16 + 5/16 + 1/8 = 15/16$.

To and $P(X = Y = 2)$ we must look at the individual outcomes where X and Y are both equal to 2. There are 2 such outcomes, namely (TTHH) and (HHTT), so $P(X = Y = 2) = 1/8$.

If we have a random variable X defined on the sample space, then since X is a function on outcomes, we can define various new random variables in terms of X.

If g is any real-valued function, we can define a new random variable g(X) by evaluating g on all of the results of X. Some possibilities include $g(X) = 2X$, which doubles every value of X, or
$g(X) = X^2$, which squares every value of X.

More generally, if we have a collection of random variables $X_1, X_2, \ldots , Xn$ defined on the same sample space, we can construct new functions in terms of them, such as the sum $X_1 +X_2 +\cdot \cdot \cdot+Xn$ that returns the sum of the values of $X_1, X_2 \ldots , Xn$ on any given outcome.

A particular random variable is the random variable identifying whether an event has occurred:

Definition: If E is any event, we dene the Bernoulli random variable for E to be X, E = (1 if E occurs 0 if E does not occur .

The name for this random variable comes from the idea of a Bernoulli trial, which is an experiment having only two possible outcomes, success (with probability p) and failure (with probability 1 − p). We think of E as being the event of success, while E^c is the event of failure.

Many experiments consist of a sequence of independent Bernoulli trials, in which the outcome of each trial is independent from the outcomes of all of the others. For example, flipping a coin 10 times and testing whether heads is obtained for each flip is an example of a Bernoulli trial.

Using our results on independence of events, we can describe explicitly the probability distribution of the random variable X giving the total number of successes when n independent Bernoulli trials are performed, each with a probability p of success.

Example: (Roll a die). The random variable *X* = number of dots showing.

x	1	2	3	4	5	6
P(*x*)	1/6	1/6	1/6	1/6	1/6	1/6

Example: (Toss 2 coins). The r.v. *X* = number of heads showing.

x	0	1	2
P(*x*)	¼	½	¼

For <u>any</u> probability distribution:

(1) P(*x*) is between 0 and 1 for any value of *x*.

(2) $\sum_x P(x)$ = 1. That is, the sum of the probabilities for all possible *x* values is

5.10 Rules of Counting

The Fundamental Counting Principle, sometimes referred to as the fundamental counting rule, is a way to figure out the number of possible outcomes for a given situation.

While there are five basic counting principles: addition, multiplication, Permutation and Combination. The one that is most closely associated with the title of “fundamental counting principle” is the multiplication rule, where if there are p ways to do one task and q ways to another task, then there are pxq ways to do both.

When selecting elements of a set, the number of possible outcomes depends on the conditions under which the selection has taken place.

Some times counting the "number of ways an Event E can occur" or the "total number of possible outcomes" can be fairly complicated. In this section, we'll learn several counting techniques, which will help us calculate some of the more complicated probabilities.

Addition Principle

The Sum Rule states that if a task can be performed in two ways, where the two methods cannot be performed simultaneously, then completing the job can be done by the sum of the ways to perform the task.
Example: if an experiment can proceed in one of two ways, with experiment-I have n_1 outcomes for the first way, and Experiment II have n_2 outcomes for the second, then the total number of outcomes for the experiment is $n_1 + n_2$

Sum rule 2: if an experiment can proceed in one of m ways, with Experiment-I n_1 outcomes for the first way, Experiment-II have n_2 outcomes for the second, . . ., and Experiment-n have n_m outcomes for the ^{m}th, then the total number of outcomes for the experiment is $n_1 + n_2 + . . . + n_m$

Example

For instance, suppose a bakery has a selection of 20 different cupcakes, 10 different donuts, and 15 different muffins. If you are to select a tasty treat, how many different choices of sweets can you choose from?

Because we have to choose from either a **cupcake** or **donut** or **muffin** (notice the “OR”), we have **20** + **10** + **15** = 45 treats to choose from.

Multiplication Principle

The **Product Rule** states that if a task can be performed in a **sequence of tasks**, one after the other, then completing the job can be done by the **product of the ways** to perform the task.

Example

Continuing our story from above, suppose a bakery has a selection of 20 different cupcakes, 10 different donuts, and 15 different muffins — how many different orders are there?

Solution

What makes this question different from the first problem is that we are **not** asking how many total choices there are. We are asking how many different ways we can select a treat.
It's possible that you only want one treat, but you can quite easily want more than one.
So how many different orders can you create, if you're allowed to choose as few or as many as you like?

This is the job for the product rule!
Because we can choose treats from a selection of **cupcakes** and **donuts** and **muffins** (notice the "AND"), we **20** x **10** x **15** = 3,000 ordering options.

Example

Now let's look at another example. Suppose a mathematics faculty and 83 mathematics majors, and no one is both a faculty member and a student.

Solution: By the sum rule, it follows that there are 37 + 83 = 120 possible ways to pick a representative.

Remember, the product rule states that if there are p ways to do one task and q ways to another task, then there are p x q ways to do both.

Example

A restaurant menu offers 4 starters, 7 main courses and 3 different desserts. How many different three-course meals can be selected from the menu?

Solution:

Multiplying together the number of choices for each course gives 4×7×3=84 different three-course meals.

Permutation and Combination

Both combination and permutation are concerned with the number of ways of selecting and arranging of objects. Combination is simply concerned with selection while permutation is concerned with arrangement. There is therefore a slight difference between the two.

Combinations

The term, combination refers to the number of ways of selecting objects from a group of objects at a time without considering the order in which they are selected. In other words, the combination of *n* different items taking *r* objects at a time is the selection of *r* out of the *n* objects with no attention paid to the order of selection. The number of possible combinations of *n* objects taking *r* at a time is denoted by ^nC_r and is expanded as follows:

$$^nC_r = \frac{n!}{(n-r)!r!}$$

[n! = n(n-1)(n-2)(n-3) ---(1); e.g. 5! = $5 \times 4 \times 3 \times 2 \times 1$ = 120]

For example, consider the selection of two numbers at a time from the set {1, 2, 3, 4, 5}. The possible selections by combination are:

(1, 2) (1, 3) (1, 4) (1, 5) (2, 3) (2, 4) (2, 5) (3, 4) (3, 5) and (4, 5). Ten possible selections are therefore made.

The total number of objects in the set = 5
Number of items selected at a time = 2
: . Number of ways of making the selection is given by:

$$^5C_2 = \frac{5!}{(5-2)!2!} = \frac{5!}{3!2!} = \frac{5\times4\times3\times2\times1}{3\times2\times1\times2\times1} = 10 \text{ ways.}$$

Ten possible ways of arrangement can be made as we can see from the illustration above.

Another example can be chosen from a lottery in which out of all the numbers from 1 to 90, five are selected as the winning numbers for the National lottery. The selection of the five numbers out of the ninety is by combinational arrangements since the other in which the winning numbers are picked is not necessary. The number of possible arrangements in this case is given by:

$$^{90}C_5 = \frac{90!}{(90-5)!5!} = \frac{90!}{85!5!} = 43{,}949{,}268 \text{ ways.}$$

Hence, the chance for one winning the lottery is too low since 43,949,268 different sets of five winning numbers can be selected.

Permutations

The term, ***permutation***, on the other hand, refers to the number of ways of arranging objects from a group of objects at a time with attention given to the order of arrangement. In order words, a permutation of *n* objects taking *r* at a time is number of arrangement of *r* objects out of the *n* objects with attention paid to the order of arrangements. Thus, if *n* is the total number of objects in the group and *r* is to be selected at a time taking into consideration the order of arrangement, the possible number of ways is given by:

$$^{n}P_r = \frac{n!}{(n-r)!}$$

For example, consider the selection of two numbers at a time from the set {1, 2, 3, 4, 5}. The possible selections by permutation are:

(1, 2) (2, 1) (1, 3) (3, 1) (1, 4) (4, 1) (1, 5) (5, 1) (2, 3) (3, 2) (2, 4) (4, 2) (2,5) (5, 2) (3, 4) (4, 3) (3, 5) (5, 3) (4, 5) and (5, 4). Twenty possible selections are therefore made.

The total number of objects in the set =5; Number of items selected at a time=2
: . Number of ways of making the selection is given by:

$$^{5}P_2 = \frac{5!}{(5-2)!} = \frac{5!}{3!} = \frac{5\times 4\times 3\times 2\times 1}{3\times 2\times 1} = 20 \text{ ways.}$$

Twenty possible ways of arrangement can be made as we can see from the illustration above.

The Concept of Exclusion and Inclusion in Combinations

The concept of exclusion and inclusion are cardinal importance in combinations. Let us tackle the two in turn.

Exclusion

If some objects are to be selected by combinational means in such a way that some particular objects are to be excluded, the number to be excluded should be deducted from the total. The experiment is then conducted on the remaining objects.

For example, assuming there are 8 boys in a class from which a committee of 3 boys is to be formed. The number of ways of forming the committee so that 2 particular boys are *excluded* can be determined as follows:

Total number of boys = 8
To number for experiment = 8 – 2 = 6

: . Number of ways the selection can be made is:

$$^{8-2}C_3 = ^6C_3 = \frac{6!}{(6-3)!3!} = \frac{6!}{3!3!} = 20 \text{ ways}$$

Inclusion

For an object or objects to be included, it has to affect both the total number of objects and the number of objects to be selected. The number to be included should be deducted from the total, and also from the number to be selected. Then, the number of combinations of the remaining objects gives the number of ways required.

For example, the number of ways of a committee of 4 girls can be formed from a grouped of 10 girls if:

(a) One particular girl is to be included in the committee;
(b) Two particular girls are to be included in the committee can be determined as follows:

Solution:

(a) Total number of girls = 10
Number of girls for experiment = 10–1 = 9; Number of girls to be selected= 4–1= 3

: . Number of ways $= {}^{9}C_3 = \frac{9!}{(9-3)!3!} = \frac{9!}{6!3!} = 84$ ways

(c) Number of ways $= {}^{10-2}C_{4-2} = {}^{8}C_2 = \frac{8!}{(8-2)!2!} = \frac{8!}{6!2!} = 56$ ways

Example. There are 5 boys and 8 girls in a club. A committee of 5 is to be formed. Find the number of ways of forming the committee if

(a) No consideration is given to sex, (b) Two boys and three girls should be on the committee

Solution

(a) Number of boys = 5; number of girls = 8; Total = 5 + 8 = 13

Since no consideration is given to sex, anybody at all in the group can be selected.

: . Number of ways $= {}^{13}C_5 = \frac{13!}{(13-5)!5!} = 1{,}287$ ways

(b) Out of 5 boys 2 are to be selected and out of 8 girls 3 are to be selected

: . Number of ways $= {}^{5}C_2 \times {}^{8}C_3 = 10 \times 56 = 560$ ways

Acitvity: A committee of 4 men and 3 women is to be formed from 10 men and 8 women so that one particular man and two particular women are excluded. Find the number of ways the committee can be formed.

Activity. There are 12 men and 15 women in an association. A committee of 3 men and 4 women is to be formed. Find the number of ways of forming the committee if

(a) One particular man and one particular woman are to be included

(b) Two particular men are to be excluded and one particular woman is to be included

Example. Out of 5 union members and 7 non-union members, a standing committee consisting of 2 union members and 3 non-union members is to be formed by a company. How many different ways can the committee be constituted if one particular member is to be excluded from the committee?

Solution

Either one is to be excluded from union members or one to be excluded from the non-union members. Therefore, number of ways the committee can be constituted

$$\left({}^{4}C_{2}\times{}^{7}C_{3}\right)or\left({}^{5}C_{2}\times{}^{6}C_{3}\right)=(6\times 35)+(10\times 20)=410\ ways$$

Example. A committee of 5 members is to be formed from a teaching staff of 7 men and 5 women.

Find the number of ways of (a) forming the committee (b) including only men (c) including at least one man (d) including 2 women.

Solution

Total number of people = 7 + 5 = 12

(a) Since no condition is given as to what number of men or women, we have to treat them as one group. Thus, number of ways = ${}^{12}C_{5}=792$ ways

(b) Number of ways of including only men = $\left({}^{7}C_{5}\times{}^{5}C_{0}\right)={}^{7}C_{5}=21\ ways$

(c) Number of ways of including at least one man is given by:

(Total no. of possible ways) – (No. of ways of selecting no man) = $\left({}^{12}C_{5}-{}^{5}C_{5}\right)=792-1=791\ ways$

(d) Number of ways of including two men means we select two men and three women (i.e. to make up the total (5) to be selected).

Therefore, the number of ways = $\left({}^{7}C_{2}\times{}^{5}C_{3}\right)=210\ ways$.

Activity. A student must answer 4 out of 7 questions in an examination.

(a) How many choices does he have?

(b) If he must answer the first two questions, how many choices does he have?

Example. There are 6 men and 9 women in a club. A committee of 5 is to be formed. Find the number of ways of selecting at least one woman.

Solution

Number of ways = (Total number of all possible ways) – (number of ways of excluding all women)

Total number of all possible ways = ${}^{6+9}C_{5}={}^{15}C_{5}=3003$

Number of ways of excluding all women = ${}^{6}C_{5}=6$

Therefore, number of ways of selecting at least one woman = 3003 – 6 = 2997 ways.

Examples on Permutations

Example. In how many ways can 6 marbles coloured differently be arranged in a row?

Solution

Since the marbles are coloured differently, the order of arrangement is important. Therefore the number of ways = ${}^{6}P_{6} = 6! = 720\, ways$.

Activity: In how many ways can 8 people be seated on a bench if only 3 seats are available?

Example: Six men and five women are to be seated in a row so that women occupy the even places. How many such arrangements are possible?

Solution

Number of seating arrangement of men = ${}^{6}P_{6}$

Number of seating arrangement of women = ${}^{5}P_{5}$

Therefore number of arrangements = ${}^{6}P_{6} \times {}^{5}P_{5}$ = 720 × 120 = 86,400 arrangements

Example: In how many ways can 5 people be seated at a round table if (a) They sit anywhere, (b) Two particular people must sit together, (c) Two particular people must not sit together

Solution

(a) Since they are to sit around a table, one of them should be made fixed.
Thus the number of ways is given by: ${}^{5-1}P_{5-1} = {}^{4}P_{4} = 24\, ways$

(b) The two particular people to be seated together should be considered as one person so that there would apparently be 4 people altogether and they can be arranged in ${}^{4-1}P_{4-1} \times 2! = {}^{3}P_{3} \times 2 = 12\, ways.$

(c) Number of ways of arranging 5 people at a round table so that 2 people do not sit together is
24 – 12 = 12 ways.

Activity: Six different Mathematics books, three different English books, and four different

Literature books are arranged on a shelf. How many different arrangements are possible if (a) The books on each particular subject must all stand together

(b) Only the Mathematics books should stand together

Example: A manager is to give three productivity awards to employees in the three sections of his department. If the total number of employees is 20 and no one person should win more than one prize, in how many ways can the awards be won?

Solution

Because no one person should win more than one, the order of arrangement is important

: . Number of ways = ${}^{20}P_3 = \frac{20!}{(20-3)!} = 6{,}840 \ \textit{ways}$

Example: In how many ways can the numbers 5, 6, 7, 8 and 9 be arranged?

Solution

Number of given numbers = 5.

Since all the numbers are different, the number of ways = ${}^{5}P_5 = 5! = 120 \ \textit{ways}$

Example: In how many ways can the numbers 5, 6, 6, 7, 7, 7, 8 and 9 be arranged?

Solution

The given numbers are eight in number. Since 6's are two and 7's are three and the remaining numbers do not repeat themselves, the number of ways = $\frac{8!}{2! \times 3!} = 3{,}360 \ \textit{ways}$

Example: In how many ways can the number 3, 4, 5, 6, 7 and 8 be arranged so that:

(a) 6, 7 and 8 should be together

(b) 3 and 5 are separated

Solution

(a) We consider the three numbers 6, 7, 8 as one number to give n=4. Number of items

(6, 7, 8) to be put together, *x = 3*

Number of ways for 6, 7 and 8 to be to be together = $3! \times \frac{4!}{(3-1)!} = 6 \times \frac{4!}{2!} = 36\ ways$

(b) 3 and 5 are two numbers. We need to find the total possible arrangements and also the

number of arrangements when the two numbers are together. We then subtract the latter

from the former to get the expected answer.

Total number of possible arrangements = n! = 6! = 720 ways

Number of ways when two numbers are together = $2! \times \frac{5!}{(2-1)!} = 2 \times \frac{5!}{1!} = 240\ ways$

Activity. Find the number of ways by which the letters in the word "STATISTICS" be arranged?

Activity: Given that $6\left({}^{7}C_{n}\right) = {}^{7}P_{n}$, find the value of n.

5.11 SELF ASSESSMENT QUESTIONS

1. On your route to work, there are two traffic lights. You are 20% likely to be stopped at the first and 40% likely to be stopped at the second.

a) USE THE DEFINITION OF EXPECTED VALUE to compute the expected number of traffic light stops you'll make on your way to work. Interpret this number in English. You may assume that the lights are not synchronized with one another in any way

b) Use the LAWS OF RANDOM VARIABLES to verify that the answer that you obtained in part a was correct.

2. How do we compute the probability of an event using the relative frequency approach?

3. Let **R** = the person interviewed is Republican and **F** = the person interviewed is female. Then how would one express the following events in terms of our probability notation? a) How likely is it the person interviewed is a male Republican? b) How likely is it that a woman we interview is a Republican? c) What fraction of Republicans are women?

4. Give an example of two independent random variables. Give an example of two dependent random variables. Under what circumstances does E(**A**) + E(**B**) = E(**A + B)**?

5. You may hear a statistic like "30% of all highway fatalities involve drunk drivers." From a statistical point of view, why is this the wrong statistic upon which to base a MADD (Mothers Against Drunk Drivers) lobbying effort? What probability involving the same events would be relevant? Hint: Compare to the statistic, "Over 50% of all highway fatalities involve male drivers.

6. We choose a number at random from 1 to 10. Let $\mathbf{D_i}$ = the number is evenly divisible by **i**, so $\mathbf{D_2}$ = the number is even, etc. Let **X** = the number selected.

a) Is $\mathbf{D_i}$ an event or a random variable? Is X an event or a random variable?

b) Find P($\mathbf{D_3}$), P($\mathbf{D_6}$), P($\mathbf{D_3} \mid \mathbf{D_6}$), P($\mathbf{D_6} \mid \mathbf{D_3}$), and P($\mathbf{D_5} \mid \mathbf{D_7}$).

c) Find P(**X =** 6), P(**X** = 6 | $\mathbf{D_3}$), P(**X** = 10 | X > 8), and P($\mathbf{D_4}$ **| X < 4**).

7. We write the number "1" one head of a coin and the number "–1" on the tail. We then flip the coin. Let **N** = the number appearing on the top of the coin, and **B** be the number on the bottom of the coin. Find E(**N**), E(**B**), E(**N** + **B**), and E(**NB**). Interpret each.

8. A standard six sided die is made so that the opposite faces always add to seven. Hence, the “1” face is always opposite the “6” face, and so on. Let **T** = the number that appears on the top face of a die that we roll, and **B** = the number appearing on the bottom face.

a) What does **T** + **B** mean? Is it correct to write **T** + **B** = 7?
b) Find E(**T** + **B**), E(**T**) and E(**B**).
c) Find E(**T** □ **B**). Does it equal E(**T**) E(**B**)?
d) Suppose our six sided die were made in a nonstandard way. The faces are still labeled with the numbers from 1 to 6, but opposite faces no longer necessarily add to 7. Answer questions a) – c) in this case, if you can.

9. The chairman of board of directors of a company wishes to set up a small finance committee comprising three directors from a group of seven. Calculate the possible alternative ways in which he can make this selection.

10. A standing committee of 4 males and 5 females is to be formed from 7 males and 8 females. How many different ways can this committee be formed?

11. There are 10 women in a mini market. A committee of 4 women is to be formed.
Find the number of ways if?
(a) one particular woman is to be excluded, (b) two particular women are to be included.

12. A committee of 4 boys and 3 girls is to be formed from a group of 9 boys and 7 girls.
Find the number of ways of forming the committee if?

(a) One particular boy and one particular girl are to be included.
(b) Three particular boys are to be included and two particular girls are to be excluded.

13. In how many ways can 3 prizes be awarded to a class of 10 boys, one for English, one for

Mathematics and for French if? (a) No boy should win more than one prize (b) there is no condition.

14. In how many ways can 10 story books be arranged on a straight shelf?

15. Find the number of arrangement of 6 pebbles coloured differently around a circle.

16. There are 5 different mathematics books and 3 different English book on a shelf.
Find the number of ways the arrangement can be made if
(a) the books on each particular subject must stand together. (b) the books should stand anyhow.

17. In how many ways can 10 girls be seated on a bench if only 5 seats are available?

18. If ${}^{n}C_{2} = 21$, find the value of n.

19. Given that $24\left({}^{8}C_{n}\right) = {}^{8}P_{n}$, find the value of n.

SUGGESTED READINGS

Bluman, A.G. (2004). Elementary Statistics. A Step by Step Approach. 5th Edition. McGraw-Hill Companies Incorporated. London.

Chaudhary, S.M. & Kamal, S. (2017). Introduction to Statistical Theory Part-I. 8th Edition. Ilmi Kitab Khana. Lahore.

Chaudhary, S.M. & Kamal, S. (2017). Introduction to Statistical Theory Part-II. 8th Edition. Ilmi Kitab Khana. Lahore.

Daniel, W.W. (1995). Biostatistics: A foundation for Analysis in Health sciences. Sixth Edition. John Wiley and sons Incorporated. USA.

Harper, W.M. (1991). Statistics. Sixth Edition. Pitman Publishing, Longman Group, United Kingdom.

Hoel, P.G. (1976). Elementary Statistics. 4th Edition. John Wiley and Sons Incorporated, NewYork.

Kiani, G. H., & Akhtar, M. S. (2012). Basic statistics, Majeed Book Depot.

Khan, A. A., Mirza, S. H., Ahmad, M. I., Baig, I. & Yaqoob, M. (2011). Business Statistics, Qureshi Brothers Publishers.

UNIT 06

PROBABILITY

Written By: **Dr. Zahid Iqbal**
Reviewed By: **Dr. Muhammad Ilyas**

CONTENTS

Pages

Introduction

Probability formulas and technique developed by Jacob Bernoulli (1654- 1705), Reverend Thomas Bays (1702- 1761), Abraham de Moivre (1667- 1754) and Joseph Lagrange (1736- 1813). In the nineteenth century Pierre Simon and Marquis de Laplace gather all these early ideas and compiled the first general theory of probability. Probability theory is a part of our daily life. In many personal and managerial decision we face uncertainty and ultimately use probability theory like weather forecasting, sale forecasting and so on.

Objectives

After studying this unit, you will be able to.

- Define experiment, outcome, event, probability and equally likely.
- Restate the formula for finding the probability of an event.
- Determine the outcomes and probabilities for experiments.
- Interact with die rolls and spinners to help predict the outcome of experiments.
- Distinguish between an event and an outcome for an experiment.
- Recognize the difference between outcomes that are equally likely and not equally likely to occur.
- Apply probability concepts.

6.1 Sets

The concept of sets is very useful in statistics because it is one of the basics of understanding the principle of probability, the subject of the next chapter which is a vital topic in statistics.
A set is a well-defined collection of objects. Any group of objects of the same kind can be considered as set. The object in a set are called *elements or members* of the set and may be anything whatsoever. We may have a set of goats, a set of cars, a set of tables, or even a set of sets sometimes called a *class* of sets. A set is usually denoted by a capital letter and an element is represented by a small letter. Thus, if a is an element of set *A*, then we write;

$$a \in A.$$

If a is not an element of A, we write;

$$a \notin A.$$

A set is specified by the content of two braces or curly brackets: { }. There are two methods for specifying the content of a set. These are:

(i) The ***Tabular Method*** in which case the elements are enumerated explicitly. For example, the set of all even numbers between *1* to *10* will be: {2, 4, 6, 8}

(ii) The ***Rule Method*** in which case the content of a set is determined by some rule, such as: {even numbers between 1 and 10}. The rule method is usually more convenient to use when the set is large. For example, it would be tedious to write explicitly using the tabular method for the set: {even numbers between 1 and 10, 000}.

Countable, Uncountable and Empty Sets

A set is said to be *countable* if its elements can be put in one-to-one correspondence with the natural numbers, which are the positive integers, 1, 2, 3, etc. In other words, elements in a countable set can be enumerated.

A set is said to be *uncountable* if the elements in it cannot be counted. For example, a set of colours: i.e. {colours}.

An *empty* set is a set which has no element(s).

The empty set is represented with a symbol ϕ or { }. It is often called a *null set*.

Finite, Infinite and Countably Infinite Sets

A set is said to be a *finite* set if it is either empty or has elements which can be counted, with the counting process starting and ending at certain stages; that is, the set has a finite or definite number of elements.
On the other hand, an *infinite* set is one whose elements are not finite. An infinite set having countable elements is known as a *countably infinite* set. For example, a set of all integers or {integers}.

Some Important Symbols for Mathematical Operations

The following symbols should be well noted in studying sets.

$\in$ ≡ “is a member of”
$\subset$ ≡ “is a subset of”
$\cap$ ≡ “is intersection of”

$\notin$ ≡ “is not a member of”
$\supset$ ≡ “is a set containing the set
$\cup$ ≡ “is union of”

Subsets

The set ***A***, is said to be a subset of another set ***B***, if all the members of ***A*** are also members of ***B*** [i.e. $A \subseteq B$ or $B \supseteq A$]. In other words, ***A*** is said to be contained in ***B***. If at least one element exists in a set ***B*** which is not in set ***A***, we say ***A*** is a proper set of ***B***.

For example if ***A*** = {1, 3, 5} and ***B*** = {1,2, 3, 4, 5, 6}, then since 1, 3 and 5 are all contained in the set ***B*** we can say ***A*** is a subset of ***B***. [i.e. **$A \subset B$]**. The null set is a subset of all other sets.

Disjoint or Mutually Exclusive Sets

If two sets, ***A*** and ***B***, have no common set or elements at all, they are called disjoint or mutually exclusive sets. For example, if **A** = {1, 3, 5, 7} and **B** = {2, 4, 6, 8, 10}, then ***A*** and ***B*** are disjoint sets.

The Universal Set (U or ε)

The *universal set*, also known as the *entity set*, is the largest possible set containing all the members in any experiment. In other words, it contains all the possible subsets. For example, the sets {natural numbers} and {integers} can be considered as universal sets.

Mathematical Operations on Sets.

Various operations are carried out in sets. These operations are explained below.

Complement sets

If ***A*** is a set, then the complement of ***A***, written as $\boldsymbol{A^1}$, is the set containing all the other elements in the universal set which are not found in the set *A*. For example, if U = {1, 2, 3, ---, 10} and A = {1, 2, 3, 5, 7}, then $\boldsymbol{A^1}$ is given by; $\boldsymbol{A^1}$ = {4, 6, 8, 9, 10}

***The intersection of Sets* (∩) [i.e. *A cap B*]**

The intersection of two sets ***A*** and ***B***, is the set containing the common elements of ***A*** and ***B***. It means the set that contains the elements which can be seen in both ***A*** and ***B***. For example, if ***A***={1, 2, 3, 5. 7} and ***B*** = {2, 5, 7, 8, 9}, then ***A* ∩ *B* =** {2, 5, 7}.

***The union of Sets* (∪)**

The union of sets is the set whose elements include the elements of all sets under consideration. Thus, if ***A*** = {1, 2, 3}, ***B*** = {2, 3, 4, 7} and ***C*** = {6, 8, 9}, then:

***A* ∪ *B* =** {1, 2, 3, 4, 7}; B ∪ **C =** {2, 3, 4, 6, 7, 8, 9};
***A* ∪ *C* =** {1, 2, 3, 6, 8, 9} and ***A* ∪ *B* ∪ *C* =** {1, 2, 3, 4, 6, 7, 8, 9}.

Venn Diagram

In working with sets, it is useful to introduce a geometrical representation that enables us to associate a physical picture with sets.

A Venn Diagram is a diagrammatic representation of sets including the universal set inside which all the available subsets are appropriately drawn.

Two-Set Problems

If two sets, ***A*** and ***B,*** intersect, the following relation should hold.

$$n(A \cup B) = n(A) + n(B) – n(A \cap B)$$

Where n(A) means the number of members in Set A, n(B) means the number of members in Set B.

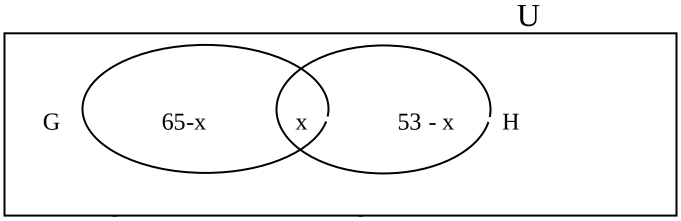
It must be noted that n(A $\cap$ B) is subtracted from the sum of members in A and B, because the intersection region is added twice.

On the other hand, if the two sets are disjoints or mutually exclusive (i.e. do not intersect), the relation reduces to: n(A $\cup$ B) = n(A) + n(B)

For example, assume that there are 100 students in a school who are going to take Geography (G) and History (H) examinations. If it is found that 65 students are to take Geography whilst 53 are to take History, the number taking both papers can be found as follows:

n(G $\cup$ H) = n(G) + n(H) – n(G $\cap$ H), 100 = 65 + 53 - n(G $\cap$ H) , n(G $\cap$ H) = 118 – 100 = 18

Solve the Problem Given in Diagram: we can solve the problem using a Venn diagram as shown in Figure 6.1 below in which x represents the number of students taking both Geography and History.



Since the total number of students is 100, we can have:

$$100 = (65 - x) + x + (53 - x) \Rightarrow x = 18.$$

As a further example, assume that in a class, the number of students studying French or History is 40. Twenty study both subjects and the number of students who study French is 10 more than the number of students who study History. Let us calculate:

(a) The number of students studying French
(b) The number of students studying History as follows:

The solution is given as follows:

(a) Let F = {Students studying French}, H = {Students studying History}, x = number who study French.

n(F $\cup$ H) = 40; n(F) = x ; n(H) = x - 10; n(F $\cap$ H) = 20 ; n(F $\cup$ H) = n(F) + n(H) – n(F $\cap$ H)

$40 = x + (x - 10) - 20 \Rightarrow 2x = 70 \Rightarrow x = 35.$

Therefore the number studying French is 35.

n(H) = 35 – 10 = 25. Therefore the number studying French is 25.

Solve the Venn Diagram Problem: we can solve the problem above using a Venn diagram as shown in Figure 6.2 below.

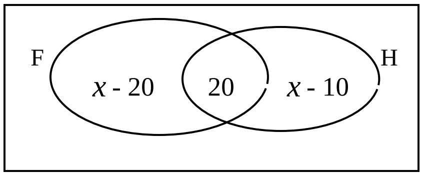


Figure 6.2: Venn diagram

$x - 20 + 20 + x - 10 = 40$. $2x = 70 \Rightarrow x = 35.$

Therefore the number studying French is 35

Finally, let us assume that in a sports contingent, there are 40 players in the football team and 36 players in the volleyball team. Eight players play both football and volleyball. Let us find:

(a) The number of players in the contingent
(b) The number who play only football or only volleyball

Let F = {Football team} and V = {Volleyball team}
$n(F) = 40$; $n(V) = 36$; $n(F \cap V) = 8$

(a) $n(F \cup H) = n(F) + n(H) - n(F \cap H) = 40 + 36 - 8 = 68$

(b) Number who play only football or only volleyball is (40 – 8) + (36 – 8) = 60

Three-Set Problems

Let A, B and C be any three intersecting sets. The following relation can be obtained if at least one of any event is to be achieved.

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

It must be noted that when the members of the three sets are exclusively added, the intersections of the members of any two sets are added twice and therefore one each has to be subtracted. When the subtractions are done, n($A \cap B \cap C$) is subtracted thrice while addition of that part has been added twice and hence, has to be added once.

If the three sets are disjoint, then

$$n(A \cup B \cup C) = n(A) + n(B) + n(C)$$

Note that no regions of intersections are encountered here.

Let us illustrate the concept with the following sample problem:

In a group of 300 traders, 210 sell Wheat, 195 sell Maize and 180 sell rice. Ninety sell both Wheat and Maize, 100 sell Wheat and Rice, and 115 sell both Rice and Maize. If each trader sells at least one of the three items, the number of traders who sell all three items can be derived as follows:
Let U = {Traders}; G = {Wheat sellers}; R = {Rice sellers}; M = {Maize sellers} and let x be number of traders selling all the three items.

$$n(G \cup R \cup M) = n(G) + n(R) + n(M) - n(G \cap R) - n(G \cap M) - n(R \cap M) + n(G \cap R \cap M)$$
$$300 = 210 + 180 + 195 - 100 - 90 - 115 + x \Rightarrow x = 20.$$

EXAMPLES

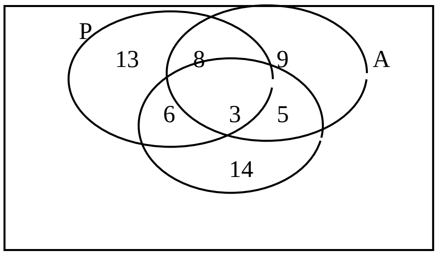
Example:. A company has a large number of typists. A survey shows that 30 can use a word processor, 25 are audio-typists and 28 are short-hand writers. Of the typists who are short-hand writers, 3 are audio-typists, and can use word processor, 5 are audio-typists but cannot use a word processor, 6 can use a word processor but are not audio typists. Eight can use word processor and are audio-typists but are not short hand typists.

(a) Present this information on a Venn diagram.
(b) How many typists were involved in the survey?
(c) How many typists have only one skill?

Solution

Let P = {Word processor typists}; A = {Audio-typists}; S = {Short-hand typists}

Using the Venn diagram below we can solve the problem as follows:



(b) The total number of typists involved = 13 + 8 + 3 + 6 + 9 + 5 + 14 = 58
(c) Number of typists with only one skill = 13 + 9 + 14 = 36

Adding all members in the various regions and solving for x gives the value of x as 55.
The number of students who passed

(a) all the three subjects = x = 55
(b) exactly one subject = (275 + x) + (205 + x) + (20+ x) = 330 + 260 + 75 = 665
(c) exactly two subjects = (175 - x) + (150 - x) + (120 - x) = 120 + 95 + 65 = 280

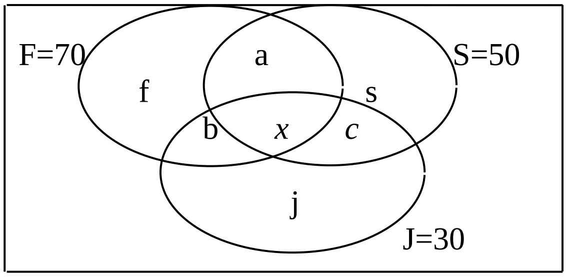
Example:**.** In a survey of the 100 out-patients who reported at a hospital one day, it was found that 70 complained of fever, 50 had stomach trouble and 30 were injured. Each of the 100 out-patients had one or other of these complaints, and 44 had exactly two of them. How many patients had all three complaints?

Solution

Let U= $\{out - patients\ reported\ that\ day\}$**;** **F=** $\{Those\ who\ had\ fever\}$
S = $\{Those\ who\ had\ stomach\ trouble\}$**;** **J =** $\{Those\ who\ were\ injured\}$
x **=** $\{Those\ who\ had\ all\ three\ complains\}$

Let appropriate letters for the various regions in the Venn diagram is on next page

U



$f = 70 – (a + b + x)$; $s = 50 – (a + c + x)$; $j = 30 – (b + c + x)$

The term ''*44 had exactly two of them*'' $\Rightarrow a + b + c = 44$.

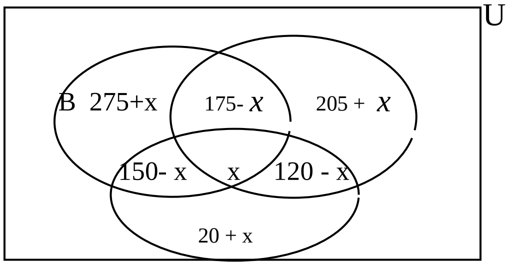
Since the union of the three sets adds up to 50, we can have;

$n(F \cup S \cup J) = f + s + j + a + b + c + x$
$= [70 – (a + b + x)] + [50 – (a + c + x)] + [30 – (b + c + x)] + a + b + c + x$
$= 150 – 2(a + b + c) + a + b + c – 3x + x = 150 – (a + b + c) - 2x$.
But $a + b + c = 44$ and therefore, $n(F \cup S \cup J)=150 – (a + b + c) - 2x =150 – 44 - 2x = 106 - 2x$
But $n(F \cup S \cup J) = 100$ and therefore
$106 - 2x = 100 \Rightarrow - 2x = - 6 \Rightarrow x = 3$, Hence, 3 people had all the three complaints.

Example: In an examination, each of the 1,000 students sat for Biology, Chemistry and Physics. All the Students passed at least one subject, 600 passed Biology, 500 passed Chemistry, and 290 passed Physics, 175 passed both Biology and Chemistry, 150 passed both Biology and Physics, and 120 passed both Chemistry and Physics. How many students passed
(a) all the three subjects (b) exactly one subject (c) exactly two subjects

Solution

Let U ={All students}; B={Students who passed Biology}; C={Students who passed Chemistry}
P = {Students who passed Physics} and x = Number who passed all the three subjects.
Using the formula:
$n(B \cup C \cup P) = n(B) + n(C) + n(P)–n(B \cap C) – n(B \cap P)– n(C \cap P) + n(B \cap C \cap P)$
$1,000 = 600 + 500 + 290 – 175 – 150 – 120 + x \Rightarrow x = 55$.



(a) Since $x = 55$, students passed all the three subjects

(b) Number who passed exactly one subject = $(275 + x) + (205 + x) + (20 + x)$
= (275 + 55) + (205 + 55) + (20 + 55) = 665

(c) Number who passed exactly two subject = $(175 - x) + (120 - x) + (150 - x)$
= (175 -55) + (120 -55) + (150 -55)= 280

6.2 Probability

Quite basic to the theory of probability is the idea of physical experiment. An experiment is any action that has a number of possible outcomes (or events). For example, the casting of a die once is an experiment of six possible outcomes which are: 1, 2, 3, 4, 5 or 6; while the tossing of a coin is an experiment of two outcomes – *head* or *tail*. It is however those experiments that are regulated in some probabilistic way that is helpful. A single performance of an experiment is called a trial for which there is a given set of outcomes.

Definition of Probability

To every event defined on a sample space S, we assign a non-negative number called a *probability*. We can therefore think of probability as a function (i.e. a function of the event) defined by the notation; *P(A),* for the probability of event *A* occurring. However, in the case of event explicitly stated as a set by the use of braces or curly brackets, we employ the notation P{A} rather than P({A}).

Probability is therefore, a measure of chance. It is a measure of likelihood of occurrence of an event. It indicates how much probable an event or an outcome can occur. If the total number of outcomes in the experiment is say ***n*** and an event from the experiment is ***a*** then the probability that the event occurs is given by:

$$\mathbf{P}(\boldsymbol{a}) = \frac{\boldsymbol{a}}{\boldsymbol{n}}$$

Thus, for example, in a toss of a fair die, the probability that 6 appears is $^1/_6$.

Trial, Outcome, Event and Sample space

A trial is any process which when repeated generates a set of results or observations. An outcome is the result of carrying out a trial. Thus, selecting a student at a random from a class is a trial while selecting a particular student say, Grace, is an outcome.

An event is a set which consists of one or more of the possible outcomes of a trial. A sample space is the set of all possible outcomes in any experiment. It is normally denoted by the letter *S* or the symbol **Ω**. Hence, the sample space is the universal set for any given experiment while an event is just a subset. All the outcomes in the sample space are mutually exclusive which, as has been explained in Section 7.7, means the occurrence of one of the outcomes rules out all the others. For example, one cannot have both a head (H) and a tail (T) in a single throw of a fair die. The probability of a sample space is equal to 1. Thus, P(S) = 1 or P(Ω) = 1.

Since, an event is a set, all our earlier definitions and operations applicable to sets are also applicable to events. For example, if two events have no common outcomes they are said to be mutually exclusive as has been explained in 7.7 above. The probability of any event *A* lies between zero and one. That is: $0 \leq \mathrm{P}(A) \leq 1$.

We can summarize therefore that, any trial has a number of possible outcomes, and the set of all possible outcomes is called the sample space. An event is defined to be a subset of sample space.

Probability space

Probability space corresponds to a given experiment comprising three items. An experiment is a course of action whose consequences is not predetermined. The three items of the probability space include:

(a) The set of all possible outcomes of the experiment which is usually called sample space.
(b) A list of all events which may possibly occur as a consequence of the experiment.
(c) An assessment of the likelihood of these events.

6.3 Types of Probability

The following are some of the various types of probability each of which plays a very important role in a specific activity.

Prior Probability

This is the probability which is concerned with estimating the likelihood that an event will occur. These probabilities are calculated prior to observing the results of an experiment. It is the type of probability which can be specified by common logic. An example is the throwing a fair die or a coin. This is an exact probability based on an objective approach.

Posterior Probability

The probability calculated after the outcome of an experiment has been observed which cannot be associated with common logic is called posterior probability. For example, if we want to find the probability of average number of workers who are punctual to work daily, will need to observe the attendance of workers for say one month and find the average number of workers who were punctual in a day. The result divided by the total number of workers is a posterior probability.

Empirical Probability

Any probability calculated from information gathered, is an empirical probability. Thus, if we want to know the probability of how many mangoes in a basket are bruised, we need to count the bruised mangoes and divide the result by the total number of mangoes in the basket. Hence, if the total number of mangoes in the basket is 50 out of which 30 are bruised, the probability of bruised mangoes is given by: P(bruised mangoes)=$^{30}/_{50}$= 0.6

Subjective Probability

At times, to find the probability of an event becomes impossible or impracticable. This is because it is unlikely to make situations exactly the same. Subjective probabilities are based on past experience of similar situations. They are therefore based on our own judgment. For example, if we want to find the probability of how many women will give birth in a locality for the next two years, the past records are studied to determine the trend of this event. The possible outcomes are for the period are then forecast which divided by the expected total to give the required probability.

6.4 Random Experiment

Any experiment conducted in such a way that each of the outcomes from the experiment has equal chance of being considered is termed as a *random experiment*. For example, in a toss of a fair coin the *head* or the *tail* has equal chance of showing up.

Equally Likely Events

Any set of events in the sample space which has all its members having equal chance of being drawn are said to be equally likely events. An example of such events is the outcomes from throwing a fair die. The event of getting 1, 2, 3, 4, 5, or 6 has a probability of $^1/_6$ for each score.

Unequally Likely Events

A set of events in the sample space whose members do not have equal chance of being drawn are said to be unequally likely. An example is throwing an unfair die. The chances of some faces showing up will be more probable than other faces.

Discrete and Continuous Variable

A variable can either be discrete or continuous. A variable is discrete if it assumes values which are usually whole numbers like 1, 2, 3, ---. A variable is usually represented by a letter or a symbol. Thus, if *x* represents the marks scored by 6 students in a class given as 18, 19, 20, 21, 19, and 22, then *x* is termed as a discrete variable because it assumes values which indicate disjoint points of whole numbers.

A continuous variable on the other hand, represents all measurements of intervals of points. A decimal or fractional value can be obtained for a continuous variable. The lifetime of a light bulb can be a continuous variable. Weight of students can also represent a continuous variable. It is therefore not restricted to whole numbers.

6.5 Probability Distribution

This is the list of all possible outcomes of an experiment and their corresponding probabilities. An example is the relative Frequency distribution given in Table 8.1 below. Another example of probability distribution is provided in Table 8.4.

Age (years)	Frequency	Relative Freq. or Probability
15	2	$^2/_{25} = 0.08$
16	5	$^5/_{25} = 0.20$
17	9	$^9/_{25} = 0.36$
19	6	$^6/_{25} = 0.24$
20	3	$^3/_{25} = 0.12$
Total	25	$^{25}/_{25} = 1.00$

6.6 Discrete and Continuous Probability

As with the sample space, events may be either discrete or continuous. The probability of any finite number of an infinite sequence of points is said to be a discrete probability. An example is the probability of throwing a fair coin or die. On the other hand, a continuous probability is the probability of the set of one or more intervals of points. An example is to find the probability of ages of children between 8 and 10 years.

Probability and Everyday Life

In many everyday situations, people are not too sure of certain events and therefore have to take precautions. For example, during Christmas period in Ghana, there are numerous lorry accidents. Hence, the probability of a person involving in an accident when traveling during a Christmas time is high. Many people therefore avoid travelling during this period.

A weather forecast on radio may state the chance of rain as 10% tomorrow but for another day it may be 90%. Thus, one advises himself as to whether or not carry a rain coat or an umbrella along.

We also apply probability at work places during planning and budgeting. How much to produce, what to produce, and when to produce, derive a great recognition from probability.

An insurance company will have to find out how long a person can live before accepting his life assurance policy to be processed. This is rightly done by considering the probability of how long the person will live. A vehicle is usually granted a comprehensive insurance policy after carefully examining its age and road worthy certificate to determine how probable it can exist and for what period. All these are well determined by the help of probability.

Probability is therefore an indisputable tool for all doctors, lawyers, managers, judges etc, in executing their day-to-day activities.

Set Notation of Events

Let us now consider problems on sets and probability.

Two-Set Problems

Problems on probability involving two sets are explained in Figure below. The sets A and B are presented as follows:

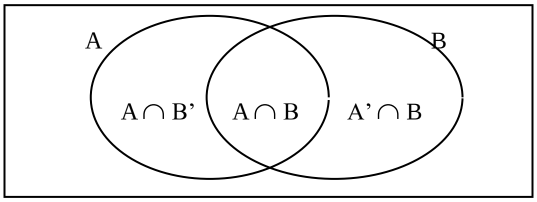


Figure : *Two-set Venn Diagram*

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

A and B are not mutually exclusive. Hence, to find $P(A \cup B)$ from the values given in the Venn diagram, the problem can be solved as follow:

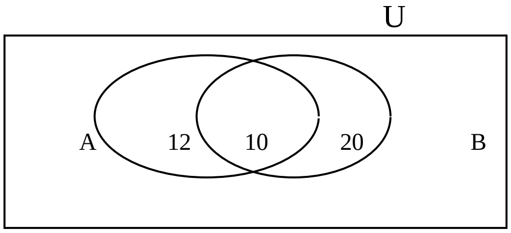


Figure : *Two-set Venn Diagram*

$P(A \cup B) = P(A)+P(B)–P(A \cap B)$; $n(A) = 12+10=22$; $n(B)=20+10=30$; $n(U)=48$
$n(A \cap B) = 10$

$$:.\ P(A \cup B) = \frac{22}{48} + \frac{30}{48} - \frac{10}{48} = \frac{42}{48} = \frac{7}{8}$$

Mutually Exclusive Events

Two or more events are said to be mutually exclusive if they have no outcome in common. The events are said to be disjoint. Examples of such events scoring a ***6*** on a fair die and getting a ***head*** on fair die when thrown once.

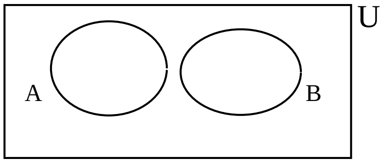
For any set of events to be mutually exclusive, it must satisfy the following conditions.

i. The probability of the intersection events must be zero. E.g. $P(A \cap B) = 0$

ii. The probability of the union events is the sum of the probabilities of the individual events e.g.

$P(A \cup B) = P(A) + P(B)$; $P(A \cup B \cup C) = P(A) + P(B) + P(C)$.

These conditions can be explained with diagrams as shown in below. The set A, B and C are farmers growing each of the products: tomato, pepper and onion. Since, no farmer grows more than one product; the probability of the intersection events is zero.



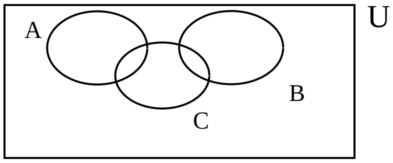


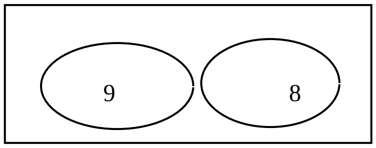
Figure: *Sets A and B*

Figure: *Sets A, B and C*

In above Figure, events A, B and C are mutually exclusive. Hence,

$P(A \cup B) = P(A \text{ or } B) = P(A) + P(B)$; and $P(A \cup B \cup C) = P(A \text{ or } B \text{ or } C) = P(A) + P(B) + P(C)$

In Figure below, events A and B are mutually exclusive. Therefore, the probability of the union event, $P(A \cup B)$, is calculated as follows.



$$P(A \cup B)=P(A \text{ or } B)=P(A) + P(B)= \frac{9}{20}+\frac{8}{20}=\frac{17}{20}$$

Figure : *Venn Diagram*

Three-set Problems

The problems in probability involving three sets are explained with the help of Figures Below.

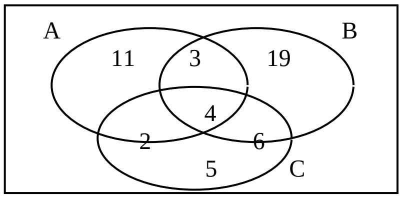


Figure A: *Three Intersecting Sets*

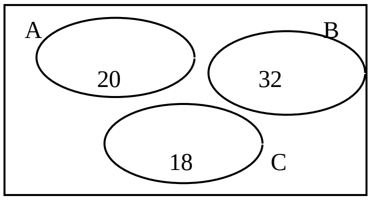


Figure B: *Three Disjoint Sets*

From Figure A above, the probability of the union of the three events is calculated as follows:

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

$$= \frac{20}{80} + \frac{32}{80} + \frac{17}{80} - \frac{3+4}{80} - \frac{2+4}{80} - \frac{4+6}{80} + \frac{4}{80} = \frac{73-23}{80} = \frac{5}{8}$$

From Figure 8.5B, since A, B and C are mutually exclusive; the probability of the union of the three events is calculated as follows:

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) = \frac{20}{80} + \frac{32}{80} + \frac{18}{80} = \frac{7}{8}$$

Complement Events

Let A' be the complement event of *A*. Then;

$$P(A) + P(A') = 1; \text{ or } P(A') = 1 – P(A) \text{; and also } P(A) = 1 –P(A')$$

Thus, the complement of an event is the set of outcomes in the sample space which are not members of outcomes of the given event.

For example, if the probability that Ben can win a game is 0.8, then the probability that Ben cannot win the game is: 1 – 0.8 = 0.2.

EXAMPLES

Example. Two boys, A_1 and A_2, play a game of chance. The probabilities of A_1 and A_2 winning the game are $^3/_5$ and $^5/_6$ respectively. Find the probability that

(a) Both of them win the game
(b) Only A_1 wins the game
(c) Only one wins the game

Solution

$P(A_1) = {}^3/_5$ $P(A_2) = {}^5/_6$ $P(A_1') = 1 – {}^3/_5 = {}^2/_5$ $P(A_2') = 1 – {}^5/_6 = {}^1/_6$

(a) $P(A_1 \text{ and } A_2) = P(A_1) \times P(A_2) = {}^3/_5 \times {}^5/_6 = ½$

(b) $P(A_1 \text{ and } A_2') = P(A_1) \times P(A_2') = {}^3/_5 \times {}^1/_6 = {}^1/_{10}$

(c) $P(A_1 \text{ and } A_2') \text{ or } P(A_1' \text{ and } A_2) = {}^3/_5 \times {}^1/_6 + {}^2/_5 \times {}^5/_6 = {}^{13}/_{30}$

Example. Ali, Amna and Farid solve a problem on Mathematics. The probability that Ali, Amna and Farid, can solve the problem are 0.7, 0.4 and 0.8 respectively. What is the probability that:

(a) All the three can solve the problem? (b) Only Amna can solve the problem?

(c) Only Amna cannot solve the problem? (d) None of them can solve the problem?

(e) At least one of them can solve the problem?

Solution

Let K = event of Ali solving the problem; K' = complement of K
A = event of Amna solving the problem; A' = complement of A
F = event of Farid solving the problem; F' = complement of F

P(K) = 0.7; P(A) = 0.4; P(F) = 0.8
P(K') = 1 – 0.7= 0.3; P(A') = 1 – 0.4= 0.6; P(F') = 1 – 0.8 = 0.2

(a) P(K and A and F) = $P(K)\times P(A)\times P(F) = 0.7\times 0.4\times 0.8 = 0.224$
(b) P(K' and A and F') = $P(K')\times P(A)\times P(F') = 0.3\times 0.4\times 0.2 = 0.024$
(c) P(K and $A'$and F) = $P(K)\times P(A')\times P(F) = 0.7\times 0.6\times 0.8 = 0.336$
(d) P(K' and$A'$and F') = $P(K')\times P(A')\times P(F') = 0.3\times 0.6\times 0.2 = 0.036$
(e) P(at least one can solve) = 1 – P(none can solve) = 1 – 0.036 = 0.964

Example. Three statistically independent events X, Y and Z are such that P(X) = 0.85;
P(Y = 0.72; P(Z) = 0.60, Find the probability of:
(a) X and Y occurring together (b) X and Z occurring together
(c) X, Y, and Z occurring together (d) None of them occurring

Solution

P(X) = 0.85; P(Y) = 0.72; P(Z) = 0.60
P(X') = 1 – 0.85= 0.15; P(Y') = 1 – 0.72= 0.28; P(Z') = 1 – 0.60 = 0.40

(a) P(X and Y)=$P(X)\times P(Y)=0.85\times 0.72 = 0.612$, (b) P(X and Z) = $P(X)\times P(Z) = 0.85\times 0.60 = 0.51$
(c) P(X and Y and Z) = $P(X)\times P(Y)\times P(Z) = 0.85\times 0.72\times 0.60 = 0.3672$
(d) P($X'$and $Y'$and Z') = $P(X')\times P(Y')\times P(Z') = 0.15\times 0.28\times 0.40 = 0.0168$

Relative Frequency Interpretation of Probability

Consider the frequency distribution table below and the Relative Frequency Table can be constructed as shown below:

Table: Frequency and Relative Frequency Distribution

Age (Years)	Frequency	Relative Frequency
2	3	$^{3}/_{20}$ = 0.15 = 15%
3	4	$^{4}/_{20}$ = 0.20 = 20%
4	8	$^{8}/_{20}$ = 0.40 = 40%
5	3	$^{3}/_{20}$ = 0.15 = 15%
6	2	$^{2}/_{20}$ = 0.15 = 10%
Total	20	$^{20}/_{20}$ = 1.00 = 100%

From Table, it could be seen that the sum of the frequencies is 20 and the sum of the corresponding relative frequencies is one (or 100%).

Let *X* denotes a random variable showing the age of boys from 2 years to 6 years. With the frequency table above, the probability distribution will be deduced as follows:

Table: Probability Distribution Table

X	2	3	4	5	6	Total
Frequency	3	4	8	3	2	20
P(*X*)	$^{3}/_{20}$	$^{4}/_{20}$	$^{8}/_{20}$	$^{3}/_{20}$	$^{2}/_{20}$	$^{20}/_{20} = 1$

From the foregoing therefore, the probability distribution of a random variable *X* is the list of the relative frequencies of the variable *X*.

6.7 Probability Tree Diagram

The theory of probability can be expanded with the probability tree diagram. For example, if a fair coin is tossed once, the sample space, S = [H, T]. It therefore consists of two possible outcomes. This can be represented in a *Tree Diagram* as shown in Figure.

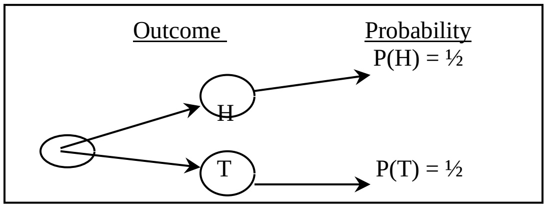


Figure : Probability Tree Diagram

Let us consider the coin when thrown twice. The sample space, *S* = [HH, HT, TH, TT], given us four possible outcomes. The tree diagram can be constructed as follows:

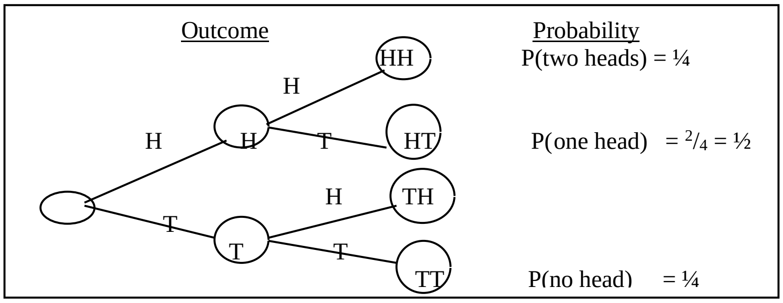


Figure: Probability Tree Diagram

If the coin is tossed thrice, the sample space S = [HHH, HHT, HTH, THH, HTT, THT, TTH, TTT]. Thus, eight possible outcomes are to be realized. The probability tree diagram is given as follows:

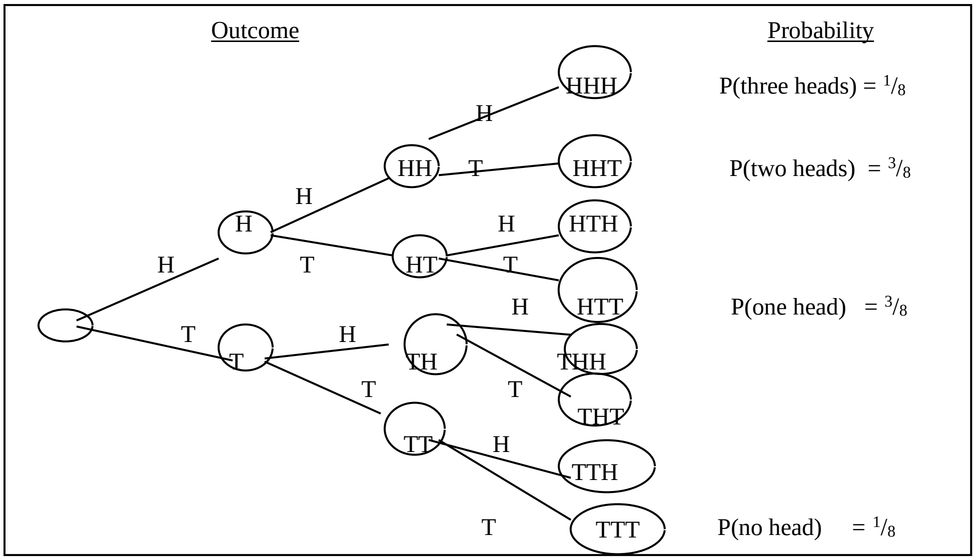


Figure: Probability Tree Diagram

6.8 Laws or Axioms of Probability

The various laws (or axioms) of probability are the ; ***P(A)≥0***
This means that the probability of any event ***A***, is non-negative. That is, it is either ***positive*** or ***zero***. Negative values are meaningless, and in fact, do not occur in the theory of probability. If the probability of the occurrence of an event is ***zero***, then

that event will not occur; and if it is ***one***, then that event will certainly occur**.** ***P(A)*** should lie between ***0*** and ***1***.

P(all possible outcomes) = P(S) = 1

This law recognizes the fact that, the sample space itself is an event which however, encompasses all events in that experiment. Hence, the sample space should have the highest possible probability of one.

Addition Law Probability (or)

The addition law is applied to the calculation of probability of two or more mutually exclusive events. Under this law, all individual probabilities are added together. The word '***or***' and the union sign, ' $\cup$ ' are concerned with addition of probabilities.

Let A_1, A_2, A_3, ---, A_n be events in the sample space which are mutually exclusive. Then:
$P(A_1$ or A_2 or A_3 or---or $A_n)=(P(A_1 \cup A_2 \cup A_3 \cup$ ----- $\cup A_n)=$ $P(A_1)+P(A_2)+P(A_3)+$--+ $P(A_n)$

$$P(\cup_n A_n) = \sum P(A_n)$$

This means the events, A_1, A_2, A_3, ---, A_n, are disjoint and therefore the union of their probabilities is the sum of the individual probabilities.

For example, to find the probability of scoring a '6' with a fair die or a 'Head' with a fair coin after tossing the die and the coin once, we proceed as follows:

$$P(6 \text{ or } H) = P(6) + P(H) = {}^1/_6 + ½ = {}^2/_3$$

Multiplication Law of Probability (and)

The law here is applied to a string of independent events of which individual probabilities are known and it is required to know the overall probability. The multiplication law of any two given events, A and B, is given by:

$$P(A \text{ and } B) = P(A \cap B) = P(A)\times P(B)$$

For example, to find the probability of scoring a '6' with a throw of a die and a 'Head' with a throw of a coin, we proceed as follows:

$$P(6 \text{ and } H) = P(6 \cap H) = P(6) \times P(H) = {}^1/_6 \times ½ = {}^1/_{12}$$

Selection *with* Replacement and Selection *without* Replacement

Selection with replacement is the selection procedure which requires that an item(s) selected is/are replaced before subsequent selections. This type of selection procedure corresponds to independent events. In this case, because an item is put back into the system before subsequent selection, the probability of any selection of a particular event and the subsequent ones of the same event, will not change.

As an example, let us find the probability of selecting two red balls from a bag containing 5 red, 6 blue and 7 green identical balls at random, one after the other, **with** replacement.

If R, B and G are the events of selecting red, blue and green balls respectively, then since the total number of balls is 18 and n(R) = 5; n(B) = 6 and n(G) = 7: $P(R) = {}^{5}/_{18}$; $P(B) = {}^{6}/_{18}$ and $P(G) = {}^{7}/_{18}$

Hence, the required probability will be calculated as follows:

P(1st is red and 2nd is red) = $P(R_1 \text{ and } R_2) = P(R_1 \cap R_2) = P(R_1) \times P(R_2) = {}^{5}/_{18} \times {}^{5}/_{18} = {}^{25}/_{324}$

Example: let us assume that a bag contains 8 white, 5 brown and 7 green marbles. Three of them are selected at random with replacement. Let us find the probability that :

(a) They are all white, b) They are of the same colour and c) The first two are brown, and the third green.

Solution:

The problem is solved as follows:

Let W = event of selecting a white marble, B = event of selecting a brown marble
G = event of selecting a green marble

n(W) = 8; n(B) = 5 and n(G) = 7: Total number of marbles = 8 + 5 + 7 = 20

(a) $P(W_1 \cap W_2 \cap W_3) = P(W_1) \times P(W_2) \times P(W_3) = {}^{8}/_{20} \times {}^{8}/_{20} \times {}^{8}/_{20} = {}^{8}/_{125}$

(b) $P(W_1 \cap W_2 \cap W_3 \text{ or } B_1 \cap B_2 \cap B_3 \text{ or } G_1 \cap G_2 \cap G_3)$
$= P(W_1 \cap W_2 \cap W_3) + P(B_1 \cap B_2 \cap B_3) + P(G_1 \cap G_2 \cap G_3)$
$= {}^{8}/_{20} \times {}^{8}/_{20} \times {}^{8}/_{20} + {}^{5}/_{20} \times {}^{5}/_{20} \times {}^{5}/_{20} + {}^{7}/_{20} \times {}^{7}/_{20} \times {}^{7}/_{20} = {}^{49}/_{400}$

On the other hand, *selection without replacement* is the selection procedure in which every item selected is *not replaced* before subsequent selections. This type of selection corresponds to dependent events. For example, let us consider the previous illustration where this time, the two red balls are selected at random, one after the other without replacement. When the first red ball is selected, the number of red balls in the bag will reduce by one and likewise, the total number of balls in the bag will reduce by one. The required probability will then be given by:

$$P(1^{st} \text{ is red and } 2^{nd} \text{ is red}) = P(R_1 \text{ and } R_2/R_1) = P(R_1) \times P(R_2/R_1) = {}^{5}/_{18} \times {}^{4}/_{17} = {}^{10}/_{153}$$

Example: consider a box containing 7 blue and 5 green marbles of the same sizes only for colour. Two marbles are selected at random, one after the other without a replacement. Let us find the probability that: They are of the same colour,
(b) Each colour is selected.

Solution:

These can be calculated as follows:
n(B) = 7; n(G) = 5. The total number of marbles = 7 + 5 = 12.

(a) $P(B_1 \text{ and } B_2 \text{ or } G_1 \text{ and } G_2) = P(B_1)P(B_2/B_1) + P(G_1)P(G_2/G_1)$
$= {}^{7}/_{12} \times {}^{6}/_{11} + {}^{5}/_{12} \times {}^{4}/_{11} = {}^{31}/_{66}$

(b) $P(B_1 \text{ and } G_2 \text{ or } G_1 \text{ and } B_2) = P(B_1)P(G_2/B_1) + P(G_1)P(B_2/G_1)$
$= {}^{7}/_{12} \times {}^{5}/_{11} + {}^{5}/_{12} \times {}^{7}/_{11} = {}^{35}/_{66}$

Statistically Independent Events

We want to introduce the concept of statistically independent events. In general, any given experiment may involve a number of events but we will first consider the simplest possible case of two events.

Let A_1 and A_2 be any two events which have nonzero probabilities of occurrence; that is, $P(A_1) \neq 0$ and $P(A_2) \neq 0$. The two events, A_1 and A_2, are said to be statistically independent if the probability of occurrence of one event is not affected by the occurrence of the other event. Thus,

$$P(A_1/A_2) = P(A_1) \text{ and, } \quad P(A_2/A_1) = P(A_2)$$

As we shall see later from conditional probability, the two events above can have a joint probability equal to the product of the probabilities of the events given by:

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

It has already been stated earlier in this chapter that the joint probability of two mutually exclusive events is zero. That is; $P(A_1 \cap A_2) = 0$. Thus, if two events have nonzero probabilities, they cannot be both mutually exclusive and statistically independent. Therefore, for any two events to be independent, they must have an intersection. That is;

$$A_1 \cap A_2 \neq \phi$$

As an illustration, let us consider two statistically independent events, *A and B*, with $P(A) = 0.4$ and $P(B) = 0.6$. Let us find the probability of both events occurring together.

Since *A and B* are statistically independent,

$$P(A \cap B) = P(A)\,P(B) = 0.4 \times 0.6 = 0.24$$

Another example can be given about a 52-card deck in which *A* is the event of selecting a *King*; *B* the event of selecting a *jack* or *queen*; and *C*, the event of selecting a *heart*. The corresponding probabilities of the three events are:

$$P(A) = {}^{4}/_{52}; \quad P(B) = {}^{8}/_{52}; \quad P(C) = {}^{13}/_{52}$$

The following joint probabilities can be computed can be computed from the above information.

$P(A \cap B) = 0$; since it is not possible to select a *king* and a *jack* or *queen* at the same time.

Since the other pairs are independent:

$P(A \cap C) = P(A)\,P(C) = {}^{4}/_{52} \times {}^{13}/_{52} = {}^{1}/_{52}$, $\quad P(B \cap C) = P(B)\,P(C) = {}^{13}/_{52} \times {}^{13}/_{52} = {}^{1}/_{52}$

Multiple Events

The set of events A_1, A_2, A_3, ---, A_n, are said to be independent if only and only if they are independent by pairs and also independent as a joint, of all the *n* possible events. Thus, for three given events A_1, A_2 and A_3, which are independent, the following conditions must be satisfied.

$$P(A_1 \cap A_2) = P(A_1)\,P(A_2), \qquad P(A_1 \cap A_3) = P(A_1)\,P(A_3)$$
$$P(A_2 \cap A_3) = P(A_2)\,P(A_3)\,, \qquad P(A_1 \cap A_2 \cap A_3) = P(A_1)\,P(A_2)\,P(A_3)$$

More generally, for *n* statistically independent events, it is required that all the conditions below must be satisfied for all $1 < i < j < \text{---} < n$

$P(A_i \cap A_j) = P(A_i)\, P(A_j)$
$P(A_1 \cap A_2 \cap A_3 \cap \text{---} \cap A_n) = P(A_1)\, P(A_2)\, P(A_3)\text{---}P(A_n).$

As an example, let us consider three boys *A, B, and C* who play a game of chance. The probabilities that *A, B, and C* win the game are 0.5, 0.7, and 0.9 respectively.

The probability that the three boys will win the game can be calculated as follows:

Let A = event of A winning the game, B = event of B winning the game
C = event of C winning the game

Then, P(A) = 0.5; P(B) = 0.7; P(C) = 0.9 and since the three event are independent,

$$P(A \text{ and } B \text{ and } C) = P(A \cap B \cap C) = P(A)\, P(B)\, P(C) = 0.5 \times 0.7 \times 0.9 = 0.315$$

Statistically Dependent (or not-independent) Events

Any set of events are said to be non-independent if the occurrence of the given event is affected by the occurrence of the previous event or events of the Same Sample Space. Thus, joint probabilities events under this are just like the problems under this are just like the problems under selection without replacement

Hence, if the events *A and B***,** are not independent, then

$$P(B/A) \neq P(B), \qquad \text{and} \qquad P(A \cap B) = P(A)\, P(B/A)$$

Joint and Conditional Probability.

The probability ***P(A ∩ B)*** is called the *joint probability* for two events ***A*** and ***B*** which represent the intersection of the sample space. As we saw from *equation 8.6.01* above,

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

which is equivalent to

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

Thus, it should be noted that the probability of the union of two events can never exceed the sum of the probabilities of the individual events. The equality holds only for mutually exclusive events since in this case, $A \cap B = \phi$ and therefore, $P(A \cap B) = P(\phi) = 0$

On the other hand, given some event *B* with nonzero probability, $P(B) > 0$**,** we define the conditional probability of an event *A*, given that *B* has occurred, by

$$P(A/B) = \frac{P(A \cap B)}{P(B)} \quad \text{and} \quad P(A/B)\,P(B) = P(A \cap B)$$

Similarly, for a nonzero event *A*, $P(B/A) = \frac{P(A \cap B)}{P(A)}$, $P(B/A)\,P(A) = P(A \cap B)$

Total Probability

Let P(A) of an event A be any probability defined on a sample space S. P(A) can be expressed in terms of conditional probabilities on the sample space S which has been partitioned into n mutually exclusive events D_i, $i = 1, 2, 3, ---,n$; whose union equals *S*.

The intersection of any pair or any group of the partitioned events is an empty set. That is:

$$Bi \cap Bj = \phi; \qquad i \neq j = 1, 2, 3, ---, n, \quad \text{and;} \quad \bigcup_{i=1}^{n} Bi = S$$

Since $A \cap S = A$, it follows that; $A \cap S = A \cap (\bigcup_{i=1}^{n} Bi) = \bigcup_{i=1}^{n} (A \cap Bi)$

Since the events, $A \cap Bi; I = 1, 2, 3, ---, n$ are mutually exclusive, as seen from the axiom above, it follows:

$$P(A) = P(A \cap S) = P[\bigcup_{i=1}^{n} (A \cap Bi)] = \sum_{i=1}^{n} P(A \cap Bi)$$

But from above, we can write: $P(A \cap B_1) = P(A/B_1)\,P(B_1)$; $P(A \cap B_2) = P(A/B_2)\,P(B_2)$;
$P(A \cap B_3) = P(A/B_3)\,P(B_3)$; -------------; $P(A \cap B_n) = P(A/B_n)\,P(B_n)$

Thus, we can write: $\sum_{i=1}^{n} P(A \cap Bi) = \sum_{i=1}^{n} P(A / B_i) P(Bi)$

From above equation it is known as the total probability of event *A*.

Bayes Theorem

The definition of conditional probability, as given by 8.24.02 and 8.24.04, applies to any two events in the sample space. Thus, if *Bi* is any one of the events defined in 8.24.05, we can write:

P(B_i/A) = $\frac{P(A \cap B_i)}{P(A)} \Rightarrow$ P(A) P(B_i/A) = P(A $\cap$ B_i) ; P(A)$\neq 0$

Alternatively,

P(A/ B_i) = $\frac{P(A \cap B_i)}{P(B_i)} \Rightarrow$ P(B_i) P(A/ B_i) = P(A $\cap$ B_i) ; P(B_i)$\neq 0$

Equations 8.26.01 and 8.26.02 one form of Bayes' theorem as:

$$\mathbf{P(B_i/A) =} \frac{P(A/B_i)P(Bi)}{P(A)}$$

But from Equations,

P(A) = $\sum_{i=1}^{n} P(A/B_i)P(Bi)$**. Thus, for any partitioned event *B_i*, to occur given any event *A*, we can write:**

P(B_i/A) = $\frac{P(A/B_i)P(Bi)}{P(A)} = \frac{P(A/B_i)P(Bi)}{P(A/B_1)P(B_1)+P(A/B_2)P(B_2)+....+P(A/B_n)P(B_n)}$

$$= \frac{P(A/B_i)P(Bi)}{\sum P(A/B_1)P(B_1)}$$

Thus, in general, if we have *n* independent events A_1, A_2, A_3, ---, A_n, and W is any other event which is common to the mutually exclusive events, A_1, A_2, A_3, ---, A_n, then by Bayes' theorem:

$$P(A_i/W) = \frac{P(W/A_i)P(A_i)}{P(W/A_1)P(A_1)+P(W/A_2)P(A_2)+....+P(W/A_n)P(A_n)}$$

$$= \frac{P(W/A_{i\,i})P(A_i)}{\sum P(W/A_i)P(A_i)}$$

Let us illustrate the above theorem with the following example.

One box contains two red balls and a second box of identical appearance contains one red and one white balls. If a box is selected at random and one ball is drawn from it, let us find the probability that the first box was the selected one if the drawn ball is red.

To solve such a problem:

Let B_1 = event of selecting the first box , B_2 = event of selecting the second box

R = event of selecting a red ball

$P(B_1) = ½$; $P(B_2) = ½$, $P(R/ B_1) = {}^2/_2 = 1$; $P(R/ B_2) = ½$

$P(B_1 \cap R) = P(B_1)P(P(R/ B_1) = ½ \times 1 = ½$, $P(B_2 \cap R) = P(B_2)P(P(R/ B_2) = ½ \times ½ = ¼$

$P(R) = P(B_1 \cap R) + P(B_2 \cap R) = ½ + ¼ = ¾$

But, $P(B_1/R) = \dfrac{P(B_1)P(R/B_1)}{P(B_1)P(R/B_1)+P(B_2)P(R/B_2)} = \dfrac{\frac{1}{2}}{\frac{1}{2}+\frac{1}{4}} = \dfrac{2}{3}$

Example: There are four different machines A, B, C and D with their respective degrees of accuracy being 90%, 70%, 50% and 40%. The probabilities that the machines will give wrong results are given as 2%, 5%, 7% and 9% respectively. If a machine is operating wrongly we can find the probability that it is machine C as follows:

Let W = event of a machine operating wrongly.

$P(A) = 0.9$; $P(B) = 0.7$; $P(C) = 0.5$; and $P(D) = 0.4$
$P(W/A) = 0.02$; $P(W/B) = 0.05$; $P(W/C) = 0.07$; and $P(W/D) = 0.09$

The required probability is given by:

$$P(C/W) = \frac{P(W/C)P(C)}{P(W/A)P(A)+P(W/B)P(B)+P(W/C)P(C)+P(W/D)P(D)}$$

$$P(C/W) = \frac{(0.5)(0.07)}{(0.9)(0.02)+(0.7)(0.05)+(0.5)(0.07)+(0.4)(0.09)} = 0.282$$

Combinational Analysis Application to Probability

In Chapter Six, we learnt that the two types of arrangements – combination and permutation. We are going to learn further how questions on dependent events can be solved using combination and permutation. The following examples can be used to illustrate this.

Example: A box containing 6 red and 9 blue balls. Two balls are selected at random, one after the other without replacement. Let us find the probabilities of the following events:

(a) They are both red b) They are of the same colour c) Each colour is selected.

Solution:

Let R = event of selecting a red ball, B= event of selecting a blue ball
n(R) = 6; n(B) = 9. Total number of balls = 6 + 9 = 15

The total number of ways of selecting any two balls out of the fifteen is given by:

$$^{15}C_2 = \frac{15!}{(15-2)!2!} = 105$$

(a) The total number of ways of selecting two red balls out of the six red balls is given by: $P(R_1R_1) = \frac{^6C_2}{105} = \frac{15}{105} = \frac{1}{7}$

(b) P(same colour) = P(R_1R_2) + P(B_1B_2) = $\frac{^6C_2 + ^9C_2}{105} = \frac{15+36}{105} = \frac{17}{35}$

(c) P(each colour selected) = P(RB) = $\frac{^6C_1 \times ^9C_1}{105} = \frac{6\times 9}{105} = \frac{54}{105} = \frac{18}{35}$

EXAMPLES

Example. A bag contains 5 red, 4 blue and 3 white marbles. Three of them are selected without replacement. Find the probability that:
(a) They are all blue, (b) Each of the colours is selected, (c) Two blue and one white are selected
(d) At least one red was drawn, (e) Each colour is selected in order red, blue and white.

Solution

Let R = event of selecting a red marble, B = event of selecting a blue marble
W = event of selecting a white marble and Total number of balls = 5 + 4 + 3 = 12
No. of way of selecting 3 out of 12 marbles = $^{12}C_3 = 220$

(a) P(all are blue) = $P(B_1B_2B_3) = \frac{^4C_3}{220} = \frac{4}{220} = \frac{1}{55}$,

b) P(each colour selected) = $\frac{^5C_1 \times ^4C_1 \times ^3C_1}{220} = \frac{5\times 4\times 3}{220} = \frac{3}{11}$

(c) P(2 blue and one white) = P(BBW) = $\frac{{}^4C_2 \times {}^3C_1}{220} = \frac{18}{220} = \frac{9}{110}$

(d) P(at least one red) = 1 – P(no red) = $1 - \frac{{}^{12-5}C_3}{220} = 1 - \frac{{}^7C_3}{220} = 1 - \frac{35}{220} = \frac{185}{220} = \frac{37}{44}$

(e) P(each colour in Order R, B, W) $\frac{{}^5P_1 \times {}^4P_1 \times {}^3P_1}{{}^{12}P_3} = \frac{5 \times 4 \times 3}{1320} = \frac{1}{22}$

Example. If A and B are mutually exclusive with P(A) = 0.3 and P(B) = 0.3, find (a) P(A $\cup$ B) , (b) P(A $\cap$ B)

Solution

(a) P(A $\cup$ B) = P(A) + P(B) = 0.3 + 0.4 = 0.7 [A and B are mutually exclusive]

(b) P(A $\cap$ B) = 0 [Since A and B are disjoint sets]

Example. If A and B are independent events with P(A) = 0.2 and P(B) = 0.5, find: (a) P(A $\cap$ B) (b) P(A $\cup$ B) (c) P($A^{/} \cap B^{/}$)

Solution

(a) P(A $\cap$ B) = P(A)P(B) = 0.2 $\times$ 0.5 = 0.10, (b) P(A $\cup$ B) = P(A) + P(B) – P(A $\cap$ B)
= P(A) + P(B) – P(A)P(B) = 0.2 + 0.5 – (0.2)(0.5) = 0.60

(c) P($A^{/} \cap B^{/}$) = P($A^{/}$)P($B^{/}$) = (1 – 0.2)(1 – 0.5) = 0.8 $\times$ 0.5 = 0.40

Note that by De Morgan's Law: P($A^{/} \cap B^{/}$) = P(A $\cup$ B)$^{/}$ = 1 – 0.6 = 0.4 [from (b)]

Example. If P(A) = *x*, P(B) = 0.35 and P(A $\cup$ B) = 0.83, find *x* if:
(a) A and B are mutually exclusive, (b) A and B are independent

Solution

(a) P(A $\cup$ B) = P(A) + P(B)
0.83 = x + 0.35
x = 0.83 – 0.35
x = 0.48

(b) P(A $\cup$ B) = P(A) + P(B) – P(A)P(B)
0.83 = x + 0.35 – (x)(0.35)
x – 0.35x = 0.83 – 0.35
0.65x = 0.48 $\Rightarrow$ x = 0.74

Example. If P(A) = *x*, P(B) = ½*x* and P(A $\cup$ B) = 0.8, find the value of *x* if A and B are independent.

Solution

P(A $\cup$ B) = P(A) + P(B) – P(A)P(B), by putting value 0.8 = x + ½x – (x)(½x)

0.8 = $^{3}/_{2}x$ – ½x^2 (Multiplying by 2 and re-arranging)

$x^2 - 3x + 1.6 = 0$

$x = \frac{3 \pm \sqrt{(-3)^2 - 4(1)(1.6)}}{2(1)} \Rightarrow x = 2.3; \quad x = 0.7$, Since x should lie between 0 and 1, x = 0.7.

Example. a)Two events A and B, are independent with P(A) = 0.4 and P(B) = 0.7. What is P($A^{/} \cap$ B)?

(b) Two events E and F are such that P(E $\cup$ F) = 0.8, P(E) = 0.7 and P(F) = 0.6. Find

(i) P($E^{/}$/F) (ii) P($F^{/}$/$E^{/}$)

Solution

(a) P($A^{/} \cap$ B) = P(B) - P(A $\cap$ B) = P(B) – P(A)P(B) = 0.7 – (0.4)(0.7) = 0.42

(b) (i) P($F^{/}$/E) = $\frac{P(F^{/} \cap E)}{P(E)} = \frac{P(E) - P(F \cap E)}{P(E)}$

But P(F $\cap$ E) = P(F) + P(E) – P(F $\cup$ E)$\Rightarrow$P(F $\cap$ E) = 0.7 + 0.6 – 0.8 = 0.5; and

P($E^{/}$) = 1 – P(E) = 1 – 0.7 = 0.3

P($F^{/}$/E)= $\frac{P(E) - P(F \cap E)}{P(E)} = \frac{0.7 - 0.5}{0.7} = 0.286$ (ii) P($F^{/}$/$E^{/}$)=

$\frac{P(F^{/} \cap E^{/})}{P(E^{/})} = \frac{P(F \cup E)^{/}}{P(E^{/})} = \frac{1 - 0.8}{1 - 0.7} = \frac{0.2}{0.3} = 0.667$

Example. The probability that a certain beginner at golf gets a good shot if he uses the correct club is $^{1}/_{3}$, and the probability of a good shot with an incorrect club is ¼ . In his bag are 5 different clubs only one of which is correct for the shot in question. If he chooses a club at random and takes a stroke what is the probability that:

(a) He gets a good shot, (b) The correct club had a good shot?

Solution

Let A = event of choosing a correct club   D = event of getting a good shot
B = event of choosing an incorrect club
P(D/A) = $^1/_3$;   P(D/B) = $^1/_4$; P(A) = $^1/_5$ and P(B) = 1- $^1/_5$ = $^4/_5$

(a) P(good shot) = P(D) = P(good shot due to A) + P(good shot due to B)
= P(A $\cap$ D) + P(B $\cap$ D) = P(A)P(D/A) + P(B)P(D/B) = $^1/_5 \times {}^1/_3 + {}^4/_5 \times {}^1/_4 = {}^4/_{15}$

(b) P(A/D) = $P(A/D) = \dfrac{P(A)P(D/A)}{P(D)} = \dfrac{1/5 \times 1/3}{4/5} = \dfrac{1}{4}$

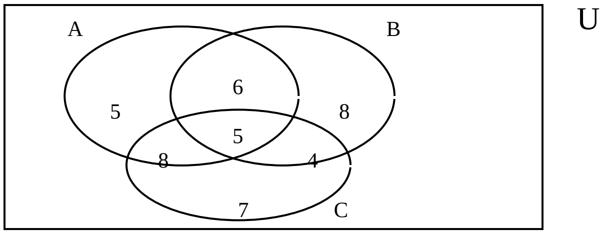
Example. On a visit to a dentist, a patient is told that his mouth contains 20 of his original teeth of which 5 are required to be drilled, 3 extracted and the rest left. What is the probability that if two teeth are chosen at random (a) They would both be required to be drilled?
(b) One will have to be drilled and one extracted?

Solution

Total number of teeth = 20; Number to be drilled (D) = 5; Number to be extracted (E) = 3
(a) P(both drilled) = P(D_1 and D_2) = P(D_1)P(D_2 / D_1) = $^5/_{20} \times {}^4/_{19} = {}^1/_{19}$
(b) P(one drilled and one extracted) = P(D_1 and E_2) or P(E_1 and D_2) = $^5/_{20} \times {}^3/_{19} + {}^3/_{20} \times {}^5/_{19}$ =3/38

6.9 SELF ASSESSMENT QUESTIONS

Q1. Consider the Venn Diagram below in which A = {Footballers}; B = {Hockey Players}; C = {Volleyball Players}



Find the proportion of players in (a) Set A (b) all three sets (c)sets A and B (d)only one set (e) none of the three Sets [Ans; (a)12/25 (b)1/10 (c)11/50 (d)2/5 (e)7/50]

Q2. A survey of reading habits of 130 students showed that 30 read both Comics and Novels, 10 read neither Comics nor Novels and twice as many read Comics as read Novels. How many read (a)Comics (b)Novels (c)Only Comics or only Novels [Ans;(a)100 (b)50 (c)90]

Q3. In a class of 50 students, 27 study French, 24 study History and 30 study Geography. Each student studies at least one of the three subjects. Five study all the three subjects while 11 study French and Geography. How many study (a)One of the three subjects (b) exactly two subjects [Ans; (a) 24 (b) 21]

Q4. Three girls are to write professional examinations. They are Amna, Bernice and Mabel. The probability that they will pass the examinations are; 0.5, 0.7 and 0.8 respectively. What is the probability that (a) The three girls will pass the examinations? (b) None of them will pass the examinations? (c) Only Mabel will pass the examinations? (d) Only one of them will pass the examinations?(e) At least one of them will pass the examinations?

[Ans: (a)0.28 (b)0.03 (c) 0.03 d)0.22 (e)0.97

Q5. A box contains 8 red, 3 white and 9 blue balls. If 3 balls are drawn at random determine the probability that all 3 are red (b) all are white (c) 2 are red and 1 is white (d) at least 1 is white (e) one of each colour is drawn (f) the balls are drawn in other; red white, blue.

$$\left[Ans:(a)\frac{14}{285}\quad (b)\frac{1}{1140}\quad (c)\frac{21}{95}\quad (d)\frac{23}{75}\quad (e)\frac{18}{95}\quad (f)\frac{3}{95}\right]$$

Q6. A diagnostic test for a new disease has the following characteristics: A person with disease if given the test certainly show positive reactions, while 10% of persons without the disease who are administered the test show positive reaction. If in a population sampled, one percent of the people have the disease, what percentage of those who reacted to the test actually has the disease? [Ans: 9%]

Q7. If two dice are tossed together once, what is the probability of a) getting a total of 7? b) Each one of them shows at least 5 points? [Ans: (a) ${}^{2}/_{9}$ (b) ${}^{1}/_{9}$]

Q8. Three fair coin are tossed together. i). List the members of the sample space ii). Find the probability of getting: (a) At least one head (b) no tail (c) one head and two tails (d) three tails or two tails [Ans: (a) ${}^{7}/_{8}$ (b) ${}^{1}/_{8}$ (c) ${}^{9}/_{64}$ (d) ½]

Q9. The events A, B and C satisfy these conditions:P(A) = 0.6 P(B) = 0.8 P(B/A) = 0.45 P(B and C) = 0.28 Calculate: (a) P(A and B) (b) P(C/B) (c) P(A/B) [Ans:(a) 0.27(b) 0.35 (c) 0.3375]

Q10. Given that P(A)=0.75,P(B/A)=0.8 and P(B/A^c)=0.6;Calculate P(B) and P(A/B) [Ans: 0.75; 0.8]

Q11. The probability that an event A occurs is P(A) = 0.3. The event B is independent of A and P(B) = 0.4. a) Calculate P(A or B or both occur) Event C is defined to be event that neither A nor B occurs. Calculate P(C/A'), where A' is the event that A does not occur. [Ans: (a) 0.58 (b) 0.6]

SUGGESTED READINGS

Bluman, A.G. (2004). Elementary Statistics. A Step by Step Approach. 5th Edition. McGraw-Hill Companies Incorporated. London.

Chaudhary, S.M. & Kamal, S. (2017). Introduction to Statistical Theory Part-I. 8th Edition. Ilmi Kitab Khana. Lahore.

Chaudhary, S.M. & Kamal, S. (2017). Introduction to Statistical Theory Part-II. 8th Edition. Ilmi Kitab Khana. Lahore.

Daniel, W.W. (1995). Biostatistics: A foundation for Analysis in Health sciences. Sixth Edition. John Wiley and sons Incorporated. USA.

Harper, W.M. (1991). Statistics. Sixth Edition. Pitman Publishing, Longman Group, United Kingdom.

Hoel, P.G. (1976). Elementary Statistics. 4th Edition. John Wiley and Sons Incorporated, NewYork.

Kiani, G. H., & Akhtar, M. S. (2012). Basic statistics, Majeed Book Depot.

Khan, A. A., Mirza, S. H., Ahmad, M. I., Baig, I. & Yaqoob, M. (2011). Business Statistics, Qureshi Brothers Publishers.

UNIT 07

PROBABILITY DISTRIBUTIONS

Written By: Dr. Zahid Iqbal
Reviewed By: Dr. Muhammad Ilyas

CONTENTS

Pages

Introduction

In unit 6, we have shown frequency distribution as a useful way of summarizing variations in observed data. Probability distribution could be thought of as the theoretical frequency distribution rather than observed one. A theoretical frequency distribution is a probability distribution that describes how outcomes are expected to vary. Because these distributions deal with expectations, they are useful models in making inferences and decisions under conditions of uncertainty.

Objectives

After studying this unit, you will be able to.

- Understand discrete distribution.
- Understand the difference between a discrete and continuous probability distribution.
- Understand the binomial distribution (discrete) and calculate probabilities of discrete outcomes.
- Understand and calculate probabilities of the Poisson (discrete) distribution.
- What probability distribution depicts the expected outcomes of possible values for a given data generating process.
- Probability distributions come in many shapes with different characteristics, as defined by the mean, standard deviation, skewness, and kurtosis.
- Investors use probability distributions to anticipate returns on assets such as stocks over time and to hedge their risk.

7.1 Binomial Random Variable

Many experiments have responses with Two possibilities (Yes/No, Pass/Fail, True/False).
Certain experiments called binomial experiments yield a type of random variable (r.v.) called a binomial random variable.

Characteristics of a binomial experiment:

(1) The experiment consists of a fixed number (denoted *n*) of identical trials.
(2) There are only two possible outcomes for each trial – denoted "Success" (S) or "Failure" (F)
(3) The probability of success (denoted *p*) is constant for each trial.
(4) The trials are independent.

Then the binomial r.v. denoted by *X* is the number of successes in the *n* trials.

Example: A fair coin is flipped 5 times. Define "success" as "head". *X* = total number of heads.
Then *X* is binomial random variable

Binomial Probability Distribution

(*n* = number of trials, *p* = probability of success.)
The probability there will be exactly *x* successes is:
$P(x) = \binom{n}{x} p^x q^{n-x}$ $(x = 0, 1, 2, \dots , n)$ where $\binom{n}{x}$ = "*n* choose *x*" = $\frac{n!}{x\,!(n-x)!}$

Properties of Binomial Distribution

The following are the properties of binomial distribution.

1. It is a discrete distribution of the occurrences of an event with outcomes–*Success* or *Failure*–of a single trial out of a number of ***n*** trials
2. The trials must be independent of one another.
3. As the number of trials increases, and as ***p*** approaches 0.5 the Binomial distribution approaches the normal distribution.
4. For larger values of ***n*** and for very small value of ***p***, the Binomial distribution approaches the Poisson distribution.

Example: A box contains a large number of screws. The screws are very similar in appearance but are, in fact, of three different types A, B, C which are present in equal numbers. For a given job, only screws of type A are suitable. If 4 screws are chosen at random, find the probability that i. exactly two are suitable, ii. at least two are suitable.

If twenty screws are chosen at random, find the expected value and variance of the number of suitable screws.

Solution:

Screws are present in equal numbers means: $P(A) = P(B) = P(C) = ⅓$
(i.e. Probability of a single trial, $p = ⅓$); $n = 4$ (i) $P(r = 2) = {}^4C_2(⅓)^2(1-⅓)^{4-2} = 6 \times {}^1/_9 \times ({}^2/_3)^2 = {}^8/_{27}$

(ii) P(at least two) $= P(r \geq 2) = 1 - P(r = 1) - P(r = 2) = 1 - {}^4C_0(⅓)^0(⅔)^3 - {}^4C_1(⅓)^1(⅔)^3$
$= 1 - {}^{16}/_{81} - {}^{32}/_{81} = {}^{11}/_{27}$

If $n = 20$, the Expected Value $= np = 20 \text{ x } ⅓ = {}^{20}/_3$ and Variance $= np(1 - p) = 20 \text{ x } ⅓ \text{ x } ⅔ = {}^{40}/_9$

Example: In a factory, 10% of the products are generally found to be defective. In a random sample of 5 products, find the probability that all are defective , b) at least two are defective and c) three are defective.

Solution :

The solution can be derived as follows: $p = 10\% = 0.10$; $q = 1 - p = 1 - 0.10 = 0.90$ and $P(x = r) = {}^nC_r\, p^r (1 - p)^{n-r}$

(a) $P(r = 5) = {}^5C_5(0.1)^5(0.9)^0 = 0.00001$, and $P(r \leq 2) = P(r = 0) + P(r = 1) + P(r = 2)$
$= {}^5C_0(0.1)_0(0.9)^5 + {}^5C_1(0.1)^1(0.9)^4 + {}^5C_2(0.1)^2(0.9)^3 = 0.99114$

(b) $P(r = 3) = {}^5C_3(0.1)^3(0.9)^2 = 0.0081$

7.2 Normal Approximation to the Binomial Distribution

Several of the various statistical distributions are closely related to one another in one way or the other. Hence, many problems can be solved by different methods using different distributions. However, usually one of them tends to be more suitable and convenient than the others. The relationship between the Normal and the Binomial distributions illustrates this important point.

It must be recalled that if a random variable *r* follows the Binomial distribution, then:

$$r \sim B(n, p)$$

and the mean of the distribution is *np*, while the variance is *np(1 – p*). It has been observed that as the sample size *n* gets larger, the Binomial distribution becomes approximately equal to the Normal distribution with mean *np* and variance *np(1 – p).* The approximation is quite accurate so far as $np > 5$ and $n(1 - p) > 5$. Hence, the approximation may not be good enough even if *n* is large so far as *p* is very close to zero or one.

To illustrate this important point, let us solve the following problem using both Binomial and Normal distributions and observe that results are relatively close.

Twenty students take an examination in statistics which is simply graded: *pass* and *fail*. If the probability, *p,* of any individual student passing is 60%, let us find the probability of at least 19 students passing the examination.

From the problem, p = 0.6; 1 – p = 0.4; n = 20.

Binomial Distribution Method

To solve the problem using the Binomial distribution, we have to find the probability of exactly 19 students passing, plus the probability of 20 passing. Since the events are mutually exclusive, the Binomial distribution is allowed. Let *r* represent the number passing. Then, the required probability will be given by:

$$P(r = 19) + P(r = 20) = {}^{20}C_{19} \text{ x } 0.6^{19} \text{ x } .04^{1} + {}^{20}C_{20} \text{ x } 0.6^{20} \text{ x } .04^{0}$$
$$= 16 \text{ x } 0.6^{19} + 0.4 \text{ x } 0.6^{20} = 0.000\,024.$$

Normal Distribution Method

If *x* represents the number of successes in *n* independent trials of an event for which *p* is the probability of success in a single trial, then the variable

$$z = \frac{x - np}{\sqrt{np(1-p)}}$$

has a distribution that approaches the normal with mean zero and standard deviation one as the number of trials *n* tends larger and larger.

Since Binomial Distribution measures discrete probabilities, the ends of the values should be corrected to make the intervals continuous. This is because Normal distribution is a continuous probability. We can now solve the above problem using the Normal distribution.

$$np = 20(0.6) = 12; \quad \sqrt{np(1-p)} = \sqrt{20(0.6)(1-0.6)} = \sqrt{4.8} = 2.19$$

$$P(r \geq 19) = P\left(\frac{r-12}{2.19} \geq \frac{19-12}{2.19}\right) = P(z \geq 3.2) = 0.5 - P(0 \leq z \leq 3.2)$$

$$= 0.5 - 0.499979 = 0.000\,021.$$

EXAMPLE

Example: One percent of the product in a factory is always generally defective. Out of a sample of 10,000 find the probability that (a) less than 120 will be defective (b) between 90 and 120 will be defective (c) only 80 will be defective (d) more than 115 will be defective.

Solution: P = 1% = 0.01; q = 1 – p; n = 10,000 which is very large (i.e. n >30)

= 1 – 0.01; np = 10,000 x 0.01= 100

$$SD = \sqrt{np(1-p)} = \sqrt{10{,}000(0.01)(0.99)} = 9.95$$

Let *x* denote defective products.

(a) $P(x < 120) = P\left(\frac{x-100}{9.95} < \frac{119.5-100}{9.95}\right) = P(z < 1.96) = 0.5 + P(0 < z < 1.96)$

$= 0.5 + 0.475 = 0.975$

(b) $P(90 < x < 120) = P\left(\frac{90.5-100}{9.95} < \frac{x-100}{9.95} < \frac{119.5-100}{9.95}\right) = P(-1 < z < 1.96)$

$= P(-1 < z < 0) + P(0 < z < 1.96) = 0.3413 + 0.475 = 0.8163$

(c) $P(x = 80) = P(79.5 < x < 80.5) = P\left(\frac{79.5-100}{9.95} < \frac{x-100}{9.95} < \frac{80.5-100}{9.95}\right)$

$= P(-2.06 < z < -1.96) = P(-2.06 < z < 0) - P(-1.96 < z < 0) = 0.4803 - 0.475 = 0.0053$

(d) $P(x > 115) = P\left(\frac{x-100}{9.95} > \frac{115.5-100}{9.95}\right) = P(z > 1.56) = 0.5 - P(0 < z < 1.56)$

$= 0.5 - 0.4406 = 0.0554$

Normal Approximation to Proportions

If *x/n* represents the proportion of successes *x,* in *n* independent trials of an event of which p is a proportion of a success in a single trial, then the variable

$$z = \frac{\frac{x}{n} - p}{\sqrt{\{p(1-p)\}/n}}$$

has a distribution that approaches the normal with mean *zero* and standard deviation *one* as the number of trials increases. This is just similar to the Normal approximation to the Binomial distribution. When both the numerator and the denominator of equation 9.3.02 are divided by *n,* we get equation above (ie. z-score for proportion).

As an illustration, let us consider a sample of size 100 of some fruits. If in general the proportions of fruits bought in a day is 5%, let us estimate the probability that the proportion will

(a) exceed 10% and b) lie between 2% and 8%

Solution:

P = 5% or 0.05; 1 – p = 1 – 0.05=0.95 ;

$\sigma = \sqrt{\{p(1-p)\}/n} = \sqrt{\{0.05(1-0.05)\}/100} = 0.022$

Let *x/n* denote any proportion

(a) P(*x/n* > 10%) = P(*x/n* > 0.10) $= \text{P}\left(\frac{x/n - 0.05}{0.022} > \frac{0.10 - 0{,}05}{0{,}022}\right)$

= P(z > 2.27) = 0.5 – P(0 < z < 2.27) = 0.5 – 0.4884 = 0.0116

And P(2% < x/n < 8%) = P(0.02 < *x/n* < 0.08)

$= \text{P}\left(\frac{0.02 - 0.05}{0.022} < \frac{x/n - 0.05}{0.022} < \frac{0.10 - 0{,}05}{0{,}022}\right)$

= P(-1.36 < z < 1.36) = 2{P(0 < z < 1.36) = 2{P(0 < z < 1.36)

= 2(0.4141) = 0.8282

7.3 Poisson Random Variables

The Poisson distribution is a common distribution used to model “count” data:

- Number of telephone calls received per hour
- Number of claims received per day by an insurance company
- Number of accidents per month at an intersection

Poisson Distribution:

Which values can a Poisson r.v. take?

Probability distribution for X (if X is Poisson with mean λ), P(x) = $\frac{\lambda^{x}\, e^{-}}{x!}\lambda$ (for x = 0, 1, 2, …)

Mean of Poisson probability distribution: λ and Variance of Poisson probability distribution: λ

Let X be a binomial random variable with probability distribution b(x; n, p). When $n \to \infty$, $p \to 0$, and $np \xrightarrow{n\to\infty} \mu$ remains constant, $b(x; n, p) \xrightarrow{n\to\infty} p(x;\mu)$.

It has been demonstrated in that the Binomial distribution could be approximated to the Normal distribution under some conditions. However, this approximation does not work well for very small values of *p,* when *np* is less than 5. In these cases, the Binomial may be approximated by the Poisson rather than the Normal distribution. Poisson distribution is used as a model for the number x, of events in a given space or time.

The Poisson distribution is defined by the formula: P(x = r) = $\frac{\lambda^{r} e^{-\lambda}}{r!}$, r=0,1,2,3,4,…………n

Where λ= np [mean of Poisson Distribution]; r = the number of successes; e =constant

Properties of Poisson Distribution

The Poisson distribution is distinguished by the following characteristics.

i. It is a discrete distribution as in the case of Binomial distribution occurs singly, independently and not simultaneously.

ii. It is a limiting form of Binomial Distribution and occurs randomly in space or time.

iii. The events occur at constant rate, mean of events, and variance and mean equal.

iv. It is positively skewed.

v. The standard deviation is the square root of its mean.

vi. As the sample size tends larger, the distribution approximates to the normal distribution. The distribution is proportional to the space or time interval.

Let us solve a couple of examples involving Poisson distribution.

First, let us consider a firm of wholesale fruit distributor who found that on the average, one apple in fifty is bruised on arrival from the growers. If the apples arrive in cartons of 100, calculate the probabilities of a carton having 0, 1, 2, 3, or more than 3 bruised apples.

Solution: The solution can be derived as follows:

Let r be the number of bruised apples.

$$\lambda = np = 100(^1/_{50}) = 2; \qquad P(x = r) = \frac{\lambda^r e^{-\lambda}}{r!}$$

$$P(r = 0) = \frac{2^0 e^{-2}}{2!} = e^{-2}; \quad P(r = 1) = \frac{2^1 e^{-2}}{1!} = 2e^{-2}; \quad P(r = 2) = \frac{2^2 e^{-2}}{2!} = 2e^{-2}$$

$$P(r = 3) = \frac{2^3 e^{-2}}{3!} = {^4/_3}e^{-2}; \qquad P(r > 3) = 1 - P(r \le 3)$$

$$= 1 - P(r = 0) - P(r = 1) - P(r = 2) - P(r = 3)$$

$$= 1 - e^{-2} - 2\,e^{-2} - 2e^{-2} - 4/3\,e^{-2} = 1 - {^{19}/_3}e^{-2}$$

Example: It is known that 0.1% of all people react adversely to certain type of drug. What is the probability that out of a sample of 1,000 people a) none will react to the drug ? b) just one person will react to the drug ? c). more than two will react to the drug ? d) less than three will react to the drug ?

Solution: We derive the solution by the following procedure.

Let r = the number of people who react to the drug;

$$P(x = r) = \frac{\lambda^r e^{-\lambda}}{r!}; \quad \lambda = np = 1000 \times 0.001 = 1$$

$$\text{(a) } P(r = 0) = \frac{1^0 e^{-1}}{0!} = e^{-1}, \qquad \text{(b) } P(r = 1) = \frac{1^1 e^{-1}}{1!} = e^{-1}$$

$$\text{(c) } P(r > 2) = 1 - P(r = 0) - P(x = 1) - P(r = 2); \quad P(r = 2) = \frac{1^2 e^{-1}}{2!} = \frac{e^{-1}}{2}$$

$$= 1 - e^{-1} - e^{-1} - e^{-1}/2 = 1 - {^5/_2}\,e^{-1}$$

$$\text{(d) } P(r < 3) = P(r \le 2) = P(r = 0) + P(r = 1) + P(r = 2)$$

$$= \frac{1^0 e^{-1}}{0!} + \frac{1^1 e^{-1}}{1!} + \frac{1^2 e^{-1}}{2!} = e^{-1} + e^{-1} + e^{-1}/2 = 2\tfrac{1}{2}\,e^{-1}$$

7.4 Binomial an Normal Approximation to the Poisson Distribution

If the probability of a single trial p approaches zero while the number n of trials becomes infinitely large in such a manner that the mean $\lambda = np$ remains fixed, then the Binomial Distribution will approach the Poisson Distribution with mean $\lambda = np$.

This can be illustrated with the following couple of sample problems.

Example: Given that a factory has 100 machines in stock for sale. Five percent of the machines were found faulty. Find the probability that a) None will be faulty, b) two will be faulty, c) at most two will be faulty and d) at least three will be faulty.

Solution: We proceed as following.

$\lambda = np = 100(5\%) = 100(0.05) = 5$, By Poisson's approximation, $P(x = r) = \frac{\lambda^r e^{-\lambda}}{r!}$

(a) $P(r = 0) = \frac{5^0 e^{-5}}{0!} = e^{-5} = 1/e^5 = 0.00674$ (b) $P(r = 2) = \frac{5^2 e^{-5}}{2!} = {}^{25}/_{2}e^5 = 0.084$

(c) $P(r \leq 2) = P(r = 0) + P(r = 1) + P(r = 2) = e^{-5} + \frac{5^1 e^{-5}}{1!} + {}^{25}/_{2}\, e^{-5} = {}^{37}/_{2}e^5$ $=0.1246$

e) $P(r \geq 3) = 1 - P(x \leq 2) = 1 - {}^{37}/_{2}\, e^5 = 0.875$

Example: 10 percent of edible oil produced by a company is defective. In a random sample of fifty gallons, we can find the probability that a) none is defective; b) three are defective; and c) at least two are defective;

Solution: Since $n = 50$ is large and $p = 0.1$ is small, we have to employ Poisson Distribution. $np = 50(0.1) = 5$; $P(x = r) = \frac{\lambda^r e^{-\lambda}}{r!}$, (a) $P(r = 0) = \frac{5^0 e^{-5}}{5!} = \frac{e^{-5}}{5!}$ $=0.000056$ (b) $P(r = 3) = \frac{5^3 e^{-5}}{3!} = 0.1404$ (c) P(at least two defectives)=1– $P(r=0)-P(r=1) = 1 - e^{-5} - \frac{5^1 e^{-5}}{1!} = 1 - 6e^{-5} = 0.9595$

Example: Assume that cars pass under a bridge at a rate of 100 per hour and that a Poisson distribution is appropriate. (a) What is the probability that during a 3-minute period no cars will pass under the bridge? (b) What time interval is such that the probability is at least 0.25 that no car will pass under the bridge during that interval?

Solution:

Rate of 100 per hour means rate of $\lambda = \frac{100}{60} = \frac{5}{3}$ per minute.

(a) For a 3-minute period, $\lambda = \frac{5}{3} \times 3 = 5$. Let X denote number of cars passing.

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!} \Rightarrow P(X = 0) = \frac{5^0 e^{-5}}{0!} = e^{-5} = 0.00674$$

(b) Let the required time interval be t. Probability of at least 0.25 for no car passing

$\Rightarrow P(X = 0) = \frac{t^0 e^{-t}}{0!} = e^{-t} \geq 0.25$. Taking natural logs on both sides gives

$-t \geq \ln 0.25 \Rightarrow -t \geq -1.386 \Rightarrow t \geq 1.4$ minutes

Normal Approximation to the Poisson Distribution

If n, the sample size is very large and p, the probability of a single trial is small and x the number of successes, then the variable

$$z = \frac{x - np}{\sqrt{np}}$$

has approximately a normal distribution with mean zero and standard deviation one.

As an illustration, let us consider the following example.

There are 10,000 tins of milk in a firm to be tested of quality. The selection of defective ones follows Poisson distribution. Let us find the probability that at least 190 are defective.

Solution

The problem can be solve as follows:

n = 10,000 is very large; p = 2% = 0.02; np = 10,000(0.02) = 200 ;

$\sigma = \sqrt{np} = \sqrt{200} = 14.142$

Let x denote any defective number of tins of milk.

$P(x \geq 190) = P\left(\frac{x - 200}{14.142} \geq \frac{190.5 - 200}{14.142}\right) = P(z \geq -0.707) = 0.5 + P(0 \leq z \leq -0.707) = 0.5 + 0.2612 = 0.7612$

EXAMPLES

Example. A machine fills millet flour in nominally 500-gram bags. The actual weight of the filled bags varies, being approximately normally distributed with standard deviation 10 grams.

(a) Find the mean weight of bags filled by the machine if 15% filled bags are underweight.

(b) Calculate the proportion of bags whose weight is between 495 grams and 535 grams.

(c) Bags weighing less than 500 grams are sold at a loss of Rs.3,000. Calculate the the loss associated with the sale of 150 bags.

(d) If the mean weight of filled bags is adjusted to 521.2 grams and the standard deviation remains unchanged, what percentage of bags would be sold at a loss?

Solution

Let ***x*** represent the weight of any filled bag and μ be the mean weight filled by the machine

$\bar{x}$ = 500 grams; σ = 10 gram

(a) P(x < 500) = 0.15, $P\left(\frac{\bar{x}-u}{10} < \frac{500-u}{10}\right) = P\left(z < \frac{500-u}{10}\right)$ = 0.15

$P\left(0 < z < \frac{500-u}{10}\right)$ = 0.5 – 0.15 = 0.35 by using normal table

= $\frac{500-u}{10}$ = - 1.04 and μ = 500 + 10(1.04) = 510.4 grams

(b) P(495 < x < 535) = $P\left(\frac{495-510.4}{10} < \frac{x-510.4}{10} < \frac{535-510.4}{10}\right)$

= P(- 1.54 < z < 2.46) = 0.4382 + 0.4931 = 0.9313

(c) P(x < 500) = $\left(\frac{x-510.4}{10} < \frac{500-510.4}{10}\right)$=P(z< -1.04)= 0.5–P(-1.04< z< 0)=0.5 – 0.35 = 0.15

: . The total number of bags associated with loss = 0.15 x 150= 22.5

: . Total loss = 22.5 x Rs.3,000 = Rs. 67,500

(d) P(x < 500) = $P\left(\frac{x-521.2}{10} < \frac{500-521.2}{10}\right)$ = P(z < -2.12) = 0.5 – P(-2.12 < z < 0)

= 0.5 – 0.4830 = 0.017, The required percentage is 1.7%

Example 2: A computer firm orders 20 personal computers (PCs). After shipment, the manufacturer detects that 5 of the PCs are faulty. If 5 PCs are selected at random from the batch of 20, what is the probability of obtaining at least 2 defective PCs?

Solution

Probability of a single trial $p = {}^5/_{20} = ¼$. Let x be number of defective PCs
This a binomial distribution, with $n = 5$, $p = ¼$.

$P(x \geq 2) = P(x = 2) + P(x = 3) + P(x = 4) + P(x = 5) = 1 - P(x = 0) – P(x = 1)$
$= 1 - {}^5C_0 (¼)^0 (¾)^5 – {}^5C_1 (¼)^1(¾)^4$
$= 1-[(¾)^5 + {}^5/_4(¾)^4] = 1- (¾)^4[{}^3/_4 + {}^5/_4] = 0.367$

Example 3. If a typist makes an average of two errors per page of a book, use the Poisson distribution to find the probability that (a) exactly four errors will be found on a page, (b) at least two errors will be found on a given page.

Solution

The mean of Poisson distribution $\lambda = 2$, Let x represent any number of errors made per page.

(a) $P(x = 4) = \frac{2^4 e^{-2}}{4!} = \frac{2}{3}e^{-2}$ (b) $P(x \geq 2) = 1 – [P(x = 0) + P(x = 1)]$

$= 1 - \frac{2^0 e^{-2}}{0!} + \frac{2^1 e^{-2}}{1!} = 1 - 3e^{-2}$

Example 4. The lifetime of batteries produced by a company are normally distributed with mean 110 hours and variance σ2. The probability that a battery has a lifetime more than 113 hours is 0.3821. (a) Find the variance σ^2.
(b) Use the variance in (a) to determine the probability a battery will last between 90 and 102 hours.

Solution

Let x denote the lifetime of any battery (a) $P(x > 113) = 0.3821$

$$P\left(\frac{x-110}{\sigma} < \frac{113-110}{\sigma}\right) = 0.3821$$

$P(z > 3/\sigma) = 0.3821$, $P(0 < z < 3/\sigma) = 0.5 – 0.3821 = 0.1179$

: . $3/\sigma = 0.3$ [normal table value for 0.1179 is 0.3], $\sigma = 10$.
Hence, the variance is $\sigma^2 = 10^2$ or 100 hours.

(b) $P(90 < x < 102) = P\left(\frac{90-110}{10} < \frac{x-110}{10} < \frac{102-110}{10}\right)$
$= P(- 2.0 < z < - 0.8) = P(-2.0 < Z < 0) – (-0.8<Z<0)$
$= 0.4772 – 0.2881 = 0.1891$

Activity: A call center averages 10 calls per hour. Assume *X* (the number of calls in an hour) follows a Poisson distribution. What is the probability that the call center receives exactly 3 calls in the next hour?

7.5 Hypergeometric Distribution

The simplest way to view the distinction between the binomial distribution of Section 5.2 and the hypergeometric distribution is to note the way the sampling is done. The types of applications for the hypergeometric are very similar to those for the binomial distribution. We are interested in computing probabilities for the number of observations that fall into a particular category. But in the case of the binomial distribution, independence among trials is required. As a result, if that distribution is applied to, say, sampling from a lot of items (deck of cards, batch of production items), the sampling must be done with replacement of each item after it is observed. On the other hand, the hypergeometric distribution does not require independence and is based on sampling done without replacement. Applications for the hypergeometric distribution are found in many areas, with heavy use in acceptance sampling, electronic testing, and quality assurance. Obviously, in many of these fields, testing is done at the expense of the item being tested. That is, the item is destroyed and hence cannot be replaced in the sample. Thus, sampling without replacement is necessary. A simple example with playing cards will serve as our first illustration. If we wish to find the probability of observing 3 red cards in 5 draws from an ordinary deck of 52 playing cards, the binomial distribution does not apply unless each card is replaced and the deck reshuffled before the next draw is made. To solve the problem of sampling without replacement, let us restate the problem. If 5 cards are drawn at random, we are interested in the probability of selecting 3 red cards from the 26 available in the deck and 2 black cards from the 26 available in the deck. There are ${}^{26}C_3$ ways of selecting 3 red cards, and for each of these ways we can choose 2 black cards in ${}^{26}C_2$ ways. Therefore, the total number of ways to select 3 red and 2 black cards in 5 draws is the product (${}^{26}C_3$) (${}^{26}C_2$). The total number of ways to select any 5 cards from the 52 that are available is ${}^{52}C_5$. Hence, the probability of selecting 5 cards without replacement of which 3 are red and 2 are black is given by

${}^{26}C_3$ X ${}^{26}C_2$ /${}^{52}C_5$ = $(\frac{26!}{3!\,23!}\ \boldsymbol{X}\ \frac{26!}{2!\,24!})$ / $\frac{52!}{5!\,47}$ =(26! X 26! X 5! X 47!)/(3! X 23! X 2! X 24! X 52!) = 0.3251

In general, we are interested in the probability of selecting x successes from the k items labeled successes and n − x failures from the N − k items labeled failures when a random sample of size n is selected from N items. This is known as a

hypergeometric experiment, that is, one that possesses the following two properties:

1. A random sample of size n is selected without replacement from N items.

2. Of the N items, k may be classified as successes and N − k are classified as failures.

The number X of successes of a hypergeometric experiment is called a hypergeometric random variable. Accordingly, the probability distribution of the hypergeometric variable is called the hypergeometric distribution, and its values are denoted by h(x; N, n, k), since they depend on the number of successes k in the set N from which we select n items. Hypergeometric Distribution in Acceptance Sampling Like the binomial distribution, the hypergeometric distribution finds applications in acceptance sampling, where lots of materials or parts are sampled in order to determine whether or not the entire lot is accepted.

Example: A particular part that is used as an injection device is sold in lots of 10. The producer deems a lot acceptable if no more than one defective is in the lot. A sampling plan involves random sampling and testing 3 of the parts out of 10. If none of the 3 is defective, the lot is accepted. Comment on the utility of this plan.
Solution: Let us assume that the lot is truly unacceptable (i.e., that 2 out of 10 parts are defective). The probability that the sampling plan finds the lot acceptable is

$P(X = 0) = {}^{2}C_{0} \text{ X } {}^{8}C_{3} / {}^{10}C_{3} = 0.467$

Thus, if the lot is truly unacceptable, with 2 defective parts, this sampling plan will allow acceptance roughly 47% of the time. As a result, this plan should be considered faulty. Let us now generalize in order to find a formula for h(x; N, n, k). The total number of samples of size n chosen from N items is ${}^{N}C_{n}$. These samples are assumed to be equally likely. There are ${}^{k}C_{x}$ ways of selecting x successes from the k that are available, and for each of these ways we can choose the n − x failures in ${}^{N-k}C_{n-x}$ ways. Thus, the total number of favorable samples among the ${}^{N}C_{n}$ possible samples is given by ${}^{k}C_{x} \text{ X } {}^{N-k}C_{n-x} / {}^{N}C_{n}$. Hence, we have the following definition

Hypergeometric Distribution The probability distribution of the hypergeometric random variable X, the number of successes in a random sample of size n selected from N items of which k are labeled success and N − k labeled failure, is h(x; N, n, k) = $({}^{k}C_{x})({}^{N-k}C_{n-x})/{}^{N}C_{n}$, $\max\{0, n - (N - k)\} \le x \le \min\{n, k\}$. The range of x can be determined by the three binomial coefficients in the definition, where x and n−x are no more than k and N −k, respectively, and both of them cannot be less than 0. Usually, when both k (the number of successes) and N − k (the number of failures) are larger than the sample size n, the range of a hypergeometric random variable will be x = 0, 1,...,n.

Example: Lots of 40 components each are deemed unacceptable if they contain 3 or more defectives. The procedure for sampling a lot is to select 5 components at random and to reject the lot if a defective is found. What is the probability that exactly 1 defective is found in the sample if there are 3 defectives in the entire lot?

Solution: Using the hypergeometric distribution with n = 5, N = 40, k = 3, and x = 1, we find the probability of obtaining 1 defective to be

$h(1; 40, 5, 3) = ({}^{3}C_{1})({}^{37}C_{4})/{}^{40}C_{5} = (3!/1!2!)(37!/4!33!) / (40!/5!35!)= 0.3011$

Once again, this plan is not desirable since it detects a bad lot (3 defectives) only about 30% of the time.

7.6 Negative Binomial Distribution

Let us consider an experiment where the properties are the same as those listed for a binomial experiment, with the exception that the trials will be repeated until a fixed number of successes occur. Therefore, instead of the probability of x successes in n trials, where n is fixed, we are now interested in the probability that the kth success occurs on the xth trial. Experiments of this kind are called negative binomial experiments.

Consider the use of a drug that is known to be effective in 60% of the cases where it is used. The drug will be considered a success if it is effective in bringing some degree of relief to the patient. We are interested in finding the probability that the fifth patient to experience relief is the seventh patient to receive the drug during a given week. Designating a success by S and a failure by F, a possible order of achieving the desired result is SFSSSFS, which occurs with probability $(0.6)(0.4)(0.6)(0.6)(0.6)(0.4)(0.6) = (0.6)^5(0.4)^2$. We could list all possible orders by rearranging the F's and S's except for the last outcome, which must be the fifth success. The total number of possible orders is equal to the number of partitions of the first six trials into two groups with 2 failures assigned to the one group and 4 successes assigned to the other group. This can be done in ${}^{6}C_{4} = 15$ mutually exclusive ways. Hence, if X represents the outcome on which the fifth success occurs, then $P(X = 7) = ({}^{6}C_{4})\ (0.6)^5(0.4)^2 = 0.1866$.

The number X of trials required to produce k successes in a negative binomial experiment is called a negative binomial random variable, and its probability distribution is called the negative binomial distribution. Since its probabilities depend on the number of successes desired and the probability of a success on a given trial, we shall denote them by $b*(x; k, p)$. To obtain the general formula for $b*(x; k, p)$, consider the probability of a success on the xth trial preceded by $k - 1$ successes and $x - k$ failures in some specified order. Since the trials are independent, we can multiply all the probabilities corresponding to each desired outcome. Each success occurs with probability p and each failure with probability $q = 1 - p$. Therefore, the probability for the specified order ending in success is

$$p^{k-1}q^{x-k}p = p^k q^{x-k}.$$

The total number of sample points in the experiment ending in a success, after the occurrence of $k-1$ successes and $x-k$ failures in any order, is equal to the number of partitions of $x-1$ trials into two groups with $k-1$ successes corresponding to one group and $x-k$ failures corresponding to the other group. This number is specified by the term ${}^{x-1}C_{k-1}$, each mutually exclusive and occurring with equal probability $p^k q^{x-k}$. We obtain the general formula by multiplying $p^k q^{x-k}$ by ${}^{x-1}C_{k-1}$.

If repeated independent trials can result in a success with probability p and a failure with probability $q = 1 - p$, then the probability distribution of the random variable X, the number of the trial on which the kth success occurs, is

$$b*(x; k, p) = ({}^{x-1}C_{k-1})\, p^k q^{x-k},\ x = k, k + 1, k + 2, \ldots\ldots\ldots$$

Example : In an National Football Association championship series, the team that wins four games out of seven is the winner. Suppose that teams A and B face each other in the championship games and that team A has probability 0.55 of winning a game over team B.

(a) What is the probability that team A will win the series in 6 games?
(b) What is the probability that team A will win the series?
(c) If teams A and B were facing each other in a regional playoff series, which is decided by winning three out of five games, what is the probability that team A would win the series?

Solution: (a) $b*(6; 4, 0.55) = ({}^5C_3)(0.55)^4(1 - 0.55)^{6-4} = 0.1853$

(a) P(team A wins the series) is $b*(4; 4, 0.55) + b*(5; 4, 0.55) + b*(6; 4, 0.55) + b*(7; 4, 0.55)$

$= 0.0915 + 0.1647 + 0.1853 + 0.1668 = 0.6083.$

(b) P(team A wins the playoff) is $b*(3; 3, 0.55) + b*(4; 3, 0.55) + b*(5; 3, 0.55)$
$= 0.1664 + 0.2246 + 0.2021 = 0.5931.$

The negative binomial distribution derives its name from the fact that each term in the expansion of $p^k(1 - q)^{-k}$ corresponds to the values of $b*(x; k, p)$ for $x = k, k + 1, k + 2, \ldots$. If we consider the special case of the negative binomial distribution where $k = 1$, we have a probability distribution for the number of trials required for a single success. An example would be the tossing of a coin until a head occurs. We might be interested in the probability that the first head occurs on the fourth toss. The negative binomial distribution reduces to the form

$$b*(x; 1, p) = pqx-1, x = 1, 2, 3, \ldots\ldots\ldots\ldots$$

Since the successive terms constitute a geometric progression, it is customary to refer to this special case as the geometric distribution and denote its values by g(x; p).

7.7 Geometric Distribution

If repeated independent trials can result in a success with probability p and a failure with probability $q = 1 - p$, then the probability distribution of the random variable X, the number of the trial on which the first success occurs, is

$$g(x; p) = pq^{x-1}, x = 1, 2, 3, \ldots\ldots\ldots\ldots$$

Example: For a certain manufacturing process, it is known that, on the average, 1 in every 100 items is defective. What is the probability that the fifth item inspected is the first defective item found?

Solution: Using the geometric distribution with $x = 5$ and $p = 0.01$, we have
$g(5; 0.01) = (0.01)(0.99)^4 = 0.0096$.

Example: At a “busy time,” a telephone exchange is very near capacity, so callers have difficulty placing their calls. It may be of interest to know the number of attempts necessary in order to make a connection. Suppose that we let $p = 0.05$ be the probability of a connection during a busy time. We are interested in knowing the probability that 5 attempts are necessary for a successful call.

Solution: Using the geometric distribution with x = 5 and p = 0.05 yields

$$P(X = x) = g(5; 0.05) = (0.05)(0.95)^4 = 0.041.$$

Quite often, in applications dealing with the geometric distribution, the mean and variance are important. For example, the expected number of calls necessary to make a connection is quite important.

The mean and variance of a random variable following the geometric distribution are

$$\mu = 1/ p \text{ and } \sigma^2 = (1 – p)/ p^2$$

Applications of Negative Binomial and Geometric Distributions

Areas of application for the negative binomial and geometric distributions become obvious when one focuses on the examples in this section and the exercises devoted to these distributions. In the case of the geometric distribution, depicts a situation where engineers or managers are attempting to determine how inefficient a telephone exchange system is during busy times. Clearly, in this case, trials occurring prior to a success represent a cost. If there is a high probability of several attempts being required prior to making a connection, then plans should be made to redesign the system. Applications of the negative binomial distribution are similar in nature. Suppose attempts are costly in some sense and are occurring in sequence. A high probability of needing a "large" number of attempts to experience a fixed number of successes is not beneficial to the scientist or engineer.

7.8 Normal Distribution

The most important continuous probability distribution in the entire field of statistics is the normal distribution. Its graph, called the normal curve, is the bell-shaped curve of Figure, which approximately describes many phenomena that occur in nature, industry, and research. For example, physical measurements in areas such as meteorological experiments, rainfall studies, and measurements of manufactured parts are often more than adequately explained with a normal distribution. In addition, errors in scientific measurements are extremely well approximated by a normal distribution. In 1733, Abraham DeMoivre developed the mathematical equation of the normal curve. It provided a basis from which much of the theory of inductive statistics is founded. The normal distribution is often referred to as the Gaussian distribution, in honor of Karl Friedrich Gauss(1777–1855), who also derived its equation from a study of errors in repeated measurements of the same quantity. A continuous random variable X having the bell-shaped distribution is called a normal random variable. The mathematical equation for the probability distribution of the normal variable depends on the two parameters μ and σ, its mean and standard deviation, respectively. Hence, we denote the values of the density of X by n(x; μ, σ).

The density of the normal random variable X, with mean μ and variance σ2, is
n(x; μ, σ) = $e- 1/2\sigma^2\ (x-\mu)^2\ /\sqrt{}2\pi\sigma$, $-\infty < x < \infty$, where $\pi = 3.14159$ ... and e = 2.71828 ... .

Once μ and σ are specified, the normal curve is completely determined. For example, if $\mu = 50$ and $\sigma = 5$, then the ordinates n(x; 50, 5) can be computed for various values of x and the curve drawn. we have sketched two normal curves having the same standard deviation but different means. The two curves are identical in form but are centered at different positions along the horizontal axis.
Based on inspection of Figures and examination of the first and second derivatives of n(x; μ, σ), we list the following properties of the normal curve:

1. The mode, which is the point on the horizontal axis where the curve is a maximum, occurs at $x = \mu$.

2. The curve is symmetric about a vertical axis through the mean μ.

3. The curve has its points of inflection at $x = \mu \pm \sigma$; it is concave downward if $\mu - \sigma < X < \mu - \sigma$ and it is concave upward otherwise.

4. The normal curve approaches the horizontal axis asymptotically as we proceed in either direction away from the mean.

5. The total area under the curve and above the horizontal axis is equal to 1.

The distribution of a normal random variable with mean 0 and variance 1 is called a standard normal distribution.

Example: Given a standard normal distribution, find the area under the curve that lies (a) to the right of $z = 1.84$ and (b) between $z = -1.97$ and $z = 0.86$.

Solution: (a) The area (a) to the right of $z = 1.84$ is equal to 1 minus the area in Table A. to the left of $z = 1.84$, namely, $1 - 0.9671 = 0.0329$.
(b) The area) between $z = -1.97$ and $z = 0.86$ is equal to the area to the left of $z = 0.86$ minus the area to the left of $z = -1.97$. From Table A. we find the desired area to be $0.8051 - 0.0244 = 0.7807$

Example: Given a standard normal distribution, find the value of k such that (a) $P(Z>k)=0.3015$ and (b) $P(k<Z< -0.18) = 0.4197$.

Solution: Distributions and the desired areas are shown. (a) we see that the k value leaving an area of 0.3015 to the right must then leave an area of 0.6985 to the left. From Table A. it follows that k = 0.52. (b) From Table A. we note that the total area to the left of −0.18 is equal to 0.4286. We see that the area between k and −0.18 is 0.4197, so the area to the left of k must be 0.4286 − 0.4197 = 0.0089. Hence, from Table A.3, we have k = −2.37.

Example: A certain type of storage battery lasts, on average, 3.0 years with a standard deviation of 0.5 year. Assuming that battery life is normally distributed, find the probability that a given battery will last less than 2.3 years.

Solution: First construct a diagram, showing the given distribution of battery lives and the desired area. To find $P(X < 2.3)$, we need to evaluate the area under the normal curve to the left of 2.3. This is accomplished by finding the area to the left of the corresponding z value. Hence, we find that $z = (2.3 – 3)/ 0.5 = -1.4$, and then, using Table A., we have $P(X < 2.3) = P(Z < -1.4) = 0.0808$.

Example: An electrical firm manufactures light bulbs that have a life, before burn-out, that is normally distributed with mean equal to 800 hours and a standard deviation of 40 hours. Find the probability that a bulb burns between 778 and 834 hours.

Solution: The distribution of light bulb life is illustrated. The z values corresponding to $x_1 = 778$ and $x_2 = 834$ are $z_1 = (778 – 800)/ 40 = -0.55$ and $z_2 = (834 – 800)/ 40 = 0.85$. Hence,

$P(778 <X< 834) = P(-0.55 <Z< 0.85) = P(Z < 0.85) - P(Z < -0.55) = 0.8023 - 0.2912 = 0.5111$.

Example: In an industrial process, the diameter of a ball bearing is an important measurement. The buyer sets specifications for the diameter to be 3.0 ± 0.01 cm. The implication is that no part falling outside these specifications will be accepted. It is known that in the process the diameter of a ball bearing has a normal distribution with mean $\mu = 3.0$ and standard deviation $\sigma = 0.005$. On average, how many manufactured ball bearings will be scrapped?

Solution: The distribution of diameters is illustrated. The values corresponding to the specification limits are $x_1 = 2.99$ and $x_2 = 3.01$. The corresponding z values are $z_1 = (2.99 - 3.0)/ 0.005 = -2.0$ and $z_2 = (3.01 - 3.0)/0.005 = +2.0$.
Hence, $P(2.99 < X < 3.01) = P(-2.0 < Z < 2.0)$.

From Table, $P(Z < -2.0) = 0.0228$. Due to symmetry of the normal distribution, we find that $P(Z < -2.0) + P(Z > 2.0) = 2(0.0228) = 0.0456$. As a result, it is anticipated that, on average, 4.56% of manufactured ball bearings will be scrapped.

Example: An electrical firm manufactures light bulbs that have a life, before burn-out, that is normally distributed with mean equal to 800 hours and a standard deviation of 40 hours. Find the probability that a bulb burns between 778 and 834 hours.

Solution: The distribution of light bulb life is illustrated. The z values corresponding to $x_1 = 778$ and $x_2 = 834$ are $z_1 = (778 – 800)/ 40 = -0.55$ and $z_2 = (834 – 800)/ 40 = 0.85$.

Hence, $P(778 < X < 834) = P(-0.55 < Z < 0.85) = P(Z < 0.85) - P(Z < -0.55) = 0.8023 - 0.2912 = 0.5111$.

Example: In an industrial process, the diameter of a ball bearing is an important measurement. The buyer sets specifications for the diameter to be 3.0 ± 0.01 cm. The implication is that no part falling outside these specifications will be accepted. It is known that in the process the diameter of a ball bearing has a normal distribution with mean $\mu = 3.0$ and standard deviation $\sigma = 0.005$. On average, how many manufactured ball bearings will be scrapped?

Solution: The distribution of diameters is illustrated. The values corresponding to the specification limits are $x_1 = 2.99$ and $x_2 = 3.01$. The corresponding z values are $z_1 = (2.99 - 3.0)/ 0.005 = -2.0$ and $z_2 = (3.01 - 3.0)/ 0.005 = +2.0$.

Hence, $P(2.99 < X < 3.01) = P(-2.0 < Z < 2.0)$. From Table, $P(Z < -2.0) = 0.0228$.

Due to symmetry of the normal distribution, we find that $P(Z < -2.0) + P(Z > 2.0) = 2(0.0228) = 0.0456$.

As a result, it is anticipated that, on average, 4.56% of manufactured ball bearings will be scrapped.

7.9 SELF ASSESSMENT QUESTIONS

Q.1 An employee is selected from a staff of 10 to supervise a certain project by selecting a tag at random from a box containing 10 tags numbered from 1 to 10. Find the formula for the probability distribution of X representing the number on the tag that is drawn. What is the probability that the number drawn is less than 4?

Q.2 In a certain city district, the need for money to buy drugs is stated as the reason for 75% of all thefts. Find the probability that among the next 5 theft cases reported in this district, (a) exactly 2 resulted from the need for money to buy drugs; (b) at most 3 resulted from the need for money to buy drugs.

Q.3 According to Chemical Engineering Progress (November 1990), approximately 30% of all pipework failures in chemical plants are caused by operator error. (a) What is the probability that out of the next 20 pipework failures at least 10 are due to operator error? (b) What is the probability that no more than 4 out of 20 such failures are due to operator error? (c) Suppose, for a particular plant, that out of the random sample of 20 such failures, exactly 5 are due to operator error. Do you feel that the 30% figure stated above applies to this plant? Comment.

Q.4 According to a survey by the Administrative Management Society, one-half of U.S. companies give employees 4 weeks of vacation after they have been with the company for 15 years. Find the probability that among 6 companies surveyed at random, the number that give employees 4 weeks of vacation after 15 years of employment is (a) anywhere from 2 to 5; (b) fewer than 3.

Q.5 A homeowner plants 6 bulbs selected at random from a box containing 5 tulip bulbs and 4 daffodil bulbs. What is the probability that he planted 2 daffodil bulbs and 4 tulip bulbs?

Q.6 To avoid detection at customs, a traveler places 6 narcotic tablets in a bottle containing 9 vitamin tablets that are similar in appearance. If the customs official selects 3 of the tablets at random for analysis, what is the probability that the traveler will be arrested for illegal possession of narcotics?

Q.7 A random committee of size 3 is selected from 4 doctors and 2 nurses. Write a formula for the probability distribution of the random variable X representing the number of doctors on the committee. Find $P(2 \leq X \leq 3)$.

Q.8 From a lot of 10 missiles, 4 are selected at random and fired. If the lot contains 3 defective missiles that will not fire, what is the probability that (a) all 4 will fire? (b) at most 2 will not fire?

Q.9 If 7 cards are dealt from an ordinary deck of 52 playing cards, what is the probability that (a) exactly 2 of them will be face cards? (b) at least 1 of them will be a queen?

Q.10 The probability that a person living in a certain city owns a dog is estimated to be 0.3. Find the probability that the tenth person randomly interviewed in that city is the fifth one to own a dog.

Q.11 Find the probability that a person flipping a coin gets (a) the third head on the seventh flip; (b) the first head on the fourth flip.

Q.12 Three people toss a fair coin and the odd one pays for coffee. If the coins all turn up the same, they are tossed again. Find the probability that fewer than 4 tosses are needed.

Q.13 A scientist inoculates mice, one at a time, with a disease germ until he finds 2 that have contracted the disease. If the probability of contracting the disease is 1/6, what is the probability that 8 mice are required?

Q.14 An inventory study determines that, on average, demands for a particular item at a warehouse are made 5 times per day. What is the probability that on a given day this item is requested (a) more than 5 times? (b) not at all?

Q.15 On average, 3 traffic accidents per month occur at a certain intersection. What is the probability that in any given month at this intersection (a) exactly 5 accidents will occur? (b) fewer than 3 accidents will occur? (c) at least 2 accidents will occur.

Q.16 On average, a textbook author makes two word-processing errors per page on the first draft of her textbook. What is the probability that on the next page she will make (a) 4 or more errors? (b) no errors?

Q.17 A certain area of the eastern United States is, on average, hit by 6 hurricanes a year. Find the probability that each year that area will be hit by (a) fewer than 4 hurricanes; (b) anywhere from 6 to 8 hurricanes.

Q.18 Suppose the probability that any given person will believe a tale about the transgressions of a famous actress is 0.8. What is the probability that (a) the sixth person to hear this tale is the fourth one to believe it? (b) the third person to hear this tale is the first one to believe it.

Q.19 The average number of field mice per acre in a 5-acre wheat field is estimated to be 12. Find the probability that fewer than 7 field mice are found (a) on a given acre; (b) on 2 of the next 3 acres inspected.

Q.20 The number of customers arriving per hour at a certain automobile service facility is assumed to follow a Poisson distribution with mean $\lambda = 7$. (a) Compute the probability that more than 10 customers will arrive in a 2-hour period. (b) What is the mean number of arrivals during a 2-hour period?

Q.21 The probability that a student at a local high school fails the screening test for scoliosis (curvature of the spine) is known to be 0.004. Of the next 1875 students at the school who are screened for scoliosis, find the probability that (a) fewer than 5 fail the test; (b) 8, 9, or 10 fail the test. What is the mean number of students who fail the test?

Q.22 The probability that a person will die when he or she contracts a virus infection is 0.001. Of the next 4000 people infected, what is the mean number who will die?

Q.23 The potential buyer of a particular engine requires (among other things) that the engine successfully start 10 consecutive times. Suppose the probability of a successful start is 0.990. Let us assume that the outcomes of attempted starts are independent. (a) What is the probability that the engine is accepted after only 10 starts? (b) What is the probability that 12 attempted starts are made during the acceptance process?

Q.24 A couple decides to continue to have children until they have two males. Assuming that P(male) = 0.5, what is the probability that their second male is their fourth child?

Q.25 The manufacturer of a tricycle for children has received complaints about defective brakes in the product. According to the design of the product and considerable preliminary testing, it had been determined that the probability of the kind of defect in the complaint was 1 in 10,000 (i.e., 0.0001). After a thorough investigation of the complaints, it was determined that during a certain period of time, 200 products were randomly chosen from production and 5 had defective brakes. (a) Comment on the “1 in 10,000” claim by the manufacturer. Use a probabilistic argument. Use the binomial distribution for your calculations. (b) Repeat part (a) using the Poisson approximation?

Q.26 A soft-drink machine is regulated so that it discharges an average of 200 milliliters per cup. If the amount of drink is normally distributed with a standard deviation equal to 15 milliliters, (a) what fraction of the cups will contain more than 224 milliliters? (b) what is the probability that a cup contains between 191 and 209 milliliters? (c) how many cups will probably overflow if 230- milliliter cups are used for the next 1000 drinks? (d) below what value do we get the smallest 25% of the drinks?

Q.27 The loaves of rye bread distributed to local stores by a certain bakery have an average length of 30 centimeters and a standard deviation of 2 centimeters. Assuming that the lengths are normally distributed, what percentage of the loaves are (a) longer than 31.7 centimeters? (b) between 29.3 and 33.5 centimeters in length? (c) shorter than 25.5 centimeters?

Q.28 A research scientist reports that mice will live an average of 40 months when their diets are sharply restricted and then enriched with vitamins and proteins. Assuming that the lifetimes of such mice are normally distributed with a standard deviation of 6.3 months, find the probability that a given mouse will live (a) more than 32 months; (b) less than 28 months; (c) between 37 and 49 months.

Q.29 The finished inside diameter of a piston ring is normally distributed with a mean of 10 centimeters and a standard deviation of 0.03 centimeter. (a) What proportion of rings will have inside diameters exceeding 10.075 centimeters? (b) What is the probability that a piston ring will have an inside diameter between 9.97 and 10.03 centimeters? (c) Below what value of inside diameter will 15% of the piston rings fall?

Q.30 A lawyer commutes daily from his suburban home to his main city office. The average time for a one-way trip is 24 minutes, with a standard deviation of 3.8 minutes. Assume the distribution of trip times to be normally distributed. (a) What is the probability that a trip will take at least 1/2 hour? (b) If the office opens at 9:00 A.M. and the lawyer leaves his house at 8:45 A.M. daily, what percentage of the time is he late for work?

SUGGESTED READINGS

Bluman, A.G. (2004). *Elementary Statistics. A Step by Step Approach*. 5th Edition. McGraw-Hill Companies Incorporated. London.

Chaudhary, S.M. & Kamal, S. (2017). *Introduction to Statistical Theory Part-I*. 8th Edition. Ilmi Kitab Khana. Lahore.

Chaudhary, S.M. & Kamal, S. (2017). *Introduction to Statistical Theory Part-II*. 8th Edition. Ilmi Kitab Khana. Lahore.

Daniel, W.W. (1995). *Biostatistics: A foundation for Analysis in Health sciences*. Sixth Edition. John Wiley and sons Incorporated. USA.

Harper, W.M. (1991). *Statistics*. Sixth Edition. Pitman Publishing, Longman Group, United Kingdom.

Hoel, P.G. (1976). *Elementary Statistics*. 4th Edition. John Wiley and Sons Incorporated, NewYork.

Kiani, G. H., & Akhtar, M. S. (2012). *Basic statistics*, Majeed Book Depot.

Khan, A. A., Mirza, S. H., Ahmad, M. I., Baig, I. & Yaqoob, M. (2011). *Business Statistics*, Qureshi Brothers Publishers.

UNIT 08

SIMPLE LINEAR CORRELATION AND REGRESSION

Written By: Dr. Zahid Iqbal
Reviewed By: Dr. Muhammad Ilyas

CONTENTS

Pages

Introduction

The term regression was first used in 1877 by Francis Galton. He made a study that showed that the height of children born to tall parents tends to move back or regress towards the mean height of the population. He coined the word regression as the name of the general process of predicting one variable (the height of the children) from another (the height of the parents). Later, the term multiple regression came into existence by which several variables are used to predict another.

Objectives

After studying this unit, you will be able to;

- Learn about the Pearson Product-Moment Correlation Coefficient (r)
- Learn about the uses and abuses of correlation.
- Learn how to calculate and interpret r.
- Identify the direction and strength of a linear correlation between two factors.
- Interpret the Pearson correlation coefficient and the coefficient of determination, and test for significance.
- Identify and explain three assumptions and three limitations for evaluating a correlation coefficient.
- Distinguish between a predictor variable and a criterion variable.
- Learn the essential elements of simple regression analysis.
- Learn how to interpret the results of simple regression.

8.1 Correlation

How can we explore the relationship between two quantitative variables? Graphically, we can construct a scatterplot. Numerically, we can calculate a correlation coefficient and a regression equation.

The Pearson correlation coefficient, *r*, measures *the degree of association* , ***strength*** *and the* **direction** *of a straight-line relationship*.

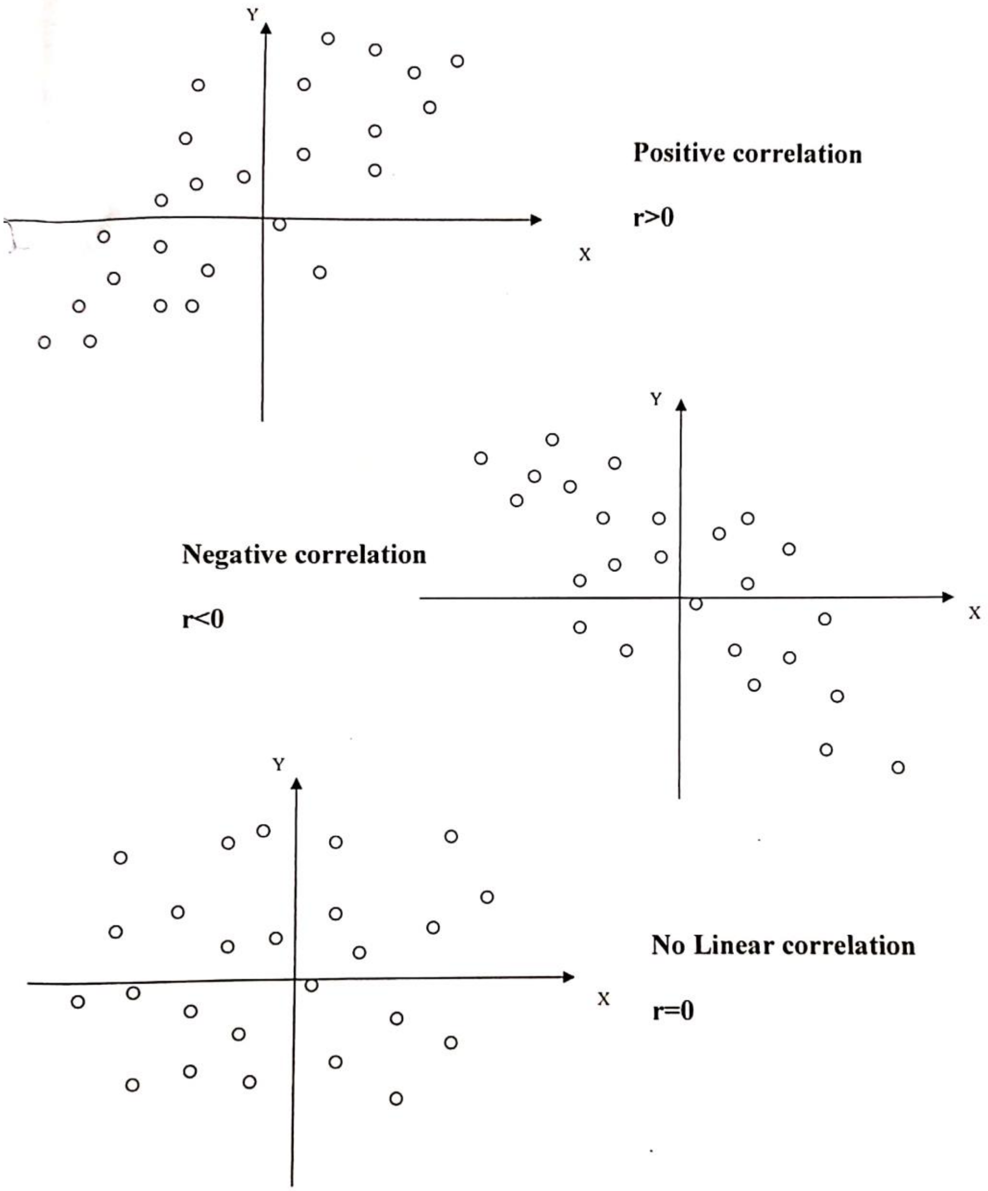
- The ***strength*** of the relationship is determined by the *closeness of the points to a straight line.*
- The ***direction*** is determined by whether one variable generally increases or generally decreases when the other variable increases.
- ***r*** is always between –1 and +1
- **magnitude** indicates the strength
- ***r* = –1 or +1** indicates a perfect linear relationship
- **sign** indicates the direction
- ***r* = 0** indicates no linear relationship

Activity: Among all elementary school children, the relationship between the number of cavities in a child’s teeth and the size of his or her vocabulary is strong and positive.

Activity: Consumption of hot chocolate is negatively correlated with crime rate. Both are responses to cold weather.

8.2 Observation Cloud

Let us consider the data of on two interdependent variables namely X and Y.



The following data were collected to study the relationship between the sale price, y and the total appraised value, x, of a residential property located in an upscale neighborhood.

Property	X	y	x^2	y^2	Xy
1	2	2	4	4	4
2	3	5	9	25	15
3	4	7	16	49	28
4	5	10	25	100	50
5	6	11	36	121	66
Σ(Sum)	20	35	90	299	163
	$\uparrow$ $\sum x$	$\uparrow$ $\sum y$	$\uparrow$ $\sum x^2$	$\uparrow$ $\sum y^2$	$\uparrow$ $\sum xy$

Pearson correlation coefficient, r.

With n=5

$$r=\frac{n\sum xy-(\sum x)(\sum y)}{\sqrt{n(\sum x^2)-(\sum x)^2}\sqrt{n(\sum y^2)-(\sum y)^2}}$$

$$r=\frac{5X163-20X35}{\sqrt{5(90)-(20)2}\sqrt{5(299)-(35)2}} = \frac{815-700}{\sqrt{450)-(400)}\sqrt{1495-(1225)}} = \frac{115}{\sqrt{50}\sqrt{270}} =$$

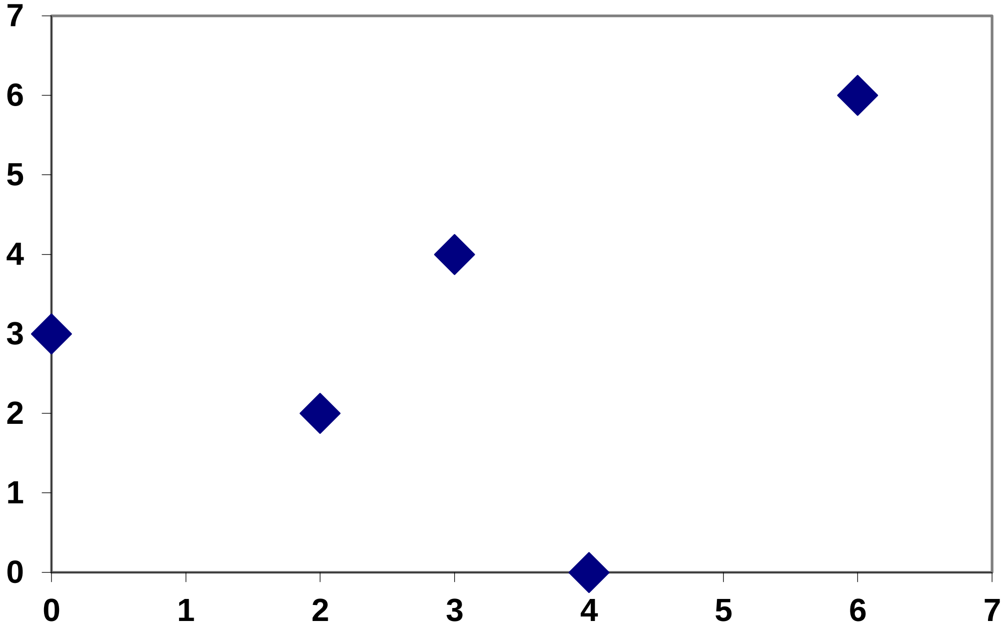
$$\frac{115}{7.071x16.432}$$

r= $\frac{115}{7.071x16.432} = \frac{115}{116.174} =$ **0.98,** X and Y are strongly Positively correlated.

Association Does Not Imply Causation

8.3 Scatter Diagram

Let us consider the scatter diagram of X and Y.



x	y	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
0	3	-3	0	0
2	2	-1	-1	1
3	4	0	1	0
4	0	1	-3	-3
6	6	3	3	9
15	15	0	0	
$\bar{x} = 3$	$\bar{y} = 3$	0	0	$\sum = 7$

$$\operatorname{cov}(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}))}{n} = \frac{7}{5} = 1.4$$

But what does this number tell us?

Nothing, So we can only compare covariances between different variables to see which is greater. Really, as

$$-\infty \leq \operatorname{cov}(x, y) \leq \infty$$

Or, we could standardize this measure, thus obtaining a more intuitive measure of correlation magnitude.

Correlation: Pearson's r

Standardize by adding the standard deviations to the equation:

$$\operatorname{cov}(x, y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n} \rightarrow r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n s_x s_y}$$

Where S_x= Standard Deviation of X and S_y= Standard deviation of Y

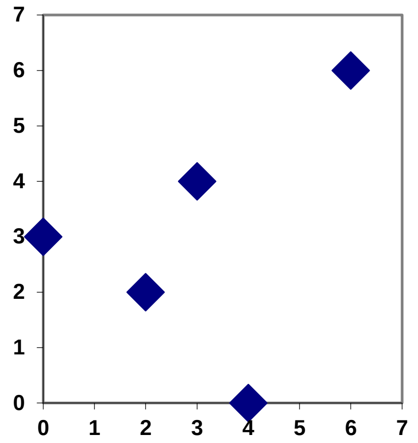
$$r_{xy} = \frac{\operatorname{cov}(x, y)}{s_x s_y}$$

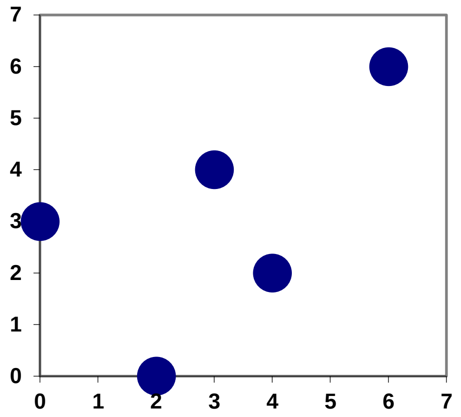
$-1 \leq r \leq 1$ The distance of r from 0 indicates strength of correlation r = 1 or r = (-1) means that we can predict y from x and vice versa with certainty; all data points are on a straight line. i.e., y = ax + b

$$r_{xy} = \frac{\sum_{i=1}^{n} Z_{x_i} Z_{y_i}}{n}$$

Important: each x_i goes with a specific y_i Why?

Example: By changing just two points of Y variable the correlation result is different…





x	y	Z_x	Z_y	$Z_{x*}Z_y$
0	3	-1.5	0	0
2	**2**	**-0.5**	**-0.5**	**0.25**
3	4	0	0.5	0
4	**0**	**0.5**	**-1.5**	**-0.75**
6	6	1.5	1.5	2.25
$\bar{x}=3$ $s_x=2$	$\bar{y}=3$ $s_y=2$			$\sum ZxZy$ =1.75

x	y	Z_x	Z_y	$Z_{x*}Z_y$
0	3	-1.5	0	0
2	**0**	**-0.5**	**-1.5**	**0.75**
3	4	0	0.5	0
4	**2**	**0.5**	**-0.5**	**-0.25**
6	6	1.5	1.5	2.25
$\bar{x}=3$ $s_x=2$	$\bar{y}=3$ $s_y=2$			$\sum ZxZy$ =2.75

$$r_{xy} = \frac{\sum_{i=1}^{n} Z_{x_i} * Z_{y_i}}{n} = \frac{1.75}{5} = 0.35$$

$$r_{xy} = \frac{\sum_{i=1}^{n} Z_{x_i} * Z_{y_i}}{n} = \frac{2.75}{5} = 0.55$$

A limitation of r: it is very sensitive to extreme values.

Example: Calculate the correlation between X and Y

X	Y
1	1
2	2
3	3
4	4
5	0

The correlation seems strong – but if we calculate it we'll find that r = 0

Note: r is actually $\hat{r}$.

So when r = 1 or r = (-1) we have a perfect linear relationship: y = ax + b
r=+1 (Perfect Positive correlation), = -1 (perfect negative correlation), r=0 (No Linear Correlation)

8.4 Regression

First recorded in 1510–20, regression is from the Latin word regression- (stem of regression).

What is regression analysis?

Umbrella selling company offers this example scenario: Suppose you're a sales manager trying to predict next month's numbers. You know that dozens, perhaps even hundreds of factors from the weather to a competitor's promotion to the rumor of a new and improved model can impact the number. Perhaps people in your organization even have a theory about what will have the biggest effect on sales. "Trust me. The more rain we have, the more we sell." "Six weeks after the competitor's promotion, sales jump."

Regression analysis is a way of mathematically sorting out which of those variables does indeed have an impact. It answers the questions: Which factors matter most? Which can we ignore? How do those factors interact with each other? And, perhaps most importantly, how certain are we about all of these factors?

We have seen how to explore the relationship between two quantitative variables graphically with a scatterplot. When the relationship has a straight-line pattern, the Pearson correlation coefficient describes it numerically. We can analyze the data further by finding an equation for the straight line that best describes the pattern. This equation predicts the value of the response(y) variable from the value of the explanatory variable.

Much of mathematics is devoted to studying variables that are deterministically related. Saying that x and y are related in this manner means that once we are told the value of x, the value of y is completely specified. For example, suppose the cost for a small pizza at a restaurant if Rs.100/- plus Rs.75 per topping. If we let x= # toppings and y = price of pizza, then y=100+75x. If we order a 3-topping pizza, then y=100+75(3)=325

There are two variables x and y which are appear to be related to one another, but not in a deterministic fashion. Suppose we examine the relationship between x=high school GPA and Y=college GPA. The value of y cannot be determined just from knowledge of x, and two different students could have the same x value but have very different y values. Yet there is a tendency for those students who have high (low) high school GPAs also to have high(low) college GPAs. Knowledge of a student's high school GPA should be quite helpful in enabling us to predict how that person will do in college.

Regression analysis is the part of statistics that deals with investigation of the relationship between two or more variables related in a nondeterministic fashion.

The statistical use of the word regression dates back to Francis Galton, who studied heredity in the late 1800's. One of Galton's interests was whether or not a man's height as an adult could be predicted by his parents' heights. He discovered that it could, but the relationship was such that very tall parents tended to have children who were shorter than they were, and very short parents tended to have children taller than themselves. He initially described this phenomenon by saying that there was a "reversion to mediocrity" but later changed to the terminology "regression to mediocrity".

The least-squares line is the line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible.

Simple Linear regression model equation for Least Squares (Regression) Line

$$Y=\beta_0 + \beta X + \in$$

When talking about regression equations, the following are terms used for X and Y
X: predictor variable, explanatory variable, or independent variable

Y: response variable or dependent variable

And the Estimated Line $\hat{y} = \hat{\beta}_o + \hat{\beta}_1 x$

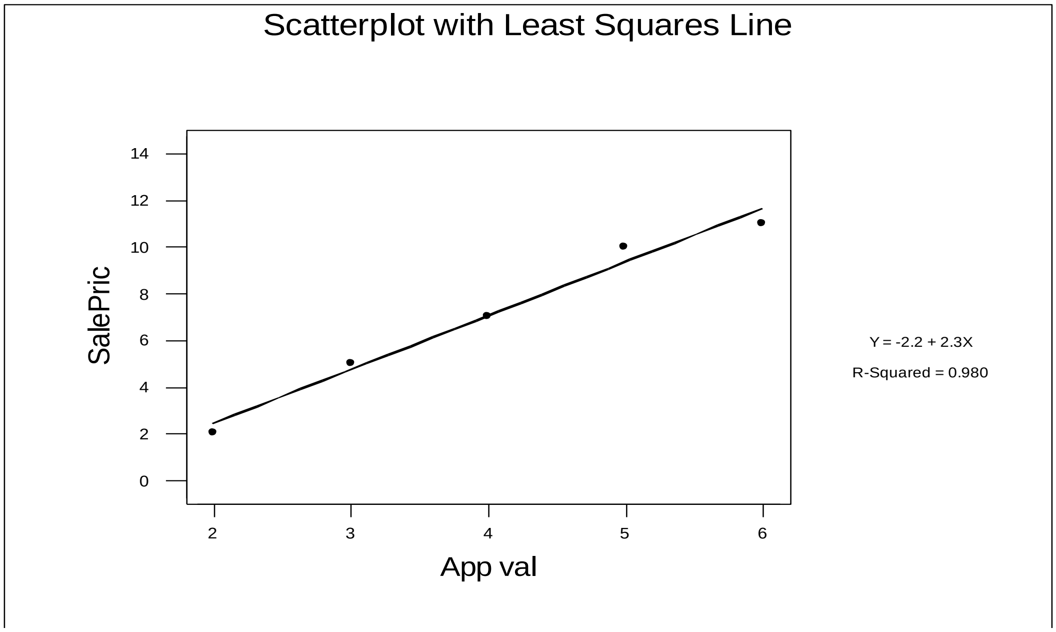
$\hat{\beta}_1$ denotes the estimated slope. The slope in the equation equals the amount that $\hat{y}$ changes when x increases by one unit.

$$\hat{\beta}_1 = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$\hat{\beta}_0$ denotes the estimated y-intercept. The y-intercept is the predicted value of y when x=0. The y-intercept may not have any interpretive value. If the answer to either of the two questions below is no, we do not interpret the y-intercept.

1. Is $\hat{\beta}_0$ a reasonable value for the explanatory variable?
2. Do any observations near x=0 exist in the data set?

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$



Equation for Least Squares Line : $\hat{y}$ = -2.2 + 2.3x

Appraisal Value, x $100,000	Sale Price, y $100,000	$\hat{y}$	$(y - \hat{y})$	$(y - \hat{y})^2$
2	2	2.4	-.4	.16
3	5	4.7	.3	.09
4	7	7	0	0
5	10	9.3	.7	.49
6	11	11.6	-.6	.36

$\mathbf{\Sigma (y - \hat{y})^2 = 1.1}$

The method of least squares chooses the prediction line $\hat{y} = \hat{B}_o + \hat{B}_1 x$ that minimizes the sum of the squared errors of prediction $\Sigma (y - \hat{y})^2$ for all sample points.

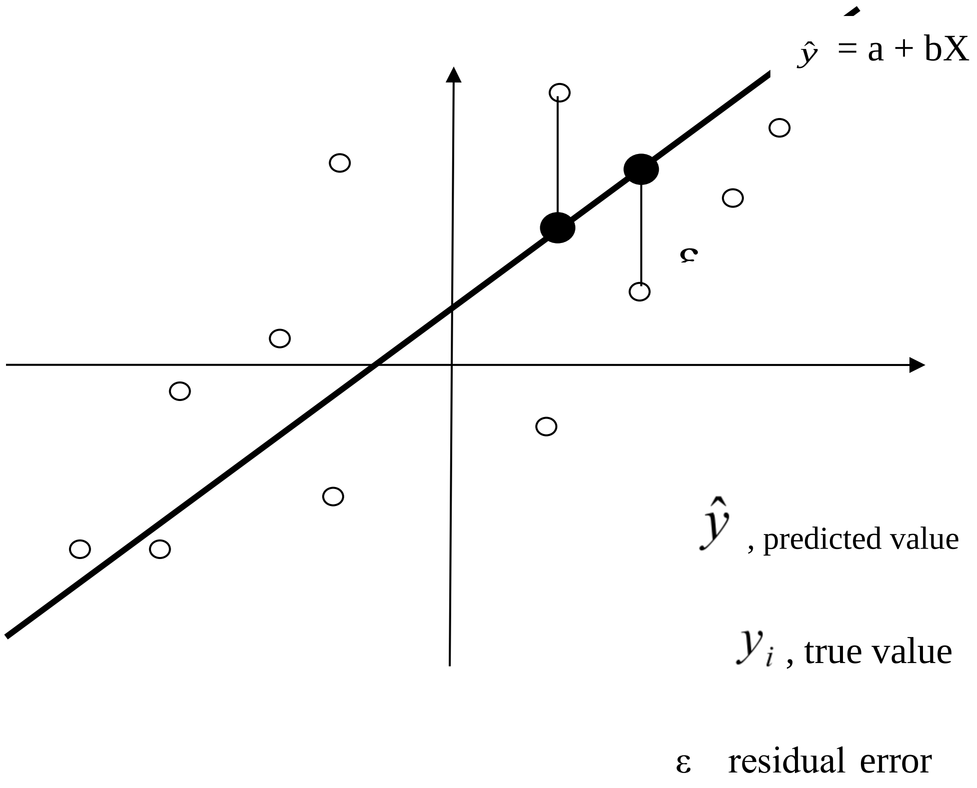
The estimated regression line from the is given by $\hat{y}$ = -2.2 + 2.3x

The slope in the equation equals the 2.3 that $\hat{y}$ changes when x i.e. price increases by one unit.

$\hat{\beta}_0$ denotes the estimated y-intercept. The y-intercept is the predicted value of y when x=0. i.e.

$\hat{y}$ = -2.2 i.e. on average sale price is -2.2 when appraisal value is zero.

Regression



Reference. Introduction to Statistical Theory Part-I page 398

The least squares principle:

$$\frac{\sum_{i=1}^{n}(y_i-\hat{y})^2}{n} \to \min$$

Example

From the data we calculate the following:

Σxy=150605 S_x=19.3679 , ΣY/n=66.93 and   ΣX/n=144.6. Run a Regression Y (height of anatomical dead space) on X (range of measurements).

Solution:

Applying these figures to the formulae for the regression coefficients, we have:

$$\hat{\beta}_1 = \frac{\sum xy - n(\overline{X}\overline{Y})}{(n-1)S^2{}_x} = \frac{150605 - 15X\,66.93X144.6}{14X(19.3679)^2}$$

$$\hat{\beta}_1 = \frac{150605 - 145171.17}{5251.6177} = \frac{5433.83}{5251.6177} = 1.0347$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

$$\hat{\beta}_0 = 66.93 - 1.0347\,X144.6$$

=66.93 – 149.6176= -82.687

Therefore, in this case, the equation for the regression of y on x becomes

$\hat{y}$ = -82.687 + 1.0367 x

This means that, on average, for every increase in height of 1 cm the increase in anatomical dead space is 1.067 ml over the range of measurements made.

The line representing the equation is shown superimposed on the scatter diagram of the data in figure. The way to draw the line is to take three values of x, one on the left side of the scatter diagram, one in the middle and one on the right, and substitute these in the equation, as follows:

If x = 110, y = (1.0367 x 110) – 82.687 = 31.35, and if x = 140, y = (1.033 x 140) – 82.4 = 62.45

If x = 170, y = (1.033 x 170) – 82.4 = 93.55

Although two points are enough to define the line, three are better as a check. Having put them on a scatter diagram, we simply draw the line through them. $\hat{y}$ = a + bx   This is true for a *sample*.

Like in all statistical methods, we want to make inferences about the *population.* So,

$$y_i = a + bx_i + \varepsilon_i$$

Then Estimated Equation is

$$\hat{y}_i = \hat{a} + \hat{b}x_i$$

Obviously, the stronger the correlation between x and y, the better the prediction; this is expressed in both parameters:

$$\hat{b} = \frac{\hat{r}s_y}{s_x} \qquad \hat{a} = \overline{y} - \frac{\hat{r}s_y}{s_x}\overline{x}$$

by putting values of a and b

$$\hat{y}_i = \hat{a} + \hat{b}x_i = \boxed{\overline{y} - \frac{\hat{r}s_y}{s_x}\overline{x}} + \boxed{\frac{\hat{r}s_y}{s_x}} x_i$$

After rearranging, we can write this:

$$\hat{y}_i = \hat{a} + \hat{b}x_i = \frac{\hat{r}s_y}{s_x}x_i - \frac{\hat{r}s_y}{s_x}\overline{x} + \overline{y}$$

$$\hat{y}_i = \frac{\hat{r}s_y}{s_x}(x_i - \overline{x}) + \overline{y}$$

It's easy to see why if there's no correlation, we will simply predict the average of y for any x. The larger the correlation, the greater the regression line's slope.
In any case, the average of the predicted values will always equal the average of

the true values: $\overline{\hat{y}} = \overline{y}$ (so $\hat{y}$ is an unbiased estimator of $\overline{y}$). The variance of the predicted values:

$$s_{\hat{y}}^2 = \frac{\sum(\hat{y}_i - \overline{y})^2}{n} = \dots\dots\dots\dots\dots\dots\dots = \hat{r}^2 s_y^2$$

So this variance is always smaller than the true variance (as the true variance is multiplied by a fraction).
Furthermore:

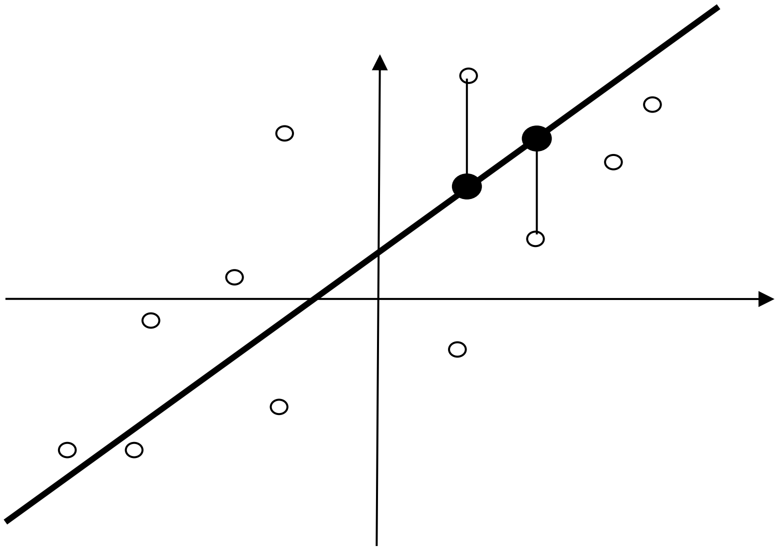
$$s_{\hat{y}}^2 = \hat{r}^2 s_y^2 \Longrightarrow \hat{r}^2 = \frac{s_{\hat{y}}^2}{s_y^2}$$

r-squared is the ***explained variance***!
It tells us what fraction of the general variance can be attributed to the model.
Therefore:
True variance = predicted variance + error variance

$$s_y^2 = s_{\hat{y}}^2 + s_{(y_i - \hat{y}_i)}^2$$

or:

$$s_y^2 = \hat{r}^2 s_y^2 + (1 - \hat{r}^2) s_y^2$$



Is the model significant?

(do we get a significantly better prediction using it than we do by just predicting the mean?)

This is where we see why it is similar to ANOVA*:

SS Total = **SS R**egression + **SS E**rror

$$\sum (y_i - \overline{y})^2 = \sum (\hat{y}_i - \overline{y})^2 + \sum (y_i - \hat{y})^2$$

In a one-way ANOVA, we have

SS Total = **SS B**etween + **SS W**ithin

$$\sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \overline{\overline{y}})^2 = \sum_{j=1}^{k} n_j (\overline{y}_j - \overline{\overline{y}})^2 + \sum_{j=1}^{k} \sum_{i=1}^{n_i} (y_{ij} - \overline{y}_j)^2$$

From the **SS** we can derive **MS** – dividing each SS by it's **degrees of freedom**:
MS Regression = **SS R**egression / 1 and **MS E**rror = **SS E**rror / (n-2)

Statistical significance test:

$$F_{(df\,\mathrm{mod}el,dferror)} = \frac{MS\,\mathrm{Re}\,g}{MSErr} = \ldots = \frac{\hat{r}^2 (N-2)^2}{1-\hat{r}^2}$$

Alternatively (as F is the square of t):

$$t_{(n-2)} = \frac{\hat{r}(n-2)}{\sqrt{1-\hat{r}^2}}$$

Assumptions

- Normal distributions, Constant variances, Independent sampling – no autocorrelations
- $\varepsilon \sim N(0,\sigma^2)$, No errors in the values of the independent variable
- All causation in the model is one-way (not necessary mathematically, but essential for prediction)

The regression model:

$$y_i = a + bx_i + \varepsilon_i$$

The regression model in GLM terms:

$$y_i = \mu_y + \beta x_i + \varepsilon_i$$

So:

$$y_1 = \beta x_1 + \mu_y * 1 + \varepsilon_1$$
$$y_2 = \beta x_2 + \mu_y * 1 + \varepsilon_2$$
$$y_3 = \beta x_3 + \mu_y * 1 + \varepsilon_3$$

And in matrix notation:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ x_3 & 1 \end{bmatrix} \begin{bmatrix} \beta \\ \mu_y \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{bmatrix}$$

In matrix Form in general

$$\underline{\boldsymbol{Y}} = \underline{\boldsymbol{X}}\,\underline{\beta} + \underline{\varepsilon}$$

Extrapolation is the use of the least-squares line for prediction outside the range of values of the explanatory variable x that you used to obtain the line. Extrapolation should not be done!

When the correlation coefficient indicates no linear relation between the explanatory and response variables, and the scatterplot indicates no relation at all between the variables, then we use the mean value of the response variable as the predicted value so that $\hat{y} = \bar{y}$.

8.5 Measuring the Contribution of x in Predicting y

We can consider how much the errors of prediction of y were reduced by using the information provided by x.

R^2 (Coefficient of Determination) = $\dfrac{\sum(\mathrm{y}-\bar{\mathrm{y}})^2 \;-\; \sum(\mathrm{y}-\hat{\mathrm{y}})^2}{\sum(\mathrm{y}-\bar{\mathrm{y}})^2}$

The coefficient of determination can also be obtained by squaring the Pearson correlation coefficient. This method works only for the linear regression model $\hat{y} = \hat{\beta}_o + \hat{\beta}_1 x$. The method does not work in general.

The coefficient of determination, r^2, represents the proportion of the total sample variation in y (measured by the sum of squares of deviations of the sample y values about their mean $\bar{y}$) that is explained by (or attributed to) the linear relationship between x and y.

Appraisal Value, x $100,000	Sale Price, y $100,000 Y	$\hat{y}$	$y-\hat{y}$	$(y-\hat{y})^2$	$(y-\bar{y})^2$
2	2	2.4	-0.4	0.16	25
3	5	4.7	0.3	0.09	4
4	7	7	0.0	0.00	0
5	10	9.3	0.7	0.49	9
6	11	11.6	-0.6	0.36	16
		Total		1.1	54

R^2(Coefficient of Determination)= $\dfrac{\sum(\mathrm{y}-\bar{\mathrm{y}})^2 \ - \ \sum(\mathrm{y}-\hat{\mathrm{y}})^2}{\sum(\mathrm{y}-\bar{\mathrm{y}})^2} = \dfrac{54-1.1}{54} = 0.98$

Interpretation: 98% of the total sample variation in y is explained by the straight-line relationship between y and x, with the total sample variation in y being measured by the sum of squares of deviations of the sample y values about their mean $\bar{y}$.

Interpretation: An R^2 of 0.98 means that the sum of squares of deviations of the y values about their predicted values has been reduced 98% by the use of the least squares equation $\hat{y}$ = -2.2 + 2.3x, instead of $\bar{y}$, to predict y.

The coefficient of determination is a number between 0 and 1, inclusive. That is, $0 \le r^2 \le 1$ If $r^2 = 0$, the least squares regression line has no explanatory value. If $r^2 = 1$, the least-squares regression line explains 100% of the variation in the response variable.

8.6 SELF ASSESSMENT QUESTIONS

Q.1: The grades of a class of 9 students on a midterm report (x) and on the final examination (y) are as follows:

(a) Calculate Correlation Coefficient between Midterm and Final examination. (b) Estimate the linear regression line. (c) Estimate the final examination grade of a student who received a grade of 85 on the midterm report.

Q.2: The amounts of a chemical compound y that dissolved in 100 grams of water at various temperatures x were recorded as follows:

X (^{0}C)						
(Grams)		, 14	, 24	, 28	, 42	, 44

a) Find the equation of the regression line. (b) Graph the line on a scatter diagram. (c) Estimate the amount of chemical that will dissolve in 100 grams of water at 50◦C.

Q.3: The following data were collected to determine the relationship between pressure and the corresponding scale reading for the purpose of calibration.

Pressure (Lb/Sq.In)										
Reading										

(a) Find the equation of the regression line.

(b) Find the Correlation coefficient between pressure and readings

(c) Draw a scatter Diagram of readings and pressure

Q.4: A study was made on the amount of converted sugar in a certain process at various temperatures. The data were coded and recorded as follows:

Temperature (X)											
Converted Sugar (Y)											

(a) Estimate the linear regression line. (b) Estimate the mean amount of converted sugar produced when the coded temperature is 1.75. (c) Plot the residuals versus temperature. (d) Find correlation coefficient e) Draw a scatter diagram.

SUGGESTED READINGS

Bluman, A.G. (2004). Elementary Statistics. A Step by Step Approach. 5th Edition. McGraw-Hill Companies Incorporated. London.

Chaudhary, S.M. & Kamal, S. (2017). Introduction to Statistical Theory Part-I. 8th Edition. Ilmi Kitab Khana. Lahore.

Chaudhary, S.M. & Kamal, S. (2017). Introduction to Statistical Theory Part-II. 8th Edition. Ilmi Kitab Khana. Lahore.

Daniel, W.W. (1995). Biostatistics: A foundation for Analysis in Health sciences. Sixth Edition. John Wiley and sons Incorporated. USA.

Harper, W.M. (1991). Statistics. Sixth Edition. Pitman Publishing, Longman Group, United Kingdom.

Hoel, P.G. (1976). Elementary Statistics. 4th Edition. John Wiley and Sons Incorporated, NewYork.

Kiani, G. H., & Akhtar, M. S. (2012). Basic statistics, Majeed Book Depot.

Khan, A. A., Mirza, S. H., Ahmad, M. I., Baig, I. & Yaqoob, M. (2011). Business Statistics, Qureshi Brothers Publishers.

UNIT 09

TIME SERIES ANALYSIS

Written By: Dr. Zahid Iqbal
Reviewed By: Dr. Muhammad Ilyas

CONTENTS

Pages

Introduction

When modeling relationships between variables, the nature of the data that have been collected has an important bearing on the appropriate choice of an econometric model. In particular, it is important to distinguish between cross-section data (data on a number of economic units at a particular point in time) and time-series data (data collected over time on one particular economic unit). Examples of both types of data. When we say ‘‘economic units’’ we could be referring to individuals, households, firms, geographical regions, countries, or some other entity on which data is collected. Because cross-section observations on a number of economic units at a given time are often generated by way of a random sample, they are typically uncorrelated. The level of income observed in the Smiths’ household, for example, does not affect, nor is it affected by, the level of income in the Jones’s household. On the other hand, time-series observations on a given economic unit, observed over a number of time periods, are likely to be correlated. The level of income observed in the Smiths’ household in one year is likely to be related to the level of income in the Smiths’ household in the year before. Thus, one feature that distinguishes time-series data from cross-section data is the likely correlation between different observations. Our challenges for this chapter include testing for and modeling such correlation. A second distinguishing feature of time-series data is its natural ordering according to time. With cross-section data there is no particular ordering of the observations that is better or more natural than another. One could shuffle the observations and then proceed with estimation without losing any information. If one shuffles time-series observations, there is a danger of confounding what is their most important distinguishing feature: the possible existence of dynamic relationships between variables. A dynamic relationship is one in which the change in a variable now has an impact on that same variable, or other variables, in one or more future time periods. For example, it is common for a change in the level of an explanatory variable to have behavioral implications for other variables beyond the time period in which it occurred. The consequences of economic decisions that result in changes in economic variables can last a long time. When the income tax rate is increased, consumers have less disposable income, reducing their expenditures on goods and services, which reduces profits of suppliers, which reduces the demand for productive inputs, which reduces the profits of the input suppliers, and so on. The effect of the tax increase ripples through the economy. These effects do not occur instantaneously but are spread, or distributed, over future time periods. As shown in Figure 9.1, economic actions or decisions taken at one point in time, t, have effects on the economy at time t, but also at times t + 1, t + 2, and so on.

Objectives

After studying this unit, you will be able to;

- Explain why lags are important in models that use time-series data, and the ways in which lags can be included in dynamic econometric models.
- Explain what is meant by a serially correlated time series, and how we measure serial correlation.
- Specify, estimate, and interpret the estimates from a finite distribute lag model.
- Explain the nature of regressions that involve lagged variables and the number of observations that are available.
- Specify and explain how the multiple regression assumptions are modified to accommodate time series data.
- Compute the autocorrelations for a time-series, graph the corresponding correlogram, and use it to test for serial correlation.

9.1 Dynamic Nature of Relationships

Given that the effects of changes in variables are not always instantaneous, we need to ask how to model the dynamic nature of relationships. We begin by recognizing three different ways of doing so.

One way is to specify that a dependent variable y is a function of current and past values of an explanatory variable x. That is,

$$y_t = f(x_t; x_{t-1}; x_{t-2}; \ldots\ldots\ldots\ldots\ldots\ldots\ldots) \qquad (9.1)$$

We can think of (y_t, x_t) as denoting the values for y and x in the current period; x_{t-1} means the value of x in the previous period; x_{t-2} is the value of x two periods ago, and so on. For the moment f (.) is used to denote any general function. Later we replace f (.) by a linear function. Equations such as (9.1) say, for example, that the current rate of inflation y_t depends not just on the current interest rate x_t, but also on the rates in previous time periods x_{t-1}, x_{t-2}, ... ….. Turning this interpretation around as in Figure 9.1, it means that a change in the interest rate now will have an impact on inflation now and in future periods; it takes time for the effect of an interest rate change to fully work its way through the economy. Because of the existence of these lagged effects, (9.1) is called a **distributed lag model**.

A second way of capturing the dynamic characteristics of time-series data is to specify a model with a **lagged dependent variable** as one of the explanatory variables. For example,

$$y_t=f(y_{t-1}; x_t) \qquad (9.2)$$

Where again f(.) is a general function that we later replace with a linear function. In this case we are saying that the inflation rate in one period y_t will depend (among other things) on what it was in the previous period, y_{t-1}. Assuming a positive relationship, periods of high inflation will tend to follow periods of high inflation and periods of low inflation will tend to follow periods of low inflation. Or, in other words, inflation is positively correlated with its value lagged one period. A model of this nature is one way of modeling correlation between current and past values of a dependent variable. Also, we can combine the features of (9.1) and (9.2) so that we have a dynamic model with lagged values of both the dependent and explanatory variables, such as

$$y_t= f(y_{t-1}; x_t; x_{t-1}; x_{t-2}) \qquad (9.3)$$

Such models are called **autoregressive distributed lag (ARDL)** models, with ‘‘autoregressive’’ meaning a regression of y_t on its own lag or lags.

A third way of modeling the continuing impact of change over several periods is via the error term. For example, using general functions f(.) and g(.), both of which are replaced later with linear functions, we can write

$$y_t = f(x_t) + e_t \qquad e_t = g(e_{t-1}) \qquad (9.4)$$

Where the function $e_t = g(e_{t-1})$ is used to denote the dependence of the error on its value in the previous period. In this case e_t is correlated with e_{t-1}; we say the errors are serially correlated or auto-correlated. Because (9.3) implies $e_{t+1} = g(et)$, the dynamic nature of this relationship is such that the impact of any unpredictable shock that feeds into the error term will be felt not just in period t, but also in future periods. The current error e_t affects not just the current value of the dependent variable y_t, but also its future values y_{t+1}; y_{t+2}; ... . As an example, suppose that a terrorist act creates fear of an oil shortage, driving up the price of oil. The terrorist act is an unpredictable shock that forms part of the error term e_t. It is likely to affect the price of oil in the future as well as during the current period.

We consider these three ways in which dynamics can enter a regression relationship—lagged values of the explanatory variable, lagged values of the dependent variable, and lagged values of the error term. What we discover is that these three ways are not as distinct as one might at first think. Including a lagged dependent variable y_{t1} can capture similar effects to those obtained by including a lagged error et1, or a long history of past values of an explanatory variable, x_{t1}; x_{t2}; …..... . Thus, we not only consider the three kinds of dynamic relationships, we explore the relationships between them. Related to the idea of modeling dynamic relationships between time series variables is the important concept of forecasting. We are not only interested in tracing the impact of a change in an explanatory variable or an error shock through time. Forecasting future values of economic time series, such as inflation, unemployment, and exchange rates, is something that attracts the attention of business, governments, and the general public. Describing how dynamic models can be used for forecasting is another objective.

9.2 Least Squares Assumptions

An important consequence of using time series data to estimate dynamic relationships is the possible violation of one of our least squares assumptions. Assumption, states that different observations on y and on e are uncorrelated. That is,

$$\mathrm{Cov}(y_i; y_j) = \mathrm{cov}(e_i; e_j) = 0 \text{ for } i \neq j$$

To emphasize that we are using time-series observations, we drop the i and j subscripts and use t and s instead, with t and s referring to two different time periods such as days, months, quarters, or years. Thus, the above assumption becomes

$$\mathrm{cov}(y_t; y_{s)} = \mathrm{cov}(e_t; e_{s)} = 0 \text{ for } t \neq s$$

The dynamic models in (9.2), (9.3) and (9.4) imply correlation between y_t and y_{t-1} or e_t and e_{t-1} or both, so they clearly violate assumption, that different observations on y and on e are uncorrelated. As mentioned below (9.4), when a variable is correlated with its past values, we say that it is autocorrelated or serially correlated. How to test for serial correlation, and its implications for estimation.

9.3 Stationarity

An assumption that we maintain throughout the time series is that the variables in our equations are stationary. This assumption will take on more meaning when it is relaxed. For the moment we note that a stationary variable is one that is not explosive, nor trending, and nor wandering aimlessly without returning to its mean. These features can be illustrated with some graphs. Plots of this kind are routinely considered when examining time-series variables. The variable Y that appears is considered stationary because it tends to fluctuate around a constant mean without wandering or trending. On the other hand, X and Z in possess characteristics of nonstationary variables. In X tends to wander, or is ‘‘slow turning,’’ while Z is trending. These concepts will be defined. For now the important thing to remember is that with modeling and estimating dynamic relationships between stationary variables whose time series have similar characteristics to those of Y. That is, they neither wander nor trend.

9.4 Alternative Paths

This starting point has the advantage of beginning with a model that is closest to those studied so far. From there we recommend covering serial correlation—relevant definitions, concepts, and testing. At this point some instructors might like to proceed with the AR(1) error model; others might prefer to jump straight to ARDL models. The second path is designed for instructors who wish to start the chapter with serial correlation. After covering definitions, concepts, and testing, they can proceed to the AR(1) error model or straight to ARDL models. Finite distributed lag models can be covered as a special case of ARDL models or omitted.

Finite Distributed Lags The first dynamic relationship that we consider is that given in (9.1),

$y_t = f(x_t; x_{t-1}; x_{t-2}; \ldots\ldots.)$, with the additional assumptions that the relationship is linear, and, after q time periods, changes in x no longer have an impact on y. Under these conditions we have the multiple regression model

$$y_t = \alpha + \beta_0 x_t + \beta_1 x_{t-1} + \beta_2 x_{t-2} \ldots\ldots\ldots \beta_q x_{t-q} + e_t \qquad (9.5)$$

The model in (9.5) can be treated in the same way as the multiple regression model. Instead of having a number of explanatory variables, we have a number of different lags of the same explanatory variable. However, for the purpose of estimation, these different lags can be treated in the same way as different explanatory variables. It is convenient to change subscript notation on the coefficients: bs is used to denote the coefficient of xts and a is introduced to denote the intercept. Other explanatory variables can be added if relevant, in which case other symbols are needed to denote their coefficients. Models such as (9.5) have two special uses. The first is forecasting future values of y. To introduce notation for future values, suppose our sample period is for t = 1, 2, ... , T. We use t for the index (rather than i) and T for the sample size (rather than N) to emphasize the time series nature of the data. Given that the last observation in our sample is at t = T, the first post sample observation that we want to forecast is at t = T + 1. The equation for this observation is given by

$$y_{T+1} = \alpha + \beta_0 x_{T+1} + \beta_1 x_T + \beta_2 x_{T+1} + \ldots\ldots\ldots.. + \beta q x_{Tq+1} + e_{T+1} \qquad (9.6)$$

The forecasting problem is how to use the time series of x-values, x_{T+1}; x_T ; x_{T-1}; ... ; x_{T-q+1} to forecast the value y_{T+1}, with special attention needed to obtain a value for x_{T+1}.

The second special use of models like (9.5) is for policy analysis. Examples of policy analysis where the distributed-lag effect is important are the effects of changes in government expenditure or taxation on unemployment and inflation (fiscal policy), the effects of changes in the interest rate on unemployment and inflation (monetary policy), and the effect of advertising on sales of a firm’s products. The timing of the effect of a change in the interest rate or a change in taxation on unemployment, inflation, and the general health of the economy can be critical. Suppose the government (or a firm or business) controls the values of x, and would like to set x to achieve a given value, or a given sequence of values, for y. The coefficient βs gives the change in $E(y_t)$ when x_{t-s} changes by one unit, but x is held constant in other periods. Alternatively, if we look forward instead of backward, βs gives the change in $E(y_{t+s})$ when x_t changes by one unit, but x in other periods is held constant. In terms of derivatives

$$a(yt)/a(x_{t-s}\ )= a(y_{t+s})/\ a(x_t) = \beta s \qquad (9.7)$$

To further appreciate this interpretation, suppose that x and y have been constant for at least the last q periods and that xt is increased by one unit, then returned to its original level. Then, using (9.5) but ignoring the error term, the immediate effect will be an increase in y_t by β0 units. One period later, y_{t+1} will increase by β_1 units, then y_{t+2} will increase by β_2 units and so on, up to period t + q, when y_{t+q} will increase by β q units. In period t+ q + 1 the value of y will return to its original level. The effect of a one-unit change in xt is distributed over the current and next q periods, from which we get the term ''**distributed lag model.**'' It is called a finite distributed lag model of order q because it is assumed that after a finite number of periods q, changes in x no longer have an impact on y. The coefficient β s is called a distributed-lag weight or an s-period delay multiplier. The coefficient β_0 (s = 0) is called the impact multiplier. It is also relevant to ask what happens if xt is increased by one unit and then maintained at its new level in subsequent periods (t + 1), (t þ+2), ... In this case, the immediate impact will again be β_0; the total effect in period t +1 will be $\beta_0 + \beta_1$, in period t+2 it will be $\beta_0 + \beta_1 + \beta_2$, and so on. We add together the effects from the changes in all preceding periods. These quantities are called **interim multipliers**. For example, the two-period **interim multiplier** is $\beta_0 + \beta_1 + \beta_2$. The **total multiplier** is the final effect on y of the sustained increase after q or more periods have elapsed; it is given by $\sum_{s=0}^{q} \boldsymbol{\beta}$s.

9.5 Assumptions

When the simple regression model was first introduced in Chapter 8, it was written in terms of the mean of y conditional on x. Specifically, E(y/x)= $\beta_1 + \beta_2$X, which led to the error term assumption E(e/x)= 0. Then, so that we could avoid the need to condition on x, and hence ease the notational burden, we made the simplifying assumption that the x's are not random. We maintained this assumption through Chapters 8, recognizing that although it is unrealistic for most data sets, relaxing it in a limited but realistic way would have had little impact on our results and on our choice of estimators and test statistics. However, because the time-series variables used in the examples in this chapter are random, it is useful to mention alternative assumptions under which we can consider the properties of least squares and other estimators. In distributed lag models both y and x are typically random. The variables used in the example that follows are unemployment and output growth. They are both random. They are observed at the same time; we do not know their values prior to ‘‘sampling.’’ We do not ‘‘set’’ output growth and then observe the resulting level of unemployment. To accommodate this randomness we assume that the x’s are random and that et is independent of all x’s in the sample—past, current, and future. This assumption, in conjunction with the other multiple regression assumptions, is sufficient for the least squares estimator to be unbiased and to be best linear unbiased conditional on the x’s in the sample. With the added assumption of normally distributed error terms, our usual t and F tests have finite sample justification. Accordingly, the multiple regression assumptions given can be modified for the distributed lag model as follows:

Assumptions of The Distributed Lag Model

Time Series Assumption1. $y_t = \alpha + \beta_0 x_t + \beta_1 x_{t-1} + \beta_2 x_{t-2} \ldots\ldots\ldots \beta_q x_{t-q} + e_t$
$t = q+1; \ldots ; T$
Time Series Assumption 2. y and x are stationary random variables, and et is independent of current, past and future values of x.
Time Series Assumption3. E(et) = 0
Time Series Assumption 4. Var(et) = σ^2
Time Series Assumption 5. Cov(et; es) = 0 $t \neq s$
Time series Assumption 6. et $\sim$ N(0; σ^2)
The least squares-estimated Phillips curve

$$INFt = \beta_1 + \beta_2 DUt + e_t$$

with both sets of standard errors—the incorrect least squares ones that ignore autocorrelation, and the correct HAC ones that recognize the autocorrelation—are as follows:

$$\widehat{\boldsymbol{INF}}_t = 0.7776 - 0.5279DU_t$$
$$(0.065) \quad (0.2294) \quad \text{incorrect standard error}$$
$$(0.1030)\ (0.3127)\ \text{HAC standard error.}$$

The HAC standard errors are larger than those from least squares, implying that if we ignore the autocorrelation, we will overstate the reliability of the least squares estimates. The t and p-values for testing H0 : $\beta_2 = 0$ are

t = - 0.5279/0:2294 = - 2:301, p = 0.0238 (from LS standard errors)

$t = - 0.5279/0.3127 = -1{:}688$, $p = 0.0950$ (from HAC standard errors)

An autoregressive distributed lag (ARDL) model is one that contains both lagged xt's and lagged yt's. In its general form, with p lags of y and q lags of x, an ARDL(p, q) model can be written as

$$yt = \alpha_0 + \beta_1 yt\text{-}1 + \beta_2 y_{t-2} + \ldots\ldots\ldots + \beta_p y_{t-p} + \gamma_0 xt + \gamma_1 x_{t-1} + \ldots\ldots\ldots + \gamma_q xtq + vt$$

The AR component of the name ARDL comes from the regression of y on lagged values of itself; the DL component comes from the distributed lag effect of the lagged x's. Two examples that we) are

ARDL(1,1): $\widehat{\boldsymbol{INF}}_t = 0.3336 + 0.5593INFt\text{-}1 – 0.6882DUt + 0{:}3200DUt\text{-}1$

ARDL(1,0): $\widehat{\boldsymbol{INF}}_t = 0.3548 + 0.5282INFt\text{-}1 – 0.4909DUt$

The ARDL model has several advantages. It captures dynamic effects from lagged x's and lagged y's, and by including a sufficient number of lags of y and x, we can eliminate serial correlation in the errors.

9.6 SELF ASSESSMENT QUESTIONS

Q.1 Consider the following distributed lag model relating the percentage growth in private investment (INVGWTH) to the federal funds rate of interest (FFRATE):

$\widehat{\boldsymbol{INVGWTH}}$t = 4 – 0.4FFRATEt – 0.8FFRATEt-1 – 0.6FFRATEt-2 - 0.2FFRATEt-3

(a) Suppose FFRATE = 1% for t ¼ 1, 2, 3, 4. Use the above equation to forecast INVGWTH for t =4.

(b) Suppose FFRATE is raised to 1.5% in period t = 5 and then returned to its original level of 1% for t =6, 7, 8, 9. Use the equation to forecast INVGWTH for periods t = 5, 6, 7, 8, 9. Relate the changes in your forecasts to the values of the coefficients. What are the delay multipliers?

(c) Suppose FFRATE is raised to 1.5% for periods t = 5, 6, 7, 8, 9. Use the equation to forecast INVGWTH for periods t = 5, 6, 7, 8, 9. Relate the changes in your forecasts to the values of the coefficients. What are the interim multipliers? What is the total multiplier?

Q.2 The contains 105 weekly observations on sales revenue (SALES) and advertising expenditure (ADV) in millions of Rupees for a large midwest department store in 2008 and 2009. The following relationship was estimated:

$\widehat{\boldsymbol{SALES}}$t =25.34 + 1.842 ADVt + 3.802 ADVt-1 + 2.265 ADVt-2

(a) Describe the relationship between sales and advertising expenditure. Include an explanation of the lagged relationship. When does advertising have its greatest impact? What is s the total effect of a sustained Rs.1 million increase in advertising expenditure?

(b) The estimated covariance matrix of the coefficients is

	C	ADV_t	ADV_{t-1}	ADV_{t-2}
C	2.5598	-0.7099	-0.1317	-0.7661
ADV_t	-0.7099	1.3964	-1.0406	0.0984
ADV_{t-1}	-0.1317	-1.0406	-2.1606	-1.067
ADV_{t-2}	-0.7661	0.0984	-1.067	-1.4214

Using a one-tail test and a 5% significance level, which lag coefficients are significantly different from zero? Do your conclusions change if you use a one tail test? Do they change if you use a 10% significance level?

(c) Find 95% confidence intervals for the impact multiplier, the one-period interim multiplier, and the total multiplier.

Q.3 Reconsider the estimated equation and covariance matrix in Question no 2. Suppose, as a marketing executive for the department store, that you have a total of Rs.6 million to spend on advertising over the next three weeks, t = 106, 107, and 108. Consider the following allocations of the Rs. 6 million:

Case No. 1. ADV106 = 6; ADV107 = 0; ADV108 = 0
Case No. 2. ADV106 = 0; ADV107 = 6; ADV108 = 0
Case No. 3. ADV106 = 2; ADV107 = 4; ADV108 = 0

(a) For each allocation of the Rs. 6 million, forecast sales revenue for t = 106, 107, and 108. Which allocation leads to the largest forecast for total sales revenue over the three weeks? Which allocation leads to the largest forecast for sales in week t = 108? Explain why these outcomes were obtained.

(b) Find 95% forecast intervals for ADV108 for each of the three allocations. If maximize ADV108 is your objective, which allocation would you choose? Why?

Q.4 In question no.1, the following Phillips curve was estimated:

$\widehat{\boldsymbol{INFt}}$ = 0.1001 + 0.2354 INFt-1 + 0.1213 INFt-2 + 0.1677 INFt-3 + 0:2819 INFt-4 – 0.7902 DUt

The last four sample values for inflation are INF_{2019Q3} = 1.0; INF_{2019Q2} = 0.5; INF_{2019Q1} = 0.1; and INF_{2018Q4} = 0.3. The unemployment rate in 2019Q3 was 5.8%. The estimated error variance for the above equation is $\hat{\boldsymbol{\sigma}}^2$ =0.225103.

(a) Given that the unemployment rates in the first three post-sample quarters are U2019Q4 = 5.6; U2020Q1= 5.4; and U2020Q2 = 5.0, use the estimated equation to forecast inflation for 2019Q4, 2020Q1 and 2020Q2.

(b) Find the standard errors of the forecast errors for your forecasts in (a).

(c) Find 95% forecast intervals for INF2019Q4; INF2020Q1; and INF2020Q2. How reliable are the forecasts you found in part (a)

SUGGESTED READINGS

Bluman, A.G. (2004). *Elementary Statistics. A Step by Step Approach*. 5th Edition. McGraw-Hill Companies Incorporated. London.

Chaudhary, S.M. & Kamal, S. (2017). *Introduction to Statistical Theory Part-I*. 8th Edition. Ilmi Kitab Khana. Lahore.

Chaudhary, S.M. & Kamal, S. (2017). *Introduction to Statistical Theory Part-II*. 8th Edition. Ilmi Kitab Khana. Lahore.

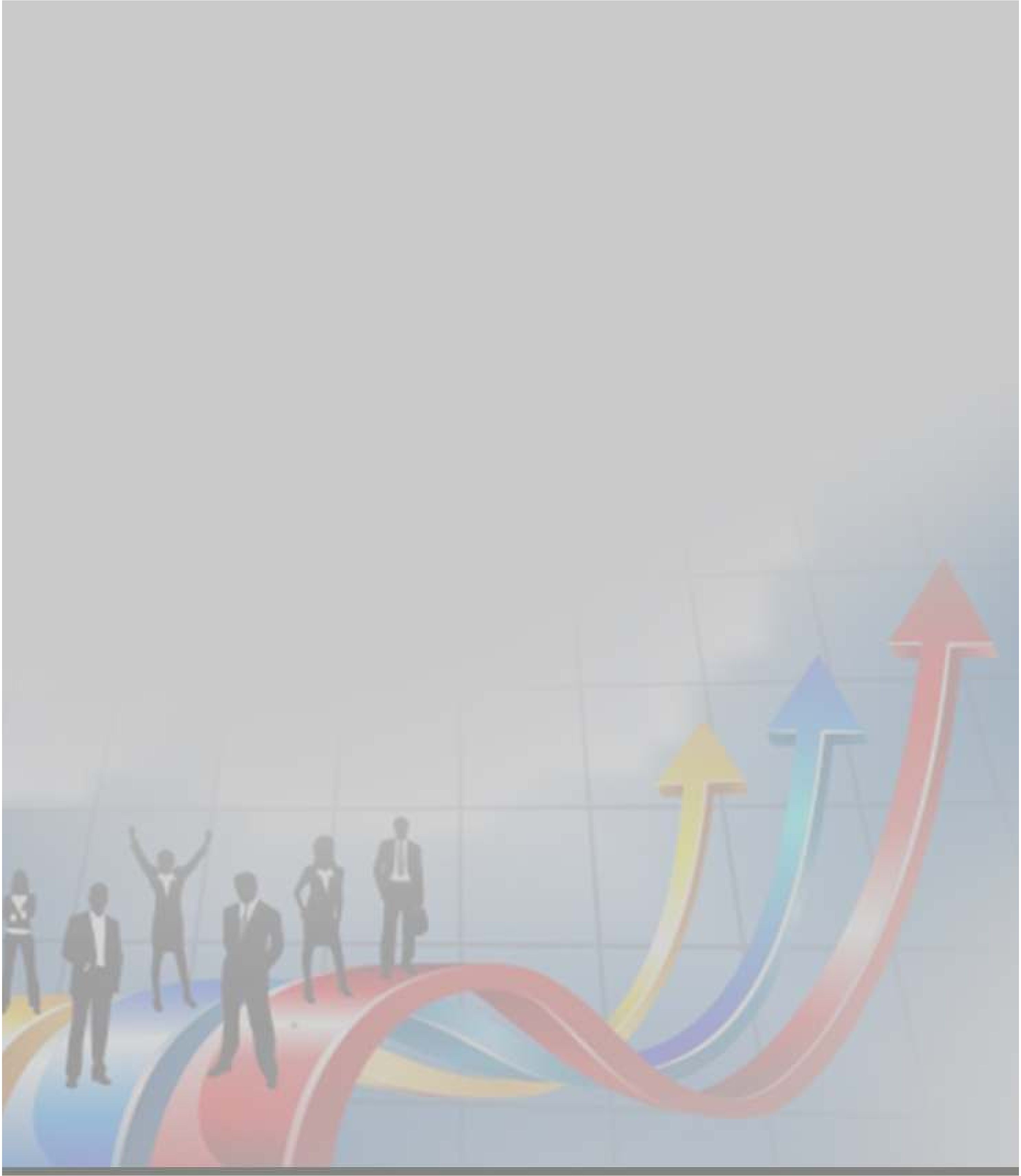
Daniel, W.W. (1995). *Biostatistics: A foundation for Analysis in Health sciences*. Sixth Edition. John Wiley and sons Incorporated. USA.

Harper, W.M. (1991). *Statistics.* Sixth Edition. Pitman Publishing, Longman Group, United Kingdom.

Hoel, P.G. (1976). *Elementary Statistics*. 4th Edition. John Wiley and Sons Incorporated, NewYork.

Kiani, G. H., & Akhtar, M. S. (2012). *Basic statistics*, Majeed Book Depot.

Khan, A. A., Mirza, S. H., Ahmad, M. I., Baig, I. & Yaqoob, M. (2011). *Business Statistics*, Qureshi Brothers Publishers.



INTRODUCTION TO STATISTICS
FOR ECONOMISTS